

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 EMOSTEER-TTS: FINE-GRAINED AND TRAINING-FREE EMOTION-CONTROLLABLE TEXT-TO-SPEECH VIA ACTIVATION STEERING

Anonymous authors

Paper under double-blind review

ABSTRACT

Text-to-speech (TTS) has shown great progress in recent years. However, most existing TTS systems offer only coarse and rigid emotion control, typically via discrete emotion labels or a carefully crafted and detailed emotional text prompt, making fine-grained emotion manipulation either inaccessible or unstable. These models also require extensive, high-quality datasets for training. To address these limitations, we propose **EmoSteer-TTS**, a novel **training-free** approach, to achieve **fine-grained** speech emotion control (conversion, interpolation, erasure) by **activation steering**. We first empirically observe that modifying a subset of the internal activations within a flow matching-based TTS model can effectively alter the emotional tone of synthesized speech. Building on this insight, we then develop a training-free and efficient algorithm, including activation extraction, emotional token searching, and inference-time steering, which can be seamlessly integrated into a wide range of pretrained models (e.g., F5-TTS, CosyVoice2, and E2-TTS). In addition, to derive effective steering vectors, we construct a curated emotional speech dataset with diverse speakers. Extensive experiments demonstrate that EmoSteer-TTS enables fine-grained, interpretable, and continuous control over speech emotion, outperforming the state-of-the-art (SOTA). To the best of our knowledge, this is the first method that achieves training-free and continuous fine-grained emotion control in TTS. Demo samples are available at <https://emosteer-tts-demo.pages.dev/>.

1 INTRODUCTION

Text-to-speech (TTS) aims to generate natural-sounding human speech from textual input (Tan et al., 2021; Xie et al., 2025). It has been widely adopted in various domains, including voice assistants, robotics, and podcast production. Emotion-controllable TTS (EC-TTS) enhances this capability by enabling control over the emotional tone of synthesized speech, making it more expressive and engaging. Fine-grained EC-TTS takes this further by allowing precise modulation of the conveyed emotion intensity in synthesized speech. Such detailed control is vital for applications requiring nuanced expressiveness, e.g., personalized storytelling (Rong et al., 2025), empathetic human-computer interaction (Wadley et al., 2022), and precise speech editing (Peng et al., 2024).

Controlling the emotional tone of synthesized speech typically requires the simultaneous manipulation of multiple characteristics, such as pitch, energy, and prosody. Independently adjusting any of these attributes often leads to undesirable artifacts. Therefore, in the literature, existing methods commonly adopt a conditional generation paradigm, including **label-based** methods that incorporate discrete emotion labels (Cho et al., 2025) and **description-based** methods that use textual emotion descriptions (Yang et al., 2025) as additional inputs to guide the speech synthesis process.

Label-based EC-TTS approaches use categorical labels (e.g., anger, happiness, fear) as an additional input to control the emotional expression during training and inference. For example, StyleTagging-TTS (Kim et al., 2021b) uses Sentence BERT (Reimers & Gurevych, 2019) to encode short phrases or keywords as emotion labels to guide the synthesis. However, such methods rely on fixed emotion labels, offering limited flexibility in control (Cong et al., 2025). Recent studies apply strength control to emotion labels. For instance, EmoSphere++ (Cho et al., 2025) converts discrete labels into the

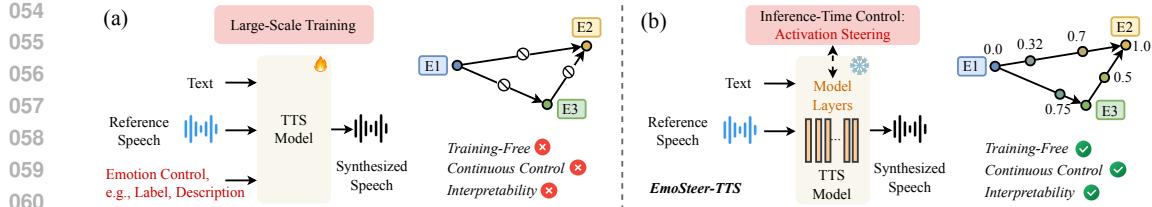


Figure 1: Motivations of our work. (a) Existing paradigm for speech emotion control. (b) EmoSteer-TTS offers training-free, fine-grained continuous emotion control with improved interpretability.

Valence-Arousal-Dominance (VAD) vector space (Mehrabian, 1980), where the origin represents a neutral state. Both the type and intensity of emotion can be controlled by adjusting the direction and magnitude of the emotional vector. However, these methods **rely on large emotion-labeled datasets** and often **struggle to generalize** to unseen reference speech (Inoue et al., 2025).

On the other hand, description-based EC-TTS methods use textual prompts, such as “*A girl says welcome in a happy tone*”, to describe the target emotion, guiding the TTS model to generate speech that aligns with the given description. For example, CosyVoice2 (Du et al., 2024) leverages textual prompts to control emotional expressiveness, enhanced via instruction fine-tuning. Similarly, EmoVoice (Yang et al., 2025) incorporates emotion descriptions into the text context to enable fine-grained emotion control. However, such methods (Guo et al., 2023; Shimizu et al., 2024; Ji et al., 2025; Li et al., 2023b) require large-scale datasets and carefully designed training procedures. Although these methods enable finer emotion manipulation, their **controllability is fundamentally limited** by the finite set of human language expressions, imposing an upper bound on control granularity. Moreover, they **exhibit instability** due to the inherent variability of textual descriptions and the stochastic nature of token sampling in the language models used for encoding.

In summary, existing methods have two limitations, i.e., **instability/poor generalization** and **coarse controllability**. The first arises from the lack of large-scale emotional speech datasets required for effective model training. The second stems from the control strategies employed in existing methods, which restrict the precision of emotion manipulation. Furthermore, the absence of exploration in emotion representations within TTS models poses challenges for researchers seeking to understand how speech emotions are encoded.

To address these limitations, we present **EmoSteer-TTS**, a training-free approach that enables fine-grained, continuous emotion control, as illustrated in Fig. 1. Specifically, we begin by analyzing the internal emotion representations of pretrained zero-shot TTS models, such as F5-TTS (Chen et al., 2025) and CosyVoice2. These models use a Diffusion Transformer (DiT) (Peebles & Xie, 2023) as the backbone and employ flow matching (Lipman et al., 2023) to generate high-fidelity mel-spectrograms. As shown in Fig. 2, we observe that only a subset of tokens, i.e., activations, within the model significantly influences the emotional tone of the synthesized speech. Building on this insight, we propose a simple yet effective algorithm to extract emotionally salient tokens, such as those associated with “sad.” After identifying these tokens, we then use the difference between emotional tokens and neutral tokens to construct steering vectors for six basic emotions (Ekman, 1992). These steering vectors, combined with an adjustable strength parameter, are then used to control the synthesized emotional tone.

In summary, EmoSteer-TTS enables training-free and fine-grained emotion control, offering improved interpretability over existing approaches. The contributions of our method are:

- We present the first fine-grained and training-free EC-TTS approach by identifying and modulating internal emotion representations within existing TTS models.
- We provide new insights and enhanced interpretability for continuous EC-TTS by uncovering the emotion steering dynamics in pretrained TTS models, offering practical guidance for the design of the proposed algorithm.
- Extensive objective and subjective evaluations demonstrate the effectiveness of EmoSteer-TTS in fine-grained speech emotion control, showing its potential applicability across a wide range of pretrained TTS models.

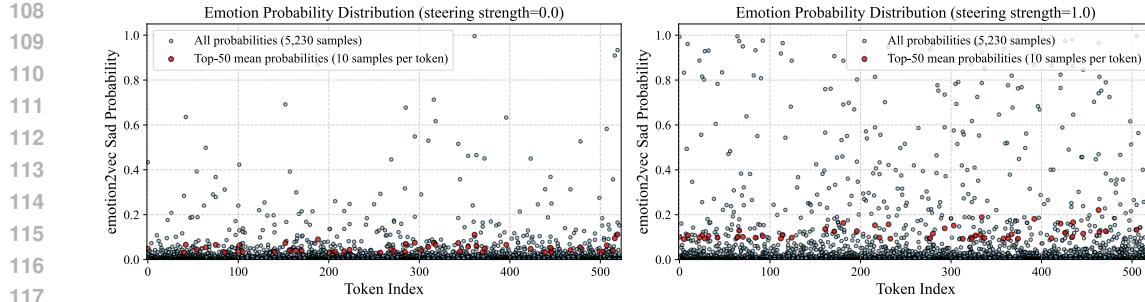


Figure 2: Adding a sadness steering vector to the activations in five DiT layers (1, 6, 11, 16, 21) of F5-TTS, conditioned on neutral speech, substantially increases the predicted sadness probability.

2 RELATED WORK

Emotion-Controllable Text-to-Speech. Unlike traditional TTS systems, e.g., VITS (Kim et al., 2021a) and VALL-E (Wang et al., 2023), that produce neutral or monotone speech, EC-TTS systems allow users to specify speech emotions, enabling more expressive and natural-sounding voices. **Label-based methods** control emotion using discrete labels (Cho et al., 2025). For instance, EmoD-ubber (Cong et al., 2025) uses a flow-based framework with positive/negative emotion guidance and a classifier to adjust emotion intensity. HED-TTS (Inoue et al., 2025) models hierarchical emotion distributions across speech segments, allowing multi-level intensity control. **Description-based methods** use textual prompts to specify emotions (Shimizu et al., 2024; Li et al., 2025; Ji et al., 2024; Zhou et al., 2025). PromptTTS (Guo et al., 2023) employs a BERT-based encoder to extract style from prompts and guide synthesis. VoxInstruct (Zhou et al., 2024) introduces semantic speech tokens and classifier-free guidance for fine-grained control from emotion descriptions. ControlSpeech (Ji et al., 2025) models emotional styles as Gaussian mixtures, aligning text and audio via KL divergence to enable zero-shot, controllable synthesis. Some zero-shot methods, e.g., MaskGCT (Wang et al., 2025b) and Vevo (Zhang et al., 2025), can also synthesize emotional speech, but they lack direct control and instead rely on reference speech. While these approaches have significantly advanced expressive speech synthesis, they require large-scale datasets and training.

Activation Steering. Activation steering aims to directly modulate the internal activations of neural networks, providing a means to exert fine-grained control over the behavior of pretrained models. Activation steering has shown great potential in the realm of LLMs. For example, it can be used to **control the behavior of LLMs**, such as enhancing the truthfulness of responses (Xiao et al., 2024; Wang et al., 2025a). Researchers can identify the mapping between the activation distributions associated with false or misleading statements and those of accurate information (Rodriguez et al., 2024). Then, during the generation process, the model’s activations are steered towards the distribution representing truth, encouraging LLMs to produce more factually correct outputs (Li et al., 2023a). Activation steering can also be used to **control text-to-image (T2I) diffusion models** (Li et al., 2024; Nair et al., 2023). By modifying the activations of the diffusion model towards the distribution that corresponds to a particular style, e.g., impressionist or cubist, the model can generate images with the desired aesthetic qualities (Rodriguez et al., 2024; Brack et al., 2022). Inspired by these advances, we explore emotion representations in pretrained zero-shot TTS models and apply activation steering, offering a stable and interpretable EC-TTS method.

3 METHOD

3.1 OVERVIEW

As shown in Fig. 3, the proposed EmoSteer-TTS approach consists of three key stages. First, we compute activation differences using pairs of neutral and emotional reference speeches. Second, we identify top- k emotion-relevant tokens (e.g., for “happy”) to construct a steering vector and its associated weight vector. At inference time, given any unseen reference speech and text, we control the emotion of the synthesized speech by applying the steering vector with a strength parameter to modify internal activations. The proposed method is detailed in the following subsections.

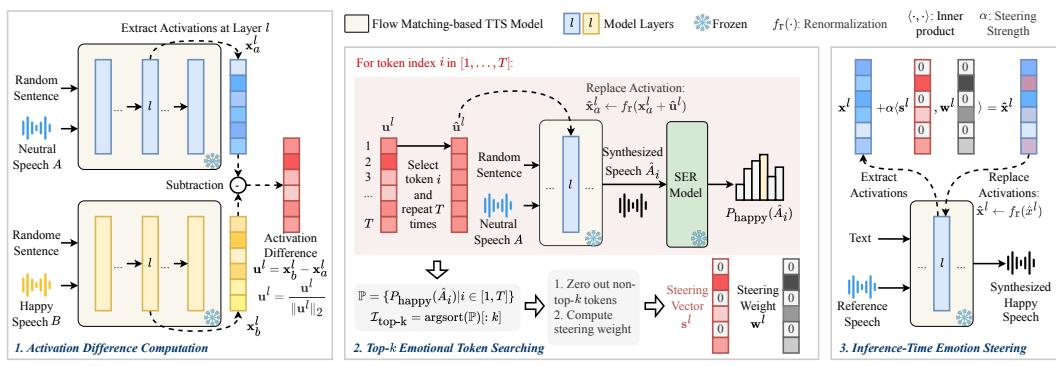


Figure 3: Overview of EmoSteer-TTS. Steering vectors and steering weights are derived from pairs of neutral and emotional reference speech. During inference, these vectors are used to modulate the activations in a TTS model, guiding it to synthesize speech that reflects the desired emotion.

3.2 ACTIVATION EXTRACTION

Our method focuses on zero-shot TTS models that use flow matching to synthesize mel-spectrograms. Given a pretrained TTS model with $|\mathcal{L}|$ DiT layers, we use random sentence texts along with M neutral speech samples (denoted as \mathcal{A}) and N emotional speech samples (denoted as \mathcal{B}) as inputs to synthesize a total of $M + N$ speech samples. For each model layer (a DiT block) $l \in \mathcal{L}$, we extract the first residual activations $\mathbf{x}_{a,i}^l$ and $\mathbf{x}_{b,j}^l$ for the synthesized speech conditioned on reference samples $A_i \in \mathcal{A}$ and $B_j \in \mathcal{B}$, respectively. The activation difference between neutral and target emotional speech at layer l is defined as:

$$\mathbf{u}^l = \frac{1}{N} \sum_{j=1}^N \mathbf{x}_{b,j}^l - \frac{1}{M} \sum_{i=1}^M \mathbf{x}_{a,i}^l. \quad (1)$$

This activation difference is also known as the *difference-in-means* (Belrose et al., 2023), which can effectively extract robust feature directions. To ensure stable steering, we normalize \mathbf{u}^l by dividing it by its L2 norm, resulting in a unit vector: $\mathbf{u}^l \leftarrow \frac{\mathbf{u}^l}{\|\mathbf{u}^l\|_2}$. The activation differences for all target layers to be steered are defined as $\mathcal{U} = \{\mathbf{u}^l \mid l \in \hat{\mathcal{L}}\}$, where $\hat{\mathcal{L}} \subseteq \mathcal{L}$ denotes the set of selected layers. It is worth noting that the direction of \mathbf{u}^l indicates the trajectory of emotional change in the feature space, while its original magnitude reflects the extent of the transition between emotions.

