EMOSCALE: LOCALIZED EMOTION INTENSITY CONTROL FOR SPEECH SYNTHESIS VIA TWIN-FORWARD FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent flow-matching and diffusion-transformer (DiT)-based text-to-speech (TTS) models achieve high naturalness, yet precise and localized control of emotional expression remains challenging. Existing instruction- or encoder-conditioned methods inject emotion through labels, embeddings, or prompts, while training-free steering methods extract emotion directions post hoc without explicitly enforcing disentanglement. These approaches provide limited control over the intensity and temporal extent of emotional expression, as the realised emotion can be influenced by the model's internal representations and interactions between emotion conditioning and linguistic content. Moreover, conventional categorical emotion representations and natural-language prompts provide limited means to precisely specify how strongly and when an emotion should be expressed. To address these challenges, we propose **EmoScale**, a training paradigm that explicitly disentangles the velocity field of a DiT-based flow-matching model into semantic and emotion components without requiring an external emotion embedding. First, we introduce a disentanglement regularizer that encourages the emotion velocity to be orthogonal to the semantic velocity. We further show that the resulting emotion representation enables continuous control of emotional intensity through linear interpolation of its corresponding acoustic attributes, including F0, periodicity and envelope, followed by speech reconstruction. This allows us to adjust the intensity of an emotional speech while preserving perceptual naturalness. We evaluate EmoScale against encoder-conditioned and training-free steering baselines in terms of naturalness, emotion recognition accuracy, content preservation, and localized emotion control. Results demonstrate that explicitly disentangling the velocity field enables precise and interpretable control over emotional intensity without requiring additional trainable auxiliary modules. The resulting RA scores exhibit a clear and consistent trend as the emotion intensity is varied, confirming the effectiveness of the proposed control mechanism.

1 INTRODUCTION

Emotional expressiveness is an important capability of modern speech generation systems, particularly for applications such as conversational agents and interactive virtual characters (Fraser et al., 2018). Recent advances in text-to-speech (TTS) have increasingly explored fine-grained and controllable emotional expression. A dominant paradigm incorporates emotional information through input-side conditioning, using discrete emotion labels, reference speech, or natural-language instructions. Early systems represent emotion as a fixed set of categorical labels (e.g., neutral, happy, sad, and angry), which are converted into one-hot or learned embedding vectors and injected into the synthesis backbone, for example, by concatenating them with encoder or decoder hidden states (Lee et al., 2017; Um et al., 2020). More recent systems, such as PromptTTS (Guo et al., 2023b), InstructTTS (Yang et al., 2024), and VoxInstruct (Zhou et al., 2024), move beyond fixed labels by specifying emotional attributes through natural-language descriptions or free-form instructions. In parallel, reference-conditioned TTS systems such as VALL-E (Wang et al., 2023a) and CosyVoice (Du et al., 2024) derive expressive characteristics implicitly from short acoustic prompts. For example, VALL-E has been shown to preserve the emotional characteristics of prompt speech without being

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

fine-tuned on emotional data, but provides no explicit mechanism for specifying or independently modulating emotion beyond the reference prompt. Despite these, these conditioning mechanisms typically influence the generation process globally or over the duration of the provided reference or instruction, offering limited control over when an emotion should be expressed.

While many emotional TTS systems operate at the utterance level, recent work has explored finer-grained emotional control at the phoneme, word, and intra-utterance segment levels (Im et al., 2022; Wang et al., 2023c; 2026; Liang et al., 2026). These approaches enable different linguistic units or temporal segments to exhibit different emotional expressions or intensities, providing more localised control than global utterance-level conditioning. However, these methods primarily realise localised control by manipulating emotion-related conditioning signals, attention mechanisms, or learned representations, which can entangle emotion modulation with the underlying linguistic content and make the magnitude of emotional change difficult to predict or control precisely.

A more recent line of work extracts explicit and controllable emotion directions from pretrained models, either through activation steering on frozen hidden states (Xie et al., 2025; Zhang et al., 2026) or through arithmetic in externally defined affective spaces (Liang et al., 2025). These approaches enable forms of emotion arithmetic, including direction extraction and intensity scaling. However, these controls are typically applied globally to the utterance or representation being steered, rather than allowing the emotional intensity to be selectively modified over a user-specified temporal region. While a few studies have explored emotional intensity control (Zhou et al., 2023), they rely on external modules to define the intensity scale, limiting the flexibility of the control.

Previous work has introduced orthogonality-based disentanglement during training, such as DiCLET-TTS, EmoSphere++, and Marco-Voice, which impose orthogonal or contrastive objectives to separate emotion from speaker identity, motivated by the speaker–emotion entanglement inherent in cross-speaker emotion transfer (Li et al., 2023; Cho et al., 2025; Tian et al., 2026). Even in these flow-matching-based systems, these constraints are imposed on static speaker and emotion embeddings provided as conditioning, rather than being enforced directly during generation. Moreover, the interaction between emotion and linguistic content over the temporal evolution of the generative process remains unaddressed.

Building on these gaps, we propose **EmoScale**, a training paradigm that decomposes the velocity field of a DiT-based (Peebles & Xie, 2023) flow-matching model into semantic and emotion components. The two components are obtained by querying twin DiTs, while a disentanglement regularizer encourages the emotion velocity to remain orthogonal to the semantic velocity. This explicit decomposition provides a direct control interface for emotion, allowing its magnitude to be scaled independently of the semantic component. At inference time, the disentangled emotion component is converted into acoustic attributes, whose F0, periodicity, and envelope can be linearly interpolated within user-specified temporal regions to adjust emotional intensity while preserving semantic content and perceptual naturalness.

Our **contributions** are as follows. First, we introduce **EmoScale**, a training-time framework for disentangled and fine-grained emotion intensity control in flow-matching models. Second, we introduce a disentanglement regularizer that orthogonalizes the emotion velocity against the semantic velocity, directly encouraging content–emotion separation within the generative dynamics. Third, we demonstrate that the resulting explicit emotion component enables continuous and interpretable intensity scaling, including localised modification over user-specified temporal regions, while maintaining linguistic content and speech naturalness.

2 Related Work

Emotional TTS: Conditioning and Steering. Emotional speech synthesis has traditionally relied on input-side conditioning. Early systems represent emotion as categorical labels (Lee et al., 2017), or as learned emotion embeddings shaped by inter-to-intra-class distance objectives, with interpolation between target and neutral embeddings for coarse intensity control (Um et al., 2020). Subsequent work replaces fixed label sets with free-form text. PromptTTS (Guo et al., 2023b) conditions synthesis on style descriptions through a dedicated style encoder, InstructTTS (Yang et al., 2024) maps natural-language style prompts into a discrete latent acoustic space, and VoxInstruct (Zhou et al., 2024) unifies transcript and style description within a human-instruction-to-speech framework

via codec language modelling. In parallel, reference-conditioned systems such as VALL-E (Wang et al., 2023a) and CosyVoice (Du et al., 2024) implicitly inherit emotional characteristics from short acoustic prompts, without an explicit interface for independently specifying or modulating emotion. However, these conditioning-based approaches typically introduce emotion as a global conditioning signal, offering limited flexibility for in-process control.

