PUSHING THE FRONTIER OF LIGHTWEIGHT DIFFUSION FOR ZERO-SHOT TEXT-TO-SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

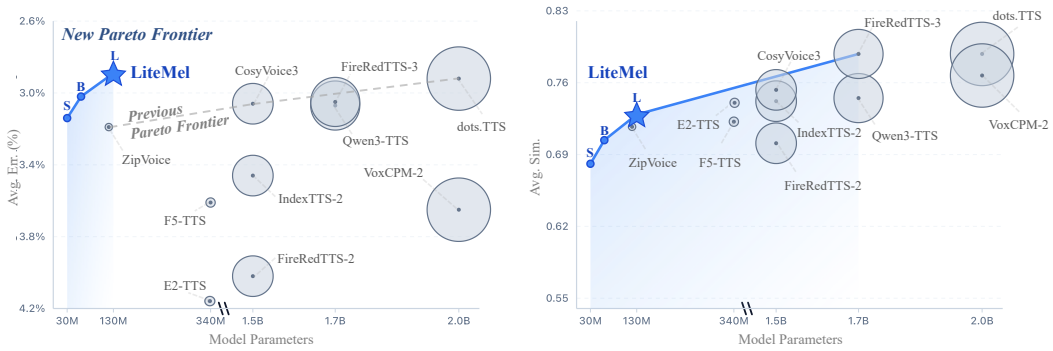
Zero-shot text-to-speech (TTS), which clones an unseen voice from a short prompt, has largely relied on domain-specific priors such as forced alignment or phoneme duration, whose supervision is costly and hard to scale. Recently, large diffusion models trade these priors for scale, letting the alignment emerge from data, while lightweight models for on-device deployment cannot afford this trade. In the absence of a prior, such models fail to learn the alignment reliably, and the explicit priors would reimpose the supervision that scaling was meant to eliminate. In this paper, we present **LITEMEL**, a family of mel-spectrogram diffusion models from 0.03B (Small) to 0.13B (Large) parameters that achieves a Pareto-optimal trade-off between model size and performance. Specifically, we first propose a dual-path framework that fuses plain Diffusion Transformers (DiT) favouring intelligibility with hierarchical DiTs favouring speaker fidelity. We also introduce speaker identity that enters through the time embedding and reaches every block by adaptive normalization. Moreover, we propose a learned soft text-speech aligner that restores the monotonic guidance without any supervision. It contains a duration head with a differentiable Gaussian expansion trained by the flow-matching objective alone. To further enhance, we introduce on-policy distillation (OPD) for each student model against our 0.13B member as teacher along the student's own stochastic rollouts, so that the teacher corrects the states the sampler actually visits. Our extensive experiments show that LITEMEL-Large attains the SOTA intelligibility, and OPD improves in every metric at both student sizes, cutting the average error by 23% relative to pre-training. The three models together occupy the entire Pareto frontier of average error against parameter count.

1 INTRODUCTION

Diffusion models have shown strong generative ability across modalities, including image (Geng et al., 2025; 2026), audio (Guan et al., 2024; Kong et al., 2021) and video (Wei et al., 2024; Jin et al., 2025) generation. Their flow-matching formulation (Lipman et al., 2023; Liu et al., 2023) regresses a velocity field along a straight probability path between noise and data, which shortens the sampling trajectory and brings high-quality generation of continuous signals. In speech, flow-matching has enabled a challenging generation task: *zero-shot text-to-speech (TTS)*, in which a model reads arbitrary text aloud in the voice of an unseen speaker, conditioned on only a few seconds of audio from the speaker. Scaled to hundreds of millions and billions of parameters, such flow systems (Eskimez et al., 2024; Chen et al., 2025; Shen et al., 2024) now approach human parity on standard benchmarks (Anastassiou et al., 2024).

However, this progress has come at the cost of parameter scale. While large flow-matching models (Xin et al., 2026; Chen et al., 2025) discard traditional domain-specific priors, such as forced alignments, and allow text-speech alignment to emerge naturally from scale, on-device deployments cannot afford such massive capacity. This creates a fundamental dilemma for lightweight TTS systems. If without sufficient parameter budget, implicit alignment fails to converge reliably, yet reverting to explicit alignment priors reinstates the heavy supervision pipelines and data scalability bottlenecks. A truly practical lightweight system demands an architecture that internally enforces monotonic alignment without relying on external labels, paired with simple optimization strategies. Although recent advancing efforts (Han et al., 2025; OpenMOSS, 2026; Chen et al., 2025) have explored compact architectures and complex optimization strategies, achieving an optimal trade-off

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.



(a) Intelligibility (Avg. Word Error Rate ↓)

(b) Fidelity (Avg. Speaker Similarity ↑)

Figure 1: **Pareto frontier comparison** of TTS models in terms of **(a)** Intelligibility and **(b)** Fidelity. The blue curve and shaded region represent the new Pareto frontier established by our LITEMEL series, which comprises the largest teacher model (LITEMEL-L) and its smaller variants (LITEMEL-B and LITEMEL-S) developed via OPD. LITEMEL demonstrates superior parameter efficiency, achieving competitive or better performance than large-scale baselines with significantly fewer parameters.

between speaker fidelity, intelligibility, and capacity remains an open challenge. This raises the following critical questions:

> *Can a lightweight model match the generation quality of systems an order of magnitude larger, and what fundamental mechanisms must be built to replace the emergent capabilities of scale?*

To address this question, we present **LITEMEL**, a family of **Li**ght**te**ight **Mel**-spectrogram flow-matching models (from 0.03B to 0.13B parameters) that replaces scale-emergent alignment with deliberate architecture and simple optimization design. Our framework resolves the lightweight impasse through three key mechanisms. First, to achieve monotonic text-speech alignment without external supervision, we propose a learned soft alignment mechanism. By allocating mel frames via normalized token logits, it algebraically guarantees that frame counts match the target length. We also ensure that a minimum frame per token renders monotonicity a structural guarantee rather than a learned heuristic, allowing the head to learn alignment smoothly under just the flow-matching loss alone. Second, to balance competing attributes under a tight capacity budget, we propose a dual-path framework that fuses plain DiT pathway (*favouring intelligibility*) and a hierarchical DiT pathway (*favouring speaker fidelity*) in parallel, fusing their representations to estimate velocity vector fields. Third, to further eliminate capacity-induced trajectory drift, we introduce On-Policy Distillation (OPD) with our larger teacher to directly correct the out-of-distribution states the sampler actually visits following the student's own stochastic rollout.

As shown in Figure 1, our extensive evaluations on Seed-TTS-eval demonstrate that LITEMEL pushes the Pareto frontier of zero-shot TTS quality across model scales, especially on intelligibility. Notably, our 0.13B variant achieves state-of-the-art intelligibility, outperforming baseline models up to 15 times its parameter count. Through OPD, our 0.03B student model already surpasses systems 4 to 11 times its size, reducing pre-training error by 24% and proving that trajectory post-correction is vital for lightweight generative speech models.

Contributions. Our contributions can be summarized as follows:

- By incorporating deliberate structural inductive biases, we propose a dual-path framework with post-training distillation that achieves a Pareto-optimal trade-off between model scale and generation quality for lightweight TTS.
- We introduce an unsupervised soft monotonic alignment that supplies the text-frame correspondence without external duration labels or forced aligners.
- We show that LITEMEL builds the new Pareto frontier of intelligibility against parameter scale on Seed-TTS-eval, and reduces student errors by 23%, driving substantial gains in speaker similarity.

2 Related Work

Flow-matching for zero-shot TTS. Voicebox (Le et al., 2023) cast zero-shot TTS as speech infilling under a conditional flow-matching objective, relying on an external forced aligner (McAuliffe et al., 2017) for frame durations. E2-TTS (Eskimez et al., 2024) eliminate the aligner by padding character sequences with filler tokens to the mel length, allowing attention to discover alignment implicitly, and F5-TTS (Chen et al., 2025) maintain this interface while upgrading the backbone to a DiT (Peebles & Xie, 2023). Moreover, Longcat-AudioDiT (Xin et al., 2026) injects text via cross-attention and scales up the DiT to the 3.5B parameter tier, further pushing zero-shot TTS to a new frontier. However, existing DiT or U-Net architectures rarely incorporate domain-specific priors of TTS, and scaling them down while balancing intelligibility and speaker fidelity remains remarkably challenging. In contrast, we perform extensive architectural ablations and propose a dual-path framework with a post-training strategy that significantly boosts lightweight model performance.