Synthesized speech may vary in length. Therefore, we use nearest interpolation to align the extracted activations (token sequences) to a fixed length, which is the average activation sequence length across all $M + N$ samples. As a result, each activation has the shape $[avg_seq_length, hidden_dim]$.

3.3 STEERING VECTOR CONSTRUCTION

After obtaining the activation difference \mathbf{u}^l , we select the top- k tokens most relevant to the target emotion to construct the steering vector. As illustrated in Fig. 3, for each token in \mathbf{u}^l , we repeat token $i \in [1, 2, \dots, T]$ T times to form a new vector $\hat{\mathbf{u}}^l$. We then modify the activation \mathbf{x}_a^l as follows:

$$\hat{\mathbf{x}}_a^l \leftarrow f_r(\mathbf{x}_a^l + \hat{\mathbf{u}}^l), \quad f_r = \frac{\|\mathbf{x}_a^l\|_2}{\|\mathbf{x}_a^l + \hat{\mathbf{u}}^l\|_2}, \quad (2)$$

where \mathbf{x}_a^l is the activation corresponding to a random sentence and a reference speech sample different from those used to compute \mathbf{u}^l , and f_r is a function that renormalizes the modified activation to preserve the original L2 norm, which ensures more stable modification Gaintseva et al. (2025).

After the activation modification, the model synthesizes the output sample \hat{A}_i corresponding to token i . We then use a pre-trained speech emotion recognition (SER) model, i.e., emotion2vec (Ma et al., 2024), to predict the probability that \hat{A}_i corresponds to the target emotion, denoted as $P_{\text{emotion}}(\hat{A}_i)$. By computing $P_{\text{emotion}}(\hat{A}_i)$ for all tokens, we obtain the probability set:

$$\mathbb{P} = \{P_{\text{emotion}}(\hat{A}_i) \mid i \in [1, T]\}, \quad (3)$$

216 and the indices of the top- k emotional tokens:
 217

$$218 \quad \mathcal{I}_{\text{top-}k} = \text{argsort}(\mathbb{P})[: k]. \quad (4)$$

219 Next, we zero out all non-top- k tokens in \mathbf{u}^l to derive the steering vector \mathbf{s}^l :
 220

$$221 \quad \mathbf{s}^l \leftarrow \mathbf{u}^l \odot \mathbf{m}, \quad \mathbf{m}_i = \begin{cases} 1, & \text{if } i \in \mathcal{I}_{\text{top-}k} \\ 0, & \text{otherwise} \end{cases}, \quad (5)$$

223 where \mathbf{m} is a mask vector, and \odot is element-wise multiplication. To apply adaptive steering strength
 224 to each token, we compute a steering weight vector \mathbf{w}^l as follows:
 225

$$226 \quad \mathbf{w}^l = \delta(\hat{\mathbb{P}}), \quad \hat{\mathbb{P}} = \{P_{\text{emotion}}(\hat{A}_i) | i \in \mathcal{I}_{\text{top-}k}\}, \quad (6)$$

228 where δ is the Softmax function: $\delta(z_i) = \frac{e^{z_i}}{\sum_{j=1}^k e^{z_j}}$. Finally, we get the weighted steering vector $\hat{\mathbf{s}}^l$:
 229

$$230 \quad \hat{\mathbf{s}}^l = \langle \mathbf{s}^l, \mathbf{w}^l \rangle = \mathbf{w}_1^l \mathbf{s}_1^l + \mathbf{w}_2^l \mathbf{s}_2^l + \dots + \mathbf{w}_T^l \mathbf{s}_T^l, \quad (7)$$

231 which can be used to steer speech emotions. Since most elements of the weighted steering vector are
 232 zero, $\hat{\mathbf{s}}^l$ lies within a subspace of the TTS model’s feature space that is specifically responsible for
 233 modeling emotional tone. To ensure the efficiency of the token searching process, we simultaneously
 234 modify all selected layers at the same token indices, which can reduce the computational complexity
 235 from $\mathcal{O}(|\hat{\mathcal{L}}| \times \text{avg_seq_length})$ to $\mathcal{O}(\text{avg_seq_length})$.
 236

237 3.4 FINE-GRAINED EMOTION CONTROL

239 In this subsection, we show how the proposed method enables fine-grained emotion control, including
 240 emotion conversion, interpolation, erasure, and composite manipulation.

241 **Emotion Conversion and Interpolation.** As shown in Fig. 3, given the text and reference speech,
 242 we can use the steering vector \mathbf{s}^l and weight \mathbf{w}^l to modify the activations in layer $l \in \hat{\mathcal{L}}$ as follows:
 243

$$244 \quad \hat{\mathbf{x}}^l = f_r(\mathbf{x}^l + \alpha \hat{\mathbf{s}}^l), \quad (8)$$

245 where α controls the steering strength. Note that $\hat{\mathbf{s}}^l$ has the same shape as a token, i.e.,
 246 $[\text{hidden_dim}]$. Thus, the plus sign in Eq. 8 involves an implicit broadcasting operation. Fine-
 247 grained emotion control, e.g., conversion and interpolation, can be achieved by tuning the parameter
 248 α : when $\alpha = 0$, the emotional tone of the synthesized speech remains unchanged; when $\alpha > 0$, the
 249 emotional tone is steered toward the target emotion; and when $\alpha < 0$, it is steered in the opposite
 250 direction of the target emotion.

251 **Emotion Erasure.** One may wish to synthesize new speech samples using the speaking style or
 252 timbre from the reference speech while disregarding the emotional tone. Suppose the weighted
 253 steering vector $\hat{\mathbf{s}}^l$ corresponds to the emotion conveyed by the reference speech, our method achieves
 254 this by subtracting the weighted steering vector $\hat{\mathbf{s}}^l$ from the original activation \mathbf{x}^l , multiplied by the
 255 projection of $\hat{\mathbf{s}}^l$ onto \mathbf{x}^l , which can be expressed as follows:
 256

$$257 \quad \hat{\mathbf{x}}^l = f_r(\mathbf{x}^l - \beta(\hat{\mathbf{s}}^l \cdot \mathbf{x}^l)\hat{\mathbf{s}}^l), \quad (9)$$

258 where β is the erasing strength. Eq. 9 also involves implicit broadcasting operations because $\hat{\mathbf{s}}^l$
 259 is a single vector while \mathbf{x}^l is a token sequence. Explanation of Eq. 9: Different reference speech
 260 samples may contain multiple emotions, including the target emotion at varying intensities. Our
 261 goal is to remove only the target emotion. The projection operation quantifies the intensity of the
 262 target emotion in the reference speech, while preserving all other speech characteristics.

263 **Composite Control.** EmoSteer-TTS also enables composite control over the emotional tone of
 264 synthesized speech. For example, given a reference speech sample, **emotion replacement** can be
 265 achieved through the following operation ($\hat{\mathbf{s}}_{\text{emo}_1}^l$ is the weighted steering vector of emotion “ emo_1 ”):
 266

$$267 \quad \hat{\mathbf{x}}^l = f_r(\mathbf{x}^l - \beta(\hat{\mathbf{s}}_{\text{emo}_1}^l \cdot \mathbf{x}^l)\hat{\mathbf{s}}_{\text{emo}_1}^l + \alpha \hat{\mathbf{s}}_{\text{emo}_2}^l), \quad (10)$$

268 which replaces emotion “ emo_1 ” with “ emo_2 ”. We can also realize **multiple emotion steering**:
 269

$$269 \quad \hat{\mathbf{x}}^l = f_r(\mathbf{x}^l + \alpha_1 \hat{\mathbf{s}}_{\text{emo}_1}^l + \alpha_2 \hat{\mathbf{s}}_{\text{emo}_2}^l + \dots + \alpha_E \hat{\mathbf{s}}_{\text{emo}_E}^l), \quad (11)$$

270 which is particularly useful for synthesizing speech with compound emotions, such as “contempt”
 271 (disgust combined with mild anger), “pleasant surprise” (a mix of happiness and surprise), as well
 272 as more nuanced emotions like “happiness tinged with sadness” or “anger intertwined with fear”.
 273

274 EmoSteer-TTS enables fine-grained, continuous emotional control and supports multiple control
 275 strategies, representing the first training-free EC-TTS approach. **Appendix A** provides code snippets
 276 for the operations described above.

277 **4 EXPERIMENT**

279 **4.1 DATASETS AND MODELS**

281 **Datasets for Steering Vector Construction.** To obtain effective steering vectors, we construct a cu-
 282 rated emotional speech dataset by collecting samples with clear emotional expression from multiple
 283 corpora: MSP-Podcast (Lotfian & Busso, 2017), IEMOCAP (Busso et al., 2008), RAVDESS (Liv-
 284 ingtonstone & Russo, 2018), CREMA-D (Cao et al., 2014), TESS (Pichora-Fuller & Dupuis, 2020),
 285 SAVEE (Jackson & Haq, 2014), ASVP-ESD (Landry et al., 2020), CASIA (CASIA, 2023),
 286 M3ED (Zhao et al., 2022), ESD (Zhou et al., 2022), and Emo-Emilia (Zhao et al., 2025). The
 287 resulting dataset contains 6,900 utterances covering six basic emotions (anger, happiness, sadness,
 288 disgust, surprise, fear) and neutrality. Each emotion includes 1,000 samples, 500 in English and 500
 289 in Chinese, except for fear, which has 400. The dataset includes diverse speakers with a balanced
 290 gender distribution. The construction details are provided in **Appendix B**. This dataset is used to
 291 compute activation differences between neutral and emotional speech, as defined in Eq. 1. To iden-
 292 tify the top- k tokens for each emotion, we synthesize speech using 10 random neutral ESD samples
 293 as references (5 English and 5 Chinese).

294 **Datasets for Inference-Time Emotion Steering.** 1) In-distribution evaluation: We sample neutral
 295 and emotional reference speeches from MSP-Podcast and ESD, which are excluded from steer-
 296 ing vector computation. 2) Out-of-distribution (OOD) evaluation: We sample neutral speech from
 297 SeedTTS Anastassiou et al. (2024) test sets and emotional speech from EMNS Noriy et al. (2023).

298 **Models.** We enhance three SOTA flow matching-based TTS models (F5-TTS, CosyVoice2, E2-
 299 TTS (Eskimez et al., 2024)) using our proposed method, and compare their controllability with
 300 that of leading EC-TTS baselines, including both label-based methods with adjustable control
 301 strength (EmoSphere++, EmoDubber, HED-TTS (Inoue et al., 2025)) and description-based meth-
 302 ods (EmoVoice, CosyVoice2, FleSpeech (Li et al., 2025)). **Appendix C** provides detailed rater
 303 information, model and hardware configurations for all experiments.

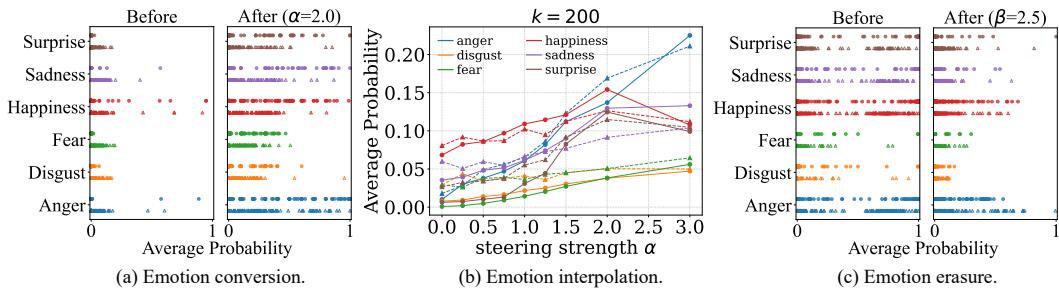
304 **4.2 EMOTION CONVERSION AND INTERPOLATION**

306 **Emotion Conversion.** We conduct emotion conversion using 100 neutral reference speech sam-
 307 ples (50 English from MSP-Podcast and 50 Chinese from ESD), with $\alpha=2.0$ and $k=200$. We report
 308 Word Error Rate (WER), Speaker Similarity (S-SIM), Emotion Similarity (E-SIM), and Natural-
 309 ness Mean Opinion Score (N-MOS, 1–5 scale, see **Appendix D** for details). WER is derived from
 310 Whisper-Large V3 (Radford et al., 2023) transcriptions. S-SIM is the cosine similarity between the
 311 embeddings of synthesized and neutral reference from a speaker embedding model (Bredin et al.,
 312 2020). E-SIM is computed as the cosine similarity between emotion2vec embeddings of synthe-
 313 sized speech and 100 anchor emotional samples (per emotion) from MSP-Podcast and ESD. To
 314 mitigate potential metric overfitting from emotion2vec, we also report E-SIM scores computed with
 315 SenseVoice An et al. (2024) embeddings. Since we cannot guarantee the synthesis quality of repro-
 316 duced baselines, we compute their scores using demo samples for fairness. The reproduced baseline
 317 results are additionally reported in **Appendix E**. As shown in Table 1, EmoSteer-TTS achieves
 318 superior performance across multiple methods. Integrated with F5-TTS, it yields a low WER of
 319 2.79, close to CosyVoice2 (2.53) and far better than label-based baselines. It also maintains high S-
 320 SIMs (0.66, 0.65), indicating strong speaker preservation. F5-TTS, E2-TTS, and CosyVoice2 with
 321 EmoSteer-TTS reach the top E-SIM scores, outperforming all baselines and matching FleSpeech. In
 322 N-MOS, “EmoSteer-TTS+CosyVoice2” (3.65) is close to the best (EmoVoice, 3.81), and our method
 323 consistently outperforms label-based systems. Fig. 4(a) also shows the shift in emotion probability
 324 distribution (averaged across three models) for 100 synthesized samples per emotional tone before
 325 ($\alpha=0$) and after ($\alpha=2$) emotion conversion.

