Training-Free and Hardware-Friendly Acceleration for Diffusion Models via Similarity-based Token Pruning

Evelyn Zhang[1], **Jiayi Tang**[2], **Xuefei Ning**[3], **Linfeng Zhang**[1*]
[1]School of Artificial Intelligence, Shanghai Jiao Tong University
[2]School of Computer Science and Technology, China University of Mining and Technology
[3]Department of Electronic Engineering, Tsinghua University
zhanglinfeng@sjtu.edu.cn

Abstract

The excellent performance of diffusion models in image generation is always accompanied by overlarge computation costs, which have prevented the application of diffusion models in edge devices and interactive applications. Previous works mainly focus on using fewer sampling steps and compressing the denoising network of diffusion models, while this paper proposes to accelerate diffusion models by introducing SiTo, a similarity-based token pruning method that adaptive prunes the redundant tokens in the input data. SiTo is designed to maximize the similarity between model prediction with and without token pruning by using cheap and hardware-friendly operations, leading to significant acceleration ratios without performance drop, and even sometimes improvements in the generation quality. For instance, the zero-shot evaluation shows SiTo leads to 1.90x and 1.75x acceleration on COCO30K and ImageNet with 1.33 and 1.15 FID reduction at the same time. Besides, SiTo has no training requirements and does not require any calibration data, making it plug-and-play in real-world applications.

Code — https://github.com/EvelynZhang-epiclab/SiTo

Introduction

Diffusion models (Dhariwal and Nichol 2021; Ho, Jain, and Abbeel 2020) have led significant breakthroughs in content generation in the last several years, which boost diverse downstream applications such as image-to-image translation (Choi et al. 2021; Saharia et al. 2022a), text-to-image generation (Saharia et al. 2022b; Rombach et al. 2022; Podell et al. 2023), text-to-video generation (Ho et al. 2022; Luo et al. 2023), image editing (Avrahami, Lischinski, and Fried 2022; Kawar et al. 2023). However, most diffusion models contain a huge number of parameters in the denoising network and require multiple timesteps in the sampling period, significantly increasing their computation costs and making them unaffordable in real-world applications.

Abundant recent works have been introduced to accelerate the sampling of diffusion models. For instance, step distillation and fast sampling methods such as DDIM (Song, Meng, and Ermon 2020) have been introduced to reduce the number of sampling steps. Quantization (Shang et al. 2023a,b; Zhao et al. 2024), pruning (Li et al. 2023), and distillation (Kim et al. 2023a) methods have been utilized to compress the denoising network in the diffusion models. These methods focus on eliminating the computational redundancy in the sampling process and the parameter space. However, the size of the input data for diffusion models, which is also an important contributor to the computation complexity, still has not been well-studied in previous works.

Token reduction, which aims to reduce the size of the input data, has shown great acceleration performance in classification models (Kim et al. 2024). For image generation, ToMeSD (Bolya and Hoffman 2023) and AT-EDM (Wang et al. 2024) are two pioneering works that focus on merging and pruning the redundant tokens, respectively. However, ToMeSD is a direct application of traditional token reduction methods (Bolya et al. 2022) for classification and does not consider the property for diffusion models and generative tasks. AT-EDM, which leverages the graph algorithm to select the redundant tokens, has uncontrolled converge time, is unfriendly to hardware, and does not support batch-wise computation, making it not practical in real-world applications. The problems of previous methods introduce our target: *performing high-ratio token reduction while preserving the generation quality, by using low-cost, hardware-friendly, and training-free operations*.

To this end, we start by analyzing the difference between token reduction for classification and generation and find that their major difference is that *the pruned tokens in generative tasks are required to be recovered.* Different from the classification task where only one or several class tokens are utilized for the final prediction, all the tokens in generative tasks are indispensable since each token corresponds to each patch in the image, which raises the requirement of pruned token recovery. Specifically, in generative tasks, after pruning the redundant tokens in the input data and feeding the unpruned tokens to the neural layers, we have to recover these pruned tokens by copying their most similar unpruned tokens. The recovery error between the real computation results and the recovered results on the pruned tokens represents the error introduced by token pruning.

To minimize the recovery errors, we propose **Si**milarity-based **To**ken pruning (SiTo), which aims to maximize the similarity between the pruned tokens and the tokens used to recover them. Specifically, SiTo has a three-stage pipeline.

*corresponding author

- SiTo carefully selects a set of *base tokens* which are utilized as the base to select and recover the pruned tokens.
- SiTo selects the tokens that have the highest similarity to the base tokens as the pruned tokens.
- SiTo feeds the unpruned tokens to the neural layers and recovers the pruned tokens by directly copying their most similar base tokens.

The major contribution of SiTo is to introduce the concept of *base tokens*, which play essential roles in both pruned token selection and recovery. The base tokens are selected based the following principles.

(I) Maximal Similarity: The base tokens are selected as a set of tokens that have the highest similarity with all the other tokens. Besides, after the selection of base tokens, we further choose the tokens that have the highest similarity to base tokens as the tokens for pruning. This selection strategy for base tokens and pruning tokens ensures the high similarity, *i.e.* small difference between the base tokens and the pruned tokens, further minimizing the error of recovering pruned tokens from base tokens.