Lightweight TTS systems. Compressing the parameter scale of zero-shot TTS is critical for edge deployment and low-latency interactive applications. While early non-autoregressive flow-matching models relied on heavy parameter scaling (0.5B to over 3.5B) to implicitly learn text-speech alignment and high-fidelity acoustics, recent works have begun pushing into the compact regime. Notable works include ZipVoice (Han et al., 2025) and MOSS-TTS-Nano (OpenMOSS, 2026), which demonstrate the feasibility of around 0.1B architectures. However, these compact systems reveal clear trade-offs: ZipVoice relies on complex optimization strategies (*e.g.*, ScaledAdam, Balancer, and Whitener (Yao et al., 2024)) to maintain performance, whereas MOSS-TTS-Nano achieves small scale at the cost of noticeable generation quality degradation. In this work, we systematically scale the backbone down to at least 0.03B parameters without sacrificing optimization simplicity. By substituting scale-emergent properties with delicate structural inductive biases and post-distillation, LiteMel establishes a new Pareto frontier for lightweight speech synthesis.

3 LiteMel: Lightweight Mel-Spectrogram Flow-Matching

In this section, we present LiteMel in the following order: Sec. 3.1 introduces the preliminary concepts necessary for understanding our work. We then detail our dual-path framework with soft monotonic guidance in Sec. 3.2 and post-training in Sec. 3.3.

3.1 Preliminaries

Let $\boldsymbol{x}_1$ denote the target mel-spectrogram and $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ Gaussian noise. Following the Optimal Transport Conditional Flow Matching (OT-CFM) framework (Lipman et al., 2023), the optimal-transport probability path $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1$, $t \in [0,1]$, a neural network $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$ is trained to regress the target velocity field $\boldsymbol{x}_1 - \boldsymbol{x}_0$ by minimizing the objective:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1} \left\| \left(\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\right) \odot \boldsymbol{m} \right\|^2, \tag{1}$$

where the binary mask $\boldsymbol{m}$ restricts the regression loss to the target generation region, while the unmasked regions serve as context prompts to condition the vector field. And the condition $\boldsymbol{c}$ comprises the text sequence, the prompt mel-spectrogram and a speaker embedding extracted by a CAM++ encoder (Wang et al., 2023) in our work. During inference, the target $\boldsymbol{x}_1$ is synthesized by numerically integrating the learned ordinary differential equation (ODE) $\mathrm{d}\boldsymbol{x}_t/\mathrm{d}t = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$ from $t = 0$ to $t = 1$ using an ODE solver, incorporating classifier-free guidance (CFG) (Ho & Salimans, 2022). Finally, the generated mel-spectrogram is converted into a high-fidelity waveform using a neural vocoder (*i.e.*, Vocos (Siuzdak, 2024) in our work).

3.2 Dual-Path Framework for Intelligibility-Fidelity Trade-off

Current flow-matching TTS models typically rely either on plain-DiT (*e.g.,* F5-TTS) for scaling efficiency (Peebles & Xie, 2023) or U-shaped structures (*e.g.,* E2-TTS) for multi-scale feature aggregation (Ronneberger et al., 2015). However, when model under strict parameter constraints, architectural inductive bias becomes decisive. Empirically, we find that full-resolution DiT designs prioritize fine-grained phonetic alignment with high intelligibility, while hierarchical U-shaped

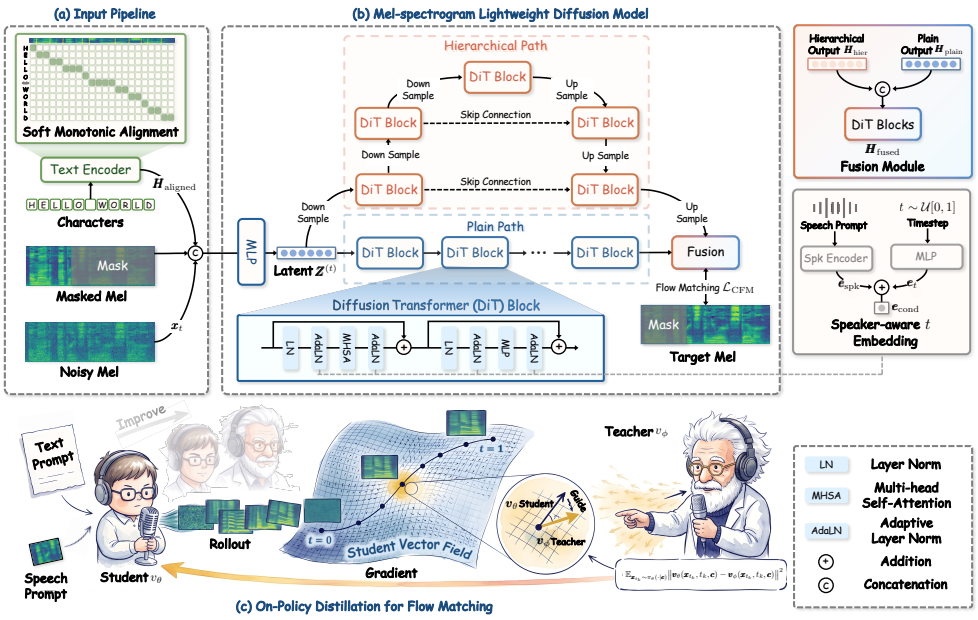


Figure 2: **Overview of LITEMEL. (a)** The input pipeline places characters on the mel frame grid by soft monotonic alignment and concatenates them with the masked and noisy mel-spectrograms. **(b)** The vector field estimator takes both plain and hierarchical paths in parallel, and a fusion module with DiT fuses their features into a target velocity prediction. **(c)** The student rolls out a trajectory with its own stochastic sampler, and each update shrinks the velocity gap to the teacher at the rollout states.

networks favor speaker fidelity. To synergize these advantages, in Figure 2, LITEMEL introduces a dual-path framework that explicitly bridges the intelligibility-fidelity trade-off via latent fusion.

Input pipeline. Given a text character sequence of length N_{text}, and a target mel-spectrogram $\boldsymbol{x}_1 \in \mathbb{R}^{N_{\text{mel}} \times d_{\text{mel}}}$, the input pipeline maps text characters and mel into a unified sequence embedding $\boldsymbol{Z}^{(t)} \in \mathbb{R}^{T \times d}$ for the backbone, where $T = N_{\text{mel}}$ and d is the hidden width, and the pipeline operates through two sequential stages:

1) Soft Monotonic Alignment. To map N_{text} text characters to N_{mel} mel frames without external aligners (*e.g.*, MFA), a lightweight multi-layer perceptron (MLP) head emits one logit l_i per character. A softmax over the text characters turns these into proportions p_i, and each character is allocated

$$d_i = 1 + p_i \cdot (N_{\text{mel}} - N_{\text{text}}), \qquad p_i = \frac{\exp(l_i)}{\sum_j \exp(l_j)}, \tag{2}$$

so every character retains at least one frame and $\sum_i d_i = N_{\text{mel}}$ holds by construction, with no separate length prediction. Because $d_i \geq 1$, the cumulative centres $\mu_k = \sum_{i<k} d_i + d_k/2$ are strictly increasing, and the alignment cannot reorder characters. Text-aligned features $\boldsymbol{H}_{\text{aligned}} \in \mathbb{R}^{N_{\text{mel}} \times d}$ are then read from text features $\boldsymbol{H}_{\text{text}} \in \mathbb{R}^{N_{\text{text}} \times d}$ using a soft alignment matrix $\boldsymbol{A} \in \mathbb{R}^{N_{\text{mel}} \times N_{\text{text}}}$:

$$A_{n,k} = \frac{\exp\left(-\frac{(n+\frac{1}{2}-\mu_k)^2}{2\sigma_k^2}\right)}{\sum_j \exp\left(-\frac{(n+\frac{1}{2}-\mu_j)^2}{2\sigma_j^2}\right)}, \qquad \boldsymbol{H}_{\text{aligned}} = \boldsymbol{A}\boldsymbol{H}_{\text{text}}, \tag{3}$$

where $n \in [0, N_{\text{mel}} - 1]$ denotes the target mel-frame index, $k \in [1, N_{\text{text}}]$ represents the source text character index, and $n + \frac{1}{2}$ specifies the continuous temporal center of frame n. The kernel scale $\sigma_k = \max(\beta d_k, 1)$ controls the temporal bandwidth for token k, where the lower bound of 1 frame guarantees sufficient overlap between neighboring kernels to preserve differentiability and softness across the sequence. The gradient reaches the head through μ under the flow-matching objective alone, with no duration loss and no forced alignment.

