Rectifying the Emotional Flow: Aligning Priors and Dynamic Guidance for High-Arousal Text-to-Speech

Fangming Feng[†] , **Dongjie Fu**[†] , **Zequn Xie** , **Yu Zhang** ,
Yangyang Wu , **Zhou Zhao** , **Tao Jin**[*]
Zhejiang University
{fangmingfeng, jint_zju}@zju.edu.cn

Abstract

While diffusion and flow-matching models have advanced TTS, generating high-arousal emotions remains a persistent challenge due to the trade-off between stability and expressiveness. Existing systems often suffer from linguistic collapse when pursuing high intensity or fail to meet target emotional levels under stable settings. In this work, we identify that standard Gaussian initialization inevitably introduces a neutral prosody bias, while uniform Classifier-Free Guidance often distorts the acoustic manifold, leading to artifacts. To address this, we propose an inference framework that rectifies the emotional trajectory. An Emotion-Rectified Noise Prior injects a semantic gradient at initialization to align sampling with the target emotional manifold, and Likelihood-Inverse Guidance adaptively schedules guidance via a conditional/unconditional likelihood ratio, strengthening guidance only when the trajectory drifts toward a neutral fallback. Extensive experiments demonstrate that our method effectively resolves the stability bottleneck in high-intensity scenarios, achieving superior linguistic accuracy and emotional fidelity without model retraining. Code is available at https://github.com/MM-Speech/emo-tts.

1 Introduction

While text to speech models deliver near-human naturalness for reading-style speech, they face a persistent stability-expressiveness bottleneck (Mehta et al., 2024; Zhou et al., 2025). Current systems excel at neutral prosody but often collapse linguistically when attempting high-arousal emotions such as screaming or sobbing. In this work, we systematically analyze this trade-off between stability and expressiveness. Specifically, we conducted an empirical evaluation using CosyVoice2 (Du et al., 2024b) on a high-intensity emotion dataset. As visualized in Figure 1, the results reveal a dilemma:

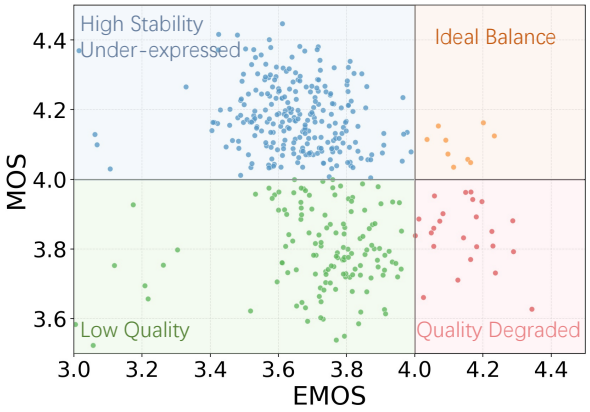

Figure 1: The stability-expressiveness trade-off in high-arousal TTS. We visualize the distribution of samples generated by CosyVoice2 on the High-Intensity Emotion Dataset. Due to the large scale of evaluation data, we utilized Gemini 2.5 Pro as a proxy for human evaluation to compute Emotion MOS (EMOS) and Overall MOS (MOS).

The majority of generated samples cluster in the blue and green regions, exhibiting relatively low Emotion MOS (EMOS), which indicates a failure to meet the target intensity. Furthermore, among the few samples that do achieve high arousal, a significant portion suffers from quality degradation (low MOS). Consequently, samples that achieve an "Ideal Balance" (orange region) are extremely scarce, highlighting the difficulty of maintaining stability while pursuing expressiveness.

To address this dilemma, we begin by revisiting the initialization of diffusion-based generation. As recent studies reveal that initial noise encodes implicit generative bias (Wang et al., 2024a; Yan et al., 2025; Qi et al., 2024; Xie et al., 2025b; Mao et al., 2024), standard isotropic Gaussian initialization inevitably misaligns with the sparse manifolds of high-arousal emotions. Consequently, sampling trajectories readily drift toward local optima associated with neutral prosody. To optimize the initial noise toward the target manifold, we introduce

[*]Corresponding author.
[†]Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21874–21888

Emotion-Rectified Noise Prior (ERNP). Rather than starting from unconstrained noise, ERNP injects a semantic gradient into the initial state. This optimization explicitly steers the trajectory into the basin of attraction of the desired emotional manifold, enabling high-intensity expression while reducing structural artifacts induced by trajectory conflicts.

However, while rectified initialization secures emotional intensity, maintaining pronunciation accuracy under high arousal remains difficult. Traditional Classifier-Free Guidance (CFG) (Ho and Salimans, 2022) relies on linear extrapolation; however, such global, uniform forcing often distorts the acoustic manifold, leading to artifacts. Moreover, research indicates that guidance is not required at every timestep, as high-level attributes emerge only within specific critical windows (Kynkäänniemi et al., 2024). To address this, we propose Likelihood-Inverse Guidance (LIG). Modeling emotional generation as disentangling the target distribution from neutral interference, we derive a state-dependent schedule based on the likelihood ratio between conditional and unconditional predictions. This establishes a self-regulating mechanism where guidance is amplified only when the trajectory drifts toward a neutral fallback, maximizing expressiveness while preserving acoustic stability.

By synergizing these two strategies, we effectively rectify the emotional flow, ensuring the generative trajectory remains consistently aligned with the high-arousal manifold to resolve the stability-expressiveness dilemma. To summarize, our main contributions are:

- We propose a plug-and-play inference paradigm that resolves the stability-expressiveness trade-off without requiring any model retraining or fine-tuning.
- We introduce Emotion-Rectified Noise Prior to align the initial state with the target emotional manifold , and Likelihood-Inverse Guidance to dynamically regulate the generation path based on real-time drift.
- Extensive experiments demonstrate that our method effectively resolves the stability bottleneck in high-arousal scenarios. We achieve superior linguistic accuracy while maintaining strong emotional intensity, with geometric analysis further confirming the successful rectification of generative trajectories.

2 Related Work

2.1 Emotional Text-to-Speech

Text-to-Speech (TTS) has evolved from stepwise regression to end-to-end generative modeling. Early paradigms, such as two-stage models (e.g., Tacotron (Wang et al., 2017; Shen et al., 2018)) and non-autoregressive frameworks (e.g., FastSpeech series (Ren et al., 2019, 2020)), improved inference speed but remained limited in flexibility. Recently, generative modeling has emerged as the mainstream solution. Among the various generative families, Flow Matching (Lipman et al., 2022) has established itself as a dominant framework, surpassing standard diffusion models in training and inference efficiency through the use of Ordinary Differential Equations (ODEs). Representative methods, such as Matcha-TTS (Mehta et al., 2024), F5-TTS (Chen et al., 2025), and the CosyVoice series (Du et al., 2024a,b, 2025), widely adopt conditional flow matching architectures, effectively transitioning TTS from discrete regression to continuous trajectory generation.

