

# Importance-Based Token Merging for Efficient Image and Video Generation

## Supplementary Material

In Appendix A, we present more qualitative comparisons, empirical evidence on the relationship between CFG and token importance, results on consistency models, results from combining our method with orthogonal diffusion acceleration techniques, experiments with varying diffusion inference steps, explorations of a dynamic number of independent tokens or replacing bipartite matching with clustering, and findings from a user study. In Appendix B, we provide more details about our experimental settings. In Appendix C, we discuss the limitations of our method. In Appendix D, we provide the prompts used to generate qualitative results. We also include a supplementary video for comparisons on text-to-video generation.

### A. Additional Results

**Additional Qualitative Results.** In Fig. 1, we provide additional qualitative comparisons between ToFu [6], ToMeSD [1], our token merging method, and the variant of our method using cross-attention maps as importance signals. Additional visual comparisons for token merging applied to diffusion transformer are shown in Fig. 2. Additional results for multi-view diffusion are presented in Fig. 3. In the supplementary video, we include comparisons on text-to-video generation, using AnimateDiff [4] as the base diffusion model and a merging ratio of 0.7. Furthermore, we provide visual comparisons between ToMeSD [1] and our method across various merging ratios for text-to-image generation in Fig. 4.

**Token Importance via Classifier-Free Guidance.** The absolute value of classifier-free guidance (CFG) can be interpreted as a token-level saliency measure, highlighting the tokens that play a crucial role in steering the output toward the given prompt or condition. Empirically, as shown in Tab. 1, removing the top 30% of high-CFG tokens leads to a significant degradation in generation results, whereas removing the bottom 30% has little impact.

	No pruning	Top 30%	Bottom 30%
FID ↓	11.88	15.48	12.78
CLIP ↑	31.83	31.58	31.88

Table 1. Comparison of pruning (dropping) the top 30% of tokens with the highest CFG values versus the bottom 30% during text-to-image generation using Stable Diffusion [15].

**Results on the Consistency Model.** Consistency models [8, 9, 18] typically distill classifier-free guidance (CFG), making the explicit guidance term inaccessible, as they approximate the final guided noise function within a single forward pass. However, our method is not limited to CFG and can leverage any reliable per-token importance signal. As shown in Tab. 2 and Fig. 5, our token merging, using cross-attention maps as importance signals, remains effective and outperforms the baseline model when applied to the latent consistency model [9].

$r$	FID ↓		CLIP ↑		Time (s) ↓	Mem. (GB) ↓
	ToMe.	Ours	ToMe.	Ours		
0	25.38		31.05		0.65	6.75
0.30	25.63	25.63	31.03	31.03	0.60	4.22
0.50	27.80	<b>27.51</b>	30.98	<b>30.99</b>	0.51	3.56
0.60	30.05	<b>29.61</b>	30.90	<b>30.92</b>	0.50	3.37
0.70	35.01	<b>32.49</b>	30.59	<b>30.77</b>	0.48	3.21
0.75	43.28	<b>36.66</b>	30.38	<b>30.60</b>	0.46	3.12

Table 2. **Results on the Consistency Model.** We show the comparison of token merging methods applied to a 4-step latent consistency model (LCM\_Dreamshaper\_v7) [9] across various merging ratios  $r$ .

**Combination with Orthogonal Diffusion Acceleration Methods.** In Tab. 3, we demonstrate that our method can be combined with orthogonal diffusion acceleration methods [7, 10, 17] to further accelerate inference while preserving generation quality.

	FID ↓	CLIP ↑	Time (s) ↓	Mem. (GB) ↓
Ours	16.22	31.79	5.8	3.55
+ DeepCache [10]	15.46	31.81	2.6	3.56
+ FasterDiff. [7]	12.48	31.80	4.2	6.33
+ FRDiff [17]	15.25	31.83	3.5	3.59

Table 3. **Combination with Diffusion Acceleration Methods.** We show text-to-image generation results by integrating our method with orthogonal diffusion acceleration techniques, using Stable Diffusion [15] as the base model and a merging ratio of 0.7.

**Number of Diffusion Inference Steps.** In Tab. 4, we compare ToMeSD [1] and our method across different numbers of diffusion inference steps for text-to-image gener-

ation, demonstrating that our method consistently outperforms ToMeSD.

