FreqTS: Frequency-Aware Token Selection for Accelerating Diffusion Models

Xinye Yang[1], Yuxin Yang[2], Haoran Pang[3], Aaron Xuxiang Tian[4], Luking Li[4]

[1]Newcastle University,
[2]University of Hong Kong,
[3]National University of Singapore,
[4]Independent Researcher
c0078451@newcastle.ac.uk, yuxyang@connect.hku.hk, e1068502@u.nus.edu

Abstract

In this paper, we propose FreqTS, a novel Frequency-Aware Token Selection approach for accelerating diffusion models without requiring retraining. Diffusion models have gained significant attention in the field of image synthesis due to their impressive generative capabilities. However, these models often suffer from high computational costs, primarily due to the sequential denoising process and large model size. Additionally, diffusion models tend to prioritize low-frequency features, leading to sub-optimal quantitative results. To address these challenges, FreqTS introduces an amplitude-based sorting method that separates Token features in the frequency domain of diffusion models into high-frequency and low-frequency subsets. It then utilizes fast Token Selection to reduce the presence of low-frequency features, effectively reducing the computational overhead. Moreover, FreqTS incorporates a Bayesian hyper-parameter search to dynamically assign different selection strategies for various denoising processes. Extensive experiments conducted on Stable Diffusion series models, PixArt-Alpha, LCM, and other models demonstrate that FreqTS achieves a minimum acceleration of 2.3× without the need for retraining. Furthermore, FreqTS showcases its versatility by being applicable to different sampling techniques and compatible with other dimension-specific acceleration algorithms.

Introduction

Diffusion models (Ho, Jain, and Abbeel 2020) have emerged as a powerful generative modeling framework, achieving remarkable success in various domains, including image synthesis (Rombach et al. 2022a), audio generation (Kong et al. 2021), and 3D generation (Poole et al. 2023). However, despite their impressive performance, these models often suffer from substantial computational redundancy, leading to inefficient inference and hindering their widespread deployment, particularly on resource-constrained devices. This computational overhead stems from the iterative nature of the diffusion process, which typically involves numerous denoising steps, as well as the inherent complexity of the model architectures employed. Addressing these efficiency challenges is crucial for enabling the practical application of diffusion models in real-world scenarios. There is a growing demand for efficient techniques that can accelerate the training and inference processes without compromising the model's generative performance.

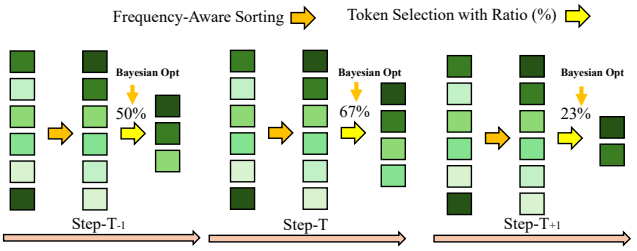


Figure 1: Our FreqTS framework begins by applying an amplitude-based sorting technique in the frequency domain, effectively dividing Token features into high-frequency and low-frequency subsets. Then By utilizing fast Token Selection, FreqTS selectively reduces the presence of low-frequency features with Token Selection , Additionally, FreqTS incorporates a Bayesian hyper-parameter search, enabling the dynamic assignment of different selection strategies tailored to specific denoising steps.

Addressing the efficiency challenges of diffusion models is crucial for enabling their practical application in real-world scenarios. There is a growing demand for efficient techniques that can accelerate the training and inference processes without compromising the model's generative performance, including knowledge distillation (Liu et al. 2023; Li and Jin 2022; Li 2022; Li et al. 2024a), pruning (Li et al. 2024c,f; Dong et al. 2024), quantization (Dong et al. 2023, 2025) and decomposition (Li et al. 2025; Gu et al. 2025). Knowledge distillation techniques have the potential to significantly accelerate inference while preserving generation quality by transferring the knowledge from a larger, computationally expensive teacher model to a smaller, more efficient student model. Pruning and sparsity techniques aim to reduce the computational complexity and memory footprint by identifying and removing redundant parameters or activations within the model architecture, thereby enabling faster inference and deployment on resource-limited devices. Quantization techniques, which involve reducing the precision of model parameters and computations, have also garnered significant attention as a means to enhance the efficiency of diffusion models. These techniques aim to reduce the computational

Figure 2: Visual comparison between SDXL Turbo, SDXL Lightning, SDXL with our FreqTS.

complexity and memory requirements of diffusion models, making them more suitable for deployment in real-world scenarios. Recent advances in these techniques have shown promising results, enabling faster and more efficient generation of high-quality images and other data types. Despite these advances, there is still much to be learned about the application of knowledge distillation, pruning, and quantization to diffusion models. The complexity of diffusion models and the diversity of techniques proposed to improve their efficiency make it challenging to understand and optimize their performance.

