

DiTReducio: A Training-Free Acceleration for DiT-Based TTS via Progressive Calibration

Yanru Huo^{1,2}, Ziyue Jiang¹, Zuoli Tang³, Qingyang Hong², Zhou Zhao^{1*}

¹Zhejiang University, ²Xiamen University, ³Wuhan University

hyrrrr0661@xmu.edu.cn

zhaozhou@zju.edu.cn

Abstract

While Diffusion Transformers (DiT) have advanced non-autoregressive (NAR) speech synthesis, their high computational demands remain a limitation. Existing DiT-based text-to-speech (TTS) model acceleration approaches mainly focus on reducing sampling steps through distillation techniques, yet they remain constrained by training costs. We introduce DiTReducio, a training-free acceleration framework that compresses computations in DiT-based TTS models via progressive calibration. We propose two compression methods, **Temporal Skipping** and **Branch Skipping**, to eliminate redundant computations during inference. Moreover, based on two characteristic attention patterns identified within DiT layers, we devise a pattern-guided strategy to selectively apply the compression methods. Our method allows flexible modulation between generation quality and computational efficiency through adjustable compression thresholds. Experimental evaluations conducted on F5-TTS and MegaTTS 3 demonstrate that DiTReducio achieves a 75.4% reduction in FLOPs and improves the Real-Time Factor (RTF) by 37.1%, while preserving generation quality. The code is available at <https://github.com/iHateTheWorld555/DiTReducio>.

1 Introduction

Recent advances in TTS synthesis have enabled the generation of highly realistic and natural speech, with applications including virtual assistants, audiobooks, and digital avatars. The autoregressive (AR) models (Chen et al., 2024a; Du et al., 2024; Zhou et al., 2024; Wang et al., 2025b; Deng et al., 2025; Zhou et al., 2025; Peng et al., 2025) and NAR TTS (Mehta et al., 2024; Wang et al., 2024; Zhu et al., 2025; Huynh-Nguyen et al., 2025) have demonstrated robust zero-shot capabilities, especially those based on DiT (Ju et al., 2024; Chen

et al., 2024b; Eskimez et al., 2024; Jiang et al., 2025; Qiang et al., 2025) achieve accelerated inference while maintaining audio generation quality through high-performance computational parallelization. These benefits have facilitated their widespread adoption in real-world applications.

Despite their benefits, DiT-based models fundamentally face architectural limitations. While effectively capturing long-range dependencies, the Transformer’s self-attention mechanism results in quadratic time and space complexity. This computational burden is further exacerbated by the multi-step denoising process and Classifier-Free Guidance (CFG) techniques (Ho and Salimans, 2022) employed in these models. While some lightweight DiT-based TTS systems have achieved efficient inference, their computational requirements still exceed the limits of on-device deployment and real-time interactive applications, highlighting the need for inference acceleration methods.

Prior works address these computational challenges through three main approaches: (1) optimizing sampling using advanced samplers or distillation to reduce inference steps (Lu et al., 2022a,b; Salimans and Ho, 2022); (2) applying model compression techniques such as quantization and pruning (Shang et al., 2023; Wan et al., 2025); and (3) implementing attention mechanism optimizations, such as sparse attention (Zaheer et al., 2020; Hassani et al., 2023) and token-wise methods (Bolya and Hoffman, 2023; Saghatchian et al., 2025) that may face compatibility issues with FlashAttention (Dao et al., 2022). While numerous innovations in efficient inference (Yuan et al., 2024; Zhang et al., 2025b,a) have emerged in image and video generation, comparable advancements in TTS remain limited, highlighting a critical research gap.

Our goal is to develop a training-free approach for quality-controllable acceleration in DiT-based TTS models. Inspired by prior observations of computational redundancies in the DiT architec-

*Corresponding author

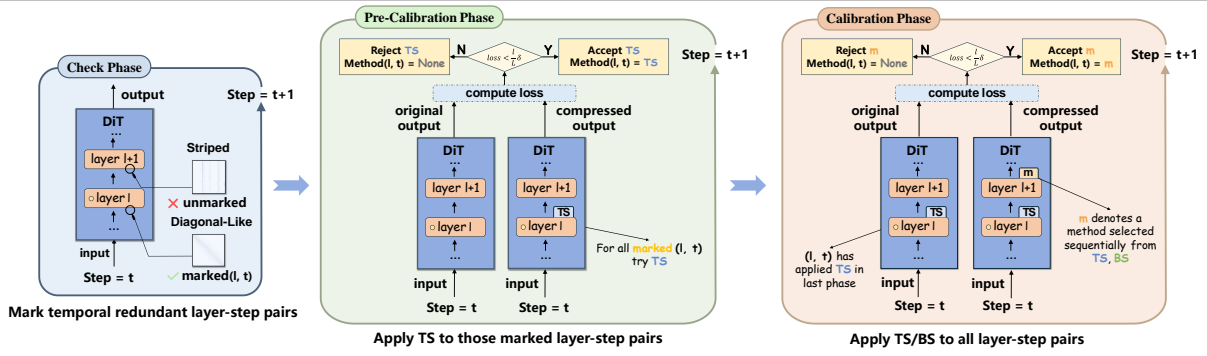


Figure 1: **Overview of DiTReducio.** In the Check Phase, we identify a subset of highly temporally redundant layer-step pairs by detecting diagonal-like attention patterns. In the Pre-Calibration Phase, we apply TS to those identified pairs and retain only those for which the resulting output loss remains below a dynamical threshold. Finally, in the Calibration Phase, both TS and BS are applied across all layer-step pairs under the same loss constraint. This procedure yields a model-specific inference acceleration strategy.

tures (Yuan et al., 2024), we systematically investigate two key phenomena.

Firstly, through detailed analysis of DiT architectures, we identify two notable forms of computational redundancy manifested during model inference at specific layer-timestep combinations (referred to as layer-step pairs below): **Temporal Redundancy**, found across adjacent timesteps, and **Branch Redundancy**, existing between conditional and unconditional branches. We then introduce two tailored strategies: **Temporal Skipping (TS)** and **Branch Skipping (BS)**, to exploit these redundancies by reusing cached computations.

To further investigate the underlying mechanisms of these redundancies, we analyze attention heatmaps in specific layer-step pairs and uncover two distinct patterns, diagonal-like patterns and striped patterns. The diagonal-like patterns are characterized by tokens primarily attending to their neighboring tokens, demonstrate high temporal redundancy, Correspondingly, we conjecture that the striped patterns represent a mechanism for global information aggregation. In contrast, striped patterns appear to establish dependencies across the entire sequence. Our empirical analysis indicates these patterns are less redundant and crucial for generation quality.