A complementary line of work explores post-hoc steering of pretrained TTS models. EmoSteer-TTS (Xie et al., 2025) derives training-free steering vectors from internal activations for emotion conversion, interpolation, and erasure, while DUET (Zhang et al., 2026) identifies emotion directions in hidden states and combines hidden-state and mel-space guidance. ECE-TTS (Liang et al., 2025) instead modulates emotion intensity in an external affective space through vector arithmetic. While these approaches provide effective global emotion control, they primarily manipulate conditioning or static representations, limiting continuous intensity scaling and temporally localised control.

Emotion Intensity Control. Distinct from selecting an emotion category, a substantial body of work has explored continuous emotion intensity control. RAR-based methods (Zhou et al., 2023) learn ordinal relationships between emotions and use attribute vectors to modulate utterance-level intensity, but treat intensity as an externally defined embedding produced by an additional module that must be trained separately. MsEmoTTS (Lei et al., 2022) further introduces syllable-level strength scores derived from handcrafted acoustic features, enabling finer-grained temporal control but requiring HMM-based alignment, an external ranker, and explicit strength assignment for each syllable. A parallel line incorporates intensity more directly into the generative process, such as EmoDiff (Guo et al., 2023a), but its intensity coefficient scales gradients from an independently trained emotion classifier. EmoSphere-TTS (Cho et al., 2024) further constructs an explicit intensity axis by defining the radius of a VAD-based spherical representation as emotion strength. However, this axis is fixed a priori by an external SER model and a hand-designed geometric assumption, with no guarantee that varying the radius preserves linguistic content. Overall, these approaches rely on externally defined intensity scales or guidance signals, making their controllability inherently dependent on external resources, rather than allowing the intensity dimension to be intrinsically learned and adapted within the TTS generative process.

Localized Emotion Control. Beyond utterance-level control, several works pursue fine-grained emotion rendering at the sub-utterance level. EmoQ-TTS (Im et al., 2022) conditions on phoneme-level emotion information with distance-based intensity quantization, while FEC (Wang et al., 2023c) learns continuous phoneme-level emotion using a ranking model trained on Mixup-augmented neutral–emotional pairs. HED (Inoue et al., 2025) further extends this paradigm to a hierarchical phoneme–word–utterance emotion distribution extractor within a flow-matching backbone. Most recently, TED-TTS (Liang et al., 2026) introduces training-free intra-utterance emotion shifts by combining causal masking with monotonic stream alignment on a pretrained zero-shot TTS model. Overall, these methods rely on predefined linguistic units for emotion localisation, while flexibility in emotion control can also extend to the temporal dimension.

Disentanglement in Emotional TTS. Several recent systems introduce explicit disentanglement objectives to separate emotion from other features. DiCLET-TTS (Li et al., 2023) uses an orthogonal projection-based module to learn speaker-irrelevant emotion embeddings, while EmoSphere++ (Cho et al., 2025) further disentangles speaker and emotion representations for zero-shot synthesis. Marco-Voice (Tian et al., 2026) combines a cross-orthogonal constraint with contrastive learning to separate speaker and emotion representations. However, these methods focus on disentangling speaker–emotion representations, rather than the interaction between emotion and linguistic content, and impose constraints on static representations rather than the generation dynamics. These methods also involve relatively complex mathematical formulations and compound objectives.

Evaluation of Emotion Intensity. Evaluating continuous emotion intensity has largely followed two paradigms. The first uses relative-attribute ranking (RAR) (Parikh & Grauman, 2011), where listeners perform pairwise judgements of which utterance expresses stronger emotion. This underlies the Best-Worst Scaling (BWS) tests in (Zhou et al., 2022b), as well as the Emotion Intensity Recognition protocol of EmoSphere-TTS (Cho et al., 2024) and the Emotion Intensity Test (EIT) of EMORL-TTS (Li et al., 2026). The second uses a pretrained SER classifier to verify whether the target-emotion posterior increases monotonically with the intensity coefficient (Zhou et al., 2023;

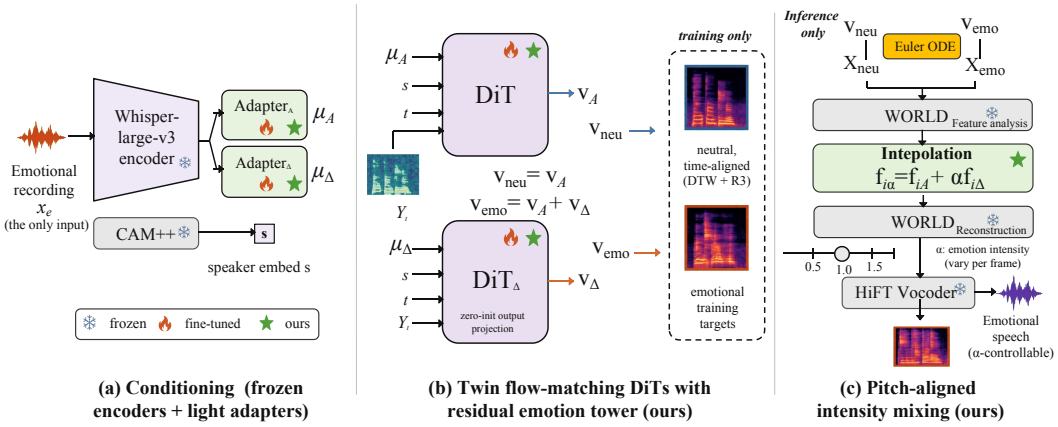


Figure 1: Overview of the EmoScale architecture. (a) Conditioning: a frozen Whisper encoder with two lightweight adapters that capture semantic and emotional information, while a frozen CAM++ encoder provides the speaker embedding. (b) Twin flow-matching DiTs: one DiT predicts the neutral semantic velocity, while the other predicts the additive emotional velocity. The two DiTs are trained using paired neutral and emotional targets. (c) Pitch-aligned intensity mixing for inference: the two velocity fields are integrated separately to obtain neutral and emotional trajectories. Their corresponding acoustic features are then interpolated according to per-frame emotion intensity estimated from the WORLD domain, followed by WORLD-based Mel-spectrogram reconstruction and a frozen HiFT vocoder to generate speech with controllable emotional intensity.

Guo et al., 2023a), but its reliability is inherently limited by the biases of the external SER model, particularly at low intensities. Therefore, we primarily evaluate EmoScale through RAR.

3 Methodology: Twin-Forward Velocity Decomposition Flow Matching

We formulate emotion-controllable speech synthesis as a decomposition of the flow-matching velocity field into a semantic component and an emotion component, learned jointly with the acoustic generator. Unlike prior work that either injects emotion as an input-side condition or extracts steering directions from a frozen backbone post hoc, our formulation makes the disentanglement of the emotion direction an explicit training objective of a single generator, and realises emotion arithmetic directly in the space in which the model predicts. The overview pipeline is shown in Figure 1.

Paired training data. We train on the Emotional Speech Dataset (ESD) (Zhou et al., 2022a), which natively provides, for each speaker and each utterance transcript, a neutral rendering paired with renderings in four non-neutral emotions (*happy*, *sad*, *angry*, *surprise*). Formally, we treat the training set as a collection of pairs $\mathcal{D} = \{(x_{\mathrm{n}}^{(i)}, x_{\mathrm{e}}^{(i)}, e^{(i)})\}_i$, where x_{n} and x_{e} share the same text and the same speaker, and $e \in \mathcal{E} = \{\mathrm{happy}, \mathrm{sad}, \mathrm{angry}, \mathrm{surprise}\}$ is the emotion label of x_{e}.

