

EmoTra-TTS: Smooth Intra-Utterance Emotion Transitions for Speech Synthesis

Tianchi Liu^{1,†}, Zeyang Song^{2,†}, Tianrui Wang³, Zhipeng Li¹,
Chenglin Xu¹, Yiwen Guo⁴

¹LIGHTSPEED, ²National University of Singapore,
³Nanyang Technological University, ⁴Independent Researcher

[†]Corresponding Authors: {tianchi_liu, zeyang_song}@u.nus.edu

Abstract

Psychological research on emotion dynamics has established that human affect is a continuous, evolving process: emotions rise, decay, and transition within seconds. Current emotional text-to-speech (TTS) systems, however, condition on a single discrete label or static embedding per utterance, fundamentally misaligning with the temporal nature of affect. While recent LLM-based TTS systems may implicitly vary prosody through text understanding, such variation is neither explicitly controllable nor precise enough for targeted intra-utterance transitions. We address three challenges: (1) a *multi-pass flow blending* pipeline synthesizes frame-aligned transition audio, circumventing the scarcity of natural intra-utterance transitions; (2) *dual-stage Valence–Arousal–Dominance (VAD) conditioning* guides prosodic planning in the LLM and acoustic realization in the flow decoder via frame-level VAD embeddings; (3) *direction–magnitude decoupled injection* structurally separates emotion direction from injection magnitude, preventing content degradation. **EmoTra-TTS¹** adds only +0.43% parameters with no latency overhead, achieves **30%–87%** relative improvement on emotion transition quality, corroborated by **64.4%–79.5%** overall win rates in pairwise preference tests against four SOTA baselines and two commercial systems.

1 Introduction

Human emotion is not a static label but a *dynamic process* that unfolds over time. Research on *emotion dynamics* (Davidson, 1998; Kuppens and Verduyn, 2017) has established that affective states fluctuate continuously, exhibiting inertia, gradual decay, and smooth transitions between categories (Verduyn et al., 2009; Cowen and Keltner,

2017), with temporal envelopes that carry communicative intent (Trampe et al., 2015). Russell’s circumplex model (Russell, 1980) and the PAD framework (Mehrabian, 1996) formalize this continuity by representing emotions in a continuous VAD space, supported by neuroscientific evidence (Posner et al., 2005) and continuous gradients between categories (Cowen and Keltner, 2017).

These findings motivate emotion control that is continuous in both emotion space and time, especially for expressive speech scenarios such as storytelling, dubbing, and audiobook narration, where abrupt transitions are highly perceptible.

Despite this foundation, emotional TTS predominantly assigns one categorical label per utterance (Lei et al., 2022). Recent systems offer finer control: EmoSphere-TTS (Cho et al., 2024) uses spherical VAD vectors, EmoKnob (Chen et al., 2024) provides disentangled style knobs, and WeSCon (Wang et al., 2025a) explores discrete word-level control. Yet these utterance- and word-level systems do not model the continuous temporal dynamics of emotion: the rise, transition, and decay that define natural affective experience. Recent LLM-based TTS systems (Hu et al., 2026; Du et al., 2025; Hussain et al., 2025; Leng et al., 2024; Zhou et al., 2025; Li et al., 2026; Wang et al., 2026c; Song et al., 2026) may implicitly vary prosody through text understanding, but such variation is neither user-controllable nor sufficiently precise for targeted emotion transitions within an utterance.

Bridging this gap requires addressing three interconnected challenges: **(1) Data scarcity:** naturally occurring intra-utterance emotion transitions are severely underrepresented in emotional speech corpora, and eliciting them from actors is notoriously difficult (Busso et al., 2008; Zhou et al., 2022; Michel et al., 2026). **(2) Dual-level conditioning:** modern LLM-based TTS (Du et al., 2024; Wang et al., 2023; Anastassiou et al., 2024) decomposes synthesis into prosodic planning (LLM)

¹Demo page: https://liu-tianchi.github.io/EmoTra_DemoPage/; The complete code, including data synthesis and training pipelines: <https://github.com/Liu-Tianchi/EmoTra-TTS>

and acoustic realization (flow decoder), each requiring temporally aligned emotion signals (Scherer, 2003). (3) **Magnitude control**: injecting conditioning into a pretrained decoder risks pushing it outside its operating range; learnable constraints are circumvented by co-adaptation (Locatello et al., 2019). To address these challenges, we propose **EmoTra-TTS** built upon three key components:

- **Synthetic emotion transition data** (§3.2): a multi-pass flow blending pipeline generates frame-aligned transition audio, circumventing the difficulty of recording natural transitions.
- **Dual-stage VAD conditioning** (§3.3, §3.4): VAD tokens condition the LLM for prosodic planning; frame-level VAD embeddings modulate the flow decoder’s speaker pathway.
- **Direction–magnitude decoupled injection** (§3.4.4): *LayerNorm* + *fixed scale* decouples emotion direction (learned) from injection magnitude (kept bounded via re-normalization), mitigating the content-expressiveness trade-off.

2 Related Work

2.1 Emotion Dynamics and Emotional TTS

Emotion dynamics research (Davidson, 1998; Kuppens and Verduyn, 2017) has established that affect possesses an intrinsic temporal microstructure. Dimensional models (Mehrabian, 1996) with acoustic correlates (Scherer, 2003; Banse and Scherer, 1996) motivate VAD transitions, not merely VAD points, as conditioning signals.

Emotional TTS has evolved from discrete emotion embeddings (Wang et al., 2018) and prosody transfer (Skerry-Ryan et al., 2018) to continuous VAD control (Cho et al., 2024; Inoue et al., 2025; Zhou et al., 2024). EmoShift (Zhou et al., 2026) and CoCoEmo (Wang et al., 2026b) explore activation steering, while EMORL-TTS (Li et al., 2025) applies reinforcement learning for emotion control. Recent efforts further explore finer-grained control: Wu et al. (2024) condition a flow-matching decoder on frame-level arousal–valence from reference audio, Liang et al. (2026) explore segment-level modeling for intra-utterance emotion variation, and WeSCon (Wang et al., 2025a) introduces word-level discrete control. While these works address related forms of fine-grained control, explicit and controllable intra-utterance transitions along

continuous VAD dimensions, specified as a user-defined trajectory rather than reference-driven or discrete control, remain underexplored (Cui et al., 2025b; Xie et al., 2025). Recent benchmarks report the same limitation for commercial systems: InstructTTSEval (Huang et al., 2025) shows Gemini and VoxInstruct handle only rudimentary changes (e.g., voice rise), not genuine emotion change; EmergentTTS-Eval (Manku et al., 2025) lists natural shifts to contrasting emotions among its most challenging cases; EmoS (Wang et al., 2026a) finds commercial models (e.g., GPT-4o-Audio, Gemini 2.5 Pro) lag far behind humans on cross-turn emotion transitions; and MINT-Bench (Chen et al., 2026) identifies paralinguistic control, including explicit intra-utterance emotion transitions, as a major bottleneck for commercial systems (e.g., Gemini, GPT-4o mini TTS, ElevenLabs). In contrast, EmoTra-TTS models such continuous transitions.

2.2 Synthetic Data for Speech Tasks

Synthetic data has proven effective for speech tasks: TTS-augmented ASR training (Yang et al., 2025c), speech translation (Zhao et al., 2023), speech security (Hwang et al., 2026; Liu et al., 2024, 2025b), speaker recognition (Tao et al., 2025; Liu et al., 2025a), and emotion recognition (Ma et al., 2024a,b). We extend this paradigm to *within-utterance* transition synthesis via mel-space flow blending, which existing corpora (Busso et al., 2008; Zhou et al., 2022) cannot provide.

2.3 LLM-Based TTS and Flow Matching

VALL-E (Wang et al., 2023) pioneered LLM-based zero-shot TTS, followed by Seed-TTS (Anastasiou et al., 2024), NaturalSpeech 3 (Ju et al., 2024), MaskGCT (Wang et al., 2025c), Llasa (Ye et al., 2025), Spark-TTS (Wang et al., 2025b), Sticker-TTS (Chen et al., 2025a), and GLM-TTS (Cui et al., 2025a). CosyVoice2 (Du et al., 2024), our backbone, employs supervised semantic tokens with a causal conditional flow matching (CFM) (Lipman et al., 2023) decoder. CFM has demonstrated strong performance (Mehta et al., 2024; Chen et al., 2025b; Yang et al., 2025a). Our work extends CFM conditioning to support time-varying emotion.

2.4 Conditioning Injection in Frozen Models

Injecting auxiliary signals into frozen models is a growing paradigm. In audio, FiLM (Perez et al., 2018) conditioning appears in WaveGrad (Chen et al., 2021) and DiffWave (Kong et al., 2021);

LoRA (Hu et al., 2022) and DoRA (yang Liu et al., 2024) enable efficient adaptation. A common thread is the need to control influence strength: e.g., ControlNet (Zhang et al., 2023) uses zero-initialization. Our direction–magnitude decoupling takes a different route: unlike soft penalties that can be circumvented (Locatello et al., 2019), *Layer-Norm + fixed scale* constrains the input-dependent MLP output through re-normalization, so that the injection magnitude remains bounded and close to the speaker-embedding norm.

3 Method

Research on emotion dynamics shows that affective episodes unfold in phases: from rising intensity to transition or decay (Davidson, 1998; Verduyn et al., 2009; Kuppens and Verduyn, 2017). Within a single utterance, this manifests as a shift from one affective state to another. We model this minimal temporal unit as a transition from an *initial emotion* v_{init} to a *resultant emotion* v_{res} , representing the non-trivial dynamic pattern that existing single-emotion TTS systems cannot express.

3.1 Problem Formulation

Given text $\mathcal{T}$, an initial emotion $v_{\text{init}} = (V_{\text{init}}, A_{\text{init}}, D_{\text{init}}) \in [0, 1]^3$, and a resultant emotion $v_{\text{res}} \in [0, 1]^3$, we seek to synthesize speech where emotion transitions smoothly from v_{init} to v_{res} over the utterance duration. Framed in terms of emotion dynamics (Davidson, 1998; Kuppens and Verduyn, 2017), the system must model the temporal transition of an emotional episode, not merely its static endpoint.

We adopt the VAD dimensional model (Russell, 1980; Mehrabian, 1996) for three reasons: (a) it subsumes discrete emotion categories as specific points in the continuous space (Posner et al., 2005); (b) it naturally supports interpolation for constructing smooth transitions, mirroring gradual emotion shifts in emotion dynamics research (Verduyn et al., 2009; Cowen and Keltner, 2017); (c) its axes have well-established acoustic correlates (Scherer, 2003; Banse and Scherer, 1996), ensuring that changes in VAD space correspond to perceptually meaningful acoustic variations.