Building on these insights, in this paper, we propose DiTReducio, a systematic progressive calibration framework for efficient DiT-based TTS inference. As Figure 1 shows, the framework operates through the following sequential phases:

1. **Check Phase:** Identifying layer-step pairs exhibiting diagonal-like attention patterns as highly temporally redundant.

2. **Pre-Calibration Phase:** Applying TS selectively to the marked layer-step pairs.

3. **Calibration Phase:** Building upon the pre-calibrated model, applying both TS and BS across all layer-step pairs while preserving generation quality.

Experimental results demonstrate that our approach can achieve a 1.6 \times improvement in RTF while maintaining controllable generation quality. As a training-free and plug-and-play solution, DiTReducio can seamlessly integrate with existing acceleration methods. Also, the framework’s adaptability in balancing acceleration and quality preservation makes it particularly advantageous for large-scale TTS deployments.

2 Related Work

2.1 DiT-based Speech Synthesis

Diffusion Models have emerged as a powerful paradigm in speech synthesis, offering high-fidelity generation distinct from autoregressive approaches. With the introduction of DiT (Peebles and Xie, 2023), diffusion-based speech synthesis has entered a new and promising phase. These models (Lee et al., 2024; Eskimez et al., 2024; Chen et al., 2024b; Jiang et al., 2025; Du et al., 2025) fully exploit the structural parallelism of Transformer architectures within the latent diffusion framework, achieving efficient training and inference, along with robust zero-shot speech generation. This dual advantage of computational efficiency and reliable generation has made them well-suited for integration with multi-modal large

language models (Xu et al., 2025; AI et al., 2025) in practical applications.

2.2 Acceleration of Diffusion Model

Despite the excellent performance of diffusion models in generation tasks (Xie et al., 2025), their inherent multi-step denoising nature leads to high computational costs, limiting their deployment in real-time interactive scenarios.

Current mainstream acceleration methods primarily include model distillation, sampler optimization, quantization, and pruning. Among these, model distillation (Salimans and Ho, 2022; Sauer et al., 2024) techniques transfer the capabilities of complex teacher models to lightweight student models and reduces sampling steps, which is also the dominant acceleration strategy in speech synthesis (Huang et al., 2022; Li et al., 2024; Guan et al., 2024; Zhu et al., 2025; Wang et al., 2025a). However, such methods require additional training overhead and depend on the performance of the teacher model. Improvements to samplers (Lu et al., 2022a,b) optimize noise schedules and achieve generation quality comparable to that of thousand-step sampling with significantly fewer sampling steps. These methods have the advantage of being training-free and are relatively mature. While quantization (Li et al., 2023; He et al., 2023) and pruning (Castells et al., 2024) can improve inference efficiency, they lead to unpredictable generation quality degradation.

Some works in image and video generation focus on optimizing the attention mechanism and propose token-wise acceleration algorithms (Bolya et al., 2022; Kong et al., 2022; Xing et al., 2024). However, these methods require training a token selector or calculating attention heatmaps at each inference step, which causes compatibility issues with efficient computation libraries such as FlashAttention, and their high implementation complexity limits practical applications.

While training-free and plug-and-play acceleration has flourished in vision, comparable advancements remain limited in speech synthesis. Closest to our work are DiTFastAttn (Yuan et al., 2024) and BlockDance (Zhang et al., 2025b). However, DiTFastAttn employs a greedy search strategy mainly focusing on attention modules, which tends to yield suboptimal compression strategies. BlockDance operates at a coarse granularity (layer-skipping) and introduces an auxiliary reinforcement learning network, which is even larger than

some lightweight TTS models, thus failing to be strictly training-free. In contrast, DiTReducio targets both attention and feed-forward modules via a fine-grained layer-step pruning strategy. By introducing a pattern-guided pre-calibration phase, it avoids the local optima inherent in greedy methods while remaining training-free.

3 Method

3.1 Overview

In this section, we present the methodology of DiTReducio. We systematically investigate the **Temporal Redundancy** and the **Branch Redundancy**, and introduce corresponding compression methods (Sec. 3.2). We then analyze the attention patterns within DiT layers and demonstrate their correlation with temporal redundancy across layer-step pairs (Sec. 3.3). Finally, we propose DiTReducio, a progressive calibration framework that iteratively explores compression opportunities through three inference passes (Sec. 3.4). The framework records effective acceleration strategies based on observed redundancy, enabling plug-and-play deployment in DiT-based speech synthesis.

3.2 Compression Methods

Redundant computations during diffusion model inference pose a major bottleneck to inference speed. Previous studies (Ma et al., 2024a; Zhao et al., 2024) have shown that diffusion models for image and video generation exhibit substantial temporal redundancy across multiple timesteps. In our work, we demonstrate that diffusion models for speech synthesis similarly exhibit extensive temporal redundancy and, under CFG, also display measurable branch redundancy.

Temporal redundancy Temporal redundancy refers to a high similarity between the outputs of a given module at adjacent timesteps during model inference. Figure 7 in Appendix A.3 shows cosine similarity heatmaps for F5-TTS and MegaTTS 3 reveal that outputs from both attention and feed-forward modules remain highly consistent across adjacent timesteps.

Based on this, we introduce the **Temporal Skipping** (TS) strategy. As illustrated in Figure 2, TS caches an output from a specific module at the preceding timestep and reuses it in subsequent steps to avoid temporally redundant computation. Formally, let O_t denote the output of a module at timestep t .

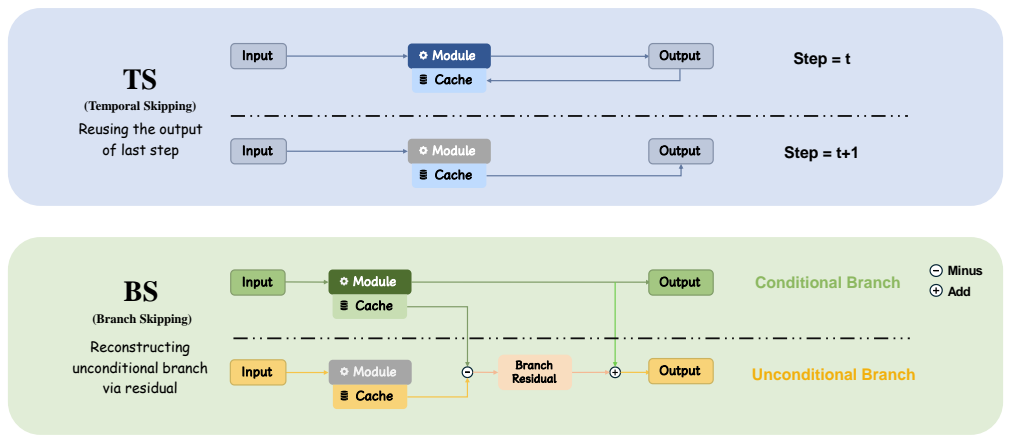


Figure 2: **Comparison of the workflows of TS, BS.** The cache is updated only when TS is not applied by the corresponding module.