Building on these generative backbones, Emotional TTS (Yang et al., 2025; Diatlova and Shutov, 2023; Xie, 2026; Cho et al., 2025) focuses on empowering synthesis systems with explicit control over affective prosody to render diverse emotional expressions. To enhance this controllability, existing works primarily employ label/embedding-driven methods (Cho et al., 2024; Fu et al., 2025; Huijuan et al., 2023; Zhang et al., 2019; Xie et al., 2026a; Jing et al., 2025), reference-based zero-shot transfer (Anastassiou et al., 2024; Chen et al., 2025; Cheng et al., 2025; Feng et al., 2026; Jiang et al., 2023b; Fu et al., 2026; Jiang et al., 2025), or LLM-integrated natural language prompting (Zhou et al., 2025; Du et al., 2024b; Cao et al., 2026; Du et al., 2025; Xie et al., 2026b; Fu et al., 2024; Yang et al., 2025). However, despite these advancements, a persistent "stability-expressiveness trade-off" remains in high-arousal emotion generation. Excessive pursuit of expressive intensity often leads to pronunciation collapse or acoustic artifacts, making the simultaneous achievement of high fidelity and strong expressiveness an unresolved challenge in current flow-matching frameworks.

2.2 Inference-Time Optimization for Diffusion Models

As the marginal benefits of scaling laws in training stage diminish, inference-time optimization

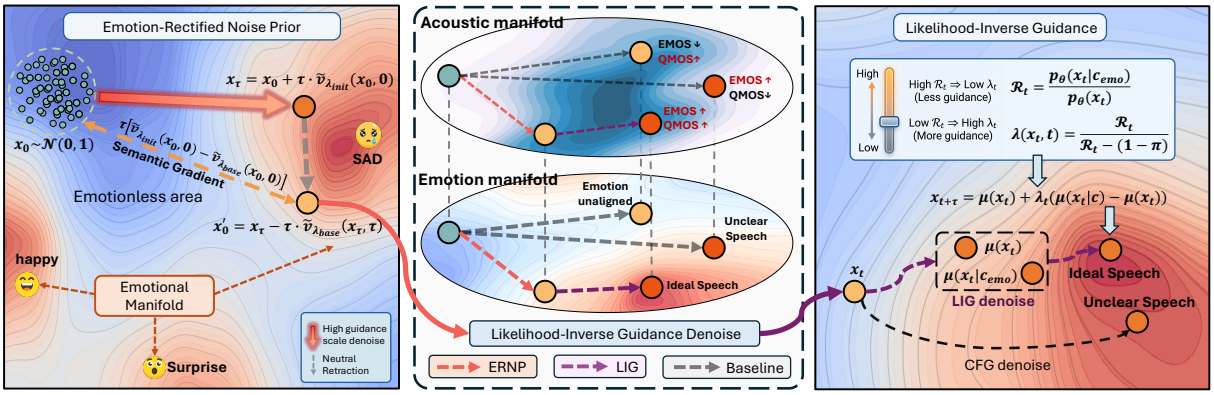

Figure 2: Overview of our method. **Left: ERNP** injects a semantic gradient to align the initial state with the target emotional manifold. **Right: LIG** dynamically regulates the guidance scale λ_t based on the likelihood ratio $\mathcal{R}_t$ to prevent neutral drift. **Middle:** The combined approach rectifies the generative flow (purple line), ensuring high-intensity expressiveness while preserving acoustic stability compared to the baseline (gray dashed line).

has become pivotal for exploiting model potential. In diffusion models, The Silent Prompt (Wang et al., 2024a) reveals that initial noise encodes implicit generative bias, enabling precise control over semantics and details via optimal noise retrieval. One More Step (Hu et al., 2024) addresses the training-inference distribution mismatch caused by non-zero terminal SNR by introducing an auxiliary pre-inference mapping to restore image quality. LI-CFG (Kynkäänniemi et al., 2024) tackles the defects of constant-weight CFG by activating guidance only within specific intervals to balance quality and diversity. However, directly transferring these priors from the image domain to speech emotion tasks is challenging, and most methods fail to resolve the collapse of high-dimensional emotional manifolds. In the speech domain, EmoSteer-TTS (Xie et al., 2025a) introduces training-free activation steering for emotion intervention but exhibits instability at high intensities. These limitations underscore the urgent need for speech-manifold-oriented inference optimization to resolve stability issues in high-expressiveness synthesis without retraining.

3 Method

3.1 Emotion-Rectified Noise Prior

3.1.1 Problem Formulation: Initial Noise as an Informational Prior

In conditional flow matching, synthesis follows an ODE driven by a learned vector field conditioned on inputs.

$$dx_t = v_\theta(x_t, t, c)dt \tag{1}$$

Although inference usually starts from isotropic Gaussian noise, prior work suggests this initial state x_0 is not purely random—it can encode a generative bias that affects the final sample (Qi et al., 2024; Mao et al., 2024). For high-arousal emotions, the target region is a sparse sub-manifold, so a standard Gaussian start tends to carry a neutral prosody bias and makes the solver fight inertia, causing curved trajectories and emotional washout. Therefore, our goal is to rectify the initial noise to explicitly encode the target emotional semantics before generation begins.

3.1.2 Semantic Gradient Injection and Geometric Interpretation

ERNP aims to embed emotion semantics into the initial state before generation. First, we sample a standard anchor from the Gaussian prior. Then, we perform a Lookahead Step, moving forward by a pseudo-step τ using a high guidance scale λ_{init} via Forward Euler:

$$\mathbf{x}_\tau = \mathbf{x}_0 + \tau \cdot \tilde{v}_{\lambda_{init}}(\mathbf{x}_0, 0) \tag{2}$$

Finally, we perform a Calibration Step, moving backward from the advanced position to the origin using a base guidance scale λ_{base}:

$$\mathbf{x}_0^* = \mathbf{x}_\tau - \tau \cdot \tilde{v}_{\lambda_{base}}(\mathbf{x}_\tau, \tau) \tag{3}$$

Substituting the lookahead step into the calibration step yields the relation:

$$\mathbf{x}_0^* = \mathbf{x}_0 + \tau(\tilde{v}_{\lambda_{init}}(\mathbf{x}_0, 0) - \tilde{v}_{\lambda_{base}}(\mathbf{x}_\tau, \tau)) \tag{4}$$

Using flow-matching's tendency toward straight and consistent trajectories, the retraction velocity

at x_τ can be approximated by that at x_0, yielding:

$$\mathbf{x}_0^* \approx \mathbf{x}_0 + \tau \left[\tilde{v}_{\lambda_{init}}(\mathbf{x}_0, 0) - \tilde{v}_{\lambda_{base}}(\mathbf{x}_0, 0)\right] \quad (5)$$

Expanding CFG shows the update is equivalent to adding an "emotional semantic gradient" with strength $\tau(\lambda_{init} - \lambda_{base})$, which geometrically points toward the high-probability manifold of the target emotion. The complete pseudo-code for ERNP is presented in Algorithm 1 in Appendix.

3.2 Likelihood-Inverse Guidance

3.2.1 Additive Mixture Formulation

Standard CFG operates under the implicit assumption that the target conditional distribution can be approximated by linearly extrapolating from a neutral baseline in the log-probability space. However, for expressive emotional speech synthesis, this linear extrapolation often distorts the acoustic manifold as high-arousal emotions are not simple extensions of neutral speech. Instead, we argue that the learned distribution is more accurately modeled as a probabilistic mixture in the density space. This formulation models the generated distribution as an additive mixture of the neutral prosody distribution and the ideal emotional distribution, enabling us to mathematically disentangle and purify the authentic emotional target from neutral interference, rather than merely amplifying feature differences.