324 Table 1: In-distribution and OOD comparison with emotion-controllable baselines.
325

326	327	Method	Conversion ($\alpha = 2.0$)			Interpolation		Erasure ($\beta = 2.5$)	
			WER(\downarrow)	S-SIM(\uparrow)	E-SIM(\uparrow)	N-MOS(\uparrow)	EI-MOS(\uparrow)	E-SIM(\uparrow)	EE-MOS(\uparrow)
329	330	Label-based*	EmoSphere++	16.25	0.44	0.25 / 0.24 _{avg=0.245}	3.23 _{±0.81}	3.50 _{±1.05}	-
			EmoDubber	18.61	0.41	0.25 / 0.22 _{avg=0.235}	2.47 _{±1.22}	2.21 _{±1.08}	-
			HED-TTS	13.27	0.52	0.22 / 0.26 _{avg=0.240}	3.31 _{±0.79}	2.59 _{±0.76}	-
331	332	Description-based*	EmoVoice	2.91	0.58	0.27 / 0.25 _{avg=0.260}	3.81 _{±0.86}	-	-
			CosyVoice2	2.53	0.73	0.24 / 0.27 _{avg=0.255}	3.69 _{±1.07}	-	-
			FleSpeech	9.34	0.54	0.29 / 0.26 _{avg=0.275}	3.07 _{±0.75}	-	-
333	334	Unsteered	F5-TTS	2.14	0.66	0.07 / 0.04 _{avg=0.055}	3.79 _{±0.89}	-	0.03 / 0.05 _{avg=0.040}
			E2-TTS	2.71	0.64	0.05 / 0.08 _{avg=0.065}	3.51 _{±0.94}	-	0.06 / 0.02 _{avg=0.040}
In-distribution evaluation on MSP-Podcast (25% en) and ESD (25% en, 50% zh)									
335	336	(Ours)	+ F5-TTS	2.79	0.64	0.29 / 0.26 _{avg=0.275}	3.29 _{±1.05}	4.00 _{±0.89}	0.27 / 0.25 _{avg=0.260}
			+ E2-TTS	3.28	0.59	0.28 / 0.28 _{avg=0.280}	3.31 _{±0.97}	3.38 _{±1.09}	0.24 / 0.26 _{avg=0.250}
			+ CosyVoice2	2.83	0.65	0.26 / 0.29 _{avg=0.275}	3.65 _{±1.08}	3.56 _{±1.15}	0.26 / 0.25 _{avg=0.255}
Cross-datasets (OOD) evaluation on EMNS (25% en) and SeedTT test sets (25% en, 50% zh)									
339	340	(Ours)	+ F5-TTS	2.65	0.65	0.25 / 0.27 _{avg=0.260}	3.58 _{±1.04}	3.46 _{±1.08}	0.25 / 0.22 _{avg=0.235}
			+ E2-TTS	3.41	0.55	0.26 / 0.25 _{avg=0.255}	3.44 _{±1.07}	3.50 _{±0.97}	0.24 / 0.27 _{avg=0.255}
			+ CosyVoice2	2.86	0.66	0.28 / 0.25 _{avg=0.265}	3.49 _{±1.01}	3.48 _{±1.27}	0.23 / 0.21 _{avg=0.220}
*: Training-based, #: Training-free, -: Neither label-based, description-based, nor unsteered methods support interpolation or erasure. The top three results are indicated in boldface. Unsteered backbones are shown in gray for reference.									

341 *: Training-based, #: Training-free, -: Neither label-based, description-based, nor unsteered methods support interpolation or erasure.
342 The top three results are indicated in boldface. Unsteered backbones are shown in gray for reference.

353 Figure 4: Emotion steering results on MSP-Podcast and ESD. ● emotion2vec, ▲ SenseVoice.
354

355 **Emotion Interpolation.** We reuse the speech samples from the emotion conversion experiments
356 to perform interpolation ($k=200$), gradually shifting emotional tone from neutrality to a target emotion.
357 To assess fine-grained controllability, we report the Emotion Interpolation MOS (EI-MOS; 1–5
358 scale), which evaluates the alignment between target intensity and synthesized speech. Detailed cri-
359 teria for EI-MOS are provided in **Appendix D**. Label-based baselines use intensity levels (e.g., 0.5 or
360 1.0) to control, while description-based methods, lacking intensity control, are excluded in this ex-
361 periment. For fairness, baseline metrics are computed using their official demo samples. As shown
362 in Table 1, EmoSteer-TTS achieves higher EI-MOS than label-based baselines, indicating superior
363 capability in controlling emotional intensity. Notably, “EmoSteer-TTS+F5-TTS” obtains the high-
364 est EI-MOS of 4.00, outperforming EmoSphere++ and HED-TTS, showing better alignment with
365 intended emotion levels. E2-TTS and CosyVoice2 variants also perform well, suggesting EmoSteer-
366 TTS generalizes across different models. As shown in Fig. 4(b), the average predicted emotion
367 probabilities (via emotion2vec and SenseVoice) vary smoothly with α , illustrating EmoSteer-TTS’s
368 fine-grained controllability. However, we find that large α values (e.g., 3) may lead to unintelli-
369 gible speech. Fig. 5(a) also illustrates smooth F0 transitions with increasing anger intensity. More
370 examples are provided in **Appendix F**.

371

4.3 EMOTION ERASURE

372

373 We randomly select 100 unseen emotional speech samples for each type of emotion from MSP-
374 Podcast (50 English) and ESD (50 Chinese), and erase the emotional tone using Eq. 9. We report the
375 average E-SIM between the emotionally erased samples and 100 randomly selected neutral samples
376 from MSP-Podcast (50 English) and ESD (50 Chinese). We also report Emotion-Erasure MOS (EE-
377 MOS, 1–5 scale), which indicates how well the synthesized speech reflects the intended emotion
erasure. Higher EE-MOS reflects better erasure performance. The standard for EE-MOS is detailed

378 in **Appendix D**. We set $\beta=2.5$, $k=200$ for this
 379 experiment. As shown in Table 1, our method
 380 achieves a fairly high EE-MOS score, indicating
 381 effective removal of target emotions. The de-
 382 creased target emotion scores shown in Fig. 4(c)
 383 further demonstrate the emotion erasing ability.
 384 Fig. 5(b) illustrates the variation of F0 contours
 385 when gradually erasing an emotional tone. **Ap-
 386 pendix F** provides more visualizations.

387 4.4 COMPOSITE CONTROL

391 **Emotion Replacement.** We use the same emo-
 392 tional samples from the emotion erasure experi-
 393 ment as reference speech for three TTS models.
 394 As defined by Eq. 10, we first remove the emo-
 395 tional tone of the original activation and add a
 396 target emotion. We perform six groups of re-
 397 placement with $\alpha=2$, $\beta=2.5$, and $k=200$. The
 398 values in Fig. 6(a) are computed by subtracting
 399 the emotion2vec probabilities before emotion re-
 400 placement from those after replacement. Each
 401 row represents a specific replacement operation
 402 (e.g., F→H denotes replacing fear with hap-
 403 piness), while each column indicates the pre-
 404 dicted probability change for a given emotion.
 405 The diagonal patterns validate the success of em-
 406otion transfer, e.g., F→H shows an increase in hap-
 407 piness (+0.28) and a marked decrease in fear
 408 (-0.33). Similar trends are observed for other
 409 pairs, such as Su→A and H→Sa, confirming that
 410 EmoSteer-TTS effectively suppresses the original
 411 emotion and enhances the target one.

412 **Multi-Emotion Steering.** We use the same neu-
 413 tral samples from the emotion conversion experi-
 414 ment as reference speech. For simplicity, this ex-
 415 periment simultaneously adds two emotions to the
 416 synthesized speech ($\alpha_1=\alpha_2=2$, $k=200$). As shown in Fig. 6(b), the predicted emotion2vec dis-
 417 tributions align closely with the intended emotion pairs.
 418 For example, the row labeled “F, H” shows ele-
 419 vated probabilities for both fear (0.22) and hap-
 420 piness (0.33), while “Sa, Su” leads to strong ac-
 421 tivations for sadness (0.28) and surprise (0.42).
 422 These results indicate that EmoSteer-TTS can blend
 423 multiple emotions, enabling expressive and nu-
 424 anced speech synthesis beyond single-label control.

425 4.5 CROSS-DATASETS EVALUATION

426 Since some samples used for computing steering vectors come from the same datasets (e.g., MSP-
 427 Podcast, ESD), we also evaluate EmoSteer-TTS in an OOD setting. For emotion conversion and
 428 interpolation, we sample 100 neutral utterances from SeedTTS and 100 emotional anchors per emotion
 429 from EMNS; for emotion erasure, we use 100 emotional utterances from EMNS as references
 430 and 100 neutral anchors per emotion from SeedTTS. As shown in Table 1 (lower section), EmoSteer-
 431 TTS maintains robust performance on unseen datasets, with minimal degradation across metrics,
 432 demonstrating strong generalization beyond the steering data.

433 In addition to the main experiments, we report an ablation on steering corpus composition in **Ap-
 434 pendix H.1** to investigate the influence of data quantity. We further provide correlation analyses
 435 between E-SIM and N-MOS/EE-MOS in **Appendix H.2**. Confidence intervals and significance
 436 tests for the subjective evaluations are included in **Appendix H.3** for completeness.

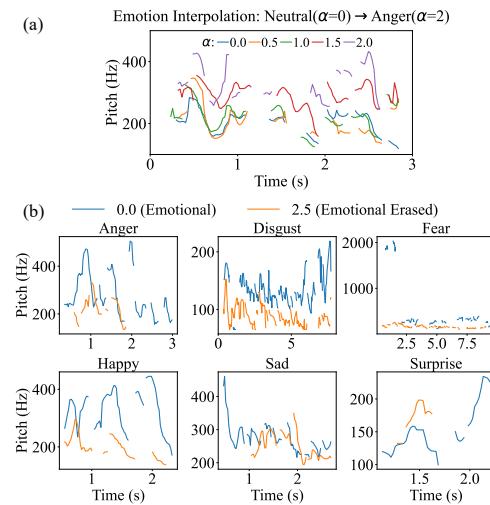


Figure 5: Visualization of F0 contours. (a) An example showing how the F0 contour varies with steering intensity; (b) The speech tone (F0 contour) becomes calmer after emotion erasure.

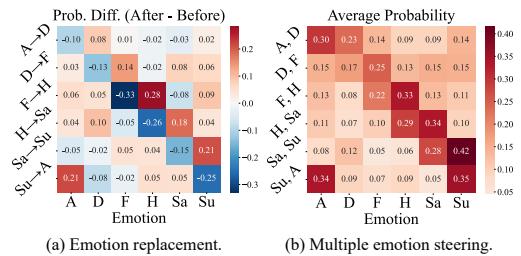


Figure 6: Results of composite control: (a) emotion replacement and (b) multi-emotion steering (Abbreviations: Anger, Disgust, Fear, Happiness, Sadness, Surprise)

8

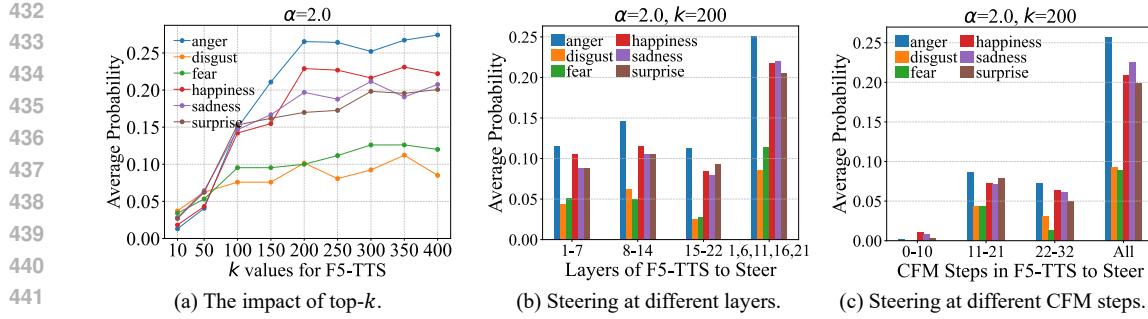


Figure 7: Analysis of emotion steering dynamics using emotion2vec predictions.

446 4.6 ANALYSIS OF EMOTION STEERING DYNAMICS

448 In this subsection, we analyze the emotion steering dynamics of our method. All analyses are con-
 449 ducted on F5-TTS, which consists of 22 DiT block layers and performs 32 flow matching steps to
 450 generate mel-spectrograms. We use the same neutral samples from the emotion conversion experi-
 451 ment as reference speech. We report emotion2vec emotion probabilities for all analyses.

452 **The Impact of Top- k .** The parameter k determines the number of emotion-related tokens used to
 453 construct the steering vectors. A larger k introduces more tokens into the steering signal, potentially
 454 capturing a broader range of emotional nuances, while a smaller k focuses on the most dominant
 455 emotion features. We conduct emotion conversion with varying k values (e.g., $k \in 10, 50, \dots, 400$)
 456 and evaluate their impact on the synthesized emotion. Fig. 7(a) shows that increasing k generally
 457 leads to higher average emotion probabilities across all categories, particularly for anger and hap-
 458 piness, which peak around $k=200$. Incorporating more emotion-relevant tokens enriches the steering
 459 signal, but gains plateau beyond $k = 200$ for most emotions. We therefore use $k = 200$ in all main
 460 experiments to balance expressiveness and efficiency.

461 **Steering Different Layers.** We examine how controlled layers affect emotion conversion by apply-
 462 ing steering vectors s^l at shallow (1–7), middle (8–14), deep (15–22), and spaced layers (1, 6, 11,
 463 16, 21). As shown in Fig. 7(b), shallow layers yield moderate emotional influence, middle layers
 464 provide slightly stronger control, and deep layers show a decline, likely focusing on acoustic details
 465 rather than emotion. Steering multiple spaced layers, however, significantly boosts probabilities
 466 across all six emotions. Overall, shallow-to-deep layers provide progressively refined control, and
 467 multi-layer steering enables the most effective emotion modulation.

468 **Steering Different Flow Matching Steps.** F5-TTS generates mel-spectrograms through 32 con-
 469 ditional flow-matching (CFM) steps. To assess the impact of steering at different stages, we apply
 470 emotion control to early (0–10), middle (11–21), late (22–32), or all steps. As shown in Fig. 7(c),
 471 early steering has little effect, while middle and late stages exert stronger influence as the spectro-
 472 gram takes shape. The strongest emotion emerges when steering spans all steps, consistent with
 473 CFM’s stepwise conditioning on reference speech. Therefore, we apply emotion steering across all
 474 steps in the main experiments: 32 for F5-TTS and E2-TTS, and 10 for CosyVoice2.

475 **Safe Steering Range.** Understanding the trade-off between steering strength α and audio quality is
 476 crucial for practical use. We have already reported the E-SIM variations in Fig. 4(b) for the emotion
 477 interpolation experiment. Using the same synthesized samples and newly synthesized samples with
 478 $\alpha = 2.5$, we further present the averaged in-distribution N-MOS and WER variations as a function of
 479 α . The detailed results are shown in Tables 10, 11, and 12 in **Appendix H.4**. In summary, increasing
 480 α produces a highly consistent pattern across all emotions and models. For small to moderate values
 481 (up to about 1.0–1.5), N-MOS and WER remain close to the baseline. As α increases further, N-
 482 MOS declines and WER rises, and very large values (≥ 2.5) push the models outside their normal
 483 operating range, leading to distortion. This trend is nearly identical across F5-TTS, E2-TTS, and
 484 CosyVoice2, suggesting a general effect of excessive steering on model representations, likely due
 485 to shared training practices such as normalization and gradient clipping. Therefore, we recommend
 486 the following ranges for choosing α : 1) Stable region (mild emotion): $\alpha \leq 1.0$; 2) Controlled region
 487 (stronger emotion): $1.0 < \alpha \leq 2.0$; 3) Unstable region (risk of distortion): $\alpha > 2.0$.