$T$	FID ↓		CLIP ↑	
	ToMeSD	Ours	ToMeSD	Ours
20	18.57	<b>17.03</b>	31.77	<b>31.80</b>
30	17.82	<b>16.51</b>	31.78	<b>31.80</b>
50	17.46	<b>16.22</b>	31.78	<b>31.79</b>

Table 4. **Number of Diffusion Inference Steps.** We compare ToMeSD [1] and our token merging method for the text-to-image generation task with different diffusion inference steps, using Stable Diffusion [15] as the base model and a merging ratio of 0.7.

**A Dynamic Number of Independent Tokens.** We experimented with dynamically allocating the number of independent tokens while keeping the total number of tokens after merging fixed. Specifically, we determine the number of independent tokens based on how many important tokens lack merging candidates that meet a similarity threshold. Interestingly, this dynamic allocation scheme does yield comparable or improved results compared to the fixed allocation, as shown in Tab. 5.

	Fixed	Dynamic (sim. threshold)		
		0.90	0.95	0.98
CLIP ↑	31.83	31.83	31.83	31.82
FID ↓	13.42	13.47	<b>13.34</b>	13.37

Table 5. Token merging results with a dynamic number of independent tokens for image generation. The merging ratio is 0.5.

**Replacing Bipartite Soft Matching with Clustering.** We explored replacing bipartite soft matching with Agglomerative Token Clustering (ATC) [5]. We use ATC to merge important tokens and also obtain cluster centers. We then merge unimportant tokens into their nearest cluster centers. Full ATC yields lower FID but significantly slows inference. When combined with our approach, which restricts clustering to important tokens, it achieves both improved performance and a substantial speedup (Tab. 6).

Method	CLIP ↑	FID ↓	Speed ↓
Ours	<b>31.79</b>	69.37	3.7 s
ATC	31.70	68.64	>10 mins
Ours + ATC	31.76	<b>68.38</b>	76.5 s

Table 6. Comparison of merging approaches for image generation, evaluated on 1K images with a token merging ratio of 0.7.

**User Study.** We conducted a randomized A/B user study to directly assess prompt alignment and visual quality in text-to-image generation. The questionnaire consists of 44 questions. In total, 66 users provided 2,904 responses. As shown in Tab. 7, our method was consistently preferred over the baselines.

Metric	Ours vs ToFu [6]			Ours vs ToMe. [1]		
	Ours	ToFu	Tie	Ours	ToMe	Tie
Alignment	<b>77.8</b>	10.1	12.1	<b>84.3</b>	6.7	8.9
Quality	<b>84.7</b>	10.6	4.7	<b>87.0</b>	8.4	4.5

Table 7. User study (in %) comparing prompt alignment and visual quality over baseline methods.

## B. Additional Experimental Settings

**Metrics.** We use the *deepspeed* [14] library to estimate TFLOPs, the *clean-fid* [12] library to calculate FID scores, and the *openai/clip-vit-base-patch16* model from OpenAI-CLIP [13] to calculate CLIP scores.

**Additional Implementation Details.** For text-to-image generation using Stable Diffusion [15], the diffusion process consists of 50 sampling steps, with the CFG scale set to 7.5. For PixArt- $\alpha$  [3], we perform diffusion sampling for 20 steps, with a CFG scale of 4.5. When computing token similarity for token merging in PixArt- $\alpha$ , we find that using pixel location distance of tokens yields better results than feature similarity, and we adopt this approach. For multi-view diffusion using Zero123++ [16], the process involves 50 sampling steps, with the CFG scale set to 4. For video diffusion using AnimateDiff [4], the sampling consists of 30 steps, with a CFG scale of 7.5. To ensure fairness, these settings are consistently applied to both our method and the baselines. For the ablation study investigating the use of cross-attention maps as importance signals, we utilize the averaged attention map from the final model block in Stable Diffusion.

**Details on Multi-view Diffusion.** The base multi-view diffusion model used in our experiments is Zero123++ [16], which fine-tunes Stable Diffusion 2 [15] to generate six novel views from an input image. During denoising, the model appends the self-attention key and value matrices from the reference input image to the attention layers for conditioning. The novel view poses are defined by a fixed set of absolute elevation and relative azimuth angles. Specifically, the elevation and azimuth angles (in degrees) are set as follows: (30, 30), (-20, 90), (30, 150), (-20, 210), (30, 270), (-20, 330). We consistently use this sequence of

novel views to present visual results in multi-view diffusion experiments.