3.2 Synthetic Emotion Transition Data

3.2.1 Motivation

Existing emotional speech corpora are predominantly annotated with single emotion labels (Busso

et al., 2008; Zhou et al., 2022). Eliciting natural within-utterance transitions from actors requires precisely controlling onset timing, direction, and intensity while maintaining natural prosody and content. These are demands that professional actors find challenging. This data bottleneck has limited prior work on intra-utterance emotion dynamics.

We circumvent this limitation by generating *synthetic emotion transition data* from a pretrained zero-shot TTS system. The key insight is that flow-based decoders condition on speaker/emotion embeddings independently of linguistic content: the same speech token sequence can be decoded under different emotion conditionings. This produces mel spectrograms that are temporally aligned but differ in emotion. This content-invariance property enables smooth blending in the mel spectrogram space without temporal misalignment.

3.2.2 Multi-Pass Flow Blending

For a given text and emotion pair (initial, resultant), the pipeline proceeds as follows:

(1) **Shared token generation.** The LLM generates a single speech token sequence $s_{\text{init}} = [s_1, \dots, s_M]$ conditioned on the initial emotion. This sequence encodes the prosodic structure and is shared across all subsequent flow passes, ensuring frame-level temporal alignment.

(2) **Multi-pass flow decoding.** Three independent flow passes produce emotion-specific mel spectrograms (Fig. 1, Part 1 to Part 3):

$$\text{mel}_{\text{init}} = \text{Flow}(s_{\text{init}}, e_{\text{init}}, c_{\text{init}}) \quad (1)$$

$$\text{mel}_{\text{tra}} = \text{Flow}(s_{\text{init}}, e_{\text{res}}, c_{\text{init}}) \quad (2)$$

$$\text{mel}_{\text{res}} = \text{Flow}(s_{\text{init}}, e_{\text{res}}, c_{\text{res}}) \quad (3)$$

where e denotes the speaker embedding and c the prompt condition, both derived from the target speaker. The intermediate mel_{tra} uses the initial emotion’s masked mel spectrogram context but the resultant emotion’s speaker embedding, forming a transition bridge between the two emotional states.

(3) **Sigmoid crossfade blending.** The three mel spectrograms are blended via sigmoid crossfades:

$$\sigma(t) = \frac{1}{1 + e^{-\kappa(t-t_c)}} \quad (4)$$

with steepness κ and center t_c , providing near-binary blending at the edges (preserving pure emotion regions) with a smooth crossfade in the center, avoiding audible artifacts from linear blending or hard concatenation. The choice of the crossfade

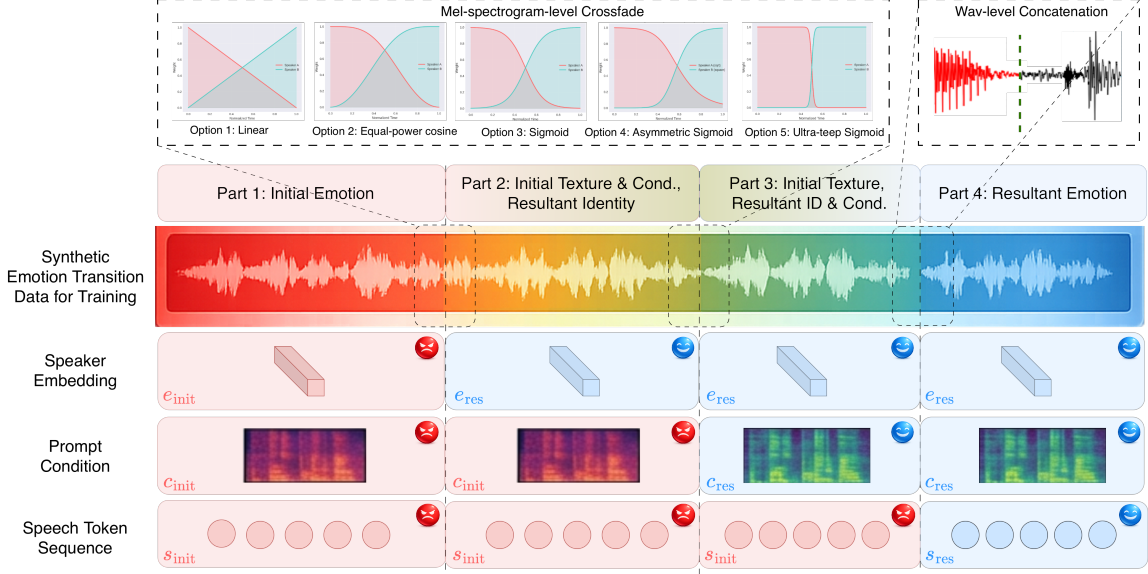


Figure 1: Synthetic emotion transition data generation pipeline. Samples are available on the demo page.

curve among several candidates, as shown in the upper part of Fig. 1, is analyzed in Appendix F. After vocoding, a fourth part appends a pure resultant-emotion utterance (Part 4 in Fig. 1).

(4) Quality filtering. Whisper ASR verifies content integrity; samples exceeding CER threshold τ_{CER} or WER threshold τ_{WER} are discarded, as the blending process can occasionally produce phonetically ambiguous regions near transition boundaries.

3.2.3 Advantages over Recordings

This pipeline offers three advantages. First, it produces arbitrary emotion pairs: any combination of available discrete emotions can be synthesized. Second, transition timing and shape are precisely controlled via segment ratios and crossfade parameters, providing consistent supervision. Third, it is scalable: multi-GPU parallel synthesis generates thousands of transition samples per hour, whereas studio recording of emotion transitions requires extensive actor preparation and multiple takes.

3.2.4 VAD Annotation

VAD values for the initial and resultant emotions are obtained by annotating their reference speech with a pretrained wav2vec2-based speech emotion recognition model (Wagner et al., 2023).

3.3 LLM with Temporal VAD Tokens

3.3.1 VAD Projection

A projection network $\phi_{\text{proj}} : \mathbb{R}^3 \rightarrow \mathbb{R}^{d_{\text{LLM}}}$ maps VAD vectors to the LLM embedding space:

$$\phi_{\text{proj}}(\mathbf{v}) = W_2 \cdot \text{ReLU}(\text{LN}(W_1 \mathbf{v} + b_1)) + b_2 \quad (5)$$

where $W_1 \in \mathbb{R}^{d_h \times 3}$, $W_2 \in \mathbb{R}^{d_{\text{LLM}} \times d_h}$, and LN denotes LayerNorm. The intermediate normalization stabilizes training given the large dimensionality expansion from 3 to d_{LLM} , following best practices for cross-modal projection in multimodal learning (Li et al., 2023).

3.3.2 Temporal Token Design

A key design question is how to encode a continuous VAD transition in a discrete token sequence. A single VAD embedding provides only a static global label, while using only the two endpoints $\mathbf{v}_{\text{init}}$ and $\mathbf{v}_{\text{res}}$ leaves the transition trajectory implicit. In our preliminary experiments, this produced perceptibly abrupt transitions. We therefore uniformly sample N_{tra} intermediate points along the line segment between them:

$$\mathbf{q}_{\text{tra},k} = \phi_{\text{proj}}((1-\alpha_k)\mathbf{v}_{\text{init}} + \alpha_k\mathbf{v}_{\text{res}}), \alpha_k = \frac{k}{N_{\text{tra}} + 1} \quad (6)$$

for $k = 1, 2, \dots, N_{\text{tra}}$. Together with the endpoint tokens $\mathbf{q}_{\text{init}}$ and $\mathbf{q}_{\text{res}}$, the $N_{\text{tra}} + 2$ VAD tokens are prepended to the text tokens. This balances transition resolution with avoiding excessive conditioning that may dilute the LLM’s attention to text.

3.3.3 Training Objective

The input sequence is $[\text{SOS}][\mathbf{q}_{\text{init}} \quad \mathbf{q}_{\text{tra},1} \dots \mathbf{q}_{\text{tra},N_{\text{tra}}} \quad \mathbf{q}_{\text{res}}][\text{text}][\text{task}][\text{speech}][\text{EOS}]$, with the loss:

$$\mathcal{L}_{\text{LLM}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{hidden}} \quad (7)$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss on speech tokens. The auxiliary hidden-state reconstruction

3.4.4 Direction–Magnitude Decoupled Injection

The choice of f_{inject} is the central design challenge. An MLP conditioning vector entangles two semantically distinct factors: a *direction* (specifying *which* emotion) and a *magnitude* (specifying *how strongly* it modulates the decoder). Our ablation (§5.1 and Appendix E) shows that unconstrained training allows MLP output norms to exceed the speaker embedding norm, causing content degradation. Learnable constraints (e.g., FiLM (Perez et al., 2018)) fail because the MLP co-adapts to circumvent them, paralleling the well-documented insufficiency of soft penalties for disentanglement (Locatello et al., 2019). The flow matching loss, averaging over all mel frames, provides no gradient signal that explicitly constrains the emotion norm, creating exactly the conditions under which soft constraints fail.

We enforce a *hard* separation through architectural constraints:

$$f_{\text{inject}}(\mathbf{x}) = \epsilon \cdot \text{LayerNorm}(\mathbf{x}) \quad (12)$$

LayerNorm re-normalizes the MLP output, so the MLP learns the emotion direction while the injection magnitude stays controlled; the trainable affine parameters (γ, β) are few (160) and identity-initialized, and the measured post-training norm is reported in §5.4. The non-trainable scalar ϵ independently controls injection strength, with no gradient path connecting the two factors. Zero-initialized output layers ensure $\mathbf{e}_{\text{emo}}(t) = \mathbf{0}$ at training start (Zhang et al., 2023). This design mirrors the direction–magnitude decomposition in DoRA (yang Liu et al., 2024), which shows that fine-tuning primarily changes weight *directions*. The complete injection is:

$$\hat{\mathbf{e}}(t) = \mathbf{e} + \epsilon \cdot \text{LN}(\psi_{\text{down}}(\phi_{\text{recon}}(\phi_{\text{proj}}(\mathbf{v}(t)))) \quad (13)$$

All decoder parameters are frozen; only ψ_{down} and LayerNorm affine parameters are trainable. Speech tokens are pre-computed from Stage 1, fully decoupling the two training stages.

