

M3-TTS: Native Two-Stream Multi-Modal Diffusion Transformer for Zero-Shot Speech Synthesis

Xiaopeng Wang^{1,†} Chunyu Qiang^{2,†} Ruibo Fu^{3,*} Zhengqi Wen⁴
Yuxin Guo³ Xuefei Liu³ Yukun Liu³ Yuzhe Liang² Kang Yin²
Teng Ma² Yushen Chen² Yuankun Xie⁵ Chenxing Li³ Changsheng Li^{1,*}

¹School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

²Kling AI, Kuaishou Technology, Beijing, China

³Institute of Automation, Chinese Academy of Sciences, Beijing, China

⁴Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing, China

⁵Department of EEE, The Hong Kong Polytechnic University, Hong Kong SAR, China

ruibofu@126.com lcs@bit.edu.cn

Abstract

Zero-shot text-to-speech (TTS) must generate speech from text while preserving the speaker characteristics of a short reference utterance. Recent non-autoregressive systems improve decoding efficiency, but they often bridge the text–speech length mismatch with duration predictors, filler-token padding, or text upsampling, and inject reference speech through sequence-level conditioning. We propose M3-TTS, a multi-modal, multi-stream, and modulation-based diffusion Transformer for reference-conditioned masked acoustic generation. Its core design is a native two-stream architecture, termed Self-Adaptive Joint Representation (SAJR), where text tokens and speech representations keep their original lengths and interact through packed joint self-attention. M3-TTS further uses a joint-to-audio generation pipeline to separate cross-modal correspondence learning from acoustic refinement, together with AdaLN-style acoustic prompting to inject speaker information in a unified acoustic representation space. Experiments show that the FBank-based M3-TTS reaches 1.52% WER on English and 1.35% CER on Chinese in the 24 kHz Seed-TTS comparison, while the Mel-VAE variant reduces RTF to 0.12. Code and demos will be available at <https://github.com/WWWxp/M3-TTS>.

1 Introduction

Zero-shot text-to-speech (TTS) synthesis aims to generate a target utterance from text while preserving speaker characteristics from a short reference clip. Recent systems have substantially improved naturalness and speaker transfer, but efficient zero-shot synthesis remains difficult because the model

must learn text–speech alignment, speaker conditioning, and acoustic generation in one process. Autoregressive (AR) models (Du et al., 2024a,b; Wang et al., 2025c,b; Qiang et al., 2024) generate speech sequentially and can capture rich prosodic variation (Wang et al., 2025a; Anastassiou et al., 2024), but token-by-token decoding limits inference efficiency and can accumulate generation errors. Non-autoregressive (NAR) models (Chen et al., 2025; Zheng et al., 2025; Zhu et al., 2025; Eskimez et al., 2024; Qiang et al., 2025b) improve decoding speed by generating acoustic trajectories in parallel. Their main challenge is that linguistic tokens and acoustic frames have different native lengths, so the model needs a way to exchange information across modalities without forcing one modality into the time scale of the other.

Existing NAR systems usually handle this mismatch through an auxiliary alignment structure. Duration-predictor-based models (Ren et al., 2019, 2021; Kim et al., 2021; Kong et al., 2023; Casanova et al., 2022) first estimate how long each linguistic unit should last, while recent diffusion or flow-based systems use weaker surrogates such as alignment preprocessing (Mehta et al., 2024), filler-token padding (Chen et al., 2025; Eskimez et al., 2024), or average text upsampling (Zhu et al., 2025). These designs make optimization tractable, but they convert native text tokens into frame-level or padded sequences before the main generator models text–speech interaction. As a result, the generator inherits redundant tokens, additional computation, or a predefined alignment bias. Another line injects linguistic or acoustic information through feature concatenation (Arik et al., 2017; Jia et al., 2018) or cross-attention (Wang et al., 2018;

[†]Equal contribution. ^{*}Corresponding authors.

Le et al., 2023). Such conditioning is expressive, but it treats one modality as an external condition to the other rather than allowing both modalities to participate symmetrically in the same attention space. These observations motivate a NAR TTS architecture in which text and speech retain their native resolutions, interact directly, and leave the remaining model capacity for acoustic refinement.

To this end, we introduce M3-TTS, a non-autoregressive framework built on a dual-stage multi-modal diffusion Transformer (MMDiT). The three M’s in M3-TTS denote **M**ulti-modal text–speech modeling, native **M**ulti-stream interaction, and **M**odulation-based acoustic prompting. Rather than converting text into a frame-level scaffold, M3-TTS models zero-shot synthesis as acoustic representation generation conditioned jointly on target text and reference acoustic context. It keeps text tokens and speech representations as two native streams and connects them through a Self-Adaptive Joint Representation (SAJR), where Joint-DiT layers pack valid tokens from both streams into one self-attention sequence while preserving modality-specific embeddings, projections, and positional structure. Audio-only DiT layers then refine the speech trajectory, and an AdaLN-style acoustic prompting interface injects speaker information in the acoustic representation space without expanding the generation sequence. In our implementation, this space can be instantiated by either FBank features or compact Mel-VAE latents, enabling us to study the efficiency–quality trade-off of different acoustic representations. On Seed-TTS, M3-TTS improves intelligibility in the 24 kHz comparison, while the Mel-VAE setting improves inference efficiency. This work makes three primary contributions:

- We propose M3-TTS, a native two-stream MMDiT architecture that preserves text and speech at their original lengths and enables SAJR-based packed joint self-attention without filler-token padding or text upsampling.
- We introduce a joint-to-audio generation pipeline, where Joint-DiT layers learn text–speech correspondence and audio-only DiT layers refine the speech trajectory for acoustic consistency.
- We formulate zero-shot TTS as reference-conditioned masked acoustic generation and use AdaLN-style acoustic prompting to inject

frame-level reference cues without increasing the generation sequence length.