To address these challenges, we propose FreqTS, tackles the computational inefficiencies and sub-optimal performance of diffusion models by employing a novel frequency-aware approach (see Fig. 1). At the core of FreqTS lies an amplitude-based sorting method that operates in the frequency domain of the diffusion model. This sorting method separates the Token features into two distinct subsets: high-frequency and low-frequency components. The key insight behind FreqTS is that low-frequency features often contribute less to the overall quality of the generated samples, while incurring significant computational costs during the iterative denoising process. By identifying and selectively reducing the presence of these low-frequency features, FreqTS can effectively reduce the computational overhead without compromising the generative performance of the diffusion model. To achieve this, FreqTS employs a fast Token Selection technique that utilizes the separated high-frequency and low-frequency subsets. By prioritizing the high-frequency components and selectively retaining a subset of the low-frequency features, FreqTS can accelerate the denoising process while preserving the essential information required for high-quality generation. However, different denoising processes within the diffusion model may exhibit varying sensitivity to the selection of high- and low-frequency features. To adaptively handle these variations, FreqTS incorporates a Bayesian hyper-parameter search mechanism. This mechanism dynamically assigns different selection strategies for various denoising processes, ensuring that the appropriate balance between computational efficiency and generative performance is achieved throughout the diffusion process. The Bayesian hyper-parameter search explores the parameter space and identifies the optimal selection strategies for each denoising step, considering factors such as the relative importance of high- and low-frequency components, the trade-off between computational complexity and output quality, and the specific characteristics of the diffusion model and the targeted application domain. By combining the amplitude-based sorting, fast Token Selection, and dynamic selection strategy assignment through Bayesian hyper-parameter search, FreqTS offers an effective and adaptive approach to accelerating diffusion models without the need for retraining or compromising generative performance.

We evaluate the effectiveness of FreqTS on various diffusion models, including the Stable Diffusion series, PixArt-Alpha, and LCM. Our extensive experiments demonstrate that FreqTS achieves a significant speedup of at least 2.3× without the need for retraining. Furthermore, we show that FreqTS can be seamlessly integrated with different sampling techniques and orthogonal to other dimension-based acceleration algorithms, providing flexibility and compatibility in practical applications.

- We introduce the FreqTS framework, a novel Frequency-Aware Token Selection approach for accelerating diffu-

sion models without requiring retraining.

- We propose an amplitude-based sorting method that separates token features in the frequency domain, enabling the selective reduction of low-frequency features to improve computational efficiency.
- We incorporate a Bayesian hyper-parameter search to dynamically assign different selection strategies for various denoising processes, further enhancing the adaptability of our approach.
- We extensively evaluate FreqTS on various diffusion models, such as Stable Diffusion, PixArt-Alpha, and LCM. Our experiments show that FreqTS achieves at least a 2.3× speedup without retraining. We also demonstrate FreqTS's compatibility with different sampling techniques and its orthogonality to other acceleration algorithms, confirming its practicality and versatility for efficient image synthesis.

Related Work

Acceleration methods for Diffusion Models

Diffusion models often suffer from substantial computational redundancy, leading to inefficient inference and hindering their widespread deployment. To address this challenge, researchers have proposed various techniques to improve the efficiency and speed of diffusion models, including knowledge distillation, and quantization. Knowledge distillation approaches include progressive distillation (Salimans and Ho 2022), consistency trainings (Song et al. 2023), and adversarial distillation, among others. For example, Progressive Distillation (Salimans and Ho 2022) proposes a progressive distillation framework to accelerate the sampling process and improve generation quality in large and complex diffusion models. Consistency Models (Song et al. 2023) introduce a novel framework that leverages consistency training to improve the generation quality and diversity of diffusion models. Adversarial Diffusion Distillation proposes (Sauer et al. 2023) an adversarial distillation employs an adversarial discriminator to guide the distillation process, encouraging the student to generate diverse and high-quality samples. Hyper-SD (Ren et al. 2024) introduces a trajectory-segmented consistency model that improves the efficiency of image synthesis. Bidirectional Consistency Models (Li and He 2024) introduce a novel framework that leverages bidirectional consistency training to improve the generation quality and diversity of diffusion models. Latent Consistency Models (Luo et al. 2023) propose a novel framework that leverages latent consistency to synthesize high-resolution images with fewer inference steps. SDXL-Lightning (Lin, Wang, and Yang 2024) introduces a progressive adversarial distillation framework for diffusion models. It gradually transfers knowledge from a pre-trained teacher model to the student, improving the student's generation quality and inference speed. Despite the advancements in knowledge distillation for diffusion models, these methods still exhibit limitations when applied to very few-step diffusion models. One major constraint is the requirement for a pre-trained teacher model, which can be computationally expensive and may not always be available. Furthermore, the design of the distillation method itself can be responsible for the performance gap between the teacher and student models, highlighting the need for careful and responsible method design. Other quantization methods reduce the precision of model parameters and computations, making diffusion models more memory-efficient and faster. For instance, Q-Diffusion (Li et al. 2023c) provides a comprehensive quantization framework for diffusion models, significantly reducing memory usage and improving inference speed. Q-DM introduces an efficient low-bit quantized diffusion model, improving inference speed and memory efficiency. PTQD preserves generation quality while improving efficiency through accurate post-training quantization. Temporal Dynamic Quantization (So et al. 2023) leverages the temporal dynamics of diffusion models, adapting the quantization scheme to the changing distribution of feature representations. Data-free Quantization (Wang et al. 2023) eliminates the need for calibration data by utilizing the statistical properties of the model's parameters. Finite Scalar Quantization simplifies vector-quantized variational autoencoders by reducing memory requirements and computational complexity. EfficientDM (He et al. 2024) proposes a quantization-aware fine-tuning framework for low-bit diffusion models, aiming to improve generation quality. It employs a differentiable quantization scheme during fine-tuning, enabling the model to adapt to the quantized representation space. However, challenges remain, including preserving generation quality, adapting to diverse architectures, and ensuring efficient fine-tuning.