2) Feature Noising and Projection. The full target mel $\boldsymbol{x}_1$ is interpolated towards noise as in (1), yielding the noisy state $\boldsymbol{x}_t$ across the entire sequence at a continuous timestep $t \in [0, 1]$. Simultaneously, a

binary temporal mask $\boldsymbol{m} \in \{0,1\}^{N_{\text{mel}}}$ is applied to the clean mel $\boldsymbol{x}_1$ to preserve only the uncorrupted prompt region $\boldsymbol{m} \odot \boldsymbol{x}_1$, providing context for conditional generation. Finally, the frame-aligned text representations, the masked clean mel $\boldsymbol{m} \odot \boldsymbol{x}_1$, and the noisy state $\boldsymbol{x}_t$ are concatenated channel-wise and projected by an MLP into the sequence embedding $\boldsymbol{Z}^{(t)} \in \mathbb{R}^{T \times d}$.

Speaker-aware time embedding for adaptive normalization. To maintain high speaker fidelity under a lightweight capacity, we employ a compact pre-trained speaker encoder, CAM++ (∼7M parameters) (Wang et al., 2023), to extract a global speaker embedding $\boldsymbol{e}_{\text{spk}}$. The timestep t is mapped to a time embedding $\boldsymbol{e}_t$ via an MLP. We construct a unified conditioning vector $\boldsymbol{e}_{\text{cond}} = \boldsymbol{e}_t + \boldsymbol{e}_{\text{spk}}$, which modulates the attention and feed-forward blocks across both parallel branches and the fusion head strictly via Adaptive Layer Normalization (AdaLN) (Peebles & Xie, 2023).

Dual-path backbone for velocity prediction. To capture both fine-grained acoustic details and multi-scale contextual patterns, LITEMEL models the velocity field through a dual-path backbone followed by latent fusion. Given the input sequence embedding $\boldsymbol{Z}^{(t)} \in \mathbb{R}^{T \times d}$ at timestep t and $\boldsymbol{e}_{\text{cond}}$, representations are extracted concurrently via two structurally distinct branches: a plain path $\boldsymbol{H}_{\text{plain}} = f_{\text{plain}}(\boldsymbol{Z}_t, \boldsymbol{e}_{\text{cond}}) \in \mathbb{R}^{T \times d_{\text{p}}}$ and a hierarchical path $\boldsymbol{H}_{\text{hier}} = f_{\text{hier}}(\boldsymbol{Z}_t, \boldsymbol{e}_{\text{cond}}) \in \mathbb{R}^{T \times d_{\text{h}}}$. Specifically, $f_{\text{plain}}(\cdot)$ is a plain DiT stack that preserves the full temporal resolution T across all layers at a constant width d_{p}. In parallel, $f_{\text{hier}}(\cdot)$ is a 5-stage U-shaped DiT operating over three temporal scales, downsampling the resolution twice to $T/4$ to aggregate global context before restoring it through upsampling blocks and skip connections. Rather than decoding velocity fields independently, our fusion module is introduced directly in the latent space by concatenating the representations along the channel dimension $\boldsymbol{H}_{\text{fused}} = [\boldsymbol{H}_{\text{plain}} \,\|\, \boldsymbol{H}_{\text{hier}}] \in \mathbb{R}^{T \times (d_{\text{p}} + d_{\text{h}})}$. A lightweight fusion head consisting of shallow DiT layers, conditioned on $\boldsymbol{e}_{\text{cond}}$ through the same AdaLN mechanism, then transforms the joint representation into the estimated velocity field:

$$\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) = f_{\text{fuse}}\big(\boldsymbol{H}_{\text{fused}}, \boldsymbol{e}_{\text{cond}}\big) \in \mathbb{R}^{T \times d_{\text{mel}}}. \tag{4}$$

3.3 POST-TRAINING VIA ON-POLICY DISTILLATION

While pre-training establishes foundational generative capabilities, post-training is crucial for boosting intelligibility and refining acoustic fidelity. A primary culprit is *exposure bias*: standard training supervises models exclusively along ground-truth data-coupled paths, leaving them ill-equipped to handle self-generated off-manifold states.

To address this challenge, we introduce *On-Policy Distillation* (OPD) (Agarwal et al., 2024; Li et al., 2026). Instead of relying on static, pre-defined trajectories, OPD aligns a compact student model $\boldsymbol{v}_\theta$ with a high-capacity teacher $\boldsymbol{v}_\phi$ by shifting state selection directly to the student's own generative rollouts. Concretely, the student generates trajectories by discretizing an SDE of the marginal-preserving form. Using an Euler–Maruyama discretization over a sampling grid $t_0 > t_1 > \cdots > t_K$, the one-step state transition from step k to $k+1$ is formulated as:

$$\begin{aligned} \boldsymbol{x}_{t_{k+1}} &\sim \mathcal{N}\big(\boldsymbol{\mu}_\theta(\boldsymbol{x}_{t_k}, t_k),\ g^2\,\Delta t_k\,\boldsymbol{I}\big), \\ \boldsymbol{\mu}_\theta(\boldsymbol{x}_{t_k}, t_k) &= \boldsymbol{x}_{t_k} + \Big[\boldsymbol{v}_\theta(\boldsymbol{x}_{t_k}, t_k, \boldsymbol{c}) - \tfrac{g^2}{2(1-t_k)}\big(\boldsymbol{x}_{t_k} - t_k \boldsymbol{v}_\theta(\boldsymbol{x}_{t_k}, t_k, \boldsymbol{c})\big)\Big]\Delta t_k, \end{aligned} \tag{5}$$

where $\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, g denotes the constant noise scale, and $\Delta t_k = t_{k+1} - t_k < 0$. Notably, $\boldsymbol{v}_\theta$ represents the standard CFG-combined velocity field in our work, ensuring that the training rollout strictly matches the inference decode path. The teacher's velocity field is then matched directly on these visited rollout states:

$$\mathcal{L}_{\text{OPD}} = \mathbb{E}_{\boldsymbol{c}\sim\mathcal{D},\ \boldsymbol{x}_{t_k}\sim\pi_\theta(\cdot|\boldsymbol{c})}\big\|\boldsymbol{v}_\theta(\boldsymbol{x}_{t_k}, t_k, \boldsymbol{c}) - \boldsymbol{v}_\phi(\boldsymbol{x}_{t_k}, t_k, \boldsymbol{c})\big\|^2, \tag{6}$$

where π_θ represents the trajectory distribution induced by the student policy. To optimize Eq. (6), gradients are computed independently per step by holding the visited states $\boldsymbol{x}_{t_k}$ fixed (*i.e.*, applying a stop-gradient). This avoids backpropagation through time, ensuring constant memory overhead with respect to the rollout depth and preventing gradient instability.

