

Trainable Log-linear Sparse Attention for Efficient Diffusion Transformers

Supplementary Material

Table 1. Hyperparameters of Pixel DiT trained on FFHQ and ImageNet of various resolutions. Models with different attention implementations have identical configurations. FFHQ models are trained on one H200 GPU and ImageNet models are trained on four H200 GPUs.

Model	FFHQ-32	FFHQ-128	FFHQ-256	FFHQ-512	ImageNet-128	ImageNet-256
Patch Size	1 × 1	1 × 1	1 × 1	1 × 1	4 × 4	4 × 4
DiT Config	DiT-S	DiT-S	DiT-S	DiT-S	PixelFlow-L	PixelFlow-L
Pretrained Model	-	FFHQ-32	FFHQ-128	FFHQ-256	-	ImageNet-128
SNR Rescale	1	2	4	8	1	1
Epochs	40	20	10	2	40	10
Batch Size	64	16	4	1	32	8
Learning Rate	1 × 10 ⁻⁴	1 × 10 ⁻⁴	1 × 10 ⁻⁴	1 × 10 ⁻⁴	1 × 10 ⁻⁴	1 × 10 ⁻⁴

Table 2. Ablation study results of Log-linear Sparse Attention

(a) Enrichment Levels			
Configuration	L_e	FID	Throughput
LLSA FFHQ-128 ($L = 2$)	0	27.98	500.38
LLSA FFHQ-128 ($L = 2$)	1	25.49	467.35
LLSA FFHQ-128 ($L = 2$)	2	24.37	436.40
(b) Extension to 512 × 512 Resolution			
Configuration	FID	Throughput	
LLSA FFHQ-256 ($L = 2$)	39.29	375.34	
LLSA FFHQ-512 ($L = 1$)	-	44.90	
LLSA FFHQ-512 ($L = 2$)	39.26	292.66	
LLSA FFHQ-512 ($L = 3$)	40.93	357.70	

1. Implementation Details

1.1. Clarification of Kernel Implementation

Both SLA [6] and VSA [7] use inefficient sparse backward implementations. SLA applies the standard mask-based sparse block sparse attention backward, skipping the unused queries for each key. VSA implements a preprocessing kernel that extracts query indices for each key from the binary sparse mask, but it still requires $O(N^2)$ complexity to construct the binary mask.

In Table 2 of the paper, to fairly evaluate the algorithms' throughput rather than their GPU implementations, we reimplement their sparse backward kernels with our efficient sparse index transpose kernel. This conservative experimental setting significantly improves the baselines' throughput, but results indicate that LLSA is still more efficient than SLA and VSA for an equal number of effective sparse tokens.

In Fig. 3 of the paper, we compare the kernel efficiency of LLSA with the official implementations of SLA and VSA. We modify their block size configurations to support $B = 16$ inference.

1.2. Training Configuration

We provide the detailed training configurations on FFHQ [4] and ImageNet [2] in Table 1.

2. Additional Experiment Results

2.1. Additional Ablation Results

More enrichment levels lead to better quality. We train three two-level LLSA DiT on FFHQ-128 with different KV enrichment layers L_e , shown in Table 2a. More enrichment levels increase the effective token number and generation quality, while slightly reducing throughput.

512 × 512 pixel token sequence generation. To assess the scalability of LLSA on substantially longer token sequences, we train DiT-S on FFHQ-512 using different num-

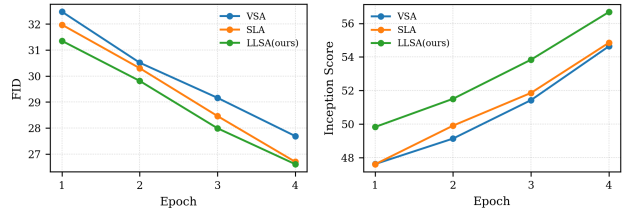


Figure 1. The FID and Inception Score curves of the first 4 epochs using VSA, SLA, and LLSA on PixelFlow ImageNet-256 benchmark.

bers of hierarchical levels L (Table 2b). The single-level LLSA ($L = 1$) fails to converge within a reasonable time budget due to its $O(N^2)$ selection cost and coarse tokens. Increasing to $L = 2$ dramatically improves throughput. Further extending to $L = 3$ yields additional speed gains. The scaling of per-token throughput, from 375.34 (at 256 × 256) to 357.70 (at 512 × 512), closely follows the $O(N \log N)$ complexity of LLSA.

2.2. Training Curves

We show the training FID [3] and Inception Score [5] curves on PixelFlow [1] ImageNet-256 benchmark in Fig. 1. The generation quality of LLSA is consistently better than that of baseline attention approaches throughout the training process.

