CONTROLANYTHING: UNIFIED WORD-LEVEL, INTENSITY-AWARE, AND COMPOSABLE CONTROL FOR EXPRESSIVE TEXT-TO-SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotional expression in human speech varies across words, exhibits varying intensity levels, and often blends multiple affective states. Reproducing this flexibility in text-to-speech (TTS) requires joint control over where emotions occur, how strongly they are expressed, and how they are combined, while applications such as voice assistants further demand fast generation for responsive interaction. However, existing methods typically address these control dimensions separately, and fine-grained word-level control has been explored primarily in autoregressive (AR) TTS, leaving unified control in efficient non-autoregressive (NAR) architectures largely underexplored. We present ControlAnything, a unified framework for word-level, intensity-aware, and composable emotional control that preserves NAR inference efficiency. First, Soft-boundary Anchor Intervention (SAI) maps target text spans to acoustic regions and progressively reinforces temporal anchoring as acoustic representations evolve during generation. Second, Emotion Concept Modulation (ECM) learns distinct emotion-specific transformations within a shared modulation mechanism, enabling fine-grained intensity adjustment and flexible multi-emotion composition. Together, SAI and ECM support localized intensity adjustment and emotion composition across different pretrained TTS models. Extensive experiments demonstrate state-of-the-art controllability in word-level localization, emotion-intensity adjustment, and multi-emotion composition, while preserving speech quality and generation efficiency comparable to the baselines. Audio demonstrations are available at `https://controlanything-review-demo.vercel.app/`.

1 INTRODUCTION

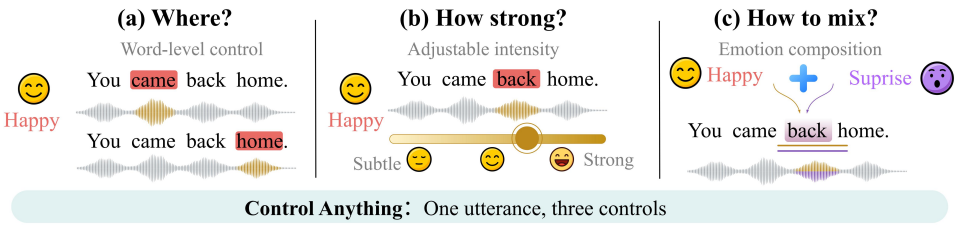


Figure 1: ControlAnything unifies (a) word-level localization, (b) adjustable intensity, and (c) multi-emotion composition for expressive TTS.

Text-to-speech (TTS) synthesizes natural and intelligible speech from text for applications including voice assistants, narration, and human–computer interaction (Xie et al., 2025a; Liang et al., 2026). Beyond intelligibility and speaker similarity, realistic synthesis requires emotional control across complementary dimensions. *Word-level control* determines where an emotion occurs within an utterance. *Intensity-aware control* regulates how strongly it is conveyed, capturing subtle variations beyond discrete emotion labels. *Composable control* allows multiple emotions to coexist, reflecting the complexity of natural expression. For example, "I'm glad you came, but I wish it weren't so late" may convey warmth in the opening and mild frustration in the closing, with affection softening

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.

the complaint. Joint control could enrich storytelling (Rong et al., 2026) and interaction (Triantafyllopoulos et al., 2023), while complementing localized speech editing (Liu et al., 2025).

Emotion-controllable TTS has advanced along these dimensions. Label-based approaches use predefined emotion categories (Xie et al., 2025a), while continuous affective representations enable finer style and intensity regulation (Cho et al., 2025). Prompt-based methods specify expression through natural-language descriptions or reference speech (Guan et al., 2024; Yang et al., 2025). Activation steering directly manipulates internal emotion representations for conversion, interpolation, and erasure (Xie et al., 2025b), as well as mixed-emotion synthesis (Wang et al., 2026). However, these methods typically address individual dimensions, leaving unified control underexplored. For example, recent compositional steering applies emotion vectors uniformly across tokens, leaving segment-level steering for future work (Wang et al., 2026). Thus, intensity modulation and emotion composition do not necessarily determine where emotions occur.

This gap is evident when fine-grained localization is combined with efficient generation. Existing efforts toward word- and span-level emotional control have focused on autoregressive (AR) TTS, where sequential token generation supports applying or switching emotional conditions. Comparable control remains less explored in non-autoregressive (NAR) TTS, despite its potential for fast inference through parallel acoustic generation. In alignment-free NAR models such as F5-TTS (Chen et al., 2025), target text spans must be mapped to jointly evolving acoustic regions during flow matching, without explicit phoneme alignment or duration modeling. This makes precise localization with natural transitions challenging. These limitations motivate a central question:

- *Would it be possible to unify precise word-level control, fine-grained intensity adjustment, and flexible emotion composition in a single TTS framework while retaining the inference speed of NAR generation?*

To this end, we propose **ControlAnything**, a unified framework for word-level, intensity-aware, and composable emotional control in non-autoregressive TTS. As illustrated in Figure 1, ControlAnything can localize emotions, adjust their intensity, and combine them within selected spans. First, we develop a **Soft-boundary Anchor Intervention (SAI)** strategy that repeatedly incorporates temporal anchors into DiT blocks during flow-matching generation. SAI localizes emotional intervention to the acoustic region corresponding to a target word or span while accommodating local changes and preserving smooth boundary transitions. Second, we introduce the **Emotion Concept Modulation (ECM)** module, which learns distinct emotion-specific modulation directions and calibrates their strength for continuous intensity control. Combined with SAI, ECM supports multi-emotion composition within selected spans with adjustable overall intensity. Finally, we comprehensively evaluate word-level localization, intensity adjustment, multi-emotion composition, speech quality, and generation efficiency using objective and subjective measures, demonstrating our method's effectiveness and generalizability.

2 Related Work

Emotionally Controllable TTS Emotionally controllable TTS has progressed from categorical labels and learned representations to textual prompts and latent manipulation (Lei et al., 2021; Im et al., 2022; Guo et al., 2023a; Cho et al., 2025). PromptTTS and InstructTTS provide flexible text-based conditioning (Guo et al., 2023b; Yang et al., 2024), while EmoSteer-TTS and CoCoEmo enable emotion conversion, intensity scaling, and mixed-emotion control through activation steering (Xie et al., 2025b; Wang et al., 2026). Despite this progress, these approaches predominantly impose utterance-level conditions, offering limited control over emotional variation within an utterance. Their growing expressive capabilities highlight the need to move beyond global emotion selection toward coordinated control over the location, strength, and composition of emotions.

Span, Intensity, and Composable Control Fine-grained emotional control specifies *where* emotions occur, *how strongly* they are conveyed, and *how they coexist*. Existing methods localize emotions through affective prediction (Luo et al., 2021), expressive-event modeling (Kanda et al., 2024; Wu et al., 2024), or autoregressive condition switching (Wang et al., 2025a; Liang et al., 2026). Intensity control uses soft labels (Guo et al., 2023a), geometric representations (Cho et al., 2025), or direction scaling (Xie et al., 2025b), while composition relies on emotion interpolation (Zhou et al.,

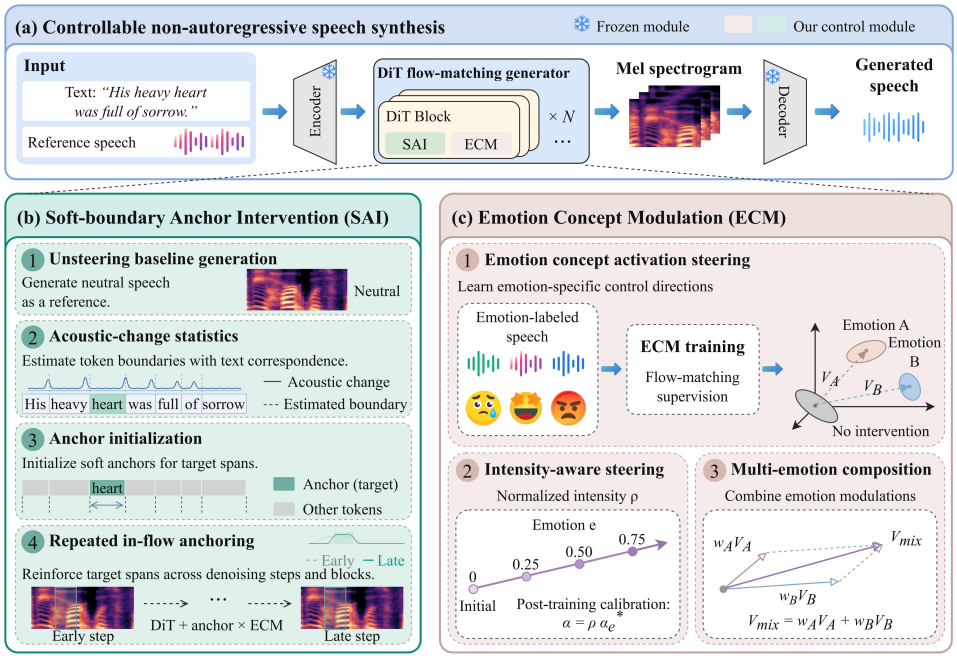


Figure 2: Overview of ControlAnything. SAI localizes word-level control with smooth transitions, while ECM enables adjustable intensity and multi-emotion composition.