The velocity-matching objective in Eq. (6) can be strictly grounded by examining the one-step transition kernels $p_\theta(\cdot|\boldsymbol{x}_{t_k})$ and $p_\phi(\cdot|\boldsymbol{x}_{t_k})$ induced by the student and teacher under the SDE discretization in Eq. (5). Because both models evaluate at the exact same visited state $\boldsymbol{x}_{t_k}$ with shared isotropic covariance $g^2\,\Delta t_k\,\boldsymbol{I}$, the per-step reverse KL divergence between their Gaussian kernels admits an

exact closed-form velocity field gap:

$$\mathrm{KL}\big(p_\theta \,\|\, p_\phi\big) = \underbrace{\frac{c(t_k)^2\,\Delta t_k}{2g^2}}_{w(t_k)} \big\|\boldsymbol{v}_\theta(\boldsymbol{x}_{t_k}, t_k, \boldsymbol{c}) - \boldsymbol{v}_\phi(\boldsymbol{x}_{t_k}, t_k, \boldsymbol{c})\big\|^2, \qquad c(t) = 1 + \frac{g^2 t}{2(1-t)}. \tag{7}$$

Eq. (7) reveals that minimizing the reverse KL along on-policy rollouts reduces to a weighted velocity-matching loss with a temporal coefficient $w(t_k)$. Crucially, because both kernels share the identical Gaussian covariance in Eq. (5), the sample noise variance cancels out analytically in Eq. (7). This exact closed-form evaluation avoids the high-variance Monte-Carlo gradient estimation typically required by reinforce-style policy gradient methods, enabling computationally efficient and gradient-stable on-policy training. Finally, as $g \to 0$, the SDE step in Eq. (5) gracefully collapses to the deterministic Euler ODE update $\boldsymbol{x}_{t_k} + \boldsymbol{v}_\theta \Delta t_k$. This guarantees mathematical consistency between ODE and SDE sampling, enabling seamless deployment across deterministic and stochastic samplers at test time.

4 Experimental Setup

Data. All models are trained on the Chinese and English portions of Emilia (He et al., 2024). After simply filtering out multi-speaker and transcription failure, we retain approximately 100K hours over 46M utterances. We evaluate our model on three benchmarks: *1) Seed-TTS test-en* (Anastassiou et al., 2024) (1,088 Common Voice samples), *2) Seed-TTS test-zh* (2,020 DiDiSpeech samples), *3) Seed-TTS test-zh-hard* (420 stress-testing samples).

Supervised training. Our supervised training consists of two sequential stages: pre-training and supervised fine-tuning (SFT). During pre-training, models are trained on the full dataset to establish foundational generative capabilities. For zero-shot SFT, we construct cross-utterance pairs (5k hours) by extracting distinct audio segments belonging to the same speaker from the Emilia dataset, using one segment as the prompt and another as the target. All models are pre-trained for 800k updates with a per-GPU batch size of 88k mel frames across 8 GPUs, yielding a global budget of 704k frames per update. For SFT, models are fine-tuned for 1 epoch on constructed cross-utterance datas. We use AdamW with a learning rate of 1×10^{-4} linearly warmed up for 8K updates and linearly decayed to 4×10^{-5}. Diffusion time steps are sampled per utterance from a logit-normal distribution ($\mu = 0, \sigma = 1$) with a shift of 3.0, which concentrates the density toward the high-noise end. We randomly mask lengths between 60% and 100% of the utterance. For CFG training, the text input and unmasked mel are randomly dropped at a 0.2 to construct the unconditional branch.

Post-training. Both 0.03B and 0.06B student models are distilled from the same frozen 0.13B SFT teacher via OPD. All models are trained for 18k updates with a learning rate of 2×10^{-5} and a global batch size of 400k mel frames. During post-training, prompts are sampled from Emilia dataset of 20k utterances (10k English, 10k Chinese), with one utterance per speaker. Each update samples a global constant of 16 independent trajectories across 8 prompts using the stochastic sampler in Eq. (5). Note that these trajectories serve as i.i.d. draws from π_θ to average the gradient.

Inference pipeline. All evaluations are conducted using Exponential Moving Averaged (EMA) model weights. For inference, the learned ODE (Sec. 3.1) is integrated using a first-order Euler solver. We decode for 32 sampling steps with classifier-free guidance at a strength of 4.0. The generated log-mel spectrograms are subsequently converted into time-domain waveforms using a pretrained Vocos vocoder (Siuzdak, 2024).

Evaluation metrics. We evaluate performance across objective, subjective, and efficiency dimensions. For *objective evaluation*, following the Seed-TTS-eval protocol (Anastassiou et al., 2024), intelligibility is measured via Word Error Rate (WER) using Whisper-large-v3 (Radford et al., 2023) for English and Character Error Rate (CER) via Paraformer-zh (Gao et al., 2022) for Chinese. Speaker similarity (SIM) is computed as the cosine similarity between WavLM-large embeddings (Chen et al., 2022) of synthesized speech and acoustic prompts. For *subjective evaluation*, we report Naturalness MOS (NMOS) and Speaker Similarity MOS (SMOS) on 5-point scales, alongside Comparative MOS (CMOS) against the teacher (details in Appendix E).

Baselines. We compare our approach against two tiers of published models: (1) *Large-scale generative models* (e.g., dots.TTS, CosyVoice3/2, Qwen3-TTS), which serve as quality upper bounds

Table 1: **Performances on Seed-TTS-eval across scales.** WER/CER (%) and speaker similarity (SIM) are reported, and lower is better for error rates, higher for SIM. **Params** denotes parameter count in Billions (B); **Data** denotes training audio volume in Million hours (M). **Bold** indicates the best entry among lightweight (<0.35B) systems. Our models are evaluated before and after OPD.

Model	Params	Data	test-en		test-zh		test-hard		Avg	
			WER↓	SIM↑	CER↓	SIM↑	CER↓	SIM↑	Err↓	SIM↑
Large Generative Model										
Seed-TTS (Anastassiou et al., 2024)	–	–	2.25	0.762	1.12	0.796	7.59	0.776	3.65	0.778
Qwen2.5-Omni (Xu et al., 2025)	7.00B	–	2.72	0.632	1.70	0.752	7.97	0.747	4.13	0.710
dots.tts (Lian et al., 2026)	2.00B	1.50M	1.34	0.768	0.96	0.805	6.46	0.792	2.92	0.788
VoxCPM2 (Zhou et al., 2026b)	2.00B	2.00M	1.84	0.753	0.97	0.795	8.13	0.753	3.65	0.767
Qwen3-TTS (Team, 2026)	1.70B	–	1.23	0.717	1.22	0.770	6.76	0.748	3.07	0.745
FireRedTTS-3 (Shen et al., 2026)	1.70B	3.10M	1.64	0.772	1.01	0.809	6.50	0.784	3.05	0.788
FireRedTTS-2 (Xie et al., 2025)	1.50B	1.40M	1.95	0.665	1.14	0.736	8.98	0.703	4.02	0.701
IndexTTS-2 (Zhou et al., 2026a)	1.50B	0.55M	2.23	0.706	1.03	0.765	7.12	0.755	3.46	0.742
CosyVoice3 (Du et al., 2025)	1.50B	1.00M	2.22	0.720	1.12	0.781	5.83	0.758	3.06	0.753
LongCat-AudioDiT-1B ()	1.00B	1.00M	1.18	0.812	1.78	0.762	6.33	0.787	3.10	0.787
CosyVoice2 (Du et al., 2024)	0.50B	0.17M	2.57	0.654	1.45	0.753	6.83	0.724	3.62	0.710
Lightweight Generative Model										
F5-TTS (Chen et al., 2025)	0.34B	0.10M	1.83	0.668	1.56	0.759	7.44	0.738	3.61	0.722
E2-TTS (Eskimez et al., 2024)	0.33B	0.10M	2.24	**0.723**	1.97	0.742	10.34	**0.752**	4.85	**0.739**
ZipVoice (Han et al., 2025)	0.12B	0.10M	1.70	0.697	1.40	0.751	6.46	0.704	3.19	0.717
MOSSTTS-Nano (OpenMOSS, 2026)	0.10B	1.00M	4.62	0.611	6.35	0.456	17.36	0.592	9.44	0.553
LITEMEL-Large (***Teacher***)	0.13B		**1.44**	0.699	**1.37**	**0.760**	**5.79**	0.729	**2.87**	0.729
LITEMEL-Base	0.06B		1.63	0.641	1.85	0.726	6.78	0.695	3.42	0.687
+OPD	0.06B	0.10M	1.48	0.655	1.40	0.740	6.19	0.718	3.02	0.704
LITEMEL-Small	0.03B		1.78	0.576	2.31	0.662	8.23	0.643	4.11	0.627
+OPD	0.03B		1.42	0.630	1.54	0.717	6.47	0.695	3.14	0.681

but are orders of magnitude larger; and (2) *Lightweight generative models* with parameter budgets comparable to ours (e.g., F5-TTS, E2-TTS, and MOSSTTS-Nano).