Table 2: Cross-lingual emotion conversion ($\alpha=2.0$, F5-TTS, token probing: emotion2vec).

Method	WER↓	S-SIM↑	E-SIM↑ emotion2vec / SenseVoice	UTMOS↑
English→Chinese	92.74	0.21	0.13 / 0.08	2.45
Chinese→English	85.51	0.36	0.09 / 0.11	3.07

Table 3: Inference time overhead brought by EmoSteer-TTS.

Backbone	w/o Steering (s)	Conversion (s)	Interpolation (s)	Erasure (s)
F5-TTS	1.867	2.415 (+0.548)	2.504 (+0.637)	2.746 (+0.879)
E2-TTS	0.942	1.258 (+0.316)	1.244 (+0.302)	1.451 (+0.509)
CosyVoice2	3.598	4.143 (+0.545)	4.261 (+0.663)	4.464 (+0.866)

4.7 GENERALIZATION ANALYSIS

The Sensitivity to SER Model for Probing. We further evaluate the sensitivity of token probing to the choice of SER model. We replace emotion2vec with SenseVoice and report E-SIM under both embeddings to assess potential overfitting. Using the same neutral and emotional samples in the main experiments as speech prompts, we report WER, S-SIM, and E-SIM for emotion conversion and erasure. We also use UTMOS (Saeki et al., 2022) instead of N-MOS to avoid labor-intensive human evaluation. As shown in Tables 13 and 14 in [Appendix H.5](#), EmoSteer-TTS shows only a very slight preference for the SenseVoice embedding space, with marginally higher E-SIM scores than under emotion2vec, indicating only mild overfitting to the SER model used for token probing. Nonetheless, human subjective results in Table 1 align with the objective metrics, confirming that EmoSteer-TTS is genuinely effective rather than overfitting a specific embedding space.

Cross-lingual Transfer. To assess whether a steering vector learned in one language transfers to another, we apply the precomputed English and Chinese steering vectors to the same reference samples in our in-distribution emotion conversion experiment, using the F5-TTS backbone. As shown in Table 2, cross-lingual transfer is highly limited. Applying the English vector to Chinese speech yields large WER degradation (92.74) and notably reduced S-SIM and E-SIM, indicating poor linguistic and emotional consistency. The reverse direction shows similar trends. These results suggest that emotion steering directions are largely language-specific, likely due to differences in phoneme–token mappings, prosody, and language-dependent emotional expression patterns.

We also analyze EmoSteer-TTS's robustness to noisy and reverberant prompts in [Appendix H.6](#).

4.8 INFERENCE-TIME EFFICIENCY

To measure the computational efficiency of our method, we use the same settings as in our main experiments (conversion, interpolation, and erasure). For each type of activation steering, we employ PyTorch hooks to modify the activations during the forward pass. The additional average (per sample) inference-time overhead introduced by our method is shown in Table 3. The computational overhead is almost negligible, demonstrating the high efficiency of our methods.

5 CONCLUSION

We presented EmoSteer-TTS, the first training-free framework for fine-grained, continuous, and interpretable emotion control in speech synthesis. By steering a subset of internal activations in a TTS model, our method enables flexible emotional manipulation, including emotion conversion, interpolation, and erasure, without modifying or fine-tuning the pretrained TTS model. We also constructed a curated emotional speech dataset to support steering vector construction. Extensive experiments confirm that EmoSteer-TTS achieves robust, zero-shot emotion control with broad applicability, outperforming SOTA methods. The analysis also offers deeper insights into the emotion steering dynamics of flow matching-based TTS. **To the best of our knowledge, this is the first fine-grained EC-TTS approach that can transform previously uncontrollable TTS models into emotionally controllable ones without any retraining, fine-tuning, and model architecture redesign.**

Limitations and Future Work. A limitation of our method is the reliance on high-quality emotional speech samples, albeit in modest quantities, to extract effective steering vectors. In addition, strong activation steering may introduce artifacts. Future work will explore combining activation steering with learning-based approaches to mitigate these issues. We also acknowledge that whether the assumption of a linearly steerable emotion subspace holds for other architectures (e.g., VITS, VAEs, or AR models) remains an open and exciting question, which will be investigated in our future work.

540 **6 REPRODUCIBILITY STATEMENT**
541

542 To ensure the reproducibility of our work, we have provided comprehensive details throughout the
 543 paper and its appendices. Our proposed methodology, EmoSteer-TTS, is thoroughly described in
 544 **Section 4**, including the key algorithms for activation extraction, steering vector construction, and
 545 fine-grained control, accompanied by precise mathematical formulations. **Appendix A** further of-
 546 fers detailed code snippets illustrating the implementation of our core steering operations. Details
 547 regarding the datasets used for both steering vector construction and evaluation are presented in **Sec-
 548 tion 4.1**, with the specific curation and filtering process for our emotional speech dataset outlined
 549 in **Appendix B**. We also provide the code for dataset preprocessing and the processed dataset in
 550 the **Supplementary Materials**. The configurations for the TTS models (F5-TTS, E2-TTS, Cosy
 551 Voice2), including the specific layers and steps selected for steering, are detailed in **Appendix C**.
 552 The hyperparameters and experimental settings for all evaluations are specified within the relevant
 553 subsections of **Section 4**, and the criteria for our subjective evaluation metrics are defined in **Ap-
 554 pendix D**. We will release the fully runnable code and curated dataset upon the paper’s acceptance
 555 to facilitate further research.

556 **7 ETHICS STATEMENT**
557

558 **Possible Bias and Fairness.** Our steering vectors rely on the representations learned by SER models
 559 (emotion2vec) and the demographic distribution of our curated dataset, which may raise bias and
 560 fairness concerns. While we utilized 11 diverse corpora to ensure gender balance, the steering
 561 vectors are currently language-specific (i.e., for English and Chinese only). Future work will focus
 562 on developing language-agnostic steering vectors to ensure equitable performance across accents
 563 and dialects.

564 **Privacy and Data Usage.** All data used to construct the steering vectors are derived from publicly
 565 available, consented academic datasets. As a training-free method, EmoSteer-TTS does not modify
 566 model weights, eliminating the risk of accidental memorization of inference-time user data.

568 **Misuse and Mitigation.** We acknowledge that fine-grained emotion control increases the realism
 569 of synthesized speech, potentially raising the risk of misuse in deepfakes or social engineering.
 570 However, our method’s interpretability offers a unique advantage: the steering vectors themselves
 571 act as known “signatures” of manipulation. To mitigate risks, we strongly advocate for the use
 572 of invisible audio watermarking in downstream applications. Furthermore, the “emotion erasure”
 573 capability, while potentially misuseable, also serves as a tool for removing toxic emotional cues
 574 from speech data used in training safety-aligned models.

575 **REFERENCES**
576

577 Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He,
 578 Hangrui Hu, Kai Hu, et al. Funaudiollm: Voice understanding and generation foundation models
 579 for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*, 2024.

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

584 Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella
 585 Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information
 586 Processing Systems*, 36:66044–66063, 2023.

587 Manuel Brack, Patrick Schramowski, Felix Friedrich, Dominik Hintersdorf, and Kristian Kersting.
 588 The stable artist: Steering semantics in diffusion latent space. *arXiv preprint arXiv:2212.06013*,
 589 2022.

591 Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin,
 592 Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote.audio: Neural
 593 building blocks for speaker diarization. In *IEEE International Conference on Acoustics, Speech
 and Signal Processing*, pp. 7124–7128, 2020.

594 Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jean-
 595 nette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic
 596 motion capture database. *Language Resources and Evaluation*, 42:335–359, 2008.

597

598 Huawei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini
 599 Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on*
 600 *Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

601 CASIA. Casia chinese emotional audio corpus. [https://aistudio.baidu.com/](https://aistudio.baidu.com/datasetdetail/209512)
 602 datasetdetail/209512, 2023. Accessed: 2025-07-01.

603

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

608 Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee. EmoSphere++:
 609 Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. *IEEE Trans-*
 610 *actions on Affective Computing*, pp. 1–16, 2025.

611 Gaoxiang Cong, Jiadong Pan, Liang Li, Yuankai Qi, Yuxin Peng, Anton van den Hengel, Jian Yang,
 612 and Qingming Huang. Emodubber: Towards high quality and emotion controllable movie dub-
 613 bing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15863–
 614 15873, 2025.

615

616 Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang,
 617 Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen
 618 Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. Cosyvoice 2: Scalable streaming speech
 619 synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

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

623

624 Paul Ekman. Facial expressions of emotion: New findings, new questions. *Psychological Science*,
 625 3(1):34–38, 1992.

626

627 Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao,
 628 Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. E2 tts: Embarrassingly easy fully non-
 629 autoregressive zero-shot tts. In *IEEE Spoken Language Technology Workshop*, pp. 682–689, 2024.

630

631 Tatiana Gaintseva, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang
 632 Deng, and Ismail Elezi. Casteer: Steering diffusion models for controllable generation. *arXiv*
 633 *preprint arXiv:2503.09630*, 2025.

634

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

638

639 Kilem L Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of*
 640 *agreement among raters*. Advanced Analytics, LLC, 2014.

641

642 Sho Inoue, Kun Zhou, Shuai Wang, and Haizhou Li. Hierarchical control of emotion rendering in
 643 speech synthesis. *IEEE Transactions on Affective Computing*, pp. 1–13, 2025. doi: 10.1109/
 644 TAFFC.2025.3582715.

645

646 Philip Jackson and SJUoSG Haq. Surrey audio-visual expressed emotion (savee) database. *Univer-*
 647 *sity of Surrey: Guildford, UK*, 2014.

648

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

648 Shengpeng Ji, Qian Chen, Wen Wang, Jialong Zuo, Minghui Fang, Ziyue Jiang, Hai Huang, Ze-
 649 han Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. Controlspeech: Towards simultaneous
 650 and independent zero-shot speaker cloning and zero-shot language style control. *arXiv preprint*
 651 *arXiv:2406.01205*, 2025.

652 Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial
 653 learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on*
 654 *Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5530–5540,
 655 2021a.

656 Minchan Kim, Sung Jun Cheon, Byoung Jin Choi, Jong Jin Kim, and Nam Soo Kim. Expressive
 657 text-to-speech using style tag. In *Interspeech 2021*, pp. 4663–4667, 2021b. doi: 10.21437/
 658 Interspeech.2021-465.

659 Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Emanuel Habets, Rein-
 660 hold Haeb-Umbach, Volker Leutnant, Armin Sehr, Walter Kellermann, Roland Maas, et al. The
 661 reverb challenge: A common evaluation framework for dereverberation and recognition of re-
 662 verberant speech. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and*
 663 *Acoustics*, pp. 1–4, 2013.

664 DTT Landry, Qianhua He, Haikang Yan, and Yanxiong Li. Asvp-esd: A dataset and its benchmark
 665 for emotion recognition using both speech and non-speech utterances. *Global Scientific Journals*,
 666 8:1793–1798, 2020.

667 Hanzhao Li, Yuke Li, Xinsheng Wang, Jingbin Hu, Qicong Xie, Shan Yang, and Lei Xie. Flespeech:
 668 Flexibly controllable speech generation with various prompts. *arXiv preprint arXiv:2501.04644*,
 669 2025.

670 Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time
 671 intervention: Eliciting truthful answers from a language model. *Advances in Neural Information*
 672 *Processing Systems*, 36:41451–41530, 2023a.

673 Senmao Li, Joost van de Weijer, taihang Hu, Fahad Khan, Qibin Hou, Yaxing Wang, and jian Yang.
 674 Get what you want, not what you don't: Image content suppression for text-to-image diffusion
 675 models. In *The Twelfth International Conference on Learning Representations*, 2024.

676 Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. Styletts 2:
 677 Towards human-level text-to-speech through style diffusion and adversarial training with large
 678 speech language models. In *Advances in Neural Information Processing Systems*, volume 36, pp.
 679 19594–19621, 2023b.

680 Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow
 681 matching for generative modeling. In *The Eleventh International Conference on Learning Repre-*
 682 *sentations*, pp. 1–28, 2023.

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

686 Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retriev-
 687 ing emotional speech from existing podcast recordings. *IEEE Transactions on Affective Comput-*
 688 *ing*, 10(4):471–483, 2017.

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

692 Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and
 693 Oriol Nieto. librosa: Audio and music signal analysis in python. *SciPy*, 2015:18–24, 2015.

694 Albert Mehrabian. Basic dimensions for a general psychological theory. In *Basic Dimensions for*
 695 *a General Psychological Theory*, pp. 39–53. Oelgeschlager, Gunn & Hain, 1980. ISBN 978-0-
 696 89946-004-8.

702 Nithin Gopalakrishnan Nair, Anoop Cherian, Suhas Lohit, Ye Wang, Toshiaki Koike-Aokino,
 703 Vishal M Patel, and Tim K Marks. Steered diffusion: A generalized framework for plug-and-
 704 play conditional image synthesis. In *Proceedings of the IEEE/CVF International Conference on*
 705 *Computer Vision*, pp. 20850–20860, 2023.

706 Kari Ali Noriy, Xiaosong Yang, and Jian Jun Zhang. Emns/imz/corpus: An emotive single-
 707 speaker dataset for narrative storytelling in games, television and graphic novels. *arXiv preprint*
 708 *arXiv:2305.13137*, 2023.

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

711 Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. Voice-
 712 Craft: Zero-shot speech editing and text-to-speech in the wild. In *Proceedings of the 62nd Annual*
 713 *Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12442–
 714 12462, 2024.

715 M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (TESS), 2020. URL
 716 <https://doi.org/10.5683/SP2/E8H2MF>.

717 Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever.
 718 Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th Interna-*
 719 *tional Conference on Machine Learning*, 2023.

720 Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-
 721 networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Process-*
 722 *ing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992,
 723 2019.

724 Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi,
 725 and Xavier Suau. Controlling language and diffusion models by transporting activations. *arXiv*
 726 *preprint arXiv:2410.23054*, 2024.

727 Yan Rong, Jinting Wang, Shan Yang, Guangzhi Lei, and Li Liu. Audiogenie: A training-
 728 free multi-agent framework for diverse multimodality-to-mutiaudio generation. *arXiv preprint*
 729 *arXiv:2505.22053*, 2025.

730 Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi
 731 Saruwatari. UTOKYO-SaruLab System for VoiceMOS Challenge 2022. In *Interspeech*
 732 2022, pp. 4521–4525, 2022.

733 Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Ko-
 734 matsu, and Kentaro Tachibana. Prompttts++: Controlling speaker identity in prompt-based text-
 735 to-speech using natural language descriptions. In *IEEE International Conference on Acoustics,*
 736 *Speech and Signal Processing*, pp. 12672–12676, 2024.

737 Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *arXiv*
 738 *preprint arXiv:2106.15561*, 2021.

739 Greg Wadley, Vassilis Kostakos, Peter Koval, Wally Smith, Sarah Webber, Anna Cox, James J Gross,
 740 Kristina Höök, Regan Mandryk, and Petr Slovák. The future of emotion in human-computer
 741 interaction. In *CHI Conference on human factors in computing systems extended abstracts*, pp.
 742 1–6, 2022.

