

A Appendix

A.1 Comparison with other methods

Comparison with ToMeSD. Recent studies [36, 66] have evaluated ToMeSD [5] on DiT models and reported significant performance drops, which are much lower than those achieved by our method. Our approach is fundamentally different from ToMeSD. While ToMeSD reduces and recovers tokens generically at every layer, our method specifically analyzes the local-global relationships unique to DiT models. Based on these insights, we strategically reduce and recover tokens. We have evaluated ToMeSD on DiT and compared it with our method, as shown in Table 8.

Comparison with DyDiT. DyDiT [66] represents the current state-of-the-art approach for enhancing the efficiency of Diffusion Transformers (DiTs). In contrast to our method, which emphasizes token reduction, DyDiT implements pruning across tokens, attention heads, and channels. A comparative analysis of our method against DyDiT, as detailed in Table 9 using ImageNet 512×512 images, illustrates a substantial advantage in computational efficiency for our approach. Specifically, our method achieves a 46% reduction in FLOPs, markedly surpassing DyDiT’s 29% reduction, coupled with a 145% increase in inference speed, significantly outperforming the 31% improvement achieved by DyDiT. The only trade-off is a minor increase in the Fréchet Inception Distance (FID), where our method’s FID is 0.08 higher than that of DyDiT. Nevertheless, this difference in FID is nearly imperceptible to human observers in practical applications.

A.2 More experimental settings in Sparse-Latte

Latte comprises two distinct types of transformer blocks: spatial transformers, which focus on capturing spatial information, and temporal transformers, which capture temporal information. To accommodate this separation of spatial and temporal feature extraction, we adjust our model’s schedule within the sparse-dense token module. Specifically, we employ two transformers to prune spatial tokens and temporal tokens separately. The sparse tokens are then processed by two sparse transformers, and finally two transformers are used to recover the temporal and spatial tokens. Dense transformers are discarded in this setup. All other model configurations are the same as in DiT, and we follow the original training settings and hyperparameters used for Latte.

A.3 Captions in Figure 5

Column 1: The image depicts a beach scene with a large body of water, such as a lake or ocean, and a sandy shoreline. The beach is filled with people, including a group of people swimming in the water.

Column 2: The image depicts a beautiful outdoor dining area with a large number of tables and chairs arranged in a row, overlooking a picturesque lake. The tables are covered with white tablecloths, and there are several umbrellas providing shade for the guests. The scene is set in a lush green field, with a large building in the background, possibly a hotel or a restaurant. The tables are arranged in a way that allows for an unobstructed view of the lake, creating a serene and relaxing atmosphere for the diners. The image has a stylish and elegant feel, with the attention to detail in the table arrangement and the choice of location contributing to a memorable dining experience.

Column 3: The image depicts a busy city street at night, with a group of people standing outside a restaurant and a bar. The scene is set in a European city, and the atmosphere is lively and bustling.

Column 4: The image features a vintage-style train, parked on a track surrounded by trees and grass. The train appears to be an old-fashioned steam engine, which is a type of locomotive powered by steam. The train is positioned in a park-like setting, with a tree-lined path nearby. The scene is set in a sunny day, creating a pleasant atmosphere. The image has a nostalgic and historical feel, evoking a sense of the past and the charm of old-time trains.

A.4 Additional visualization

We provide additional visualizations at a resolution of 512×512 on SparseDiT-XL from Figure 4 to Figure 15. The class labels, including “arctic wolf”, “volcano”, “cliff drop-off”, “balloon”, “sulphur-crested cockatoo”, “lion”, “otter”, “coral reef”, “macaw”, “red panda”, “husky”, and “panda”,

Table 8: Comparison of our method with ToMeSD on DiT models.

Model	Method	FID	Speed-up
DiT-B	ToMeSD	29.24	+20%
	Ours	8.23	+68%
DiT-XL	ToMeSD	14.74	+66%
	Ours	2.38	+87%

Table 9: Comparison of our method with DyDiT on DiT models.

