Importance-Based Token Merging for Efficient Image and Video Generation

Haoyu Wu[1] Jingyi Xu[1] Hieu Le[†2,3] Dimitris Samaras[†1]
[1]Stony Brook University [2]EPFL [3]UNC-Charlotte

Abstract

Token merging can effectively accelerate various vision systems by processing groups of similar tokens only once and sharing the results across them. However, existing token grouping methods are often ad hoc and random, disregarding the actual content of the samples. We show that preserving high-information tokens during merging—those essential for semantic fidelity and structural details—significantly improves sample quality, producing finer details and more coherent, realistic generations. To do so, we propose an importance-based token merging method that prioritizes the most critical tokens in computational resource allocation, leveraging readily available importance scores, such as those from classifier-free guidance in diffusion models. Experiments show that our approach significantly outperforms baseline methods across multiple applications, including text-to-image synthesis, multi-view image generation, and video generation with various model architectures such as Stable Diffusion, Zero123++, AnimateDiff, or PixArt-α.

1. Introduction

The rise of powerful diffusion models such as DALL-E [76], Stable Diffusion (SD) [79], or Imagen [80] has dramatically changed the landscape of generative AI [2, 17, 26, 79, 80, 90]. At their core, these models operate by iteratively denoising through multiple passes of a backbone network, processing a substantial number of tokens, which are particularly computationally intensive. To reduce the computational demands, prior work [3, 4, 36] has explored merging similar tokens and sharing the results across member tokens.

However, such a merging process can also significantly degrade image quality, losing some important details and structure. This typically happens when merging occurs in highly informative image regions, where critical visual details are compressed or discarded. One major reason is that existing methods rely on ad hoc heuristics for grouping tokens, often based on spatial proximity or fixed patterns.

[†]equal contribution

Prompt: "*A cute owl wearing a wizard hat.*"

(a) Spatial [3] (b) Importance-based (Ours)

Figure 1. **Importance-based Token Merging.** Our method prioritizes important tokens during token merging, resulting in images with greater details in essential areas compared to ToMeSD [3]. In the second row, we show the regions (in white) where computation (*e.g.*, attention) will take place after token merging.

Thus, the destination tokens, where the other tokens will merge into, are chosen randomly within predefined regions, sometimes pushing important visual elements into less relevant areas. This leads to suboptimal resource allocation, where crucial textures, edges, or fine-grained structures are lost, while regions that are less important retain more detail than necessary, as can be seen in Fig. 1 (a).

To address this limitation, we propose a novel token selection method that prioritizes computational resources in areas of high visual and semantic importance. Compared to existing approaches, our method ensures more effective merging decisions by using guidance from actual content relevance. Rather than treating all tokens equally, we use per-token importance signals to identify key or salient regions and ensure that the merging procedure prioritizes preserving more important tokens. Note that there are many proxies for importance, such as attention scores, saliency maps, or user-provided bounding boxes, all of which can be integrated into our merging method. Specifically in our case, we point out that an excellent choice is the classifier-free guidance (CFG) [25], as it inherently highlights regions that strongly influence model outputs and is obtainable with no additional cost.

More specifically, instead of merging tokens arbitrarily across the image, we first construct a pool of high-importance tokens. Then, token partitioning is only performed within this pool using a soft-matching strategy. This ensures that the most relevant tokens serve as anchors for merging, preserving key details while also maintaining diversity among the anchor tokens. Doing so facilitates a more efficient allocation of computational resources, preventing redundancy and ensuring that merging decisions are both meaningful and effective. Compared to simply selecting the top-k important tokens—which can lead to redundancy and suboptimal token assignments—our method strikes a balance between relevance and diversity, resulting in more coherent and detailed generations.

We apply our token merging strategy across three key applications: text-to-image generation, multi-view generation, and text-to-video generation, as well as across various model structures such as U-Net or transformers. Compared to baseline methods, our method consistently demonstrates superior performance, delivering results with higher fidelity and significantly improves image details across all tested scenarios.

Our contributions are as follows:

- We propose a novel importance-based token merging paradigm for diffusion models, designed to preserve crucial image content. The importance score can be easily obtained from CFG at no additional cost.
- We design a novel token-partitioning strategy based on a pool of important tokens to improve generation quality.
- Our method achieves state-of-the-art performance across various diffusion model tasks, including text-to-image synthesis, multi-view generation, and video generation, as well as across model architectures including U-Net and transformers.

2. Related Work

2.1. Diffusion Models

Diffusion models [12, 26, 88, 90] are a class of generative models that iteratively transform random noise into complex data structures, such as images, by gradually reversing a diffusion process. In these models, data is progressively corrupted by adding noise, and the model learns to recover the original data through iterative denoising with learned parameters. This approach has demonstrated impressive results in generating high-quality, realistic images [1, 6, 16, 33, 67, 69, 72, 76, 79, 80, 105], videos [2, 5, 17, 39, 84, 114], 3D content [42, 53, 58, 62, 70, 74, 83], and audio [27, 37]. A popular diffusion model architecture is introduced by Stable Diffusion [79], which uses a U-Net with transformer blocks. It first encodes a noisy image as latent tokens, which are processed through transformer blocks comprising self-attention, MLP, and cross-attention layers. This design has been extended for multi-view generation with multi-view attention [83] and for video generation with temporal layers [17]. More recent approaches [6, 114] replace the U-Net with a fully transformer-based architecture for improved scalability.

In diffusion models, classifier-free guidance (CFG) [25] is a technique that enhances fidelity and detail in generation by guiding the model toward conditioned inputs without the need for an external classifier. Recent work [98, 110] suggests a connection between CFG and saliency. Our study advances this by identifying CFG as an indicator of token importance, enabling more effective token merging.

To reduce the cost of diffusion inferences, various techniques have been proposed. Some approaches require retraining, including better model structures [35, 73, 79], model pruning [14], model compression [48, 106, 113], step distillation [18, 54, 68, 81, 82], and consistency regularization [63, 91]. Others avoid retraining, such as improved sampling to reduce inference steps [52, 59, 60, 90], caching [8, 31, 44, 64, 65, 86, 101, 112], model quantization [7, 11, 22, 46, 87, 95], and token reduction [3, 4, 32]. In this work, we focus on improving token reduction. Notably, our method is compatible with other diffusion acceleration techniques, enabling combined use for further speedup.