(II) Uniform Spatial Distribution: Classical research in image analysis reveals that the patches in the same region of the images carry similar information, implying that the tokens that are adjacent in the spatial dimension may have similar representation, and hence recovering the pruned tokens with their spatially adjacent tokens is more suitable than recovering them with distant tokens. Hence, instead of directly selecting base tokens from all the tokens in the image, we select one base token in each local region of the image to guarantee that the base tokens are uniformly distributed in different spatial positions of the image. Besides, since the base tokens will not be pruned, this solution also ensures that there will be at least one token not pruned in each region, thus ensuring the error introduced by token pruning is not overly concentrated.

(III) Selection with Randomness: The previous two principles effectively minimize the recovery error led by token pruning in a single denoising step. However, the sampling process of diffusion models contains multiple denoising steps and the adjacent steps have similar token representations (Ma, Fang, and Wang 2024). As a result, the choice of base tokens and pruned tokens can be very similar and even identical in adjacent timesteps. Moreover, as shown in Fig. 1(b), since the pruned tokens are recovered by directly copying from their most similar base tokens, the pruned tokens tend to maintain a higher similarity to the base tokens in all the future denoising steps, and hence they are very likely to be pruned in almost all the timesteps. This extremely unbalanced token pruning may lead to a significant performance drop in generation quality. To solve this problem, we propose to add Gaussian noise to the similarity of different tokens, which introduces randomness during base token selection. As shown in Fig. 1(a), this solution reduces the ratio of pruning the same tokens in adjacent two timesteps from 97% to 72%, effectively avoiding unbalanced pruning.

The main difference between SiTo and previous methods (Bolya and Hoffman 2023; Wang et al. 2024) is twofold.

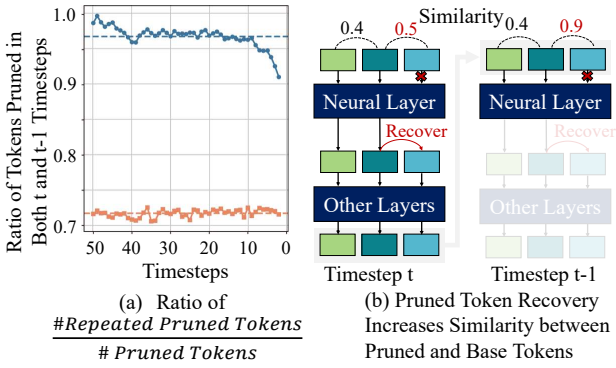

Figure 1: Motivation of **(III) Selection with Randomness**. (a) The ratio of tokens repeated pruned in adjacent timesteps: randomness introduced by Gaussian noise solves the unbalanced pruning problem. (b) Explanation on why unbalanced pruning happens: recovering the pruned tokens from the base tokens leads to higher similarity, making the pruned tokens still be pruned in the next timestep.

Firstly, SiTo carefully selects a set of base tokens that are responsible for the pruned token selection and the pruned token recovery, which helps reduce the error introduced by token pruning. Secondly, SiTo does not introduce any complex computation, especially not leveraging the attention scores in self-attention and cross-attention that are utilized in previous work (Wang et al. 2024). This strategy reduces the additional computation costs introduced by SiTo and brings benefits in memory footprint reduction. Extensive experiment results demonstrate the effectiveness of SiTo on Stable Diffusion (SD) on both ImageNet and COCO30K, offering the following advantages:

- **Lossless Acceleration:** SD v1.5 applied with SiTo achieves 1.9 times acceleration and 2.7 times memory compression measured on hardware with 1.33 FID reduction on ImageNet, which indicates acceleration, compression, and generation quality improvements at the same time. Besides, SiTo cooperates well with other acceleration methods such as fast sampler.
- **Hardware-Friendliness:** The operations introduced by SiTo contain extremely low computational costs and no additional memory footprint. They are well-supported by CUDA and practical for parallel computation.
- **Training-Free and Data-Free:** The operations introduced by SiTo are non-parametric, resulting in no requirements for training. Besides, SiTo generalizes well in different datasets, models, and sampling settings and does not require any validation data.

Related Work

Fast Stable Diffusion Models

Diffusion models (Sohl-Dickstein et al. 2015; Song and Ermon 2019; Dhariwal and Nichol 2021; Ho, Jain, and Abbeel 2020) create images by progressively removing noise from an initial noisy input through a series of diffusion steps. The latent diffusion model (Rombach et al. 2022), like many

