

Token Merging for Fast Stable Diffusion

Daniel Bolya Judy Hoffman
 Georgia Tech
 {dbolya, judy}@gatech.edu

Abstract

The landscape of image generation has been forever changed by open vocabulary diffusion models. However, at their core these models use transformers, which makes generation slow. Better implementations to increase the throughput of these transformers have emerged, but they still evaluate the entire model. In this paper, we instead speed up diffusion models by exploiting natural redundancy in generated images by merging redundant tokens. After making some diffusion-specific improvements to Token Merging (ToMe), our ToMe for Stable Diffusion can reduce the number of tokens in an existing Stable Diffusion model by up to 60% while still producing high quality images without any extra training. In the process, we speed up image generation by up to $2\times$ and reduce memory consumption by up to $5.6\times$. Furthermore, this speed-up stacks with efficient implementations such as xFormers, minimally impacting quality while being up to $5.4\times$ faster for large images. Code is available at <https://github.com/dbolya/tomesd>.

1. Introduction

With the rise of powerful diffusion [4, 21] models such as DALL-E 2 [13], Imagen [18], and Stable Diffusion [15], generating high quality images has never been easier. However, running these models can be expensive, especially for large images. All of these methods function by denoising images through several evaluations of a transformer [22] backbone, meaning that computation scales with the *square* of the number of tokens (and thus also the square of pixels).

Several existing methods to speed up transformers have already been successfully applied to open-source diffusion models such as Stable Diffusion. Flash Attention [2] computes attention efficiently by cleverly accounting for memory bandwidth. XFormers [8] contains several optimized implementation of transformer components. And as of PyTorch 2.0, these optimizations are natively available [12].

However, none of these approaches *reduce* the amount of work necessary—they still evaluate the transformer on every token. Most images, including those generated by



Figure 1. **Token Merging for Stable Diffusion.** When applied properly, ToMe [1] can significantly increase the speed of image generation without jeopardizing quality. Moreover, its benefits *stack* with existing methods such as xFormers [8]. With ToMe and xFormers together, this 2048×2048 image generated in just 28 seconds on a 4090, which is $5.4\times$ faster than the original model.

diffusion models, have a high amount of *redundancy*. And thus, performing computation on *every token* is a waste of resources. Recent work in token reduction such as token pruning [7, 11, 14, 23] and token merging [1, 10, 17] have shown the ability to *remove* these redundant tokens in transformers to speed up evaluation with a small accuracy drop.

Though most of these methods require re-training the model (which would be prohibitively expensive for e.g., Stable Diffusion), Token Merging (ToMe) [1] stands out in particular by *not requiring any additional training*. While the authors only apply it to ViT [5] for classification, they claim that it should also work for downstream tasks.

In this paper, we put that to the test by applying ToMe to Stable Diffusion. Out of the box, a naïve application can speed up diffusion by up to $2\times$ and reduce memory consumption by $4\times$ (Tab. 1), but the resulting image quality suffers greatly (Fig. 3). To address this, we introduce new techniques for token partitioning (Fig. 5) and perform several experiments to decide how to apply ToMe (Tab. 3). As a result, we can keep the speed and improve the memory benefits of ToMe, while producing images extremely close to the original model (Fig. 6, Tab. 4). Furthermore, this speed-up *stacks* with implementations such as xFormers (Fig. 1).

2. Background

In this work, our goal is to speed up an off-the-shelf Stable Diffusion [15] model *without training* using ToMe [1].

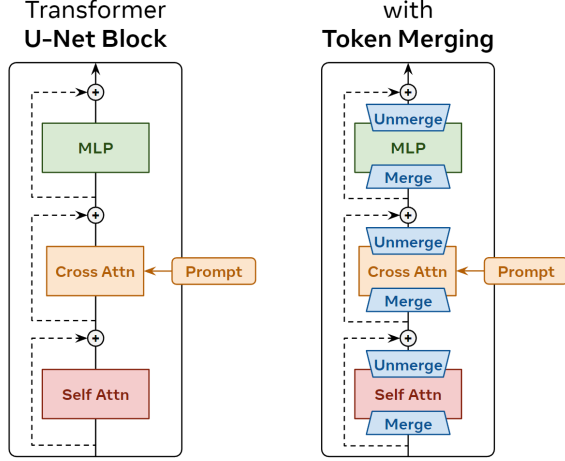


Figure 2. **A U-Net Block** with ToMe applied. Stable Diffusion [15] uses a U-Net [16] model with transformer-based blocks [22]. We apply ToMe [1] by merging tokens before each component of the block and unmerging after to reduce compute costs (Sec. 3).