To facilitate frame-level correspondence for subsequent interpolation and intensity scaling, we employ Dynamic Time Warping (DTW) (Sakoe & Chiba, 1978) and RubberBandStretcher[1] to temporally scale the neutral audio to the same duration as the emotional utterance, with the technical details provided in Appendix A. The resulting neutral audio therefore has the same duration as the emotional counterpart while providing closer temporal alignment at the sentence level. Importantly, the temporal scaling is applied only to the neutral audio and does not explicitly modify its emotional category. Thus, the aligned neutral audio remains a neutral target, and this preprocessing step does not introduce an additional factor into the subsequent emotion-intensity scaling.

Conditioning. We take a single emotional recording x_e as input and reuse two frozen pretrained encoders to extract complementary conditions. A frozen Whisper-large-v3 (Radford et al., 2023)

[1] https://github.com/pawel-glomski/pylibrb

encoder produces frame-level semantic features that are routed through two lightweight, fine-tunable adapters. We design adapter A to extract the semantic condition μ_A, while Adapter adapter Δ yields a parallel condition μ_Δ dedicated to the emotion-residual component. Speaker identity is captured by a frozen CAM++ (Wang et al., 2023b) speaker encoder, which produces a global embedding s.

Twin flow-matching DiTs. The above conditioning signals are then fed into the core generative module, which consists of two twin DiT-based flow-matching modules. The two modules share the same time step t, noised mel state Y_t (the flow-matching state at time t, derived from the original emotional audio), and speaker embedding s, with μ_A and μ_Δ providing the respective conditioning signals. One DiT consumes μ_A and produces the semantic velocity v_A. The other DiT Δ progressively learns an emotion-specific velocity component v_Δ from μ_Δ.

During training, the neutral branch is supervised with the aligned neutral target to learn the **semantic direction** $v_{\text{neu}} = v_A$, while the emotional branch is supervised with x_e to learn the combined velocity, **emotion direction** $v_{\text{emo}} = v_A + v_\Delta$. By sharing the same speaker, temporal, and conditioning information across the two branches, the residual branch is explicitly encouraged to capture the dynamics that distinguish emotional generation from its neutral counterpart.

Orthogonality on the predicted velocity. To better split semantic and emotion direction, we impose an angular regularisation on the predicted velocity fields, encouraging the temporal trajectories of v_A and v_Δ to occupy distinct directions in the trajectory space. Let $\hat{v}A, i$ and $\hat{v}\Delta, j$ denote the ℓ_2-normalised trajectories of the i-th and j-th mel channels, respectively, where the normalisation is performed along the temporal dimension. The regularisation is defined as

$$\mathcal{L}_{\text{ort}}(\theta) \;=\; \mathbb{E}_{t,\,s,\,\mu_A,\,\mu_\Delta,\,Y_t}\left[\,\frac{1}{F^2}\sum_{i=1}^{F}\sum_{j=1}^{F}\langle \hat{v}_{A,i},\,\hat{v}_{\Delta,j}\rangle^2\,\right], \tag{1}$$

where $F = 80$ is the number of mel channels. Two properties are worth noting. First, normalising each trajectory makes the penalty equivalent to squared cosine similarity, making it invariant to the magnitude of the velocity fields and preventing the regularisation from implicitly suppressing the emotional component. Second, the regularisation is applied directly to the outputs of f_θ, so gradients propagate through both conditioning passes and directly shape the DiT's velocity fields rather than an external projection. This yields a learnable emotion-velocity component without requiring an auxiliary emotion classifier, reference-based emotion extractor, or predefined affective geometry.

Training objective. For reconstruction, both branches are trained with the standard flow-matching regression objective against their respective target velocities, $u_t^{\text{n}} = x_{\text{n}} - \epsilon, u_t^{\text{e}} = x_{\text{emo}} - \epsilon$, where $(\epsilon \sim \mathcal{N}(0, I))$ denotes the initial noise sample. The reconstruction loss is defined as:

$$\mathcal{L}_{\text{FM}}^{\text{neu}}(\theta) \;=\; \mathbb{E}\big[\,\|v_{\text{neu}} - u_t^{\text{n}}\|_2^2\,\big], \qquad \mathcal{L}_{\text{FM}}^{\text{emo}}(\theta) \;=\; \mathbb{E}\big[\,\|v_{\text{emo}} - u_t^{\text{e}}\|_2^2\,\big]. \tag{2}$$

Retaining *both* reconstruction terms is essential, as the neutral reconstruction constraint encourages v_Δ to preserve the necessary semantic content without absorbing all information required for emotional reconstruction, thereby preventing v_A from collapsing into an uninformative residual.

The three terms are combined with a single scalar coefficient $\lambda = 1$:

$$\mathcal{L}(\theta) \;=\; \mathcal{L}_{\text{FM}}^{\text{neu}}(\theta) \;+\; \mathcal{L}_{\text{FM}}^{\text{emo}}(\theta) \;+ \lambda\mathcal{L}_{\text{ort}}(\theta). \tag{3}$$

We use two branches to jointly learn a coherent reconstruction, while $\mathcal{L}_{\text{ort}}$ discourages the branches from becoming entangled over time.

Pitch-aligned intensity interpolation at inference. At inference, the two branches are synchronously integrated with Euler ODE solvers to obtain a neutral candidate and an emotional candidate, which are then synthesised into waveforms using the HiFT vocoder (Kong et al., 2020). Rather than directly interpolating the two waveforms, which can introduce multimodal spectral components and lead to undesirable spectral artefacts and unnatural timbre as proved later in experiments, EmoScale performs **pitch-aligned intensity mixing** in an interpretable acoustic space. We employ WORLD (Morise et al., 2016) to decompose the two waveforms into frame-level fundamental frequency (F_0), spectral envelope (s_i), and aperiodicity representations (a_i), all defined on the same

temporal grid. For each frame i, we perform interpolation in the corresponding acoustic domains. Specifically, F_0 is interpolated in the logarithmic domain, while the spectral envelope and aperiodicity representations are interpolated linearly as follows:

$$\begin{aligned} \log F_{0,i} &= \log F_{0,i}^{A} + \alpha_i \log F_{0,i}^{\Delta}, \\ s_i &= s_i^{A} + \alpha_i s_i^{\Delta}, \\ a_i &= a_i^{A} + \alpha_i a_i^{\Delta}, \end{aligned} \tag{4}$$

where α_i specifies the emotion intensity at frame i. Log-domain interpolation of F_0 provides a more perceptually appropriate treatment of pitch, as pitch perception is approximately logarithmic with respect to frequency, whereas the spectral envelope and aperiodicity are interpolated directly in their respective acoustic representations.

The interpolated acoustic features are subsequently resynthesised into a waveform using WORLD. It ensures a consistent analysis and reconstruction pipeline, allowing the manipulated acoustic representations to be resynthesised into high-quality waveforms without introducing noticeable degradation. This feature-level interpolation provides a continuous and interpretable mechanism for controlling emotion strength while preserving the shared temporal and pitch structure of the two candidates, without requiring an auxiliary discriminator.