Under TS, we have:

$$O_t = O_{t-1}. \quad (1)$$

Branch redundancy Branch redundancy arises in CFG-based diffusion models when the conditional branch output O_t^c and the unconditional branch output O_t^u of a module at a given timestep are highly similar. Figure 8 in Appendix A.3 confirms that high branch similarity manifests at specific layer-step pairs across multiple modules.

To address this, we propose the **Branch Skipping (BS)** strategy. While the high similarity suggests the potential for acceleration, simply replacing the unconditional branch with the conditional one can aggravate CFG-induced oversaturation and artifacts (Sadat et al., 2024), manifesting as audible clipping in synthesized speech. This issue is further compounded by the diverse conditional signals inherent to TTS (e.g., text and timbre). Therefore, unlike direct replacement, BS employs a Branch Residual mechanism to preserve these essential details. As shown in Figure 2, under BS, only the conditional branch is executed, and the unconditional branch output for the current timestep is reconstructed using the conditional branch output and the branch residual (computed as the difference between the cached unconditional and conditional branch outputs). Let the output of a module at timestep t be $O_t = \text{Concat}(O_t^c, O_t^u)$, under BS, the output becomes

$$O_t = \text{Concat}(O_t^c, O_{t-1}^u - O_{t-1}^c + O_t^c), \quad (2)$$

This formulation skips the redundant branch computation while ensuring that the resulting output closely approximates the original.

3.3 Attention Pattern

Early identification and application of TS on temporally redundant layer-step pairs avoids forcing them into suboptimal compression strategies during the subsequent greedy-based calibration phase. Therefore, we focus on developing accurate methods for detecting such redundancy. In fact, the degree of temporal redundancy depends not only on the interval between timesteps but also on the distinct functional roles of internal layers in generation tasks. DeepCache (Ma et al., 2024b) observed that shallow layers in diffusion models construct the overall outline, whereas deeper layers are responsible for synthesizing fine details. In our work, we further explore the relationship between the attention patterns of each layer in the DiT and the distinct functional roles of those layers in the inference process, leveraging this connection to efficiently identify highly temporally redundant layer-step pairs.

The attention heatmaps in the DiT layers exhibit two distinct patterns, diagonal-like and striped. These patterns indicate potential functional distinctions among layer-step pairs. Figure 3 presents these patterns during inference in the F5-TTS model: (a) shows the diagonal-like pattern for layer 5 at timestep 0, while (b) displays the striped pattern for layer 2 at timestep 26. In the context of speech synthesis, the diagonal-like patterns suggest a focus on local acoustic refinement, such as prosody and spectral details, thereby enhancing speech fidelity and clarity. Conversely, we hypothesize that striped patterns function as attention anchors, broadcasting information across the sequence to establish long-range dependencies, which are crucial for preserving the structural co-

herence of the synthesized speech.

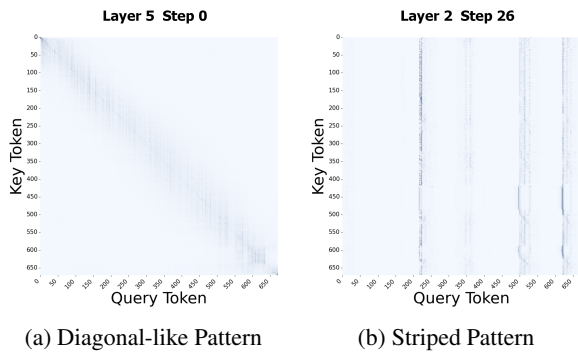


Figure 3: **Attention Patterns in F5-TTS inference:** (a) Diagonal-like patterns in both conditional and unconditional branches. (b) Striped patterns in both branches.

To further analyze the diagonal-like attention pattern, we collected the cosine similarity between attention heatmaps and diagonal matrices, as well as the corresponding temporal redundancy across all layer-step pairs in F5-TTS. The greater the similarity is, the more closely the layer-step pair aligns with a diagonal-like pattern. For layer l at timestep t , let O denote the original output of the model and $O_{l,t}$ represent the final model output when layer l uses the cached output from the timestep $t - 1$ instead of recomputing it at the current timestep t . Based on these outputs, we define the temporal redundancy as:

$$R_{l,t} = 1 - \frac{1}{b \cdot n \cdot d} \sum_{k=1}^b \sum_{i=1}^n \sum_{j=1}^d \|O_{k,i,j} - O'_{k,i,j}\|_1, \quad (3)$$

where b denotes the batch size, n the sequence length, and d the feature dimension. This metric measures the mean absolute error between O and $O_{l,t}$. Layer-step pairs with redundancy above 0.9 are marked as highly temporally redundant. Figure 4 illustrates the percentage of temporally redundant layer-step pairs in each similarity interval in F5-TTS. At lower similarity, specifically below 0.1, the percentage of redundant pairs remains relatively low, fluctuating between 75% and 90%. As the similarity increases from 0.1 to 0.35, the redundancy percentage exhibits a notable upward trend, reaching a peak of nearly 100% between 0.15 and 0.35. The phenomenon suggests a tendency that: (1) DiT layers exhibiting the diagonal-like patterns are more likely to demonstrate higher temporal redundancy during inference; (2) conversely, layers with striped patterns tend to exhibit lower temporal redundancy.

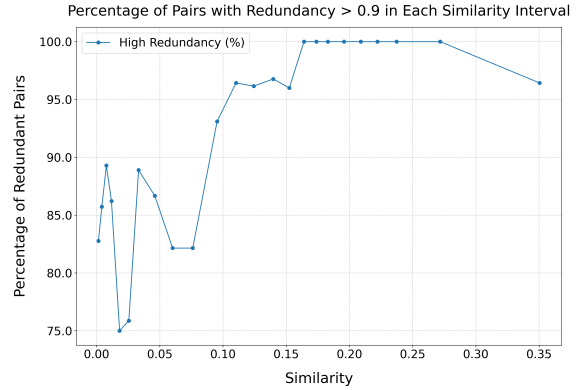


Figure 4: **Percentage of temporally redundant layer-step pairs in each similarity interval in F5-TTS.**

3.4 DiTReducio

Previous studies (Sun et al., 2024) analyzed the computational redundancy of prevalent DiT relative to their inputs, revealing that internal redundancy patterns are input-agnostic, depending primarily on model architecture and model parameters rather than specific input content. This property enables the construction of model-agnostic acceleration strategies via a limited number of inference runs. Accordingly, we introduce DiTReducio, which achieves remarkable acceleration by performing only three iterative inference passes and trades off quality and speed with a predefined compression threshold. DiTReducio comprises three phases: **Check Phase**, **Pre-Calibration Phase**, and **Calibration Phase**.