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

746 Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha
 747 Wang, and Liantao Ma. Adaptive activation steering: A tuning-free ILM truthfulness improvement
 748 method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference*, pp.
 749 2562–2578, 2025a.

756 Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang
 757 Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. MaskGCT: Zero-shot text-to-speech
 758 with masked generative codec transformer. In *The Thirteenth International Conference on Learning
 759 Representations*, pp. 1–24, 2025b.

760

761 Yuxin Xiao, Wan Chaoqun, Yonggang Zhang, Wenxiao Wang, Binbin Lin, Xiaofei He, Xu Shen,
 762 and Jieping Ye. Enhancing multiple dimensions of trustworthiness in LLMs via sparse activation
 763 control. *Advances in Neural Information Processing Systems*, 37:15730–15764, 2024.

764

765 Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. Towards controllable speech
 766 synthesis in the era of large language models: A survey. *arXiv preprint arXiv:2412.06602*, 2025.

767

768 Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yi-
 769 fan Yang, Zhikang Niu, Wenrui Liu, Fan Yu, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie
 770 Chen. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. *arXiv
 771 preprint arXiv:2504.12867*, 2025.

772

773 Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff
 774 Hwang, Dangna Li, Yuhao Wang, Julian Chan, Yuan Huang, Zhizheng Wu, and Mingbo Ma.
 775 Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *The Thir-
 776 teenth International Conference on Learning Representations*, pp. 1–24, 2025.

777

778 Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Jinchao Wang, and Haizhou Li.
 779 M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the
 780 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,
 781 pp. 5699–5710, 2022.

782

783 Zhixian Zhao, Xinfu Zhu, Xinsheng Wang, Shuiyuan Wang, Xuelong Geng, Wenjie Tian, and Lei
 784 Xie. Steering language model to stable speech emotion recognition via contextual perception and
 785 chain of thought. *arXiv preprint arXiv:2502.18186*, 2025.

786

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

789

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

793

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

798

799 **A APPENDIX A: CODE SNIPPETS FOR FINE-GRAINED EMOTION CONTROL**

800

801 For all three TTS models used in our study, i.e., F5-TTS (Chen et al., 2025), E2-TTS (Eskimez et al.,
 802 2024), and CosyVoice2 (Du et al., 2024), steering operations are implemented as hook functions.
 803 These hooks are registered either before or after the forward pass of the first residual stream in each
 804 DiT block. Code will be released upon acceptance.

805

806 **A.1 EMOTION CONVERSION AND INTERPOLATION**

807

808 For emotion conversion and interpolation, we steer the activation of the first residual stream in
 809 selected DiT blocks by registering a `forward_pre_hook` that modifies the inputs before they enter
 810 the linear residual stream module. The weighted steering vector, i.e., \hat{s}^l in Eq. 8, is stored in
 811 variable `steering_activations`. The steering intensity, i.e., α in Eq. 8, is controlled via `args.
 812 steering_strength`. The code for emotion conversion and interpolation is shown in Listing 1.

```

810
811
812     1 def act_steering_hook(block_idx, name=None):
813         2 """
814             3     Create a hook function for steering activations.
815             4 """
816
817         5
818         6 def hook(module, input_args):
819             7     if input_args and len(input_args) > 1:
820                 8         # Get the input
821                 9         (
822                     10            x,
823                     11            t,
824                     12            time,
825                     13            mask,
826                     14            rope,
827                     15            drop_audio_cond,
828                     16            drop_text,
829                     17            ref_audio_len,
830                     18            ) = input_args
831                     19             if (
832                         20                             not drop_audio_cond
833                         21                             ): # If drop_audio_cond is True, no manipulation of activation values.
834                         22                             step = int(time * 32) # time is a floating - point number between 0 and 1
835                         23                             act = steering_activations[block_idx // 5, step, :] # (1024)
836                         24                             act = act.unsqueeze(0).repeat(
837                             25                               ref_audio_len, 1
838                             26                             ) # (ref_audio_len, 1024)
839                         27                             act = act.unsqueeze(0) # (1, ref_audio_len, 1024)
840                         28                             act = act.to(x.device)
841
842                         29             # Normalize act to unit vector
843                         30             act = act / (act.norm(p=2) + 1e-8)
844
845                         31             pad_len = x.size(1) - act.size(1)
846                         32             pad_tensor = torch.zeros(
847                             33                               x.size(0),
848                             34                               pad_len,
849                             35                               x.size(2),
850                             36                               dtype=x.dtype,
851                             37                               device=x.device,
852                             38                         )
853                         39             act = torch.cat([act, pad_tensor], dim=1).to(x.dtype)
854
855                         40             # Save original norm for each sample in batch
856                         41             orig_norm = x.norm(p=2, dim=(1, 2), keepdim=True) # (B, 1, 1)
857
858                         42             x = x + args.steering_strength * act
859
860                         43             # Rescale x to have the same norm as original x
861                         44             new_norm = x.norm(p=2, dim=(1, 2), keepdim=True) + 1e-8
862                         45             x = x * (orig_norm / new_norm)
863
864                         46             return (
865                             47                               x,
866                             48                               t,
867                             49                               time,
868                             50                               mask,
869                             51                               rope,
870                             52                               drop_audio_cond,
871                             53                               drop_text,
872                             54                               ref_audio_len,
873                             55                         )
874
875         63     return hook
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1987
1988
1989
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2087
2088
2089
2089
2090
2091
2092
2093
2094
2095
2096
2097
2097
2098
2099
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2108
2109
2110
2111
2112
2113
2114
2115
2116
2116
2117
2118
2119
2119
2120
2121
2122
2123
2124
2125
2126
2127
2127
2128
2129
2129
2130
2131
2132
2133
2134
2135
2136
2137
2137
2138
2139
2139
2140
2141
2142
2143
2144
2145
2146
2147
2147
2148
2149
2149
2150
2151
2152
2153
2154
2155
2156
2156
2157
2158
2159
2159
2160
2161
2162
2163
2164
2165
2166
2166
2167
2168
2168
2169
2170
2171
2172
2173
2174
2175
2175
2176
2177
2177
2178
2179
2179
2180
2181
2182
2183
2184
2185
2185
2186
2187
2187
2188
2189
2189
2190
2191
2192
2193
2194
2195
2195
2196
2197
2197
2198
2199
2199
2200
2201
2202
2203
2204
2204
2205
2206
2206
2207
2208
2208
2209
2210
2211
2212
2213
2213
2214
2215
2215
2216
2217
2217
2218
2219
2219
2220
2221
2222
2223
2224
2225
2225
2226
2227
2227
2228
2229
2229
2230
2231
2232
2233
2234
2234
2235
2236
2236
2237
2238
2238
2239
2240
2240
2241
2242
2242
2243
2244
2244
2245
2246
2246
2247
2248
2248
2249
2250
2250
2251
2252
2252
2253
2254
2254
2255
2256
2256
2257
2258
2258
2259
2260
2260
2261
2262
2262
2263
2264
2264
2265
2266
2266
2267
2268
2268
2269
2270
2270
2271
2272
2272
2273
2274
2274
2275
2276
2276
2277
2278
2278
2279
2280
2280
2281
2282
2282
2283
2284
2284
2285
2286
2286
2287
2288
2288
2289
2290
2290
2291
2292
2292
2293
2294
2294
2295
2296
2296
2297
2298
2298
2299
2300
2300
2301
2302
2302
2303
2304
2304
2305
2306
2306
2307
2308
2308
2309
2310
2310
2311
2312
2312
2313
2314
2314
2315
2316
2316
2317
2318
2318
2319
2320
2320
2321
2322
2322
2323
2324
2324
2325
2326
2326
2327
2328
2328
2329
2330
2330
2331
2332
2332
2333
2334
2334
2335
2336
2336
2337
2338
2338
2339
2340
2340
2341
2342
2342
2343
2344
2344
2345
2346
2346
2347
2348
2348
2349
2350
2350
2351
2352
2352
2353
2354
2354
2355
2356
2356
2357
2358
2358
2359
2360
2360
2361
2362
2362
2363
2364
2364
2365
2366
2366
2367
2368
2368
2369
2370
2370
2371
2372
2372
2373
2374
2374
2375
2376
2376
2377
2378
2378
2379
2380
2380
2381
2382
2382
2383
2384
2384
2385
2386
2386
2387
2388
2388
2389
2390
2390
2391
2392
2392
2393
2394
2394
2395
2396
2396
2397
2398
2398
2399
2400
2400
2401
2402
2402
2403
2404
2404
2405
2406
2406
2407
2408
2408
2409
2410
2410
2411
2412
2412
2413
2414
2414
2415
2416
2416
2417
2418
2418
2419
2420
2420
2421
2422
2422
2423
2424
2424
2425
2426
2426
2427
2428
2428
2429
2430
2430
2431
2432
2432
2433
2434
2434
2435
2436
2436
2437
2438
2438
2439
2440
2440
2441
2442
2442
2443
2444
2444
2445
2446
2446
2447
2448
2448
2449
2450
2450
2451
2452
2452
2453
2454
2454
2455
2456
2456
2457
2458
2458
2459
2460
2460
2461
2462
2462
2463
2464
2464
2465
2466
2466
2467
2468
2468
2469
2470
2470
2471
2472
2472
2473
2474
2474
2475
2476
2476
2477
2478
2478
2479
2480
2480
2481
2482
2482
2483
2484
2484
2485
2486
2486
2487
2488
2488
2489
2490
2490
2491
2492
2492
2493
2494
2494
2495
2496
2496
2497
2498
2498
2499
2500
2500
2501
2502
2502
2503
2504
2504
2505
2506
2506
2507
2508
2508
2509
2510
2510
2511
2512
2512
2513
2514
2514
2515
2516
2516
2517
2518
2518
2519
2520
2520
2521
2522
2522
2523
2524
2524
2525
2526
2526
2527
2528
2528
2529
2530
2530
2531
2532
2532
2533
2534
2534
2535
2536
2536
2537
2538
2538
2539
2540
2540
2541
2542
2542
2543
2544
2544
2545
2546
2546
2547
2548
2548
2549
2550
2550
2551
2552
2552
2553
2554
2554
2555
2556
2556
2557
2558
2558
2559
2560
2560
2561
2562
2562
2563
2564
2564
2565
2566
2566
2567
2568
2568
2569
2570
2570
2571
2572
2572
2573
2574
2574
2575
2576
2576
2577
2578
2578
2579
2580
2580
2581
2582
2582
2583
2584
2584
2585
2586
2586
2587
2588
2588
2589
2590
2590
2591
2592
2592
2593
2594
2594
2595
2596
2596
2597
2598
2598
2599
2600
2600
2601
2602
2602
2603
2604
2604
2605
2606
2606
2607
2608
2608
2609
2610
2610
2611
2612
2612
2613
2614
2614
2615
2616
2616
2617
2618
2618
2619
2620
2620
2621
2622
2622
2623
2624
2624
2625
2626
2626
2627
2628
2628
2629
2630
2630
2631
2632
2632
2633
2634
2634
2635
2636
2636
2637
2638
2638
2639
2640
2640
2641
2642
2642
2643
2644
2644
2645
2646
2646
2647
2648
2648
2649
2650
2650
2651
2652
2652
2653
2654
2654
2655
2656
2656
2657
2658
2658
2659
2660
2660
2661
2662
2662
2663
2664
2664
2665
2666
2666
2667
2668
2668
2669
2670
2670
2671
2672
2672
2673
2674
2674
2675
2676
2676
2677
2678
2678
2679
2680
2680
2681
2682
268
```

```

864
865     """
866     Create a hook function for emotion erasure.
867     """
868
869     def hook(module, input_args):
870         if input_args and len(input_args) > 1:
871             (
872                 x, # (B, L, 1024)
873                 t,
874                 time,
875                 mask,
876                 rope,
877                 drop_audio_cond,
878                 drop_text,
879                 ref_audio_len,
880             ) = input_args
881         if (
882             not drop_audio_cond
883         ):
884             step = int(time * 32)
885             act = steering_activations[block_idx // 5, step, :] # (1024)
886             act = act.to(x.dtype).to(x.device)
887
888             # Normalize act to unit vector
889             act = act / (act.norm(p=2) + 1e-8)
890
891             projection = torch.matmul(
892                 act.unsqueeze(0), # (1, 1024)
893                 x[:, :ref_audio_len, :].transpose(
894                     1, 2
895                     ), # (B, ref_audio_len, 1024)
896                 .transpose(
897                     1, 2
898                     ) # (B, ref_audio_len, 1)
899
900             pad_len = x.size(1) - ref_audio_len
901             padded_projection = torch.cat(
902                 [
903                     projection,
904                     torch.zeros(
905                         x.size(0),
906                         pad_len,
907                         1,
908                         dtype=x.dtype,
909                         device=x.device,
910                     ),
911                 ],
912                 dim=1,
913             )
914
915             act = act.unsqueeze(0).repeat(
916                 ref_audio_len, 1
917             ) # (ref_audio_len, 1024)
918             act = act.unsqueeze(0) # (1, ref_audio_len, 1024)
919
920             pad_tensor = torch.zeros(
921                 x.size(0),
922                 pad_len,
923                 x.size(2),
924                 dtype=x.dtype,
925                 device=x.device,
926             )
927             act = torch.cat([act, pad_tensor], dim=1)
928
929             # Save original norm for each sample in batch
930             orig_norm = x.norm(p=2, dim=(1, 2), keepdim=True) # (B, 1, 1)
931
932             x = x - args.erasing_strength * padded_projection * act
933
934             # Rescale x to have the same norm as original x
935             new_norm = x.norm(p=2, dim=(1, 2), keepdim=True) + 1e-8
936             x = x * (orig_norm / new_norm)
937
938             return (
939                 x,
940                 t,
941                 time,
942                 mask,
943                 rope,
944                 drop_audio_cond,
945                 drop_text,
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982

```

```

918     ref_audio_len,
919     )
920
921     return hook
922
923

```

A.3 EMOTION REPLACEMENT

For emotion replacement, we steer the activation of the first residual stream in selected DiT blocks by registering a `forward_pre_hook`, which modifies the inputs before they enter the linear residual stream module. The weighted steering vectors for emotion 1 and emotion 2, i.e., \hat{s}_{emo1}^l and \hat{s}_{emo2}^l in Eq. 10, are stored in the variable `steering_activations_1` and variable `steering_activations_2`, respectively. The erasing and steering intensities, i.e., β and α in Eq. 10, are controlled by variables `args.erasing_strength` and `args.steering_strength`, respectively. The implementation of emotion replacement is provided in Listing 3.