2.2. Token Reduction

Token reduction [4, 20, 93] decreases the number of tokens to process by pruning or merging them. It is widely applied in tasks such as classification [4, 20, 50, 66], segmentation [34, 103], detection [56], video understanding [10], and large language models [30, 45, 89]. Token pruning methods include removing tokens based on attention scores [15, 20, 50, 57, 100, 102, 104], using the Gumbel-Softmax trick for selective pruning [29, 38, 56, 77, 100], developing sampling methods [15, 107], integrating sparsity in the model [9, 43], and reinforcement learning-based methods [71]. For token merging [4, 19, 40, 78, 92, 97, 103, 115], approaches include bipartite soft matching [3, 4], K-Means [66], K-Medoids clustering [66], and Density-Peak Clustering with K-Nearest Neighbors (DPC-KNN) [108]. Additionally, soft merging methods assign tokens to multiple clusters before merging them [19, 78, 115].

Recent studies have applied token reduction to diffusion model inference [3, 21, 32, 36, 47, 49, 61, 85, 94, 96, 111]. ToMeSD [3] introduced token merging to Stable Diffusion [79] with a training-free method. The typical token merging procedure for diffusion models selects destination tokens from input feature tokens and utilizes bipartite soft matching to merge redundant tokens. The reduced token set is processed by operations like attention, and is then copied back to the merged token locations, ensuring the final token count matches the input. Although this token merging approach performs well, its selection of destination tokens is not optimal, *i.e.*, spatially random across the image. ToFu [36] combines token pruning and merging but still follows similarly suboptimal destination token selection strategy. In our work, we propose that selecting destination tokens based on their importance improves the fidelity and quality in generation, and this importance can be easily obtained via classifier-free guidance without additional cost. AT-EDM [94] uses self-attention maps to derive token importance, but this requires first performing full self-attention and increases peak memory usage. ATC [21] applies bottom-up hierarchical clustering for better token merging, but we show it is costly for generative tasks.

3. Method

We propose a novel importance-based token-merging method, summarized in Fig. 2, allowing for a more efficient allocation of computing resources. We highlight that classifier-free guidance serves as an effective importance indicator. Further, we propose using a dynamic pool of important tokens, where the pool size adapts to the token merging ratio, optimizing resource allocation and reducing redundancy.

Important Tokens. Selecting which tokens to serve as anchors (destination tokens) for merging is critical for generating high-quality content. This is because these tokens correspond to the primary image regions where subsequent computations, such as attention, are applied. In principle, all reliable per token importance signals, such as cross attention maps, user provided bounding boxes, or saliency maps, can be integrated with our method. We find that classifier-free guidance (CFG) [25] serves as an excellent indicator. CFG modifies the noise prediction to improve sample control. It is designed to steer the predicted noise in a direction more aligned with the condition:

$$\tilde{\epsilon}_\theta(\mathbf{x}_t, t) = \epsilon_\theta(\mathbf{x}_t, t) + w \cdot (\epsilon_\theta(\mathbf{x}_t \mid y, t) - \epsilon_\theta(\mathbf{x}_t, t)), \quad (1)$$

where $\epsilon_\theta(\mathbf{x}_t \mid y, t)$ is the noise prediction conditioned on input y, $\epsilon_\theta(\mathbf{x}_t, t)$ is the unconditional noise prediction, and w is the guidance weight. It is widely used in diffusion models and incurs no additional computational cost. For each token, we calculate its importance as the absolute value

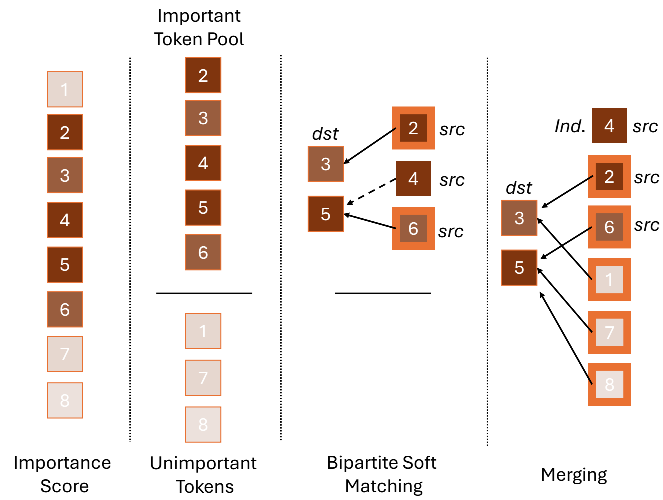


Figure 2. **Overview.** We propose an importance-based token merging method. The importance of each token can be determined using classifier-free guidance. These scores, visualized with colors ranging from light to dark (indicating less to more important tokens), are used to construct a pool of important tokens. We randomly select a set of destination (*dst*) tokens from this pool and the remaining important tokens become source (*src*) tokens. Bipartite soft matching is then performed between the *dst* tokens and *src* tokens. *src* tokens without a suitable match are considered independent tokens (*ind.*). All other *src* tokens and unimportant tokens are merged with the destination tokens for subsequent computational steps.

of its CFG score:

$$\begin{aligned} \text{importance} &= |\epsilon_\theta(\mathbf{x}_t \mid y, t) - \epsilon_\theta(\mathbf{x}_t, t)| \\ &\approx |-\sigma_t \nabla_{\mathbf{x}_t} \log p(y \mid \mathbf{x}_t)|, \end{aligned} \quad (2)$$

where σ_t is the noise scale. The guidance term effectively estimates the gradient of the log-likelihood of the conditioning variable y (e.g., a text prompt) with respect to the noisy sample $\mathbf{x}_t$ [25]. Thus, the CFG magnitude can be interpreted as a saliency or importance measure: tokens with a high CFG magnitude have stronger influences in steering the generation toward satisfying the condition y. In Fig. 3, we provide an example of the resulting importance maps.