2023) or direction combination (Xie et al., 2025b; Wang et al., 2026). However, these dimensions remain largely separate, limiting intensity adjustment and composition within selected spans, especially in NAR TTS, whose joint acoustic generation lacks explicit temporal anchors. Unifying all three dimensions within NAR generation remains an open challenge.

3 The ControlAnything Framework

Task Definition. Given a word sequence $\mathbf{y} = (y_1, \ldots, y_N)$ and speaker conditioning $\mathbf{s}$, we synthesize speech with word-level, intensity-aware, and composable emotional expression. We define a control request as $\mathcal{C} = (\mathcal{S}, \rho, \mathbf{w})$, where $\mathcal{S} = [a, b]$ specifies a target word span, with $1 \le a \le b \le N$, $\rho \in [0, 1]$ denotes normalized intervention intensity, and $\mathbf{w} = (w_1, \ldots, w_E)$ assigns nonnegative weights to E emotion concepts, with $\sum_{e=1}^{E} w_e = 1$. Setting $a = b$ targets a single word, while $a < b$ selects a multi-word span. A one-hot $\mathbf{w}$ selects one emotion, whereas multiple nonzero entries specify an emotional mixture. Synthesis is formulated as $\hat{\mathbf{x}} = F(\mathbf{y}, \mathbf{s}; \mathcal{C})$, aiming to realize the requested expression within the acoustic region corresponding to $\mathcal{S}$ while preserving linguistic content, speaker identity, and natural transitions across span boundaries. Here, $\mathcal{S}$, ρ, and $\mathbf{w}$ respectively specify *where* emotional control acts, *how strongly* it acts, and *how* emotions are composed. Setting $\mathcal{S} = [1, N]$ extends control to the entire utterance, while multiple requests $\mathbb{C} = \{\mathcal{C}_k\}_{k=1}^{K}$ allow distinct controls across spans.

To realize this control formulation, **ControlAnything** integrates SAI with ECM modules, as illustrated in Figure 2. SAI in panel (b) enables word-level control by using local acoustic changes to estimate token boundaries and define the intervention region corresponding to $\mathcal{S}$, with smooth boundary transitions. ECM modules embedded within DiT blocks translate ρ and $\mathbf{w}$ into intensity-aware, composable interventions along the acoustic generation trajectory.

3.1 Word-level Control via Soft-Boundary Anchor Intervention

Where to intervene. Word-level control has mainly been explored in AR TTS through segment-wise generation or decoding-time condition scheduling Wang et al. (2025a); Liang et al. (2026), with less exploration in NAR TTS. As shown in Figure 2(a), NAR TTS encodes text and reference speech to condition a DiT generator that synthesizes the full utterance. Boundary information can

be introduced at three locations: (1) the encoded conditioning before DiT, (2) the output through segment synthesis and concatenation, and (3) the intermediate representations within DiT blocks.

To investigate intervention placement, we conduct a pilot study using two statistical metrics. First, Relative acoustic-response centroid offset (ARC offset) measures $|\mu - m|$ in ten milliseconds, where m is the target midpoint and μ is the temporal centroid within the analysis window, weighted by absolute log-energy differences between controlled and uncontrolled speech. Second, Boundary step (BS) measures $|\bar{r}^{+} - \bar{r}^{-}|$, averaged over both target boundaries, where $\bar{r}^{-}$ and $\bar{r}^{+}$ denote mean acoustic responses within 50 ms before and after each boundary. The response is the framewise mean absolute log-mel difference between controlled and uncontrolled speech. Lower values indicate closer temporal alignment and smoother boundary transitions, respectively.

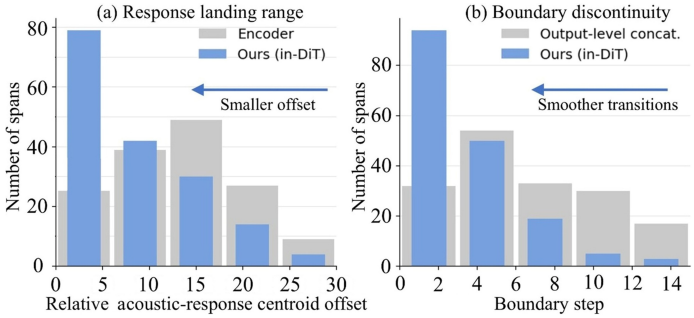


Figure 3: Pilot study on intervention placement.

As shown in Figure 3(a), In-DiT anchoring concentrates centroid offsets near zero, while Encoder anchoring yields a broader distribution with larger offsets. In-DiT exceeds Encoder by 31.6% in the proportion with a centroid offset below 50 ms. In panel (b), In-DiT anchoring produces predominantly small boundary steps, whereas output-level concatenation exhibits more large discontinuities, and exceeds output-level concatenation by 35.7% in the proportion with a boundary step below 2 dB. These trends support In-DiT anchoring for localized control with smooth transitions.

To complement these distributional observations, we quantify the trade-offs among localization, continuity, emotional expression, and efficiency. Table 1 reports Emo2v., ARC offset, BS, and real-time factor (RTF). Output-level splicing yields the smallest ARC offset but the largest BS, indicating a trade-off between localization and continuity. Encoder achieves the lowest BS but the largest ARC offset and lowest Emo2v. In-DiT reduces ARC offset by 43.6% relative to Encoder and BS by 3.43 dB relative to output-level splicing, while matching the latter's Emo2v. It also achieves the lowest RTF. These results support boundary anchoring within DiT blocks as an effective way to balance localization, continuity, emotional expression, and efficiency.

Table 1: Comparison of intervention settings.

Setting	Emo2v. ↑	ARC offset ↓	BS ↓	RTF ↓
Output splice	**0.617**	**2.517**	6.163	0.267
Encoder	0.558	12.374	**0.392**	0.144
In-DiT	**0.617**	6.976	2.736	**0.137**

Takeaway. Embedding boundary anchors throughout the DiT generation process better balances target-span control with continuity across span boundaries.

Soft Boundary Anchoring. Given the input text $\mathbf{y}$ and speaker conditioning $\mathbf{s}$, we first synthesize a unsteered reference utterance $\mathbf{x}^0$ and estimate the start and end times of the words in the target span $\mathcal{S} = [a, b]$. These word boundaries provide initial temporal anchors for locating emotional intervention. We convert the timestamps to acoustic-frame indices, accounting for the reference prefix, and construct a temporal anchor $\mathbf{m} \in [0, 1]^T$ that identifies the target region within the acoustic sequence. Further details of anchor initialization are provided in Appendix A.4.

Based on the initialized temporal anchors, we introduce emotional modulation at the output of each DiT block throughout flow-matching generation, as illustrated in Figure 2. This progressively integrates local emotional control into the evolving acoustic representations. Let $\mathbf{H}_{\tau}^{\ell}$ denote the output of block ℓ at flow time τ, and let $\mathcal{I}^{\ell}$ denote the emotional transformation defined by the ECM module in Section 3.2. The anchored update is

$$\widetilde{\mathbf{H}}_{\tau}^{\ell} = (\mathbf{1} - \mathbf{m}) \odot \mathbf{H}_{\tau}^{\ell} + \mathbf{m} \odot \mathcal{I}^{\ell}\big(\mathbf{H}_{\tau}^{\ell}\big)\,, \tag{1}$$

where $\mathbf{m}$ is broadcast across hidden channels and specifies the temporal support of direct modulation. While the anchor fixes the intervention region, contextual interactions allow the target and

surrounding acoustic representations to evolve jointly as generation proceeds. This joint evolution helps soften acoustic transitions across the anchored boundaries, supporting word-level emotional control while preserving overall speech fluency.