5 EXPERIMENTAL RESULTS

5.1 QUANTITATIVE EVALUATION

Objective evaluation. As shown in Table 1, we first evaluate LITEMEL against leading large-scale and lightweight zero-shot TTS models on Seed-TTS-eval. Despite using only 0.10M hours of training data, LITEMEL-Large (0.13B) achieves SOTA average intelligibility (2.89%) among all models, even outperforming billion-parameter systems such as dots.tts (2.92%) and CosyVoice3 (3.06%), while attaining competitive speaker fidelity, underscoring the extreme parameter and data efficiency of our architecture. Although a capacity gap remains in speaker similarity compared to larger systems like E2-TTS, this highlights a finding that model capacity constraints primarily bottleneck high-dimensional acoustic rather than linguistic content. Despite this intrinsic capacity floor, LITEMEL achieves highly competitive speaker similarity metrics while operating at a fraction of the parameters.

We further explore sub-0.1B models (0.06B and 0.03B). Even before post-training, our pre-trained backbones achieve lower or comparable average error against baselines 5–10$\times$ their size (e.g., F5-TTS and E2-TTS). Applying OPD from our 0.13B teacher unlocks significant further gains. It yields relative error reductions of 11.7% in 0.06B and 23.6% in 0.03B, while substantially boosting speaker similarity across all domains. Crucially, post-trained with OPD, our 0.03B model outperforms both F5-TTS and ZipVoice, demonstrating superior inherent parameter efficiency.

Subjective evaluation. To evaluate human-perceptual speech quality and speaker fidelity, we conduct double-blind subjective user studies using Comparative Mean Opinion Score (CMOS) for overall naturalness/intelligibility and Speaker Similarity MOS (SMOS) for voice similarity. We recruit 25 native bilingual evaluators to assess 30 randomly selected test utterances per benchmark. Evaluators rate pairwise CMOS on a 7-point scale from -3 (much worse) to $+3$ (much better) relative

Table 2: **CMOS and absolute SMOS evaluation results.** The second column lists the SMOS scores for baselines, while the top row shows SMOS scores for our variants. The bottom-right block contains the pairwise CMOS comparisons. All values are mean ± 95% confidence interval.

Baseline	SMOS↑	LiteMel Ablation Variants				
		Large (0.13B)	Base (0.06B)	Base+OPD	Small (0.03B)	Small+OPD
SMOS of our variants →		3.85±0.08	3.47±0.13	3.94±0.08	2.92±0.16	3.70±0.14
		CMOS ↑ (Ours vs. baseline)				
Ground Truth	3.59±0.15	**+0.57±0.36**	-0.57±0.64	**+1.14±0.79**	-1.43±0.70	**+0.05±0.65**
FireRedTTS-3 (1.7B)	3.98±0.08	-0.50±0.51	-0.14±0.68	**+0.29±0.52**	-1.31±0.64	-0.43±0.40
VoxCPM2 (2.0B)	3.88±0.11	**+0.62±0.53**	-0.94±0.55	**+0.56±0.47**	-1.79±0.55	-0.81±0.67
F5-TTS (0.34B)	3.66±0.13	**+0.71±0.28**	-1.71±0.25	**+0.71±0.75**	-1.57±0.70	-0.29±0.56
ZipVoice (0.12B)	3.64±0.09	**+0.62±0.36**	-0.44±0.84	**+0.56±0.24**	-0.56±0.93	-0.22±0.44

to the baseline in each row, and assign SMOS on an absolute 5-point scale with 0.5-point increments against reference speaker audio. All metrics are reported with 95% confidence intervals in Table 2.

As shown in Table 2, our OPD strategy delivers significant gains across all scales, boosting SMOS by up to a relative 26.7% and driving a massive +1.71 CMOS shift over Ground Truth for *Base*. Remarkably, with up to 33× parameter reduction, our compact *Base+OPD* and *Large* (0.13B) models achieve speaker similarity and naturalness competitive with or outperforming massive baselines like FireRedTTS-3 and VoxCPM2. Notably, both *Large* (+0.57) and *Base+OPD* (+1.14) achieve positive CMOS over Ground Truth, confirming superior naturalness.

5.2 Ablation Studies and Discussions

Impact of backbone components. To demonstrate the impact of our components under a strictly matched parameter budget (∼0.13B), Table 3 presents a component-wise incremental ablation study on Seed-TTS-eval: 1) Dual-Path DiT Backbone effectively enhances speaker identity modeling, boosting the average speaker similarity from 0.693 to 0.704; 2) Soft Monotonic Alignment serves as the pivotal engine for robust speech generation, drastically suppressing deletion and hallucination errors by cutting the average error rate from 5.03% to 3.37% and reducing `test-hard` CER by 4.44%; and 3) Speaker-aware AdaLN further refines global acoustic modulation without disrupting local text-frame alignment, yielding our full LiteMel which achieves the optimal synthesis performance with the lowest overall error rate and highest average similarity.

Table 3: **Ablation summary on Seed-TTS-eval.**

Ablation Component	Avg. Err↓	Avg. SIM↑
(1) Plain DiT Baseline	4.47	0.693
(2) + Dual-Path Backbone	5.03	0.704
(3) + Soft Monotonic Align	3.37	0.692
(4) + Speaker AdaLN	**3.11**	**0.705**

Impact of post-training mechanisms. To investigate the efficacy of post-training paradigms on speech generation, given our 0.13B teacher model, we evaluate offline KL distillation (Offline KD), online Group Relative Policy Optimization (GRPO), and our introduced On-Policy Distillation (OPD) across both 0.03B and 0.06B student models. As illustrated in Figure 5, Offline KD suffers from inherent exposure bias, leading to premature saturation in speaker similarity (SIM ≈ 0.637 for 0.03B). While GRPO introduces online feedback by explicitly optimizing composite scalar rewards based on WER and SIM, relying purely on coarse, sequence-level rewards without dense distributional guidance makes it challenging to navigate high-dimensional continuous speech representations. In contrast, our OPD seamlessly bridges on-policy state exploration with fine-grained teacher density guidance. By continuously aligning the student's real-time generation distribution against the teacher model, OPD eliminates exposure bias while providing stable, dense supervisory signals. Consequently, OPD achieves performance gains across both scales, yielding 8.6% and 2.5% relative SIM improvements alongside 18.8% and 10.2% relative WER reductions on 0.03B and 0.06B models, respectively. These empirical results confirm that OPD is an exceptionally effective post-training strategy, successfully harmonizing phonetic intelligibility and speaker identity preservation.

Representational analysis of text encoder. To evaluate text-to-speech alignment without explicit duration supervisors, we analyze predicted duration of our text encoder and mel-spectrograms across 0.5x, 1.0x, and 2.0x global durations. As shown in Figure 4, LiteMel achieves a clean, monotonic

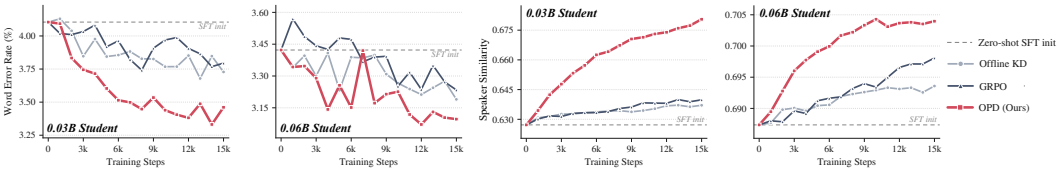


Figure 3: **Training trajectories of post-training paradigms** evaluated on 0.03B and 0.06B student models in terms of *WER*↓ (left) and *SIM*↑ (right). Our OPD (red curve) consistently outperforms Offline KD and GRPO, achieving superior performance.