4 Experiments and results

We initialise the twin DiTs from the pretrained CosyVoice3 model (Du et al., 2024) to retain the acoustic knowledge acquired during pretraining and thereby improve generalisation. Specifically, DiT is fully initialised with the pretrained weights, while the projector in DiT_{Δ} is initialised to zero. For dataset, we use around 20K ESD pairs for training and 3K pairs for testing. Unlike previous emotion-related studies that explicitly adopt speaker-disjoint splits, we retain the speakers across the training and test sets because speaker identity is not a target factor in our study. The resulting ESD test set therefore evaluates in-domain performance under the same speaker and corpus distribution as training. We further assess cross-speaker and cross-corpus generalisation using CREMA-D (Cao et al., 2014), RAVDESS (Livingstone & Russo, 2018) and IEMOCAP (Busso et al., 2008) as out of domain (OOD) evaluation sets. The remaining training hyperparameters and dataset details are shown in Appendix B. The codes and dataset are shown in Anonymous Link.

Main result We evaluate automatically naturalness, content correctness, speaker consistency, and emotion intensity using the UTMOS (UTokyo-SpeechMOS) (Saeki et al., 2022), CER (Character Error Rate) (Morris et al., 2004), SECS (Speaker Embedding Cosine Similarity) (Desplanques et al., 2020), and RA (Relative Attribute) (Parikh & Grauman, 2011) metrics, respectively. SECS is calculated between the generated one with the input. The RA scorer is trained independently on ESD and further validated on other datasets against their original annotations, showing consistent agreement, shown in Appendix C. Since the other datasets do not necessarily contain paired neutral and emotional utterances, we report the average score over the neutral samples as the Real Neutral score. Results are reported in Table 1.

Overall, the results show two desirable properties. First, speech quality and content preservation remain broadly stable across intensity levels, with UTMOS mostly above 2 and no systematic degradation in CER as α increases. Lower UTMOS scores on some datasets are largely attributable to the quality of the original recordings. Second, RA increases monotonically with α across all five test sets, including OOD datasets, showing consistent and predictable control of emotional intensity.

We additionally recruited 10 participants (5 male and 5 female), aged 18–30 years, all of whom had prior experience with audio machine learning. For each of the five datasets, we randomly sampled 40 utterances, with 10 samples for each intensity level ($\alpha \in 0, 0.5, 1.0, 1.5$). Each participant evaluated 60 utterances, resulting in three independent ratings per utterance on average. Participants performed two subjective evaluation tasks. First, they conducted a BWS (Zhou et al., 2022b) test to assess relative emotional intensity. Second, they rated the overall speech quality of each utterance on a 5-point Likert scale, ranging from 1 (very poor) to 5 (excellent). The results are shown in Table 2, and samples are shown in Anonymous Link. The results echo with automatic objective test scores.

Table 1: Full evaluation results across five test sets. α controls the emotion intensity, with $\alpha = 0$ corresponding to the neutral output and larger values producing stronger emotional expression. Darker red denotes higher RA.

		EmoScale at emotion intensity α							Real recordings	
Dataset	**Metric**	0	0.25	0.5	0.75	1.0	1.25	1.5	Neutral	Emotion
ESD Chinese	UTMOS ↑	2.32	2.15	2.07	2.19	2.29	2.00	1.63	3.68	2.74
	CER ↓	0.264	0.222	0.161	0.085	0.064	0.074	0.100	0.016	0.030
	SECS ↑	0.855	0.867	0.883	0.902	0.917	0.909	0.883	–	1.000
	RA	0.16	0.29	0.47	0.69	0.97	1.26	1.63	0.07	1.03
ESD English	UTMOS ↑	2.87	2.83	2.82	2.93	2.98	2.66	2.17	4.19	3.65
	CER ↓	0.124	0.097	0.080	0.060	0.053	0.059	0.081	0.026	0.028
	SECS ↑	0.827	0.837	0.851	0.871	0.892	0.895	0.883	–	1.000
	RA	0.37	0.42	0.49	0.62	0.76	0.92	1.10	0.27	0.78
CREMA-D	UTMOS ↑	2.67	2.45	2.23	2.33	2.44	2.23	1.94	1.57	1.55
	CER ↓	0.542	0.527	0.489	0.445	0.418	0.431	0.438	0.400	0.398
	SECS ↑	0.696	0.702	0.720	0.742	0.757	0.758	0.747	–	1.000
	RA	0.33	0.44	0.57	0.73	0.93	1.09	1.34	-0.20	0.56
RAVDESS	UTMOS ↑	2.59	2.39	2.40	2.57	2.61	2.28	1.86	4.08	3.32
	CER ↓	0.346	0.283	0.180	0.094	0.064	0.067	0.094	0.000	0.003
	SECS ↑	0.719	0.743	0.769	0.799	0.826	0.835	0.825	–	1.000
	RA	0.15	0.27	0.40	0.58	0.79	0.97	1.20	0.26	0.88
IEMOCAP	UTMOS ↑	2.13	1.90	1.78	1.88	1.98	1.85	1.62	2.40	2.21
	CER ↓	0.707	0.621	0.483	0.379	0.346	0.349	0.395	0.194	0.139
	SECS ↑	0.690	0.695	0.711	0.740	0.770	0.783	0.781	–	1.000
	RA	0.30	0.37	0.46	0.61	0.79	0.95	1.16	-0.10	0.42

Comparison with other baseline models To assess the effectiveness of our method against existing emotion-intensity control approaches under OOD conditions, we evaluate three types of emotion-intensity control strategies on RAVDESS. We consider three categories of emotion-intensity control: **explicit control (EC)**, where the model provides a mechanism inherently to adjust the emotion intensity; **implicit control (IC)**, where the intensity is manipulated through an emotion representation which we could manipulate via parameters such as α; and **condition-based control (CC)**, where emotion intensity is controlled through conditioning such as prompts. The results are shown in Table 3, excluding methods that cannot be reproduced, such as InstructTTS, for which no checkpoint is publicly available. The details for running these baseline models are in Appendix E. Our method achieves both predictable emotion-intensity control and good speech quality, with RA increasing consistently from weak to strong while maintaining competitive UTMOS scores. In contrast, existing methods either provide less reliable intensity control (IC and CC) or sacrifice speech quality (EC).

Table 2: Subjective evaluation results across five datasets. BWS measures relative emotional intensity, while 5-star ratings assess overall speech quality. α denotes the emotion intensity control parameter. Across all five datasets, perceived emotional intensity increased significantly ($p<0.001$). Darker red denotes higher BWS.

		Emotion intensity α			
Dataset	**Metric**	0	0.5	1.0	1.5
ESD-ZH	BWS	-0.92	-0.03	+0.07	+0.88
	5-star	3.68	3.52	3.91	3.27
ESD-EN	BWS	-0.64	-0.11	+0.05	+0.70
	5-star	3.61	3.48	3.82	3.25
CREMA-D	BWS	-0.69	-0.04	+0.02	+0.72
	5-star	3.27	3.03	3.56	2.99
RAVDESS	BWS	-0.82	-0.10	+0.07	+0.85
	5-star	3.15	3.10	3.58	2.72
IEMOCAP	BWS	-0.64	-0.14	+0.06	+0.72
	5-star	3.79	3.12	3.49	2.84

Ablation study with λ value Our method introduces an orthogonality loss to disentangle the emotion and semantic directions. As shown in Table 4, without this loss, the two directions remain insufficiently separated despite exhibiting some inherent separation. As λ increases, orthogonality improves, but overly strong regularisation reduces the relative contribution of v_Δ, with Δ_{share} decreasing. This is accompanied by a narrower RA range, indicating reduced emotion-intensity controllability. Thus, excessively strong orthogonality constraints can suppress the emotion direction.

