DiT-Pruner: Pruning Diffusion Transformer Models for Text-to-Image Synthesis Using Human Preference Scores

Youngwan Lee[1,2], Yong-Ju Lee[1], and Sung Ju Hwang[2,3]

[1] ETRI, South Korea
[2] KAIST, South Korea
[3] DeepAuto.ai, South Korea

Abstract. Despite their remarkable performance compared to U-Net-based text-to-image (T2I) models, Diffusion Transformer (DiT)-based T2I models incur substantial inference costs due to their large model size and computational requirements. While recent efforts in layer pruning for large language models (LLMs) have found redundancy in Transformers, attempts to prune DiT models have not yet been explored. In this work, we propose a simple layer-pruning method specifically for DiT-based T2I models. Unlike pruning methods for LLMs that identify unimportant layers based on the similarity across layers or between input/output features of each layer, our approach prunes layers using a *direct* quality metric based on human preference scores, which more precisely reflects the overall generated image quality. In experiments using the Pixart-Σ model, our method outperforms similarity-based methods across different pruning ratios. Additionally, we find that fine-tuning with a knowledge distillation objective can further restore performance.

Keywords: Diffusion Transformer · Text-to-image synthesis · Pruning

1 Introduction

Recent diffusion transformer (DiT)-based text-to-image (T2I) synthesis models [3–6] have demonstrated superior generation quality compared to conventional U-Net-based models. However, this improvement comes at the cost of significantly increased computational requirements and larger model sizes, leading to substantial inference costs for high-resolution image generation. Several attempts [1,7,12,14–16,23] to prune large language models (LLMs) using Transformers [19] have been proposed very recently, while pruning for DiT-based T2I models has not yet been explored.

It is crucial to determine the importance of each layer during the pruning process. In recent LLM literature, relative magnitude [16], angular distance [7], and cosine similarity [15] are used to measure the similarity across layers or between the input and output of each layer. The underlying assumption is that if the similarity between features is high, the layer can be considered unimportant and thus be removed. However, these similarity-based metrics may not guarantee

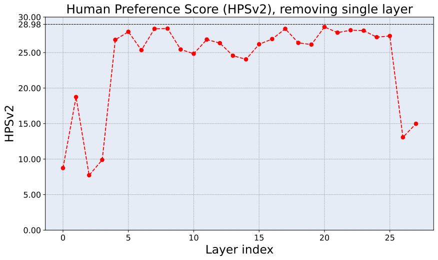

Fig. 1: Performance of removing a single layer.

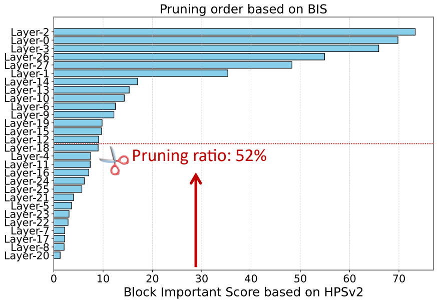

Fig. 2: Pruning order by Block Importance Score based on Human Preference.

the final generation quality because solely comparing features between the input and output of individual layers does not reflect the overall visual quality in the T2I domain.

To address these challenges, this work empirically studies a simple layer-pruning strategy specifically suited for DiT-based T2I models. We define the block importance score (BIS) using a more direct quality metric, Human Preference Score (HPS) [20, 21], which reflects the actual visual quality. We compute the BIS for each layer by removing a single layer and evaluating the model's performance with that layer removed using HPSv2 [20]. In this manner, we obtain a pruning order list by BIS for the DiT model and perform layer pruning in ascending order of BIS. For experiments, we analyze the redundancy in Pixart-Σ [3] as a DiT-based T2I model and find that our method maintains 85% performance while reducing approximately 52% of model parameters and computation, outperforming other similarity-based methods. We can summarize two key lessons in this study as follows:

– For identifying redundant layers in DiT-based T2I models, a direct quality metric based on human preference is superior to similarity-based metrics.
– For healing the pruned DiT model, fine-tuning with knowledge distillation is more effective.