Stable Diffusion. Diffusion models [4, 20, 21] generate images by repeatedly denoising some initial noise over some number of diffusion steps. Like most modern large diffusion models, Stable Diffusion uses a U-Net [16] with transformer-based blocks. Thus, it first encodes the current noised image as a set of tokens, then passes it through a series of transformer blocks. Each transformer block has the standard self attention [22] and multi-layer perceptron (mlp) modules, with the addition of a cross attention module to condition on the prompt (see Fig. 2).

Token Merging. Token Merging (ToMe) [1] reduces the number of tokens in a transformer gradually by merging r tokens in each block. To do this efficiently, it partitions the tokens into a **source (src)** and **destination (dst)** set. Then, it merges the r most similar tokens from **src** into **dst**, reducing the number of tokens by r , making the next block faster.

3. Token Merging for Stable Diffusion

While ToMe as described in Sec. 2 works well for classification, it’s not entirely straightforward to apply it to a dense prediction task like diffusion. While classification only needs a single token to make a prediction, diffusion needs to know the noise to remove for *every token*. Thus, we need to introduce the concept of *unmerging*.

3.1. Defining Unmerging

While other token reduction methods such as pruning (e.g., [14]) *remove* tokens, ToMe is different in that it *merges* them. And if we have information about what tokens we merged, we have enough information to then *unmerge* those same tokens. This is crucial for a dense prediction task, where we really do need every token.

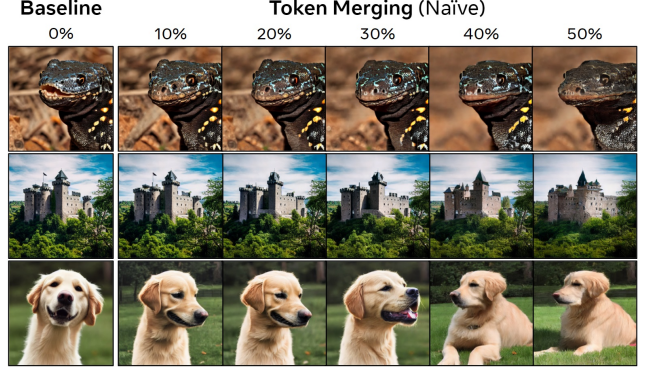


Figure 3. **ToMe applied naively** as described in Sec. 3 works to maintain image coherence, but the content inside the image can change drastically with high amounts of reduction.



Figure 4. **Importance of Merging.** If we just prune tokens instead (and replace them with 0), the resulting images quickly degrade.

In this work, we’ll define unmerging in the simplest possible way. Given two tokens with c channels $x_1, x_2 \in \mathbb{R}^c$ s.t. $x_1 \approx x_2$, if we merge them into a single token $x_{1,2}^*$, e.g.,

$$x_{1,2}^* = (x_1 + x_2)/2 \quad (1)$$

we can “unmerge” them back into x'_1 and x'_2 by setting

$$x'_1 = x_{1,2}^* \quad x'_2 = x_{1,2}^* \quad (2)$$

While this loses information, the tokens were already similar so the error is small. We find this works well in our case, but other unmerging methods might be worth exploring.

3.2. An Initial Naïve Approach

Merging tokens and then immediately unmerging them doesn’t help us though. Instead, we’d like to merge tokens, do some (now reduced) computation, and then unmerge them afterward so we don’t lose any tokens. Naïvely, we can just apply ToMe *before* each component of each block (i.e., self attn, cross attn, mlp), and then unmerge the outputs before adding the skip connection (see Fig. 2).