2 Methodology

2.1 Overview of M3-TTS

M3-TTS is a NAR framework built around native two-stream acoustic generation. The model first maps speech into an acoustic representation space, instantiated as either FBank features or compact Mel-VAE latents, and then predicts a conditional flow in that space with a dual-stage Diffusion Transformer (DiT). The first stage contains Joint-DiT layers, where text tokens and speech representations remain separate streams with modality-specific embeddings, projections, and positional encodings, but their valid tokens are packed into one attention sequence for cross-modal interaction. The second stage contains audio-only DiT layers, where only the speech stream is refined under masked reference representations and the diffusion timestep. This design lets text guide alignment without forcing text tokens to be expanded to the speech frame rate, while keeping a dedicated refinement stage for acoustic consistency.

Figure 1 summarizes the training and inference pipelines. During training, M3-TTS uses a single utterance: the transcript is encoded into text tokens, the speech waveform is converted into acoustic representations, and a random target span is masked. The model learns to infer the masked span from the text stream, the unmasked speech representations, and the diffusion timestep. During inference, the model receives reference speech, its transcript, and target text. The reference speech provides prompt representations for speaker and acoustic context, while the concatenated reference and target transcript supplies the linguistic condition for the prompt-plus-target trajectory. This formulation matches the zero-shot setting used in evaluation, where reference transcripts are available in the test set and are assumed to be manually transcribed in normal use.

2.2 Multi-modal Diffusion Transformer Architecture

The core generation engine of M3-TTS is a dual-stage Diffusion Transformer that explicitly separates native-length cross-modal interaction from audio-only acoustic refinement.

Overview of M3-TTS

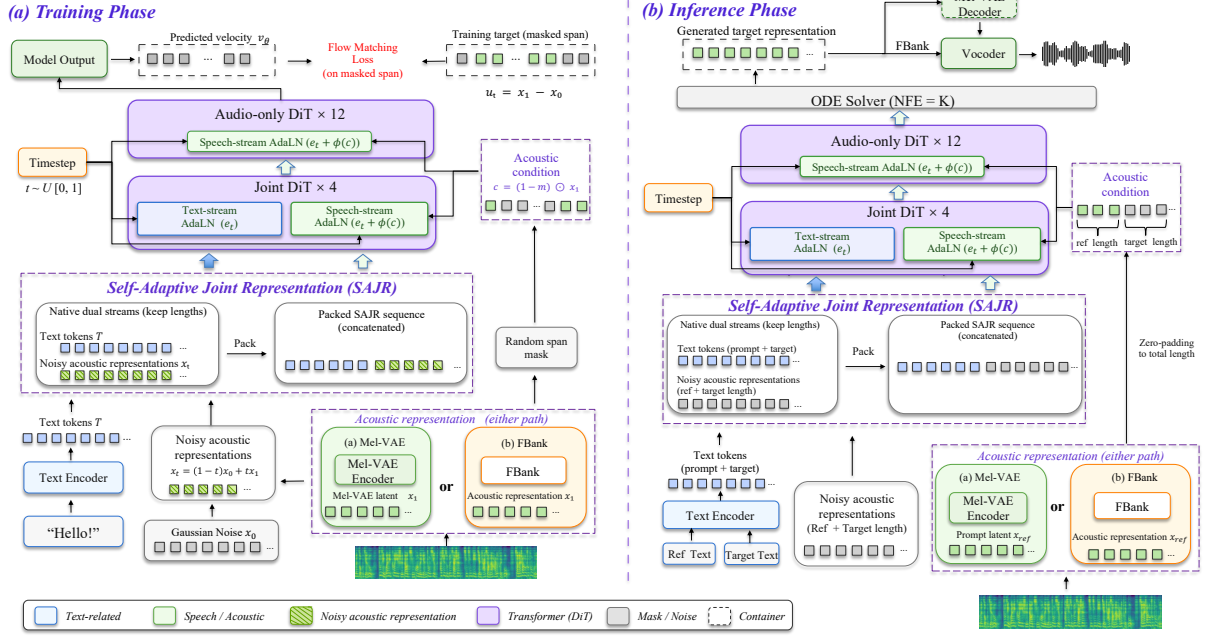


Figure 1: Overview of M3-TTS: training (left) and inference (right).

2.2.1 Self-Adaptive Joint Representation for Alignment

The Self-Adaptive Joint Representation (SAJR) realizes native two-stream interaction. Let $T \in \mathbb{R}^{T_t \times D}$ denote text tokens and $A \in \mathbb{R}^{T_s \times D}$ denote speech representations, where T_t and T_s are allowed to differ. Before attention, the two streams are embedded separately and projected by modality-specific QKV layers. For each sample, Joint-DiT packs only the valid text tokens and speech representations into a unified attention sequence $Z = \text{Concat}_{\text{time}}(T, A) \in \mathbb{R}^{(T_t + T_s) \times D}$ and applies shared scaled dot-product attention over this packed sequence. RoPE is applied separately to the text and speech streams before packing, preserving within-modality order without forcing the two modalities onto one absolute positional axis. This design gives speech representations a direct pathway to query linguistic tokens while retaining modality-specific parameterization.