Pruning and Sparsity for Diffusion Models

Pruning and sparsity techniques aim to improve the efficiency and speed of diffusion models by reducing their computational complexity and memory footprint through the removal of redundant parameters or activations. For instance, Token Merging (Bolya et al. 2023) proposes an efficient inference technique for Stable Diffusion models. It introduces a token merging strategy that reduces the number of tokens during the generation process, resulting in faster inference and reduced memory usage. Diff-Pruning (Fang, Ma, and Wang 2023) introduces a structural pruning technique specifically designed for diffusion models. DeepCache (Ma, Fang, and Wang 2023) introduces a caching mechanism to accelerate diffusion models by reusing intermediate computations. It stores and reuses the outputs of computationally expensive layers, reducing redundant calculations and improving inference speed. The caching mechanism maintains image quality while significantly speeding up the generation process. Faster Diffusion (Li et al. 2023b) reconsiders the role of the UNet encoder in diffusion models, proposing an alternative architecture. By removing the encoder, the model achieves faster inference speed and reduced memory requirements. The approach maintains generation quality by directly processing the input and leveraging the decoder's capabilities. While these sparsity techniques for improving the efficiency and speed of diffusion models offer promising results, they do have certain limitations, particularly when applied to very few-step diffusion models. These methods may not be as effective in scenarios where the generation process requires some steps. In addition, these methods involve hands-on tuning of the settings without automated ML techniques (Li et al.

2024d; Sun et al. 2024; Li et al. 2024e, 2023a; Dong, Li, and Wei 2023; Li et al. 2024b).

Methodology

Recap of Diffusion Model Fundamentals

Diffusion models are a class of generative models that operate by gradually adding noise to a data sample and then learning to reverse this process. The forward diffusion process involves progressively corrupting a clean data sample $\mathbf{x}_0$ by adding Gaussian noise over a series of T timesteps, resulting in a sequence of increasingly noisy samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$. The forward diffusion process can be described by the following formula:

$$\mathbf{x}_t \sim \sqrt{\bar{\alpha}_t}\mathbf{x}_{t-1} + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \quad \text{where} \quad \epsilon_t \sim \mathcal{N}(0, \mathbf{I}) \quad (1)$$

Here, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ is the cumulative product of pre-defined noise scales $\alpha_i \in (0, 1)$, and ϵ_t is a Gaussian noise term. The goal of the diffusion model is to learn the reverse process, known as the inverse denoising process, which maps the noisy samples back to the clean data distribution. This is achieved by training a denoising model $\epsilon_\theta(\mathbf{x}_t, t)$ to predict the noise ϵ_t added at timestep t, given the noisy sample $\mathbf{x}_t$ and the timestep t. In the DDPM framework, the denoising model is trained to minimize the simple ℓ_2 loss between the predicted noise and the actual noise:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\mathbf{x}_0,t,\epsilon}\left[\|\epsilon_\theta(\mathbf{x}_t, t) - \epsilon\|_2^2\right] \quad (2)$$

During the inference phase, the denoised sample $\mathbf{x}_{t-1}$ is computed iteratively using the following formula:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \frac{1-\bar{\alpha}_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}_t, \quad (3)$$

where $\mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$, σ_t is a pre-defined noise scale, and $\mathbf{z}_t$ is an auxiliary Gaussian noise term. The DDIM framework introduces a different optimization objective and inference procedure. During training, the denoising model is optimized to predict the noise-perturbed data sample $\mathbf{x}_{t-1}$ directly, rather than the noise ϵ_t:

$$\mathcal{L}_{\text{DDIM}} = \mathbb{E}_{\mathbf{x}_0,t,\epsilon}\left[\|\epsilon_\theta(\mathbf{x}_t, t) - \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0 - \sqrt{1-\bar{\alpha}_{t-1}}\epsilon\|_2^2\right] \quad (4)$$

During inference, the denoised sample $\mathbf{x}_{t-1}$ is computed using the following formula:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\epsilon_\theta(\mathbf{x}_t, t) + \sqrt{1-\bar{\alpha}_{t-1}}\mathbf{x}_t \quad (5)$$

This inference procedure does not require auxiliary Gaussian noise, leading to deterministic sampling. In both DDPM and DDIM, the denoising model ϵ_θ is typically implemented as a deep neural network, such as a U-Net or a Transformer-based architecture. The choice of architecture and training strategy can significantly impact the model's performance and computational efficiency.