Check Phase The goal of this phase is to mark highly temporally redundant layer-step pairs before calibration. Leveraging the correlation between attention patterns and redundancy revealed in Sec. 3.3, we compute the cosine similarity $s_{l,t}$ between each layer’s attention heatmap and the identity matrix before the forward of layer l at step t . After inference, all $s_{l,t}$ values are sorted in descending order, and the top $q\%$ highest-similarity layer-step pairs are marked.

Pre-Calibration Phase This phase performs a preliminary calibration based on the Check Phase results. Its procedure mirrors that of the Calibration Phase, except that (1) only the marked layer-step pairs are considered, and (2) only the TS is applied, without the BS. The ablation studies in Sec. 4.3 verify that this phase substantially improves the quality of the resulting acceleration strategy.

Algorithm 1: Calibration Phase of DiTReducio

Input : Model M , Calibration threshold δ , total number of layers L

Output : Strategy table $\text{strategy_table}[T][L]$

Initialize strategy_table as a $T \times L$ matrix filled with NONE;

During the sampling process of the model, at each timestep t :

for layer $l \leftarrow 1$ **to** L **do**

for method $m \in \{TS, BS\}$ **do**

$O \leftarrow$ model output at timestep t without compression at layer l ;

$O' \leftarrow$ model output at timestep t with method m applied to layer l ;

$\epsilon \leftarrow \frac{1}{n} \sum |O - O'|$, where n is the number of elements in the tensor;

if $\epsilon < \frac{l}{L} \cdot \delta$ **then**

 Update $\text{strategy_table}[t][l] \leftarrow m$;

Calibration Phase This phase performs a comprehensive calibration following the Pre-Calibration Phase. Since the impact of compressing deep layers is greater than that of compressing shallow layers, under a given global calibration threshold δ , we introduce a dynamic threshold that controls the computational compression for the layer-step pair (l, t) using the threshold $\frac{l}{L}\delta$, where L is the total number of layers in the DiT. The algorithm is outlined in Algorithm 1. During the model’s inference at step t , for layer l , we first compute the entire model output O without applying any acceleration strategy to pair (l, t) . Then we sequentially apply both TS and BS strategies to the attention and feed-forward modules of the layer to obtain the compressed output O' , prioritizing TS due to its higher computational savings. Next, we compute the mean absolute error between O and O' . If this value is less than the threshold $\frac{l}{L}\delta$, we record the applied strategy used for (l, t) ; otherwise, we record that no acceleration strategy is used for (l, t) . Notably, if (l, t) has been selected for TS application during the Pre-Calibration Phase, we skip the calibration for that pair.

4 Experiments

4.1 Settings

Model We evaluate DiTReducio on F5-TTS (Chen et al., 2024b) and MegaTTS 3 (Jiang et al., 2025), both of which employ DiT for speech synthesis. To implement our compression method, we make several modifications to the model. Implementation details are provided in the Appendix A.1.

Dataset & Task We utilize the LibriSpeech-PC-test-clean subset (Meister et al., 2023) comprising 1,127 samples, following F5-TTS. We assess the performance of the model under DiTReducio within the cross-sentence task paradigm, where models generate speech with consistent speaker characteristics based on a reference text, a speaker prompt, and a corresponding transcription.

Metric We adopt the evaluation metrics from F5-TTS to assess generation quality. The speaker similarity-objective (SIM-o) is computed using the WavLM-large-based model (Chen et al., 2022) to compute cosine similarity between features extracted from synthesized and reference audio. The Word Error Rate (WER) is calculated by comparing the reference text against transcriptions generated by Whisper-large-v3 (Radford et al., 2023). For perceptual quality, we conducted a Mean Opinion Score (MOS) evaluation focusing on the baseline (T0) and the optimal compression threshold (T4) for both F5-TTS and MegaTTS 3. Detailed experimental protocols regarding the subjective evaluation are provided in Appendix A.2. To evaluate acceleration performance, we adopt the RTF to quantify inference speed, and use the total floating-point operations (FLOPs) as an indicator of computational costs. For F5-TTS, the RTF is measured based on the model’s total inference time. For MegaTTS 3, only the DiT inference time is considered, as DiT is not the sole bottleneck in the overall inference pipeline.

Evaluation Results are averaged over 5 random seeds. For both F5-TTS and MegaTTS 3, we evaluate 6 distinct compression thresholds (denoted as T1 through T6 in ascending order, where T0 repre-

Model	Metric	Threshold						
		T0	T1	T2	T3	T4	T5	T6
F5-TTS	SIM-o	0.640	0.640	0.637	0.629	0.618	0.610	0.590
	WER (%)	2.636	2.655	2.564	2.643	2.634	2.661	2.900
	RTF	0.178	0.165	0.149	0.138	0.129	0.120	0.112
	Ops Ratio (%)	100.00	82.59	66.38	55.09	45.58	39.26	34.42
MegaTTS 3	SIM-o	0.750	0.750	0.748	0.743	0.734	0.691	0.626
	WER (%)	3.112	3.112	3.110	3.073	3.095	3.133	3.030
	RTF	0.396	0.395	0.359	0.287	0.224	0.176	0.156
	Ops Ratio (%)	100.00	98.87	88.02	68.19	48.94	33.88	27.52

Table 1: **Performance comparison between F5-TTS and MegaTTS 3 under varying compression thresholds.** The bold column (T4) represents the optimal threshold, at which the models achieve substantial acceleration while maintaining generation quality within acceptable bounds. *Ops Ratio* denotes the ratio of FLOPs after compression relative to the baseline (uncompressed) model, indicating the extent of computational reduction. The data is evaluated on a single Nvidia 3090 GPU.

sents the uncompressed baseline). The thresholds are uniformly distributed, with maximum values of 0.3 and 1.2 for F5-TTS and MegaTTS 3, and the maximum threshold resulting in approximately 10% degradation in SIM-o. In the Check Phase, we identify the top 10% of layer-step pairs as highly temporally redundant pairs.

Table 2: **Subjective Mean Opinion Score (MOS) evaluation.** The Acceleration Ratio is calculated based on the RTF reduction.

Model	Threshold	MOS (\uparrow)	Speedup
F5-TTS	T0	3.93 \pm 0.12	1.00 \times
	T4	3.90 \pm 0.11	1.37 \times
MegaTTS 3	T0	3.98 \pm 0.09	1.00 \times
	T4	3.94 \pm 0.08	1.76 \times

4.2 Results

DiTReducio significantly accelerates inference while maintaining high fidelity in both objective and perceptual metrics. As shown in Table 1, for F5-TTS, applying lower thresholds yields substantial RTF reductions (e.g., 16.5% at T2) with acceptable quality loss. Even at maximum compression (37.0% reduction), degradation remains moderate. MegaTTS 3 exhibits a similar robust trajectory, preserving generation quality across most thresholds while notably improving efficiency. Table 2 confirms that at the optimal threshold (T4), DiTReducio achieves speedups of 1.37 \times for F5-TTS and 1.76 \times for MegaTTS 3, with no significant drop in listening quality compared to the baseline.