3.4.5 Flow Matching Training Objective

The flow decoder is trained with a conditional flow matching (CFM) (Lipman et al., 2023) objective. Given target mel spectrogram $\mathbf{x}_1 \in \mathbb{R}^{d_{\text{mel}} \times T}$, noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and time $t \sim \mathcal{U}(0, 1)$, we construct the interpolated sample $\mathbf{x}_t = (1 - (1 - \sigma_{\min})t) \mathbf{z} +$

$t \mathbf{x}_1$ and ground-truth velocity $\mathbf{u} = \mathbf{x}_1 - (1 - \sigma_{\min}) \mathbf{z}$, with $\sigma_{\min} = 10^{-6}$. The loss minimizes:

$$\mathcal{L}_{\text{Flow}} = \mathbb{E}_{t, \mathbf{x}_1, \mathbf{z}} \left[\|\hat{\mathbf{u}}_{\theta}(\mathbf{x}_t, t, \boldsymbol{\mu}, \hat{\mathbf{e}}(t)) - \mathbf{u}\|^2 \right] \quad (14)$$

where $\hat{\mathbf{u}}_{\theta}$ is the velocity estimator and $\hat{\mathbf{e}}(t)$ is from Eq. (13). Classifier-free guidance (Ho and Salimans, 2021) is applied by randomly dropping conditioning with probability $p_{\text{cfg}} = 0.2$.

4 Experimental Setup

Base model. CosyVoice2-0.5B (Du et al., 2024).

Source corpus. EmoVoice-DB (Yang et al., 2025b), containing five speakers and seven emotion categories (*neutral, happy, sad, angry, fearful, disgusted, surprised*). We apply cross-validated filtering: a pretrained wav2vec2-based VAD predictor (Wagner et al., 2023)² estimates continuous scores, and only samples falling within literature-informed acceptance ranges (Russell, 1980; Mehra-bian, 1996; Fontaine et al., 2007) (Table 6 in Appendix C) are retained, removing mislabeled or acoustically ambiguous samples.

Synthetic data. Emotion transition pairs are generated via multi-pass flow blending (§3.2), covering all emotion pair combinations ($\sim 100\text{K}$ utterances). Quality filtering uses Whisper large-v3 (Radford et al., 2023) with $\tau_{\text{CER}} = 0.10$; minimum VAD change threshold $\delta_{\text{vad}} = 0.35$ (Chebyshev distance). Data split: 90/10 train/validation.

Training. Stage 1 (LLM) fine-tunes the VAD projection and reconstruction modules; Stage 2 (Flow) freezes all decoder parameters and trains only the injection MLP and LayerNorm affine ($\sim 280\text{K}$ parameters). Fixed emotion scale $\epsilon = 0.07$ yields an effective norm matched to the speaker embedding norm ($\|\mathbf{e}\|_2 \approx 0.622$). Full hyperparameters are provided in Appendix A.

Evaluation. We conduct a blind listening test with a professional panel of 21 raters. To enable a more fine-grained and independent evaluation of different aspects, we decompose MOS into three metrics: **MOS-Qua**, for speech naturalness and quality (ignoring emotional correctness); **MOS-Emo** for emotion adequacy; and **MOS-Tra** for emotion transition smoothness. We compute **speaker similarity (SIM)**³, with the reference defined as the average embedding of 200 samples

²<https://github.com/audeering/w2v2-how-to>

³<https://github.com/BytedanceSpeech/seed-tts-eval>

covering all emotions. **Word error rate (WER)** is evaluated by Whisper large-v3⁴ (Radford et al., 2023). We further include two pairwise preference tests with partially overlapping rater pools: one against the open-source baselines (11 raters, 160 pairs, $\sim 6\times$ coverage; §5.5, Appendix B.2) and a supplementary one against two commercial systems (12 raters, 80 pairs, $\sim 8\times$ coverage; §5.6, Appendix B.3); both report **win rates**, **95% CIs**, and **inter-rater agreement**. In addition, we compute the **Prosodic Jerk Ratio (JR-F0)** in Appendix D as an objective metric for transitions.

5 Results and Analysis

5.1 Injection Architecture Comparison

Table 1: Injection architecture comparison.

Variant	Quality Evaluation			Emotion Evaluation	
	MOS-Qua	SIM	WER	MOS-Emo	MOS-Tra
Linear	3.60 $\pm$ 0.12	0.707	1.65	3.39 $\pm$ 0.15	3.33 $\pm$ 0.16
Dir-Mag (ours)	3.54 $\pm$ 0.13	0.701	1.77	3.52$\pm$0.15	3.63$\pm$0.17

We evaluate seven injection architectures for Stage 2 (Table 1; full results and failure analysis in Appendix E, Table 8). Linear yields the best speech quality (MOS-Qua 3.60, SIM 0.707, WER 1.65%) but weaker emotion (MOS-Emo 3.39, MOS-Tra 3.33), suggesting that simple additive injection preserves content but lacks expressiveness. Our Dir-Mag design structurally decouples emotion direction from injection magnitude, achieving the best emotion performance (MOS-Tra **3.63**, MOS-Emo **3.52**) with acceptable speech quality trade-off (MOS-Qua 3.54, SIM 0.701, WER 1.77%).

5.2 Comparison with Open-source Systems

Table 2 compares EmoTra-TTS against baselines spanning seven conditioning strategies. These baselines are all recent open-source systems, including CosyVoice2 (Du et al., 2024), EmoVoice (Yang et al., 2025b), WeSCon (Wang et al., 2025a), Qwen3-TTS (Hu et al., 2026), and MOSS-TTS (MOSS-TTS Team, 2026), which together reasonably represent the current instruction-following capability of open-source TTS. Closed-source commercial systems are treated separately: recent benchmarks report the same intra-utterance transition limitation (§2), and we provide a direct comparison against two of them (GPT-4o mini TTS and ElevenLabs v3) in §5.6.

⁴<https://huggingface.co/openai/whisper-large-v3>

Neutral baselines (a) achieve the highest quality (up to MOS-Qua 4.02), but produce minimal emotion rendering and transition quality. MOS-Qua primarily reflects speech quality rather than emotional correctness; therefore, high MOS-Qua scores under neutral conditions are expected.

Instruction-based systems (b, c) show that current TTS systems cannot translate natural-language emotion instructions into precise intra-utterance emotion transitions. The instruction-following capability determines emotion rendering (MOS-Emo ranging from 2.48 to 3.33); adding a neutral prompt (c) preserves speaker similarity (SIM 0.634–0.640) but anchors the decoder to neutral prosody, further limiting emotion expressiveness.

Prompt-based concatenation (d) achieves the strongest baseline emotion rendering (up to MOS-Emo 3.89) since each segment is cloned from a matching emotional reference. However, stronger emotional expressiveness increases perceptual discontinuity at boundaries. MOSS-TTS exemplifies this: it achieves the highest baseline MOS-Emo (3.89) yet the lowest MOS-Tra (1.94), as vivid emotion makes boundary artifacts more noticeable.

High Failure Cost. We further quantify this effect. When discrete baselines such as MOSS-TTS produce abrupt emotion shifts (Category d), naturalness drops significantly compared to neutral conditions (Category a), showing that a single failed transition can degrade overall quality. Moreover, this degradation increases with higher emotion expressiveness (MOS-Emo), indicating that failure cost grows with expressiveness. This emphasizes the importance of smooth transition modeling in expressive TTS due to its high perceptual impact.

WeSCon (e) controls intra-utterance emotion variation in a single decoding pass via word-level prompt conditioning. While this enables finer-grained control, its discrete per-word assignment lacks smooth transition modeling between segments. As a result, it introduces boundary discontinuities, leading to lower MOS-Tra (2.39).

All above systems achieve MOS-Tra ≤ 2.80 , indicating that controllable smooth intra-utterance emotion transition remains an open problem.

Our synthetic data (f) achieves the highest emotion rendering performance (MOS-Emo **4.05**) and a MOS-Tra of 3.55, validating the effectiveness of the synthetic data. While outperforming all baselines, the multi-pass flow blending design substantially reduces synthesis efficiency (see §5.7), motivating the training of our proposed EmoTra-TTS.

Table 2: Comparison with existing TTS systems on the test set. SIM: speaker similarity. WER: word error rate (%). We decompose MOS into MOS-Qua, MOS-Emo, and MOS-Tra for a more fine-grained and independent evaluation of speech quality, emotion rendering adequacy, and intra-utterance emotion transition smoothness, respectively. Subjective evaluation results are reported as means with 95% CIs. Our base system is CosyVoice2.

Model	Emotion Conditioning	Quality Evaluation			Emotion Rendering	Emotion Transition
		MOS-Qua↑	SIM↑	WER↓	MOS-Emo↑	MOS-Tra↑
(a) Neutral prompt speech; no emotion conditioning						
Qwen3-TTS (Hu et al., 2026)	—	4.02 ±0.09	0.726	1.05	2.72±0.17	2.40±0.18
MOSS-TTS (MOSS-TTS Team, 2026)	—	3.98±0.09	0.754	1.01	2.76±0.18	2.36±0.17
CosyVoice2 (Du et al., 2024)	—	3.87±0.10	0.711	1.21	2.53±0.16	2.32±0.17
(b) Natural-language emotion instruction (e.g., “speak happily then sadly”)						
Qwen3-TTS (Hu et al., 2026)	Instruct	3.79±0.10	N/A	1.28	3.33±0.15	2.67±0.19
MOSS-TTS (MOSS-TTS Team, 2026)	Instruct	3.22±0.15	N/A	2.55	2.80±0.17	2.13±0.16
(c) Natural-language emotion instruction with neutral prompt speech						
CosyVoice2 (Du et al., 2024)	Instruct	3.63±0.12	0.634	1.60	2.64±0.14	2.15±0.16
EmoVoice (Yang et al., 2025b)	Instruct	3.59±0.13	0.640	1.42	2.48±0.15	2.31±0.17
(d) Emotional prompt speech: generate each emotion segment separately, then concatenate						
Qwen3-TTS (Hu et al., 2026)	Prompt	3.46±0.14	0.730	0.92	3.70±0.14	2.80±0.18
MOSS-TTS (MOSS-TTS Team, 2026)	Prompt	2.85±0.17	0.735	2.64	3.89±0.12	1.94±0.16
CosyVoice2 (Du et al., 2024)	Prompt	2.97±0.15	0.703	1.83	3.82±0.12	2.26±0.17
(e) Emotional prompt speech with word-level emotion generation control						
WeSCon (Wang et al., 2025a)	Prompt	2.80±0.15	0.649	2.92	3.53±0.14	2.39±0.17
(f) Our synthetic training data: multi-pass flow blending (§3.2), decoded by base CosyVoice2						
Synth. Data (Ours)	Prompt	3.58±0.11	0.707	1.49	4.05 ±0.12	3.55±0.15
(g) Proposed: continuous VAD conditioning with smooth intra-utterance transition						
EmoTra-TTS (Ours)	VAD	3.54±0.13	0.701	1.77	3.52±0.15	3.63 ±0.17
– w/o flow SFT	VAD	3.62±0.12	0.709	1.65	2.70±0.17	2.58±0.18

EmoTra-TTS (g) synthesizes the utterance in a single decoding pass with continuous frame-level VAD interpolation, enabling smooth rather than abrupt transitions, yielding the best MOS-Tra (**3.63**) with competitive results on other metrics. Among systems with explicit emotion rendering (Categories d, e), EmoTra-TTS achieves the highest MOS-Qua, corroborated by an objective metric (Appendix D) showing lower F0 jerk at transitions.