Interpretability of the emotion direction To examine whether the emotion direction is successfully learned, we visualise generated utterances with different speakers, sentence contents, and emo-

Table 3: Comparison of different emotion-intensity control strategies on the OOD RAVDESS dataset. **EC** denotes explicit control, where the model provides a native intensity parameter; **IC** denotes implicit control, where intensity is manipulated through the internal emotion representation; and **CC** denotes condition-based control, where emotion intensity is controlled through textual or acoustic conditioning. Bold values indicate the results closest to the corresponding real-emotion intensity. For CosyVoice3 and EmotiVoice, we evaluate both prompt-based conditioning and embedding interpolation. W/M/S: weak/medium/strong target intensity. Our method is highlighted in blue.

		RA ↑			CER ↓			UTMOS ↑		
Type	**Model**	W	M	S	W	M	S	W	M	S
EC	RAR (Zhou et al., 2023)	0.67	0.97	1.17	0.067	0.062	0.068	1.36	1.34	1.33
	EmoKnob (Chen et al., 2024)	0.62	0.67	0.68	0.072	0.082	0.091	2.99	2.95	2.90
IC	CosyVoice3 (Du et al., 2025)	0.12	0.11	0.13	0.018	0.021	0.025	3.57	3.55	3.58
	EmotiVoice†	0.01	0.01	0.01	0.040	0.036	0.027	3.04	3.02	3.04
CC	CosyVoice3 (Du et al., 2025)	0.10	0.08	0.10	0.047	0.021	0.083	3.56	3.54	3.59
	EmotiVoice	0.01	-0.02	0.07	0.040	0.036	0.027	3.04	3.02	3.04
	PromptTTS (Guo et al., 2023b)	0.24	0.27	0.26	0.034	0.040	0.045	4.28	4.30	4.27
EC	**EmoScale (Ours)**	0.40	0.58	0.79	0.180	0.094	0.064	2.40	2.57	2.61

† `https://github.com/netease-youdao/emotivoice`

tions in Figure 2a. We apply UMAP (McInnes et al., 2018) for visualisation. The embeddings cluster primarily by emotion rather than speaker or sentence content, with cluster centres also determined by emotion.

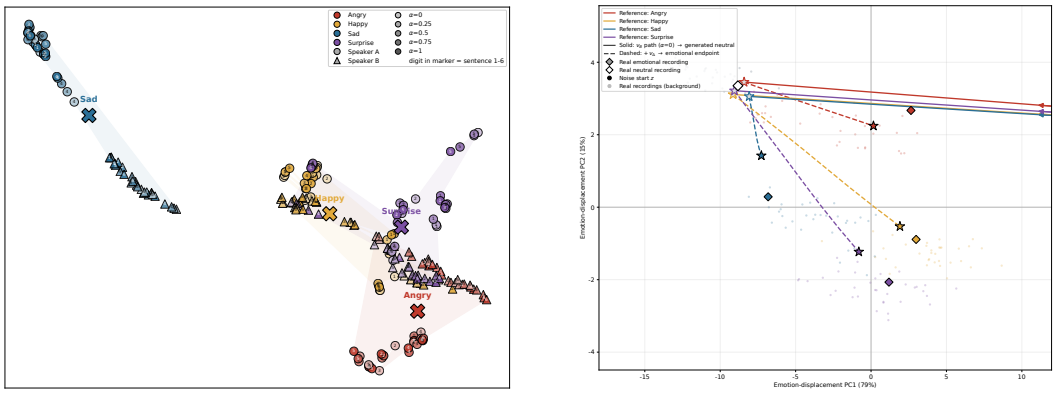
We also visualise the inference trajectories conditioned on different emotions in Figure 2b. We apply PCA (Pearson, 1901) for dimensionality reduction before visualisation. The emotion velocities v_Δ form distinct trajectories across emotions, while the semantic velocities v_A remain close to each other, as indicated by the blank star representing neutral generation. This suggests that the learned emotion direction is effectively disentangled from semantic information.

Table 4: Ablation study on the orthogonality regularisation coefficient λ. L_{ortho} and $\cos(v_\Delta, v_A)$ measure the separation between the emotion and semantic directions. Δ_{share} measures the proportion of the overall direction accounted for by v_Δ. Bold values indicate the results closest to the real emotion intensity. Darker red denotes higher RA; blue marks the default $\lambda = 1$.

	Disentanglement			RA at intensity α			
λ	L_{ortho}	cos	Δ_{share}	0	0.5	1	1.5
0	0.250	-0.30	3.9%	0.008	0.651	1.126	1.639
1	0.073	-0.17	3.0%	**0.080**	0.646	**1.041**	1.548
2	0.034	-0.11	2.9%	0.166	0.681	1.066	1.492
5	0.010	-0.06	2.5%	0.213	0.669	0.946	1.245
Real	–	–	–	**0.08**	–	**1.03**	–

Localised emotion intensity adjustment Unlike prompt-based TTS systems, which do not allow fine-grained control of emotion intensity within an utterance, our method enables frame-level control by varying α over time. To demonstrate this capability, we construct an utterance with different α values in its first and second halves and visualise the resulting mel-spectrogram in Appendix F. The spectrogram exhibits the expected gradual modulation of emotional characteristics. The $\alpha = 0$ segment resembles neutral speech, the $\alpha = 1$ segment exhibits stronger emotional characteristics, while the $\alpha = 0.5$ segment shows an intermediate pattern. The related objective scores shown in the Figure are also in expectation. The range is not restricted to either the first or second half in principle, but can occur at any point, and the human participants agreed that they could clearly perceive the change in emotion. The generated sample can be heard in Anonymous Link.

Acoustic feature change To further investigate why our mechanism is capable of controlling emotional intensity, we use openSMILE (Eyben et al., 2010) to extract the 88-dimensional eGeMAPSv02 feature set and analyse which acoustic features are modulated by varying α. For the investigated emotions, pitch, loudness, and spectral-envelope features vary monotonically with α, whereas formants and vocal perturbation show less consistent trends, as shown in Figure 3. This echoes previous findings that emotion is related to acoustic features such as pitch and loudness (Scherer, 2003), while our analysis further shows that pitch, loudness, and spectral-envelope features exhibit more predictable changes with α. Additional emotion related feature changes are analysed in Appendix D. It remains as the open future questions about whether specifically improves formants these features could better control emotion intensity.



(a) UMAP of generated utterances **(b)** PCA of inference trajectories

Figure 2: Interpretability visualisation. (a) Visualisation of randomly sampled utterances from the ESD test set. Darker colours indicate larger values of α. Four emotions, two speakers, and six sentence contents are sampled. Numbers inside the shapes denote different sentence contents. (b) Inference trajectories conditioned on different emotions. Since the initial noise points are too far apart to be clearly visualised together, we omit the initial segments and show the trajectories from points where they can be clearly visualised.