Unlike duration-predictor-based systems, SAJR does not first expand text tokens to the acoustic frame rate. Unlike channel-concatenation or padding-based designs, it also does not require text and speech to share the same temporal length before entering the Transformer. The two streams keep their native lengths and are packed as $[\text{valid}(T), \text{valid}(A)]$ for each sample. In each Joint-DiT layer, each speech representation

can attend to all valid text tokens and neighboring speech representations through the same self-attention operator, so soft text–speech correspondence is learned during denoising. After attention, the packed output is unpacked back into text and speech branches; only the speech branch is passed to the audio-only refinement stage. This design uses text tokens where cross-modal alignment is needed, then removes the text stream when the model only needs to refine acoustic details.

2.2.2 AdaLN-style Acoustic Prompting

To inject reference acoustic information without appending condition tokens, M3-TTS uses an AdaLN-style acoustic prompting interface. This component is a conditioning mechanism rather than a new normalization operator: it converts masked acoustic representations into modulation parameters for the speech stream. Given the masked acoustic condition $c = (1 - m) \odot x_1$, M3-TTS first projects it into the model dimension and combines it with the timestep embedding to form the speech-side modulation condition.

In Joint-DiT, the text branch is modulated only by the timestep embedding, since text tokens should not be directly conditioned on acoustic frames. The speech branch is modulated by the combination of the timestep and masked acoustic condition. After the packed sequence is unpacked, the audio-only DiT keeps only the speech branch

and applies the same speech-side AdaLN conditioning.

The modulation follows the AdaLN pattern used in DiT blocks. We denote the text-side and speech-side modulation conditions as

$$c_{\text{text}} = e_t, \quad c_{\text{sp}} = \phi(c) + e_t, \quad (1)$$

where e_t is the timestep embedding and $\phi(c)$ is the projected masked acoustic condition. For a hidden sequence H and condition $c_\star \in \{c_{\text{text}}, c_{\text{sp}}\}$, a small modulation network predicts shift, scale, and residual gate parameters for both the attention and feed-forward sublayers:

$$(\Delta, \Sigma, G, \Delta_{\text{ff}}, \Sigma_{\text{ff}}, G_{\text{ff}}) = \text{MLP}_{\text{mod}}(c_\star). \quad (2)$$

The attention sublayer is conditioned as

$$\begin{aligned} \tilde{H} &= \text{LN}(H) \odot (1 + \Sigma) + \Delta, \\ H' &= H + G \odot \text{Attn}(\tilde{H}), \end{aligned} \quad (3)$$

and the feed-forward sublayer is conditioned as

$$\begin{aligned} \bar{H} &= \text{LN}(H') \odot (1 + \Sigma_{\text{ff}}) + \Delta_{\text{ff}}, \\ H_{\text{out}} &= H' + G_{\text{ff}} \odot \text{FFN}(\bar{H}). \end{aligned} \quad (4)$$

This formulation matches the implementation: the scale term is applied as $(1 + \Sigma)$, and the gates control the residual contribution of the attention and feed-forward outputs. Its advantage is architectural. Compared with sequence concatenation, AdaLN-style prompting does not increase the generation length; compared with cross-attention, it avoids an additional condition-attention branch. The ablation in Section 4.3.2 therefore evaluates this interface against concatenation and cross-attention baselines.

2.2.3 Audio-only Acoustic Refinement

Following the initial cross-modal interaction, the packed sequence is unpacked back into speech and text streams. The audio-only DiT stage keeps only the speech branch H^a and refines it with self-attention and feed-forward layers modulated by the speech-side condition c_{sp} . This stage produces the final vector field $v_\theta(\cdot)$ for flow matching. Discarding the text stream after the joint layers reduces the cost of iterative sampling and reflects a division of labor: Joint-DiT layers establish text–speech correspondence, while audio-only layers focus on local acoustic consistency and timbre refinement.

2.3 Acoustic Representation Space

M3-TTS performs flow matching in an acoustic representation space rather than directly in the waveform domain. In our implementation, this space can be instantiated in two forms: FBank features or compact Mel-VAE latents. The FBank setting uses directly extracted acoustic features as the generation target and serves as the main configuration in the Seed-TTS comparison. The Mel-VAE path follows the VQ-CTAP design (Qiang et al., 2025a), where the encoder projects log-mel filterbanks into compact acoustic latents $A \in \mathbb{R}^{T_s \times D}$.

Compared with FBank-based modeling, the Mel-VAE representation provides $2\times$ temporal compression and $2.5\times$ dimensional compression, reducing the feature dimension from 100 to 40. The diffusion model is trained and sampled in the selected acoustic representation space. When the FBank setting is used, the generated acoustic features are used for waveform synthesis through the vocoder interface. When the Mel-VAE path is used, the Mel-VAE decoder maps the generated latent sequence back to acoustic features before waveform synthesis. Therefore, the Mel-VAE path serves as an efficiency-oriented instantiation of the acoustic representation space, lowering the cost of Transformer interaction and supporting the Mel-VAE setting evaluated in Section 4.2.

2.4 Learning and Inference via Flow Matching

The model is optimized with Conditional Flow Matching (CFM). During training, we sample Gaussian noise $x_0 \sim \mathcal{N}(0, I)$ and use the clean acoustic representation x_1 as the data endpoint. The interpolation path is

$$x_t = (1 - t)x_0 + tx_1, \quad t \in [0, 1], \quad (5)$$

and the target velocity is $u_t = x_1 - x_0$. Here, x_t is the noisy speech-stream input, $c = (1 - m) \odot x_1$ is the masked acoustic condition for modulation, and u_t is used only as the flow-matching target. Given the masked acoustic condition c , text tokens T , and timestep t , the dual-stage DiT predicts $\hat{u}_\theta = v_\theta(x_t, c, T, t)$. The objective is the masked mean squared error over the target span:

$$\begin{aligned} \mathcal{L}(\theta) &= \mathbb{E}_{x_0, x_1, t, m} \left[\frac{1}{\sum_i m_i} \sum_i m_i \ell_i \right], \\ \ell_i &= \|\hat{u}_{\theta, i} - u_{t, i}\|_2^2. \end{aligned} \quad (6)$$

For inference, the total prompt-plus-target length is estimated via a proportional duration heuristic based on the reference utterance. We solve the ordinary differential equation $\dot{x}_t = v_\theta(x_t, c, T, t)$ from Gaussian noise with a fixed-step solver using 32 steps, restore the prompt region from the reference condition, and crop the target region for decoding.