The selection of compression thresholds is critical. Our analysis reveals an interesting

threshold-dependent behavior: from lower to moderate thresholds (T2 to T4), increasing the threshold yields considerable speedup with minimal quality degradation. However, at higher threshold levels (T5 to T6), the marginal acceleration benefits diminish while quality degradation becomes more evident. This phenomenon suggests that compression approaches its theoretical limit around T4, beyond which further threshold increases may incorrectly identify essential computations as redundant. With appropriate threshold selection, DiTReducio effectively balances inference acceleration and generation quality. Furthermore, different models exhibit varying sensitivities to threshold settings: the F5-TTS shows quality degradation at a compression threshold of 0.3, while MegaTTS 3 retains quality up to a threshold of 0.4, where it achieves notable compression gains.

4.3 Ablation Study

In this section, we investigate the impact of the Check Phase, Pre-calibration Phase, and compression strategies on DiTReducio’s performance. Additionally, to validate the necessity of the branch residual mechanism, we compare our approach with direct branch replacement strategies as employed in DiTFastAttn to demonstrate its superiority in preserving generation quality.

The Check Phase and Pre-Calibration Phase are critical for identifying effective acceleration strategies. As evidenced in Figure 5, omitting the first two phases of DiTReducio leads to significant degradation in both generation quality and inference speed for F5-TTS under equivalent com-

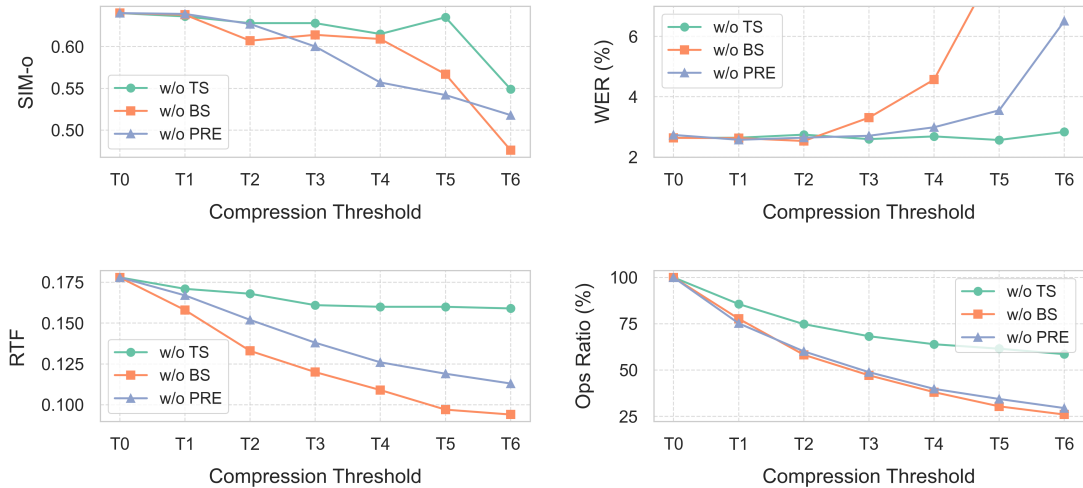


Figure 5: Ablation study on F5-TTS. PRE represents the Check Phase and the Pre-Calibration Phase.

pression thresholds. These findings confirm that these phases guide the Calibration Phase toward a superior strategy combination.

TS strategy is crucial for achieving effective acceleration. As shown in Figure 5, ablation results demonstrate that applying only the Branch Skipping (BS) strategy to F5-TTS maintains speech quality. However, further increases in the compression threshold yield diminishing returns in acceleration beyond T3. This highlights the intrinsic limitations of relying solely on the BS strategy for acceleration.

BS strategy is essential for maintaining generation quality. As shown in Figure 5, while only applying the TS strategy achieves greater acceleration of F5-TTS, it significantly degrades audio quality. For F5-TTS, employing TS alone leads to a substantial decrease in SIM-o at equivalent compression thresholds, with WER even reaching 23.06% at the maximum threshold, indicating loss of conditional information during inference.

Branch Residual Mechanism is vital for BS Strategy. As illustrated in Figure 6, although all three variants achieve comparable inference acceleration (RTF), their generation quality differs significantly. The *Uncond-Replace* variant suffers from severe degradation in both SIM-o and WER. Furthermore, while *Cond-Replace* alleviates the issue, it still yields inferior metrics compared to our approach. In contrast, our method consistently maintains the generation quality at high thresholds, verifying that the branch residual is essential for recovering guidance details that are otherwise lost in direct replacement.

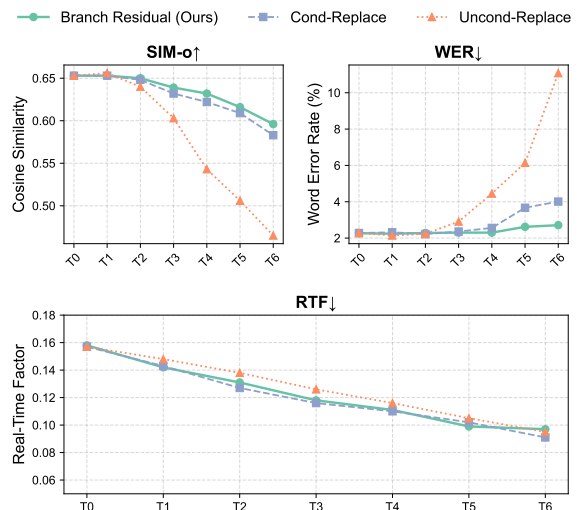


Figure 6: Ablation study of branch replacement strategies on F5-TTS.

5 Conclusion

In this paper, we present a training-free acceleration approach for DiT-based TTS models, which derives a persistent acceleration strategy through a progressive calibration process. We observe temporal redundancy and branch redundancy in model inference, and develop corresponding strategies to exploit them effectively. We analyze the relationship between attention patterns and temporal redundancy in a specific layer-step pair, and develop a calibration framework that effectively identifies and compresses internal redundancies in a model-specific manner. Our experimental results verify that DiTReducio accelerates inference at no training cost while maintaining generation quality.

Limitations

Firstly, the applicability of DiTReducio is primarily constrained to DiT-based speech synthesis models, which limits its generalization to other model architectures. Additionally, the framework demands high-quality calibration audio for optimal performance.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. U25B2064 and Grant No. 62572423.