**Preserving Structure in Early Time-steps.** Prior work [6] suggests that token pruning (directly dropping tokens) could help preserve structural details. Building on this observation, we found that incorporate token pruning during the early diffusion steps, followed by token merging in later steps, improves generation results. Specifically, we apply token pruning during the first 6, 10, and 4 diffusion inference steps for image, multi-view, and video generation, respectively. Furthermore, the low token variance in the early steps [19] reduces the effectiveness of classifier-free guidance in identifying important tokens. To mitigate this, during these initial steps that involve token pruning, we randomly select one token from each  $2 \times 2$  region of the feature map as the destination token.

In Tab. 8 and Fig. 6, we compare the results of ToMeSD [1] under two scenarios: (1) applying token pruning during the early diffusion steps followed by token merging, and (2) using token merging throughout all diffusion steps. The results show that token pruning in the early steps more effectively preserves the generation layout. In Tab. 9, we show that pruning tokens during the first 5-20% denoising steps consistently improves performance across different tasks, highlighting its robustness.

$r$	FID ↓		CLIP ↑	
	w/o pr.	w/ pr.	w/o pr.	w/ pr.
0.10	11.75	<b>11.72</b>	31.81	<b>31.82</b>
0.30	<b>12.16</b>	12.20	31.82	31.82
0.50	<b>13.49</b>	13.50	31.79	31.79
0.60	14.81	14.81	31.79	<b>31.80</b>
0.70	17.51	<b>17.46</b>	31.76	<b>31.78</b>
0.75	21.05	<b>20.89</b>	31.69	<b>31.71</b>

Table 8. Ablation studies on token pruning in early diffusion inference steps. We compare the results of ToMeSD [1] with token pruning in early diffusion inference steps followed by token merging, versus using token merging for all steps. We evaluate using the text-to-image generation task with Stable Diffusion [15] as the base model across various token merging ratios  $r$ .

## C. Discussion and Limitations

Our method demonstrates broad applicability across diffusion models. For multi-guidance scenarios (e.g., InstructPix2Pix [2]), weighted averaging of importance signals based on user preferences could be beneficial. For step-distilled diffusion models [11], which already distills classifier-free guidance into the final model, our approach can still be applied by utilizing alternative importance sig-

Metric	Pruning Steps (m%)			
	0%	5%	10%	20%
<i>Image</i>				
FID ↓	16.27	16.28	16.22	<b>16.13</b>
CLIP ↑	31.77	31.79	<b>31.79</b>	31.79
<i>Multi-View</i>				
PSNR ↑	14.73	<b>14.95</b>	14.75	14.80
SSIM ↑	0.783	0.782	<b>0.787</b>	0.785
LPIPS ↓	0.279	0.269	<b>0.268</b>	0.274
<i>Video</i>				
Score ↑	79.36	79.59	79.63	<b>79.66</b>

Table 9. We apply token pruning to the first  $m\%$  denoising steps (merging ratio = 0.7) and evaluate on three generation tasks.

nals, such as attention maps. However, an interesting direction for future research could involve refining the distillation process to enable the model to predict an additional output: a classifier-free guidance map, which could then be used for better token merging or other innovative applications.

## D. Prompts

We provide the prompts used to generate the qualitative results shown in the paper but not included in the figures.

Text prompts corresponding to the text-to-image generations in Figure 5 of the main paper:

- *Elegant teacup with a delicate floral pattern*
- *Young musician playing guitar on stage*
- *Colorful butterfly with wings fully spread*

Text prompts corresponding to the text-to-image generations in Figure 6 of the main paper:

- *A cute cat*
- *A real beautiful face*
- *A small cactus with a happy face in the Sahara desert*

Text prompts corresponding to the text-to-video generations in Figure 8 of the main paper:

- *Tower*
- *A jellyfish floating through the ocean, with bioluminescent tentacles*
- *In a still frame, the ornate Victorian streetlamp stands solemnly, adorned with intricate ironwork and stained glass panels*

In Fig. 7, we show image prompts corresponding to the image-conditioned multi-view generations in Figure 7 of the main paper.

In Fig. 8, we show image prompts corresponding to the image-conditioned multi-view generations in Fig. 3.