Training and inference procedure The training procedure follows an utterance-level infilling objective. Given an utterance with transcript y and speech s , the text encoder produces $T = \text{Enc}_{\text{text}}(y)$ and the acoustic encoder produces the clean target $x_1 = \text{Enc}_{\text{ac}}(s)$, where Enc_{ac} denotes either FBank extraction or Mel-VAE encoding. A random span mask m covers a continuous span of target frames, and the unmasked representations $(1 - m) \odot x_1$ are used as the acoustic condition. We sample x_0 from a standard Gaussian distribution, sample $t \sim \mathcal{U}(0, 1)$, construct $x_t = (1 - t)x_0 + tx_1$, and train the dual-stage DiT to predict the velocity $x_1 - x_0$ on the masked target region. This objective exposes the model to incomplete acoustic observations, making the remaining speech representations function as speaker and style context while the text stream supplies the missing linguistic content.

At inference time, the reference and target transcripts are concatenated as the linguistic condition, while the reference acoustic representation is padded to the estimated total length and used as the acoustic condition only on the prompt region. Starting from Gaussian noise over the full prompt-plus-target length, M3-TTS integrates the learned ODE with a fixed-step solver using 32 steps and classifier-free guidance, restores the reference region from the condition, and keeps only the generated target segment. If the Mel-VAE path is used, the Mel-VAE decoder maps generated latents back to acoustic features; otherwise, the generated representation is already the FBank feature sequence. A high-fidelity vocoder based on BigVGAN then converts the acoustic features into waveform samples. This pipeline keeps cross-modal generation in the acoustic representation space and invokes the vocoder only after sampling has finished.

3 Experimental Setup

Implementation Details We implement M3-TTS with a 16-layer MMDiT backbone. The backbone uses 4 Joint-DiT layers for SAJR and 12 audio-only DiT layers for acoustic refinement. Each layer uses a hidden dimension of 1024, 16 atten-

Table 1: Algorithmic summary of M3-TTS.

Algorithm 1: M3-TTS

Note: Enc_{ac} denotes either FBank extraction or Mel-VAE encoding.

Training: paired sample (τ, s)

- 1 Encode $T = \text{Enc}_{\text{text}}(\tau)$ and $x_1 = \text{Enc}_{\text{ac}}(s)$.
 - 2 Sample a random span mask m and set $c = (1 - m) \odot x_1$.
 - 3 Sample x_0, t ; set $x_t = (1 - t)x_0 + tx_1$ and $u_t = x_1 - x_0$.
 - 4 Predict $\hat{u}_\theta = v_\theta(x_t, c, T, t)$ with 4 Joint-DiT and 12 audio-only DiT layers.
 - 5 Update θ using the masked CFM loss in Eq. 6.
-

Inference: $(s_{\text{ref}}, \tau_{\text{ref}}, \tau_{\text{tar}})$

- 1 Encode $T = \text{Enc}_{\text{text}}([\tau_{\text{ref}}, \tau_{\text{tar}}])$ and $x_{\text{ref}} = \text{Enc}_{\text{ac}}(s_{\text{ref}})$.
 - 2 Estimate length L ; pad x_{ref} to length L and zero the target region to form c .
 - 3 Sample x_0 over length L and solve $\dot{x}_t = v_\theta(x_t, c, T, t)$ with fixed-step sampling using 32 steps and CFG.
 - 4 Restore reference frames, crop the target segment, apply the Mel-VAE decoder if needed, and vocode.
-

tion heads, and RMS-normalized queries and keys. Text is represented with a learned token embedding over a custom English–Chinese pinyin vocabulary and encoded by the F5-TTS text encoder (Chen et al., 2025), which uses a 4-layer ConvNeXtV2 stack with 512-dimensional token embeddings and a 1024-dimensional feed-forward dimension. The model is trained on $8 \times \text{A100}$ GPUs with a learning rate of 7.5×10^{-5} . Pre-training uses the Emilia corpus (He et al., 2024), which contains approximately 95k hours of English and Chinese speech.

Datasets and Baselines We evaluate zero-shot synthesis on Seed-TTS (Anastassiou et al., 2024) test-en/test-zh at 24 kHz. This setting is used for the main comparison with recent AR systems, including CosyVoice (Du et al., 2024a), CosyVoice 2 (Du et al., 2024b), and FireRedTTS (Guo et al., 2024), and recent NAR systems, including E2-TTS (Es-kimez et al., 2024), F5-TTS (Chen et al., 2025), and ZipVoice (Zhu et al., 2025). Unless otherwise specified, M3-TTS denotes the FBank-based acoustic representation used in the main Seed-TTS comparison. We additionally compare M3-TTS with FBank features and M3-TTS with Mel-VAE latents under the same vocoder in Section 4.2, in order to isolate the efficiency–quality trade-off introduced by compact latent generation.