Model	Method	FLOPs	FID	Speed-up
DiT-XL	DyDiT	375 (-29%)	2.88	+31%
	Ours	286 (-46%)	2.96	+145%

49 correspond to the same cases presented in DiT. Readers can compare our results with those of DiT.
 50 Our samples demonstrate comparable image quality and fidelity. The classifier-free guidance scale is
 51 set to 4.0, and all samples shown here are uncured.



Figure 4: Uncurated 512×512 SparseDiT-XL samples.
 Classifier-free guidance scale = 4.0
 Class label = "arctic wolf" (270)



Figure 5: Uncurated 512×512 SparseDiT-XL samples.
 Classifier-free guidance scale = 4.0
 Class label = "volcano" (980)



Figure 6: Uncurated 512×512 SparseDiT-XL samples.
 Classifier-free guidance scale = 4.0
 Class label = "husky" (250)



Figure 7: Uncurated 512×512 SparseDiT-XL samples.
 Classifier-free guidance scale = 4.0
 Class label = "sulphur-crested cockatoo" (89)



Figure 8: Uncurated 512×512 SparseDiT-XL samples.
 Classifier-free guidance scale = 4.0
 Class label = “cliff drop-off” (972)



Figure 9: Uncurated 512×512 SparseDiT-XL samples.
 Classifier-free guidance scale = 4.0
 Class label = “balloon” (417)

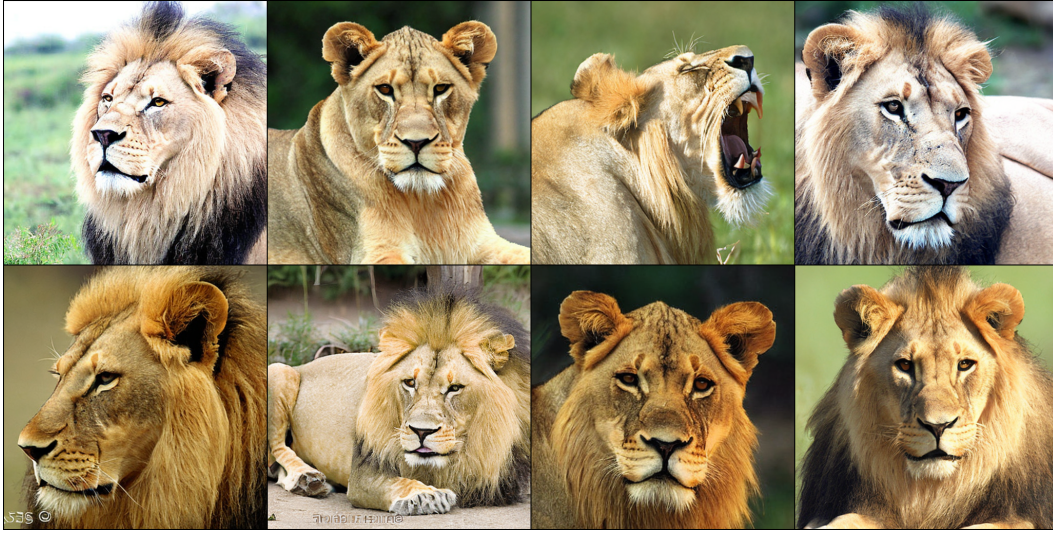


Figure 10: Uncurated 512×512 SparseDiT-XL samples.
 Classifier-free guidance scale = 4.0
 Class label = "lion" (291)



Figure 11: Uncurated 512×512 SparseDiT-XL samples.
 Classifier-free guidance scale = 4.0
 Class label = "otter" (360)



Figure 12: Uncurated 512×512 SparseDiT-XL samples.
 Classifier-free guidance scale = 4.0
 Class label = "red panda" (387)



Figure 13: Uncurated 512×512 SparseDiT-XL samples.
 Classifier-free guidance scale = 4.0
 Class label = "panda" (388)



Figure 14: Uncurated 512×512 SparseDiT-XL samples.
 Classifier-free guidance scale = 4.0
 Class label = “coral reef” (973)



Figure 15: Uncurated 512×512 SparseDiT-XL samples.
 Classifier-free guidance scale = 4.0
 Class label = “macaw” (88)