References

- Inclusion AI, Bowen Ma, Cheng Zou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, GuangMing Yao, and 1 others. 2025. Ming-flash-omni: A sparse, unified architecture for multimodal perception and generation. *arXiv preprint arXiv:2510.24821*.
- Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*.
- Daniel Bolya and Judy Hoffman. 2023. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4599–4603.
- Thibault Castells, Hyoung-Kyu Song, Bo-Kyeong Kim, and Shinkook Choi. 2024. Ld-pruner: Efficient pruning of latent diffusion models using task-agnostic insights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 821–830.
- Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024a. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.
- Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024b. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.
- Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. 2022. Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6147–6151. IEEE.
- Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.
- Wei Deng, Siyi Zhou, Jingchen Shu, Jinchao Wang, and Lu Wang. 2025. Indextts: An industrial-level controllable and efficient zero-shot text-to-speech system. *arXiv preprint arXiv:2502.05512*.
- Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.
- Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.
- Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, and 1 others. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 682–689. IEEE.
- Wenhao Guan, Qi Su, Haodong Zhou, Shiyu Miao, Xingjia Xie, Lin Li, and Qingyang Hong. 2024. Reflow-tts: A rectified flow model for high-fidelity text-to-speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10501–10505. IEEE.
- Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. 2023. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6185–6194.
- Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. 2023. Ptdq: Accurate post-training quantization for diffusion models. *arXiv preprint arXiv:2305.10657*.
- Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. 2022. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605.
- Hieu-Nghia Huynh-Nguyen, Ngoc Son Nguyen, Huynh Nguyen Dang, Thieu Vo, Truong-Son Hy, and Van Nguyen. 2025. Ozspeech: One-step zero-shot speech synthesis with learned-prior-conditioned flow matching. *arXiv preprint arXiv:2505.12800*.

- Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, and 1 others. 2025. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. *arXiv preprint arXiv:2502.18924*.
- Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, and 1 others. 2024. Naturspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.
- Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, and 1 others. 2022. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *European conference on computer vision*, pages 620–640. Springer.
- Keon Lee, Dong Won Kim, Jaehyeon Kim, and Jaewoong Cho. 2024. Ditto-tts: Efficient and scalable zero-shot text-to-speech with diffusion transformer. *arXiv preprint arXiv:2406.11427*.
- Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. 2023. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17535–17545.
- Yinghao Aaron Li, Rithesh Kumar, and Zeyu Jin. 2024. Dmdspeech: Distilled diffusion model surpassing the teacher in zero-shot speech synthesis via direct metric optimization. *arXiv preprint arXiv:2410.11097*.
- Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022a. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787.
- Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022b. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.
- Xinyin Ma, Gongfan Fang, Michael Bi Mi, and Xinchao Wang. 2024a. Learning-to-cache: Accelerating diffusion transformer via layer caching. *Advances in Neural Information Processing Systems*, 37:133282–133304.
- Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2024b. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15762–15772.
- Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2024. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11341–11345. IEEE.
- Aleksandr Meister, Matvei Novikov, Nikolay Karpov, Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. 2023. Librispeech-pc: Benchmark for evaluation of punctuation and capitalization capabilities of end-to-end asr models. In *2023 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 1–7. IEEE.
- William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.
- Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, and 1 others. 2025. Vibevoice technical report. *arXiv preprint arXiv:2508.19205*.
- Chunyu Qiang, Kang Yin, Xiaopeng Wang, Yuzhe Liang, Jiahui Zhao, Ruibo Fu, Tianrui Wang, Cheng Gong, Chen Zhang, Longbiao Wang, and 1 others. 2025. 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 *International conference on machine learning*, pages 28492–28518. PMLR.
- Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. 2024. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*.
- Omid Saghatian, Atiyeh Gh Moghadam, and Ahmad Nickabadi. 2025. Cached adaptive token merging: Dynamic token reduction and redundant computation elimination in diffusion model. *arXiv preprint arXiv:2501.00946*.
- Tim Salimans and Jonathan Ho. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.
- Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. 2024. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer.
- Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. 2023. Post-training quantization on diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1972–1981.
- Xibo Sun, Jiarui Fang, Aoyu Li, and Jinzhe Pan. 2024. Unveiling redundancy in diffusion transformers (dits): A systematic study. *arXiv preprint arXiv:2411.13588*.

- Ben Wan, Tianyi Zheng, Zhaoyu Chen, Yuxiao Wang, and Jia Wang. 2025. Pruning for sparse diffusion models based on gradient flow. *arXiv preprint arXiv:2501.09464*.
- Wei Wang, Rong Cao, Yi Guo, Zhengyang Chen, Kuan Chen, and Yuanyuan Huo. 2025a. Intmeanflow: Few-step speech generation with integral velocity distillation. *arXiv preprint arXiv:2510.07979*.
- Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, and 1 others. 2025b. Sparktts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.
- Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.
- Fan Xie, Dan Zeng, Qiaomu Shen, and Bo Tang. 2025. A comprehensive survey on text-to-video generation. *Chinese Journal of Electronics*, 34(4):1009–1036.
- Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and 1 others. 2024. Pyramidrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.
- Zhihang Yuan, Hanling Zhang, Lu Pu, Xuefei Ning, Linfeng Zhang, Tianchen Zhao, Shengen Yan, Guohao Dai, and Yu Wang. 2024. Ditfastattn: Attention compression for diffusion transformer models. *Advances in Neural Information Processing Systems*, 37:1196–1219.
- Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and 1 others. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.
- Hanling Zhang, Rundong Su, Zhihang Yuan, Pengtao Chen, Mingzhu Shen, Yibo Fan, Shengen Yan, Guohao Dai, and Yu Wang. 2025a. Ditfastattnv2: Headwise attention compression for multi-modality diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16399–16409.
- Hui Zhang, Tingwei Gao, Jie Shao, and Zuxuan Wu. 2025b. Blockdance: Reuse structurally similar spatio-temporal features to accelerate diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12891–12900.
- Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. 2024. Real-time video generation with pyramid attention broadcast. *arXiv preprint arXiv:2408.12588*.
- Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. 2024. Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 554–563.
- Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, and 1 others. 2025. Voxcpm: Tokenizer-free tts for context-aware speech generation and true-to-life voice cloning. *arXiv preprint arXiv:2509.24650*.
- Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. 2025. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. *arXiv preprint arXiv:2506.13053*.

A Appendix

A.1 Implementation Details

For F5-TTS, we optimize the DiT implementation by integrating the conditional and unconditional branch computations into a single forward pass. Specifically, we concatenate conditional and unconditional inputs into a single batch. This modification allows the BS strategy to be applied while retaining functional equivalence with the original dual-pass method.

For MegaTTS 3, which employs multi-condition classifier-free guidance, we adapted the BS strategy to accommodate its dual conditional branches by computing two separate residuals: one between the text conditional branch and the speaker conditional branch, and another between the unconditional branch and the speaker conditional branch. During inference, only the speaker conditional branch is computed explicitly, while the other two branches are reconstructed by adding their respective residuals to it.