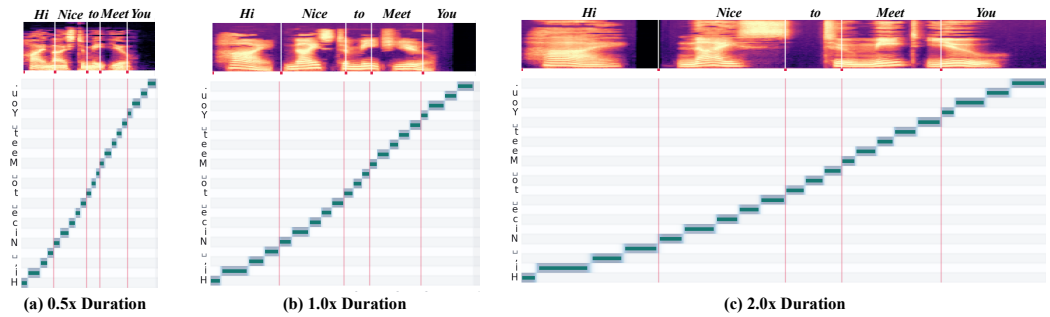


Figure 4: **Visualization of mel-to-character alignment.** It demonstrates that the text encoder implicitly learns a powerful text-speech prior purely optimized under the flow matching objective.

step-wise progression between characters and time frames, closely matching the acoustic energy boundaries in the corresponding spectrograms above. Given a fixed global duration, the encoder automatically distributes time steps across characters instead of relying on naive linear stretching, thereby assigning appropriate durations to distinct phonetic units such as vowels and consonants. Crucially, this robust text-speech prior emerges entirely from the continuous flow matching objective without requiring external duration predictors or monotonic alignment search (MAS), effectively simplifying the network architecture while ensuring reliable temporal control.

Efficiency analysis of lightweight systems. Table 4 summarizes the efficiency-quality trade-offs of lightweight speech synthesis systems. Compared to the 0.3B F5-TTS baseline, our LITEMEL models across varying scales (S/B/L) maintain competitive intelligibility (CER) and perceptual quality while offering substantially fewer parameters and a lower RTF at NFE = 32 or 16.

Table 4: **Efficiency&quality trade-off on `test-hard`.** Each cell lists NFE $= 32/16/8$ sampling steps. Real-Time Factor (RTF) is the median end-to-end real-time factor over all utterances.

Model	Params	NFE	CER (%)↓	SIM↑	UTMOS↑	RTF↓ (NFE=32)
LITEMEL-S	0.03B	32/16/8	7.12 / 7.09 / 13.14	0.6925 / 0.6934 / 0.6527	2.736 / **2.825** / **2.499**	**0.27**
LITEMEL-B	0.06B	32/16/8	6.33 / 6.32 / 11.71	0.7183 / 0.7181 / 0.6701	**2.749** / **2.825** / 2.459	0.33
LITEMEL-L	0.13B	32/16/8	**5.79** / **5.80** / **11.26**	0.7292 / 0.7293 / 0.6575	2.735 / 2.777 / 2.113	0.48
F5-TTS	0.3B	32/16/8	7.56 / 7.70 / 13.76	**0.7357** / **0.7386** / **0.6923**	2.696 / 2.773 / 2.098	0.49

6 CONCLUSION

In this work, we investigate architectural inductive bias and post-training for the lightweight TTS system. We show that when parameter capacity is constrained, generation quality is primarily limited by the degradation of acoustic quality and trajectory drift during inference. To address these challenges, LITEMEL incorporates a dual-path DiT architecture that balances intelligibility with speaker fidelity, and a learned soft monotonic aligner trained end-to-end via the flow-matching objective alone. To further mitigate inference drift, we introduce OPD to correct out-of-distribution states along the student's own stochastic rollouts, providing an effective trajectory correction mechanism without requiring complex reinforcement learning setups. Extensive evaluations demonstrate that LITEMEL achieves an improved trade-off between model scale and generation quality, establishing a competitive Pareto frontier against larger baselines.

AI USE STATEMENT

Required disclosures. We utilized generative AI tools as coding assistants for results analysis. All AI-assisted scripts, analytical findings, and text were thoroughly audited, verified, and finalized by the authors.

Not used. No component of the released models is AI-generated. The architecture of Section 3.2, the on-policy distillation objective of Section 3.3 and the derivation in Appendix A are the authors' own, and generative AI tools were not used to formulate the mathematical claims or to produce the derivation. We did not use such tools to generate, augment, filter or relabel any training or evaluation data. Our corpus is the public release of Emilia (He et al., 2024) and our benchmarks are Seed-TTS-eval and LibriSpeech-PC *test-clean*, all used as distributed, with no model-generated transcripts, captions or synthetic speech added.

Recommended disclosures. We additionally used generative AI tools to draft and edit portions of this paper for readability, and to produce and refine tables to the conference template.

Review and responsibility. The authors take full responsibility for all content, including any text or code that originated with AI assistance. All literature references were checked against the original publications.

ETHICS STATEMENT

Zero-shot TTS reproduces a speaker's voice from a few seconds of audio, so this work carries impersonation and consent risks that a benchmark or a discriminative model would not. We acknowledge four classes of consideration.

Consent and provenance of training data. Our models are trained on Emilia-v2 (He et al., 2024), a publicly released corpus derived from in-the-wild recordings. We did not collect new speech, did not target any individual speaker, and release no voice prints or speaker-identifiable artifacts beyond what that corpus already contains. We note the limitation honestly: consent for the original recordings was not obtained for voice-cloning research specifically, which is a property of the public corpora the field currently trains on rather than of this work alone.

Speaker privacy. The prompt audio in our evaluations comes from the public benchmark distributions. We do not attempt speaker re-identification, and the synthesized audio we release is limited to the fixed benchmark prompts, not to arbitrary target voices.

Human subjects. The subjective scores were collected from compensated raters under a standard research-annotation agreement. Participation was voluntary, raters were told in advance that the audio might be synthetic, and no personal information about them was retained. Appendix E gives the recruitment, screening and instruction protocol.

REPRODUCIBILITY STATEMENT

Section 4 states the corpus, front-end, optimizer schedule, batching, update counts, sampler settings and metric implementations for every arm. Appendix C gives the per-size architectural geometry and the full hyperparameter table, including the duration head and its soft clamp, the time-step distribution, the guidance dropout rates and the sampling schedule. Appendix **??** explains the parameter-counting convention, without which our counts cannot be reproduced; Appendix **??** the real-time-factor protocol and what it includes; Appendix B the measured corpus statistics; Appendix A the full derivation of the reweighted velocity gap of Eq. (7); and Appendix E the verbatim rater instructions. The alignment mechanism of Section 3.2 needs no forced aligner, duration labels or external supervision, so the data pipeline it requires is the corpus as distributed. Both distillation arms at each size start from the same supervised checkpoint, so the post-training comparison can be reproduced without retraining a teacher. Code, configurations and checkpoints will be released.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, et al. Seed-tts: A family of high-quality versatile speech generation models. *CoRR*, 2024.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16 (6):1505–1518, 2022.

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, 2025.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, 2024.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *CoRR*, 2025.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. E2 TTS: embarrassingly easy fully non-autoregressive zero-shot TTS, 2024.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pp. 2063–2067, 2022.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 30467–30476, 2026.

Wenhao Guan, Kaidi Wang, Wangjin Zhou, Yang Wang, Feng Deng, Hui Wang, Lin Li, Qingyang Hong, and Yong Qin. LAFMA: A latent flow matching model for text-to-audio generation. In *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*, 2024.

Zhu Han, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2025, Honolulu, HI, USA, December 6-10, 2025*, pp. 1–8, 2025.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2-5, 2024*, pp. 885–890, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, 2022.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. DiffusionOPD: A unified perspective of on-policy distillation in diffusion models. *CoRR*, 2026.