Details. Because we’re not accumulating any token reduction (merged tokens are quickly unmerged), we have to merge a lot more than the original ToMe. Thus instead of removing a *quantity* of tokens r , we remove a percentage ($r\%$) of all tokens. Moreover, computing token similarities for merging is expensive, so we only do it once at the start of each block. Finally, we don’t use proportional attention and use the input to the block x for similarly rather than attention keys k . More exploration is necessary to find if these techniques carry over from the classification setting.

Method	r%	FID ↓	s/im ↓	GB/im ↓
Baseline	0	33.12	3.09	3.41
ToMe (Naïve)	10	33.14	2.60	2.99
	20	33.53	2.29	2.17
	30	33.60	2.11	1.71
	40	34.67	1.81	1.26
	50	38.95	1.53	0.89

Table 1. **Quantitative evaluation** of the results in Fig. 3. While the approach in Sec. 3 can lead to $\sim 2\times$ faster image generation with $\sim 4\times$ less memory used (here for 512×512 images), it results in a significantly higher FID score. Thus, we explore further.

4. Further Exploration

Amazingly, the simple approach described in Sec. 3 works fairly well out of the box *without any training*, even for large amounts of token reduction (see Fig. 3). This is in stark contrast to if we *pruned* tokens instead, which completely destroys the image (see Fig. 4). However, we’re not done yet. While the images with ToMe applied look alright, the content within each image changes drastically (mostly for the worse). Thus, we make further improvements using Naïve ToMe with 50% reduction as our starting point.

Experimental Details. To quantify performance, we use Stable diffusion v1.5 to generate 2,000 512×512 images of ImageNet-1k [3] classes (2 per class) using 50 PLMS [9] diffusion steps with a cfg scale [4] of 7.5. We then compute FID [6] scores between those 2,000 samples and 5,000 class-balanced ImageNet-1k val examples using [19]. To test speed, we simply average the time taken over all 2,000 samples on a single 4090 GPU. Applying ToMe naïvely increases FID substantially (see Tab. 1), though evaluation is up to $2\times$ faster with up to $4\times$ less memory used.

4.1. A New Partitioning Method

By default, ToMe partitions the tokens into **src** and **dst** (see Sec. 2) by alternating between the two. This works for ViTs without unmerging, but in our case this causes **src** and **dst** to form alternating columns (see Fig. 5a). Since half of all tokens are in **src**, if we merge 50% of all tokens, then the entirety of **src** gets merged into **dst**, so we effectively halve the resolution of the images along the rows.

A simple fix would be to select tokens for **dst** with some 2d stride. This significantly improves the image both qualitatively (Fig. 5b) and quantitatively (Tab. 2a) and gives us the ability to merge more tokens if we want (i.e., the **src** set is larger), but the **dst** tokens are still always in the same place. To resolve this, we can introduce randomness.

However, if we just sample **dst** randomly, the FID jumps massively (Tab. 2b w/o fix). Crucially, we find that when using classifier-free guidance [4], the prompted and unprompted samples *need to allocate dst tokens in the same way*. We resolve this by fixing the randomness across the batch, which improves results past using a 2d stride (Fig. 5c,

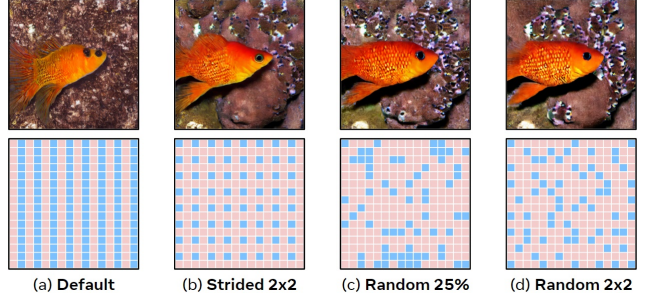


Figure 5. **Partitioning src and dst.** ToMe [1] merges tokens from **src** into **dst**. (a) By default, ToMe alternates **src** and **dst** tokens. In our case, this causes **dst** to form regular columns which leads to bad outputs (poor fish). (b) We can improve generation by sampling **dst** with a 2d stride (e.g., 2×2), but this still forms a regular grid. (c) We can introduce irregularity by sampling randomly, but this can cause undesirable clumps of **dst** tokens. (d) Thus, we sample one **dst** token randomly in each 2×2 region.