5.3 Ablation: Dual-Stage SFT

We compare three configurations in Table 2: baseline CosyVoice2 (a), LLM-only SFT (“w/o flow SFT” in group g), and full EmoTra-TTS (g). LLM-only SFT improves emotion performance (MOS-Emo 2.53 $\rightarrow$ 2.70, MOS-Tra 2.32 $\rightarrow$ 2.58) while preserving speech quality, confirming that temporal VAD tokens provide meaningful prosodic planning signals. Adding flow SFT yields a further substantial gain (MOS-Emo 2.70 $\rightarrow$ 3.52, MOS-Tra 2.58 $\rightarrow$ 3.63), demonstrating *complementary* contributions: the LLM plans emotion-appropriate prosody, while the flow stage renders acoustic realization through frame-level VAD modulation.

5.4 Ablation: Fixed Scale ϵ

The fixed scale ϵ controls the emotion-to-speaker norm ratio $r = \|e_{\text{emo}}\|_2 / \|e\|_2$ ($\|e\|_2 \approx 0.622$),

Table 3: Ablation on fixed scale ϵ . The effective emotion norm is $\epsilon \cdot \sqrt{d_{\text{mel}}}$; the speaker embedding norm is $\|e\|_2 \approx 0.622$.

ϵ	Eff. norm	Quality Evaluation			Emotion Evaluation	
		MOS-Qua	SIM	WER	MOS-Emo	MOS-Tra
0.04	0.358	3.60 ± 0.12	0.706	1.69	3.46 ± 0.15	3.31 ± 0.16
0.07	0.626	3.54 ± 0.13	0.701	1.77	3.52 ± 0.15	3.63 ± 0.17
0.10	0.894	2.59 ± 0.18	0.491	5.30	2.34 ± 0.17	1.85 ± 0.17

where $r < 1$ means speaker identity dominates, $r > 1$ means emotion dominates, and $r \approx 1$ is parity. The results reveal a sensitivity boundary: $\epsilon = 0.04$ ($r = 0.58$) preserves speech quality (MOS-Qua 3.60) but attenuates emotion; $\epsilon = 0.10$ ($r = 1.44$) triggers catastrophic degradation as emotion injection overwhelms content and speaker identity. The optimal $\epsilon = 0.07$ ($r \approx 1.0$) achieves the best transition quality (MOS-Tra **3.63**), confirming that parity between emotion and speaker norms defines the safe operating regime.

At the deployed checkpoint ($\epsilon = 0.07$), the *measured* post-training injection norm is $\|e_{\text{emo}}(t)\|_2 \approx 0.667$, about $1.07\times$ the speaker-embedding norm (≈ 0.622); the 0.626 in Table 3 is the *nominal* $\epsilon \cdot \sqrt{d_{\text{mel}}}$. That the measured norm stays close to the speaker norm confirms that the trainable Layer-Norm affine parameters (γ, β ; 160 in total, identity-initialized) do not inflate the injection magnitude, keeping it far below the failure regime, where an

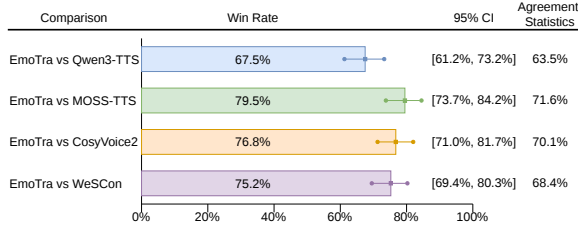


Figure 3: Paired preference of EmoTra-TTS over four baselines in Categories (d) and (e) of Table 2.

input-dependent MLP inflates the norm to $\approx 2\times$ (MLP-Add) or $3\text{--}5\times$ (FiLM) and collapses content (Appendix E).

5.5 Pairwise Preference Evaluation

Beyond the per-dimension MOS above, we conduct a paired preference test for an overall A/B comparison. Figure 3 shows EmoTra-TTS is preferred over all four baselines with win rates of 67.5%–79.5%, and every 95% CI lower bound exceeds 60%, well above the 50% chance level. Inter-rater agreement of 63.5%–71.6% supports these preferences.

5.6 Comparison with Commercial Systems

As a supplementary study added during the revision period, we additionally compare EmoTra-TTS against two closed-source commercial systems spanning the two widely adopted commercial control paradigms. This study uses a partially overlapping but not identical rater pool and set of pairs; it therefore constitutes a valid EmoTra-vs.-commercial comparison but is *not* directly comparable to the paired results in §5.5 and Figure 3. The full protocol is provided in Appendix B.3.

ElevenLabs v3 (latest), from one of the most widely used commercial TTS providers, controls emotion through discrete audio tags rather than natural-language prompts. **GPT-4o mini TTS**, from OpenAI’s GPT-4o audio family, is controlled via a natural-language instruction prompt; as it is the same family used to synthesize our source corpus EmoVoice-DB (Yang et al., 2025b), the comparison is especially stringent. The two systems thus span the natural-language instruction setting (GPT-4o mini TTS) and the discrete-tag setting (ElevenLabs v3) available in commercial products; configuration and prompting details are given in Appendix B.3.

As shown in Table 4, EmoTra-TTS is preferred over both commercial systems in overall preference (72.0% and 64.4%), with larger margins on emotion accuracy and transition naturalness (74.5%

Table 4: Comparison with commercial instruction-following systems. Values are the percentage of pairs in which EmoTra-TTS is preferred (win rate), with 95% CIs. “Emo. & Tra.” evaluates emotion rendering and transition naturalness; “Overall” is overall preference. “Agr.” denotes inter-rater agreement.

EmoTra-TTS vs.	Emo. & Tra.			Overall		
	Win	95% CI	Agr.	Win	95% CI	Agr.
GPT-4o mini TTS	74.5%	[69.5, 78.9]	67.7%	72.0%	[66.9, 76.6]	63.8%
ElevenLabs v3	70.5%	[65.2, 75.3]	63.5%	64.4%	[59.0, 69.5]	60.7%

and 70.5%); every 95% CI lower bound exceeds the 50% chance level.

5.7 Inference Efficiency

Table 5: Inference efficiency comparison. RTF: real-time factor. Mean $\pm$ std reported. Mem.: GPU memory.

System	RTF $\downarrow$	Latency (s) $\downarrow$	Params	Δ Params	Mem.
CosyVoice2	0.270 $\pm$.022	1.75 $\pm$.39	639M	–	3.04 GB
Synth. Data	0.397 $\pm$.045	2.92 $\pm$.47	639M	+0.00%	5.09 GB
EmoTra-TTS	0.244 $\pm$.010	1.74 $\pm$.42	642M	+0.43%	3.05 GB

EmoTra-TTS’s latency is close to the baseline, indicating negligible overhead from the additional 0.43% parameters. The lower RTF results from longer generated audio rather than faster synthesis. The synthetic data pipeline requires $4\times$ flow inference and is intended for offline generation. See Appendix B.4 for test set details.

6 Conclusion

Emotion dynamics research has long established that affect unfolds continuously over time, yet explicit, controllable modeling of the temporal trajectory of emotion within a single utterance is still only partially addressed by existing systems. EmoTra-TTS bridges this gap through multi-pass flow blending for synthetic transition data, dual-stage VAD conditioning for complementary prosodic and acoustic control, and direction-magnitude decoupled injection that constrains the injection magnitude to alleviate the content-expressiveness trade-off. The system adds only +0.43% parameters with no latency overhead and achieves the best perceptual scores on emotion transition smoothness among all evaluated systems, with 30%–87% relative improvement over multiple SOTA TTS baselines and commercial systems, and 64.4%–79.5% win rates in pairwise preference tests. We hope this work offers a useful step toward modeling emotion as a *temporal process* rather than a static attribute, and encourages further exploration of affective speech synthesis.

Limitations

Our work has several limitations:

First, the current experiments are conducted on EmoVoice-DB, which provides limited speaker diversity and language coverage. While we validate EmoTra-TTS on multiple speakers from EmoVoice-DB, a separate model is trained per voice, and zero-shot generalization to arbitrary unseen speakers is not yet demonstrated. This limitation, however, mainly concerns speaker and linguistic generalization rather than the proposed intra-utterance emotion transition mechanism itself. Extending the framework to larger multilingual multi-speaker emotional corpora and incorporating prompt-based speaker conditioning remain important directions for future work.

Second, emotion control stability can be inconsistent for subtle VAD differences, likely improvable with more diverse training data.

Third, the current piecewise linear VAD interpolation provides only a first-order approximation of natural emotion dynamics (Davidson, 1998). More complex nonlinear trajectories may better capture realistic affective evolution within an utterance.

Fourth, a formal characterization of conditioning tolerance boundaries in frozen generative models remains an open research problem.

Ethical Statement

Human annotation and fair compensation. All evaluators involved in the blind listening test were either formally employed researchers or graduate students supported by institutional scholarships. The blind listening test work was compensated in accordance with local labor regulations and institutional guidelines, consistent with ACL requirements regarding fair treatment and remuneration of human participants. No personally identifiable data were collected from evaluators, and the task involved only subjective quality ratings of AI-synthesized speech; our institution classifies this as minimal-risk research exempt from formal ethics board review.

Data privacy and consent. All training data are synthesized from EmoVoice-DB (Yang et al., 2025b), a publicly available emotional speech corpus released for research purposes. EmoVoice-DB is itself entirely AI-generated and contains no human recordings. The synthetic emotion transition data are generated entirely from this public dataset via our multi-pass flow blending pipeline; no pri-

vate, user-uploaded, or personally identifiable data are used at any stage. The released data and model checkpoints do not contain any personally identifiable information or sensitive user data.