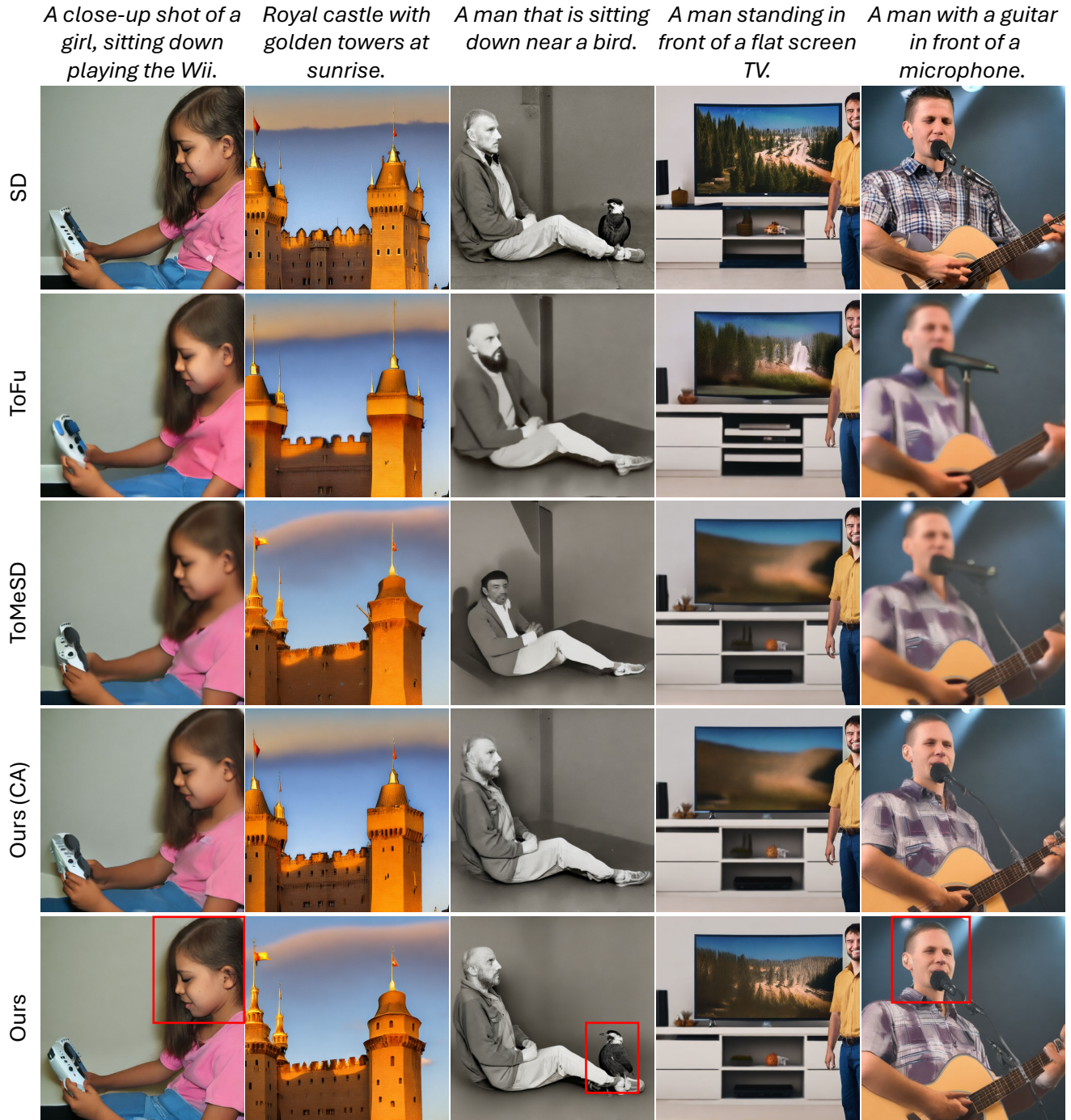
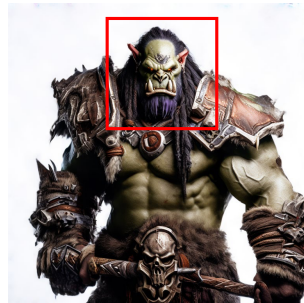
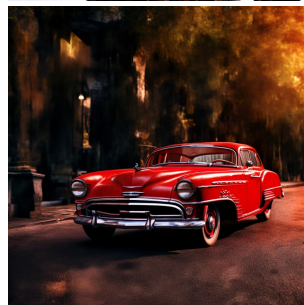
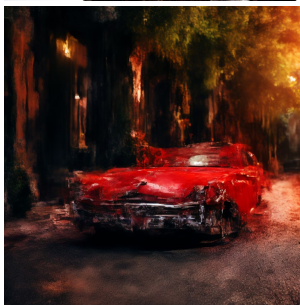