3.2 Emotion Concept Modulation

Learning emotion-specific modulation. The ECM module uses a separate branch for each emotion to generate channel-wise scaling and shifting parameters for each DiT block:

$$(\boldsymbol{\gamma}_e^\ell, \boldsymbol{\beta}_e^\ell) = \mathcal{M}_e^\ell(\boldsymbol{\phi}_e), \qquad \ell = 1, \dots, L, \tag{2}$$

where $\mathcal{M}_e^\ell$ maps the learnable parameters $\boldsymbol{\phi}_e$ of emotion branch e to modulation vectors $\boldsymbol{\gamma}_e^\ell, \boldsymbol{\beta}_e^\ell \in \mathbb{R}^d$ for block ℓ. Here, L and d denote the number of blocks and hidden dimension, respectively.

Each branch takes a constant input and consists of a multilayer perceptron with block-specific output heads. A hyperbolic tangent mapping bounds the modulation parameters, while zero-initialized heads make the intervention an identity transformation at initialization. Scaling reweights acoustic features, and shifting introduces an emotion-specific offset. Their interaction with evolving hidden representations enables state-dependent emotional modulation.

We train ECM on emotion-labeled acoustic targets using conditional flow matching (CFM). Let v denote the velocity predictor incorporating ECM, conditioned on text and acoustic information $\mathbf{c}$ and the target emotion e. The training objective is

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}\Big[\, \| v(\mathbf{z}_\tau, \tau, \mathbf{c}, e) - \mathbf{u}_\tau \|_F^2 \Big], \tag{3}$$

where $\mathbf{z}_\tau$ is an intermediate acoustic state at flow time τ, and $\mathbf{u}_\tau$ is the corresponding target velocity.

During training, we activate the branch corresponding to the target emotion and jointly optimize ECM and the DiT backbone using the backbone's original conditional flow-matching objective on emotion-labeled acoustic targets.

Intensity-aware control. Building on the learned emotion-specific modulation, we can determine the steering range of each emotion branch after training. We progressively increase its steering strength and monitor synthesis quality, using the onset of quality degradation to identify the upper operating limit $\alpha_e^\star$.

Given a normalized intensity request $\rho \in [0, 1]$, we set $\alpha = \rho\alpha_e^\star$ and apply the resulting strength to the emotion-specific modulation:

$$\mathcal{I}_e^\ell\big(\mathbf{H}_\tau^\ell\big) = \mathbf{H}_\tau^\ell + \alpha \left(\boldsymbol{\gamma}_e^\ell \odot \mathbf{H}_\tau^\ell + \boldsymbol{\beta}_e^\ell\right), \tag{4}$$

where the modulation vectors are broadcast across acoustic frames.

Setting $\rho = 0$ disables modulation, while increasing ρ scales the intervention toward the calibrated limit. Combined with the temporal anchor in Eq. 1, this enables control over emotional intensity within the target span without additional controller training.

Composable emotional steering. Finally, we extend emotion-specific steering to multi-emotion composition by combining the learned modulation parameters. Given nonnegative weights $\mathbf{w}$ with $\sum_{e=1}^E w_e = 1$, the composed transformation at block ℓ and flow time τ is

$$\mathcal{I}^\ell\big(\mathbf{H}_\tau^\ell\big) = \mathbf{H}_\tau^\ell + \alpha \sum_{e=1}^{E} w_e \left(\boldsymbol{\gamma}_e^\ell \odot \mathbf{H}_\tau^\ell + \boldsymbol{\beta}_e^\ell\right), \tag{5}$$

where w_e specifies the relative contribution of emotion e, and α controls the overall control strength.

The modulation vectors are broadcast across acoustic frames. A one-hot $\mathbf{w}$ selects a single emotion, while multiple nonzero entries produce a weighted combination of emotion-specific scaling and shifting parameters. This formulation separates emotion composition from overall steering strength, enabling flexible mixtures without training combination-specific controllers.

Combined with the temporal anchor in Eq. 1, the composed transformation supports emotional mixtures across an entire utterance or within selected word spans, unifying localization, intensity adjustment, and emotion composition within a shared modulation mechanism.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENT SETTINGS

Datasets. We combine ESD (Zhou et al., 2021), RAVDESS (Livingstone & Russo, 2018), MSP-Podcast (Busso et al., 2026), and TESS Pichora-Fuller & Dupuis (2020), following the emotion filtering and sample selection procedures of Wang et al. (2026) and Xie et al. (2025b). Where feasible, we prioritize within-speaker comparisons for each target emotion to mitigate speaker-related confounding. The combined dataset covers six basic emotions (anger, happiness, sadness, disgust, surprise, and fear) and a neutral category, with approximately 1,000 samples per category. It is split into training, validation, and test sets at a 0.5/0.2/0.3 ratio. To further assess the generalization ability of the proposed method, we conduct an out-of-distribution (OOD) evaluation using CREMA-D (Cao et al., 2014) as an additional test set. Details please refer to Appendix A.3.

Backbones and Baselines. We evaluate the proposed framework across multiple pretrained TTS models, with Matcha-TTS (Mehta et al., 2024) and F5-TTS (Chen et al., 2025) serving as the primary backbones for experiments under the NAR paradigm. We further assess the generalizability of the proposed emotion steering approach to the AR paradigm using CosyVoice2 (Du et al., 2024). For comparison, we also evaluate many representative external TTS systems, including GPT-4o-mini-TTS[1], Qwen3-TTS (Hu et al., 2026), DoubaoTTS[2], SparkTTS (Wang et al., 2025b), MaskGCT (Wang et al., 2024), and IndexTTS2 (Zhou et al., 2025), using their native inference interfaces. Whenever possible, all systems are evaluated with matched input texts, speakers, and reference conditions.

Evaluation Metrics. We evaluate multiple control settings: (1) For utterance-level control, we use SER accuracy and emotion embedding similarity (E-SIM) to assess target-emotion realization. (2) For word-level control, following Wang et al. (2025a), we use DNSV-Pro for prosodic transition consistency and AutoPCP, Emo2v., and Aro. for prosodic, emotional, and arousal similarity, respectively. (3) For intensity-aware control, following Xie et al. (2025b), we use Spearman's ρ for intensity ranking agreement and ΔEmo2v. for emotional change. In the joint-control and ablation tables, ρ_I and ρ_W denote Spearman's rank correlations for emotion intensity and composition weights, respectively. (4) For compositional control, following Wang et al. (2026), we use E-SIM, Spearman's ρ, and H-Rate for emotion similarity, composition agreement, and target-combination success, respectively.

Across settings, WER, S-SIM, and UTMOS measure intelligibility, speaker preservation, and speech quality, respectively, while RTF measures synthesis efficiency. Subjective evaluation covers naturalness (N-MOS), speaker similarity (S-MOS), and emotional expressiveness (E-MOS), with emotion interpolation MOS (EI-MOS) for the smoothness and consistency of intensity progression. Please refer to Appendix A.1.1 for metric definitions and calculation details.

4.2 UTTERANCE-LEVEL EMOTION CONTROL PERFORMANCE

As shown in Table 2, our method enables effective utterance-level emotion control on both NAR backbones. Its F5-TTS variant achieves the best SER and E-SIM under both ID and OOD evaluation, outperforming GPT-4o-mini-TTS and MaskGCT, the strongest evaluated closed-source and open-source baselines by SER, respectively. Compared with the original F5-TTS, it improves SER by 79.4 and 54.7 percentage points and E-SIM by 101.1% and 23.3%, respectively.

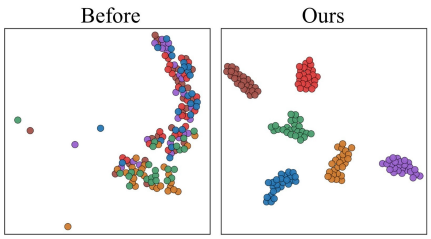


Figure 4: Emotion embedding t-SNE.

It also achieves the highest S-SIM, UTMOS, and mean N-MOS, S-MOS, and E-MOS scores in both settings, demonstrating strong emotion control while preserving speaker fidelity and speech quality, with only a small additional inference cost.

To visualize the effect of emotion control, we compare t-SNE projections of F5-TTS emotion representations before and after applying our framework (Figure 4), with colors denoting emotion cate-

[1] https://developers.openai.com/api/docs/models/gpt-4o-mini-tts

[2] https://docs.volcengine.com/docs/DoubaoVoice/unidirectional-streaming-text-to-speech-http?lang=zh

Table 2: Utterance-level emotion control under in- and out-of-distribution evaluation.