Importance-based Token Merging. With the token importance scores, a naive approach is to pick the top-k tokens as destination tokens (*dst*) and merge the rest tokens that are similar to them. However, this approach produces low-quality outputs due to merging inefficiency, as shown in Fig. 4 (a). More specifically, there are two particular issues about this:

1. **Redundancy**. The top-k tokens can be very similar - but all important tokens, leading to redundancy and less intra-variant among the selected destination tokens.
2. **Unimportant Independent Tokens**. In the token merging pipeline, some tokens lack similar destination to-

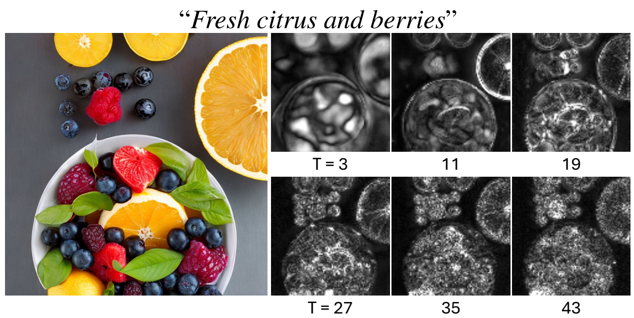


Figure 3. **Importance Maps.** We present token importance maps derived from classifier-free guidance (CFG) across diffusion inference timesteps. These maps highlight areas significantly align with the user prompt. In the early steps, they capture the semantics and structure of the image relevant to the prompt, while in later steps, they focus on finer details of the objects the user intends to generate. The generated image is shown on the left for reference.

(a) Top-k (b) Ours

Figure 4. We compare our method with an approach that uses the top-k important tokens as destination tokens (*dst*) for token merging. The computation locations after token merging are illustrated as non-black pixels in the bottom-right windows. They include locations of *dst* tokens, which are shown in white, and independent tokens (some other tokens that lack a similar *dst* token for merging), which are shown in orange. Our method produces more structured and detailed image, as highlighted in the red box.

kens, making them "independent" and remain unmerged. In the top-k approach, background or irrelevant tokens often become independent due to the absence of suitable matches among the important tokens, as shown in Fig. 4.

To avoid these issues, we propose to first create a pool of the most important tokens and then, ensure that both destination tokens and independent tokens are drawn from this set. To do so, destination tokens are randomly sampled from the pool, while independent tokens are selected as those in the important token pool that are most dissimilar to the chosen destination tokens. This approach ensures that all computations following the merging step operate on important tokens, leading to improved detail preservation, as illustrated in Fig. 4 (b).

To determine the optimal pool size, we adapt it based on the token merging ratio r. Specifically, we set the pool size as $\mathbf{P} = (1-r)\cdot(1+p)$, where p is a hyper-parameter of our method. With a constrained token processing budget, *i.e.*, a high token merging ratio r, the pool size remains small, ensuring the selection of only the most critical tokens. Conversely, with a larger compute budget, the pool size increases, reducing the likelihood of selecting redundant tokens as destination tokens.

Token Merging in Diffusion Inference. At time-step t of the diffusion inference, a diffusion model layer, *e.g.* a transformer layer, takes $\mathbf{N}$ tokens as input. The token merging ratio is r. Based on the token importance derived from the previous timestep's classifier-free guidance, we select the top $\mathbf{K} = \mathbf{N}\cdot(1-r)\cdot(1+p)$ tokens as the important token pool, denoted as $\mathbf{A}$. From $\mathbf{A}$, we randomly pick $\mathbf{D} = \mathbf{N}\cdot k$ tokens to form the destination (*dst*) set. Here, p, k are hyper-parameters. The remaining tokens in $\mathbf{A}$ become the source (*src*) set.

Next, we perform bipartite soft matching by computing pairwise cosine similarities between *src* and *dst* tokens. Each *src* token is linked to its most similar *dst*. Then, in *src* set, we select the top $\mathbf{I} = \mathbf{N}\cdot(1-k-r)$ tokens with the smallest similarity to their closest *dst* tokens to serve as independent tokens. The remaining *src* tokens and unimportant tokens are merged into their corresponding *dst* tokens. The merging is performed via averaging all grouped tokens. After merging, the number of tokens reduces to $\mathbf{I}+\mathbf{D} = \mathbf{N}\cdot(1-r)$ tokens. The diffusion model layer then processes this reduced token set, the merged locations are filled with the corresponding processed *dst* tokens, maintaining the same output shape as the input.

4. Experiments

4.1. Experimental Settings

Text-to-image Generation. We use Stable Diffusion 2 (SD) [79] as the base model, a text-to-image latent diffusion model with a U-Net architecture that generates 768x768 images. Our token merging method is compared to ToFu [36], ToMeSD [3] and ATC [21]. Due to ATC's slow inference (several minutes per image), we only display its visual results. For quantitative comparison, we follow previous studies [1, 76, 80, 105] and report FID [24] and CLIP scores [23, 75] for zero-shot image generation on the MS-COCO 2014 validation dataset [51], with 30K randomly sampled image-caption pairs.

We also experiment with a diffusion transformer, PixArt-α [6], for text-to-image generation. We compare our method with ToMeSD and similarly evaluate on the MS-COCO dataset. Unless otherwise stated, "text-to-image" in this paper refers to generation based on Stable Diffusion.

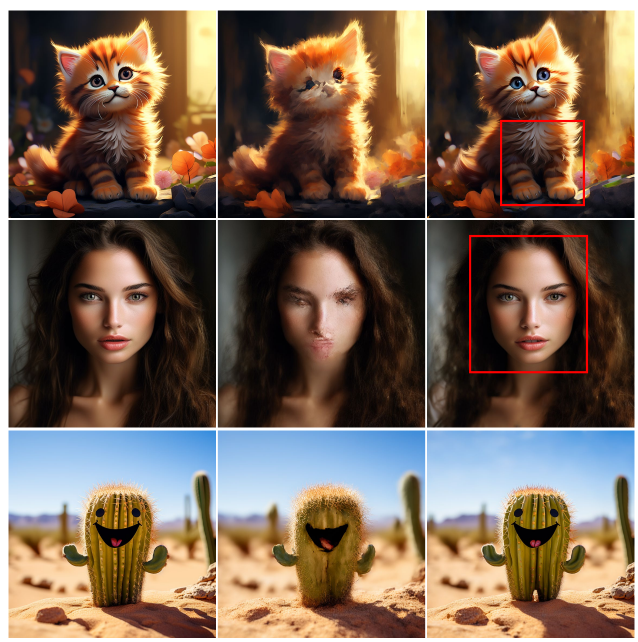
(a) SD [79] $\sim$ 8s (b) ATC [21] $\sim$ mins (c) ToFu [36] $\sim$ 5s (d) ToMe. [3] $\sim$ 5s (e) Ours $\sim$ 5s

Figure 5. **Qualitative comparison of text-to-image generation.** The first column shows results from Stable Diffusion (SD) [79], while the subsequent columns show SD combined with various token merging methods. As highlighted in red boxes, our approach consistently produces finer details with coherent structures. Note that ATC requires minutes to generate an image, whereas other methods, including ours, complete the task in seconds. The token merging ratio is 0.7. Please see the supplementary for prompts. Best viewed with zoom-in.