We hypothesize that the learned conditional distribution $p_\theta(x_t \mid c_{emo})$ is an additive mixture of a fallback neutral distribution $p(x_t)$ and the ideal emotional distribution $p_{true}(x_t|c_{emo})$[1]:

$$p_\theta(x_t|c_{emo}) = (1-\pi) \cdot p(x_t) + \pi \cdot p_{true}(x_t|c_{emo}) \quad (6)$$

Here, $\pi \in [0, 1]$ is the purity coefficient, representing the proportion of the target distribution that is authentically emotional. In our experiments, we set $\pi = 0.95$ by default. To synthesize high-fidelity expressive speech, our objective is to sample directly from the purified distribution p_{true}. By inverting the mixture equation, we disentangle the target density:

$$p_{true}(x_t|c_{emo}) = \frac{1}{\pi}\left[p_\theta(x_t|c_{emo}) - (1-\pi) \cdot p(x_t)\right] \quad (7)$$

[1]For notational convenience, $p_\theta(x_t)$ denotes conditioning on all non-emotional conditions (e.g., text and speaker), while $p_\theta(x_t \mid c_{emo})$ further conditions on emotion.

3.2.2 Derivation of the Purified Vector Field

In the Flow Matching, the generative process is governed by a time-dependent vector field $v(x)$. Since the score function is $s(x) = \nabla_x \log p(x)$ and the optimal vector field is proportional to it, $v(x) \propto \nabla_x \log p(x)$, we derive the vector field v_{true} by taking the log-gradient of Eq. (7). Using the identity $\nabla \log p = \frac{\nabla p}{p}$ and substituting the velocity counterparts, the target vector field can be expressed as (please refer to the Appendix A.1 for the detailed derivation):

$$v_{\text{true}}(x_t) = \frac{p_\theta(x_t \mid c_{emo})\, v_c - (1-\pi)\, p_\theta(x_t)\, v_u}{p_\theta(x_t \mid c_{emo}) - (1-\pi)\, p_\theta(x_t)} \quad (8)$$

where v_c and v_u denote the velocity outputs conditioned on the full conditioning information including c_{emo}, and on the same conditioning information with emotion excluded, respectively. To align this with the standard guidance structure (base + scale $\times$ difference), we rearrange the terms algebraically to obtain (please refer to the Appendix A.2 for the detailed derivation):

$$v_{true}(x_t) = v_u(x_t) + \lambda(x_t, t) \cdot (v_c(x_t) - v_u(x_t)) \quad (9)$$

Here, $\lambda(x_t, t)$ is a dynamic scaling coefficient derived as:

$$\lambda(x_t, t) = \frac{p_\theta(x_t \mid c_{emo})}{p_\theta(x_t \mid c_{emo}) - (1-\pi)p(x_t)} \quad (10)$$

By defining the Likelihood Ratio R_t as the ratio between the conditional and unconditional densities:

$$R_t = \frac{p_\theta(x_t \mid c_{emo})}{p(x_t)} \quad (11)$$

we can rewrite the guidance scale in a concise analytical form:

$$\lambda(x_t, t) = \frac{R_t}{R_t - (1-\pi)} \quad (12)$$

This constitutes the core mechanism of Likelihood-Inverse Guidance. When $R_t \gg 1$, the trajectory is already in a region where the emotional component has much higher probability mass, so $\lambda \to 1$ and the update is essentially equivalent to directly using the conditional vector field. In contrast, as $R_t \to 1 - \pi$, the conditional distribution nearly degenerates to the neutral fallback term, accordingly, λ grows to strongly correct the trajectory and reinforce emotional expression. However, as $R_t \to 1 - \pi$, the denominator

of $\lambda(x_t, t) = \frac{R_t}{R_t-(1-\pi)}$ approaches 0, causing λ to diverge. For numerical stability, we enforce $R_t \geq \frac{\lambda_{\max}(1-\pi)}{\lambda_{\max}-1}$, which is equivalent to clipping the guidance scale as $\lambda \leftarrow \min(\lambda, \lambda_{\max})$. We set $\lambda_{\max} = 30$ by default.

3.2.3 Recursive Estimation via Transition Dynamics

To track the evolution of the likelihood ratio R_t in real-time, we analyze its discrete increment over a small time step Δt. Defining the logarithmic increment $\Delta \log R_t = \log R_{t+\Delta t} - \log R_t$, we first expand it using the definition of the likelihood ratio:

$$\begin{aligned}\Delta \log R_t = &\log\left(\frac{p_\theta(x_{t+\Delta t}|c_{emo})}{p_\theta(x_{t+\Delta t})}\right) \\ &- \log\left(\frac{p_\theta(x_t|c_{emo})}{p_\theta(x_t)}\right)\end{aligned} \tag{13}$$

By regrouping the terms, we isolate the contributions from the conditional and unconditional components:

$$\begin{aligned}\Delta \log R_t = &\underbrace{[\log p_\theta(x_{t+\Delta t}|c_{emo}) - \log p_\theta(x_t|c_{emo})]}_{\text{Conditional Change}} \\ &- \underbrace{[\log p_\theta(x_{t+\Delta t}) - \log p_\theta(x_t)]}_{\text{Unconditional Change}}\end{aligned} \tag{14}$$

Since we are sampling a specific trajectory, we leverage the Markov property of the generative process. By applying the chain rule to the joint probability of consecutive states, expressed as $p(x_{t+\Delta t}, x_t) = p(x_{t+\Delta t}|x_t) \cdot p(x_t)$, the incremental change in the log-density is principally determined by the transition probability of the current step. This allows us to express the update in terms of the transition kernels (please refer to the Appendix A.2 for the detailed derivation):

$$\Delta \log R_t = \log p(x_{t+\Delta t}|x_t, c_{emo}) - \log p(x_{t+\Delta t}|x_t) \tag{15}$$

In Flow Matching, the local transition kernel $p(x_{t+\Delta t}|x_t)$ is modeled as a Gaussian distribution $\mathcal{N}(x_{t+\Delta t}; \mu, \sigma^2\mathbf{I})$. Consequently, the log-probability is proportional to the negative squared distance from the mean. Let μ_c and μ_u denote the predicted means for the conditional and unconditional models, respectively:

$$\Delta \log R_t = \frac{1}{2\sigma^2}\left(\|x_{t+\Delta t} - \mu_u\|^2 - \|x_{t+\Delta t} - \mu_c\|^2\right) \tag{16}$$

Under the Euler solver, the next state is updated as $x_{t+\Delta t} = x_t + v_t \Delta t$, where $v_t = v_u + \lambda_t (v_c - v_u)$. Meanwhile, the predicted means are $\mu_c = x_t + v_c \Delta t$ and $\mu_u = x_t + v_u \Delta t$. Substituting these into the distance terms leads to the cancellation of the current state x_t, transforming the geometric distances into velocity differences:

$$\begin{aligned}x_{t+\Delta t} - \mu_c &= (v_t - v_c)\Delta t \\ x_{t+\Delta t} - \mu_u &= (v_t - v_u)\Delta t\end{aligned} \tag{17}$$

Finally, substituting these relations back into the update rule yields:

$$\Delta \log R_t = \frac{\Delta t^2}{2\sigma^2}\left(\|v_t - v_u\|^2 - \|v_t - v_c\|^2\right) \tag{18}$$

where $\sigma = 1 - t$ represents the noise level. This derivation confirms that the update of R_t is rigorously driven by the instantaneous geometric divergence between the vector fields. The complete pseudo-code for LIG is presented in Algorithm 2 in Appendix.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We use three public benchmarks: ESD (Zhou et al., 2021), EmoVoice (Yang et al., 2025), and Expresso (Nguyen et al., 2023). Neutral speech from the target speaker is consistently used as reference audio. Since ESD does not provide an official test split, we follow the setting used by MaskGCT (Wang et al., 2024b) and MegaTTS2 (Jiang et al., 2023a), constructing an evaluation set from a subset of the dataset. Specifically, we construct our evaluation set by taking the first 100 prompts from each of speakers 0001 and 0011. For Expresso, we only use the read-speech subset for evaluation.