Licensing and responsible use. The complete code, including both the synthetic data pipeline and model training pipelines, will be released under an open-source license upon acceptance, restricted to non-commercial academic research. We acknowledge that controllable emotional speech synthesis carries potential risks of misuse, including generating deceptive or manipulative audio content. We emphasize that EmoTra-TTS is intended as a research contribution to advance affectively dynamic speech synthesis, and we encourage responsible use with appropriate human oversight in any downstream application.

Usage of AI assistants. AI language models were used in two capacities: (1) generating test set sentences with emotion transition annotations (Appendix B.5), and (2) language polishing during paper writing. All experimental design, model development, analysis, and scientific conclusions were made by the authors.

References

- Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.
- Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.
- Rainer Banse and Klaus R Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.
- Haozhe Chen, Run Chen, and Julia Hirschberg. 2024. EmoKnob: Enhance voice cloning with fine-grained emotion control. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8170–8180.
- Huakang Chen, Jingbin Hu, Liumeng Xue, Qirui Zhan, Wenhao Li, Guobin Ma, Hanke Xie, Dake Guo, Linhan Ma, Yuepeng Jiang, and 1 others. 2026. Mint-bench: A comprehensive multilingual benchmark for

- instruction-following text-to-speech. *arXiv preprint arXiv:2604.17958*.
- Jie Chen, Jinhao Jiang, Yingqian Min, Zican Dong, Shijie Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2025a. Sticker-TTS: Learn to utilize historical experience with a sticker-driven test-time scaling framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12328–12338.
- Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2021. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations (ICLR)*.
- Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025b. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Sang-Hoon Lee, and Seong-Whan Lee. 2024. EmoSphere-TTS: Emotional Style and Intensity Modeling via Spherical Emotion Vector for Controllable Emotional Text-to-Speech. In *Interspeech 2024*, pages 1810–1814.
- Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909.
- Jiayan Cui, Zhihan Yang, Naihan Li, Jiankun Tian, Xingyu Ma, Yi Zhang, Guangyu Chen, Runxuan Yang, Yuqing Cheng, Yizhi Zhou, and 1 others. 2025a. Glm-tts technical report. *arXiv preprint arXiv:2512.14291*.
- Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and Irwin King. 2025b. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 13943–13970.
- Richard J Davidson. 1998. Affective style and affective disorders: Perspectives from affective neuroscience. *Cognition & emotion*, 12(3):307–330.
- Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.
- Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.
- Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. 2007. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057.
- Jonathan Ho and Tim Salimans. 2021. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.
- Hangrui Hu, Xinfu Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, and 1 others. 2026. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*.
- Li Hu. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8153–8163.
- Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu. 2025. Instructttsval: Benchmarking complex natural-language instruction following in text-to-speech systems. *arXiv preprint arXiv:2506.16381*.
- Shehzeen Samarah Hussain, Paarth Neekhara, Xuesong Yang, Edresson Casanova, Subhankar Ghosh, Roy Fejgin, Mikyas T. Desta, Rafael Valle, and Jason Li. 2025. Koel-TTS: Enhancing LLM based speech generation with preference alignment and classifier free guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21219–21234.
- Chanwoo Hwang, Sunpill Kim, Yong Kiam Tan, Tianchi Liu, Seunghun Paik, Dongsoo Kim, Mondal Soumik, Khin Mi Mi Aung, and Jae Hong Seo. 2026. Scores know bobs voice: Speaker impersonation attack. *arXiv preprint arXiv:2603.02781*.
- Sho Inoue, Kun Zhou, Shuai Wang, and Haizhou Li. 2025. Hierarchical control of emotion rendering in speech synthesis. *IEEE Transactions on Affective Computing*, 16(4):3316–3328.
- Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. Natural-speech 3: zero-shot speech synthesis with factorized codec and diffusion models. In *Forty-first International Conference on Machine Learning (ICML)*.
- Minki Kang, Dongchan Min, and Sung Ju Hwang. 2023. Grad-stylespeech: Any-speaker adaptive text-to-speech synthesis with diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

- Zhifeng Kong, Wei Ping, Jiayi Huang, Kexin Zhao, and Bryan Catanzaro. 2021. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*.
- Peter Kuppens and Philippe Verduyn. 2017. Emotion dynamics. *Current opinion in psychology*, 17:22–26.
- Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie. 2022. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:853–864.
- Yichong Leng, Zhifang Guo, Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, Lei He, Xiangyang Li, sheng zhao, Tao Qin, and Jiang Bian. 2024. PromptTTS 2: Describing and generating voices with text prompt. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Haitao Li, Chunxiang Jin, Chenglin Li, Wenhao Guan, Zhengxing Huang, and Xie Chen. 2026. Restyle-tts: Relative and continuous style control for zero-shot speech synthesis. *arXiv preprint arXiv:2601.03632*.
- Haoxun Li, Yu Liu, Yuqing Sun, Hanlei Shi, Leyuan Qu, and Taihao Li. 2025. Emorl-tts: Reinforcement learning for fine-grained emotion control in llm-based tts. *arXiv preprint arXiv:2510.05758*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 19730–19742.
- Yingting Li, Rishabh Bhardwaj, Ambuj Mehrish, Bo Cheng, and Soujanya Poria. 2024. HYPERTTS: Parameter efficient adaptation in text to speech using hypernetworks. In *LREC-COLING*, pages 8516–8527.
- Qifan Liang, Yuansen Liu, Ruixin Wei, Nan Lu, Junchuan Zhao, and Ye Wang. 2026. Segment-aware conditioning for training-free intra-utterance emotion and duration control in text-to-speech. *arXiv preprint arXiv:2601.03170*.
- Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. 2024. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*.
- Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *ICLR*.
- Tianchi Liu, Ivan Kukanov, Zihan Pan, Qiongqiong Wang, Hardik B. Sailor, and Kong Aik Lee. 2024. Towards quantifying and reducing language mismatch effects in cross-lingual speech anti-spoofing. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 1185–1192.
- Tianchi Liu, Ruijie Tao, Qiongqiong Wang, Yidi Jiang, Hardik B. Sailor, Ke Zhang, Jingru Lin, and Haizhou Li. 2025a. Interpolating speaker identities in embedding space for data expansion. In *Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 589–594.
- Tianchi Liu, Duc-Tuan Truong, Rohan Kumar Das, Kong Aik Lee, and Haizhou Li. 2025b. Nes2net: A lightweight nested architecture for foundation model driven speech anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 20:12005–12018.
- Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, pages 4114–4124.
- Ziyang Ma, Wen Wu, Zhisheng Zheng, Yiwei Guo, Qian Chen, Shiliang Zhang, and Xie Chen. 2024a. Leveraging speech ptm, text llm, and emotional tts for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11146–11150.
- Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024b. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760.
- Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alex Smola. 2025. EmergentTTS-eval: Evaluating TTS models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. In *The Conference on Neural Information Processing Systems*.
- Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current psychology*, 14(4):261–292.
- Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2024. Matcha-tts: A fast tts architecture with conditional flow matching. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11341–11345.
- Gaspard Michel, Elena V. Epure, and Christophe Cerisara. 2026. Computational narrative understanding for expressive text-to-speech. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 6194–6215.
- MOSS-TTS Team. 2026. MOSS-TTS Family: An open-source speech and sound generation model family.
- Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of AAAI*.

- Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*, pages 28492–28518.
- James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.
- Klaus R Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256.
- RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A. Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *ICML*, pages 4693–4702.
- Zeyang Song, Tianchi Liu, Tianrui Wang, Chenglin Xu, Yiwen Guo, and Haizhou Li. 2026. Diagnose, then refine: A closed-loop TTS system with AudioLLM-guided correction. *arXiv preprint*.
- Ruijie Tao, Zhan Shi, Yidi Jiang, Tianchi Liu, and Haizhou Li. 2025. Voice conversion augmentation for speaker recognition on defective datasets. In *Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 2529–2534.
- Debra Trampe, Jordi Quoidbach, and Maxime Taquet. 2015. Emotions in everyday life. *PloS one*, 10(12):e0145450.
- Philippe Verduyn, Ellen Delvaux, Hermina Van Coillie, Francis Tuerlinckx, and Iven Van Mechelen. 2009. Predicting the duration of emotional experience: two experience sampling studies. *Emotion*, 9(1):83.
- Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2023. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759.
- Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.
- Junyu Wang, Siyuan Zhang, Peiyuan Jiang, Jian Zong, Jingyu Zhang, Tianrui Wang, Yuqin Lin, Zhenghui Chen, Shuqing Xie, Ziyang Ma, and 1 others. 2026a. Emos: A theory-grounded framework for evaluating and aligning emotional intelligence in spoken language models. *arXiv preprint arXiv:2608.09189*.
- Siyi Wang, Shihong Tan, Siyi Liu, Hong Jia, Gongping Huang, James Bailey, and Ting Dang. 2026b. Cocoemo: Composable and controllable human-like emotional tts via activation steering. *arXiv preprint arXiv:2602.03420*.
- Tianrui Wang, Ziyang Ma, Yizhou Peng, Haoyu Wang, Zhikang Niu, Zikang Huang, Yihao Wu, Yi-Wen Chao, Yu Jiang, Yuheng Lu, Guanrou Yang, Xuanchen Li, Hexin Liu, Chunyu Qiang, Cheng Gong, Yifan Yang, Tianchi Liu, Junyu Wang, Nana Hou, and 10 others. 2026c. Evaluating the expressive appropriateness of speech in rich contexts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9088–9106.
- Tianrui Wang, Haoyu Wang, Meng Ge, Cheng Gong, Chunyu Qiang, Ziyang Ma, Zikang Huang, Guanrou Yang, Xiaobao Wang, EngSiong Chng, Xie Chen, Longbiao Wang, and Jianwu Dang. 2025a. Word-level emotional expression control in zero-shot text-to-speech synthesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, and 1 others. 2025b. Sparktts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.
- Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xuexiao Zhang, Shunsi Zhang, and Zhizheng Wu. 2025c. MaskGCT: Zero-shot text-to-speech with masked generative codec transformer. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 5180–5189.
- Haibin Wu, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Daniel Tompkins, Chung-Hsien Tsai, Canrun Li, Zhen Xiao, Sheng Zhao, Jinyu Li, and Naoyuki Kanda. 2024. Laugh now cry later: Controlling time-varying emotional states of flow-matching-based zero-shot text-to-speech. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 690–697.
- Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. 2025. Towards controllable speech synthesis in the era of large language models: A systematic survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 764–791.
- Dong Yang, Yiyi Cai, Yuki Saito, Lixu Wang, and Hiroshi Saruwatari. 2025a. Shallow flow matching

for coarse-to-fine text-to-speech synthesis. *arXiv preprint arXiv:2505.12226*.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, Fan Yu, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2025b. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10748—10757.