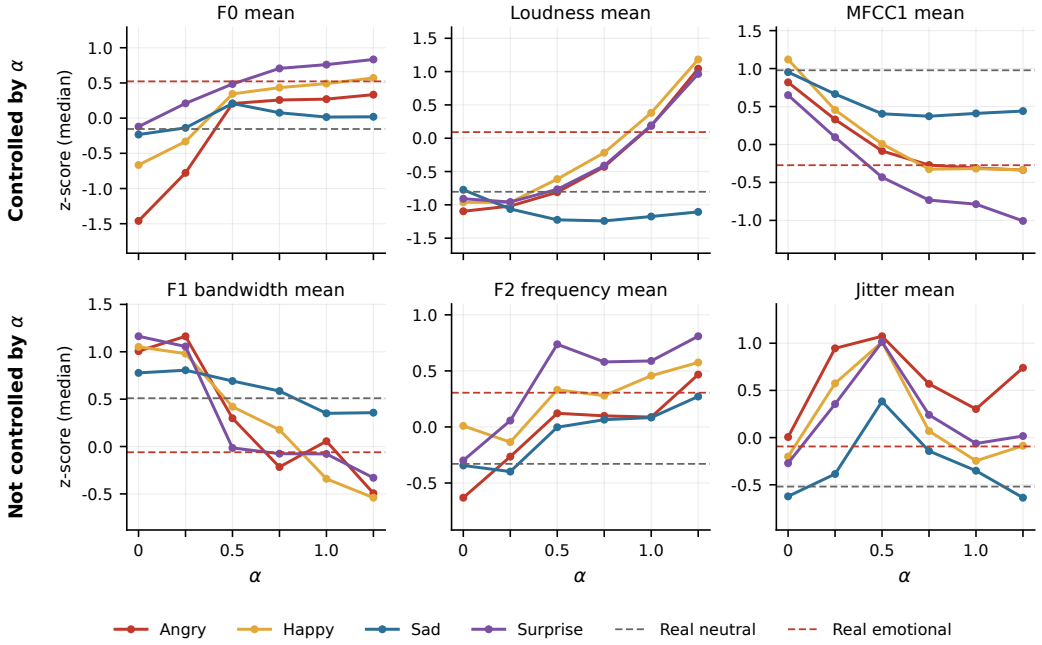


Figure 3: Top and bottom acoustic features with the most and least controllable changes as α varies, respectively.

5 Conclusion

In this work, we proposed **EmoScale**, a flow-matching DiT framework that explicitly disentangles semantic and emotional velocity fields, enabling continuous and localised control of emotional intensity without external emotion embeddings or additional trainable auxiliary modules. Experiments across multiple datasets show that EmoScale achieves predictable emotion-intensity control, providing interpretability, with RA scores increasing consistently with the specified intensity, while maintaining content preservation and competitive speech quality. In future work, we will extend EmoScale to support a broader range of emotions and more diverse emotional expressions.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with translation, supporting qualitative and thematic data analysis, and interpreting results. We have not used generative AI tools for helping, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, or implementing methods. We have reviewed all AI-assisted work. Generating synthetic data sets, cleaning, and reformatting datasets are not applicable. We checked LLM-synthetic data, data analysis, and interpreted results through a manual check and reorganisation, LLM-generated code was verified and tested for correctness by 5 authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This project does not raise any significant ethical concerns. It involves human participants for only their judgements but does not collect personal or sensitive data, or interventions that could cause physical, psychological, or social harm. The work is conducted using computational methods and publicly available or appropriately licensed data, with due consideration given to responsible use of the developed methods and the limitations of the experimental results.

REPRODUCIBILITY STATEMENT

We have provided all the data and resources codes in Anonymous Link.

REFERENCES

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Lee, Sungbok Kim, Kwang Seob Chang, Sungbok Lee, and Shrikanth Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4): 377–390, 2014. doi: 10.1109/TAFFC.2014.2336244. URL `https://doi.org/10.1109/TAFFC.2014.2336244`.

Haozhe Chen, Run Chen, and Julia Hirschberg. Emoknob: Enhance voice cloning with fine-grained emotion control. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8170–8180, 2024.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Sang-Hoon Lee, and Seong-Whan Lee. Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. *arXiv preprint arXiv:2406.07803*, 2024.

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

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. Interspeech*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650.

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

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

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

Jamie Fraser, Ioannis Papaioannou, and Oliver Lemon. Spoken conversational ai in video games: Emotional dialogue management increases user engagement. In *Proceedings of the 18th international conference on intelligent virtual agents*, pp. 179–184, 2018.

Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu. Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance. In *ICASSP 2023-2023 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

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

Chae-Bin Im, Sang-Hoon Lee, Seung-Bin Kim, and Seong-Whan Lee. Emoq-tts: Emotion intensity quantization for fine-grained controllable emotional text-to-speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6317–6321. IEEE, 2022.

Sho Inoue, Kun Zhou, Shuai Wang, and Haizhou Li. Hierarchical control of emotion rendering in speech synthesis. *IEEE Transactions on Affective Computing*, 16(4):3316–3328, 2025.

Eric Jordan, Raphaël Terrisse, Valeria Lucarini, Motasem Alrahabi, Marie-Odile Krebs, Julien Desclés, and Christophe Lemey. Speech emotion recognition in mental health: Systematic review of voice-based applications. *JMIR mental health*, 12(1):e74260, 2025.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

Younggun Lee, Azam Rabiee, and Soo-Young Lee. Emotional end-to-end neural speech synthesizer. *arXiv preprint arXiv:1711.05447*, 2017.

Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:853–864, 2022.

Haoxun Li, Yu Liu, Yuqing Sun, Hanlei Shi, Leyuan Qu, and Taihao Li. Emorl-tts: Reinforcement learning for fine-grained emotion control in llm-based tts. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17272–17276. IEEE, 2026.

Tao Li, Chenxu Hu, Jian Cong, Xinfa Zhu, Jingbei Li, Qiao Tian, Yuping Wang, and Lei Xie. Diclet-tts: Diffusion model based cross-lingual emotion transfer for text-to-speech—a study between english and mandarin. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 3418–3430, 2023.

Qifan Liang, Yuansen Liu, Ruixin Wei, Nan Lu, Junchuan Zhao, and Ye Wang. Ted-tts: Training-free intra-utterance emotion and duration control for text-to-speech synthesis. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23485–23508, 2026.

Shixiong Liang, Ruohua Zhou, and Qingsheng Yuan. Ece-tts: A zero-shot emotion text-to-speech model with simplified and precise control. *Applied Sciences*, 15(9):5108, 2025.

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

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Proc. Interspeech 2004*, pp. 2765–2768, 2004.

Devi Parikh and Kristen Grauman. Relative attributes. In *2011 International conference on computer vision*, pp. 503–510. IEEE, 2011.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/radford23a.html`.

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

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256, 2003.

Fengping Tian, Peng Bai, Xuanfan Ni, Haoqin Sun, Qingjuan Li, Zhiqiang Qian, Chenyang Lyu, Haijun Li, Longyue Wang, Zhao Xu, et al. Marco-voice: A unified framework for expressive speech synthesis with voice cloning. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3256–3260. IEEE, 2026.

Se-Yun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang. Emotional speech synthesis with rich and granularized control. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7254–7258. IEEE, 2020.

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

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*, 2023b.

Shijun Wang, Jón Gunoason, and Damian Borth. Fine-grained emotional control of text-to-speech: Learning to rank inter-and intra-class emotion intensities. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023c.

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

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

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2913–2925, 2024.

Xu Zhang, Longbing Cao, and Zhangkai Wu. Duet: Unified dual-space emotion control for diffusion and flow-matching driven text-to-speech. *arXiv preprint arXiv:2606.00066*, 2026.