Multi-view Diffusion. We use Zero123++ v1.2 [83] as the base model. Zero123++ is an image-conditioned multi-view latent diffusion model that generates six novel views at a resolution of 320×320. We compare our method to ToMeSD. Following evaluation protocols of prior work [55, 58], we test on GSO dataset [13], which comprises 30 everyday objects, and compute PSNR, SSIM [99], and LPIPS metrics [109] to evaluate the similarity between the generated images and ground truth. We also include visual comparisons using in-the-wild images as input.

Video Diffusion. We adopt AnimateDiff v3 [17] as the base model, which adds temporal attention layers to turn the text-to-image model, *i.e.* Stable Diffusion, into a video diffusion model. This model generates 16-frame videos at a resolution of 512×512. We compare our token merging method with ToMeSD and evaluate using VBench [28]. We report the semantic, quality, and total scores from VBench. The semantic score assesses alignment between the generated videos and the user prompt, focusing on entity types, attributes, and styles. The quality score evaluates the temporal consistency and visual quality of the generated videos.

Other Metrics. We report TFLOPs, latency, and GPU memory usage, with and without memory-efficient atten-



(a) PixArt-α [6] (b) ToMeSD [3] (c) Ours

Figure 6. **Token merging for diffusion transformer.** We apply ToMeSD [3] and our method to PixArt-α [6] with a merging ratio of 0.3. Detailed generations are highlighted with red boxes. Best viewed with zoom-in. Please see the supplementary for prompts.

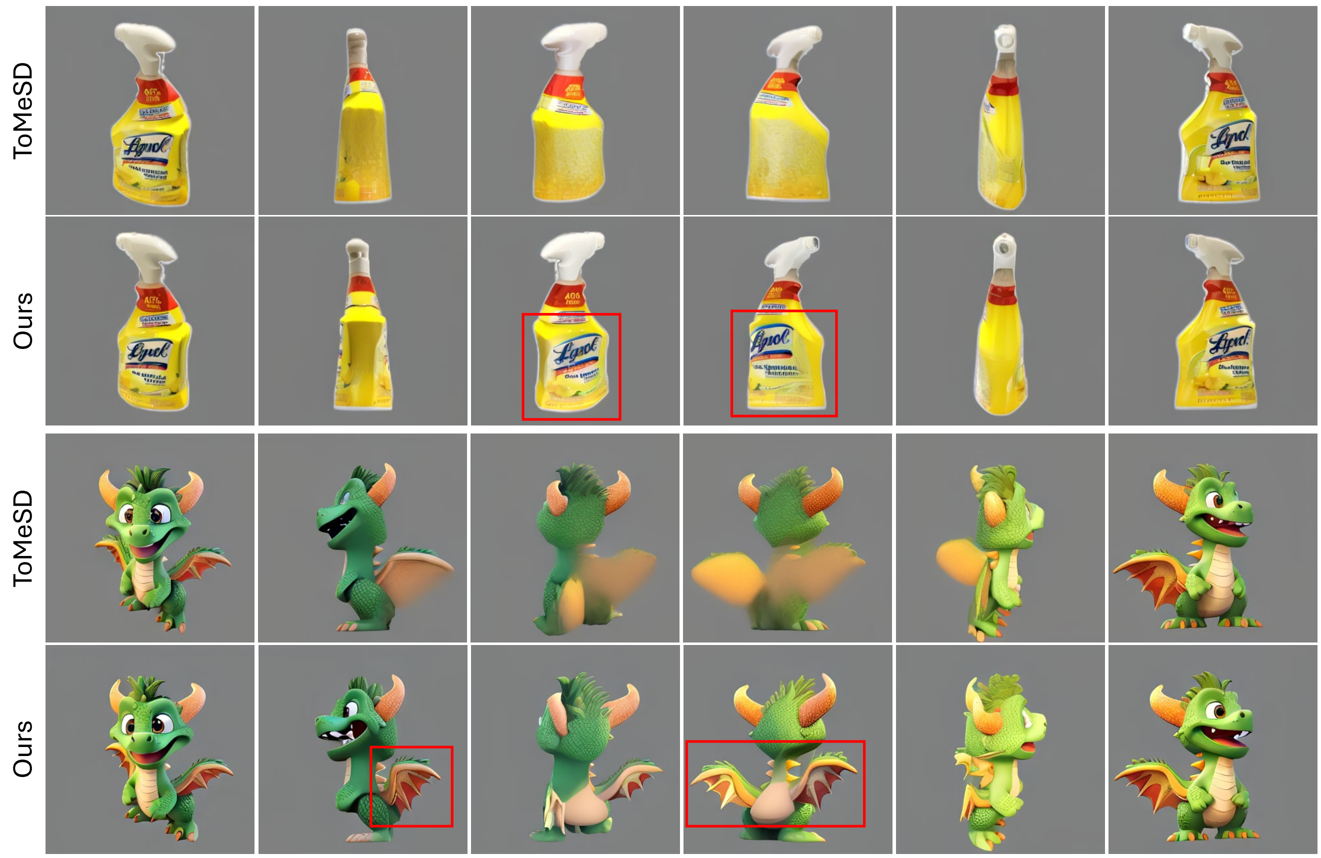


Figure 7. **Qualitative comparison of multi-view diffusion.** We apply ToMeSD [3] and our token merging method to the multi-view diffusion model. We use Zero123++ [83] as the base model and a merging ratio of 0.6. Our method outputs finer details, as highlighted in red boxes. Best viewed with zoom-in. Please refer to the supplementary for input images.

tion [41], on an NVIDIA A5000 GPU. Inference uses float16 precision. We estimate TFLOPs for a single diffusion sampling step. For latent diffusion models, the inference cost is measured exclusively in the latent space.