Guanrou Yang, Fan Yu, Ziyang Ma, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2025c. Enhancing low-resource asr through versatile tts: Bridging the data gap. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Shih yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning (ICML)*.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, and 1 others. 2025. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847.

Jinming Zhao, Gholamreza Haffari, and Ehsan Shareghi. 2023. Generating synthetic speech from SpokenVocab for speech translation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1975–1981.

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

Kun Zhou, You Zhang, Shengkui Zhao, Hao Wang, Zexu Pan, Dianwen Ng, Chong Zhang, Chongjia Ni, Yukun Ma, Trung Hieu Nguyen, and 1 others. 2024. Emotional dimension control in language model-based text-to-speech: Spanning a broad spectrum of human emotions. *arXiv preprint arXiv:2409.16681*.

Li Zhou, Hao Jiang, Junjie Li, Tianrui Wang, and Haizhou Li. 2026. Emoshift: Lightweight activation steering for enhanced emotion-aware speech synthesis. *arXiv preprint arXiv:2601.22873*.

Siya Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. In-dextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

Continuous Emotion Transition — Speech Quality Evaluation

Evaluation Instructions

You will listen to speech samples generated by multiple TTS systems that attempt to produce continuous emotion transitions within a single utterance. For each sample, rate each system across 3 evaluation dimensions.

Dimension 1: MOS — Naturalness & Audio Quality

Rate how natural and human-like the speech sounds overall.

Score	Level	Description
5	Excellent	Completely natural, human-like, no artifacts.
4	Good	Natural, very minor artifacts.
3	Fair	Understandable, somewhat robotic or buzzy.
2	Poor	Unnatural, distinct metallic or robotic sounds.
1	Bad	Completely unnatural, unintelligible.
N	N/A	Audio cannot be played or is missing.

Dimension 2: Emotion Accuracy

Does the speech convey the correct emotions (start emotion → end emotion)?

Score	Level	Description
5	Perfect	Both start and end emotions are clearly and correctly expressed.
4	Good	Emotions are mostly correct, minor deviations.
3	Partial	One emotion is correct, the other is weak or wrong.
2	Weak	Emotions are barely perceptible or mostly wrong.
1	Wrong	Emotions are completely wrong or absent.
N	N/A	Audio cannot be played or is missing.

Dimension 3: Emotion Transition Naturalness

How naturally does the emotion change from start to end within the utterance?

Score	Level	Description
5	Seamless	Transition is smooth and natural, like a real human's emotion shift.
4	Good	Transition is noticeable and mostly smooth, slightly abrupt.
3	Moderate	Transition exists but feels mechanical or forced.
2	Weak	Barely any transition, mostly stays in one emotion.
1	None	No transition at all, or abrupt unnatural jump.
N	N/A	Audio cannot be played or is missing.

Tips

- System order is **randomized** for each sample to avoid bias.
- Sample order is **randomized** for each session.
- You can **end early** at any time and download results for completed samples.

Evaluator Name (required):

Start Evaluation

Figure 4: Rating criteria for ‘Naturalness & Audio Quality’ (MOS-Qua), ‘Emotion Accuracy’ (MOS-Emo), and ‘Emotion Transition Naturalness’ (MOS-Tra).

A Training Details

Synthetic data generation: Cross-fade steepness $\kappa = 12$; segment ratios $r_{\text{init}}/r_{\text{tra}}/r_{\text{res}} = 60\%/15\%/25\%$; crossfade duration $\min(0.4s, 0.1 \times T_{\text{dur}})$. ASR quality filter: Whisper large-v3 with CER threshold $\tau_{\text{CER}} = 0.10$.

Stage 1 training (LLM): Adam optimizer, learning rate $= 1.0 \times 10^{-5}$, gradient clipping at 5.0. $N_{\text{tra}} = 3$ intermediate VAD tokens ($N_{\text{tra}} + 2 = 5$ total), $d_h = 256$, $d_{\text{LLM}} = 896$, $d_{\text{SER}} = 1024$, $\lambda = 0.1$.

Stage 2 training (Flow): All decoder parameters frozen. Adam optimizer, gradient accumulation factor of 2. Segment ratios: $r_{\text{init}} = 0.4$, $r_{\text{tra}} = 0.3$, $r_{\text{res}} = 0.3$. Fixed emotion scale $\epsilon = 0.07$, yielding effective norm $\approx \epsilon \cdot \sqrt{d_{\text{mel}}} = 0.07 \times \sqrt{80} \approx 0.626$, matched to the observed speaker embedding norm ($\|e\|_2 \approx 0.622$). ψ_{down} :

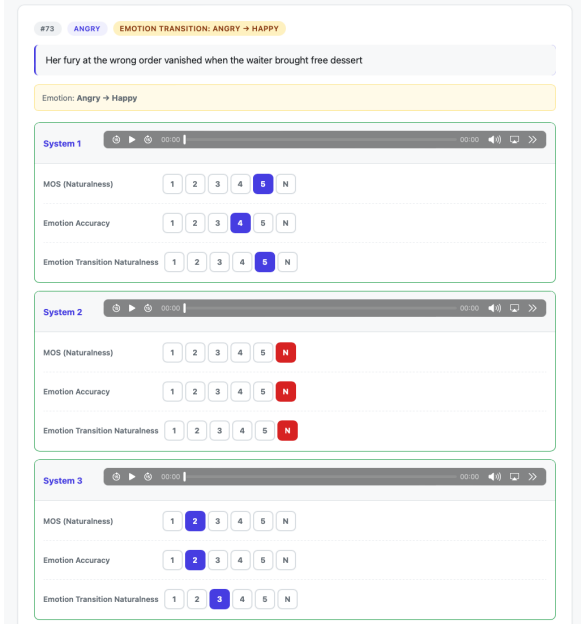


Figure 5: The blind test evaluation interface. Evaluators listen to each sample and rate ‘MOS Naturalness’ (MOS-Qua), ‘Emotion Accuracy’ (MOS-Emo), and ‘Emotion Transition Naturalness’ (MOS-Tra) on 1–5 Likert scales. System identities are hidden, and the system order is randomly shuffled for each sample.

1024 $\rightarrow$ 256 $\rightarrow$ ReLU $\rightarrow$ 80 (last layer zero-initialized). LayerNorm: $d_{\text{mel}} = 80$.

Evaluation cost. The two commercial systems used for the paired comparison in §5.6 incurred a total cost of USD 11 (ElevenLabs subscription: USD 6; GPT-4o mini TTS API: USD 5).

B Evaluation Protocol

We conduct blind listening evaluations using a professional paid blind-test team with established quality control procedures. In total, 21 raters participated across all evaluations, including researchers and non-technical listeners.

B.1 MOS-based Evaluation Protocol

For the MOS-based evaluation, 10 evaluators (3 researchers and 7 non-technical listeners) participated to capture both expert and general-listener perspectives. Each system contains 200 utterances, and each evaluator rated about 25 randomly sampled utterances per system, resulting in approximately 250 ratings per system and about 25% repeated coverage.

These test sample texts and target emotions are generated by ChatGPT-5.2⁵ using a struc-

⁵<https://help.openai.com/en/articles/6825453-chatgpt-release-notes>

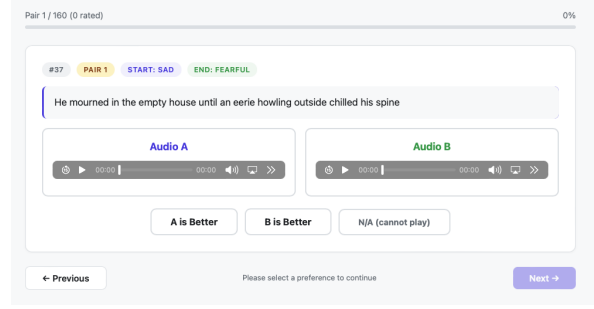


Figure 6: Screenshot of the paired preference evaluation interface. System identities are hidden, and the system order is randomly shuffled for each sample.

tured prompt (Appendix B.5) that enforces event-driven narratives, balanced emotion-pair coverage, and VAD annotations within tight per-category ranges. The resulting test set is independent of the EmoVoice-DB (Yang et al., 2025b) training set.

Notably, we prioritize the **quality** of the blind test over its **quantity**. With a carefully selected panel, we emphasize careful, attentive ratings rather than maximizing raw throughput, while still achieving a high overall completion rate of ~94%. All samples are randomly shuffled and system identities are hidden from evaluators.

Three metrics are collected on a 1–5 Likert scale:

- **MOS-Qua:** speech naturalness and quality, without considering emotional correctness.
- **MOS-Emo:** adequacy of emotion rendering relative to the target specification, evaluating whether both the initial and resultant emotions are correctly expressed.
- **MOS-Tra:** smoothness and naturalness of intra-utterance emotion transition, assessing whether the shift between emotions sounds gradual and natural rather than abrupt or discontinuous.

The detailed rating criteria for each metric are shown in Figure 4, and an example of the blind evaluation interface is provided in Figure 5.

B.2 Paired Preference Evaluation Protocol

For the paired preference evaluation in §5.5, we retain 4 raters from the original pool in Appendix B.1 and include 7 additional raters, resulting in 11 raters for this study. For each system pair, 40 utterances were randomly sampled from the 200 utterances mentioned above, forming four paired comparison groups (160 pairs in total). Each rater

Evaluation Instructions

You will listen to pairs of speech samples and decide which one is **better overall**.

Task Description

For each pair, you will see:

- **Start Emotion and End Emotion:** the intended emotion transition of the speech.
- **Text:** the content that the TTS system is reading.
- **Audio A and Audio B:** two speech samples from different systems (order is randomized).

Evaluation Criteria

Please judge which sample is **better overall**, considering the following aspects (but not limited to):

- **Audio Quality:** Is the speech clear and natural?
- **Start & End Emotion Accuracy:** Does the speech correctly express the intended start and end emotions?
- **Emotion Transition Naturalness:** Is the transition between the two emotions smooth and convincing?
- **Text Correctness:** Does the speech correctly read the given text without errors or omissions?