Listing 3: Emotion Replacement Hook

```

932
933     def act_replacement_hook(block_idx, name=None):
934         """
935             Create a hook function for emotion replacement.
936         """
937
938     def hook(module, input_args):
939         if input_args and len(input_args) > 1:
940             (
941                 x, # (B, L, 1024)
942                 t,
943                 time,
944                 mask,
945                 rope,
946                 drop_audio_cond,
947                 drop_text,
948                 ref_audio_len,
949             ) = input_args
950             if (
951                 not drop_audio_cond
952             ):
953                 step = int(time * 32)
954                 act1 = steering_activations_1[block_idx // 5, step, :] # (1024)
955                 act1 = act1.to(x.dtype).to(x.device)
956
957                 # Normalize act to unit vector
958                 act1 = act1 / (act1.norm(p=2) + 1e-8)
959
960                 projection = torch.matmul(
961                     act1.unsqueeze(0), # (1, 1024)
962                     x[:, :ref_audio_len, :].transpose(
963                         1, 2
964                         ), # (B, ref_audio_len, 1024)
965                     .transpose(
966                         1, 2
967                         ) # (B, ref_audio_len, 1)
968
969                 pad_len = x.size(1) - ref_audio_len
970                 padded_projection = torch.cat(
971                     [
972                         projection,
973                         torch.zeros(
974                             x.size(0),
975                             pad_len,
976                             1,
977                             dtype=x.dtype,
978                             device=x.device,
979                         ),
980                         ],
981                         dim=1,
982                     )
983
984                 act1 = act1.unsqueeze(0).repeat(
985                     ref_audio_len, 1
986                     ) # (ref_audio_len, 1024)
987                 act1 = act1.unsqueeze(0) # (1, ref_audio_len, 1024)
988
989                 pad_tensor = torch.zeros(
990                     x.size(0),
991                     pad_len,
992                     )
993
994

```

A.4 MULTIPLE EMOTION STEERING

For multiple emotion steering, we steer the activation of the first residual stream in selected DiT blocks by registering a `forward_pre_hook` that modifies the inputs before they enter the linear residual stream module. The weighted steering vectors for emotions 1 and 2, i.e., \hat{s}_{emo1}^l and \hat{s}_{emo2}^l in Eq. 11, are stored in variable `steering_activations_1` and variable `steering_activations_2`, respectively. The steering intensities for the two emotions, i.e., α_1 and α_2 in Eq. 11, are controlled via `args.steering_strength_1` and `steering_strength_2`, respectively. The code for multiple emotion steering is shown in Listing 4.

Listing 4: Multiple Emotion Steering Hook

```
1  def act_multi_steering_hook(block_idx, name=None):
2      """
3          Create a hook function for multiple emotion steering.
4      """
5
6      def hook(module, input_args):
7          if input_args and len(input_args) > 1:
8              # Get the input
9              (
10                  x,
11                  t,
12                  time,
13                  mask,
14                  rope,
15                  drop_audio_cond,
```

```

1026
1027     drop_text,
1028     ref_audio_len,
1029 ) = input_args
1030 if (
1031     not drop_audio_cond
1032 ) :
1033     step = int(time * 32)
1034     act1 = steering_activations_1[block_idx // 5, step, :] # (1024)
1035     act1 = act1.unsqueeze(0).repeat(
1036         ref_audio_len, 1
1037     ) # (ref_audio_len, 1024)
1038     act1 = act1.unsqueeze(0) # (1, ref_audio_len, 1024)
1039     act1 = act1.to(x.device)
1040
1041     # Normalize act to unit vector
1042     act1 = act1 / (act1.norm(p=2) + 1e-8)
1043
1044     pad_len = x.size(1) - act1.size(1)
1045     pad_tensor = torch.zeros(
1046         x.size(0),
1047         pad_len,
1048         x.size(2),
1049         dtype=x.dtype,
1050         device=x.device,
1051     )
1052     act1 = torch.cat([act1, pad_tensor], dim=1).to(x.dtype)
1053
1054     act2 = steering_activations_2[block_idx // 5, step, :] # (1024)
1055     act2 = act2.unsqueeze(0).repeat(
1056         ref_audio_len, 1
1057     ) # (ref_audio_len, 1024)
1058     act2 = act2.unsqueeze(0) # (1, ref_audio_len, 1024)
1059     act2 = act2.to(x.device)
1060
1061     # Normalize act to unit vector
1062     act2 = act2 / (act2.norm(p=2) + 1e-8)
1063
1064     pad_len = x.size(1) - act2.size(1)
1065     pad_tensor = torch.zeros(
1066         x.size(0),
1067         pad_len,
1068         x.size(2),
1069         dtype=x.dtype,
1070         device=x.device,
1071     )
1072     act2 = torch.cat([act2, pad_tensor], dim=1).to(x.dtype)
1073
1074     # Save original norm for each sample in batch
1075     orig_norm = x.norm(p=2, dim=(1, 2), keepdim=True) # (B, 1, 1)
1076
1077     x = x + args.steering_strength_1 * act1 + args.steering_strength_2 * act2
1078
1079     # Rescale x to have the same norm as original x
1080     new_norm = x.norm(p=2, dim=(1, 2), keepdim=True) + 1e-8
1081     x = x * (orig_norm / new_norm)
1082
1083     return (
1084         x,
1085         t,
1086         time,
1087         mask,
1088         rope,
1089         drop_audio_cond,
1090         drop_text,
1091         ref_audio_len,
1092     )
1093
1094     return hook

```

B APPENDIX B: DATASET CONSTRUCTION

To ensure the effectiveness of the steering vectors, we curate an emotional speech dataset by collecting and filtering audio samples with clearly distinguishable emotional tones from multiple existing corpora, including MSP-Podcast (Lotfian & Busso, 2017), IEMOCAP (Busso et al., 2008), RAVDESS (Livingstone & Russo, 2018), CREMA-D (Cao et al., 2014), TESS (Pichora-Fuller & Dupuis, 2020), SAVEE (Jackson & Haq, 2014), ASVP-ESD (Landry et al., 2020), CASIA (CASIA,

Table 4: The details of model configuration for activation steering.

Model	# Layers	# CFM Steps	Steered Layers	Steered Activations in Each Layer
F5-TTS	22	32	Every 5 layers starting from layer 1	The first residual stream
E2-TTS	8	32	Every 3 layers starting from layer 1	The first residual stream
CosyVoice2	56	10	Every 5 layers starting from layer 1	The first residual stream

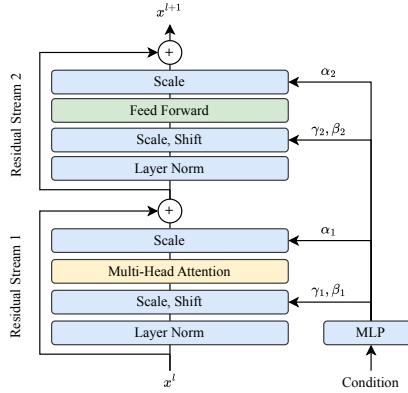


Figure 8: A DiT block in a CFM-based TTS model.

2023), M3ED (Zhao et al., 2022), ESD (Zhou et al., 2022), and Emo-Emilia (Zhao et al., 2025). The quality filtering process involves the following steps:

1. We use `librosa` (McFee et al., 2015) to remove utterances that are either too short (<2s) or too long (>20s).
2. We further filter out samples exhibiting excessive silence (>30%) or a low signal-to-noise ratio (SNR) (<10dB), also using `librosa`.
3. We use a SER model, emotion2Vec, to eliminate samples with low recognition confidence (<0.6), retaining only those predicted as ground truth labels.
4. Finally, we perform a manual inspection on 50% of the data to ensure overall dataset quality.

The resulting dataset covers a broad range of speakers, emotions, and speaking styles, providing a robust foundation for learning and evaluating fine-grained emotion steering in text-to-speech synthesis. Data and code will be released upon acceptance.

C APPENDIX C: CONFIGURATIONS

C.1 MODEL CONFIGURATIONS

We steer three pretrained conditional flow matching (CFM)-based TTS models, i.e., F5-TTS, E2-TTS, and CosyVoice2, in our main experiments. As illustrated in Fig.8, we apply steering to the first residual stream at each layer of these models. The detailed configurations of the models are provided in Table 4. emotion2vec and SenseVoice checkpoints are downloaded from their official repos¹².

C.2 HARDWARE AND SOFTWARE CONFIGURATIONS

All experiments were conducted on a server equipped with 8× NVIDIA RTX 6000 Ada GPUs (48GB each) and 2× Intel(R) Xeon(R) Platinum 8375C CPUs (2.9GHz, 32 cores each), with a total of

¹https://huggingface.co/emotion2vec/emotion2vec_plus_large

²<https://huggingface.co/FunAudioLLM/SenseVoiceSmall>

1134 256GB of RAM. The operating system is Ubuntu 20.04.6 LTS. All code was executed in Conda
 1135 environments. The relevant software libraries and frameworks for each model (F5-TTS, E2-TTS,
 1136 CosyVoice2) are described in their GitHub repositories³⁴⁵.
 1137

1138 **C.3 RATERS' INFORMATION AND INTER-RATER RELIABILITY**
 1139

1140 30 raters participated in the human evaluation for our main experiments. All raters were either mas-
 1141 ter's or PhD students. We adopt Percent Agreement (Gwet, 2014) as a more appropriate measure of
 1142 reliability for the human evaluation of synthesized samples. The results show a Top-2 Box Agree-
 1143 ment of 88.1%, meaning that the vast majority of ratings fell within the 4 (Good) or 5 (Excellent)
 1144 categories. Furthermore, the raters demonstrated high consistency in their qualitative judgment, with
 1145 negligible divergence on the acceptable range.
 1146

1147 **D APPENDIX D: OBJECTIVE EVALUATION METRICS**
 1148

1149 **D.1 NATURALNESS MEAN OPINION SCORE**
 1150

1151 The Naturalness Mean Opinion Score (N-MOS) evaluates the perceived naturalness of synthesized
 1152 speech on a 5-point Likert scale. Participants are asked to rate each utterance based solely on how
 1153 natural and human-like it sounds, regardless of its emotional expressiveness or content accuracy.
 1154 The scale is defined as follows:
 1155

- 1156 • 5 — Completely natural: indistinguishable from real human speech.
 1157
- 1158 • 4 — Mostly natural: minor artifacts but still sounds largely human.
 1159
- 1160 • 3 — Moderately natural: noticeable synthetic artifacts, but intelligible.
 1161
- 1162 • 2 — Barely natural: speech is intelligible but sounds clearly robotic.
 1163
- 1164 • 1 — Not natural at all: heavily distorted or unnatural-sounding.
 1165

1166 Each utterance is evaluated by multiple annotators, and the final N-MOS is computed as the average
 1167 score across all evaluations.
 1168

1169 **D.2 EMOTION INTERPOLATION MEAN OPINION SCORE**
 1170

1171 The Emotion Interpolation Mean Opinion Score (EI-MOS) assesses the system's ability to smoothly
 1172 interpolate between two emotional styles. For each interpolation sequence (e.g., neutral → angry),
 1173 raters listen to a series of utterances generated with gradually increasing emotion intensity and judge
 1174 how naturally and smoothly the emotional change is conveyed. Raters are instructed to focus on the
 1175 continuity and consistency of emotional expression rather than the naturalness or correctness of
 1176 individual utterances. The scoring scale is as follows:
 1177

- 1178 • 5 — Emotion transition is smooth and realistic throughout the sequence.
 1179
- 1180 • 4 — Emotion changes are mostly smooth, with minor inconsistencies.
 1181
- 1182 • 3 — Some transitions feel abrupt or inconsistent.
 1183
- 1184 • 2 — Transitions are disjointed, or emotion interpolation feels unnatural.
 1185
- 1186 • 1 — No meaningful emotion interpolation perceived.
 1187

1188 Each interpolation sequence is rated by multiple annotators, and the EI-MOS is reported as the
 1189 average of all scores.
 1190

³<https://github.com/SWivid/F5-TTS>

⁴<https://github.com/lucidrains/e2-tts-pytorch>

⁵<https://github.com/FunAudioLLM/CosyVoice>

1188 D.3 EMOTION ERASURE MEAN OPINION SCORE
1189

1190 The Emotion Erasure Mean Opinion Score (EE-MOS) evaluates the effectiveness of emotion re-
1191 moval from synthesized speech. Specifically, it measures how well the target emotion has been
1192 erased, with the desired outcome being emotionally neutral and natural-sounding speech. Annota-
1193 tors are instructed to assess whether the emotional content of the original utterance has been suc-
1194 cessfully suppressed or removed. The rating is based on a 5-point scale:

- 1196 • 5 — *Emotion fully removed*: The target emotion is completely erased; the speech sounds
1197 emotionally neutral and natural, with no detectable emotional cues.
- 1198 • 4 — *Emotion mostly removed*: Only faint traces of the original emotion remain; the speech
1199 is close to neutral.
- 1200 • 3 — *Emotion partially removed*: The emotional intensity is reduced, but the target emotion
1201 is still clearly noticeable.
- 1202 • 2 — *Emotion barely removed*: The emotional expression remains strong; only minimal
1203 reduction is observed.
- 1204 • 1 — *Emotion not removed*: The original emotional tone persists fully or is even uninten-
1205 tionally enhanced.

1209 Each utterance is evaluated independently by multiple listeners, and the EE-MOS is calculated as
1210 the average of all individual scores. A higher EE-MOS indicates a more effective erasure of the
1211 target emotion.

1214 Table 5: Unguaranteed reproduced results of open-source baselines.
1215

1216	Method	Conversion ($\alpha = 2.0$)			Interpolation		Erasure ($\beta = 2.5$)		
		WER(\downarrow)	S-SIM(\uparrow)	E-SIM(\uparrow)	N-MOS(\uparrow)	EI-MOS(\uparrow)	E-SIM(\uparrow)	EE-MOS(\uparrow)	
emotion2vec / SenseVoice									
In-distribution evaluation on MSP-Podcast (25% en) and ESD (25% en, 50% zh)									
1220	Label-based*	EmoSphere++	37.29	0.21	0.14 / 0.11 _{avg=0.125}	2.14 _{±0.91}	2.41 _{±0.83}	-	
1221		EmoDubber	65.93	0.16	0.08 / 0.05 _{avg=0.065}	1.07 _{±0.94}	1.13 _{±1.02}	-	
1222	Description -based*	EmoVoice	5.31	0.48	0.22 / 0.19 _{avg=0.205}	3.26 _{±1.22}	-	-	
1223		CosyVoice2	2.71	0.69	0.23 / 0.25 _{avg=0.240}	3.66 _{±1.17}	-	-	
1224	Unsteered	F5-TTS	2.14	0.66	0.07 / 0.04 _{avg=0.055}	3.79 _{±0.89}	-	0.03 / 0.05 _{avg=0.040}	
1225		E2-TTS	2.71	0.64	0.05 / 0.08 _{avg=0.065}	3.51 _{±0.94}	-	1.21 _{±1.17}	
1226	EmoSteer-TTS# (Ours)	+ F5-TTS	2.79	0.64	0.29 / 0.26 _{avg=0.275}	3.29 _{±1.05}	4.00 _{±0.89}	0.27 / 0.25 _{avg=0.260}	
1227		+ E2-TTS	3.28	0.59	0.28 / 0.28 _{avg=0.280}	3.31 _{±0.97}	3.38 _{±1.09}	0.24 / 0.26 _{avg=0.250}	
1228		+ CosyVoice2	2.83	0.65	0.26 / 0.29 _{avg=0.275}	3.65 _{±1.08}	3.56 _{±1.15}	0.26 / 0.25 _{avg=0.255}	
Cross-datasets (OOD) evaluation on EMNS (25% en) and SeedTT test sets (25% en, 50% zh)									
1229	EmoSteer-TTS# (Ours)	+ F5-TTS	2.65	0.65	0.25 / 0.27 _{avg=0.260}	3.58 _{±1.04}	3.46 _{±1.08}	0.25 / 0.22 _{avg=0.235}	
1230		+ E2-TTS	3.41	0.55	0.26 / 0.25 _{avg=0.255}	3.44 _{±1.07}	3.50 _{±0.97}	0.24 / 0.27 _{avg=0.255}	
1231		+ CosyVoice2	2.86	0.66	0.28 / 0.25 _{avg=0.265}	3.49 _{±1.01}	3.48 _{±1.27}	0.23 / 0.21 _{avg=0.220}	

1232 *: Training-based, #: Training-free, -: Unsupported operation.