Figure 1. **Additional comparison of text-to-image generation.** The first row shows results from Stable Diffusion (SD) [15], while the subsequent rows show SD combined with ToFu [6], ToMeSD [1], our method using cross-attention (CA) map, and our method using classifier-free guidance. The token merging ratio is 0.7. Our method outputs finer details, as highlighted in red boxes. Notably, the variant of our method that utilizes the cross-attention map also achieves better generation details compared to baseline methods, demonstrating the generalization ability of our method. Best viewed with zoom-in for clarity.

*Product  
photography,  
world of  
warcraft orc  
warrior, white  
background.*



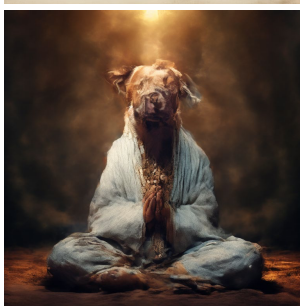
*A photo of a  
red car.*



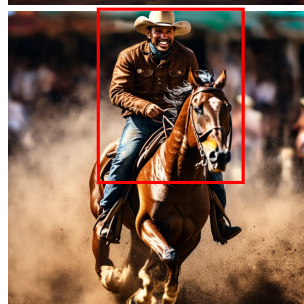
*Chinese  
painting of  
grapes.*



*A dog that has  
been  
meditating all  
the time.*



*A man riding a  
brown horse  
at a rodeo.*



(a) PixArt- $\alpha$

(b) ToMeSD

(c) Ours

Figure 2. **Additional comparison for token merging applied to diffusion transformer.** We apply ToMeSD [1] and our token merging method to PixArt- $\alpha$  [3] for text-to-image synthesis, using a token merging ratio of 0.4. We highlight our generation details with red boxes. Best viewed with zoom-in for clarity.



Figure 3. **Additional qualitative comparison of multi-view diffusion.** Token merging is applied to the multi-view diffusion model, Zero123++ [16], with merging ratio as 0.6. Our method outputs finer details, as highlighted in red boxes. Best viewed with zoom-in.