Choices

- **A is Better:** If Audio A is better overall.
- **B is Better:** If Audio B is better overall.
- **N/A:** If you cannot play the audio (technical issue). This will not count in statistics.

Tips

- The left/right position of systems is **randomized** for each pair to avoid bias.
- Pairs are presented in **random order**.
- You can go back to revise previous choices.
- Progress is **auto-saved** to your browser.
- You can **stop at any time** and download the JSON with your current results using "Save & Download JSON".
- Use **Save & Exit** to pause and resume later.

Evaluator Name (required):

Start Evaluation

Figure 7: Rating criteria for the paired evaluation.

completed about 85 comparisons on average, yielding about 6-fold repeated coverage per pair.

The detailed rating criteria for each metric are shown in Figure 7, and an example of the blind evaluation interface is provided in Figure 6.

B.3 Comparison with Commercial Systems

sage

[sad] The weather was so bad that day, and I was sadly walking home [angry] when I suddenly saw something that made me extremely angry.

+ Add speaker

Settings History

Try Flows
Node-based canvas for creating image, video, speech, music, all in one place.

Voice
sage

Model
Best practices
v3 Eleven v3

Stability
Creative Robust

Language Override
[Toggle]

Output Format
MP3 44.1 kHz (128kbps)

Figure 8: ElevenLabs v3 inference interface used in our comparison, showing the manually inserted discrete audio tags (e.g., [sad], [angry]) and the “Stability” control set to “Creative”.

This appendix details the two commercial systems compared in §5.6 and the evaluation protocol. **GPT-4o mini TTS** (OpenAI’s GPT-4o audio family) is controlled via a natural-language instruction

Evaluation Instructions

You will listen to pairs of speech samples. For each pair you answer two questions.

What you see

- **Start Emotion and End Emotion:** the intended emotion transition of the speech.
- **Text:** the content that the TTS system is reading.
- **Audio A and Audio B:** two speech samples from different systems (left/right order is randomized).

The two questions

- **1. Emotion & Transition:** which sample better expresses the intended **start and end emotions** and has a more **natural, smooth transition** between them?
- **2. Overall Preference:** all things considered (audio quality, naturalness, emotion, correct reading of the text), which sample do you **prefer overall**?

Choices

- **A is Better / B is Better:** pick one for each question.
- **N/A (cannot play):** use only if the audio cannot be played (technical issue). It marks the whole pair as N/A and it will not count in statistics.

Figure 9: Rating criteria for the comparison with commercial systems.

Auto-saved 02:19:28 **Save & Download JSON** **Save & Exit**

Pair 1 / 80 (0 done) 0%

#42 PAIR 1

She gasped at the broken vase and turned angry seeing who had caused it

Audio A **Audio B**

1. Emotion & Transition (whether the start/end emotions are accurately conveyed, and whether the emotional transition is natural and smooth)
 START: SURPRISED END: ANGRY
 A is Better B is Better

2. Overall Preference (all things considered, which one do you think is better)
 A is Better B is Better
 N/A (cannot play)

← Previous Please answer both questions to continue Next →

Figure 10: Screenshot of the paired preference evaluation interface. System identities are hidden, and the system order is randomly shuffled for each sample.

prompt. **ElevenLabs v3** does not support natural-language prompts for emotion; instead, the user inserts discrete audio tags (e.g., [happy], [sad]) at chosen positions^{6 7}, which requires an additional human or LLM step to place the tags and specifies discrete switches rather than continuous transitions. To present v3 in its best light, we placed the tags manually and set the “Stability” control to “Creative” (Figure 8), the most expressive setting per the official guidance.⁸

We ran a paired evaluation on 40 sentences per system (80 pairs in total) with 12 raters (4 newly recruited). Each rater evaluated about 53 pairs, and

⁶Prompting guide: <https://elevenlabs.io/docs/overview/capabilities/text-to-speech/best-practices#prompting-eleven-v3-alpha>

⁷(updated Jul 5, 2026): <https://elevenlabs.io/blog/eleven-v3-situational-awareness>

⁸<https://elevenlabs.io/app/speech-synthesis/text-to-speech>

each pair was rated about 8 times. To avoid conflating synthesis quality with transition quality, we collected dimension-specific preferences: (i) emotion accuracy and transition smoothness, and (ii) overall preference as shown in Figure 9 and Figure 10.

B.4 Inference Efficiency Evaluation Setup

This appendix summarizes the test set used for the inference efficiency evaluation. The evaluation set contains 100 sentences, with 5 additional warm-up sentences excluded from the statistics. The average sentence length is 13.9 words, with a standard deviation of 3.4 words. Sentence lengths range from 9 to 21 words, with a median of 13 words. In total, the test set contains 1,391 words.

B.5 Test Set Generation Prompt

The following prompt was used with ChatGPT-5.2 to generate 200 test sentences with emotion transition annotations.

Prompt for Test Set Generation

Task: Generate English sentences that describe a transition from one emotional state to another within a short narrative.

Each sample must include: the sentence, the starting emotion, the ending emotion, and the VAD (Valence, Arousal, Dominance) values for both emotions.

EMOTION CATEGORIES

angry, happy, sad, surprised, fearful, neutral

Transitions may occur between any pair of different emotions (30 possible pairs).

SENTENCE REQUIREMENTS

Each sentence should describe a scenario or event that naturally causes an emotional shift. The narrative must clearly reflect the transition so that a speaker reading it aloud could express the emotional shift through prosody and tone.

AVOID sentences that just name emotions. Instead, describe an event or realization that triggers the shift.

BAD: "I was sad then I became happy"

GOOD: "The long empty street made him feel alone until distant music drifted through the night"

Sentence length: 12–30 words. Natural pauses allowed (commas, conjunctions). May include connectors such as: but, until, then, suddenly, moments later, after that, and then, when.

Sentences should include diverse narrative contexts such as: daily life, workplace, social interaction, unexpected discoveries, environmental changes, memories, misunderstandings, surprising events, physical sensations, weather, animals, travel, etc.

Avoid repeating similar structures, scenarios, or phrases.

VAD ANNOTATION

Each emotion must include VAD values: (Valence, Arousal, Dominance). All values should be formatted to two decimal places.

Target VAD Ranges:

angry: V (0.18–0.26) A (0.82–0.90) D (0.78–0.89)

happy: V (0.80–0.92) A (0.76–0.88) D (0.70–0.88)

sad: V (0.18–0.26) A (0.18–0.28) D (0.18–0.26)

surprised: V (0.60–0.80) A (0.80–0.89) D (0.50–0.68)

fearful: V (0.18–0.26) A (0.78–0.88) D (0.18–0.28)

Important: Values should reflect emotionally saturated states for non-neutral categories. Vary the values within the range; do not always use the same numbers.

The VAD distance between starting and ending emotion must be clearly large enough to be perceptible. Avoid transitions where both emotions have very similar VAD values.

DISTRIBUTION CONSTRAINT

Cover at least 25 unique emotion pairs out of 30. No single pair more than 15 times in 200 samples. Roughly balanced distribution.

FEW-SHOT EXAMPLES

The loud crack startled him and then a creeping dread filled his chest | surprised | fearful | (0.68,0.86,0.60) | (0.20,0.84,0.22)

The cheerful laughter at the party was cut short when she saw the broken vase on the floor | happy | angry | (0.88,0.82,0.78) | (0.20,0.88,0.85)

A sudden gust of wind knocked the tent down and he scrambled in panic, but then burst into laughter | fearful | happy | (0.22,0.82,0.20) | (0.85,0.80,0.75)

She cursed the delayed train and then sat down feeling deeply miserable | angry | sad | (0.24,0.87,0.83) | (0.19,0.23,0.22)

OUTPUT FORMAT

Each line: Sentence | emotion1 | emotion2 | (V,A,D) | (V,A,D)

Use | as separator. No numbering, no headers, no blank lines.

TARGET: Generate 200 samples.

C Emotion-Specific VAD Ranges

Table 6: Emotion-specific VAD acceptance ranges for source corpus filtering. All values are on a $[0, 1]$ scale. Ranges are intentionally broad to accommodate natural within-category variability.

Emotion	Valence	Arousal	Dominance
Angry	[0.00, 0.35]	[0.75, 1.00]	[0.70, 1.00]
Happy	[0.65, 1.00]	[0.50, 1.00]	[0.55, 1.00]
Sad	[0.00, 0.35]	[0.00, 0.50]	[0.00, 0.50]
Surprised	[0.35, 0.80]	[0.60, 1.00]	[0.35, 0.75]
Fearful	[0.00, 0.45]	[0.40, 0.90]	[0.10, 0.55]
Disgusted	[0.00, 0.35]	[0.30, 0.70]	[0.30, 0.65]
Neutral	[0.40, 0.60]	[0.40, 0.60]	[0.40, 0.60]

Table 6 lists the VAD acceptance ranges used for cross-validated filtering of the source corpus (§4). These ranges are **intentionally set broader** than prototypical VAD centroids (Russell, 1980; Mehra-bian, 1996) for two reasons: (1) natural emotional speech exhibits substantial within-category variability (Scherer, 2003; Banse and Scherer, 1996), and (2) the VAD predictor (Wagner et al., 2023) has estimation uncertainty, so tight boundaries would conflate predictor error with genuine mismatch. The filtering retains samples only if their categorical label *and* predicted VAD are mutually consistent, removing mislabeled or acoustically ambiguous utterances.

D Objective Smoothness Evaluation

To complement the perceptual evaluation with an objective measure, we introduce the *Prosodic Jerk Ratio* (JR-F0), a signal-level metric computed directly from f_0 via Praat. Given a voiced f_0

Table 7: Prosodic Jerk Ratio for the proposed EmoTra-TTS and baselines in Categories (d) and (e) of Table 2. All differences between EmoTra-TTS and the baselines are statistically significant ($p < 0.001$).

System	JR-F0 ↓
EmoTra-TTS (Ours)	0.025 ± 0.009
Synthetic data (Ours)	0.030 ± 0.014
Qwen3-TTS (Hu et al., 2026)	0.036 ± 0.016
WeSCon (Wang et al., 2025a)	0.038 ± 0.017
MOSS-TTS (MOSS-TTS Team, 2026)	0.038 ± 0.017
CosyVoice2 (Du et al., 2024)	0.039 ± 0.018

contour, we measure the second-order difference $\Delta^2 f_0[n] = f_0[n+1] - 2f_0[n] + f_0[n-1]$ and define

$$\text{JR-F0} = \frac{\#\{n : |\Delta^2 f_0[n]| > \tau\}}{\#\{n : f_0[n] \text{ voiced}\}}, \quad (15)$$

i.e. the fraction of voiced frames with abrupt pitch discontinuities, where τ is an adaptive per-utterance threshold. Lower JR-F0 indicates smoother prosodic evolution.