To evaluate high-arousal scenarios, we construct the High-Intensity Emotion Dataset (HIED), a curated subset filtered from several publicly available open-source academic speech corpora. HIED consists of 400 extreme-arousal utterances across four categories: Angry, Happy, Sad, and Surprise (100 samples per category). The dataset encompasses 37 distinct speakers, including 295 male and 105 female utterances. To ensure objectivity, we utilize a pre-trained dimensional arousal predictor (Wagner et al., 2022) to score and rank candidates, selecting only the top-tier utterances for each category. For each speaker, we retain at least two samples: one as a high-arousal reference and another for the target text with ground-truth audio.

Dataset	Model	Method	WER (↓)	ES (↑)	Recall (↑)	UTMOSv2 (↑)	GMOS (↑)
ESD	CosyVoice2	Base	2.56	0.74	**39.3%**	2.95	4.02
		+Ours	**2.05**	**0.81**	39.2%	**3.27**	**4.02**
	Index-TTS2	Base	**1.71**	0.76	**48.3%**	**3.47**	4.10
		+Ours	1.90	**0.79**	47.5%	3.33	**4.13**
EmoVoice	CosyVoice2	Base	3.41	**0.77**	**42.5%**	3.28	3.88
		+Ours	**3.35**	0.75	41.8%	**3.56**	**3.91**
	Index-TTS2	Base	2.97	0.74	46.3%	3.35	4.04
		+Ours	**2.13**	**0.82**	**48.5%**	**3.98**	**4.22**
Expresso	CosyVoice2	Base	3.13	0.72	39.7%	**3.71**	3.96
		+Ours	**2.27**	**0.83**	**43.1%**	3.61	**4.05**
	Index-TTS2	Base	2.70	0.75	**45.2%**	3.02	4.02
		+Ours	**1.95**	**0.84**	44.7%	**3.26**	**4.13**

Table 1: Overall performance comparison on standard benchmarks. "Base" denotes the original model, and "+Ours" denotes the model integrated with ERNP and LIG.

Model	Method	SMOS	PMOS	QMOS	EMOS
CosyVoice2	Base	3.85	3.62	3.79	3.53
	+Ours	**3.91**	**3.87**	**3.90**	**3.82**
IndexTTS2	Base	**3.92**	3.78	3.95	3.94
	+Ours	3.77	**3.79**	**4.02**	**4.18**

Table 2: MOS on the HIED subset (50 samples).

4.1.2 Model and Implementation Details

We selected three representative emotion-controllable models as baselines: Index-TTS2 (Zhou et al., 2025), CosyVoice2 (Du et al., 2024b), and F5-TTS (Chen et al., 2025). For CosyVoice2, where emotion and text features are fused at the LLM stage, we designed a dual-branch strategy to enable LIG. Specifically, by feeding the input containing the emotion control prompt and the input without it into the LLM separately, we obtain two distinct outputs. These outputs serve as the emotion-guided and unconditional conditions for LIG, respectively. We then fix the random seed and feed these conditions into the diffusion model for generation.

4.1.3 Evaluation Metrics

We employ a combination of objective and subjective metrics:

Objective Metrics: Speech intelligibility is evaluated using word error rate (WER), with FunASR (Gao et al., 2023) for Chinese content and Whisper (Radford et al., 2023) for English. For emotional evaluation, we utilize emotion2vec (Ma et al., 2024) to extract representations and compute the Cosine Similarity (ES) and Recall between the generated audio and the target emotion. Additionally, UTMOSv2 (Baba et al., 2024) is used as a proxy metric for naturalness.

Subjective Metrics: We adopt a multi-dimensional MOS framework, including Similarity (SMOS), Prosody (PMOS), Quality (QMOS), and Emotion (EMOS). Given the high cost of human evaluation, we additionally utilized Gemini 2.5 Pro (Comanici et al., 2025) to conduct automatic MOS scoring (GMOS). This approach has been proven to exhibit a high correlation with human judgment(Manku et al., 2025). For human evaluation, 50 samples were selected from the generation results on HIED for blind testing.

4.2 Comparative Analysis

4.2.1 Overall Performance

We evaluate on ESD, EmoVoice, and Expresso with CosyVoice2 and IndexTTS2. As presented in Table 1, our method demonstrates consistent improvements in both emotional alignment and linguistic stability across most scenarios. In EmoVoice and Expresso, our method yields substantial gains. For instance, IndexTTS2 on the EmoVoice dataset achieves a significant reduction in WER (2.97 → 2.13) alongside a marked increase in Emotion Similarity (0.74 → 0.82) and Recall. Similarly, on the Expresso dataset, CosyVoice2 sees a drastic WER improvement (3.13 → 2.27) and enhanced expressiveness (ES 0.72 → 0.83). It is worth noting that

for the IndexTTS2 model on ESD dataset, the improvement in objective metrics is marginal (e.g., slight fluctuations in WER). This is primarily because ESD constitutes a portion of IndexTTS2's training data. Consequently, the base model is already highly optimized for this specific distribution, leaving limited headroom for inference-time optimization. Nevertheless, even in this well-fitted scenario, our method still enhances Emotion Similarity (0.76 $\rightarrow$ 0.79) and achieves higher subjective preference (GMOS).

Furthermore, we extracted 50 samples specifically from HIED for human evaluation using CosyVoice2 and IndexTTS2. The results, presented in Table 2, demonstrate that our method significantly improves emotional expressiveness without compromising, and often improving, speech quality.

Furthermore, we extracted 50 samples specifically from the HIED dataset for human evaluation using CosyVoice2 and IndexTTS2. The results, presented in Table 2, demonstrate that our method significantly enhances emotional expressiveness while maintaining overall speech quality. Specifically, CosyVoice2 achieves a substantial gain in EMOS, rising from 3.53 to 3.82, alongside consistent improvements in prosody and audio quality. Similarly, IndexTTS2 attains an EMOS of 4.18 with our method, surpassing the baseline's 3.94.

4.2.2 Performance on Pure Flow-matching Model

To verify the effectiveness of our method under intense emotional conditions, we employed F5-TTS, a pure non-autoregressive flow-matching model, on the HIED dataset. Given that HIED consists of curated high-arousal audio and F5-TTS employs a reference-based mechanism to directly clone prosody and style, this setup allows us to explicitly modulate the emotional intensity of the output by controlling the acoustic properties of the reference prompt.