	Control interface		*Quality / fidelity*			*Emotion accuracy*		*Subjective (MOS)*			*Efficiency*
Method	Prompt	Activation	WER↓	S-SIM↑	UTMOS↑	SER↑	E-SIM↑	N-MOS↑	S-MOS↑	E-MOS↑	RTF↓
In-distribution evaluation (ESD-EN)											
CLOSE SOURCE											
GPT-4o-mini-TTS	✓	×	7.288	0.683	3.942	0.389	0.597	$3.86_{\pm 0.17}$	$4.66_{\pm 0.20}$	$2.29_{\pm 0.29}$	–
DoubaoTTS	✓	×	6.790	0.675	3.862	0.239	0.527	$3.31_{\pm 0.36}$	$4.54_{\pm 0.24}$	$1.69_{\pm 0.24}$	–
OPEN SOURCE											
Qwen3TTS	✓	×	8.863	0.698	4.207	0.181	0.502	$3.57_{\pm 0.28}$	$4.48_{\pm 0.17}$	$2.00_{\pm 0.00}$	0.891
IndexTTS2	✓	×	4.782	0.919	4.150	0.497	0.702	$3.60_{\pm 0.23}$	$4.70_{\pm 0.16}$	$1.95_{\pm 0.18}$	0.872
CosyVoice2	✓	×	3.516	0.899	4.117	0.339	0.640	$3.69_{\pm 0.16}$	$4.21_{\pm 0.36}$	$2.33_{\pm 0.29}$	0.856
SparkTTS	✓	×	5.606	0.877	4.187	0.335	0.597	$3.83_{\pm 0.22}$	$4.35_{\pm 0.31}$	$2.52_{\pm 0.28}$	0.976
MaskGCT	✓	×	5.368	0.866	3.812	0.660	0.772	$3.50_{\pm 0.17}$	$4.50_{\pm 0.35}$	$1.93_{\pm 0.27}$	0.542
Matcha-TTS	×	×	8.756	0.876	3.724	N/A	N/A	$3.95_{\pm 0.12}$	$4.61_{\pm 0.18}$	$2.70_{\pm 0.20}$	**0.017**
F5-TTS	×	×	**3.183**	0.918	4.293	0.104	0.449	$3.60_{\pm 0.32}$	$4.75_{\pm 0.12}$	$2.05_{\pm 0.26}$	0.080
Ours (Matcha-TTS)	×	✓	8.455	0.840	3.275	0.555	0.715	$3.75_{\pm 0.31}$	$4.50_{\pm 0.31}$	$3.07_{\pm 0.18}$	0.019
Ours (F5-TTS)	×	✓	4.670	**0.924**	**4.295**	**0.898**	**0.903**	$\mathbf{4.32}_{\pm 0.21}$	$\mathbf{4.89}_{\pm 0.21}$	$\mathbf{3.24}_{\pm 0.20}$	0.087
Out-of-distribution evaluation (CREMA-D)											
CLOSE SOURCE											
GPT-4o-mini-TTS	✓	×	4.167	0.653	3.938	0.441	0.788	$3.80_{\pm 0.20}$	$4.73_{\pm 0.25}$	$1.57_{\pm 0.20}$	–
DoubaoTTS	✓	×	3.299	0.685	3.789	0.185	0.706	$3.85_{\pm 0.29}$	$4.64_{\pm 0.16}$	$1.79_{\pm 0.25}$	–
OPEN SOURCE											
Qwen3TTS	✓	×	7.431	0.555	4.208	0.226	0.722	$3.83_{\pm 0.14}$	$4.60_{\pm 0.19}$	$2.71_{\pm 0.18}$	0.891
IndexTTS2	✓	×	8.333	0.855	2.406	0.299	0.728	$3.92_{\pm 0.09}$	$4.79_{\pm 0.21}$	$2.76_{\pm 0.24}$	0.872
CosyVoice2	✓	×	4.936	0.775	3.669	0.325	0.777	$3.88_{\pm 0.16}$	$4.55_{\pm 0.31}$	$2.83_{\pm 0.17}$	0.856
SparkTTS	✓	×	4.803	0.634	3.174	0.335	0.782	$3.97_{\pm 0.19}$	$4.55_{\pm 0.12}$	$2.45_{\pm 0.34}$	0.976
MaskGCT	✓	×	4.707	0.625	2.991	0.391	0.782	$3.45_{\pm 0.25}$	$4.67_{\pm 0.18}$	$1.57_{\pm 0.28}$	0.542
Matcha-TTS	×	×	6.250	0.661	3.218	N/A	N/A	$3.85_{\pm 0.12}$	$4.31_{\pm 0.22}$	$2.45_{\pm 0.25}$	**0.017**
F5-TTS	×	×	3.075	0.857	4.208	0.249	0.696	$3.95_{\pm 0.12}$	$4.83_{\pm 0.14}$	$2.61_{\pm 0.22}$	0.080
Ours (Matcha-TTS)	×	✓	4.563	0.657	3.033	0.600	0.813	$4.03_{\pm 0.14}$	$4.77_{\pm 0.18}$	$2.89_{\pm 0.28}$	0.019
Ours (F5-TTS)	×	✓	**2.778**	**0.859**	**4.219**	**0.796**	**0.858**	$\mathbf{4.36}_{\pm 0.08}$	$\mathbf{4.92}_{\pm 0.11}$	$\mathbf{3.12}_{\pm 0.17}$	0.087

gories. Before intervention, samples from different categories overlap considerably, showing limited emotional separation in the projected space. In contrast, our method produces more compact and clearly separated clusters, suggesting greater within-category consistency and stronger control performance. For the plotted samples, local emotion purity in the projected space improves by 75.2%, further supporting enhanced emotional separation.

4.3 Word-level Emotion Control Performance

We next evaluate word-level emotion control. Following WeSCon (Wang et al., 2025a), we adapt utterance-level TTS systems into segment-wise baselines by splitting text according to emotion assignments, synthesizing each span separately, and concatenating the resulting speech.

Table 3: Word-level emotion control performance.

	Quality / transition			*Expression similarity*			*Subjective MOS*			*Efficiency*
Method	WER (%)↓	DNSV-Pro↓	S-SIM↑	AutoPCP↑	Emo2v.↑	Aro.↑	E-MOS↑	N-MOS↑	S-MOS↑	RTF↓
GPT-4o-mini-TTS	<u>0.981</u>	13.240	0.873	2.471	0.422	0.602	$2.12_{\pm 0.17}$	$3.64_{\pm 0.18}$	$4.82_{\pm 0.23}$	-
DoubaoTTS	1.037	6.671	0.873	<u>2.560</u>	0.520	0.596	$2.21_{\pm 0.17}$	$3.87_{\pm 0.18}$	$4.68_{\pm 0.09}$	-
Qwen3TTS	1.536	6.248	0.861	2.473	0.391	0.543	$2.26_{\pm 0.14}$	$3.83_{\pm 0.22}$	$4.85_{\pm 0.39}$	3.211
IndexTTS2-seg	**0.572**	<u>5.682</u>	<u>0.875</u>	2.401	<u>0.621</u>	0.661	$2.10_{\pm 0.16}$	$3.60_{\pm 0.14}$	$4.87_{\pm 0.14}$	3.128
F5-TTS-seg	5.116	9.096	0.853	2.441	0.498	0.660	$2.17_{\pm 0.22}$	$3.31_{\pm 0.24}$	$4.89_{\pm 0.17}$	1.295
SparkTTS-seg	4.291	8.985	0.864	2.406	0.522	0.646	$2.24_{\pm 0.20}$	$3.62_{\pm 0.29}$	$4.14_{\pm 0.48}$	1.582
CosyVoice2-seg	6.834	8.661	0.865	2.393	0.588	<u>0.674</u>	$2.00_{\pm 0.13}$	$3.69_{\pm 0.24}$	$4.76_{\pm 0.28}$	4.084
Ours	3.562	**5.074**	**0.883**	**2.827**	**0.646**	**0.733**	$\mathbf{2.46}_{\pm 0.08}$	$\mathbf{4.22}_{\pm 0.12}$	$\mathbf{4.97}_{\pm 0.11}$	**0.137**

Word-level evaluation. As shown in Table 3, our method achieves effective local expression with smooth transitions and strong speaker consistency. Compared with F5-TTS using segment-wise synthesis, it reduces DNSV-Pro by 4.02 and improves AutoPCP by 0.386, indicating smoother transi-

tions and improved target prosody. It also outperforms Qwen3TTS and IndexTTS2-seg, the strongest evaluated closed-source and open-source baselines by DNSV-Pro, respectively.

Our method further achieves the highest arousal, speaker similarity, and mean E-MOS, N-MOS, and S-MOS scores, supporting natural and expressive word-level control ability. Unless otherwise specified, 'Ours' appearing in the tables below refers to F5-TTS as the backbone version.