Frequency-Aware Token Selection

Diffusion models typically operate in the latent space, where the input data is projected onto a sequence of tokens or embeddings. These tokens serve as the fundamental units for the diffusion process, and their effective processing is crucial for efficient computation and high-quality generation. In this section, we introduce our proposed Frequency-Aware Token Selection (FreqTS) approach, which leverages the frequency domain properties of the tokens to reduce computational overhead while preserving generative performance.

Token Representation in Diffusion Models Let $\mathbf{x}_t \in \mathbb{R}^{D\times N}$ be the noisy sample at timestep t, where D is the dimensionality of the latent space, and N is the number of tokens. We can express $\mathbf{x}_t$ as a sequence of tokens $\{\mathbf{z}_t^1, \mathbf{z}_t^2, \ldots, \mathbf{z}_t^N\}$, where $\mathbf{z}_t^i \in \mathbb{R}^D$ represents the i-th token at timestep t.

Frequency Decomposition To analyze the frequency characteristics of the tokens, we first apply a suitable frequency transform, such as the Discrete Fourier Transform (DFT) or the Discrete Cosine Transform (DCT), to each token $\mathbf{z}_t^i$:

$$\mathcal{F}(\mathbf{z}_t^i) = \sum_{k=0}^{D-1} \mathbf{z}_t^i[k] \cdot e^{-j\frac{2\pi k}{D}n}, \quad n = 0, 1, \ldots, D-1 \quad (6)$$

Here, $\mathcal{F}$ represents the frequency transform operator, and $\mathbf{z}_t^i[k]$ denotes the k-th element of the token $\mathbf{z}_t^i$. The resulting frequency coefficients $\mathcal{F}(\mathbf{z}_t^i)$ capture the frequency content of the token, with lower indices representing lower frequencies and higher indices representing higher frequencies.

Frequency-Aware Token Sorting. We sort the tokens based on their frequency amplitudes, separating them into high-frequency and low-frequency subsets. Specifically, we compute the amplitude spectrum $\mathcal{A}(\mathbf{z}_t^i)$ of each token:

$$\mathcal{A}(\mathbf{z}_t^i) = |\mathcal{F}(\mathbf{z}_t^i)| \quad (7)$$

We then sort the tokens in descending order based on their amplitude spectra, creating two subsets: $\mathcal{S}_{\text{high}} = \{\mathbf{z}_t^{i_1}, \mathbf{z}_t^{i_2}, \ldots, \mathbf{z}_t^{i_m}\}$ containing the m tokens with the highest amplitudes (high-frequency subset), and $\mathcal{S}_{\text{low}} = \{\mathbf{z}_t^{j_1}, \mathbf{z}_t^{j_2}, \ldots, \mathbf{z}_t^{j_{N-m}}\}$ containing the remaining $N-m$ tokens (low-frequency subset).

Token Selection. FreqTS employs a token selection strategy that prioritizes the high-frequency subset $\mathcal{S}_{\text{high}}$ while selectively retaining a portion of the low-frequency subset $\mathcal{S}_{\text{low}}$. The selected tokens $\mathcal{S}_{\text{sel}}$ are then used as input for the subsequent denoising process:

$$\mathcal{S}_{\text{sel}} = \mathcal{S}_{\text{high}} \cup \mathcal{S}'_{\text{low}}, \quad \text{where} \quad \mathcal{S}'_{\text{low}} \subseteq \mathcal{S}_{\text{low}} \quad (8)$$

The selection of the low-frequency subset $\mathcal{S}'_{\text{low}}$ is governed by a hyper-parameter $\lambda \in [0, 1]$, which determines the fraction of tokens to be retained from $\mathcal{S}_{\text{low}}$. Specifically, $|\mathcal{S}'_{\text{low}}| = \lfloor\lambda(N-m)\rfloor$, where $|\cdot|$ denotes the cardinality of a set. By prioritizing the high-frequency tokens and selectively retaining a portion of the low-frequency tokens, FreqTS reduces the computational overhead associated with processing the entire token sequence while preserving the essential information required for high-quality generation.