We conducted a comprehensive evaluation on HIED. As shown in Table 3, our method significantly reduces the WER from 4.41% to 2.53%, effectively mitigating the phoneme collapse observed in the baseline. Beyond the objective metrics, human subjective evaluation (as shown in Table 4) shows that the EMOS significantly improves from 3.63 to 3.89.

Furthermore, to verify the efficacy of our method independent of LLM-generated semantic guidance, we extended the objective evaluation of F5-TTS to standard benchmarks (ESD, EmoVoice, Expresso). Even when using emotional reference audio for zero-shot transfer (where emotion is entangled with speaker timbre), our ERNP+LIG framework consistently boosts Emotion Similarity (ES) and reduces WER across all datasets (detailed in Table 5). This conclusively demonstrates that our method fundamentally improves the generative flow dynamics of the decoder itself.

Method	WER (↓)	ES (↑)	Recall (↑) (↑)	GMOS (↑)
F5-TTS (Base)	4.41	**0.78**	52.1%	3.99
F5-TTS (+Ours)	**2.53**	0.73	**61.7%**	**4.12**

Table 3: Performance of F5-TTS on the HIED.

F5-TTS	**SMOS** (↑)	**PMOS** (↑)	**QMOS** (↑)	**EMOS** (↑)
Baseline	3.32	3.08	3.34	3.63
+Ours	**3.33**	**3.10**	**3.41**	**3.89**

Table 4: Comprehensive Evaluation of F5-TTS on the HIED Benchmark. Subjective scores are reported with 95% Confidence Intervals.

Dataset	**Method**	**WER** (↓)	**ES** (↑)	**UTMOSv2** (↑)
ESD	Baseline	3.62	0.78	3.16
	+Ours	**2.38**	**0.82**	**3.19**
EmoVoice	Baseline	3.91	0.77	3.47
	+Ours	**2.67**	**0.83**	**3.65**
Expresso	Baseline	3.79	0.72	3.41
	+Ours	**2.91**	**0.79**	**3.58**

Table 5: Extended Objective Evaluation of F5-TTS across multiple datasets.

4.3 Ablation Studies

In this section, we conduct extensive ablation studies to validate the effectiveness of each proposed component. All experiments in this section are performed using IndexTTS2 on HIED.

4.3.1 Emotion-Rectified Noise Prior (ERNP)

To validate our initialization strategy, we compared generation quality with and without Rectified Noise (Table 6). Removing it causes significant degradation in emotional metrics: ES drops from 0.84 to 0.73, and Recall decreases from 54.2% to 51.5%, suggesting the model struggles to enter the target emotional mode without a semantic prior. This is confirmed by t-SNE visualization (Figure 3), where Rectified Noise forms distinct emotion-specific

Initialization	WER (↓)	ES (↑)	Recall (↑)
Standard Gaussian Noise	3.25	0.73	51.5%
Rectified Noise (Ours)	**3.05**	**0.84**	**54.2%**

Table 6: Comparison between Standard Gaussian Noise and our proposed ERNP.

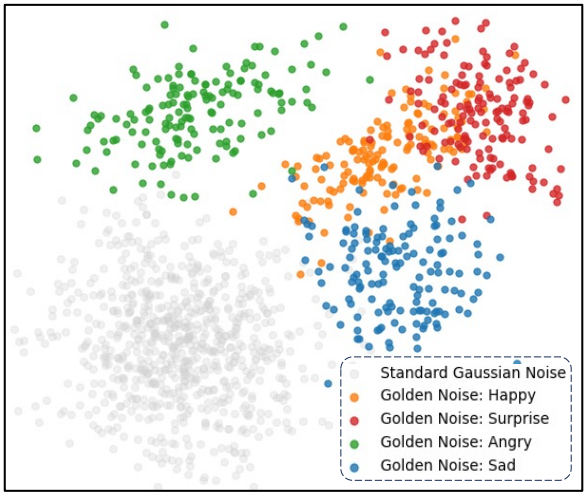

Figure 3: t-SNE visualization of Standard Gaussian Noise vs. Rectified Noise.

clusters, unlike the random distribution of standard Gaussian noise.

To understand the physical impact, we analyzed the guidance coefficient λ_t (Figure 4). With standard initialization, the model detects a large initial disparity, triggering a sharp spike ($\lambda_t > 18$) at $t \approx 0.2$. This forceful intervention causes trajectory overshoot and instability. In contrast, ERNP provides an effective warm-start, constraining the initial λ_t to a stable range (peaking $\approx 7-8$) and ensuring smooth convergence to the target manifold.

4.3.2 Likelihood-Inverse Guidance (LIG)

We compared LIG against standard Constant CFG and Limited-Interval CFG (LI-CFG) (Table 7). Constant CFG achieves high emotional similarity but suffers from high WER (4.08%) due to signal over-saturation. While LI-CFG improves stability, it fails to reach peak intensity. Our LIG strikes the optimal balance, achieving the lowest WER (3.05%) and highest Emotion Similarity (0.84).

The instability of Constant CFG arises because the Rectified Noise already provides significant emotional variance. Indiscriminate high-scale guidance forces excessive amplification, causing emotional features to interfere with phonation and producing artifacts. LIG acts as a dynamic regulator,

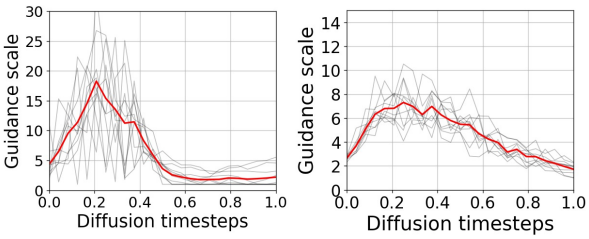

Figure 4: Curves of guidance coefficient λ_t over time. Left: Standard Gaussian Noise (unstable spikes); Right: Emotion-Rectified Noise Prior (stable trajectory).

Guidance Schedule	WER (↓)	ES (↑)	GMOS (↑)
Constant CFG	4.08	0.79	3.65
LI-CFG	3.86	0.74	3.91
LIG (Ours)	**3.05**	**0.84**	**4.22**

Table 7: Ablation study on Guidance Schedules. Comparison of Constant CFG, LI-CFG, and LIG.

relaxing guidance once the emotional trajectory is established to preserve speech quality without sacrificing expressiveness.

4.4 Hyperparameter Sensitivity

To determine the optimal configuration, we conducted sensitivity analyses on the mixture purity coefficient π and guidance scale cap λ_{max} (as shown in Table 8,). For π, which controls emotional disentanglement, varying it from 0.9 to 0.99 revealed that low values yield insufficient separation, while approaching 1.0 causes numerical instability (spiking WER to 3.82%). Thus, $\pi = 0.95$ was selected to maximize expressiveness (ES=0.84) while maintaining stability. Regarding λ_{max}, used to prevent gradient explosion, tests ranging from 10 to 50 indicated that a low cap (e.g., 10) fails to correct neutral drift (WER 4.12%), while an excessive cap (50) degrades audio quality, leading us to set $\lambda_{max} = 30$.