4.4 Intensity-Aware Emotion Control Performance

Following the emotion interpolation protocol (Xie et al., 2025b), we evaluate intensity-aware control by progressively increasing the requested intensity from neutral to the target emotion, assessing steering continuity, modulation range, and synthesis quality.

Table 4: Intensity-aware control performance.

Method	$\rho \uparrow$	ΔEmo2v.↑	WER (%)↓	S-SIM↑	EI-MOS↑
GPT-4o-mini-TTS	0.465	0.335	3.659	0.669	$3.65_{\pm 0.49}$
Qwen3TTS	0.112	0.128	6.093	0.615	$2.30_{\pm 0.52}$
DoubaoTTS	0.216	0.229	**3.011**	0.622	$2.81_{\pm 0.47}$
IndexTTS2	0.527	0.411	3.147	**0.935**	$3.88_{\pm 0.62}$
Matcha-TTS	0.465	0.332	11.142	0.846	$3.76_{\pm 0.74}$
Ours	**0.546**	**0.474**	4.236	0.918	$4.12_{\pm 0.29}$

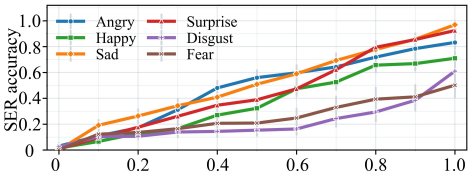


Figure 5: Steering intensity performance.

Intensity-aware evaluation. As shown in Table 4, our method achieves the best intensity monotonicity and modulation range, surpassing IndexTTS2, the strongest open-source baseline on both metrics, by 0.019 in Spearman's ρ and 0.063 in ΔEmo2v. It also outperforms the strongest evaluated closed-source system, GPT-4o-mini-TTS, on both metrics. Speaker similarity ranks second, although intelligibility remains below some baselines. Its mean EI-MOS exceeds the best baseline by 0.24 points, supporting smoother and more consistent emotional progression.

Figure 5 further illustrates how different emotions respond to the control strength. As the control strength increases, the predicted probabilities of all six target emotions show an overall upward trend, supporting gradual intensity modulation across emotion categories.

4.5 Composable Emotion Control Performance

Table 5: Composable emotion performance.

System	E-SIM↑	$\rho \uparrow$	H-Rate↑	WER (%)↓	S-SIM↑
Same-span emotion coupling					
GPT-4o-mini-TTS	0.702	0.217	0.042	3.318	**0.960**
Qwen3TTS	0.704	0.157	0.042	21.572	0.906
DoubaoTTS	0.752	–	–	1.441	0.953
IndexTTS2	0.720	0.574	0.167	2.129	0.907
Ours	**0.861**	**0.662**	**0.458**	**0.709**	0.924
Multi-span multi-emotion					
GPT-4o-mini-TTS	0.608	0.044	0.375	2.102	0.947
Qwen3TTS	0.451	0.017	0.375	16.069	0.911
DoubaoTTS	0.634	0.131	0.250	7.842	0.939
IndexTTS2-seg	0.748	0.299	0.625	5.648	0.936
F5-TTS-seg	0.644	0.118	0.750	17.919	0.915
Ours	**0.891**	**0.822**	**0.900**	**0.621**	**0.987**

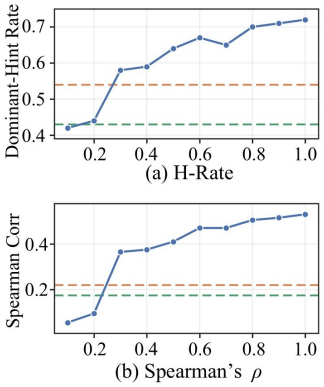


Figure 6: Composable emotion trend

Multi-emotion composable evaluation. Table 5 evaluates compositional emotion control in two settings: mixing emotions throughout an utterance and assigning distinct emotions to its two halves. In the utterance-level mixing setting, our method outperforms baselines generally, indicating closer adherence to the target emotion weights and their intended hierarchy. When distinct emotions are assigned to the two halves of an utterance, our method achieves the best results across all metrics, improving E-SIM by 0.143 over IndexTTS2-seg and H-Rate by 15.0 percentage points over F5-TTS-seg, the strongest baselines for the respective metrics. These results demonstrate effective emotion mixing throughout an utterance and localized control of distinct emotions within it.

To further investigate whether the method supports fine-grained intensity adjustment when composing multiple emotions, we progressively increase steering strength while fixing the text and emotion mixing weights. In Figure 6, the blue curve represents our method, while the green and orange dashed lines indicate the scores of CosyVoice2 and GPT-4o-mini-TTS, respectively. Both ρ and H-Rate exhibit an overall upward trend as the control intensity increases, suggesting that stronger steering brings the relative emotion gains into closer agreement with the intended mixture hierarchy.

4.6 Joint evaluation of all three control dimensions.

Table 6: Evaluation of joint control across all three dimensions.

Method	N-MOS↑	DNSV-Pro↓	ρ_I ↑	ρ_W ↑	H-Rate↑	WER (%)↓
GPT-4o-mini-TTS	$3.68_{\pm 0.52}$	22.193	0.104	0.026	0.356	0.731
Qwen3-TTS	$3.57_{\pm 0.63}$	12.847	-0.054	0.018	0.338	0.684
F5-TTS-seg	$3.11_{\pm 0.21}$	42.688	–	0.130	0.379	**0.287**
Ours	$\mathbf{4.23_{\pm 0.37}}$	**9.916**	**0.326**	**0.184**	**0.458**	0.443

Table 7: Evaluation settings.

Word	Int.	Comp.	Section
✓	✓	–	Sec.A.2.1
✓	–	✓	Sec.A.2.2
–	✓	✓	Sec.A.2.3
✓	✓	✓	Sec.4.6

Combined three dimensions control evaluation. As shown in Table 6, our method achieves the best results across all control metrics when word-level localization, intensity adjustment, and emotion composition are applied jointly. Compared with F5-TTS-seg, it improves ρ_W by 0.054 and H-Rate by 0.079 while achieving a lower DNSV-Pro, indicating smoother local emotional transitions while meeting compositional control requirements. Furthermore, our method achieves the lowest DNSV-Pro and the highest mean N-MOS, exceeding the best baseline in N-MOS by 0.55 points. These results indicate that synthesized speech retains good transition smoothness and naturalness under multidimensional control.

We also evaluate pairwise control combinations, with settings and analyses summarized in Table 7 (Int.: intensity-aware; Comp.: composable). Additional experiments on control combinations and comparisons with CosyVoice2 are provided in Appendix A.2.

Takeaway. ControlAnything uses a shared intervention mechanism to jointly support word-level, intensity-aware, and compositional control without combination-specific controllers.

4.7 Ablation Study

Table 8: Component ablation study.

Method	N-MOS↑	DNSV-Pro↓	ρ_I ↑	ρ_W ↑	H-Rate↑	WER (%)↓
Full Ours	$\mathbf{4.23_{\pm 0.37}}$	**9.916**	**0.326**	**0.184**	**0.458**	**0.443**
w/o SAI	$3.58_{\pm 0.81}$	21.809	0.279	0.162	0.451	1.659
w/o ECM	$3.18_{\pm 0.54}$	29.803	-	-	0.305	0.724

Our pilot study in Table 1 identifies intervention within DiT blocks as the most effective placement for emotional control. Building on this finding, we conduct ablation studies to assess the contribution of each component in ControlAnything, as shown in Table 8. The full model outperforms both ablated variants in transition smoothness, speech naturalness, while achieving the lowest WER. Without SAI, DNSV-Pro nearly doubles and N-MOS declines, indicating that local anchoring helps maintain smooth and natural emotional transitions. Removing ECM reduces H-Rate by 0.153 and worsens transition smoothness, therefore without ECM, emotional regulation function will almost fail.

5 Conclusion

In this work, we presented ControlAnything, a unified framework for word-level, intensity-aware, and composable emotional control in non-autoregressive TTS. By integrating Soft-boundary Anchor Intervention (SAI) with Emotion Concept Modulation (ECM), it supports localized expression, adjustable intensity, and multi-emotion composition through a shared intervention mechanism. Objective and subjective evaluations demonstrate effective control across individual and combined dimensions while maintaining speech naturalness, smooth transitions, and speaker consistency. The intervention-location experiment supports intervention within DiT blocks, while component ablations further validate the framework's design. These findings demonstrate the feasibility of flexible multidimensional emotional control with efficient NAR TTS synthesis.

AI USE STATEMENT

During this revision, we used Codex (GPT-5.6 sol) to assist with manuscript editing, LaTeX preparation, and consistency checks. The LLMs and audio language models evaluated as simulator backbones are documented separately in the experimental setup. We reviewed all AI-assisted work and take full responsibility for the final manuscript and reported results.