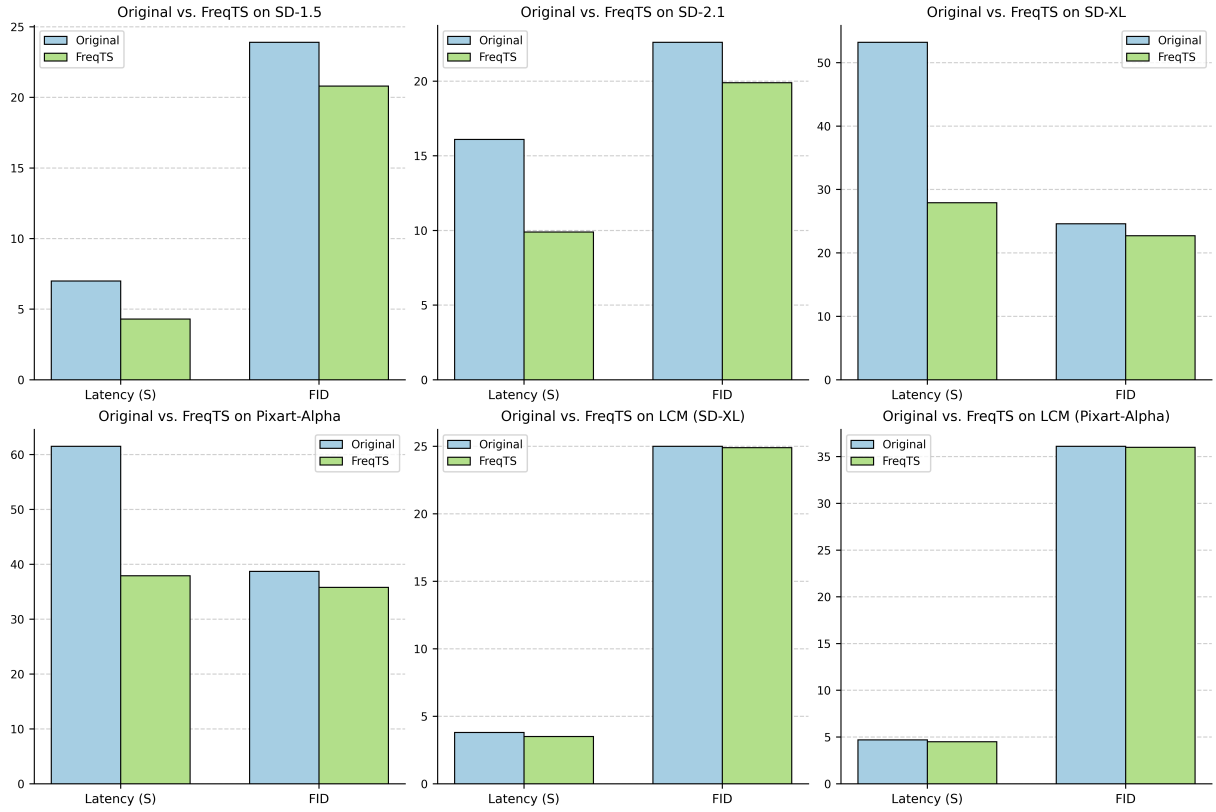


Figure 3: Quantitative Results of FreqTS.

Bayesian Selection Strategy Optimization

While the Frequency-Aware Token Selection approach effectively reduces computational overhead, different denoising processes within the diffusion model may exhibit varying sensitivity to the selection of high- and low-frequency components. To adaptively handle these variations, FreqTS incorporates a Bayesian hyper-parameter search mechanism that dynamically assigns selection strategies tailored to each denoising process.

Bayesian Hyper-parameter Formulation. Let $\lambda_t \in [0, 1]$ denote the hyper-parameter governing the fraction of low-frequency tokens to be retained at timestep t. We model λ_t as a random variable following a Beta distribution, which is a suitable choice for representing proportions:

$$\lambda_t \sim \text{Beta}(\alpha_t, \beta_t) \tag{9}$$

The Beta distribution is parameterized by two shape parameters, α_t and β_t, which control the distribution's mean and variance. Higher values of α_t bias the distribution towards retaining more low-frequency tokens, while higher values of β_t bias the distribution towards retaining fewer low-frequency tokens. To optimize the hyper-parameters α_t and β_t, we adopt a Bayesian optimization approach. Specifically, we define an acquisition function $\mathcal{A}(\alpha_t, \beta_t)$ that quantifies the expected improvement in performance by sampling λ_t from the current Beta distribution. The acquisition function balances the trade-off between exploration (sampling from regions with high uncertainty) and exploitation (sampling from regions with high expected performance). A commonly used acquisition function is the Expected Improvement (EI):

$$\mathcal{A}_{\text{EI}}(\alpha_t, \beta_t) = \mathbb{E}_{\lambda_t \sim \text{Beta}(\alpha_t, \beta_t)} \left[\max \left(0, f(\lambda_t) - f^+ \right) \right] \tag{10}$$

Here, $f(\lambda_t)$ represents the performance metric (e.g., negative loss or a quality score) for a given value of λ_t, and f^+ is the best observed performance so far. The optimization proceeds by iteratively updating the hyper-parameters α_t and β_t to maximize the acquisition function $\mathcal{A}(\alpha_t, \beta_t)$. This process can be performed using various optimization techniques, such as gradient-based methods or Bayesian optimization algorithms like Gaussian Processes.

Dynamic Selection Strategy Assignment. During the inference phase, FreqTS dynamically assigns selection strategies for each denoising process by sampling λ_t from the optimized Beta distribution:

$$\lambda_t \sim \text{Beta}(\alpha_t^*, \beta_t^*) \tag{11}$$

Here, α_t^* and β_t^* are the optimized shape parameters obtained from the Bayesian hyper-parameter optimization process. The sampled value of λ_t determines the fraction of low-frequency tokens to be retained in the Token Selection step for the current denoising process. By incorporating the Bayesian Selection Strategy Optimization, FreqTS can effectively adapt to the varying sensitivity of different denoising