4.5 Mechanism Analysis: Why it works?

To uncover the physical mechanism behind the performance gains, we analyze the geometric properties of the generated ODE trajectories. Our core finding is that the proposed method significantly smoothes the denoising path, minimizing gradient conflicts. We quantify this improvement using two complementary metrics:

1. Directional Stability (CAD) We compute the Cumulative Angular Deviation to measure local directional consistency. It aggregates the rotational

Purity Coefficient (π)			Max Guidance Scale (λ_{max})		
π	WER (↓)	ES (↑)	λ_{max}	WER (↓)	ES (↑)
0.90	3.08%	0.81	10	4.12%	0.78
0.95	**3.05%**	**0.84**	**30**	**3.05%**	**0.84**
0.99	3.82%	0.84	50	3.48%	0.83

Table 8: Hyperparameter sensitivity analysis.

Method	CAD (↓)	$\log(\mathcal{S}(Z))$ (↓)
Baseline	34.26	2.28
Ours (ERNP + LIG)	**15.83**	**1.04**

Table 9: Quantitative analysis of geometric properties.

magnitude between consecutive velocity vectors:

$$\text{CAD} = \sum_{i=1}^{N-1} \arccos\left(\frac{v_i \cdot v_{i+1}}{\|v_i\|\|v_{i+1}\|}\right) \quad (19)$$

A lower CAD indicates that the solver follows a decisive path with minimal "zigzag" oscillations, directly correlating with reduced acoustic tremors and artifacts.

2. Global Linearity ($\mathcal{S}$) We adopt the straightness metric from InstaFlow (Liu et al., 2024) to measure the deviation from an ideal uniform linear motion:

$$\mathcal{S}(Z) = \int_0^1 \|v(x_t, t) - (x_1 - x_0)\|^2 dt \quad (20)$$

Physically, $\mathcal{S} \to 0$ implies the trajectory avoids complex non-linear "detours," simplifying the mapping from noise to speech into a near-perfect linear interpolation.

Result Analysis As shown in Table 9, our method drastically reduces both CAD ($34.26 \to 15.83$) and $\log \mathcal{S}$ ($2.28 \to 1.04$). This geometric rectification minimizes discretization errors, ensuring robustness even under limited inference steps. Consequently, at the extreme setting of NFE=5 (see Table 10), the baseline suffers severe collapse (WER 15.29%), whereas our method maintains high stability (WER 7.98%).

5 Conclusion

In this work, we address the expressiveness-stability dilemma in high-arousal TTS with a training-free framework. By combining ERNP for aligned initialization and LIG for dynamic inference regulation, we rectify the emotional generative flow. Experiments show our method mitigates acoustic instability and linearizes trajectories, achieving superior expressiveness and fidelity even in low-step settings. As a plug-and-play solution, it unlocks existing diffusion models for human-level emotional synthesis without retraining.

NFE	Method	WER (↓)	ES (↑)	Recall (↑)
25	Baseline	4.36	0.79	53.1%
	+Ours	**3.05**	**0.84**	**54.2%**
10	Baseline	7.48	0.73	**58.8%**
	+Ours	**4.31**	**0.78**	49.2%
5	Baseline	15.29	**0.74**	**55.7%**
	+Ours	**7.98**	0.71	49.5%

Table 10: Performance comparison under different NFE. Our method maintains high stability even with significantly reduced inference steps.

6 Limitations

The primary limitation of our framework is its inherent reliance on the continuous probability flows and iterative nature of diffusion models. Since our mechanisms (ERNP and LIG) are derived from ODE properties, they are not directly applicable to non-iterative architectures, such as discrete Autoregressive models or single-step feed-forward networks, which lack the continuous latent trajectories required for dynamic flow rectification.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. U25B2064, the "Pioneer" and "Leading Goose" R&D Program of Zhejiang under Grant No. 2025C02110, the Public Welfare Research Program of Ningbo under Grant No. 2024S062, and the Yongjiang Talent Project of Ningbo under Grant No. 2024A-161-G.

References

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*.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. 2024. The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*.

Di Cao, Dongjie Fu, Hai Yu, Siqi Zheng, Xu Tan, and Tao Jin. 2026. X-opd: Cross-modal on-policy distillation for capability alignment in speech llms. *Preprint*, arXiv:2603.24596.

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

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. 2025. AHabench: Benchmarking audio hallucinations in large audio-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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. *arXiv preprint arXiv:2406.07803*.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Daria Diatlova and Vitaly Shutov. 2023. Emospeech: Guiding fastspeech2 towards emotional text to speech. *arXiv preprint arXiv:2307.00024*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

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. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Fangming Feng, Sihang Cai, Zequn Xie, Yangyang Wu, and Tao Jin. 2026. Scene-aware spatiotemporal generalization: Towards robust temporal action detection across domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 3903–3911.

Dongjie Fu, Xize Cheng, Linjun Li, Xiaoda Yang, Lujia Yang, and Tao Jin. 2025. PACHAT: Persona-aware speech assistant for multi-party dialogue. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29313–29330, Suzhou, China. Association for Computational Linguistics.

Dongjie Fu, Xize Cheng, Xiaoda Yang, Wang Hanting, Zhou Zhao, and Tao Jin. 2024. Boosting speech recognition robustness to modality-distortion with contrast-augmented prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 3838–3847, New York, NY, USA. Association for Computing Machinery.

Dongjie Fu, Fangming Feng, Xize Cheng, Linjun Li, Zhou Zhao, and Tao Jin. 2026. Character beyond speech: Leveraging role-playing evaluation in audio large language models via reinforcement learning. *Preprint*, arXiv:2604.13804.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *Preprint*, arXiv:2207.12598.

Minghui Hu, Jianbin Zheng, Chuanxia Zheng, Chaoyue Wang, Dacheng Tao, and Tat-Jen Cham. 2024. One more step: A versatile plug-and-play module for rectifying diffusion schedule flaws and enhancing low-frequency controls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7340.

ZHAO Huijuan, YE Ning, and WANG Ruchuan. 2023. Improved cross-corpus speech emotion recognition using deep local domain adaptation. *Chinese Journal of Electronics*, 32(3):640–646.

Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Zhenhui Ye, Shengpeng Ji, Qian Yang, Chen Zhang, Pengfei Wei, Chunfeng Wang, and 1 others. 2023a. Megatts 2: Boosting prompting mechanisms for zero-shot speech synthesis. *arXiv preprint arXiv:2307.07218*.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, and 1 others. 2025. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. *arXiv preprint arXiv:2502.18924*.

Ziyue Jiang, Yi Ren, Zhenhui Ye, Jinglin Liu, Chen Zhang, Qian Yang, Shengpeng Ji, Rongjie

Huang, Chunfeng Wang, Xiang Yin, and 1 others. 2023b. Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. *arXiv preprint arXiv:2306.03509*.

Xin Jing, Kun Zhou, Andreas Triantafyllopoulos, and Björn W Schuller. 2025. Enhancing emotional text-to-speech controllability with natural language guidance through contrastive learning and diffusion models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. 2024. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. 2024. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. 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. *Preprint*, arXiv:2505.23009.

Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. 2024. The lottery ticket hypothesis in denoising: Towards semantic-driven initialization. *Preprint*, arXiv:2312.08872.

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

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, and 1 others. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. 2024. Not all noises are created equally:diffusion noise selection and optimization. *Preprint*, arXiv:2407.14041.

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

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, and 1 others. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2022. Model for dimensional speech emotion recognition based on wav2vec 2.0.