Implementation Details. For text-to-image synthesis with Stable Diffusion [79], we merge tokens in the self-attention layers of the first and last model blocks, similar to ToMeSD [3]. For PixArt-α [6], we apply token merging to self-attention and cross-attention in the middle half of the model layers (7–20 out of 28) for simplicity. For multi-view diffusion, we merge tokens in the self-attention layers of the first two and last two blocks of Zero123++ [83]. For video diffusion, we merge tokens in the first and last blocks of AnimateDiff [17]. The hyper-parameter p, which determines the size of our important token pool, is set to 0.4, 0.6, and 0.8 for image, multi-view, and video generation tasks, respectively. The number of destination tokens (k) remains consistent across all tasks and follows the setting used in ToMeSD, utilizing 25% of the total tokens.

4.2. Results

In Tab. 1 and Fig. 5, we compare different token merging methods applied to Stable Diffusion 2 [79] for text-to-image generation. Our method consistently outperforms baselines, especially at higher token merging ratios (r). For example, at $r = 0.75$, our method achieves an FID of 17.75 versus ToMeSD [3]'s 20.89. Agglomerative Token Clustering (ATC) [21] is not included due to its prohibitive computational cost, as it is CPU-bound and non-batched, making it impractical for large-scale evaluations. Notably, our method also performs well when used with cross-attention maps (Sec. 4.3). When the merging ratio is small, such as 0.3, our important token pool becomes the whole token set, making it functionally equivalent to ToMeSD.

In Tab. 2 and Fig. 6, we show our method significantly outperforms ToMeSD when applied to a diffusion transformer, achieving an FID improvement of 17–48%. This highlights the generalizability of our approach. In Tab. 3, we compare ToMeSD and our token merging method in the context of multi-view diffusion. Our importance-based token merging method consistently shows improved perfor-

r	FID ↓			CLIP ↑		
	ToFu	ToMe.	Ours	ToFu	ToMe.	Ours
0	-	-	11.88	-	-	31.83
0.30	13.48	12.20	12.20	31.82	31.82	31.82
0.50	15.81	13.50	**13.42**	31.81	31.79	**31.83**
0.60	17.32	14.81	**14.51**	31.81	31.80	**31.81**
0.70	18.94	17.46	**16.22**	31.78	31.78	**31.79**
0.75	19.29	20.89	**17.75**	31.76	31.71	**31.76**

Table 1. **Text-to-image generation.** We apply ToFu [36], ToMeSD [3] and our token merging method to Stable Diffusion [79] across various token merging ratios r.

r	FID ↓		CLIP ↑		Latency (s) ↓
	ToMe.	Ours	ToMe.	Ours	
0	27.51		31.30		9.14
0.3	34.23	**28.50**	30.76	**31.05**	8.96
0.5	65.46	**34.02**	29.85	**30.68**	7.54
0.7	95.46	**50.76**	28.67	**29.93**	7.14

Table 2. Comparison of our method with ToMeSD [3] when applied to the diffusion transformer PixArt-α [6].

r	PSNR ↑		SSIM ↑		LPIPS ↓	
	ToMe.	Ours	ToMe.	Ours	ToMe.	Ours
0.40	14.80	**14.82**	0.775	**0.777**	0.260	**0.259**
0.60	14.71	**14.85**	0.782	**0.783**	0.272	**0.263**
0.70	14.18	**14.80**	**0.787**	0.785	0.302	**0.274**
0.75	13.12	**14.58**	**0.789**	0.784	0.349	**0.283**

Table 3. **Multi-view diffusion.** We compare ToMeSD [3] with our token merging method when applied to Zero123++ [83].

mance over ToMeSD, especially at higher merging ratios. At $r = 0.75$, our method achieves a significant improvement in PSNR (14.58 vs. 13.12) and a much lower LPIPS (0.283 vs. 0.349), highlighting its ability to maintain high output quality even under aggressive token compression. Qualitative examples in Fig. 7 further validate this, showcasing finer geometrical and textual details in the objects generated by our method. A similar trend can be observed in Tab. 4 and Fig. 8, where we extend the comparison to video diffusion. Across these tests, our method consistently performs better than ToMeSD, both in numerical metrics and in visual quality, particularly in preserving object details. We observed that merging tokens in the temporal layers significantly reduces the generated dynamics. Nonetheless, we present results with token merging applied for both spatial and temporal layers in Tab. 5.

r	Semantic ↑		Quality ↑		Total ↑	
	ToMe.	Ours	ToMe.	Ours	ToMe.	Ours
0.40	75.40	75.40	81.69	81.69	80.44	80.44
0.60	74.03	**74.51**	81.58	**81.75**	80.07	**80.30**
0.70	72.03	**73.23**	**81.52**	81.23	79.62	**79.63**
0.75	69.67	**71.58**	80.82	**81.00**	78.59	**79.12**

Table 4. **Video diffusion (spatial).** Token merging applies on only spatial attention layers of AnimateDiff [17] with various token merging ratios r. We compare our method and ToMeSD [3] with VBench scores [28].

	Semantic ↑	Quality ↑	Total ↑
ToMeSD	72.71	81.35	79.62
Ours	**73.52**	**81.69**	**80.06**

Table 5. **Video diffusion (spatial and temporal)**. Token merging applies on both spatial and temporal attention layers of AnimateDiff [17] with a spatial merging ratio of 0.7 and a temporal merging ratio of 0.2. We report VBench scores [28] of our method in comparison to ToMeSD [3].

r	Latency (s)		Memory (GB)	TFLOPs
	ToMeSD	Ours		
0	8.5 (5.3)		7.63 (3.16)	4.30
0.3	8.0 (5.0)	8.0 (5.0)	5.20 (3.16)	3.83
0.5	6.0 (4.7)	6.1 (4.8)	4.05 (3.16)	3.55
0.7	5.7 (4.5)	5.8 (4.5)	3.55 (3.16)	3.36

Table 6. Comparison of inference costs when applying our token merging method and ToMeSD [3] to Stable Diffusion 2 [79]. Numbers in parentheses indicate measurements when memory-efficient attention [41] is enabled.