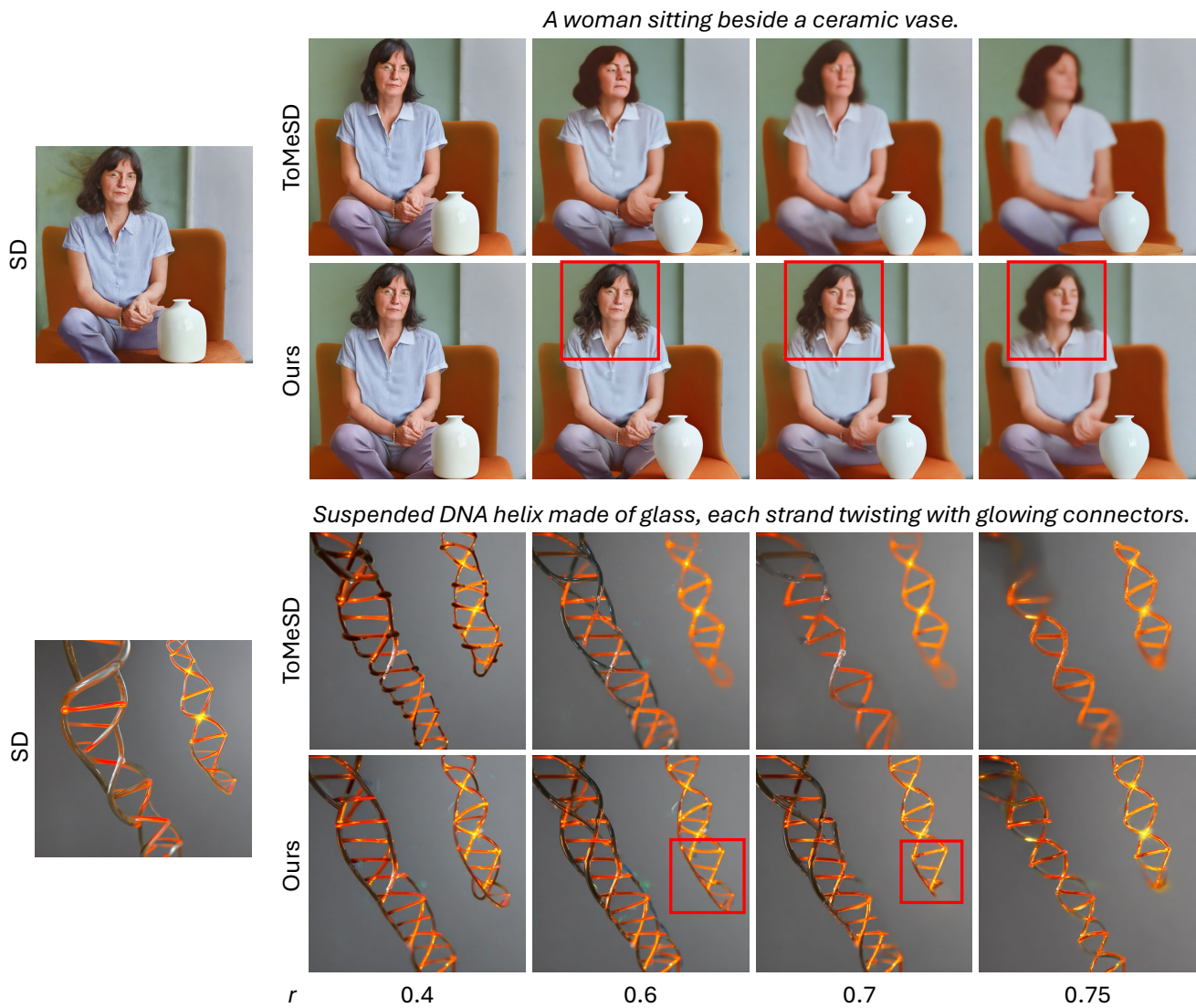


Figure 4. We provide an additional comparison between ToMeSD [1] and our method when applied to Stable Diffusion [15] across various merging ratios  $r$ . For reference, the results of Stable Diffusion without token merging are shown on the left. Our method outputs finer details, as highlighted in red boxes. Best viewed with zoom-in for clarity.

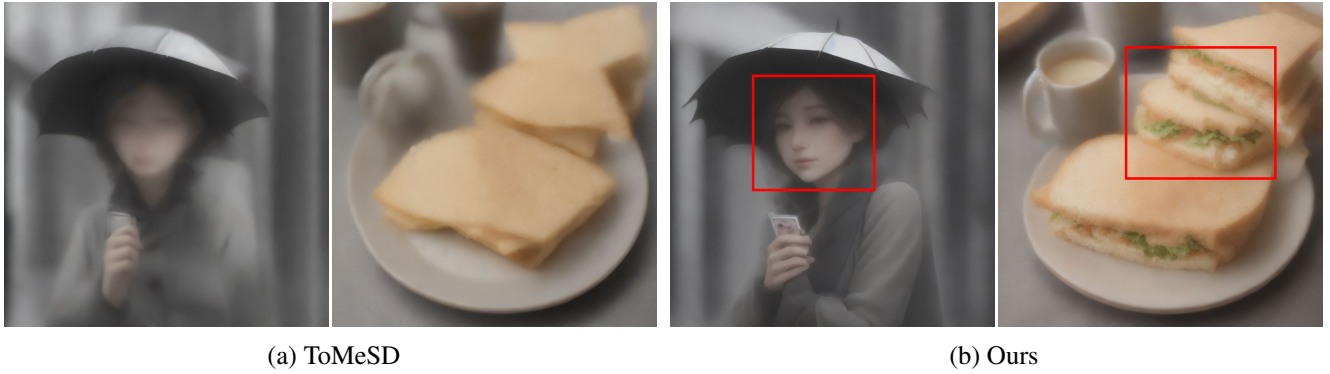


Figure 5. Token merging results on the latent consistency model (LCM\_Dreamshaper.v7) with a merging ratio of 0.7.