Metrics We report objective and subjective metrics that cover intelligibility, speaker preservation, naturalness, and efficiency. Intelligibility is mea-

sured by WER for English and CER for Chinese, computed with Whisper-large-v3 (Radford et al., 2023) and Paraformer-zh (Gao et al., 2022), respectively. Speaker similarity is measured by SIM-o using ECAPA-TDNN (Desplanques et al., 2020) embeddings. Naturalness is estimated with UTMOS (Saeki et al., 2022), and subjective NMOS/QMOS are collected through double-blind listening tests. Runtime efficiency is reported as real-time factor (RTF) under the corresponding generation setting.

4 Experimental Results

The experimental results are organized from overall effectiveness to component analysis. Section 4.1 first compares M3-TTS with recent AR and NAR zero-shot TTS systems on Seed-TTS. Section 4.2 then examines the efficiency–quality trade-off introduced by the Mel-VAE acoustic representation. Section 4.3 evaluates alignment and acoustic-conditioning alternatives.

4.1 Overall Comparison

Table 2 compares M3-TTS with recent AR and NAR systems on Seed-TTS. M3-TTS obtains the lowest English WER among the evaluated systems, reducing WER to 1.52%, while maintaining competitive speaker similarity and naturalness with 0.690 SIM-o and 3.80 UTMOS. On the Chinese subset, M3-TTS achieves the best CER of 1.35%, the highest SIM-o of 0.764, and the highest UTMOS of 3.16. These results indicate that the proposed native two-stream architecture improves intelligibility while maintaining competitive speaker preservation.

Table 3 reports subjective evaluation separately because listening-test scores are available only for a subset of systems. M3-TTS reaches an NMOS of 3.77 and a QMOS of 3.95, achieving comparable or slightly better subjective quality than the compared NAR systems within the reported confidence intervals. Since the overall comparison does not isolate individual components, the following analyses separately examine the Mel-VAE representation, AdaLN-style acoustic prompting, and learned joint attention patterns.

4.2 Effect of Mel-VAE Latent Representation

Figure 2 isolates the effect of the acoustic representation. To avoid conflating representation effects with waveform generation differences, all three systems in this comparison use the same BigVGAN-based vocoder. The Mel-VAE variant reduces RTF

to 0.12, compared with 0.23 for M3-TTS with FBank features and 0.31 for F5-TTS. It also improves Seed-TTS error rates to 1.37% / 1.32% WER/CER on English/Chinese, with speaker similarity of 0.64 / 0.70. The trade-off is that the Mel-VAE setting reduces Seed-TTS speaker similarity relative to the FBank variant. We therefore interpret the Mel-VAE latent space as an efficiency-oriented representation: it improves speed and intelligibility in our evaluation, while trading off speaker similarity.

4.3 Ablation Studies

Unless otherwise specified, all ablation variants use the same training data, acoustic representation, vocoder, sampling steps, CFG setting, and Seed-TTS evaluation protocol as the corresponding M3-TTS configuration. We report RTF to reflect the computational overhead introduced by different alignment or conditioning mechanisms.

4.3.1 Ablation on Native Two-stream Alignment

To isolate the effect of SAJR, we compare the proposed native two-stream alignment with common alternatives for bridging text and speech lengths. The text-upsampled variant expands text representations to the acoustic-representation length before generation, the filler-token variant follows the padding strategy used by flow-based NAR systems, and the cross-attention variant treats text as an external condition to the speech stream. Table 4 reports English WER, Chinese CER, and RTF under the same Seed-TTS evaluation protocol. This ablation directly tests whether preserving native text and speech lengths in packed joint self-attention provides a better intelligibility–efficiency trade-off than explicit length matching or external text injection.

4.3.2 Ablation on Acoustic Condition Injection

We compare four ways of injecting speaker and acoustic information into the DiT backbone, as illustrated in Figure 3. Sequence concatenation directly appends the acoustic condition to the input sequence together with text and target speech representations, which preserves frame-level cues but increases the generation sequence length. Cross-attention keeps the main DiT input as text and target speech representations, while introducing an additional attention branch to inject the acoustic con-

Table 2: Objective evaluation on Seed-TTS at 24 kHz. Bold indicates the best system result within each metric; Ground Truth is included as a reference rather than a model baseline.

Model	Type	English			Chinese		
		SIM-o $\uparrow$	WER $\downarrow$	UTMOS $\uparrow$	SIM-o $\uparrow$	CER $\downarrow$	UTMOS $\uparrow$
Ground Truth	–	0.734	2.14	3.52	0.755	1.25	2.78
CosyVoice	AR	0.609	4.29	–	0.723	3.63	–
CosyVoice 2	AR	0.652	2.57	–	0.748	1.45	–
FireRedTTS	AR	0.463	3.82	–	0.634	1.51	–
E2-TTS	NAR	0.706	2.32	3.21	0.713	1.91	2.26
F5-TTS	NAR	0.664	1.85	3.72	0.750	1.53	2.93
ZipVoice	NAR	0.697	1.70	3.82	0.751	1.40	3.15
M3-TTS	NAR	0.690	1.52	3.80	0.764	1.35	3.16