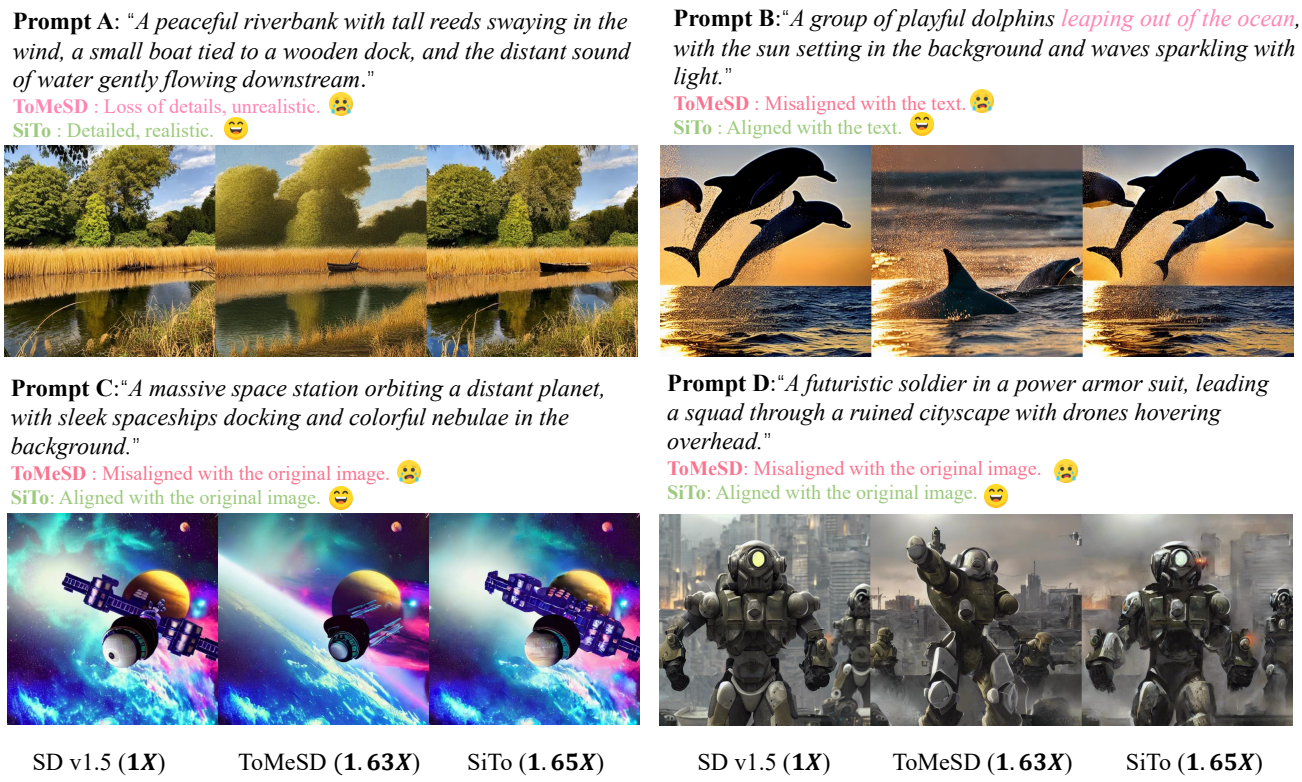

Figure 2: Visual comparisons with the manually crafted challenging prompts. We apply ToMeSD (Bolya and Hoffman 2023) and SiTo on stable diffusion v1.5, achieving similar speed-up ratios of 1.63 and 1.65, respectively. Under these comparable conditions, our method generated more realistic, detailed images that better aligned with the original images and text prompts.

modern large diffusion models, uses a U-Net (Ronneberger, Fischer, and Brox 2015) architecture with transformer-based blocks. Recently, abundant methods have been proposed to accelerate diffusion models by using a few steps for sampling and compressing the denoising network. Distillation-based methods have been introduced to distill the knowledge from multi-step teacher to few-step and even single-step student (Meng et al. 2023). Fast samplers have been introduced to directly reduce the sampling steps without training (Song, Meng, and Ermon 2020). Recent efforts to compress U-Net in diffusion models include quantization (Shang et al. 2023a,b), pruning (Li et al. 2023), and distillation (Kim et al. 2023b). Most methods, except fast samplers, require time-consuming retraining, highlighting the need for more efficient, retraining-free approaches.

Token Reduction

Token Reduction Techniques Token efficiency in Vision Transformers (ViTs) is achieved through both learned and heuristic methods. Learned approaches, such as DynamicViT (Rao et al. 2021) and A-ViT (Yin et al. 2022), use auxiliary models to rank and eliminate redundant tokens, with DynamicViT employing an MLP for pruning masks and A-ViT leveraging feature channels and auxiliary losses. In contrast, heuristic methods like Token Pooling (Marin et al. 2021) provide practical solutions without extensive training. Adaptive-Token Sampling (Fayyaz et al. 2022) selects tokens based on cls token similarity in attention maps, but this dependency hampers its effectiveness in dense prediction tasks like image generation.

Training-free Efficient ViT Token Merging (Bolya et al. 2022), as the pioneer of token merging methods, introduces a novel training-free approach that averages similar tokens based on an efficient bipartite matching algorithm. Going further, Token Fusion (Kim et al. 2024) synergizes the advantages of token merging and pruning. It divides tokens into two groups based on their linear correlation in high-dimensional space: one group undergoes pruning while the other undergoes merging. Additionally, Token Fusion optimizes the merging process with MLERP merging, addressing the inherent limitations of average merging.

Training-free Efficient DMs Training-free efficiency strategies are underexplored in diffusion models (DMs). ToMeSD (Bolya and Hoffman 2023) and AT-EDM (Wang et al. 2024) are two pioneering works that focus on merging and pruning the redundant tokens. ToMeSD relies on classification-focused token reduction methods (Bolya et al. 2022), overlooking diffusion models' generative characteristics, while AT-EDM's graph-based approach suffers from unpredictable convergence and lacks batch-wise computation, limiting real-world applicability.