ETHICS STATEMENT

This work studies fine-grained controllable emotional speech synthesis. All datasets used in our experiments are publicly available research datasets and are used in accordance with their respective licenses and intended research purposes. We acknowledge that expressive speech synthesis technologies may potentially be misused for voice impersonation or deceptive content generation. Our work is intended solely for research purposes, and we encourage responsible use of the proposed techniques.

REPRODUCIBILITY STATEMENT

The main text describes the experimental setup and evaluation protocol, while the appendix provides additional details on dataset construction, evaluation metrics, implementation settings, and supplementary experimental results. We also include case studies and additional analyses to facilitate the verification of our findings. An anonymized demo page with audio samples is available at `https://controlanything-review-demo.vercel.app/`.

REFERENCES

Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N. Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, Huang-Cheng Chou, and Pravin Mote. The msp-podcast corpus. *IEEE Transactions on Affective Computing*, 17 (3):3065–3083, 2026. doi: 10.1109/TAFFC.2026.3678489.

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

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.313. URL `https://aclanthology.org/2025.acl-long.313/`.

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. doi: 10.1109/TAFFC.2025.3561267. URL `https://doi.org/10.1109/TAFFC.2025.3561267`.

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

Wenhao Guan, Yishuang Li, Tao Li, Hukai Huang, Feng Wang, Jiayan Lin, Lingyan Huang, Lin Li, and Qingyang Hong. MM-TTS: Multi-modal prompt based style transfer for expressive text-to-speech synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18117–18125, 2024. doi: 10.1609/aaai.v38i16.29769. URL `https://ojs.aaai.org/index.php/AAAI/article/view/29769`.

Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu. EmoDiff: Intensity controllable emotional text-to-speech with soft-label guidance. In *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 *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

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 *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6317–6321. IEEE, 2022.

Naoyuki Kanda, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Hemin Yang, Zirun Zhu, Min Tang, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, et al. Making flow-matching-based zero-shot text-to-speech laugh as you like. *arXiv preprint arXiv:2402.07383*, 2024.

Yi Lei, Shan Yang, and Lei Xie. Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 423–430. IEEE, 2021.

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. *arXiv preprint arXiv:2601.03170*, 2026.

Rui Liu, Jiatian Xi, Ziyue Jiang, and Haizhou Li. FluentEditor2: Text-based speech editing by modeling multi-scale acoustic and prosody consistency. *IEEE Transactions on Audio, Speech and Language Processing*, 2025. doi: 10.1109/TASLPRO.2025.3624913. URL https://doi.org/10.1109/TASLPRO.2025.3624913.

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.

Xuan Luo, Shinnosuke Takamichi, Tomoki Koriyama, Yuki Saito, Hiroshi Saruwatari, et al. Emotion-controllable speech synthesis using emotion soft labels and fine-grained prosody factors. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 794–799. IEEE, 2021.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11341–11345, 2024. doi: 10.1109/ICASSP48485.2024.10448291.

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

Yan Rong, Shan Yang, Chenxing Li, Dong Yu, and Li Liu. Dopamine Audiobook: A training-free MLLM agent for emotional and immersive audiobook generation. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2526–2538, 2026. doi: 10.1109/TASLPRO.2026.3688415. URL https://doi.org/10.1109/TASLPRO.2026.3688415.

Andreas Triantafyllopoulos, Björn W. Schuller, Gökçe İymen, Metin Sezgin, Xiangheng He, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, Ruibo Fu, and Jianhua Tao. An overview of affective speech synthesis and conversion in the deep learning era. *Proceedings of the IEEE*, 111(10):1355–1381, 2023. doi: 10.1109/JPROC.2023.3250266. URL https://doi.org/10.1109/JPROC.2023.3250266.

Siyi Wang, Shihong Tan, Siyi Liu, Hong Jia, Gongping Huang, James Bailey, and Ting Dang. CoCoEmo: Composable and controllable human-like emotional TTS via activation steering. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=wW2LIHSCfw.

Tianrui Wang, Haoyu Wang, Meng Ge, Cheng Gong, Chunyu Qiang, Ziyang Ma, Zikang Huang, Guanrou Yang, Xiaobao Wang, Eng Siong Chng, Xie Chen, Longbiao Wang, and Jianwu Dang. Word-level emotional expression control in zero-shot text-to-speech synthesis. *arXiv preprint arXiv:2509.24629*, 2025a.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens, 2025b. URL https://arxiv.org/abs/2503.01710.

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

Haibin Wu, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Daniel Tompkins, Chung-Hsien Tsai, Canrun Li, Zhen Xiao, Sheng Zhao, Jinyu Li, et al. Laugh now cry later: Controlling time-varying emotional states of flow-matching-based zero-shot text-to-speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 690–697. IEEE, 2024.

Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. Towards controllable speech synthesis in the era of large language models: A systematic survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 764–791. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.emnlp-main.40. URL https://aclanthology.org/2025.emnlp-main.40/.

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, 2025b. URL https://arxiv.org/abs/2508.03543.

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.

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. EmoVoice: LLM-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10748–10757. Association for Computing Machinery, 2025. doi: 10.1145/3746027.3754829. URL https://doi.org/10.1145/3746027.3754829.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 920–924. IEEE, 2021.

Kun Zhou, Berrak Sisman, R. Rana, B. W. Schuller, and Haizhou Li. Speech synthesis with mixed emotions. *IEEE Transactions on Affective Computing*, 14(4):3120–3134, 2023.

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

A Appendix

This appendix provides evaluation details in Section A.1, additional experimental results in Section A.2, the dataset construction procedure in Section A.3, and implementation details in Section A.4.

A.1 Evaluation Details

This section describes the objective and perceptual metrics used in our experiments, covering speech quality, intelligibility, speaker preservation, generation efficiency, and the different dimensions of emotional control.

A.1.1 Evaluation Metrics

We evaluate speech quality, emotion controllability, localization, and speaker preservation using objective and perceptual metrics. Unless otherwise specified, higher values indicate better performance, except for WER, RTF, ARC offset, BS, and DNSV-Pro, where lower values are preferred.

(1) Content, speaker identity, quality, and efficiency. **Word error rate (WER)** measures transcription errors between ASR outputs and the reference text, and is used to evaluate speech intelligibility. **Speaker similarity (S-SIM)** measures the cosine similarity between speaker embeddings of synthesized and reference speech. **UTMOS** provides an automatic estimate of perceptual speech quality, while **real-time factor (RTF)** measures inference efficiency as the ratio between synthesis time and generated audio duration.

(2) Emotion realization. **Speech emotion recognition (SER)** reports the proportion of synthesized samples recognized as the requested emotion, while **target emotion probability (TEP)** measures the classifier confidence assigned to the requested emotion. **Emotion similarity (E-SIM)** evaluates the similarity between emotion2vec representations of synthesized speech and the target emotional reference, following prior emotional TTS evaluation protocols (Xie et al., 2025b; Wang et al., 2026).

(3) Word-level control and temporal consistency. Following Wang et al. (2025a), we use **AutoPCP** to evaluate local prosodic consistency and **Emo2v.** and **Aro.** to measure emotion- and arousal-related similarity between the generated target segment and its corresponding emotional reference. **DNSV-Pro** measures the temporal variance of sliding-window DNSMOS-Pro predictions and is used to assess the consistency of speech quality across an utterance (Wang et al., 2025a).

(4) Localization and boundary continuity. **Acoustic-response centroid offset (ARC offset)** measures the temporal distance between the center of the acoustic response induced by control and the center of the requested target span. **Boundary step (BS)** measures the acoustic discontinuity around the start and end boundaries of the controlled span. Lower values indicate more accurate localization and smoother boundaries, respectively.

(5) Intensity control. We use ρ_I, the Spearman correlation between requested emotion intensity and the corresponding target-emotion response, to measure the monotonicity of intensity control. **ΔEmo2v.** measures the change in the emotion2vec target-emotion score between the lowest and highest control strengths, reflecting the effective modulation range. For perceptual evaluation, **EI-MOS** evaluates whether emotion changes smoothly and consistently with increasing requested intensity, following Xie et al. (2025b).

(6) Composable emotion control. Following Wang et al. (2026), ρ_W measures the Spearman rank agreement between requested emotion weights and the corresponding emotion responses. **Dominant-hit rate (H-Rate)** measures the proportion of samples for which the requested dominant emotion exhibits the largest relative increase in predicted probability. Together, these metrics evaluate both the relative ordering and the dominant component of composable emotion control.