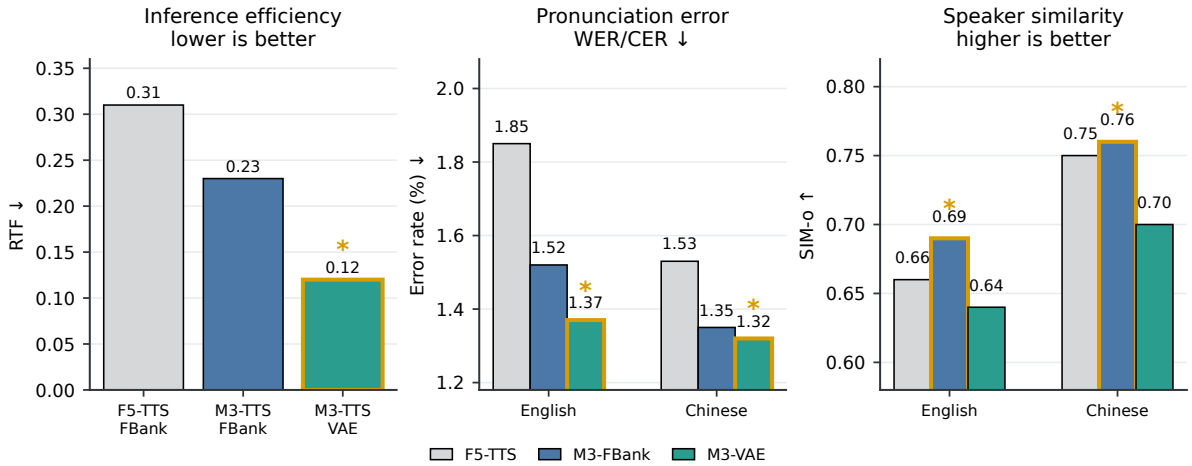


Figure 2: Representation trade-off among F5-TTS with FBank features, M3-TTS with FBank features, and M3-TTS with Mel-VAE latents. Gold outlines and stars mark the best result in each panel or language group. The Mel-VAE setting improves inference efficiency and error rates, while reducing Seed-TTS speaker similarity relative to the FBank variant.

Table 3: Subjective evaluation on Seed-TTS. Scores are reported with confidence intervals.

Model	NMOS $\uparrow$	QMOS $\uparrow$
Ground Truth	3.83 ± 0.09	3.99 ± 0.08
F5-TTS	3.72 ± 0.19	3.87 ± 0.17
ZipVoice	3.75 ± 0.14	3.94 ± 0.12
M3-TTS	3.77 ± 0.12	3.95 ± 0.11

Table 4: Ablation of text–speech alignment mechanisms under the same Seed-TTS evaluation protocol. EN and ZH report WER and CER, respectively.

Alignment	EN $\downarrow$	ZH $\downarrow$	RTF $\downarrow$
Text upsampling	1.74	1.44	0.34
Filler-token padding	1.83	1.52	0.37
Text cross-attention	1.66	1.43	0.52
SAJR (native two-stream)	1.52	1.35	0.23

dition. CAM++ global conditioning first extracts an utterance-level speaker embedding from the reference speech using the CAM++ model released in 3D-Speaker (Wang et al., 2023a; Zheng et al., 2024), and then injects this global vector through AdaLN modulation. In contrast, M3-TTS uses the masked acoustic condition as frame-level prompting and injects it through AdaLN, avoiding both sequence expansion and additional cross-attention layers while retaining detailed reference acoustic cues. This ablation tests whether global speaker

conditioning is sufficient for zero-shot acoustic prompting, or whether frame-level masked speech representations provide more useful speaker and style information. Table 5 shows that AdaLN-style prompting achieves the highest SIM scores, 0.690 / 0.764 on English/Chinese, and the lowest RTF of 0.23. Sequence concatenation increases the conditioning sequence and yields the highest RTF of 0.39, while cross-attention reduces RTF to 0.36 but remains slower than modulation-based prompting.

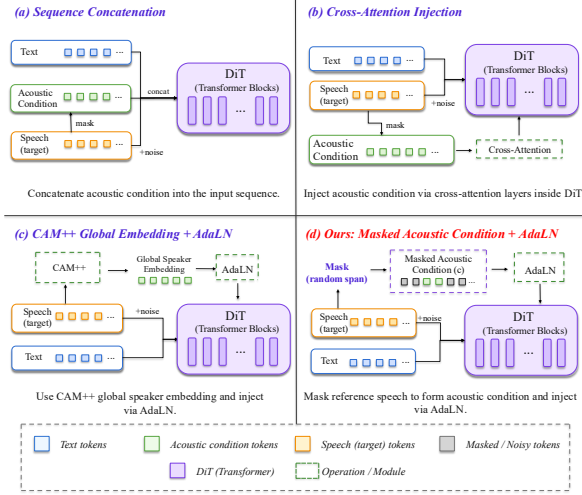


Figure 3: Illustration of four acoustic condition injection mechanisms compared in the ablation study.

CAM++ global conditioning improves over cross-attention in SIM and reduces RTF to 0.29, but still trails AdaLN-style prompting, suggesting that a single utterance-level speaker vector loses useful frame-level prompt information.

Table 5: Quantitative comparison of acoustic condition injection mechanisms on SIM and RTF. CAM++ denotes a global speaker embedding extracted from the reference speech by the 3D-Speaker CAM++ model.

Injection Method	SIM (EN / ZH) $\uparrow$	RTF $\downarrow$
Sequence Concatenation	0.659 / 0.753	0.39
Cross-attention	0.661 / 0.755	0.36
CAM++ global embedding	0.674 / 0.759	0.29
AdaLN-style prompting	0.690 / 0.764	0.23

4.4 Joint Attention Visualization

Figure 4 visualizes representative cross-modal attention maps from Joint-DiT. The x -axis corresponds to speech representations and the y -axis corresponds to text tokens; darker pixels indicate higher attention weights, and red dots mark the row-wise argmax for each text token. The maps show a predominantly monotonic diagonal structure with limited off-diagonal drift. This qualitative pattern suggests that SAJR learns an implicit text–speech correspondence in the packed attention space, without relying on explicit duration supervision or text upsampling.