Inference Costs. We compare the inference costs of our method and ToMeSD when applied to Stable Diffusion 2 in Tab. 6. As can be seen, both methods show similar improvements in inference times, GPU usage, and TFLOPs. This demonstrates that our token merging strategy can enhance performance without incurring additional costs.

4.3. Ablation Study

We conduct ablation studies on the text-to-image generation task to examine alternative design choices in our method. As shown in Tab. 7, simply using top-k important tokens as destination tokens leads to worse results. Furthermore, not choosing independent tokens exclusively from the important set (w/ global *ind.*), leads to a performance drop.

Cross-Attention Maps for Token Importance. In principle, our method can be used with any per-token importance scores. As shown in Tab. 8, we use our method with

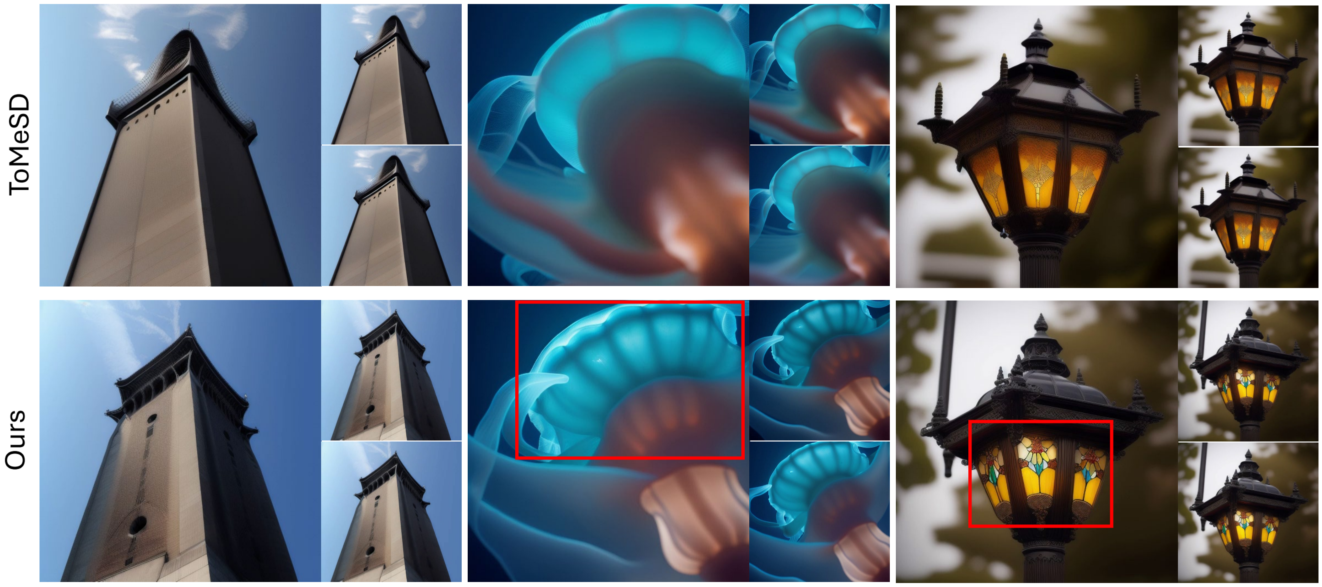


Figure 8. **Qualitative comparison of video diffusion.** We apply ToMeSD and our token merging method to the video diffusion model. For each generated video, we show three frames: the first on the left, the 8th at the top right, and the last 16th frame at the bottom right. We use AnimateDiff [17] as the base model and a merging ratio of 0.7. Best viewed with zoom-in. Please refer to supplementary for prompts.

r		Ours	w/ top-k *dst*	w/ global *ind.*
0.3	FID ↓	**12.20**	12.56	12.22
	CLIP ↑	31.82	31.82	31.82
0.7	FID ↓	**16.22**	16.29	16.43
	CLIP ↑	31.79	31.79	31.80

Table 7. Ablation studies of our method for text-to-image generation. 'w/ top-k *dst*' means top-k rather than random selection from the important token pool for destination tokens. 'w/ global *ind.*' means independent tokens may also be outside the important token pool, instead of solely within it.

r	p	0	0.2	0.4	0.8
0.3	FID ↓	12.87	12.32	**12.19**	**12.19**
	CLIP ↑	31.83	31.83	31.82	31.82
0.7	FID ↓	16.52	16.23	**16.22**	16.42
	CLIP ↑	31.78	31.78	31.79	31.79

Table 9. **Choice of p.** We show the results of our method for the text-to-image generation task with different values of p, which determines the important token pool size.

r	FID ↓			CLIP ↑		
	ToMe	Ours *(CA)*	Ours *(CFG)*	ToMe	Ours *(CA)*	Ours *(CFG)*
0.30	12.20	**12.17**	12.20	31.82	31.82	31.82
0.50	13.50	**13.38**	13.42	31.79	31.82	**31.83**
0.70	17.46	16.79	**16.22**	31.78	31.79	**31.79**
0.75	20.89	18.17	**17.75**	31.71	31.75	**31.76**

Table 8. Comparison of token merging methods applied to Stable Diffusion [79]. CA and CFG denote cross-attention and classifier-free guidance as importance signals, respectively.

cross-attention maps instead of CFG. While this may require additional memory to compute and store the attention matrices, it remains a strong alternative and also allows our method to remain applicable in scenarios where CFG is unavailable.

Choice of p. Our important token pool size is $(1-r)\cdot(1+p)$, where r is the token merging ratio. Tab. 9 shows that our method remains robust to the choice of p, provided that p is within a reasonable range.

5. Conclusions

We propose an importance-based token merging method for generation tasks, which maintains generation quality while reducing inference latency. We utilize token importance to dynamically allocate computational resources to regions of high relevance to the input condition, thereby enhancing the fidelity of the generated outputs. This novel, simple, and intuitive strategy accelerates various models for free with no modifications needed. Notably, we identify classifier-free guidance as an effective token importance indicator. Our method achieves state-of-the-art performance across diverse tasks, including text-to-image synthesis, multi-view generation, and video generation.