Both models are enhanced with FlashAttention, confirming the compatibility of our framework with efficient attention implementations.

A.2 Subjective Evaluation Details

We conducted a subjective evaluation to assess the perceptual quality of the synthesized speech. Our evaluation focused on four specific model-threshold combinations: the baseline (T0) and the optimal compression threshold (T4) for both F5-TTS and MegaTTS 3. For each combination, we generated 10 English utterances of varying lengths. We recruited 20 participants and evenly divided them into four independent groups using a between-subjects design, with each group assigned to evaluate one specific model-threshold combination. Each participant rated all 10 utterances within their assigned group, yielding 50 subjective scores per combination. Ratings were based on audio quality using a 5-point Likert scale. The final MOS is reported along with the standard deviation.

A.3 Redundancy in Model

We analyze the temporal and branch redundancy in both F5-TTS and MegaTTS 3 models. Figure 7 presents the temporal redundancy analysis results. For F5-TTS, we examine layers 10 and 20, computing the cosine similarity between outputs from adjacent denoising timesteps for both attention and feed-forward. Similar analysis is conducted for

MegaTTS 3, focusing on layers 10 and 22. The results reveal high temporal redundancy during the inference of the model.

The branch redundancy characteristics are depicted in Figure 8. For F5-TTS, we measure the cosine similarity between outputs from conditional and unconditional branches for both attention and feed-forward modules across various timesteps. For MegaTTS 3, which utilizes multi-condition classifier-free guidance with two conditional branch outputs, we analyze the similarities among all three branches, specifically comparing the speaker conditional branch (Branch 1) with both the text conditional branch (Branch 2) and the unconditional branch (Branch 3).

A.4 Method Distribution

Figure 9 shows the method distribution heatmaps from the calibration of F5-TTS under increasing compression thresholds. Each cell in the heatmaps represents the compression method applied to a specific layer-step pair. The heatmaps are arranged in a grid layout from left to right and top to bottom, with each subplot corresponding to a threshold in 0.05 increments, ranging from 0.05 to 0.30. As the compression threshold rises, we observe a progressive increase in the number of layer-step pairs employing compression strategies. Notably, layer-step pairs at later timesteps are more likely to adopt TS, whereas those at earlier timesteps tend to remain uncompressed or use BS. This suggests an underlying connection between internal model redundancy and the progression across diffusion timesteps.

A.5 Potential Risks

While DiTReducio offers an efficient, training-free approach to accelerate DiT-based TTS models, its deployment entails certain risks that require careful management. A key concern is the potential degradation of speech quality in high-stakes applications such as healthcare, legal transcription, or emergency response, where even minor distortions could lead to misinterpretation or diminished user trust. Additionally, the enhanced efficiency of DiTReducio may lower the barrier for misuse, enabling malicious actors to generate deceptive audio content such as deepfakes or impersonation attacks. Robust safeguards are therefore essential to mitigate such risks.

A.6 Supplementary Experiments

To further validate the performance of DiTReducio, we conduct additional comparative experiments on the LibriSpeech-PC-test-clean and LibriSpeech-PC-test-other subsets. Given the absence of specialized training-free acceleration baselines in the speech synthesis domain, we implement DiTFastAttn, a training-free method from the image and video generation domain, as our primary baseline. To verify performance in more challenging scenarios, we select the noisier and more difficult LibriSpeech-PC-test-other subset. We strictly follow the cross-sentence evaluation protocol of F5-TTS, filtering 1,832 samples with durations ranging from 3 to 10 seconds. All experiments are performed on $8 \times$ NVIDIA H200 GPUs. F5-TTS employs batch inference with a batch size of 128 to mitigate the impact of cache I/O overhead on RTF measurements.

As shown in Table 3, within the standard compression range T1–T4, DiTReducio consistently achieves a significantly lower RTF than DiTFastAttn across both the LibriSpeech-PC-test-clean and the LibriSpeech-PC-test-other subsets. As acceleration threshold increases, the generation quality gap between the two methods becomes particularly pronounced. DiTReducio maintains a remarkably stable WER, fluctuating marginally between 2.05% and 2.08% on the clean subset, whereas DiTFastAttn exhibits linear quality degradation. This divergence amplifies on the more challenging LibriSpeech-PC-test-other subset, where the baseline WER sharply rises at T3 and peaks at 7.01% at T6, while DiTReducio remains robust. Although the acceleration ratios of both methods converge at this stage, DiTReducio effectively preserves linguistic accuracy.

Table 3: Performance comparison between DiTFastAttn (Baseline) and DiTReducio (Ours) on LibriSpeech-PC-test-clean and LibriSpeech-PC-test-other.

Dataset	Method	Level	RTF (\downarrow)	WER (%) (\downarrow)	SIM-o (\uparrow)
LibriSpeech-PC-test-clean	Vanilla	-	0.0474	2.054	0.670
	Baseline (DiTFastAttn)	T1	0.0494	2.066	0.675
		T2	0.0462	2.118	0.680
		T3	0.0426	2.160	0.664
		T4	0.0403	2.213	0.652
		T5	0.0383	2.663	0.627
		T6	0.0369	3.481	0.585
	Ours (DiTReducio)	T1	0.0451	2.060	0.670
		T2	0.0421	2.086	0.670
		T3	0.0408	2.076	0.662
		T4	0.0392	2.052	0.661
		T5	0.0371	2.068	0.633
T6		0.0369	2.085	0.606	
LibriSpeech-PC-test-other	Vanilla	-	0.0576	4.373	0.653
	Baseline (DiTFastAttn)	T1	0.0607	4.463	0.659
		T2	0.0562	4.641	0.664
		T3	0.0514	4.963	0.659
		T4	0.0493	5.315	0.634
		T5	0.0466	5.349	0.602
		T6	0.0445	7.014	0.583
	Ours (DiTReducio)	T1	0.0551	4.543	0.653
		T2	0.0515	4.257	0.652
		T3	0.0489	4.270	0.649
		T4	0.0473	4.274	0.629
		T5	0.0447	4.542	0.619
T6		0.0440	4.654	0.602	