As shown in Table 7, EmoTra-TTS attains the lowest JR-F0, corroborating the perceptual transition-naturalness results with a fully signal-level measure. Notably, EmoTra-TTS is even smoother than its own synthetic training data (0.025 vs. 0.030): the synthetic corpus retains residual discontinuities from four-segment blending (§3.2.2), whereas frame-level interpolated VAD conditioning yields globally continuous prosody.

E Detailed Failure Analysis

This appendix documents the iterative development across all seven injection architectures (Table 8), each failure revealing a distinct phenomenon in auxiliary conditioning of frozen generative models.

E.1 Decoder Drift (Finetune)

A linear projection ($896 \rightarrow 80$; $\sim 72\text{K}$ parameters) injects emotion through the conditioning channel $\mathbf{c}$ with the decoder *jointly trained*. Strong emotion is achieved, but the decoder’s $\sim 639\text{M}$ parameters absorb the adaptation by shifting pretrained weights, degrading content fidelity. This establishes the *decoder freezing* principle for all subsequent variants.

E.2 Emotion Capacity Limit (Linear)

A single linear layer ($1024 \rightarrow 80$; 82K parameters) with frozen decoder and additive injection preserves content (MOS-Qua 3.60) but produces only moderate emotion (MOS-Emo 3.39, MOS-Tra 3.33). The linear mapping lacks capacity for the nonlinear $d_{\text{SER-to-}d_{\text{mel}}}$ transformation (Alain

and Bengio, 2016), establishing a *lower bound* on model complexity for expressive conditioning.

E.3 Content Collapse (MLP+LoRA)

A two-layer MLP ($1024 \rightarrow 256 \rightarrow 80$; $\sim 280\text{K}$) with LoRA (Hu et al., 2022) rank-8 adapters on all 336 decoder layers ($\sim 2.9\text{M}$ total) yields emotion but *zero* intelligible content (MOS-Qua 1.03, WER 107.15%, SIM 0.627). LoRA’s global perturbation shifts the flow decoder’s content–style equilibrium, catastrophically amplifying style at the expense of content, paralleling findings that adapter placement critically affects multi-conditional generation (Hu, 2024; Lin et al., 2024). This establishes that decoder weights should not be modified for single-axis adaptation, even with parameter-efficient methods (Hu et al., 2022).

E.4 Magnitude Mismatch (MLP-Add)

The same MLP without LoRA ($\sim 280\text{K}$ parameters, decoder frozen) achieves strong emotion but content failure (MOS-Qua 2.10, WER 20.09%) with a diagnostic bimodal pattern: emotion is reflected in both successful and garbled samples, indicating a distributional out-of-range problem. Norm measurements reveal the cause: $\|e_{\text{emo}}(t)\|_2 \approx 1.248 \approx 2 \times \|e\|_2$, pushing the combined embedding $3\times$ beyond the decoder’s pretrained operating range.

We isolate magnitude as the causal factor via inference-time manual scaling λ on frozen MLP weights. Content failure is monotonically related to the emotion-to-speaker norm ratio: at $\lambda = 0.5$ (ratio ≈ 1.0), content is fully preserved; at $\lambda = 1.0$ (ratio ≈ 2.0), only a small portion of samples remain intelligible. Emotion direction is correct even in garbled samples, confirming that the problem is exclusively one of magnitude. This directly validates our Dir–Mag design and generalizes: for any additive conditioning of frozen models, the conditioning-to-existing-signal norm ratio must stay within the pretrained operating range.

E.5 MLP Compensation (FiLM)

FiLM (Perez et al., 2018) modulation ($\hat{e}(t) = \gamma(t) \odot e + \beta(t)$; $\sim 560\text{K}$ parameters, identity-initialized (Zhang et al., 2023), decoder frozen) is designed to provide learnable per-dimension scale and shift. However, FiLM fails even more severely than MLP-Add (MOS-Qua 1.96, WER 53.35%, SIM 0.400): the MLP compensates by inflating output magnitude $3\text{--}5\times$ during joint optimization.

Table 8: Full injection architecture comparison across all seven variants. SIM: speaker similarity. WER: word error rate (%). We decompose MOS into MOS-Qua, MOS-Emo, and MOS-Tra for a more fine-grained and independent evaluation of speech naturalness, emotion rendering adequacy, and intra-utterance emotion transition smoothness, respectively. [†]Finetune is excluded from blind test evaluation due to severely degraded content (decoder drift renders most outputs unintelligible), precluding meaningful perceptual assessment.

Variant	Injection	Δ Params	Decoder	Quality Evaluation			Emotion Evaluation		Key finding
				MOS-Qua $\uparrow$	SIM $\uparrow$	WER $\downarrow$	MOS-Emo $\uparrow$	MOS-Tra $\uparrow$	
Finetune [†]	Linear	72K	Trainable	—	—	—	—	—	Decoder drift
Linear	Linear, additive	82K	Frozen	3.60 $\pm$ 0.12	0.707	1.65	3.39 $\pm$ 0.15	3.33 $\pm$ 0.16	Emotion Capacity limit
MLP+LoRA	MLP+LoRA	2.9M	LoRA	1.03 $\pm$ 0.05	0.627	107.15	1.69 $\pm$ 0.13	1.42 $\pm$ 0.11	Content collapse
MLP-Add	MLP, additive	280K	Frozen	2.10 $\pm$ 0.17	0.595	20.09	2.49 $\pm$ 0.17	2.17 $\pm$ 0.17	Magnitude mismatch
FiLM	FiLM modulation	560K	Frozen	1.96 $\pm$ 0.16	0.400	53.35	2.44 $\pm$ 0.17	1.82 $\pm$ 0.16	MLP compensation
Norm Clamp	Norm Clamp	280K	Frozen	3.29 $\pm$ 0.15	0.692	3.89	3.30 $\pm$ 0.15	3.02 $\pm$ 0.16	Gradient suppression
Dir-Mag (ours)	LN+fixed ϵ	280K	Frozen	3.54 $\pm$ 0.13	0.701	1.77	3.52 $\pm$ 0.15	3.63 $\pm$ 0.17	Dir-mag decoupled

This demonstrates that *soft constraints on entangled factors are insufficient* (Locatello et al., 2019). The flow loss, averaging over all frames, provides no gradient penalizing magnitude growth, and the MLP and FiLM parameters co-adapt to circumvent the constraint.

E.6 Gradient Suppression (Norm Clamp)

Hard norm clamping ($\tau = 0.70$, derived from the manual scaling experiment; $\sim 280K$ parameters, decoder frozen) substantially improves content over MLP-Add (MOS-Qua 3.29, WER 3.89%), confirming magnitude control as the key factor. However, emotion is only moderate (MOS-Emo 3.30) due to *gradient suppression*: beyond τ , clamping zeros the radial gradient component, and the discontinuous transition between clamped and unclamped regimes causes the MLP to oscillate near the boundary rather than converging smoothly. This demonstrates that even a *correct* threshold is insufficient without a *smooth* normalization mechanism.

E.7 Success: Direction-Magnitude Decoupled

The same MLP followed by LayerNorm($d_{\text{mel}} = 80$) and fixed $\epsilon = 0.07$ ($\sim 280K + 160$ parameters, decoder frozen) achieves content fidelity with strong emotion (MOS-Qua 3.54, MOS-Emo 3.52, MOS-Tra 3.63). LayerNorm locks magnitude at $\sqrt{d_{\text{mel}}} \approx 8.94$; the effective norm $\epsilon \cdot \sqrt{d_{\text{mel}}} \approx 0.626$ matches the speaker norm ($\|e\|_2 \approx 0.622$). The MLP learns emotion *directions* freely but cannot inflate *magnitude*, confirming that *structural* constraints are necessary for safe auxiliary conditioning.

F Crossfade Curve Selection

This appendix supplements the mel-spectrogram-level crossfade in §3.2.2 (Fig. 1) by documenting the curve selection process. The crossfade curve

Table 9: Crossfade curves evaluated for mel-level blending. $\sigma_\kappa(t) = 1/(1 + e^{-\kappa(t-0.5)})$, normalized to $[0, 1]$; $s(t) = \frac{1}{2}(1 - \cos \pi t)$. Act. fr. indicates active frames.

Curve	$w_A(t)$	$w_B(t)$	Act. fr.	$w_A^2 + w_B^2 = 1?$
Linear	$1 - t$	t	200	$\times$
Equal-power cosine	$\cos(\frac{\pi}{2} s(t))$	$\sin(\frac{\pi}{2} s(t))$	~ 150	$\checkmark$
Sigmoid symmetric	$1 - \sigma_{12}(t)$	$\sigma_{12}(t)$	~ 34	$\times$
Sigmoid asymmetric	$\sqrt{1 - \sigma_{12}(t)}$	$\sigma_{12}(t)^2$	~ 34	$\times$
Ultra-steep sigmoid	$\sqrt{1 - \sigma_{100}(t)}$	$\sigma_{100}(t)^{1.5}$	~ 4	$\times$

is **not** a core contribution; we include this analysis for interested readers. The selection was based on informal listening by the authors (~ 20 samples per curve, not a formal blind test).

Within the transition region, blending follows $\text{mel}(t) = w_A(t) \text{mel}_A(t) + w_B(t) \text{mel}_B(t)$ on normalized $t \in [0, 1]$. Since the quality of synthetic training data directly depends on the blending curve, we evaluated five candidates (Table 9) to find the best trade-off between transition smoothness and emotion purity for data generation.

Linear blending made the splice point audible in a subset of samples. **Equal-power cosine** removed the splice artifact but noticeably degraded quality in the transition region. **Ultra-steep sigmoid** ($\kappa=100$) compressed blending to ~ 4 frames, yielding no artifacts but sounding perceptually identical to a hard switch and losing all sense of gradual transition. Both **sigmoid** variants ($\kappa=12$) concentrate blending in ~ 34 central frames while keeping edges binary and sound largely identical in practice. The asymmetric variant exhibited mild spectral artifacts in only a few samples; the symmetric version avoided these, so we adopted it for consistency.

We adopted the **sigmoid symmetric** with $\kappa=12$ (Eq. (4)). The key finding is that κ governs the smoothness–purity trade-off: too gradual (linear, cosine) sacrifices emotion contrast; too steep ($\kappa \geq 100$) collapses to a hard switch.