(7) Perceptual evaluation. We conduct Mean Opinion Score (MOS) evaluations on a five-point scale. **N-MOS** evaluates perceived naturalness and continuity, **S-MOS** evaluates speaker similarity, and **E-MOS** evaluates agreement with the requested emotion. **EI-MOS** is used for intensity interpolation as described above. The reported scores are averaged over valid listener ratings.

A.2 More Experimental Results

This section supplements the main experiments with pairwise combinations of word-level, intensity-aware, and compositional control in Sections A.2.1–A.2.3. We further examine the adaptation of ECM to CosyVoice2 in Section A.2.4 and present qualitative examples of localized emotional control in Section A.2.5.

A.2.1 Word-Level + Intensity-Aware Control

Table 9: Joint word-level and intensity-aware emotion control performance.

Method	WER (%)↓	DNSV-Pro↓	S-SIM↑	AutoPCP↑	ρ_I ↑	ΔEmo2v.↑
GPT-4o-mini-TTS	0.544	14.111	0.861	2.526	0.211	0.184
DoubaoTTS	1.248	12.008	0.877	2.460	0.170	0.174
Qwen3TTS	0.797	11.610	0.883	2.382	0.082	0.120
IndexTTS2-seg	0.447	16.734	0.857	2.447	0.365	0.317
F5-TTS-seg	0.454	40.909	0.893	2.535	–	–
SparkTTS-seg	0.876	24.312	0.885	2.304	–	–
CosyVoice2-seg	1.254	19.770	0.894	2.311	–	–
Ours	0.642	7.650	0.890	2.681	0.487	0.436

As shown in Table 9, our method achieves the lowest DNSV-Pro and the highest AutoPCP, ρ_I, and ΔEmo2v. Compared with the strongest baseline for each metric, it reduces DNSV-Pro by 3.96 and improves ρ_I by 0.122. These results suggest that intensity adjustment within selected spans is compatible with smooth prosodic transitions, enabling stronger and more consistent emotional progression without relying on segment-wise concatenation.

A.2.2 Word-Level + Composable Emotion Control

Table 10: Joint word-level and composable emotion control performance.

Method	WER (%)↓	DNSV-Pro↓	S-SIM↑	E-SIM↑	ρ_W ↑	H-Rate↑
GPT-4o-mini-TTS	1.044	18.406	0.910	0.628	0.085	0.446
DoubaoTTS	0.916	11.769	0.902	0.722	–	–
Qwen3TTS	1.045	9.866	0.851	0.643	0.023	0.301
IndexTTS2-seg	0.477	10.134	0.885	0.757	0.308	0.422
F5-TTS-seg	0.573	35.560	0.864	0.607	0.149	0.403
SparkTTS-seg	1.107	17.916	0.899	0.699	0.034	0.439
CosyVoice2-seg	0.776	16.971	0.848	0.654	0.102	0.414
Ours	0.323	9.641	0.875	0.837	0.234	0.517

Table 10 shows that our method achieves the highest E-SIM and H-Rate, alongside the lowest DNSV-Pro and WER. It improves E-SIM by 0.080 and H-Rate by 7.1 percentage points over the strongest respective baselines. However, its ρ_W remains below IndexTTS2-seg, indicating that higher combination success does not necessarily imply more accurate emotion-weight ranking. Overall, these results support effective emotion composition in the word-level setting with smooth transitions, while leaving room to improve the relative balance among emotions.

Compositional fidelity and localization answer different questions: a mixture may be expressed accurately but extend beyond the requested span. Both aspects are therefore needed to interpret this setting.

A.2.3 INTENSITY-AWARE + COMPOSABLE EMOTION CONTROL

Table 11: Joint intensity-aware and composable emotion control performance at the utterance level.

Method	WER (%)↓	S-SIM↑	E-SIM↑	ρ_I ↑	ΔEmo2v.↑	ρ_W ↑	H-Rate↑
GPT-4o-mini-TTS	1.050	0.935	0.719	0.386	0.204	0.223	0.207
DoubaoTTS	1.371	0.860	0.739	0.200	0.219	–	–
Qwen3TTS	1.945	0.908	0.625	0.079	0.096	0.044	0.045
IndexTTS2	1.225	0.884	0.721	0.393	0.283	0.451	0.259
SparkTTS	0.933	0.857	0.697	–	–	0.177	0.285
F5-TTS	0.354	0.927	0.586	–	–	0.117	0.139
Ours	0.598	0.936	0.791	0.502	0.447	0.504	0.383

As shown in Table 11, our method leads all evaluated control metrics when intensity adjustment and emotion composition are applied jointly. Compared with IndexTTS2, the strongest baseline on both correlation metrics, it improves ρ_I by 0.109 and ρ_W by 0.053. Together with the highest E-SIM and H-Rate, these results suggest that increasing emotional intensity remains compatible with realizing the requested mixture. The highest S-SIM further supports speaker preservation, although WER remains higher than that of F5-TTS.

The main question is whether increasing control strength changes the emotional mixture as intended without introducing disproportionate shifts among its components or degrading speech quality.

A.2.4 CONTROL PERFORMANCE ON COSYVOICE2

Since SAI is primarily designed for NAR generation, our evaluation on CosyVoice2 focuses on ECM's intensity-aware and compositional emotion control capabilities.

Table 12: Emotion and intensity control performance on CosyVoice2.

Method	Target prob.↑	ρ_I ↑	ΔTarget prob.↑	WER (%)↓	S-SIM↑
CosyVoice2 (ECM off)	0.02	–	–	10.4	0.90
CosyVoice2 + ECM (per-emotion)	0.38	0.41	0.36	19.7	0.88
CosyVoice2 + ECM (joint training)	0.52	0.53	0.50	16.8	0.89

Adaptation to autoregressive generation. To examine the applicability of Emotion Concept Modulation (ECM) beyond NAR synthesis, we adapt ECM to the autoregressive decoder of CosyVoice2. In this implementation, ECM introduces emotion-conditioned additive residuals into the final eight decoder layers:

$$h'_\ell = h_\ell + \alpha\, \delta_\ell(e), \tag{6}$$

where $\delta_\ell(e)$ denotes the learned offset for emotion e and α controls the intervention strength. This implementation adds approximately 0.47M parameters, corresponding to 0.07% of the base model. Unlike the scaling-and-shifting formulation used in F5-TTS, the CosyVoice2 adaptation uses additive modulation and evaluates utterance-level emotion and intensity control.

Evaluation. We compare ECM-disabled synthesis with separately trained emotion modulators and a jointly trained four-emotion configuration. Target prob. measures the predicted probability of the requested emotion, ρ_I measures its Spearman correlation with intervention strength, and ΔTarget prob. measures the probability increase from zero intervention to the upper endpoint of a predefined strength range. WER and S-SIM assess intelligibility and speaker preservation. Scores are aggregated over angry, happy, sad, and surprise under matched evaluation conditions.

Preliminary observations. The available strength sweeps indicate that ECM can increase target-emotion responses in CosyVoice2, although its effectiveness depends on the emotion and training configuration. Joint training yields a stronger peak response for surprise, while sad remains weak and higher intervention strengths can impair intelligibility. These observations provide preliminary

evidence for adapting ECM to autoregressive decoding; word-level localization and emotion composition remain to be evaluated in this adaptation.

A.2.5 CONTROL CASES STUDY

Figure 7 compares baseline (light gray) and edited (darker) energy contours. Shaded intervals mark the requested spans.

A: Angry on *give*. The intervention changes the onset and local contour of *give*, with an in-span difference of 11.1 dB versus 1.5 dB outside. The larger local response, together with the close surrounding contours, illustrates effective concentration of the edit on the requested word.

B: Angry on *say*. The selected word changes by 3.7 dB while the out-of-span difference is only 0.5 dB. The near-overlap elsewhere demonstrates precise selectivity even for a short target. The contour is reshaped locally; the request does not simply raise energy throughout the utterance.

C: Sad on *crying*. The edited contour has a lower peak within *crying*. Its 7.7 dB in-span difference exceeds the 3.1 dB outside difference, showing a localized attenuation while retaining the broad surrounding contour. This case also shows that some contextual change remains.

D: Distinct emotions on separate spans. Angry on *give* and Sad on *cry harder* produce differences of 10.4 and 10.5 dB in the respective regions, compared with 2.6 dB outside. The localized response on *give* and attenuation on *cry harder* illustrate how the multi-request formulation accommodates different conditions within one utterance. These plots support acoustic localization; the emotion metrics and listening evaluations assess emotional realization.