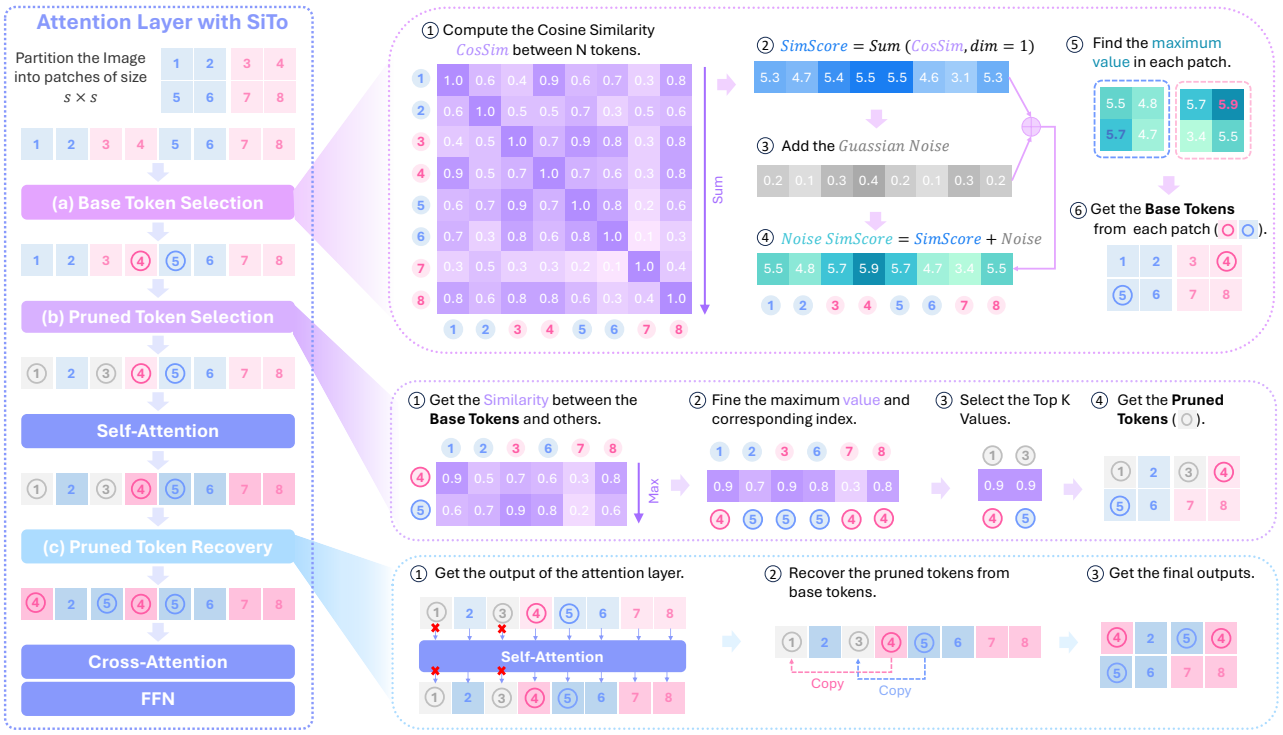

Figure 3: The pipeline of SiTo on the example of self-attention. *(a) Base Token Selection:* We compute the Cosine Similarity between all the tokens. For each token, we sum its similarity to all the tokens as the SimScore. Then, Gaussian Noise is added to the SimScore introduces randomness, preventing identical base and pruned token choices across timesteps. Finally, the token that has the highest Noise SimScore in an image patch is selected as a **base token**. *(b) Pruned Token Selection:* The tokens with the highest similarity to the base tokens are selected as **pruned tokens**. *(c) Pruned Token Recovery:* The unpruned tokens are fed to the neural layers. Then, the pruned tokens are recovered by copying from their most similar base tokens.

Methodology

Preliminary

Diffusion Models A diffusion model is defined with a diffusion process that adds Gaussian noise to a real image and a denoising process that iteratively performs image denoising from a standard Gaussian noise to a real image. The UNet-style denoising network is composed of L transformer down/up blocks and a single mid block, represented as $f = f_{down_1} \cdots \circ f_{down_L} \circ f_{mid} \circ f_{up_L} \circ \ldots f_{up_1}$, is utilized to estimate the noise during the denoising process. As discussed in the previous work (Li et al. 2024), most of the computation costs are distributed in the self-attention layers of f_{down_1} and f_{up_1}.

Token Reduction Given an input as a set of tokens $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ and a neural network f, the original computation process can be written as $f(\mathbf{X})$. Token reduction aims to reduce the length of $\mathbf{X}$ but preserve the prediction results, which can be formulated as

$$\arg\min_{\pi} \|f(\mathbf{X}) - f(\pi(\mathbf{X}))\|, \tag{1}$$

where π denotes the token reduction strategy and $\pi(\mathbf{X})$ is the reduced token set satisfying $|\pi(\mathbf{X})| \leq |\mathbf{X}|$ where $|\cdot|$ indicates the cardinality of the set. For token reduction, $\pi(\mathbf{X}) \subset \mathbf{X}$ is a subset of the original token set. For token merge, $\pi(\mathbf{X}) = \{(x_i + x_j)/2\}_{i \neq j}$ is obtained by averaging two most similar tokens in $\mathbf{X}$.

Similarity-based Token Pruning

As shown in Fig.3, our proposed similarity-based token pruning contains three stages, including base token selection, pruned token selection, and pruned token recovery.

Base Token Selection

In the base token selection stage, we identify a critical subset of tokens as base tokens for SiTo. These base tokens are preserved during computation and are used to determine which tokens to prune and how to recover them.