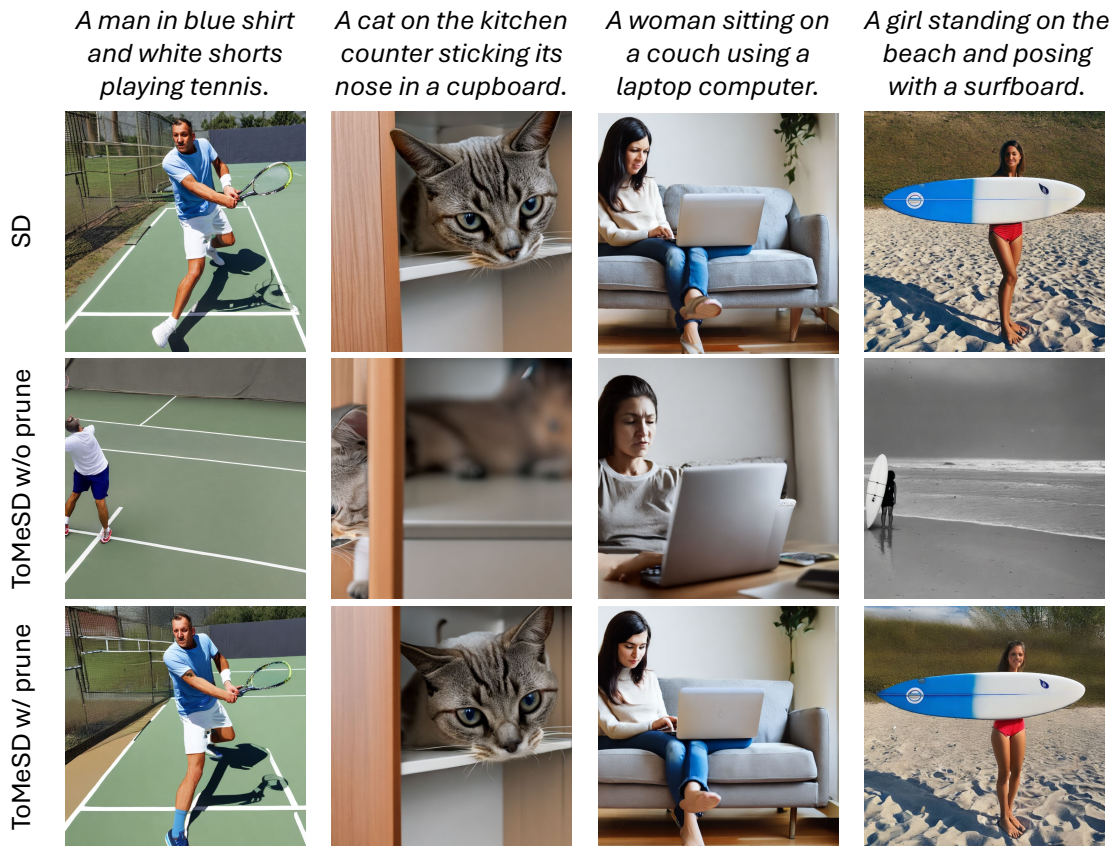


Figure 6. We compare the results of ToMeSD [1] with token pruning in early diffusion inference steps followed by token merging (w/ prune), versus using token merging for all steps (w/o prune). We use Stable Diffusion [15] as the base model and a merging ratio of 0.6.



Figure 7. Image prompts for Figure 7 of the main paper.



Figure 8. Image prompts for Fig. 3.

## References

- [1] 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](#), [3](#), [4](#), [5](#), [7](#), [8](#)
- [2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. [3](#)
- [3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\alpha$ : Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. [2](#), [5](#)
- [4] 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] 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](#)
- [6] 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](#)
- [7] 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. [1](#)
- [8] 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. [1](#)
- [9] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023. [1](#)
- [10] 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. [1](#)
- [11] 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. [3](#)
- [12] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. [2](#)
- [13] 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. [2](#)
- [14] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020. [2](#)
- [15] 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](#), [7](#), [8](#)
- [16] 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](#), [6](#)
- [17] Junhyuk So, Jungwon Lee, and Eunhyeok Park. Frdiff: Feature reuse for universal training-free acceleration of diffusion models. *arXiv preprint arXiv:2312.03517*, 2023. [1](#)
- [18] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. [1](#)
- [19] 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. [3](#)