Shi Lian, Changtao Li, Bohan Li, Hankun Wang, Da Zheng, Junfeng Tian, Yufeng Ma, Colin Zhang, and Kai Yu. dots.tts technical report. *CoRR*, abs/2606.07080, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *18th Annual Conference of the International Speech Communication Association, Interspeech 2017, Stockholm, Sweden, August 20-24, 2017*, pp. 498–502, 2017.

OpenMOSS. Moss-tts-nano. GitHub repository, 2026. URL https://github.com/OpenMOSS/MOSS-TTS-Nano.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4172–4182, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351, pp. 234–241, 2015.

Feiyu Shen, Kun Xie, Yichen Wu, Ziqi Dai, Yichen Han, Junjie Li, Xuelong Geng, Fenglong Xie, Lei Xie, Xu Tang, and Yao Hu. Fireredtts3: Unified speech generation and editing with semantically enriched speech representations. *CoRR*, 2026.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Qwen Team. Qwen3-tts technical report. *CoRR*, 2026.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. CAM++: A fast and efficient network for speaker verification using context-aware masking. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pp. 5301–5305, 2023.

Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, Yingya Zhang, and Hongming Shan. Dreamvideo-2: Zero-shot subject-driven video customization with precise motion control. *CoRR*, 2024.

Kun Xie, Feiyu Shen, Junjie Li, Fenglong Xie, Xu Tang, and Yao Hu. Fireredtts-2: Towards long conversational speech generation for podcast and chatbot. *CoRR*, 2025.

Detai Xin, Shujie Hu, Chengzuo Yang, Chen Huang, Guoqiao Yu, Guanglu Wan, and Xunliang Cai. Longcat-audiodit: High-fidelity diffusion text-to-speech in the waveform latent space. *CoRR*, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *CoRR*, 2025.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. Zipformer: A faster and better encoder for automatic speech recognition. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

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. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 35139–35148, 2026a.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Jiancheng Gui, et al. Voxcpm2 technical report. *CoRR*, 2026b.

A Derivation of the Per-Step KL

We derive the coefficient $w(t_k)$ of Eq. (7) from the transition kernels induced by the discretisation of Eq. (5). Three assumptions are used, all of which hold in our implementation: the student and the teacher share the sampler, so their per-step kernels share the covariance $g^2\Delta t_k \boldsymbol{I}$ at every step; the noise scale g is constant over the rollout; and the discretisation is Euler–Maruyama, so each step is exactly Gaussian.

One Gaussian Step. In our convention $t = 0$ is noise, $t = 1$ is data, the interpolant is $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1$ and the backbone predicts $\boldsymbol{v} = \boldsymbol{x}_1 - \boldsymbol{x}_0$. The conditional expectation of the injected noise is then

$$\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t] = \boldsymbol{x}_t - t\,\boldsymbol{v}(\boldsymbol{x}_t, t, \boldsymbol{c}), \tag{8}$$

and the score of the marginal follows from the Gaussian conditional,

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) = -\frac{\boldsymbol{x}_t - t\,\boldsymbol{v}(\boldsymbol{x}_t, t, \boldsymbol{c})}{1-t}. \tag{9}$$

The marginal-preserving SDE is $\mathrm{d}\boldsymbol{x} = [\boldsymbol{v} + \frac{g^2}{2}\nabla \log p_t(\boldsymbol{x})]\mathrm{d}t + g\,\mathrm{d}\boldsymbol{w}$, whose Fokker–Planck equation reproduces the ODE's continuity equation $\partial_t p = -\nabla\cdot(\boldsymbol{v}p)$, so the marginals are unchanged.

The Step Kernel. Substituting the score and taking one Euler–Maruyama step of size Δt_k,

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_{t_k}, t_k) = \underbrace{\left[1 - \frac{g^2\Delta t_k}{2(1-t_k)}\right]}_{\alpha_k}\boldsymbol{x}_{t_k} + \underbrace{\left[1 + \frac{g^2 t_k}{2(1-t_k)}\right]}_{c(t_k)}\Delta t_k\,\boldsymbol{v}_\theta(\boldsymbol{x}_{t_k}, t_k, \boldsymbol{c}), \tag{10}$$

so $p_\theta(\cdot \mid \boldsymbol{x}_{t_k}) = \mathcal{N}\big(\boldsymbol{\mu}_\theta, g^2\Delta t_k \boldsymbol{I}\big)$.

The KL Collapses to the Velocity Gap. For two Gaussians of equal covariance $\boldsymbol{\Sigma}$, $\mathrm{KL}(p_\theta \| p_\phi) = \frac{1}{2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_\phi)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_\phi)$. The state-dependent term $\alpha_k \boldsymbol{x}_{t_k}$ in Eq. (10) is identical for both models and therefore cancels in the difference, which is what leaves the coefficient multiplying the velocity gap alone:

$$\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_\phi = c(t_k)\,\Delta t_k\,\big[\boldsymbol{v}_\theta(\boldsymbol{x}_{t_k}, t_k, \boldsymbol{c}) - \boldsymbol{v}_\phi(\boldsymbol{x}_{t_k}, t_k, \boldsymbol{c})\big]. \tag{11}$$

Substituting $\boldsymbol{\Sigma} = g^2\Delta t_k \boldsymbol{I}$ and simplifying,

$$\mathrm{KL}\big(p_\theta \,\|\, p_\phi\big) = \frac{c(t_k)^2\,\Delta t_k}{2g^2}\big\|\boldsymbol{v}_\theta(\boldsymbol{x}_{t_k}, t_k, \boldsymbol{c}) - \boldsymbol{v}_\phi(\boldsymbol{x}_{t_k}, t_k, \boldsymbol{c})\big\|^2 = w(t_k)\,\big\|\boldsymbol{v}_\theta - \boldsymbol{v}_\phi\big\|^2. \tag{12}$$

Two consequences are worth stating. First, the coefficient carries no hyperparameter beyond g: the target is the same squared velocity gap that offline distillation minimises (Eq. (6) with teacher-forced states), so the two objectives differ in the state distribution alone. Second, because the covariance cancels analytically, the estimator carries no sampling noise — unlike a score-matching or REINFORCE estimator of the same divergence, whose variance would have to be traded against the step size.

B Data Details

The Chinese and English portions of Emilia (He et al., 2024) measure 47,034,483 utterances over 118,066 hours, mean utterance length 9.04.

Filtering. Two criteria are applied, both on the metadata that ships with the release rather than on audio we re-scored. We drop utterances whose speaker annotation marks more than one speaker, since a mel target spanning two voices is not a valid regression target for a single-speaker conditional model and would inject speaker leakage into the prompt. We also drop utterances flagged as transcription failures, which otherwise contribute text that does not correspond to the audio and would teach the aligner a schedule unrelated to what is spoken. Neither criterion uses an external model, so the filtering we apply is reproducible from the release alone and introduces no model-generated labels. Approximately 100K hours over 46M utterances survive.

Signal representation. Audio is resampled to 24 kHz and represented as 100-bin log-mel spectrograms with a 1024-point window and a 256-sample hop, the configuration expected by the Vocos vocoder (Siuzdak, 2024) used at synthesis time. No loudness normalisation or bandwidth extension is applied beyond resampling, so the mel targets retain the level statistics of the source recordings.

Text normalisation. Text is tokenised with a pinyin-based grapheme inventory shared across both languages: Latin characters are consumed one at a time, while Chinese characters are converted to toned pinyin with tone sandhi applied, so a character expands to a short sequence of tokens rather than one. The inventory is fixed before training and shared by all arms in this paper. Because the inventory contains a token for the space, pauses are represented in the token sequence and receive a learned duration like any other token.

C Implementation Details

Table 5: Geometry of the three sizes. "Fast" is the isotropic branch, "slow" the U-shaped branch (depths per stage, three resolutions), "fusion" the joint head. Parameter counts are from `parameters()` with the deployed vocabulary; Appendix **??** explains why summing `state_dict()` entries instead inflates them. The 30M model retains the teacher's depth profile at half the width, while the 60M model also trims depth.