1233 The top three results are indicated in boldface. Unsteered backbones are shown in gray for reference.

1234
1235 E APPENDIX E: REPRODUCED BASELINE RESULTS
1236

1239 This section recomputes baselines under a controlled protocol, using the same text prompts, refer-
1240 ence speeches, and evaluation scripts as in the in-distribution evaluation on MSP-Podcast and ESD.
1241 We report results only for open-source methods, as the reproduced quality cannot be guaranteed.
Therefore, the results in Table 5 are provided for reference only.

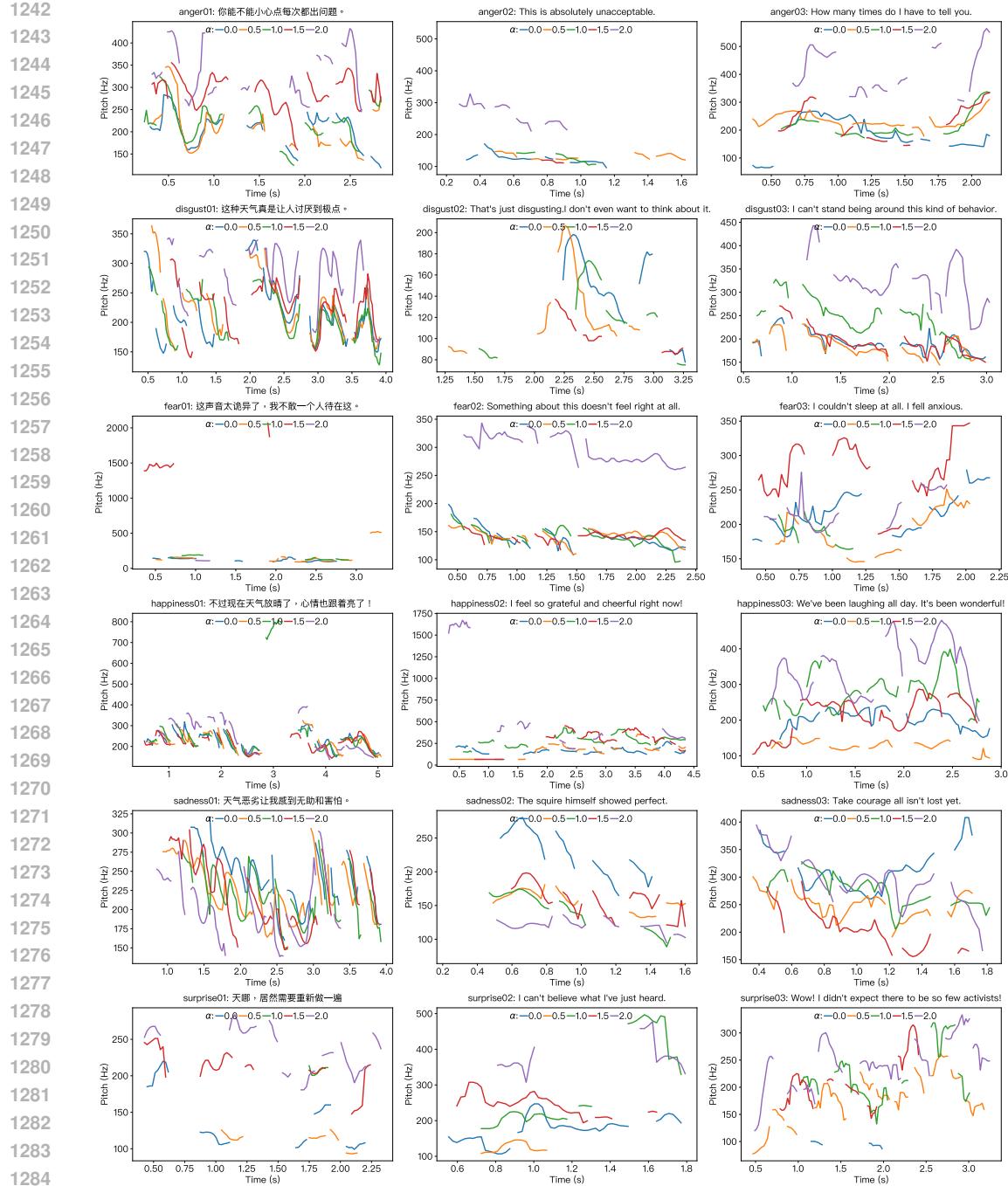


Figure 9: Visualizations of F0 contours in emotion interpolation. From left to right: F5-TTS, E2-TTS, and CosyVoice2. From top to bottom: anger, disgust, fear, happiness, sadness, and surprise. All the synthesized speech samples are interpolated between neutrality ($\alpha=0$) and a target emotion ($\alpha=2$).

F APPENDIX F: VISUALIZATION OF F0 CONTOURS

F.1 EMOTION INTERPOLATION

In this subsection, we present additional visualizations of F0 contours to illustrate the fine-grained and continuous emotion interpolation capabilities of the proposed EmoSteer-TTS. As shown in

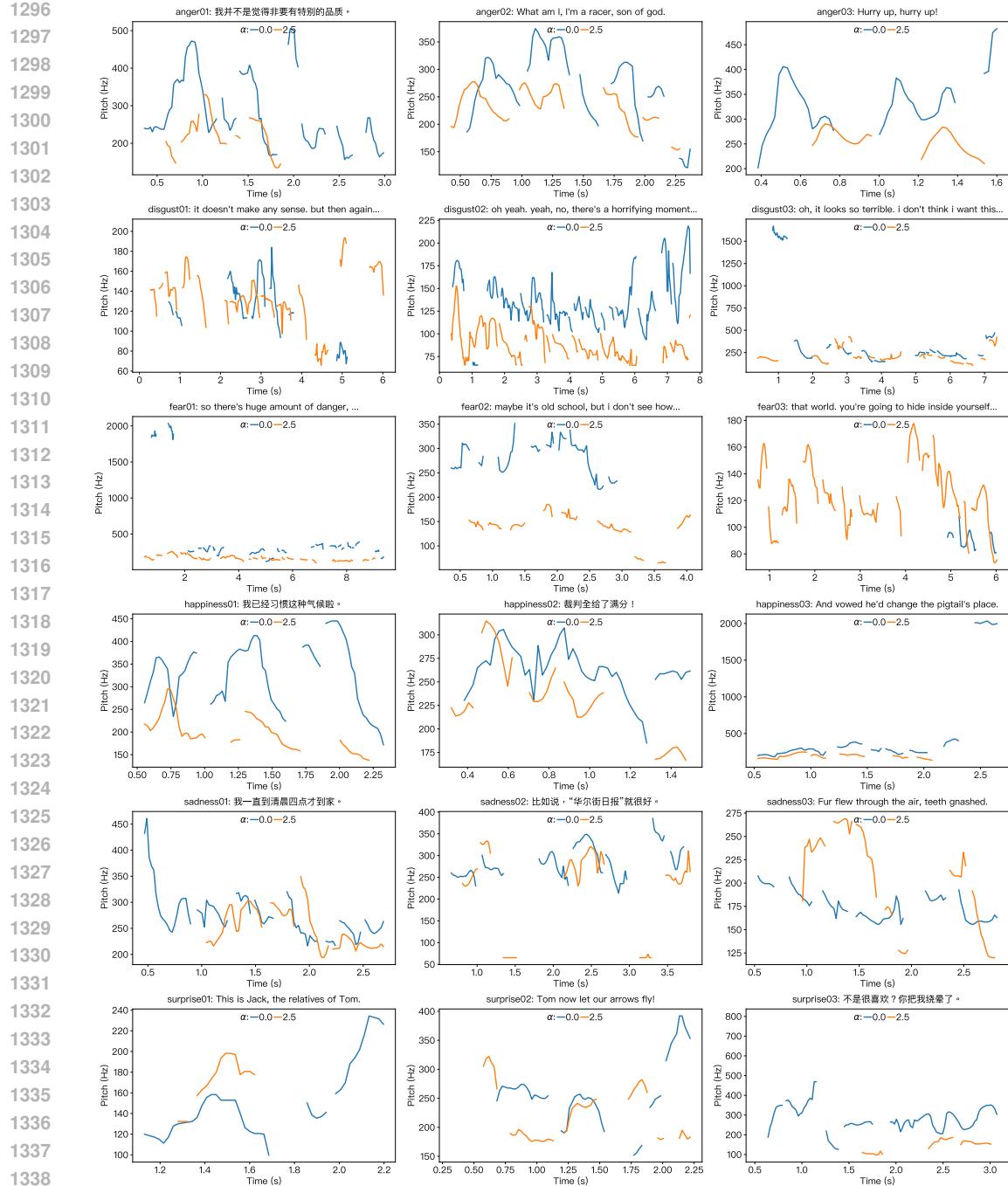


Figure 10: Visualizations of F0 contours in emotion erasure. From left to right: F5-TTS, E2-TTS, and CosyVoice2. From top to bottom: anger, disgust, fear, happiness, sadness, and surprise. All the synthesized speech samples are emotionally erased from a target emotion ($\beta=0$) towards neutrality ($\beta=2.5$).

Fig. 9, voices with angrier, happier, or more surprised tones tend to exhibit higher pitch, while sadder tones are generally associated with lower pitch. In contrast, pitch variations in disgust and fear interpolation show no clear monotonic trend, which we attribute to the fact that these emotions are more closely tied to semantic content than to acoustic characteristics.

1350 Table 6: Emotion conversion ($\alpha = 2.0$) on F5-TTS using emotion2vec for token probing.
1351

Steering Corpus Composition	WER↓	S-SIM↑	E-SIM↑ emotion2vec / SenseVoice	UTMOS↑
3 datasets (302 samples): IEMOCAP (Busso et al., 2008), SAVEE (Jackson & Haq, 2014), CREMA-D (Cao et al., 2014)	2.91	0.59	0.18 / 0.15	3.42
7 datasets (3,021 samples): + MSP-Podcast (Lotfian & Busso, 2017), RAVDESS (Livingstone & Russo, 2018), TESS (Pichora-Fuller & Dupuis, 2020), ASVP-ESD (Landry et al., 2020)	2.84	0.64	0.21 / 0.17	3.51
11 datasets (6,900 samples): + CASIA (CASIA, 2023), M3ED (Zhao et al., 2022), ESD (Zhou et al., 2022), Emo-Emilia (Zhao et al., 2025)	2.79	0.64	0.29 / 0.26	3.49

1366 F.2 EMOTION ERASURE

1367 In this subsection, we present additional F0 contour visualizations to demonstrate the emotion erasure
1368 capability of the proposed EmoSteer-TTS. As shown in Fig. 10, the pitch contours of angry,
1369 disgusted, happy, and surprised voices become noticeably flatter after emotion erasure, indicating
1370 a calmer prosodic pattern. In contrast, the changes in pitch for fear and sadness are more diverse
1371 and less predictable. This variability may stem from the fact that fear can be expressed through
1372 multiple vocal styles, such as a low, trembling voice or a high-pitched scream, making it difficult for
1373 pitch alone to capture the underlying emotional shift. Similarly, sadness may manifest as either soft
1374 weeping or loud crying, resulting in inconsistent pitch patterns that do not reliably reflect emotional
1375 intensity.

1376 In both emotion interpolation and erasure, F0 contours capture only a partial aspect of human emotional
1377 perception, as pitch alone cannot fully convey complex emotional nuances. Therefore, we
1378 encourage readers to listen to the audio samples available on our demo page.

1381 G APPENDIX G: THE USE OF LLMs

1382 Some portions of this paper were paraphrased or refined with the assistance of ChatGPT and Gemini.
1383 No content was directly generated by LLMs.

1387 H APPENDIX H: ADDITIONAL ANALYSIS OF EMOTION STEERING
1388 DYNAMICS

1390 H.1 SENSITIVITY TO STEERING CORPUS COMPOSITION

1392 We conducted an additional ablation study to further examine the sensitivity to the composition of
1393 the steering corpus. The entire corpus was constructed from 11 datasets, resulting in a huge number
1394 of possible combinations. It is infeasible to evaluate all of them exhaustively. A reasonable strat-
1395 egic is to combine the datasets in chronological order, which may partially reflect overall recording
1396 quality as recording devices and speech processing technology improve over time. Therefore, we
1397 conduct the ablation using three chronological dataset groups and report WER, S-SIM, E-SIM, and
1398 UTMOS on the F5-TTS backbone only.

1399 As shown in Tables 6 and 7, WER, S-SIM, and UTMOS remain largely stable across different
1400 steering corpus sizes, indicating that general speech quality and semantic fidelity are minimally
1401 affected. In contrast, E-SIM consistently increases with the number of datasets, suggesting that
1402 emotion similarity benefits from larger and more diverse steering corpora. Overall, these results
1403 indicate that dataset quantity primarily influences emotional control, while other aspects of synthesis
are largely insensitive to corpus composition.