Ruoyu Wang, Huayang Huang, Ye Zhu, Olga Russakovsky, and Yu Wu. 2024a. The silent prompt: Initial noise as implicit guidance for goal-driven image generation. *arXiv e-prints*, pages arXiv–2412.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024b. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, and 1 others. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

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

Zequn Xie. 2026. Conquer: Context-aware representation with query enhancement for text-based person search. *arXiv preprint arXiv:2601.18625*.

Zequn Xie, Xin Liu, Boyun Zhang, Yuxiao Lin, Sihang Cai, and Tao Jin. 2026a. Hvd: Human vision-driven video representation learning for text-video retrieval. *arXiv preprint arXiv:2601.16155*.

Zequn Xie, Chuxin Wang, Yeqiang Wang, Sihang Cai, Shulei Wang, and Tao Jin. 2025b. Chat-driven text generation and interaction for person retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5259–5270.

Zequn Xie, Boyun Zhang, Yuxiao Lin, and Tao Jin. 2026b. Delving deeper: Hierarchical visual perception for robust video-text retrieval. *arXiv preprint arXiv:2601.12768*.

Weicai Yan, Wang Lin, Zirun Guo, Ye Wang, Fangming Feng, Xiaoda Yang, Zehan Wang, and Tao Jin. 2025. Diff-prompt: Diffusion-driven prompt generator with mask supervision. *Preprint*, arXiv:2504.21423.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, and 1 others. 2025. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. *arXiv preprint arXiv:2504.12867*.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2021. 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)*, pages 920–924. IEEE.

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

A Appendix

A.1 Derivation of the Purified Vector Field

1. Definition of the Target Vector Field. According to the property of Flow Matching, the optimal vector field is proportional to the score function (the gradient of the log-density):

$$v(x_t, t) \propto s(x_t, t) := \nabla_{x_t} \log p(x_t, t).$$

Using the identity

$$\nabla_x \log p(x) = \frac{\nabla_x p(x)}{p(x)},$$

we can expand the log-gradient in terms of the density gradient.

2. Gradient of the Mixture Density. Recall the disentangled density from Eq. (7):

$$p_{\text{true}}(x_t \mid c_{\text{emo}}) = \frac{1}{\pi}\Big[p_\theta(x_t \mid c_{\text{emo}}) - (1-\pi)\, p_\theta(x_t)\Big].$$

Note: In Eq. (8), the paper denotes the neutral distribution as $p_\theta(x_t)$.

We compute the gradient ∇_{x_t} for both sides:

$$\begin{aligned}&\nabla_{x_t} p_{\text{true}}(x_t \mid c_{\text{emo}}) \\ &= \frac{1}{\pi}\Big[\nabla_{x_t} p_\theta(x_t \mid c_{\text{emo}}) - (1-\pi)\, \nabla_{x_t} p_\theta(x_t)\Big].\end{aligned}$$

3. Substituting Gradients with Velocity Fields. We apply the chain rule

$$\nabla_x p(x) = p(x)\, \nabla_x \log p(x),$$

and the proportionality $v(x) \propto \nabla_x \log p(x)$ to express gradients in terms of the velocity outputs v_c and v_u:

Conditional Term:

$$\begin{aligned}&\nabla_{x_t} p_\theta(x_t \mid c_{\text{emo}}) \\ &= p_\theta(x_t \mid c_{\text{emo}})\, \nabla_{x_t} \log p_\theta(x_t \mid c_{\text{emo}}) \\ &\propto p_\theta(x_t \mid c_{\text{emo}})\, v_c(x_t, t),\end{aligned}$$

Unconditional Term:

$$\begin{aligned}&\nabla_{x_t} p_\theta(x_t) \\ &= p_\theta(x_t)\, \nabla_{x_t} \log p_\theta(x_t) \propto p_\theta(x_t)\, v_u(x_t, t).\end{aligned}$$

Substituting these back into the gradient equation gives (up to a common proportionality constant):

$$\begin{aligned}\nabla_{x_t} p_{\text{true}}(x_t \mid c_{\text{emo}}) \propto \frac{1}{\pi}\Big[&p_\theta(x_t \mid c_{\text{emo}})\, v_c(x_t, t) \\ &- (1-\pi)\, p_\theta(x_t)\, v_u(x_t, t)\Big].\end{aligned}$$

4. Final Assembly. Substitute the numerator (from Step 3) and the denominator (from Eq. (7)) into

$$\begin{aligned}v_{\text{true}}(x_t, t) &\propto \nabla_{x_t} \log p_{\text{true}}(x_t \mid c_{\text{emo}}) \\ &= \frac{\nabla_{x_t} p_{\text{true}}(x_t \mid c_{\text{emo}})}{p_{\text{true}}(x_t \mid c_{\text{emo}})}.\end{aligned}$$

The common factor $1/\pi$ cancels out, yielding Eq. (8):

$$\begin{aligned}&v_{\text{true}}(x_t, t) = \\ &\frac{p_\theta(x_t \mid c_{\text{emo}})\, v_c(x_t, t) - (1-\pi)\, p_\theta(x_t)\, v_u(x_t, t)}{p_\theta(x_t \mid c_{\text{emo}}) - (1-\pi)\, p_\theta(x_t)}.\end{aligned}$$

A.2 Derivation of Eq. (9) from Eq. (8)

To rewrite v_{true} in the standard classifier-free guidance form (base + scale · difference), we perform algebraic manipulation on the numerator of Eq. (8).

Let $D = p_\theta(x_t|c_{emo}) - (1-\pi)p_\theta(x_t)$ be the denominator.

Starting from Eq. (8):

$$v_{true} = \frac{p_{cond}v_c - (1-\pi)p_{uncond}v_u}{D}$$

We add and subtract the term $p_{cond}v_u$ in the numerator to create the difference $(v_c - v_u)$:

$$\begin{aligned} v_{true} &= \frac{p_{cond}v_c - p_{cond}v_u + p_{cond}v_u - (1-\pi)p_{uncond}v_u}{D} \\ &= \frac{p_{cond}(v_c - v_u) + [p_{cond} - (1-\pi)p_{uncond}]v_u}{D} \end{aligned}$$

Now, we split the fraction into two parts:

$$v_{true} = \frac{[p_{cond} - (1-\pi)p_{uncond}]v_u}{D} + \frac{p_{cond}(v_c - v_u)}{D}$$

The first term simplifies to v_u because the numerator equals the denominator D:

$$v_{true} = v_u + \left(\frac{p_{cond}}{D}\right)(v_c - v_u)$$

By defining $\lambda(x_t, t) = \frac{p_{cond}}{D}$ (which matches Eq. 10), we obtain Eq. (9):

$$v_{true}(x_t) = v_u(x_t) + \lambda(x_t, t) \cdot (v_c(x_t) - v_u(x_t))$$

Derivation of the Recursive Likelihood Ratio Update

In this section, we provide the detailed derivation of Equation (15), which expresses the incremental change of the likelihood ratio in terms of transition probabilities.