6. Acknowledgments

This work was supported in part by the NASA Biodiversity Program (Award 80NSSC21K1027), and the National Science Foundation (NSF) Grant IIS-2212046.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 4

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 2

[3] Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4599–4603, 2023. 1, 2, 4, 5, 6, 7

[4] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *ArXiv*, abs/2210.09461, 2022. 1, 2

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 3, 2024. 2

[6] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 2, 4, 5, 6, 7

[7] Lei Chen, Yuan Meng, Chen Tang, Xinzhu Ma, Jingyan Jiang, Xin Wang, Zhi Wang, and Wenwu Zhu. Q-dit: Accurate post-training quantization for diffusion transformers. *arXiv preprint arXiv:2406.17343*, 2024. 2

[8] Pengtao Chen, Mingzhu Shen, Peng Ye, Jianjian Cao, Chongjun Tu, Christos-Savvas Bouganis, Yiren Zhao, and Tao Chen. Delta-dit: A training-free acceleration method tailored for diffusion transformers. *arXiv preprint arXiv:2406.01125*, 2024. 2

[9] Tianlong Chen, Yu Cheng, Zhe Gan, Lu Yuan, Lei Zhang, and Zhangyang Wang. Chasing sparsity in vision transformers: An end-to-end exploration. *Advances in Neural Information Processing Systems*, 34:19974–19988, 2021. 2

[10] Joonmyung Choi, Sanghyeok Lee, Jaewon Chu, Minhyuk Choi, and Hyunwoo J Kim. vid-tldr: Training free token merging for light-weight video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18771–18781, 2024. 2

[11] Juncan Deng, Shuaiting Li, Zeyu Wang, Hong Gu, Kedong Xu, and Kejie Huang. Vq4dit: Efficient post-training vector quantization for diffusion transformers. *arXiv preprint arXiv:2408.17131*, 2024. 2

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[13] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 5

[14] Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. In *Advances in Neural Information Processing Systems*, 2023. 2

[15] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, pages 396–414. Springer, 2022. 2

[16] Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Learned representation-guided diffusion models for large-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8532–8542, 2024. 2

[17] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1, 2, 5, 6, 7, 8

[18] Amirhossein Habibian, Amir Ghodrati, Noor Fathima, Guillaume Sautiere, Risheek Garrepalli, Fatih Porikli, and Jens Petersen. Clockwork diffusion: Efficient generation with model-step distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8352–8361, 2024. 2

[19] Joakim Bruslund Haurum, Meysam Madadi, Sergio Escalera, and Thomas B Moeslund. Multi-scale hybrid vision transformer and sinkhorn tokenizer for sewer defect classification. *Automation in Construction*, 144:104614, 2022. 2

[20] Joakim Bruslund Haurum, Sergio Escalera, Graham W Taylor, and Thomas B Moeslund. Which tokens to use? investigating token reduction in vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 773–783, 2023. 2

[21] Joakim Bruslund Haurum, Sergio Escalera, Graham W Taylor, and Thomas B Moeslund. Agglomerative token clustering. In *European Conference on Computer Vision*, pages 200–218. Springer, 2025. 2, 3, 4, 5, 6

[22] Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Ptqd: Accurate post-training quantization for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[23] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free

evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 4

[24] 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. *Advances in neural information processing systems*, 30, 2017. 4

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[27] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023. 2

[28] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 5, 7

[29] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 2

[30] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 2

[31] Kumara Kahatapitiya, Haozhe Liu, Sen He, Ding Liu, Menglin Jia, Michael S Ryoo, and Tian Xie. Adaptive caching for faster video generation with diffusion transformers. *arXiv preprint arXiv:2411.02397*, 2024. 2

[32] Kumara Kahatapitiya, Adil Karjauv, Davide Abati, Fatih Porikli, Yuki M Asano, and Amirhossein Habibian. Object-centric diffusion for efficient video editing. In *European Conference on Computer Vision*, pages 91–108. Springer, 2025. 2

[33] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2025. 2

[34] Daniel Kienzle, Marco Kantonis, Robin Schön, and Rainer Lienhart. Segformer++: Efficient token-merging strategies for high-resolution semantic segmentation. *arXiv preprint arXiv:2405.14467*, 2024. 2

[35] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. On architectural compression of text-to-image diffusion models. 2023. 2

[36] Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1383–1392, 2024. 1, 2, 3, 4, 5, 7

[37] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 2

[38] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, et al. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *European conference on computer vision*, pages 620–640. Springer, 2022. 2

[39] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 2

[40] Dong Hoon Lee and Seunghoon Hong. Learning to merge tokens via decoupled embedding for efficient vision transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[41] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xformers: A modular and hackable transformer modelling library. `https://github.com/facebookresearch/xformers`, 2022. 6, 7

[42] Gang Li, Heliang Zheng, Chaoyue Wang, Chang Li, Changwen Zheng, and Dacheng Tao. 3ddesigner: Towards photorealistic 3d object generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2211.14108*, 2022. 2

[43] Ling Li, David Thorsley, and Joseph Hassoun. Sait: Sparse vision transformers through adaptive token pruning. *arXiv preprint arXiv:2210.05832*, 2022. 2

[44] Senmao Li, Taihang Hu, Fahad Shahbaz Khan, Linxuan Li, Shiqi Yang, Yaxing Wang, Ming-Ming Cheng, and Jian Yang. Faster diffusion: Rethinking the role of unet encoder in diffusion models. *CoRR*, 2023. 2

[45] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *arXiv preprint arXiv:2407.02392*, 2024. 2

[46] Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17535–17545, 2023. 2

[47] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7486–7495, 2024. 2

[48] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[49] Feng Liang, Akio Kodaira, Chenfeng Xu, Masayoshi Tomizuka, Kurt Keutzer, and Diana Marculescu. Looking backward: Streaming video-to-video translation with feature banks. *arXiv preprint arXiv:2405.15757*, 2024. 2

[50] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need:

Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022. 2

[51] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[52] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 2

[53] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[54] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[55] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 5