Model	Shared	Fast (width × depth)	Slow depths	Fusion	Params
LiteMel-L	256	512 × 13	[2, 2, 3, 2, 2]	256 × 5	129.8M
LiteMel-B	192	384 × 11	[2, 2, 2, 2, 2]	192 × 4	60.4M
LiteMel-S	128	256 × 13	[2, 2, 3, 2, 2]	128 × 5	32.8M

Table 5 gives the three configurations. All widths are channel dimensions and all depths are DiT block counts; *heads* counts are per branch and heads × dim_head equals the branch width in every case. The three sizes were not produced by one width sweep: LiteMel-B also trims depth, which is worth knowing when reading the scaling trend of Sec. 5.1.

The duration head is the whole of the alignment mechanism's parameter cost: it adds 0.264M parameters, i.e. 0.20% of LiteMel-L, and is applied identically at training and inference. Its final layer is zero-initialised so that the proportions of Eq. (5) start uniform and the schedule departs from that prior gradually rather than from a random one. The soft clamp is the algebraic form $x/\sqrt{1+(x/c)^2}$ rather than a hyperbolic tangent: it is smooth everywhere, its gradient vanishes at saturation instead of flattening to a constant, and it is exactly linear for small x, which preserves the ranking of small logit differences. The clamp value $c = 3.0$ was introduced as part of a gradient-spike fix and is used by every arm reported here.

D Extended Architecture Ablations

Table 7 isolates the estimator topology of Sec. 3.2 at a matched parameter budget and matched updates.

E Subjective Evaluation

The subjective scores reported in Sec. 5 were collected in a double-blind listening study. Evaluators were recruited from a bilingual pool, screened with a short headphone check and two obvious-quality anchors before their ratings were accepted, and compensated at a fixed hourly rate. Participation was voluntary, no personal information about them was retained, and every evaluator was told in advance that the audio might be synthetic.

Protocol. Each session presented 30 randomly selected utterances per benchmark, with system identity hidden and the left/right or A/B ordering randomised per item. Comparative scores were collected on a 7-point scale from −3 (much worse) to +3 (much better) against the baseline named in each row of the results table. Absolute naturalness and speaker similarity were collected on 5-point scales in 0.5-point increments, the latter against the reference speaker audio rather than against the

Table 6: Shared hyperparameters. Values are identical across the three sizes except where a column is given; the learning rates differ because each size was tuned at its own rung. The duration head is the only component added by the alignment mechanism.

Component	Key	Value
Backbone	text dimension	512 / 384 / 256
	shared encoder width	256 / 192 / 128
	shared conv layers	4
	feed-forward multiplier	2
Duration head	architecture	Linear → SiLU → LayerNorm → Linear
	hidden width	equals the text dimension (512/384/256)
	output	one logit per token
	soft clamp	$x/\sqrt{1+(x/c)^2}$, $c = 3.0$
	final layer init	zero weight and bias
Flow matching	time distribution	logit-normal ($\mu = 0$, $\sigma = 1$)
	time shift	3.0
	masking length	60%–100% of the utterance
Optimisation	learning rate	1×10^{-4}
	warmup	8000 updates
	optimiser	AdamW, grad-norm clip 1.0
	dropout	0.1 in all branches and the fusion head
Mel front-end	sample rate	24 kHz
	mel channels	100
	hop / window / FFT	256 / 1024 / 1024
	vocoder	Vocos (24 kHz)

Table 7: **Architecture ablation on Seed-TTS-eval.** All model variants are trained for 4 epochs on the Emilia dataset under a matched parameter budget. **Bold** marks the best performance for each metric. UTMOS is averaged across all three test sets.

Ablation Component	test-en		test-zh		test-hard		UTMOS↑	Average	
	WER↓	SIM↑	CER↓	SIM↑	CER↓	SIM↑		Err↓	SIM↑
(1) Plain DiT Baseline	1.75	0.648	1.64	0.735	10.02	0.692	**3.25**	4.47	0.693
(2) + Dual-Path DiT Backbone	1.89	0.661	1.58	0.739	11.63	**0.713**	2.98	5.03	0.704
(3) + Soft Monotonic Alignment	1.52	0.653	**1.41**	0.732	7.19	0.692	3.09	3.37	0.692
(4) + Speaker AdaLN (LITEMEL)	**1.45**	**0.678**	1.51	**0.735**	**6.38**	0.703	3.10	**3.11**	**0.705**

target recording. Repeated anchor items were interleaved as attention checks; ratings from a session that failed them were discarded. All figures are reported as means with 95% confidence intervals.

System pool. Ratings cover our three sizes both before and after post-training, the ground-truth recordings, and the published baselines. Because the pool spans systems differing by more than an order of magnitude in parameter count, the comparison is reported as a preference matrix rather than as a single ranking.

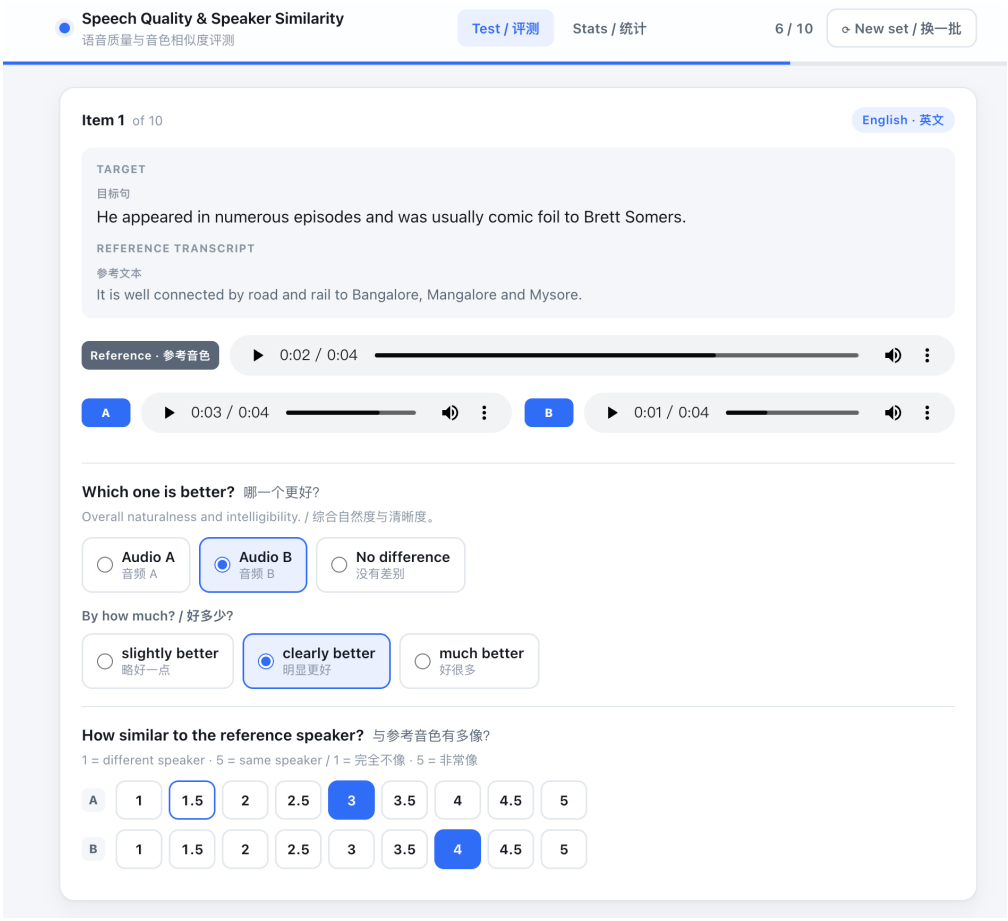


Figure 5: Screenshot of the web-based subjective evaluation interface for speech quality and speaker similarity assessment. Evaluators compare paired audio samples (Audio A and Audio B) against the target reference text and audio in terms of overall naturalness, preference margin, and speaker similarity scores.