1. Decomposition of the Log-Likelihood Ratio Increment

Recall Equation (14), where the increment of the log-likelihood ratio $\Delta \log R_t$ is decomposed into conditional and unconditional components:

$$\begin{aligned} \Delta \log R_t = &\underbrace{[\log p_\theta(x_{t+\Delta t}|c_{emo}) - \log p_\theta(x_t|c_{emo})]}_{\text{Conditional Change}} \\ &- \underbrace{[\log p_\theta(x_{t+\Delta t}) - \log p_\theta(x_t)]}_{\text{Unconditional Change}} \end{aligned} \tag{21}$$

2. Markov Property and Chain Rule Application

Since the generation process samples a specific continuous trajectory, we can leverage the Markov property of the generative process. Consider the joint probability of consecutive states x_t and $x_{t+\Delta t}$. By applying the chain rule, we have:

$$p(x_{t+\Delta t}, x_t) = p(x_{t+\Delta t}|x_t) \cdot p(x_t)$$

In the context of a small time step Δt, the probability density of the next state $p(x_{t+\Delta t})$ is principally determined by the transition from the current state x_t. Thus, we can approximate the evolution of the marginal density locally using the transition kernel:

$$p(x_{t+\Delta t}) \approx p(x_{t+\Delta t}|x_t) \cdot p(x_t)$$

Taking the logarithm and rearranging the terms yields the unconditional incremental change:

$$\log p(x_{t+\Delta t}) - \log p(x_t) \approx \log p(x_{t+\Delta t}|x_t)$$

3. Application to the Conditional Distribution

The same logic applies to the conditional distribution conditioned on emotion c_{emo}. The transition dynamics under the emotion condition satisfy:

$$p(x_{t+\Delta t}|c_{emo}) \approx p(x_{t+\Delta t}|x_t, c_{emo}) \cdot p(x_t|c_{emo})$$

Similarly, taking the logarithm gives the conditional incremental change:

$$\begin{aligned} \log p(x_{t+\Delta t}|c_{emo}) - \log p(x_t|c_{emo}) \\ \approx \log p(x_{t+\Delta t}|x_t, c_{emo}) \end{aligned} \tag{22}$$

4. Final Derivation

Substituting the results from Steps 2 and 3 back into Equation (21), we obtain the update rule expressed solely in terms of the local transition probabilities:

$$\Delta \log R_t = \log p(x_{t+\Delta t}|x_t, c_{emo}) - \log p(x_{t+\Delta t}|x_t) \tag{23}$$

This confirms Equation (15). It indicates that the evolution of the likelihood ratio is driven by the divergence between the conditional and unconditional transition probabilities at the current time step.

A.3 Interaction Analysis of ERNP Hyperparameters

To investigate the optimal intensity for the ERNP, we conduct a joint ablation study on the Lookahead guidance scale λ_{init} and the Calibration base scale λ_{base}. We define the Effective Gradient Gap as $\Delta\lambda = \lambda_{init} - \lambda_{base}$. We test various configurations by varying $\lambda_{init} \in \{10, 20, 30, 40\}$ and

Table 11: Joint ablation study on λ_{init} and λ_{base} evaluated on the HIED dataset.

λ_{init}	λ_{base}	Gap ($\Delta\lambda$)	WER (%) ↓	ES ↑	Recall ↑
10	1	9	3.43	0.81	52.1%
10	3	7	3.65	0.83	51.8%
20	1	19	3.23	0.82	53.7%
20	3	17	3.34	0.80	53.3%
30	**1**	**29**	**3.05**	**0.84**	**54.2%**
40	1	39	3.06	0.84	53.8%

$\lambda_{base} \in \{1, 3\}$. The results are shown in Table 11. It can be observed that widening the gradient gap $\Delta\lambda$ from 9 to 29 consistently enhances model performance. Specifically, the setting with $\lambda_{init} = 30$ and $\lambda_{base} = 1$ achieves the best trade-off, reaching the lowest WER (3.05%) and highest Emotion Similarity (0.84). However, compared to $\lambda_{init} = 30$, further increasing the scale to $\lambda_{init} = 40$ ($\Delta\lambda = 39$) leads to diminishing returns, with slight degradations in both intelligibility and expressiveness. This suggests that excessive rectification strength may over-perturb the initialization, distorting the linguistic structure. Therefore, we identify $\lambda_{init} = 30$ as the optimal configuration for balancing emotional intensity and acoustic stability.

A.4 Pseudo Code

Algorithm 1 Emotion-Rectified Noise Prior (ERNP)

Require: Target emotion condition c_{emo}, lookahead step size τ, initial guidance scale γ_{init}, base guidance scale γ_{base}.

Ensure: Rectified initial noise x_0^*.

Sample standard Gaussian noise $x_0 \sim \mathcal{N}(0, I)$ ▷ Standard initialization

Lookahead Step:

Calculate velocity with high guidance:

$\tilde{v} \leftarrow \text{Model}(x_0, 0, c_{emo}, \gamma_{init})$

Move forward to pseudo-state:

$x_\tau \leftarrow x_0 + \tau \cdot \tilde{v}$ ▷ Eq. (2)

Calibration Step:

Calculate velocity for retraction:

$\hat{v} \leftarrow \text{Model}(x_\tau, \tau, c_{emo}, \gamma_{base})$

Move backward to rectified origin:

$x_0^* \leftarrow x_\tau - \tau \cdot \hat{v}$ ▷ Eq. (3)

return x_0^*

Algorithm 2 Likelihood-Inverse Guidance (LIG) Sampling

Require: Rectified noise x_0, emotion condition c_{emo}, neutral condition $\emptyset$, total steps N, purity π, max scale λ_{max}.

Ensure: Generated speech sample x_1.

Initialization:

Initialize log-likelihood ratio at $t = 0$: $\log R_0 \leftarrow 0$ ▷ Assumes $p(x_0|c) \approx p(x_0)$

Step size $\Delta t \leftarrow 1/N$

for $i = 0$ **to** $N - 1$ **do**

  $t \leftarrow i \cdot \Delta t$

  $\sigma_t \leftarrow 1 - t$ ▷ Noise level at t

  1. Model Prediction:

  $v_c \leftarrow \text{Model}(x_t, t, c_{emo})$ ▷ Conditional velocity

  $v_u \leftarrow \text{Model}(x_t, t, \emptyset)$ ▷ Unconditional velocity

  2. Dynamic Schedule Calculation:

  $R_t \leftarrow \exp(\log R_t)$ ▷ Recover R_t from log-domain

  $\lambda_t \leftarrow \frac{R_t}{R_t - (1-\pi)}$ ▷ Calculate guidance scale, Eq. (12)

  $\lambda_t \leftarrow \min(\lambda_t, \lambda_{max})$ ▷ Clamp for stability

  3. State Update (Euler Step):

  $v_{true} \leftarrow v_u + \lambda_t \cdot (v_c - v_u)$ ▷ Rectified vector field, Eq. (9)

  $x_{t+\Delta t} \leftarrow x_t + v_{true} \cdot \Delta t$ ▷ Update state

  4. Recursive Likelihood Estimation (for next step):

  Calculate squared norm difference:

  $\Delta \mathcal{D}_t \leftarrow \|v_{true} - v_u\|^2 - \|v_{true} - v_c\|^2$

  Calculate log-ratio increment:

  $\Delta \log R_t \leftarrow \frac{\Delta t^2}{2\sigma_t^2} \cdot \Delta \mathcal{D}_t$ ▷ Derived from Eq. (18)

  Update log-likelihood for next step:

  $\log R_{t+\Delta t} \leftarrow \log R_t + \Delta \log R_t$

end for

return x_1