$s_y \times s_x$	dst%	FID ↓
1×2	50	38.95
2×1	50	39.28
2×2	25	36.12
2×4	12.5	37.09
4×2	12.5	37.14
4×4	6.25	38.97

(a) **Strided** w/ diff strides.

		w/o Fix	w/ Fix
method	fix	FID ↓	
rand 25%	✗	46.08	
rand 25%	✓	36.00	
rand 2×2	✓	35.66	

(b) **Random** w/ diff methods.

Table 2. **Partition Experiments.** Evaluation of the **src** and **dst** partitioning methods described in Fig. 5. 50% of tokens are merged in all experiments. Random methods (b) perform the best provided we fix the randomness across the batch (see Sec. 4.1).

Tab. 2b w/ fix). Combining the two methods by randomly choosing one **dst** token in each 2×2 region performs even better (Fig. 5d), so we make this our default going forward.

4.2. Design Experiments

In Sec. 3, we apply ToMe to every module, layer, and diffusion step. Here we search for a better design (Tab. 3).

What should we apply ToMe to? Originally, we applied ToMe to all modules (self attn, cross attn, mlp). In Tab. 3a, we test applying ToMe to different combinations of these modules and find that in terms of speed vs. FID trade-off, just applying ToMe to self attn is the clear winner. Note that FID doesn’t consider prompt adherence, which is likely why merging the cross attn module actually reduces FID.

Where should we apply ToMe? Applying ToMe to every block in the network is not ideal, since blocks at deeper U-Net scales have much fewer tokens. In Tab. 3b, we try restricting ToMe to only blocks with some minimum number of tokens and find that only the blocks with the most tokens

self attn	cross attn	mlp	FID ↓	s/im ↓	min tokens	blocks	FID ↓	s/im ↓	r% start	r% end	FID ↓	s/im ↓
✓	✓	✓	35.66	1.56	64	15 (all)	35.66	1.56	70	30	35.89	1.65
✓	✗	✓	36.10	1.57	256	14	35.71	1.55	60	40	35.53	1.58
✓	✗	✗	33.73	1.64	1024	9	34.37	1.56	50	50	35.66	1.56
✗	✗	✓	34.70	2.81	4096	4	33.29	1.63	40	60	36.09	1.58
									30	70	36.45	1.61

(a) **What should we apply ToMe to?** By only applying ToMe to self-attention modules, we can get most of the speed-up with much better FID.

(b) **Where should apply ToMe?** If we apply ToMe to only the layers with the most tokens, we can get great FID while still being fast.

(c) **When should we apply ToMe?** We can get a small boost by merging more tokens during early diffusion steps and a fewer during later steps.

Table 3. **Design Experiments.** Using the random 2×2 partitioning method from Tab. 2, we now explore how best to apply ToMe (with $r = 50\%$). Each experiment is independent, and we highlight our resulting design choice in gray.

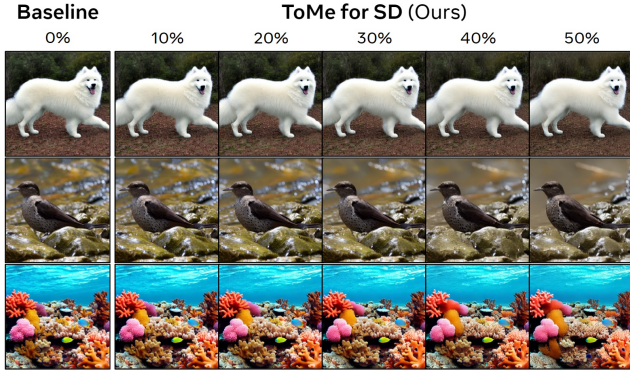


Figure 6. **Qualitative Results.** Our version of ToMe retains the image content, even at high token reduction. While some slight detail may be lost (e.g., in the background for the dog and bird), our method still handles complex scenes well (like the coral reef).

Method	r%	FID ↓	s/im ↓	GB/im ↓
Baseline	0	33.12	3.09	3.41
ToMe for SD	10	32.86	2.56	2.99
	20	32.86	2.29	2.17
	30	32.80	2.06	1.71
	40	32.87	1.85	1.26
	50	33.02	1.65	0.89
	60	33.37	1.52	0.60

Table 4. **Quantitative Results.** Using the improvements from Sec. 4, our ToMe for Stable Diffusion obtains similar or better FID compared to the baseline while still being up to $2\times$ faster and using up to $5.6\times$ less memory with 60% of tokens reduced.

need ToMe applied to get most of the speed-up.