[56] Yifei Liu, Mathias Gehrig, Nico Messikommer, Marco Cannici, and Davide Scaramuzza. Revisiting token pruning for object detection and instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2658–2668, 2024. 2

[57] Sifan Long, Zhen Zhao, Jimin Pi, Shengsheng Wang, and Jingdong Wang. Beyond attentive tokens: Incorporating token importance and diversity for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2023. 2

[58] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 2, 5

[59] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787, 2022. 2

[60] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 2

[61] Wenbo Lu, Shaoyi Zheng, Yuxuan Xia, and Shengjie Wang. Toma: Token merging with attention for diffusion models. 2024. 2

[62] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2837–2845, 2021. 2

[63] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2

[64] Zhengyao Lv, Chenyang Si, Junhao Song, Zhenyu Yang, Yu Qiao, Ziwei Liu, and Kwan-Yee K Wong. Fastercache: Training-free video diffusion model acceleration with high quality. *arXiv preprint arXiv:2410.19355*, 2024. 2

[65] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15762–15772, 2024. 2

[66] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers for image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 12–21, 2023. 2

[67] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[68] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 2

[69] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[70] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 2

[71] Bowen Pan, Rameswar Panda, Rogerio Schmidt Feris, and Aude Jeanne Oliva. Interpretability-aware redundancy reduction for vision transformers, 2023. US Patent App. 17/559,053. 2

[72] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[73] Pablo Pernias, Dominic Rampas, Mats L Richter, Christopher J Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*, 2023. 2

[74] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[75] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[76] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image

generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 1, 2, 4

[77] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34: 13937–13949, 2021. 2

[78] Cedric Renggli, André Susano Pinto, Neil Houlsby, Basil Mustafa, Joan Puigcerver, and Carlos Riquelme. Learning to merge tokens in vision transformers. *arXiv preprint arXiv:2202.12015*, 2022. 2

[79] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[80] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Raphael Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2, 4

[81] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2

[82] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2025. 2

[83] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 2, 5, 6, 7

[84] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[85] Ethan Smith, Nayan Saxena, and Aninda Saha. Todo: Token downsampling for efficient generation of high-resolution images. *arXiv preprint arXiv:2402.13573*, 2024. 2

[86] Junhyuk So, Jungwon Lee, and Eunhyeok Park. Frdiff: Feature reuse for universal training-free acceleration of diffusion models. *arXiv preprint arXiv:2312.03517*, 2023. 2

[87] Junhyuk So, Jungwon Lee, Daehyun Ahn, Hyungjun Kim, and Eunhyeok Park. Temporal dynamic quantization for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[88] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[89] Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multimodal llms. *arXiv preprint arXiv:2409.10994*, 2024. 2

[90] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2

[91] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 2

[92] Hoai-Chau Tran, Duy MH Nguyen, Duy M Nguyen, Trung-Tin Nguyen, Ngan Le, Pengtao Xie, Daniel Sonntag, James Y Zou, Binh T Nguyen, and Mathias Niepert. Accelerating transformers with spectrum-preserving token merging. *arXiv preprint arXiv:2405.16148*, 2024. 2

[93] Hongjie Wang, Bhishma Dedhia, and Niraj K Jha. Zero-tprune: Zero-shot token pruning through leveraging of the attention graph in pre-trained transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16070–16079, 2024. 2

[94] Hongjie Wang, Difan Liu, Yan Kang, Yijun Li, Zhe Lin, Niraj K Jha, and Yuchen Liu. Attention-driven training-free efficiency enhancement of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16080–16089, 2024. 2, 3

[95] Haoxuan Wang, Yuzhang Shang, Zhihang Yuan, Junyi Wu, Junchi Yan, and Yan Yan. Quest: Low-bit diffusion model quantization via efficient selective finetuning. *arXiv preprint arXiv:2402.03666*, 2024. 2

[96] Kafeng Wang, Jianfei Chen, He Li, Zhenpeng Mi, and Jun Zhu. Sparsedm: Toward sparse efficient diffusion models. *arXiv preprint arXiv:2404.10445*, 2024. 2

[97] Yancheng Wang and Yingzhen Yang. Efficient visual transformer by learnable token merging. *arXiv preprint arXiv:2407.15219*, 2024. 2

[98] Yibin Wang, Weizhong Zhang, Jianwei Zheng, and Cheng Jin. High-fidelity person-centric subject-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7675–7684, 2024. 2

[99] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[100] Siyuan Wei, Tianzhu Ye, Shen Zhang, Yao Tang, and Jiajun Liang. Joint token pruning and squeezing towards more aggressive compression of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2092–2101, 2023. 2

[101] Felix Wimbauer, Bichen Wu, Edgar Schoenfeld, Xiaoliang Dai, Ji Hou, Zijian He, Artsiom Sanakoyeu, Peizhao Zhang, Sam Tsai, Jonas Kohler, et al. Cache me if you can: Accelerating diffusion models through block caching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6211–6220, 2024. 2

[102] Xinjian Wu, Fanhu Zeng, Xiudong Wang, and Xinghao Chen. Ppt: Token pruning and pooling for efficient vision transformers. *arXiv preprint arXiv:2310.01812*, 2023. 2

[103] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 2

[104] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2964–2972, 2022. 2

[105] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4

[106] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 22552–22562, 2023. 2

[107] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10809–10818, 2022. 2

[108] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022. 2

[109] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[110] Jing Zhao, Heliang Zheng, Chaoyue Wang, Long Lan, and Wenjing Yang. Magicfusion: Boosting text-to-image generation performance by fusing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22592–22602, 2023. 2

[111] Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. Dynamic diffusion transformer. *arXiv preprint arXiv:2410.03456*, 2024. 2

[112] Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast. *arXiv preprint arXiv:2408.12588*, 2024. 2

[113] Yang Zhao, Yanwu Xu, Zhisheng Xiao, and Tingbo Hou. Mobilediffusion: Subsecond text-to-image generation on mobile devices. *arXiv preprint arXiv:2311.16567*, 2023. 2

[114] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 2

[115] Zhuofan Zong, Kunchang Li, Guanglu Song, Yali Wang, Yu Qiao, Biao Leng, and Yu Liu. Self-slimmed vision transformer. In *European Conference on Computer Vision*, pages 432–448. Springer, 2022. 2