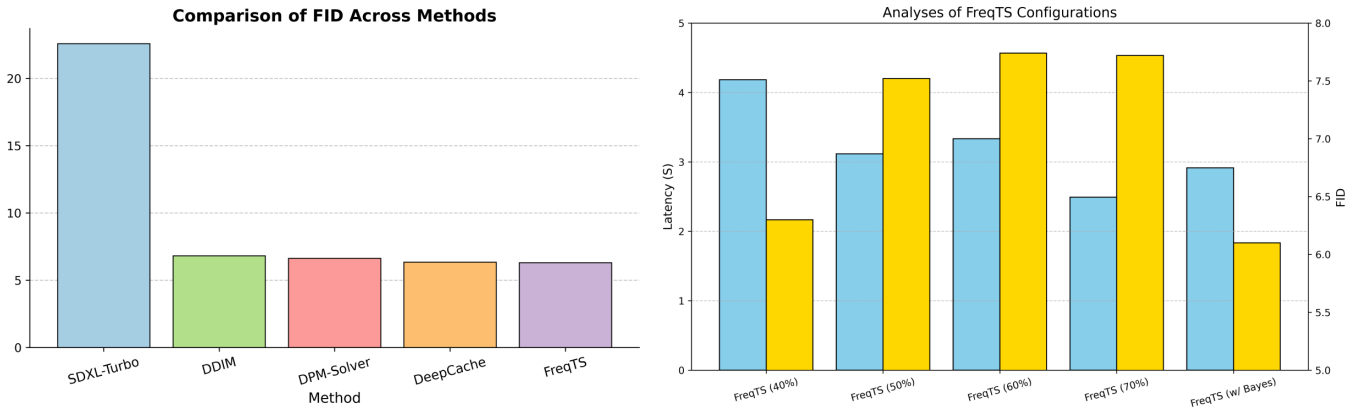


Figure 4: Left: Comparison of Existing Diffusion Acceleration Methods.; Right: Analysis of different configurations of FreqTS.

processes, ensuring an appropriate balance between computational efficiency and generative performance throughout the diffusion process. This dynamic assignment of selection strategies leverages the strengths of Bayesian optimization to explore the hyper-parameter space and identify the optimal strategies for each denoising step, considering factors such as the relative importance of high- and low-frequency components, the trade-off between computational complexity and output quality, and the specific characteristics of the diffusion model and the targeted application domain.

Experiments

Experiments Setup

To validate the effectiveness of our proposed idea, we assess its efficacy across various existing diffusion models. We intentionally choose diverse models to showcase the versatility of our approach. For example, the pixel-space model and LDM use a U-net structure. Our method is designed to be applicable across all these model architectures. We obtain pretrained weights from the official repository for all models except the pixel-space model. In evaluation, we conduct qualitative and quantitative experiments. In the qualitative assessment, we compare our generated images to the baseline images produced by DDIM with 50 steps in terms of fidelity and latency. For quantitative evaluation, we employed the Fréchet Inception Distance (FID) (Heusel et al. 2017) as a measure of image quality. Specifically, we calculated the FID for Stable Diffusion (SD) (Rombach et al. 2022b) and SDXL (Podell et al. 2023) using 5k samples from the MS-COCO dataset (Lin et al. 2014). This quantitative analysis allowed us to objectively compare the performance of our approach against the baseline models. All experiments are conducted on a GPU server equipped with an NVIDIA GeForce RTX 4090. Latency measurements are performed using PyTorch with a batch size of 1 on the GeForce RTX 4090.