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

Kun Zhou, Berrak Sisman, Rajib Rana, Björn W Schuller, and Haizhou Li. Speech synthesis with mixed emotions. *IEEE Transactions on Affective Computing*, 14(4):3120–3134, 2022b.

Kun Zhou, Berrak Sisman, Rajib Rana, Björn W. Schuller, and Haizhou Li. Emotion intensity and its control for emotional voice conversion. *IEEE Transactions on Affective Computing*, 14 (1):31–48, January 2023. ISSN 2371-9850. doi: 10.1109/taffc.2022.3175578. URL `http://dx.doi.org/10.1109/TAFFC.2022.3175578`.

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

A Temporal Alignment of the Neutral Target

The neutral and emotional targets need to be temporally aligned because they are used as frame-level counterparts in the two flow-matching branches. However, neutral and emotional recordings may differ in duration, speaking rate, and pause locations. Uniformly matching their global durations does not guarantee local frame-level correspondence. Without explicit alignment, the temporal discrepancy between the neutral target and the Whisper conditioning can reach 23–284 ms, leading to substantial content misalignment when interpolating.

DTW-based correspondence. Given a neutral waveform x_n and an emotional waveform x_e, both are resampled to 24 kHz and converted into 80-dimensional mel-spectrograms using $n_{\text{fft}} = 1920$ and a hop size of 480 (20 ms). We first estimate their frame-level correspondence using DTW with Euclidean distance. For each emotional frame, the corresponding neutral frame indices along the DTW path are averaged to obtain an initial mapping. We then apply a cumulative maximum to enforce a monotonic mapping from the emotional to the neutral time axis.

Anchor-based time-scale estimation. We sample temporal anchors every 250 ms (12 mel frames), together with the final frame. Each anchor (S_i, T_i) specifies a source position S_i on the neutral waveform and its corresponding target position T_i on the emotional time axis. The local time-scale ratio is

$$r_i = \frac{T_{i+1} - T_i}{S_{i+1} - S_i}, \tag{5}$$

which is constrained to $[0.4, 2.5]$ to prevent excessive local stretching. Non-monotonic anchors are discarded, and the final anchor is set to the end of the emotional target so that the warped neutral waveform matches its duration.

Time-varying Rubber Band R3 warping. The resulting local ratios define a piecewise-constant time-scale curve, which is applied to the neutral waveform using the R3 (Finer) engine of the Rubber Band Library via `pylibrb`. We use realtime processing with 256-sample blocks (16 ms), updating the `time_ratio` for each block according to the local DTW-derived ratio. Processing the utterance as a continuous stream allows the time-scale ratio to vary smoothly across the recording without independently stretching and concatenating individual segments. The resulting aligned neutral target has the same number of frames as the emotional target and is used for neutral flow-matching supervision and subsequent emotion-intensity interpolation. The time-scale transformation is applied only to compensate for duration and speaking-rate differences. Although it may affect temporal prosody, it does not explicitly introduce an emotional attribute into the neutral target.

B HYPERPARAMETERS IN TRAINING

The ESD dataset is split by sentence index rather than by speaker. All 20 speakers are represented in the training set, and the evaluation is not speaker-disjoint for this dataset. The resulting dataset statistics and preprocessing configuration are summarised in Table 5.

Table 5: Dataset statistics and preprocessing configuration.

Item	Configuration
Training corpus	350 utterances per speaker $\times$ 5 emotions
Training pairs	21,480: 10,097 Mandarin + 11,383 English
Test pairs	2,696: 1,348 Mandarin + 1,348 English
Whisper input dimension	1,280
Speaker representation dimension	192

The two DiTs share the same architecture and pretrained weights. For the residual branch, the output projection is zero-initialised so that the residual velocity is initially zero, i.e. $v_\Delta = 0$, and the emotional branch therefore initially reduces to the base branch.

The architecture and parameter initialisation are summarised in Table 6.

Table 6: Model architecture, parameterisation, and initialisation details.

Module	Parameters	Architecture	Initialisation	Trainable
DiT	331M	CosyVoice3 flow-matching decoder	Pretrained CosyVoice3 weights	✓
DiT_Δ	331.2M	Same architecture as DiT	Pretrained weights; output projector weights and bias zero-initialised	✓
Adapters $\times 2$	2M $\times 2$	Conv1d $\rightarrow$ GELU $\rightarrow$ Conv1d	PyTorch default random initialisation	✓
Speaker affine	15K	Linear	Pretrained CosyVoice3 weights	✓
Whisper-large-v3	1,550M	Frozen pretrained encoder	Hugging Face weights	$\times$
CAMPPlus	7M	Frozen pretrained speaker encoder	Hugging Face weights	$\times$
HiFT	21M	Frozen pretrained vocoder	Hugging Face weights	$\times$
Total trainable	**666M**			

Unless otherwise stated, all experiments use the training configuration in Table 7. The flow-matching time variable t is sampled independently for each utterance. A single Gaussian noise sample is shared by the neutral and emotional targets within each training pair, ensuring that the two branches are trained from the same initial noise state.

C RELATIVE ATTRIBUTE TRAINING

We train a RA scorer following the pairwise ranking formulation of Parikh & Grauman (2011), following its application to speech emotion intensity estimation in Zhou et al. (2023). The scorer is trained exclusively on the Chinese training split of ESD, using 150 utterances per emotion category (neutral, angry, happy, sad, and surprised) from speakers 0001–0010. The test and evaluation splits are strictly excluded from training.

We extract 88-dimensional openSMILE eGeMAPSv02 functionals and fit a `StandardScaler` on the pooled training data. For each emotion, a separate linear ranking model is trained by contrasting emotional utterances against neutral utterances. Specifically, pairwise differences are constructed as $\mathbf{D} = [\mathbf{P} - \mathbf{N}; \mathbf{N} - \mathbf{P}]$ with labels $\mathbf{y} = [+1; -1]$. We use a `LinearSVC` with $C = 0.1$, no intercept, and a maximum of 20,000 iterations.

Table 7: Training hyperparameters.

Hyperparameter	Configuration
Optimiser	AdamW
Learning rate	10^{-5}, constant
Gradient clipping	1.0
Precision	FP32
Global batch size	8
Distributed training	8
Total training steps	400,000
Flow-matching time	$t \sim U(0.05, 0.95)$, one sample per utterance
Emotion intensity	$\alpha \sim U(0, 1)$, one sample per utterance
$\sigma_{\min}$	10^{-6}
Noise	One Gaussian noise sample per utterance, shared by both targets
λ	1.0
Random seeds	1234
GPU device	NVIDIA H100 80GB HBM3

For an input utterance, its intensity score is computed as

$\mathrm{RA}(x) = \frac{\mathbf{z}^\top \mathbf{w} - l}{h - l}$, where $\mathbf{z}$ is the standardised feature vector, $\mathbf{w}$ is the learned ranking direction, and l and h are the mean scores of the neutral and corresponding emotional training samples, respectively. Thus, the calibrated scale assigns approximately 0 to neutral speech and 1 to emotional speech.

For cross-dataset evaluation, the RA scorer is applied without retraining. Because the absolute score can vary across speakers and corpora, each evaluation set is independently calibrated using its real neutral and real emotional samples. Therefore, absolute RA values are compared only within each evaluation layer, while cross-layer comparisons are based on the corresponding EIT measures.