Cosine Similarity: In the case of diffusion models, the input data $\mathbf{X}$ is a set of image tokens with shape (B, N, C), where B is the batch size, N represents the number of tokens, and C denotes the size of the channels. In the stage of base token selection, we start by computing the cosine similarity between each input token in each image as CosSim$\in \mathbb{R}^{B \times N \times N}$, where $\text{CosSim}_{b,i,j}$ is the cosine similarity between the i_{th} token and the j_{th} token for the b_{th} sample in the batch, which has

$$\text{CosSim}_{b,i,j} = \text{CosSim}(\mathbf{X}_{b,i,j}) = \frac{\mathbf{X}_{b,i} \cdot \mathbf{X}_{b,j}}{\|\mathbf{X}_{b,i}\| \|\mathbf{X}_{b,j}\|} \tag{2}$$

Similarity Score: Then, for each token, we compute the sum of its cosine similarity to the other tokens, which can be formulated as

$$\text{SimScore}_{b,i} = \sum_{j=1}^{N} \text{CosSim}_{b,i,j} \tag{3}$$

This results in a matrix $\mathbf{SimScore} \in \mathbb{R}^{B\times N}$, where each entry $\mathbf{SimScore}_{b,i}$ represents the aggregated similarity score for the i-th token in the b-th batch. A higher similarity score indicates this token is more similar to the other tokens in the image, and hence this token is more suitable to be utilized as the base for recovering the other tokens.

Selection with Randomness: The cosine similarity is a mathematically good metric to decide the choice of base tokens. However, as discussed in many previous works (Kim et al. 2023a; Ma, Fang, and Wang 2024), the diffusion models have very similar representation in different timesteps, which results in similar CosSim, and further leads to similar choices of the base tokens and the pruned tokens. Furthermore, since the pruned tokens are decided by their similarity to the base token while the pruned tokens are then recovered by directly copying from the base token, the similarity between the pruned tokens and the base tokens can be greatly increased in each sampling timestep. Consequently, certain tokens may be frequently pruned throughout all timesteps, resulting in a significant decline in generation quality.

To address this problem, we propose to introduce random noise in the selection of base tokens by adding Gaussian noise to the cosine similarity, which can be formulated as

$$\text{Noisy SimScore}(\mathbf{X}) = \text{SimScore} + \epsilon \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ indicates the Gaussian noise and σ^2 indicates its variance. The Gaussian noise introduce randomness into the selection of base tokens and pruned tokens in order to avoid over-pruning on certain tokens. This simple solution is cheap in computation but leads to significant improvements in the generation quality (see Fig.4).

Patch-based Token Selection We then identify the base tokens in each patch of the image. Instead of directly selecting the tokens with the highest SimScore in the whole input data, we further add the constraint that the base token should be uniformly distributed in the spatial positions on the image. Hence, for each patch in the image, we select the token with the highest SimScore as the base token in this patch. Specifically, we first reshape the similarity score matrix SimScore into a 2D grid format with dimensions $(B, \sqrt{N}, \sqrt{N})$. This transformation maps the sequence of tokens into their corresponding patches in an image. Subsequently, we partition the grid into non-overlapping patches of size $s \times s$. Specifically, the matrix is divided into $\sqrt{N}/s \times \sqrt{N}/s$ patches, resulting in each patch encompassing $s \times s$ tokens. Within each patch, we select the token with the maximum similarity score in SimScore as the base token in this patch. Hence, we obtain a sequence of base tokens which can be denoted as $\mathbf{X}_{\text{base}} \subset X$ with $|\mathbf{X}_{\text{base}}| = {}^{|\mathbf{X}|}/_{s^2}$, where its i_{th} element $\mathbf{X}_{\text{base},i}$ can be formulated as

$$\begin{aligned} \mathbf{X}_{\text{base},i} = &\arg\max_{\mathbf{X}_{m,n}} \text{Noisy SimScore}(\mathbf{X}_{m,n}) \\ &\text{s.t.} \quad m \in [is, (i+1)s-1], n \in [is, (i+1)s-1] \end{aligned} \tag{5}$$

Pruned Token Selection

Based on the base tokens, we then select the tokens to be pruned from the left tokens $\mathbf{X} - \mathbf{X}_{base}$. For each token inside this set, we get the highest similarity between this token and all the base tokens as the criterion for whether this token should be pruned. The tokens that are more similar to the base tokens will be pruned since recovering them from base tokens leads to lower reconstruction errors. This process can be formulated as

$$\begin{aligned} \mathbf{X}_{\text{prune}} = &\arg\operatorname{top}\!K_{\mathbf{Xi}} \max_{\mathbf{X}_j} \text{Cosine Similarity}(\mathbf{X}_i, \mathbf{X}_j) \\ &\mathbf{X}_i \in \mathbf{X} - \mathbf{X}_{\text{base}} \quad \text{and} \quad \mathbf{X}_j \in \mathbf{X}_{\text{base}}. \end{aligned} \tag{6}$$

where K denotes the number of pruned tokens. Please note that these cosine similarities have been previously computed in the stage of base token selection in CosSim. Hence, the cosine similarity in Eq.6 can be directly obtained and does not require re-computation. Besides, we prune the same number of tokens for all the images in the same batch and thus our method can be well-supported by GPU.

Pruning Token Recovery

The pruned token sequence $\mathbf{X} - \mathbf{X}_{\text{prune}}$ is then input to the attention layers for diffusion models for computing, resulting in an output sequence $f(\mathbf{X} - \mathbf{X}_{\text{prune}})$ with the same length. Then, we recover the pruned tokens by directly copying them from their most similar base tokens found in Eq.6.