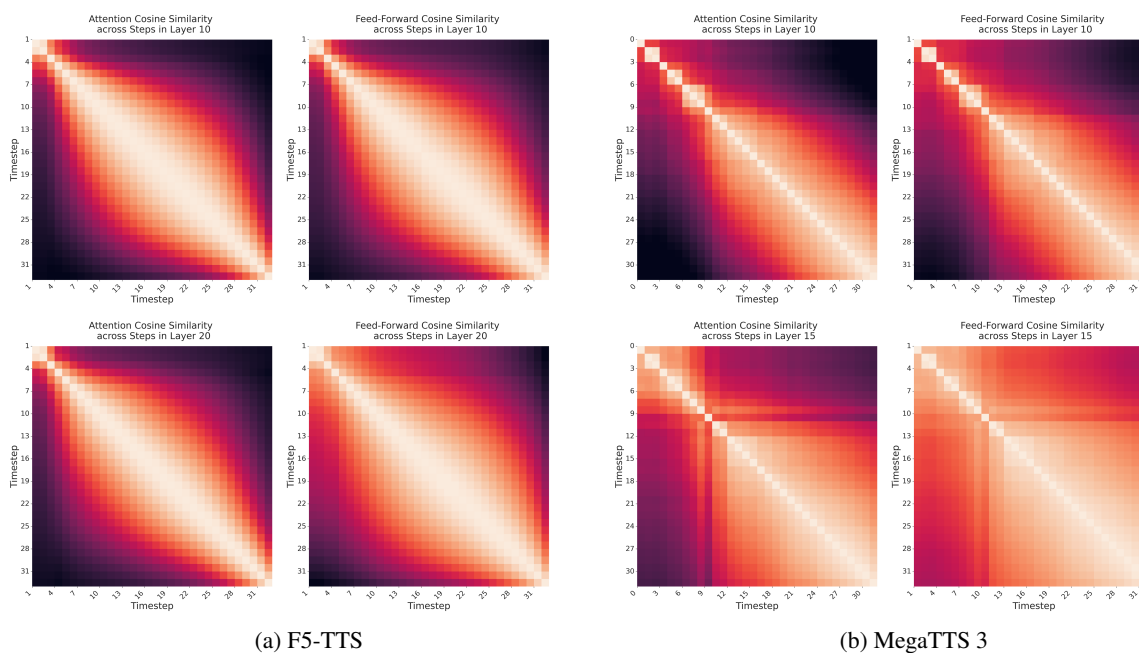


Figure 7: Temporal redundancy in F5-TTS and MegaTTS 3.

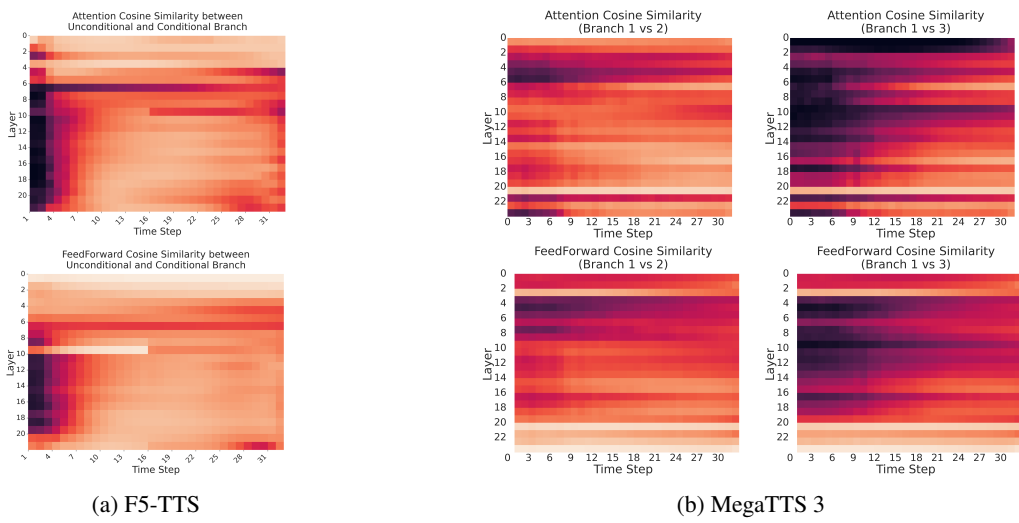


Figure 8: Branch redundancy in F5-TTS and MegaTTS 3.

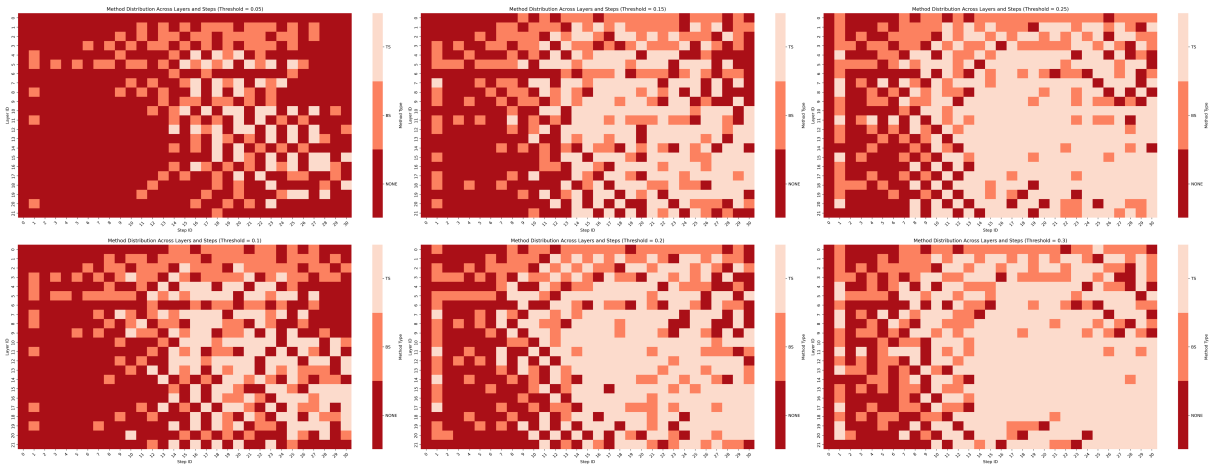


Figure 9: Method Distribution of F5-TTS across compression thresholds.

Listening Test (MOS)

General Instructions

- **Equipment:** Please use high-quality studio headphones and a professional soundcard.
- **Preparation:** Listen to the **entire audio** before rating.
- **Holistic Evaluation:** Rate the overall impression of the audio quality and its alignment with the provided text.

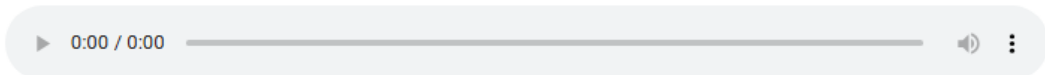
Scoring Criteria

5 (Excellent)	No defects, perfectly clear, no missing words, very comfortable.
4 (Good)	Minimal defects, clear and complete expression, comfortable.
3 (Fair)	Defects present but not distracting, understandable but mediocre.
2 (Poor)	Noticeable defects, uncomfortable, hinders full comprehension.
1 (Bad)	Severe distortion/noise, incoherent, many missing words.

Item 1 of 10

Test Sample

"The quick brown fox jumps over the lazy dog."



Rate the Quality (1.0 - 5.0):

1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0

Submit & Next

Figure 10: The page for subjective Mean Opinion Score (MOS) assessment..