5 Conclusions

We present M3-TTS, a NAR framework that combines native two-stream text–speech interaction,

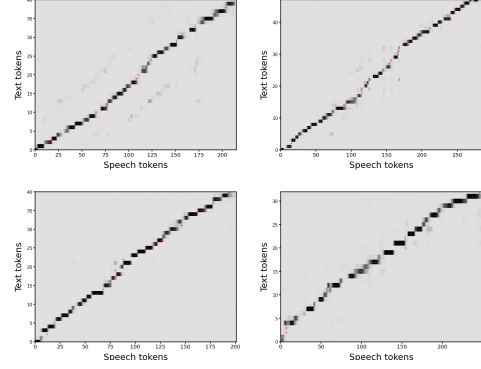


Figure 4: Visualization of cross-modal attention within Joint-DiT across four test samples. The x -axis represents speech representations, while the y -axis represents text tokens. Higher attention weights are indicated by darker pixel intensities, and red dots signify the row-wise argmax for each text token.

AdaLN-style acoustic prompting, and acoustic-representation-based generation within a dual-stage MMDiT. SAJR preserves the original lengths of text tokens and speech representations, and enables implicit cross-modal correspondence through packed joint self-attention. After joint interaction, audio-only DiT layers refine the speech trajectory, while modulation-based acoustic prompting injects frame-level reference cues without increasing the generation sequence length. Experiments on Seed-TTS show that M3-TTS improves intelligibility while maintaining competitive speaker similarity and naturalness. Further representation and conditioning ablations show that compact Mel-VAE latents provide an efficiency-oriented acoustic representation, and that AdaLN-style prompting offers a favorable quality–efficiency trade-off compared with sequence concatenation, cross-attention, and global speaker embedding-based conditioning.

6 Limitations

M3-TTS still has several limitations. First, although the experiments cover standard zero-shot TTS benchmarks and component-level ablations, the evaluation is still mainly conducted on Seed-TTS-style short utterances. Broader validation on more languages, expressive speaking styles, noisy references, and long-form generation would further verify the robustness of the proposed native two-stream design. Second, the current inference pipeline assumes that the reference transcript is available and estimates the prompt-plus-target length with a proportional duration heuristic. This may limit robustness when reference transcripts

are inaccurate or when the target utterance contains unusual speaking rates, pauses, or highly expressive prosody. Third, the compact Mel-VAE representation improves inference efficiency and pronunciation accuracy, but it reduces speaker similarity compared with the FBank variant, suggesting that stronger speaker-preserving acoustic compression is still needed. Future work will explore long-form generation and more speaker-preserving latent acoustic representations.

7 Generative AI Usage Disclosure

Following the conference guidelines regarding artificial intelligence, the authors declare the use of large language models during the preparation of this manuscript. These tools were employed strictly for the purposes of language refinement, grammatical correction, and enhancing the overall fluency and readability of the text. We affirm that all scientific hypotheses, architectural designs, experimental data analyses, and core conclusions are entirely the original work of the human authors and were not generated by AI.

References

- Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, and 27 others. 2024. Seed-TTS: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.
- Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep Voice 2: Multi-speaker neural text-to-speech. In *Advances in neural information processing systems*, volume 30.
- Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir Antonelli Ponti. 2022. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *ICLR*.
- Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *Proc. ACL*.
- Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN-based speaker verification. In *Interspeech*.
- Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024a. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.
- Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024b. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.
- 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. 2024. E2 TTS: Embarrassingly easy fully non-autoregressive zero-shot TTS. In *SLT*.
- Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Proc. Interspeech 2022*, pages 2063–2067.
- Hao-Han Guo, Yao Hu, Kun Liu, Fei-Yu Shen, Xu Tang, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kai-Tuo Xu. 2024. FireRedTTS: A foundation text-to-speech framework for industry-level generative speech applications. *arXiv preprint arXiv:2409.03283*.
- Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *SLT*.
- Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems*, volume 31.
- Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.
- Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *ICLR*.
- Jungil Kong, Jihoon Park, Beomjeong Kim, Jeongmin Kim, Dohee Kong, and Sangjin Kim. 2023. VITS2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design. In *Interspeech*.

- Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Proc. NeurIPS*.
- Keon Lee, Dong Won Kim, Jaehyeon Kim, Seungjun Chung, and Jaewoong Cho. 2024. DiTTo-TTS: Diffusion transformers for scalable text-to-speech without domain-specific factors. *arXiv preprint arXiv:2406.11427*.
- Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2024. Matcha-TTS: A fast TTS architecture with conditional flow matching. In *Proc. ICASSP*.
- Chunyu Qiang, Wang Geng, Yi Zhao, Ruibo Fu, Tao Wang, Cheng Gong, Tianrui Wang, Qiuyu Liu, Jiangyan Yi, Zhengqi Wen, Chen Zhang, Hao Che, Longbiao Wang, Jianwu Dang, and Jianhua Tao. 2025a. VQ-CTAP: Cross-modal fine-grained sequence representation learning for speech processing. *IEEE Transactions on Audio, Speech and Language Processing*, 33.
- Chunyu Qiang, Hao Li, Hao Ni, He Qu, Ruibo Fu, Tao Wang, Longbiao Wang, and Jianwu Dang. 2024. Minimally-supervised speech synthesis with conditional diffusion model and language model: A comparative study of semantic coding. In *ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10186–10190. IEEE.
- Chunyu Qiang, Kang Yin, Xiaopeng Wang, Yuzhe Liang, Jiahui Zhao, Ruibo Fu, Tianrui Wang, Cheng Gong, Chen Zhang, Longbiao Wang, and Jianwu Dang. 2025b. InstructAudio: Unified speech and music generation with natural language instruction. *arXiv preprint arXiv:2511.18487*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICLR*.
- Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.
- 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.
- Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. *Proc. Interspeech 2022*.
- Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proc. ICASSP*.
- Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2025a. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*.
- Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023a. CAM++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*.
- Wenbin Wang, Yang Song, and Sanjay Jha. 2023b. Generalizable zero-shot speaker adaptive speech synthesis with disentangled representations. *arXiv preprint arXiv:2308.13007*.
- Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfu Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, and 6 others. 2025b. Spark-TTS: An efficient LLM-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.
- Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2025c. MaskGCT: Zero-shot text-to-speech with masked generative codec transformer. *International Conference on Learning Representations*.
- Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*, pages 5180–5189. PMLR.
- Detai Xin, Shujie Hu, Chengzuo Yang, Chen Huang, Guoqiao Yu, Guanglu Wan, and Xunliang Cai. 2026. LongCat-AudioDiT: High-fidelity diffusion text-to-speech in the waveform latent space. *arXiv preprint arXiv:2603.29339*.
- Qixi Zheng, Yushen Chen, Zhikang Niu, Ziyang Ma, Xiaofei Wang, Kai Yu, and Xie Chen. 2025. Accelerating Flow-Matching-Based Text-to-Speech via Empirically Pruned Step Sampling. In *Interspeech 2025*, pages 2445–2449.
- Siqi Zheng, Luyao Cheng, Yafeng Chen, Hui Wang, and Qian Chen. 2024. 3D-Speaker: A large-scale multi-device, multi-distance, and multi-dialect corpus for speech representation disentanglement. *arXiv preprint arXiv:2306.15354*.
- Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiwei Zhuang, Long

Lin, and Daniel Povey. 2025. ZipVoice: Fast and high-quality zero-shot text-to-speech with flow matching. *ASRU*.

A Related Work

A.1 Zero-shot Text-to-Speech Synthesis

Zero-shot TTS aims to synthesize speech for unseen speakers from a short reference utterance. Autoregressive systems such as VALL-E (Wang et al., 2025a), Seed-TTS (Anastassiou et al., 2024), CosyVoice (Du et al., 2024a), CosyVoice 2 (Du et al., 2024b), and FireRedTTS (Guo et al., 2024) model speech generation sequentially and can capture rich prosodic variation, but their decoding process is inherently less efficient for long-form generation. Recent non-autoregressive systems, including E2-TTS (Eskimez et al., 2024), F5-TTS (Chen et al., 2025), and ZipVoice (Zhu et al., 2025), improve inference speed by generating acoustic trajectories in parallel. However, efficient NAR generation places more pressure on the model to learn text–speech alignment and speaker conditioning without the sequential structure used by AR models. M3-TTS follows the NAR direction, but differs from prior systems by treating alignment, acoustic conditioning, and compact speech representation as coupled design choices within one diffusion Transformer.

A.2 Alignment Modeling in Non-autoregressive TTS

Text–speech alignment is a central difficulty in NAR TTS because the text sequence and acoustic sequence have different lengths. Earlier systems often rely on explicit duration modeling or monotonic alignment components (Ren et al., 2019, 2021; Kim et al., 2021; Kong et al., 2023; Casanova et al., 2022), which provide a useful temporal scaffold but require a predefined alignment pathway. More recent diffusion and flow-based NAR systems reduce this dependency, yet they still commonly introduce pseudo-alignment surrogates. F5-TTS and E2-TTS use filler-token padding to bring text closer to the acoustic length (Chen et al., 2025; Eskimez et al., 2024), while ZipVoice uses average text upsampling before generation (Zhu et al., 2025); related systems also use alignment-oriented preprocessing (Mehta et al., 2024). These designs make training tractable, but they add redundant tokens or impose a predefined length conversion before the main generator.

Another line avoids explicit duration predictors by injecting text into the speech generator through cross-attention. DiTTo-TTS uses a DiT backbone with cross-attention between encoded text and speech representations, together with a speech length predictor (Lee et al., 2024). LongCat-AudioDiT extends diffusion TTS to waveform latent generation and represents a recent large-scale AudioDiT direction (Xin et al., 2026). Cross-attention is expressive, but it still treats text as an external condition to the speech stream rather than as a native stream participating in the same self-attention space. In contrast, M3-TTS uses SAJR to pack valid text tokens and speech representations into a shared self-attention sequence while preserving their native lengths. This formulation moves the alignment burden from a preprocessing rule or a separate condition-attention branch to the model’s joint attention mechanism.

A.3 Acoustic Conditioning and Latent Speech Representations

Zero-shot TTS also requires acoustic conditioning that preserves speaker identity while avoiding leakage from the reference content. Common conditioning strategies include feature concatenation (Arik et al., 2017; Jia et al., 2018), cross-attention-based injection (Wang et al., 2018; Le et al., 2023), and global speaker embeddings extracted by speaker verification models such as CAM++ (Wang et al., 2023a; Zheng et al., 2024). These approaches are effective, but direct use of fine-grained acoustic features may entangle speaker, content, and prosodic information (Ju et al., 2024; Wang et al., 2023b), while a single global embedding may discard frame-level acoustic cues useful for prompt-conditioned generation. M3-TTS addresses this issue with AdaLN-style acoustic prompting, which injects reference information through affine modulation rather than by expanding the generation sequence. Another related line of work studies compact speech representations for generative modeling. High-dimensional FBank features contain detailed acoustic information but increase the cost of cross-modal interaction, while latent or compressed representations can reduce redundancy (Shen et al., 2018; Ju et al., 2024; Qiang et al., 2025a).