Analysis on the Pruning Error

Without loss of generality, by denoting a base token as x_b and a set of n pruned tokens which are recovered from x_b after pruning as $\{x_i\}_{i=1}^n$, then the error e introduced by SiTo can be formulated as $e = \sum_{i=1}^n d(x_i, x_b)$, where $d(\cdot,\cdot)$ indicates the metric of distance. To minimize e, it is reasonable to make the base token x_b have the highest similarity to all the possible tokens for pruning. Hence, as introduced in Eq.5, SiTo selects the base tokens according to the sum of similarities to all the other tokens. Minimizing e with the given x_b leads to selecting tokens with the highest similarity to x_b, as shown in Eq.6.

Experimental Results

Experimental Setup

Evaluation Our experiments are conducted with SD v1.5 and SD v2 by generating 512×512 images using 50 PLMS (Liu et al. 2022) steps with a cfg scale (Dhariwal and Nichol 2021) of 7.5 and 9.0, respectively. We generate 2,000 images of ImageNet-1k (Deng et al. 2009) (2 per class) and 30,000 images of COCO30k classes (1 per caption) for evaluation. FID is utilized as the metric for generation quality. The average latency for generate an image and speedup are measured on a single 4090 GPU.

Implementation of ToMeSD and SiTo Without special notification, we apply ToMeSD and SiTo only to the first and the last self-attention layers, which account for most of the computations in the UNet. ToMeSD a-e and SiTo a-e denotes ToMeSD and SiTo with different acceleration ratios.

Qualitative Analysis

As shown in Fig.2, we evaluate ToMeSD (Bolya and Hoffman 2023) and SiTo using manually crafted challenging prompts.With similar speed-up ratios on Stable Diffusion v1.5 (1.63x for ToMeSD at a 0.5 merge ratio and 1.65x for SiTo at a 0.6 pruning ratio), our method produces **more realistic, detailed images** and **better alignment with original images and text prompts**. For example, under Prompt A, the image generated by ToMeSD changes to an oil painting style, with the tree textures lost. In Prompt B, although the prompt specifies "a dolphin leaping out of the ocean", the dolphin generated by ToMeSD is still underwater. In Prompts C and, the images generated using the SiTo show a higher similarity to those produced by the original SD v1.5.

Quantitative Evaluations

Evaluation on the ImageNet Dataset We apply ToMeSD and SiTo on SD v1.5 and SD v2. Tab.1 shows that SiTo has a lower FID score, higher speed-up ratio, and lower memory usage compared to ToMeSD across all speedup ratio settings. With a pruning ratio of 0.7, SiTo achieves 1.9 times acceleration and 2.70 times memory compression measured on hardware with 1.33 FID reduction.

Evaluation on the COCO30k Dataset To further validate the effectiveness of SiTo, we conduct experiment on the COCO30K Dataset. As reported in Tab. 1, SiTo achieves an FID score of 11.17 with a 1.75× speed-up on SD v1.5, and an FID of 11.67 with a 1.65× speed-up on SD v2, consistently surpassing ToMeSD in both speed and image quality.

Ablation Study

This section ablates key designs in SiTo, with all experiments conducted on SD v1.5 using the ImageNet dataset.

Base token selection methods We conduct ablation experiments on the following six base token selection methods: **I**. consistently selecting the top-left token within a 2×2 patch, **II**. randomly selecting 25% of tokens globally, **III**. selecting the token with the highest SimScore globally, **IV**. randomly selecting one token within a 2×2 patch, **V**. choosing the token with the highest SimScore within a 2×2 patch, and **VI**. adding Gaussian noise to SimScore and then selecting the token with the highest score within a 2×2 patch (default setting in SiTo). As presented in Tab.3, our experiments yield two key insights:

Temporal Distribution Uniformity: The base token should maintain uniform distribution across different denoising timesteps. Strategy **I**, which involves consistently selecting the same token at each timestep, performs the worst. Strategy **VI** outperforms **V** because the introduction of random noise in **VI** allowed for variation in base token selection across different timesteps. A similar reasoning explains why strategy **II** outperforms **III**. Furthermore, we conduct a visual analysis of Strategy **V** and Strategy **VI** in Fig.4. It illustrates that introducing noise prevents repetitive pruning of certain tokens, resulting in images with richer details, such as the more detailed grass texture and the increased number of flowers observed in Fig.4a.

Spatial Distribution Uniformity : The base tokens should also maintain spatial uniformity. Tab.3 shows that strategies **IV**, **V**, and **VI** outperforms **II** and **III**, indicating that local selection within patches yields better results than global selection. This is likely because global selection can lead to a concentration of base tokens in certain areas, resulting in dense pruning in other regions, leading to significant information loss that is difficult to recover.

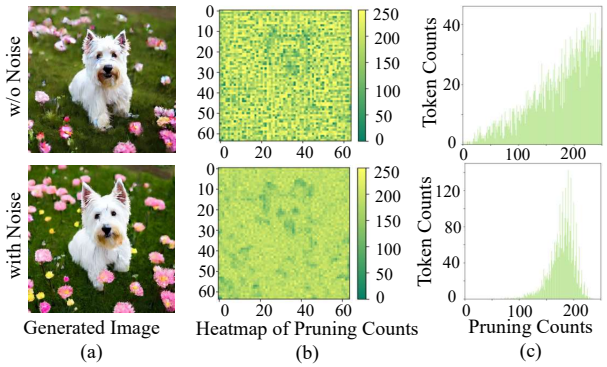

Figure 4: Comparison between our proposed SiTo with Gaussian noise and without Gaussian noise on (a) the generated images, (b) the heatmap of the pruning times for each token, and (c) the distribution pruning times for each token.