1404
1405 Table 7: Emotion erasure ($\beta = 2.5$) on F5-TTS using emotion2vec for token probing.
1406
1407

Steering Corpus Composition	WER↓	S-SIM↑	E-SIM↑ emotion2vec / SenseVoice	UTMOS↑
3 datasets (302 samples): IEMOCAP (Busso et al., 2008), SAVEE (Jackson & Haq, 2014), CREMA-D (Cao et al., 2014)	2.88	0.61	0.07 / 0.05	3.51
7 datasets (3,021 samples): + MSP-Podcast (Lotfian & Busso, 2017), RAVDESS (Livingstone & Russo, 2018), TESS (Pichora-Fuller & Dupuis, 2020), ASVP-ESD (Landry et al., 2020)	2.94	0.58	0.18 / 0.12	3.68
11 datasets (6,900 samples): + CASIA (CASIA, 2023), M3ED (Zhao et al., 2022), ESD (Zhou et al., 2022), Emo-Emilia (Zhao et al., 2025)	2.81	0.63	0.26 / 0.25	3.55

1420
1421 Table 8: The Pearson correlation coefficients between the E-SIM (emotion2vec/SenseVoice) scores
1422 and the N-MOS and EE-MOS (emotion2vec is used for token probing).
1423
1424

	E-SIM (emotion2vec)	E-SIM (SenseVoice)
N-MOS (Conversion, $\alpha = 2.0$)	-0.78	0.12
EE-MOS (Erasue, $\beta = 2.5$)	0.47	-0.08

1425
1426
1427 H.2 CORRELATION OF E-SIM METRICS WITH N-MOS AND EE-MOS

1428
1429 We report the Pearson correlation coefficients between the E-SIM (emotion2vec/SenseVoice) scores
1430 and the N-MOS, EE-MOS ratings for emotion conversion and erasure in our main experiments,
1431 respectively. As shown in Table 8, the E-SIM computed with emotion2vec exhibits a clear and
1432 consistent trend: it is negatively correlated with N-MOS (-0.78), indicating that stronger steering
1433 inevitably leads to noticeable degradation in naturalness. At the same time, it is positively correlated
1434 with EE-MOS (+0.47), suggesting that a larger E-SIM (more neutral) corresponds to more success-
1435 ful emotion erasure, as perceived by human raters. This confirms the expected trade-off between
1436 emotion controllability and naturalness.

1437
1438 In contrast, the correlations obtained using SenseVoice show almost no relationship with either
1439 N-MOS (+0.12) or EE-MOS (-0.08). We attribute this inconsistency to a mismatch between the
1440 emotion space captured by SenseVoice and that encoded by emotion2vec, which is also used in our
1441 token-probing framework.

1442
1443 H.3 CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS OF SUBJECTIVE EVALUATION

1444
1445 For the in-distribution evaluation in Table 1, the overall averaged N-MOS, EI-MOS, and EE-MOS
1446 across the three backbones, along with their corresponding confidence intervals, are summarized
1447 in Table 9. These results indicate that the naturalness of the synthesized speech, the interpolation
1448 capability, and the emotion erasure effectiveness of our method are consistently perceived by human
1449 raters as “Good” or above.

1450
1451 We conduct significance tests using the N-MOS and EI-MOS ratings from 30 raters, comparing
1452 our method with the strongest label-based baselines. We focus on these baselines because they

1453
1454 Table 9: Confidence Intervals of Subjective Evaluation

Metric	Averaged	Confidence Interval
N-MOS	3.42	95% of [3.38, 3.46]
EI-MOS	3.65	95% of [3.61, 3.69]
EE-MOS	3.86	95% of [3.82, 3.90]

1458 Table 10: Steering strength α vs. N-MOS and WER for F5-TTS.
1459

α	0.00	0.25	0.50	0.75	1.00	1.25	1.50	2.00	2.50	3.00
N-MOS (Anger)	4.27	4.24	4.25	4.18	4.02	3.93	3.64	3.41	2.60	2.15
N-MOS (Disgust)	4.19	4.20	4.08	4.02	3.86	3.65	3.37	3.16	2.29	2.08
N-MOS (Fear)	4.32	4.22	4.13	3.92	3.68	3.42	3.37	3.28	2.36	1.93
N-MOS (Happiness)	4.25	4.13	4.02	3.97	3.72	3.51	3.36	3.37	1.66	1.57
N-MOS (Sadness)	4.18	4.23	4.15	4.06	3.92	3.70	3.59	3.53	2.41	2.01
N-MOS (Surprise)	4.22	4.16	4.08	3.91	3.78	3.49	3.35	3.37	1.80	1.59
WER (Anger)	2.64	2.71	2.69	2.83	2.68	2.54	2.62	2.75	15.27	26.14
WER (Disgust)	2.81	2.47	2.92	2.58	3.11	2.73	3.05	2.66	14.83	27.42
WER (Fear)	2.55	3.18	2.69	3.04	2.88	2.41	3.22	2.79	16.44	24.91
WER (Happiness)	2.93	2.62	2.85	2.50	3.07	3.29	2.74	3.18	13.97	28.33
WER (Sadness)	2.49	2.88	3.15	2.73	2.60	3.18	2.57	3.11	15.62	25.40
WER (Surprise)	3.12	2.59	2.48	3.26	2.74	2.95	3.31	2.63	14.21	29.08

1472 Table 11: Steering strength α vs. N-MOS and WER for E2-TTS.
1473

α	0.00	0.25	0.50	0.75	1.00	1.25	1.50	2.00	2.50	3.00
N-MOS (Anger)	4.31	4.26	4.22	4.14	4.05	3.90	3.63	3.48	2.71	1.26
N-MOS (Disgust)	4.27	4.22	4.05	3.99	3.82	3.69	3.33	3.12	3.04	1.94
N-MOS (Fear)	4.34	4.20	4.16	3.88	3.71	3.39	3.41	3.24	2.97	2.17
N-MOS (Happiness)	4.23	4.15	3.98	4.00	3.63	3.56	3.31	3.35	2.62	1.85
N-MOS (Sadness)	4.20	4.21	4.18	4.04	3.90	3.73	3.60	3.50	2.98	2.00
N-MOS (Surprise)	4.25	4.12	4.11	3.89	3.81	3.46	3.37	3.32	2.63	1.58
WER (Anger)	3.24	3.21	3.19	3.33	3.18	3.04	3.12	3.25	15.77	25.62
WER (Disgust)	3.09	3.18	3.15	3.28	3.20	3.16	3.33	3.32	15.48	24.45
WER (Fear)	3.16	3.23	3.20	3.35	3.19	3.07	3.11	3.26	11.63	25.90
WER (Happiness)	3.31	3.20	3.08	3.21	3.37	2.95	3.00	3.04	19.24	35.60
WER (Sadness)	3.15	3.22	3.19	3.34	3.21	3.08	3.14	3.28	17.70	29.00
WER (Surprise)	3.42	3.19	3.26	3.40	3.18	3.16	3.21	3.13	16.93	29.70

1489 provide adjustable emotion intensity control, whereas description-based methods neither support
1490 emotion interpolation nor emotion erasure. Specifically, for N-MOS, we compare ‘‘EmoSteer-
1491 TTS+CosyVoice2’’ against HED-TTS, and for EI-MOS, we compare ‘‘EmoSteer-TTS+F5-TTS’’
1492 against EmoSphere++.

1493 A two-sided t-test indicates that our method significantly outperforms the baselines, with p-values
1494 of $0.01483 < 0.05$ for N-MOS and $0.00732 < 0.01$ for EI-MOS. These results demonstrate that
1495 our approach not only preserves naturalness but also more effectively conveys the intended emotion
1496 intensity, validating the advantages of our emotion-steering mechanism.

1498 H.4 TRADE-OFF BETWEEN α AND WER/N-MOS

1500 We have already reported the E-SIM variations in 4(b) for the emotion interpolation experiment.
1501 Therefore, using the same synthesized samples and newly synthesized samples with $\alpha = 2.5$, we
1502 further present the averaged in-distribution N-MOS and WER variations across the three backbones
1503 (F5-TTS, E2-TTS, and CosyVoice2) as a function of the steering strength α . For N-MOS, we ran-
1504 domly selected two groups of synthesized samples per emotion per model, where each group con-
1505 tains samples with varying α but identical linguistic content. This design reduces the substantial
1506 workload required for human evaluation. The tabulated results are shown in Tables 10, 11, and 12
1507 (WERs are computed using Whisper-Large V3 transcriptions, and the N-MOS scores are averaged
1508 across 12 participants).

1509 As shown in Tables 10, 11, and 12, increasing the steering strength α has a very consistent effect
1510 across all emotions and all three models. When α is small or moderate (up to about 1.0–1.5), both
1511 N-MOS and WER stay close to the baseline, meaning that the emotion direction can be applied
without harming speech quality or intelligibility. When α becomes larger, N-MOS gradually drops

1512 Table 12: Steering strength α vs. N-MOS and WER for E2-TTS.
1513

α	0.00	0.25	0.50	0.75	1.00	1.25	1.50	2.00	2.50	3.00
N-MOS (Anger)	4.40	4.35	4.33	4.26	4.18	4.07	3.84	3.80	2.96	1.43
N-MOS (Disgust)	4.37	4.28	4.20	4.13	4.03	3.85	3.62	3.48	3.25	1.92
N-MOS (Fear)	4.43	4.34	4.26	4.09	3.90	3.70	3.55	3.50	3.08	2.01
N-MOS (Happiness)	4.35	4.22	4.13	4.08	3.93	3.72	3.55	3.58	2.78	1.07
N-MOS (Sadness)	4.30	4.32	4.21	4.15	4.05	3.84	3.66	3.62	3.16	0.98
N-MOS (Surprise)	4.33	4.25	4.18	4.01	3.92	3.68	3.51	3.45	2.82	1.26
WER (Anger)	2.65	2.72	2.70	2.84	2.69	2.55	2.63	2.77	27.58	15.92
WER (Disgust)	2.51	2.90	2.68	2.81	2.67	2.56	2.63	2.73	18.46	28.37
WER (Fear)	2.67	3.14	2.71	2.87	2.71	2.59	2.64	2.78	26.25	21.73
WER (Happiness)	2.73	3.02	2.70	2.83	2.68	2.55	2.61	2.76	15.87	29.48
WER (Sadness)	2.46	2.73	2.71	2.85	2.72	2.58	2.66	2.79	24.36	16.48
WER (Surprise)	2.62	3.09	2.67	2.82	2.70	2.56	2.63	2.74	28.79	20.11

1527 Table 13: Emotion conversion ($\alpha = 2.0$) using SenseVoice for token probing.
1528

Method	WER \downarrow	S-SIM \uparrow	E-SIM \uparrow		UTMOS \uparrow
			emotion2vec / SenseVoice		
EmoSteer-TTS + F5-TTS	2.94	0.62	0.27 / 0.29		3.45
EmoSteer-TTS + E2-TTS	3.46	0.60	0.25 / 0.26		3.26
EmoSteer-TTS + CosyVoice2	2.77	0.58	0.26 / 0.28		3.57

1536 and WER starts to rise, and extremely large values (≥ 2.5) cause the model to leave its normal
1537 operating range and produce distorted speech. This pattern is nearly identical for F5-TTS, E2-TTS,
1538 and CosyVoice2, indicating that the behavior is general and that excessive steering can distort the
1539 feature representation across all models. This phenomenon may be attributed to shared training
1540 practices across the models, e.g., gradient clipping, normalization layers, and other regularization
1541 techniques. Therefore, we recommend the following guidance for choosing the steering strength α :

- 1542 • Stable region, less emotional: $\alpha \leq 1.0$
- 1543 • Controlled, minimal degradation, emotional: $1.0 < \alpha \leq 2.0$
- 1544 • Unstable region, noisy: $\alpha > 2.0$

1548 H.5 SENSITIVITY TO THE SER MODEL FOR TOKEN PROBING

1549 Different SER models are trained on different datasets, the final objective scores are therefore also
1550 influenced by the particular SER model used to guide the construction of steering vectors.
1551

1552 Specifically, we use SenseVoice for token probing, while reporting E-SIM scores under both
1553 emotion2vec and SenseVoice to reveal whether EmoSteer-TTS is overfitting to a specific SER embedding
1554 space. We use the same neutral samples from MSP-Podcast and ESD in our main experiments to
1555 construct steering vectors and report WER, S-SIM, and E-SIM for emotion conversion ($\alpha = 2.0$)
1556 and erasure ($\beta = 2.5$). We also report model-based UTMOS (Saeki et al., 2022) scores instead of
1557 N-MOS to avoid the substantial workload associated with human evaluation. The results are shown
1558 in the following table.

1559 As shown in Tables 13 and 14, our additional analysis reveals an extremely slight tendency of
1560 EmoSteer-TTS to align more closely with the SenseVoice emotion embedding space, as evidenced
1561 by the marginally higher E-SIM scores under SenseVoice compared to emotion2vec in Tables 13
1562 and 14. This suggests a mild degree of overfitting to the specific SER model used for token probing.

1563 However, our strong human subjective scores (e.g., EI-MOS and EE-MOS in Table 1) align with the
1564 objective metrics, giving us high confidence that EmoSteer-TTS is genuinely effective and not just
1565 overfitting to a specific metric’s embedding space. We leave for future work the extension to more
1566 SER embedding models for selecting the top-k tokens.

1566

Table 14: Emotion erasure ($\beta = 2.5$) using SenseVoice for token probing.

1567

1568

1569

1570

1571

1572

1573

1574

1575

1576

1577

1578

1579

1580

1581

1582

1583

1584

1585

1586

1587

1588

1589

1590

1591

1592

1593

1594

1595

1596

1597

1598

1599

1600

1601

1602

1603

1604

1605

1606

1607

1608

1609

1610

1611

1612

1613

1614

1615

1616

1617

1618

1619

Method	WER \downarrow	S-SIM \uparrow	E-SIM \uparrow emotion2vec / SenseVoice	UTMOS \uparrow
EmoSteer-TTS + F5-TTS	3.01	0.51	0.24 / 0.27	3.39
EmoSteer-TTS + E2-TTS	3.67	0.49	0.23 / 0.23	3.18
EmoSteer-TTS + CosyVoice2	2.98	0.53	0.26 / 0.29	3.46

Table 15: The robustness to noise and reverberation of emotion conversion ($\alpha = 2.0$) on F5-TTS using emotion2vec for token probing.

	WER \downarrow	S-SIM \uparrow	E-SIM \uparrow emotion2vec / SenseVoice	UTMOS \uparrow
Noise	34.27	0.47	0.18 / 0.16	2.94
Reverberation	12.58	0.53	0.22 / 0.20	3.15

H.6 ROBUSTNESS TO NOISE AND REVERBERATION

To investigate the robustness of our method to speech prompts with noise and reverberation, we collect 100 English samples with noise (from the Microsoft DNS Challenge dataset (Dubey et al., 2024)) and 100 English samples with reverberation (from the REVERB Challenge dataset (Kinoshita et al., 2013)) to report WER, S-SIM, E-SIM, and UTMOS for emotion conversion using F5-TTS backbone.

As shown in Table 15, reverberation has a much smaller impact on emotion conversion than additive noise. Noisy inputs significantly degrade intelligibility (WER = 34.27) and reduce both style and emotional similarity (S-SIM = 0.47, E-SIM = 0.18/0.16). In contrast, reverberant inputs maintain substantially better performance across all metrics (WER = 12.58, S-SIM = 0.53, E-SIM = 0.22/0.20), and also achieve higher perceptual quality (UTMOS = 3.15 vs. 2.94).

Overall, these results indicate that the precomputed steering vector remains robust under moderate reverberation, while strong additive noise introduces more noticeable degradation, although the emotional cues are still partially preserved.