3. Qualitative Results

We present qualitative results of LLSA in this section. In Fig. 2, we show the samples generated from LLSA DiT-S trained on FFHQ-128, FFHQ-256, and FFHQ-512. For FFHQ-512, the model is only trained for 2 epochs. We believe that better quality can be obtained by longer training. In Fig. 3, we compare the ImageNet-256 samples generated

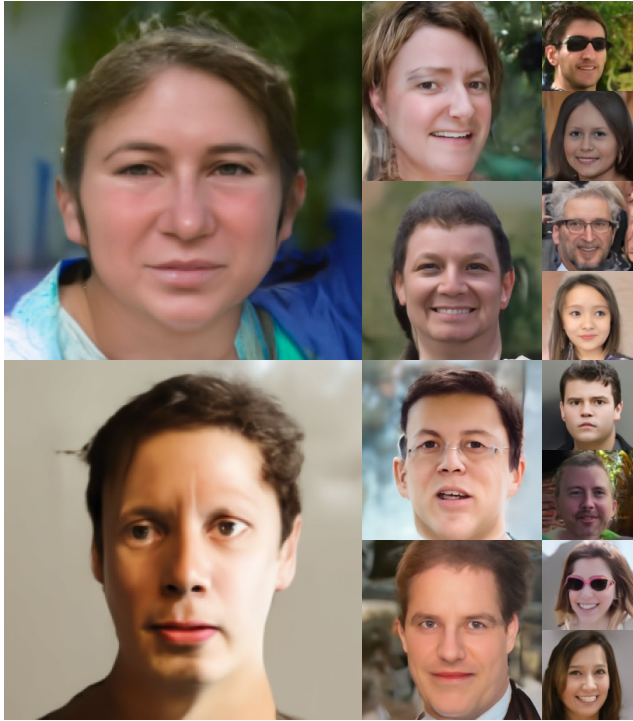


Figure 2. The qualitative results of pixel space DiT-S using LLSA trained on FFHQ-128, FFHQ-256, and FFHQ-512. For FFHQ-512, the model is only trained for 2 epochs. We believe that better quality can be obtained by longer training.

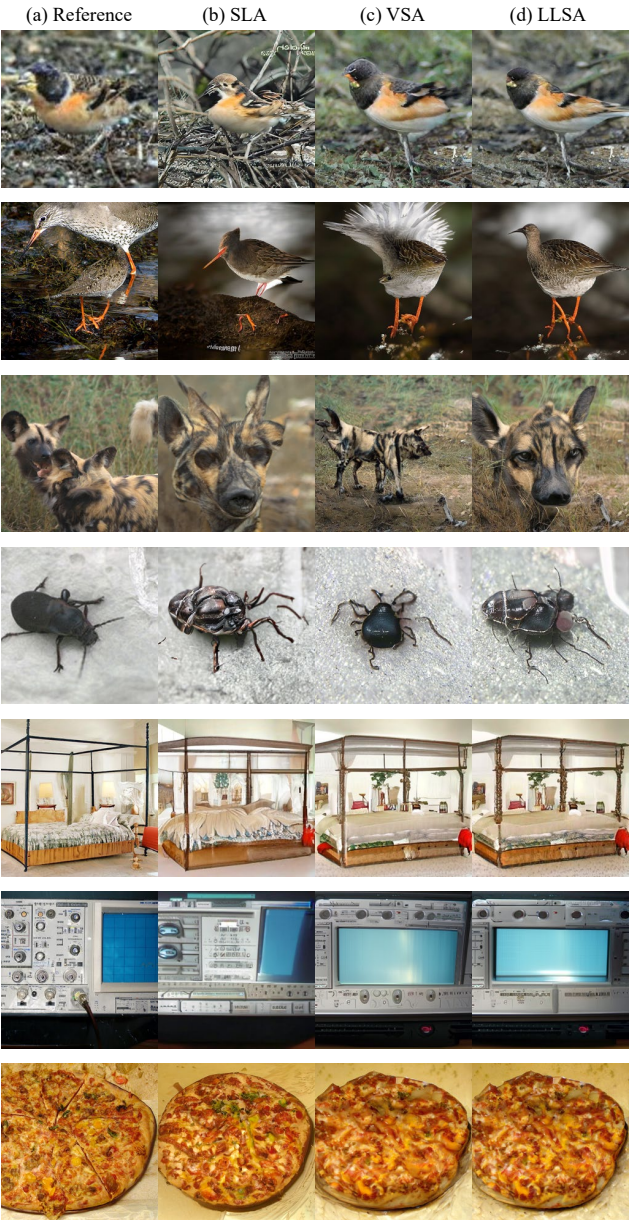


Figure 3. The qualitative comparison of SLA, VSA, and LLSA trained on PixelFlow-L ImageNet-256. The reference images are generated by a well-trained full-attention PixelFlow model from the official repository.

by LLSA PixelFlow-L with the those produced by SLA and VSA variants. For reference, we also include samples generated by the official best-performing PixelFlow model.

References

[1] Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[3] 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. 1

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[5] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1

[6] Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kai-

wen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, Joseph E. Gonzalez, Jun Zhu, and Jianfei Chen. Sla: Beyond sparsity in diffusion transformers via fine-tunable sparse-linear attention. *arXiv preprint arXiv:2509.24006*, 2025. 1

[7] Peiyuan Zhang, Haofeng Huang, Yongqi Chen, Will Lin, Zhengzhong Liu, Ion Stoica, Eric P Xing, and Hao Zhang. Faster video diffusion with trainable sparse attention. *arXiv e-prints*, pages arXiv–2505, 2025. 1