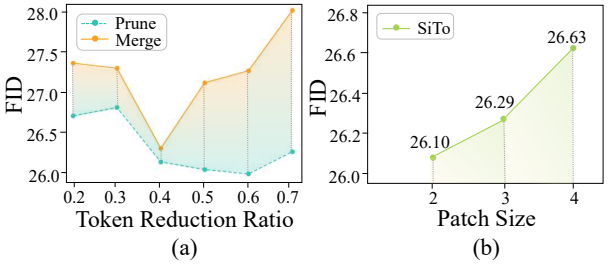

Figure 5: (a) Comparison of merging and pruning operations. (b) The influence of different patch sizes.

Pruning vs. Merging To validate our pruning operation, we replace the pruning operation with an average merging strategy similar to ToMeSD. As shown in Fig.5a, the pruning operation consistently yields lower FID scores. A high token reduction ratio forces a base token to reconstruct multiple tokens, leading to significant errors and deviation from its true distribution, which degrades image quality.

The impact of different patch sizes In the process of selecting base tokens, we partition the features into patches of size $s \times s$, with a single token chosen as the base token within each patch. As the patch size increases, the proportion of base tokens decreases correspondingly. Fig.5b illustrates

Dataset	Method	SD v1.5 FID↓	SD v1.5 Latency (Second)↓	SD v1.5 Speedup Ratio↑	SD v1.5 Memory (GB)↓	SD v1.5 Compression Ratio↑	SD v2 FID↓	SD v2 Latency (Second)↓	SD v2 Speedup Ratio↑	SD v2 Memory (GB)↓	SD v2 Compression Ratio↑
ImageNet	Origin	27.64	2.61	1.00	3.92	1.00	29.36	2.46	1.00	2.69	1.00
	ToMeSD-a	27.24	2.12	1.23	2.18	1.79	29.33	2.04	1.21	2.08	1.22
	SiTo-a	26.83	2.01	1.29	1.85	2.13	28.60	1.96	1.26	1.86	1.45
	ToMeSD-b	27.30	1.89	1.37	2.31	1.69	29.12	1.83	1.34	1.99	1.35
	SiTo-b	26.19	1.80	1.44	1.68	2.33	28.4	1.77	1.39	1.83	1.47
	ToMeSD-c	27.31	1.69	1.54	1.88	2.08	29.21	1.65	1.49	1.71	1.56
	SiTo-c	26.10	1.63	1.60	1.54	2.56	28.02	1.61	1.53	1.60	1.67
	ToMeSD-d	27.20	1.52	1.71	1.69	2.33	29.04	1.51	1.63	1.63	1.67
	SiTo-d	26.05	1.49	1.75	1.58	2.50	27.90	1.49	1.65	1.52	1.75
	ToMeSD-e	27.44	1.38	1.88	1.68	2.33	29.20	1.39	1.77	1.55	1.75
	SiTo-e	26.31	1.37	1.90	1.46	2.70	28.16	1.38	1.78	1.51	1.79
COCO30K	Origin	12.32	2.61	1.00	3.92	1.00	13.68	2.46	1.00	2.69	1.00
	ToMeSD-d	12.15	1.52	1.71	1.69	2.33	13.10	1.51	1.63	1.63	1.67
	SiTo-d	11.17	1.49	1.75	1.58	2.50	11.67	1.49	1.65	1.52	1.75

Table 1: Comparison between the proposed SiTo and ToMeSD with SD v1.5 and SD v2 on ImageNet and COCO30k.

SA	CA	FFN	FID↓	Latency (Second)↓	Speedup Ratio↑
✓	✗	✗	26.19	1.80	1.44
✓	✓	✗	26.49	1.79	1.45
✓	✗	✓	27.54	1.79	1.45
✓	✓	✓	27.46	1.78	1.46

Table 2: SiTo in layers, including self-attention (SA), cross-attention (CA), and feed-forward network (FFN).

Method	FID↓	CLIP Score↑	Latancy (Second)↓	Speedup Ratio ↑
Origin	27.64	17.48	2.61	1.00
ToMeSD	27.30	17.49	1.89	1.37
SiTo (I)	27.01	17.48	1.76	1.47
SiTo (II)	26.51	17.49	1.76	1.47
SiTo (III)	26.76	17.49	1.78	1.47
SiTo (IV)	26.40	17.52	1.77	1.47
SiTo (V)	26.40	17.51	1.79	1.45
SiTo (VI)	26.19	17.53	1.80	1.44

Table 3: Performance of six different base token selection methods in SiTo. VI is the default setting.

that while larger patch sizes yield only a marginal improvement in speed-up, they lead to a substantial increase in FID.

Performance on fewer sampling timesteps We evaluate the performance of ToMeSD and SiTo with fewer sampling timesteps. As shown in Fig.6, SiTo consistently surpasses ToMeSD in both image quality and speedup across different step counts, indicating its orthogonality to diffusion acceleration methods like DDIM (Song, Meng, and Ermon 2020).

The position to apply SiTo SiTo can be applied to all modules, including self-attention, cross-attention, and feed-forward network. We explore SiTo's application across modules to optimize the trade-off between acceleration and image quality. The results in Tab.2 shows that applying SiTo to CA and MLP offers minimal speed-up gains while compromising image generation quality. Furthermore, while SiTo and ToMeSD can be applied to deeper UNet blocks, Fig.7 shows that this results in a decline in generation quality.