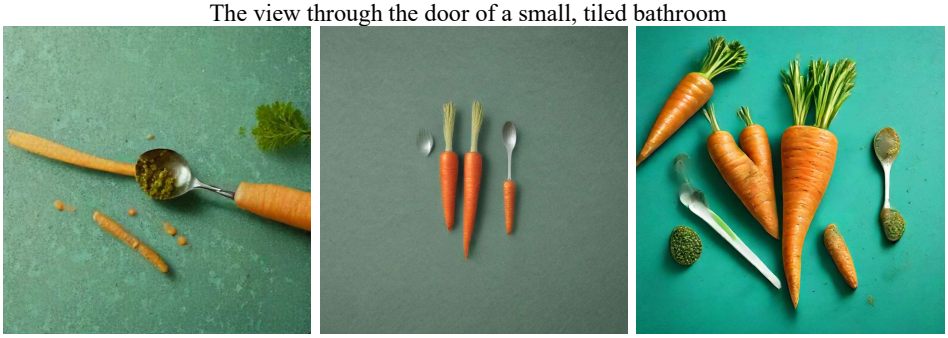
Results on Different Diffusion Models

Results presented in Figure 3 demonstrate the effectiveness of our proposed FreqTS method in reducing the computational requirements of diffusion models while maintaining their generative performance. By leveraging frequency-aware token selection and Bayesian selection strategy optimization, FreqTS achieves significant computational savings across various diffusion models, as evident from the reduced multiply-accumulate operations (MACs), parameter counts, and inference latencies. Focusing on the Stable Diffusion models, we observe that FreqTS reduces the computational complexity for SD-1.5, SD-2.1 , and SD-XL. Notably, these substantial reductions in computational requirements are accompanied by only a modest increase in the zero-shot 10K-FID metric on the MS-COCO dataset, indicating that FreqTS maintains the generative performance of the models. For instance, the FID of SD-1.5 increases from 23.9 to 20.8, SD-2.1 from 22.6 to 19.9, and SD-XL from 24.6 to 22.7, which are relatively small trade-offs considering the significant computational savings achieved. The effectiveness of FreqTS extends beyond the Stable Diffusion models, as evidenced by its performance on PixArt-Alpha and LCM acceleration methods. For PixArt-Alpha, FreqTS reduces the inference latency by 38.4% (from 61.5s to 37.9s), while maintaining a competitive FID of 35.8 compared to the baseline 38.7. When combined with LCM, FreqTS further reduces the computational requirements, demonstrating its compatibility with other acceleration techniques. These results highlight the ability of FreqTS to significantly reduce the amount of model computation required by diffusion models, leading to faster inference times and potentially enabling deployment on resource-constrained devices. Crucially, FreqTS achieves these computational savings without compromising the generative performance of the models, as evidenced by the maintained FID scores across various diffusion models and acceleration techniques.

Comparison and Analysis

In Figure 4, we compare our proposed method with existing diffusion acceleration techniques, highlighting its ability to reduce computational requirements while maintaining competitive generative performance. Notably, our method does not require retraining the diffusion model, unlike approaches such as SDXL-Turbo, which incur additional training costs. By adjusting the token sort threshold, our method offers a

The view through the door of a small, tiled bathroom



A spoon, a large carrot, a medium carrot and a small carrot, on blue-green speckled surface

Figure 5: More comparison between SDXL Turbo, SDXL Lightning, SDXL with our FreqTS.

trade-off between computational efficiency and generative quality. At a 40% token sort threshold, our method achieves a latency of 4.183, outperforming the widely-used DDIM method while maintaining a competitive FID score of 6.40. As we increase the token sort threshold to 70%, the latency is further reduced to 2.492, albeit with a slightly higher FID of 7.83. Notably, when combined with our proposed Bayesian selection strategy optimization, our method achieves a remarkable balance of performance and efficiency. With a latency of 2.914 and a low FID of 6.20, our method outperforms existing techniques like DDIM, DPM-Solver++, and DeepCache in terms of both computational efficiency and generative quality. Additionally, our method compares favorably to the SDXL-Turbo approach, which requires retraining and has a higher FID of 22.58. These results demonstrate the effectiveness of our method in reducing the computational burden of diffusion models without sacrificing generative quality. By leveraging frequency-aware token selection and Bayesian optimization, our method adaptively balances computational efficiency and generative performance, making it a promising solution for deploying diffusion models in resource-constrained environments or accelerating existing models without compromising their generative capabilities.

Visualisation Analysis The comparative analysis in Fig. 5 focuses on the visual quality of four different image processing techniques applied to the same subject matter. The image features a spoon, along with large, medium, and small carrots arranged on a blue-green speckled surface. The findings clearly demonstrate that SDXL with our FreqTS surpasses SDXL Turbo and SDXL Lightning in terms of detail and clarity. Notably, the carrots exhibit enhanced intricacies, with their features more pronounced and distinct. Moreover, the spoon's texture appears smoother, and the overall image quality is notably crisper with SDXL and our FreqTS. These observations strongly indicate that SDXL with our FreqTS excels in preserving fine details and enhancing image sharpness, solidifying its superiority over the other evaluated methods.

Conclusion

In conclusion, our proposed FreqTS approach offers a promising solution for accelerating diffusion models without retraining. The frequency-aware token selection approach can be further refined and expanded by exploring alternative sorting methods and incorporating additional metrics. Additionally, optimizing the dynamic assignment of selection strategies through the Bayesian hyper-parameter search could enhance adaptability and performance. Collaborative efforts can focus on integrating FreqTS with other acceleration techniques, fostering hybrid approaches that leverage multiple strengths. Our experiments demonstrate significant speedup across various diffusion models, showcasing the effectiveness of FreqTS. Our FreqTS represents a significant advancement in addressing computational challenges in diffusion models. Its success inspires future research, optimization strategies, and interdisciplinary collaborations. By continuing to innovate, we can enable wider adoption of diffusion models in resource-constrained scenarios, driving progress in generative modeling across diverse industries.

References

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2023. Token Merging: Your ViT but Faster. In *ICLR*.

Dong, P.; Li, L.; Tang, Z.; Liu, X.; Pan, X.; Wang, Q.; and Chu, X. 2024. Pruner-Zero: Evolving Symbolic Pruning Metric from Scratch for Large Language Models. In *ICML*.

Dong, P.; Li, L.; and Wei, Z. 2023. DisWOT: Student Architecture Search for Distillation WithOut Training. In *CVPR*.

Dong, P.; Li, L.; Wei, Z.; Niu, X.; Tian, Z.; and Pan, H. 2023. Emq: Evolving training-free proxies for automated mixed precision quantization. In *ICCV*.

Dong, P.; Li, L.; Zhong, Y.; Du, D.; Fan, R.; Chen, Y.; Tang, Z.; Wang, Q.; Xue, W.; Guo, Y.; et al. 2025. STBLLM: Breaking the 1-Bit Barrier with Structured Binary LLMs. In *ICLR*.