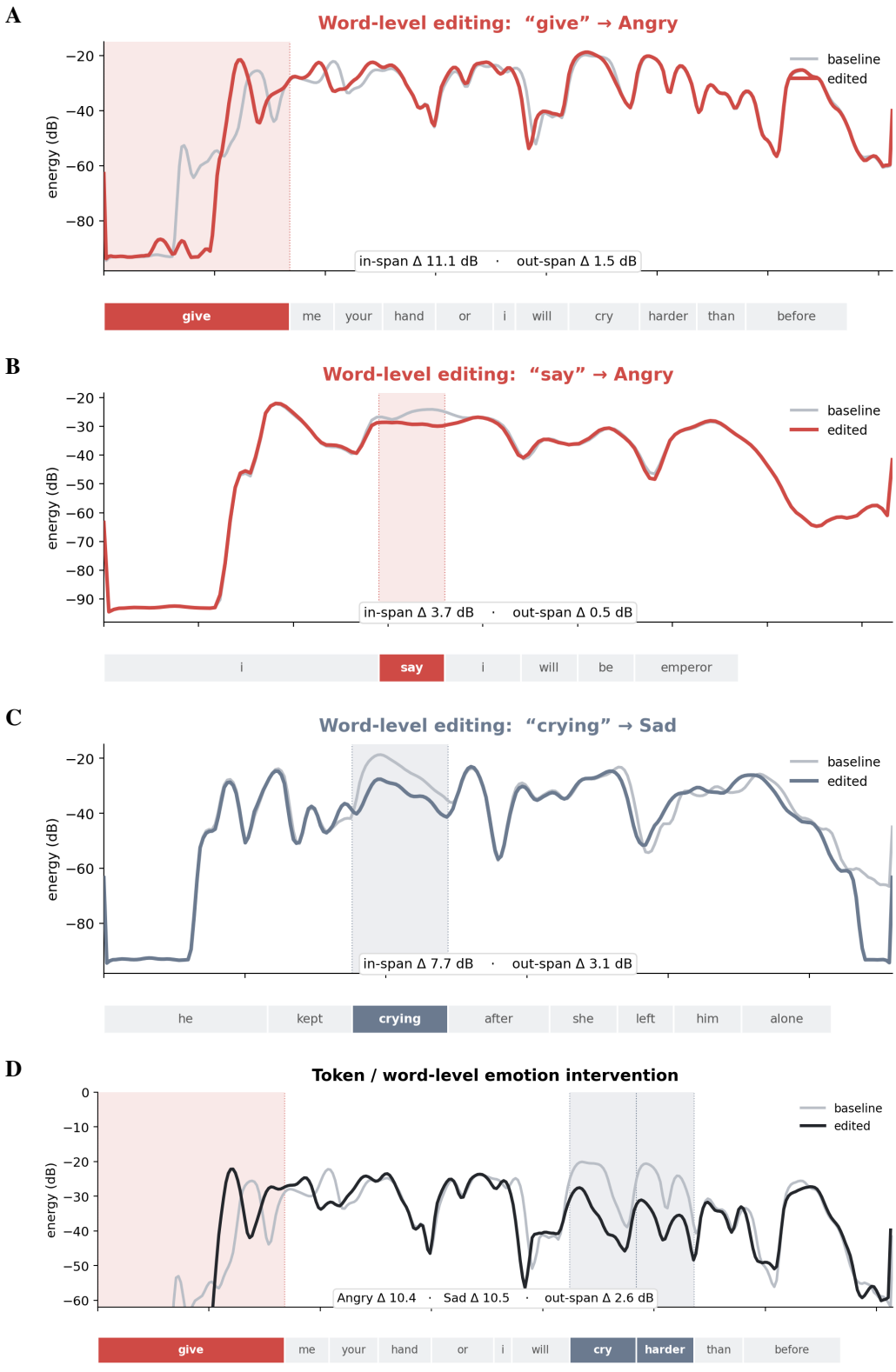


Figure 7: Word-level control cases. **A**: *give*→Angry. **B**: *say*→Angry. **C**: *crying*→Sad. **D**: *give*→Angry and *cry harder*→Sad in the same utterance. Colored word boxes and shaded intervals identify the requested spans. Curves compare baseline and edited energy (dB); annotations report differences inside and outside the target spans.

A.3 Dataset Construction

We construct the emotional speech corpus by combining ESD (Zhou et al., 2021), RAVDESS (Livingstone & Russo, 2018), MSP-Podcast (Busso et al., 2026), and TESS (Pichora-Fuller & Dupuis, 2020), following the emotion filtering and sample selection procedures of Wang et al. (2026) and Xie et al. (2025b). The corpus covers six basic emotions (anger, happiness, sadness, disgust, surprise, and fear) and a neutral category. Where feasible, we prioritize within-speaker comparisons for each target emotion to mitigate speaker-related confounding. The construction procedure consists of the following steps:

1. We enumerate candidate utterances from the source datasets, recover their transcripts and speaker identities, and normalize emotion annotations to the seven target categories. Speaker-matched neutral references are associated with emotional utterances where available.
2. We use `librosa` to extract acoustic statistics, including duration, root-mean-square (RMS) energy, peak amplitude, zero-crossing rate, spectral centroid, and, when available, fundamental frequency (F0). These statistics support quality-aware and emotion-specific ranking. Acoustically atypical samples are penalized rather than automatically discarded by rigid global thresholds.
3. We apply an utterance-level emotion2vec model to obtain target-emotion and neutral scores, the prediction margin, and the top predicted label. Emotion-specific criteria serve as soft rejection signals, with neutral-label dominance additionally required for the neutral category.
4. Candidates are ranked using acoustic statistics and, optionally, emotion2vec scores. Dataset- and emotion-specific quotas, together with speaker balancing, limit overrepresentation by individual speakers or source corpora. Shortlisted candidates are collected for manual listening.
5. Each shortlisted utterance is manually evaluated for emotion correctness, expression strength, and audio quality. Samples meeting the predefined perceptual criteria are retained, with human judgment taking precedence when it disagrees with automatic emotion predictions.
6. We construct a unified manifest containing the dataset source, speaker identity, transcript, audio path, emotion label, optional speaker-matched neutral reference, and one-hot emotion vector. Retained samples are partitioned into training, validation, and test subsets.

The resulting corpus contains approximately 1,000 samples per category and is split into training, validation, and test sets at a ratio of 0.5/0.2/0.3. CREMA-D (Cao et al., 2014) is excluded from this corpus and used exclusively as an additional out-of-distribution (OOD) test set to assess generalization.

A.4 Implementation Details

This section describes the word-to-acoustic alignment procedure used for F5-TTS word-level control, including temporal anchor initialization and the two-pass refinement that compensates for modulation-induced timing drift.

A.4.1 Word-to-Acoustic Alignment

Word-level control requires mapping the requested text span to acoustic frames. For the F5-TTS and Matcha-TTS word-level evaluation protocol, we first synthesize an unsteered reference utterance x^0 and obtain word-level timestamps using Whisper tiny. Let $[t_i^s, t_i^e]$ denote the estimated time interval of word i in the generated utterance, excluding the reference-speech prefix. For a target span $S = [a, b]$, its acoustic-frame interval is

$$[u_S, v_S) = [T_{\text{ref}} + \mathcal{Q}(f_{\text{mel}} t_a^s),\ T_{\text{ref}} + \mathcal{Q}(f_{\text{mel}} t_b^e))\,, \tag{7}$$

where f_{mel} is the acoustic frame rate, T_{ref} is the number of reference-prefix frames, and $\mathcal{Q}$ discretizes frame coordinates into integer indices. The resulting interval defines the temporal anchor used by SAI to localize ECM modulation. This procedure uses an existing ASR model and requires no additional alignment-model training.

Emotional modulation can change local speaking rate and word timing, causing the initial anchor to become misaligned with the generated speech. We therefore use a two-pass refinement procedure, referred to as *iterate-2*. The first controlled pass uses the anchor estimated from x^0 to generate $x^{(1)}$. We then re-estimate the target word boundaries from this output and update the anchor:

$$m^{(1)} = \mathcal{M}\Big(\mathcal{A}(x^{(1)}, S)\Big)\,, \tag{8}$$

where $\mathcal{A}$ estimates the word boundaries of span S in the generated speech, and $\mathcal{M}$ converts these boundaries into a temporal anchor using Eq. 7. A second controlled pass uses the updated anchor to produce the final output, with the text, speaker conditioning, and control request unchanged. The anchor remains fixed within each flow-matching generation pass and is updated only between passes. This refinement addresses modulation-induced timing drift, while contextual interactions within the DiT allow acoustic transitions around the anchored region to evolve jointly.