We further validate the trained RA ranker on external datasets with independently annotated emotion intensity levels. On CREMA-D, which provides neutral, low-, mid-, and high-intensity emotional utterances, the mean RA scores increase consistently from -0.22 to 0.15, 0.42, and 1.10 for low, mid, and high intensity, respectively. On RAVDESS, the corresponding scores increase from 0.62 for normal emotion to 1.14 for strong emotion. These consistent monotonic trends across independently annotated datasets support the reliability of the RA ranker for measuring relative emotional intensity.

D Emotion related feature changes

We visualise the acoustic features changes according to α. For the investigated emotions, pitch, loudness, and spectral-envelope features vary monotonically with α, whereas formants and vocal perturbation show less consistent trends, as shown in Figure 3.

We additionally find that sadness exhibits a distinctive pattern compared with the other emotions in the trajectory visualisation, showing the greatest separation from the remaining emotion trajectories. This may be related to its relatively low-arousal nature (Jordan et al., 2025). Sadness also shows the weakest controllability among the four emotions, which can be partly attributed to its reliance on rhythm-related characteristics that are difficult to control through α. As shown in Table 8, the two dimensions with the largest shifts towards real sadness, `VoicedSegmentsPerSec` (-0.51) and `MeanUnvoicedSegmentLength` ($+0.44$), exhibit almost no corresponding change as α increases (-0.02 and -0.07, respectively). Moreover, several other rhythm-related features change in directions opposite to those observed in real sadness. Overall, increasing α tends to produce shorter and more compact voiced segments, whereas sadness is associated with slower speech and longer pauses. This mismatch may partly explain the weaker controllability of sadness.

E Baseline model intensity adjustment

As mentioned in the paper, we have performed different strategies for EC, IC, and CC models. The implementation details are as follows.

Table 8: Rhythm- and duration-related features for sadness. Δ_{raw} denotes the median shift from neutral to real sadness, while the remaining columns show the median shift induced by increasing α.

Feature	Δ_{raw}	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1.0$	$\alpha = 1.25$
VoicedSegmentsPerSec	-0.51	-0.10	-0.20	-0.10	-0.37	-0.08	-0.12
MeanVoicedSegmentLengthSec	+0.03	+0.10	-0.26	-0.44	-0.12	-0.25	-0.29
StddevVoicedSegmentLengthSec	+0.18	-0.01	-0.19	-0.48	-0.45	-0.45	-0.54
MeanUnvoicedSegmentLength	+0.44	-0.08	+0.10	+0.13	+0.11	+0.04	-0.15
StddevUnvoicedSegmentLength	+0.29	-0.31	-0.33	-0.46	-0.49	-0.53	-0.42

RAR (EC): RAR represents emotion using a five-dimensional relative-attribute (RA) vector. Following its original formulation, the target emotion is controlled by the corresponding dimension of the attribute vector, where 0 indicates that the target emotion is fully present and larger values indicate weaker presence. We set the target-emotion dimension to 0.8, 0.4, and 0.0 to obtain weak, medium, and strong emotion, respectively, while keeping the other four dimensions fixed at 1.0. The resulting attribute vector is then provided to the two-stage seq2seq model.

EmoKnob (EC): EmoKnob controls emotion intensity by scaling a fixed emotion direction in the speaker embedding space. Specifically, given a speaker embedding $\mathbf{e}$ and an emotion direction $\mathbf{d}$, the modified representation is $\mathbf{e}' = \mathbf{e} + \gamma\mathbf{d}$, where γ controls the strength of the emotion. We use three values corresponding to one-half, three-quarters, and the full value of the default range provided by the released demo.

CosyVoice3 (IC and CC): We evaluate CosyVoice3 using two settings. In the native prompt-based setting, intensity is controlled through degree adverbs in the instruction, using "a little", "moderately", and "very" to represent weak, medium, and strong emotion, respectively. The instruction is provided in the short form required by the model rather than as a complete sentence, avoiding the possibility that the instruction itself is generated as speech. We additionally construct an embedding-interpolation setting by extracting the token embeddings of two emotion instructions and linearly interpolating the instruction embeddings. Specifically, the interpolated embedding is $\mathbf{e}_\alpha = (1-\alpha)\mathbf{e}_{\text{neutral}} + \alpha\mathbf{e}_{\text{emotion}}$, where $\alpha \in 0.5, 0.75, 1.0$. At $\alpha = 1.0$, this construction exactly reduces to the original emotion instruction.

EmotiVoice (IC and CC). We proform the same as what we did for Cosyvoice3.

PromptTTS. PromptTTS provides emotion through natural-language style prompts. We control intensity using the degree adverbs "slightly", "moderately", and "very", resulting in prompts of the form "The speaker sounds slightly/moderately/very emotion." These three prompts are used as the weak, medium, and strong conditions, respectively.

F Localised emotion intensity adjustment mel-spectrogram visualisation

We visualise a case study of localised emotion intensity adjustment mel-spectrogram here.

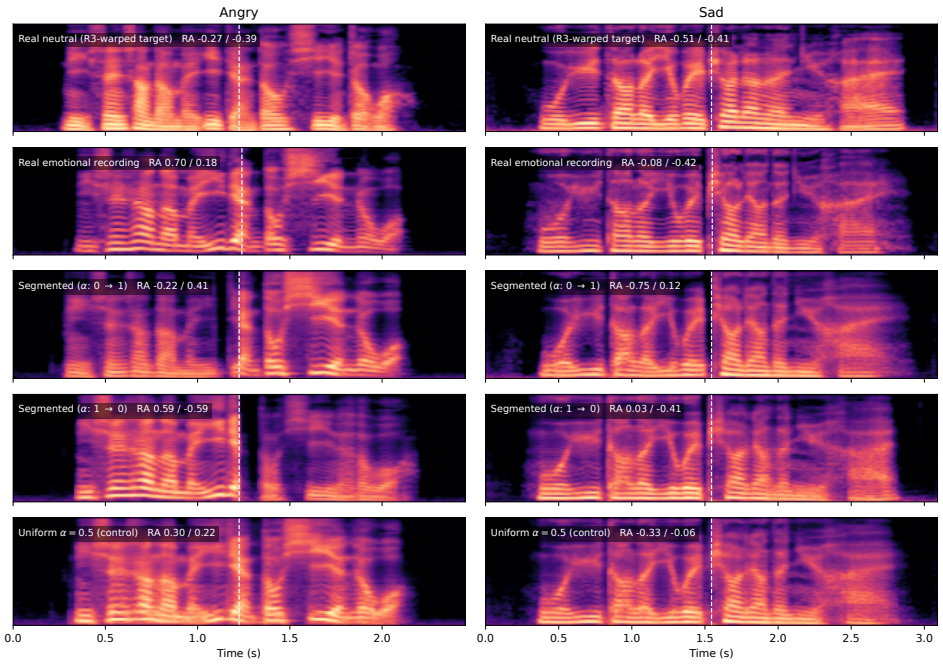


Figure 4: Mel-spectrograms of an utterance generated with different α values in its first and second halves. *Uniform* denotes that the same α value is used throughout the utterance, while *Segmented* denotes that different α values are used in the two halves. RA x/y indicates the RA scores for the first and second halves, respectively. The *Angry* and *Sad* examples correspond to `0004_text00060` and `0008_text00240`, respectively.