Fang, G.; Ma, X.; and Wang, X. 2023. Structural pruning for diffusion models. *arXiv preprint arXiv:2305.10924*.

Gu, H.; Li, W.; Li, L.; Qiyuan, Z.; Lee, M.; Sun, S.; Xue, W.; and Guo, Y. 2025. D^2-MoE: Delta Decompression for MoE-based LLMs Compression. *arXiv preprint arXiv:2502.17298*.

He, Y.; Liu, J.; Wu, W.; Zhou, H.; and Zhuang, B. 2024. EfficientDM: Efficient Quantization-Aware Fine-Tuning of Low-Bit Diffusion Models. In *ICLR*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. 6840–6851.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *ICLR*. OpenReview.net.

Li, L. 2022. Self-Regulated Feature Learning via Teacher-free Feature Distillation. In *ECCV*.

Li, L.; Bao, Y.; Dong, P.; Yang, C.; Li, A.; Luo, W.; Liu, Q.; Xue, W.; and Guo, Y. 2024a. DetKDS: Knowledge Distillation Search for Object Detectors. In *ICML*.

Li, L.; Dong, P.; Li, A.; Wei, Z.; and Yang, Y. 2024b. Kd-zero: Evolving knowledge distiller for any teacher-student pairs. *NeuIPS*.

Li, L.; Dong, P.; Wei, Z.; and Yang, Y. 2023a. Automated knowledge distillation via monte carlo tree search. In *ICCV*.

Li, L.; and He, J. 2024. Bidirectional Consistency Models. *arXiv preprint arXiv:2403.18035*.

Li, L.; and Jin, Z. 2022. Shadow Knowledge Distillation: Bridging Offline and Online Knowledge Transfer. In *NeuIPS*.

Li, L.; Peijie; Tang, Z.; Liu, X.; Wang, Q.; Luo, W.; Xue, W.; Liu, Q.; Chu, X.; and Guo, Y. 2024c. Discovering Sparsity Allocation for Layer-wise Pruning of Large Language Models. In *NeuIPS*.

Li, L.; Sun, H.; Li, S.; Dong, P.; Luo, W.; Xue, W.; Liu, Q.; and Guo, Y. 2024d. Auto-gas: Automated proxy discovery for training-free generative architecture search. In *ECCV*.

Li, L.; Wei, Z.; Dong, P.; Luo, W.; Xue, W.; Liu, Q.; and Guo, Y. 2024e. Attnzero: efficient attention discovery for vision transformers. In *European Conference on Computer Vision*, 20–37. Springer.

Li, S.; Hu, T.; Khan, F. S.; Li, L.; Yang, S.; Wang, Y.; Cheng, M.; and Yang, J. 2023b. Faster Diffusion: Rethinking the Role of UNet Encoder in Diffusion Models. *arXiv*.

Li, W.; Li, L.; Huang, Y.-L.; Lee, M. G.; Sun, S.; Xue, W.; and Guo, Y. 2025. Structured Mixture-of-Experts LLMs Compression via Singular Value Decomposition.

Li, W.; Li, L.; Lee, M.; and Sun, S. 2024f. ALS: Adaptive Layer Sparsity for Large Language Models via Activation Correlation Assessment. In *NeuIPS*.

Li, X.; Lian, L.; Liu, Y.; Yang, H.; Dong, Z.; Kang, D.; Zhang, S.; and Keutzer, K. 2023c. Q-Diffusion: Quantizing Diffusion Models. *ICCV*.

Lin, S.; Wang, A.; and Yang, X. 2024. SDXL-Lightning: Progressive Adversarial Diffusion Distillation. arXiv:2402.13929.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. 740–755.

Liu, X.; Li, L.; Li, C.; and Yao, A. 2023. NORM: Knowledge Distillation via N-to-One Representation Matching. In *ICLR*.

Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023. Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference. arXiv:2310.04378.

Ma, X.; Fang, G.; and Wang, X. 2023. DeepCache: Accelerating Diffusion Models for Free. *arXiv*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *CoRR*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *ICLR*. OpenReview.net.

Ren, Y.; Xia, X.; Lu, Y.; Zhang, J.; Wu, J.; Xie, P.; Wang, X.; and Xiao, X. 2024. Hyper-SD: Trajectory Segmented Consistency Model for Efficient Image Synthesis. arXiv:2404.13686.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Salimans, T.; and Ho, J. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

Sauer, A.; Lorenz, D.; Blattmann, A.; and Rombach, R. 2023. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*.

So, J.; Lee, J.; Ahn, D.; Kim, H.; and Park, E. 2023. Temporal Dynamic Quantization for Diffusion Models. In *NeurIPS*.

Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency Models. In *ICML*, volume 202.

Sun, H.; Li, L.; Dong, P.; Wei, Z.; and Shao, S. 2024. Auto-DAS: Automated Proxy Discovery for Training-free Distillation-aware Architecture Search. ECCV.

Wang, C.; Wang, Z.; Xu, X.; Tang, Y.; Zhou, J.; and Lu, J. 2023. Towards Accurate Data-free Quantization for Diffusion Models. *arXiv preprint arXiv:2305.18723*.