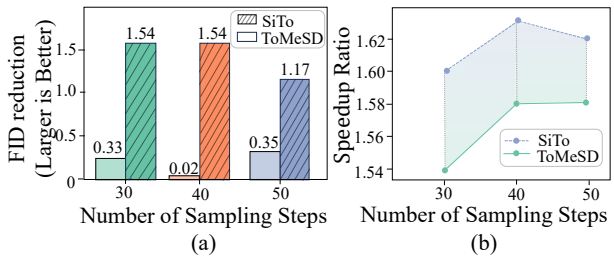

Figure 6: Performance of SiTo with fewer sampling timesteps. (a) FID reduction from SiTo and ToMeSD compared with the original SDv1.5 (larger is better). (b) Corresponding acceleration ratios.

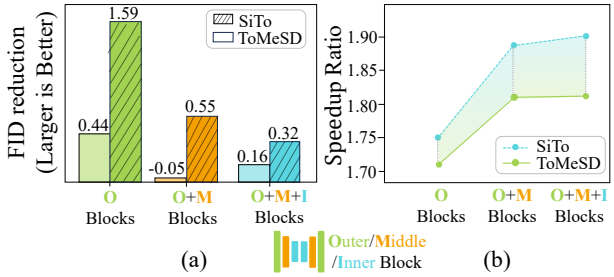

Figure 7: Performance of SiTo and ToMeSD in different blocks of the UNet in SDv1.5.

Conclusion

We present SiTo, a token pruning method for efficient diffusion models, designed to minimize pruning errors through similarity-based selection. SiTo is training-free, hardware-friendly, and improves generation quality while significantly reducing memory and computation costs. It integrates seamlessly into workflows, generalizes across models and datasets, and reveals the redundancy in pretrained diffusion models, paving the way for more efficient designs.

References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18208–18218.

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*.

Bolya, D.; and Hoffman, J. 2023. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4599–4603.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. Ilvr: Conditioning method for denoising diffusion probabilistic models. In 2021 IEEE. In *CVF international conference on computer vision (ICCV)*, volume 1, 2.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Fayyaz, M.; Koohpayegani, S. A.; Jafari, F. R.; Sengupta, S.; Joze, H. R. V.; Sommerlade, E.; Pirsiavash, H.; and Gall, J. 2022. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, 396–414. Springer.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Kim, B.-K.; Song, H.-K.; Castells, T.; and Choi, S. 2023a. BK-SDM: A Lightweight, Fast, and Cheap Version of Stable Diffusion. *arXiv preprint arXiv:2305.15798*.

Kim, B.-K.; Song, H.-K.; Castells, T.; and Choi, S. 2023b. On architectural compression of text-to-image diffusion models.

Kim, M.; Gao, S.; Hsu, Y.-C.; Shen, Y.; and Jin, H. 2024. Token fusion: Bridging the gap between token pruning and token merging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1383–1392.

Li, Y.; Wang, H.; Jin, Q.; Hu, J.; Chemerys, P.; Fu, Y.; Wang, Y.; Tulyakov, S.; and Ren, J. 2023. SnapFusion: Text-to-Image Diffusion Model on Mobile Devices within Two Seconds. *arXiv preprint arXiv:2306.00980*.

Li, Y.; Wang, H.; Jin, Q.; Hu, J.; Chemerys, P.; Fu, Y.; Wang, Y.; Tulyakov, S.; and Ren, J. 2024. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *Advances in Neural Information Processing Systems*, 36.

Liu, L.; Ren, Y.; Lin, Z.; and Zhao, Z. 2022. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*.

Luo, Z.; Chen, D.; Zhang, Y.; Huang, Y.; Wang, L.; Shen, Y.; Zhao, D.; Zhou, J.; and Tan, T. 2023. Videofusion: Decomposed diffusion models for high-quality video generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10209–10218. IEEE.

Ma, X.; Fang, G.; and Wang, X. 2024. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15762–15772.

Marin, D.; Chang, J.-H. R.; Ranjan, A.; Prabhu, A.; Rastegari, M.; and Tuzel, O. 2021. Token pooling in vision transformers. *arXiv preprint arXiv:2110.03860*.

Meng, C.; Rombach, R.; Gao, R.; Kingma, D.; Ermon, S.; Ho, J.; and Salimans, T. 2023. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14297–14306.

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. *arXiv preprint arXiv:2307.01952*.

Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C.-J. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34: 13937–13949.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, 1–10.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Shang, Y.; Yuan, Z.; Xie, B.; Wu, B.; and Yan, Y. 2023a. Post-training quantization on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1972–1981.

Shang, Y.; Yuan, Z.; Xie, B.; Wu, B.; and Yan, Y. 2023b. Post-training quantization on diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1972–1981.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Wang, H.; Liu, D.; Kang, Y.; Li, Y.; Lin, Z.; Jha, N. K.; and Liu, Y. 2024. Attention-Driven Training-Free Efficiency Enhancement of Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16080–16089.

Yin, H.; Vahdat, A.; Alvarez, J. M.; Mallya, A.; Kautz, J.; and Molchanov, P. 2022. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10809–10818.

Zhao, T.; Ning, X.; Fang, T.; Liu, E.; Huang, G.; Lin, Z.; Yan, S.; Dai, G.; and Wang, Y. 2024. MixDQ: Memory-Efficient Few-Step Text-to-Image Diffusion Models with Metric-Decoupled Mixed Precision Quantization. *arXiv preprint arXiv:2405.17873*.