When should we apply ToMe? It might not be right to reduce the same number of tokens in each diffusion step. Earlier diffusion steps are coarser and thus might be more forgiving to errors. In Tab. 3c, we test this by linearly interpolating the percent of tokens reduced and find that indeed merging more tokens earlier and fewer tokens later is slightly better, but not enough to be worth it.

5. Putting It All Together

We combine all the techniques discussed in Sec. 4 into one method, dubbed “ToMe for Stable Diffusion”. In Fig. 6

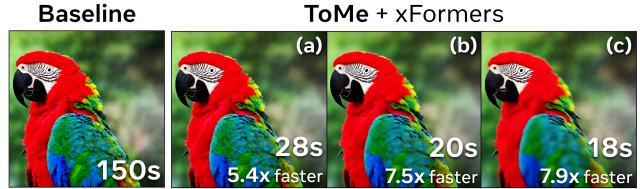


Figure 7. **Tuning for More Speed-up.** The result in Fig. 1 (a) uses the choices made in Sec. 4, which are tuned for the best quality. However, we can get even more speed-up if we are okay with some drop in fidelity (b), but there’s a limit (c) after which little time is taken by modules outside of our control, resulting in little further speed-up. (Hint: the bird’s plumage gets less detailed).

we show representative samples of how it performs visually, and in Tab. 4 we show the same qualitatively. Overall, ToMe for Stable Diffusion minimally impacts visual quality while offering up to $2\times$ faster evaluation using $5.6\times$ less memory.

ToMe + xFormers. Since ToMe just reduces the number of tokens, we can still use off the shelf fast transformer implementations to get even more benefit. In Fig. 1 we test generating a 2048×2048 image with ToMe and xFormers combined and find massive speed benefits. We can get even more speed-up if we’re okay with sacrificing more visual quality (Fig. 7). Note that with smaller images, we found this speed-up to be less pronounced, likely due to the diffusion model not being the bottleneck. Moreover, the memory benefits did not stack with xFormers.

6. Conclusion and Future Directions

Overall, we successfully apply ToMe to Stable Diffusion in a way that generates high quality images while being *significantly faster*. Notably, we do this *without training* which is rather remarkable, as any other token reduction method would require retraining. Still, these results could likely be improved by exploring 1.) better unmerging strategies or 2.) whether proportional attention or key-based similarity are useful for diffusion. Furthermore, our success motivates more exploration into using ToMe for dense prediction tasks. We hope this work can serve as both a useful tool for practitioners as well as a starting point for future research in token merging.

References

- [1] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *ICLR*, 2023. 1, 2, 3
- [2] Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *arXiv:2205.14135 [cs.LG]*, 2022. 1
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3
- [4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 1, 2, 3
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1
- [6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 3
- [7] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Bin Ren, Minghai Qin, Hao Tang, and Yanzhi Wang. Spvit: Enabling faster vision transformers via soft token pruning. In *ECCV*, 2022. 1
- [8] Benjamin Lefauveux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, and Daniel Haziza. xformers: A modular and hackable transformer modelling library. <https://github.com/facebookresearch/xformers>, 2022. 1
- [9] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv:2202.09778 [cs.CV]*, 2022. 3
- [10] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers. *arXiv:2110.03860 [cs.CV]*, 2021. 1
- [11] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. Adavit: Adaptive vision transformers for efficient image recognition. In *CVPR*, 2022. 1
- [12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 1
- [13] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125 [cs.CV]*, 2022. 1
- [14] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *NeurIPS*, 2021. 1, 2
- [15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2
- [16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2
- [17] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In *NeurIPS*, 2021. 1
- [18] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 1
- [19] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>, August 2020. Version 0.2.1. 3
- [20] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2
- [21] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 2019. 1, 2
- [22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 1, 2
- [23] Hongxu Yin, Arash Vahdat, Jose Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-ViT: Adaptive tokens for efficient vision transformer. In *CVPR*, 2022. 1