2 DiT-Pruner

In this work, we use Pixart-Σ [3] as our baseline Diffusion Transformer architecture for text-to-image synthesis. We measure block importance scores using a *direct* quality metric, the **human preference score (HPS)** [20, 21], which is learned to evaluate generation quality based on a human preference dataset. To this end, we investigate the block influence in the DiT architecture by removing a single block, generating images with the model, and evaluating the generated images using HPSv2 [20]. Specifically, we obtain the block importance score (BIS$_i$) of the i-th block in Pixart-Σ by calculating $\frac{hps_o - hps_i}{hps_o}$ on the test set (3.2K samples) of the HPSv2 dataset [20], where hps_o and hps_i are the mean scores of HPSv2 for the original Pixart-Σ and Pixart-Σ with the i-th block removed, respectively. This indicates how much the performance deteriorates when

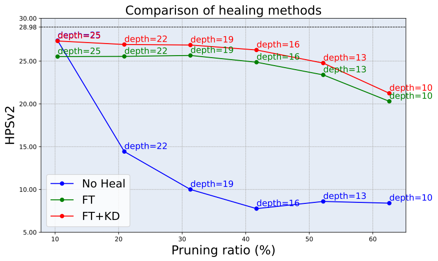

Fig. 3: Healing strategies.

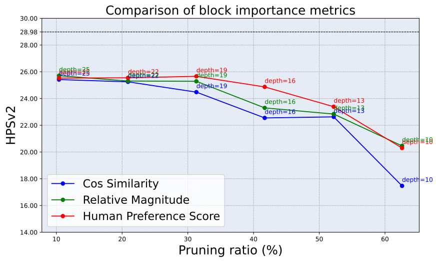

Fig. 4: block importance metrics.

one block is removed compared to the original model. For example, if hps_i is low, it signifies that the i-th block has a significant impact on performance, indicating its high importance. In this manner, we perform single-layer removal for all blocks (*e.g.*, 28 blocks) in Pixart-Σ and then obtain the block importance scores as shown in Fig. 1. We then derive the final pruning order based on BIS, as shown in Fig. 2. Using this pruning order, we can prune Pixart-Σ in ascending order of scores according to the required compression ratio.

Interestingly, we observe in Fig. 1 that some blocks in the middle groups (*e.g.*, 4th to 25th blocks) show minimal performance drop, while the very early and last blocks (*e.g.*, 0, 1, 2, 3, 26, 27) exhibit drastic performance degradation. This result is also observed qualitatively when we generate an image with the single-block removal experiments in Figs. 6 to 8. A similar phenomenon has also been observed in LLM literature [7, 15, 23].

3 Experimental Results

Implementation details. By using the official code and pre-trained weights of Pixart-Σ [3], we conduct experiments. We use the Pixart-Σ-256px model for all experiments. For fine-tuning, we train models on the LAION-POP dataset [17] for 50 epochs. More training details are described in Appendix A. For evaluation, we use the Human Preference Score v2 (HPSv2) [20] as a visual aesthetics metric instead of the FID [8] because recent works [2, 9, 21] have claimed that the FID score does not correlate well with visual quality in text-to-image synthesis tasks.
Comparison of healing methods. We evaluate the performance degradation of the DiT architecture after pruning under three scenarios: without any healing, with supervised fine-tuning (FT), and with both FT and a knowledge distillation (KD) objective for healing, as shown in Fig. 3. For the KD objective, we follow the self-attention-based KD training recipe described in [10]. As shown in Fig. 3, the case without healing shows a drastic performance drop after pruning six blocks (depth=22) with a pruning ratio of 20.8%. Contrary to findings in LLM literature [7, 15] that show sharp accuracy drops at pruning fractions around 30%-40% in LLaMA-2 models [18], our T2I model does not exhibit a flat performance region. We speculate that since Pixart-Σ has fewer layers (*e.g.*, 28) compared to LLaMA-2 models (*e.g.*, 32, 40, and 80), the importance of each layer is greater in Pixart-Σ, leading to a more significant performance drop at lower pruning ratios. Interestingly, FT restores the damaged performance, and

Table 1: Peformance comparison with different block importance metrics. All models are trained with supervised fine-tuning and knowledge-distillation objectives.

BIS measure	pruning ratio	anime	photo	painting	concept-art	HPSv2 (mean)
Dense (Pixart-Σ [3])	-	30.72	27.58	28.55	29.06	28.98
Cos similarity [15]	52.19%	24.15	22.79	23.01	23.58	23.38
Relative magnitude [16]	52.19%	25.43	23.59	24.03	24.27	24.33
Ours (HPS)	52.19%	**25.88**	**23.86**	**24.47**	**24.88**	**24.77**
Cos similarity [15]	31.31%	26.91	24.49	25.2	25.56	25.54
Relative magnitude [16]	31.31%	27.56	25.40	25.60	25.96	26.13
Ours (HPS)	31.31%	**28.32**	**25.65**	**26.72**	**26.85**	**26.88**

KD enables further improvement. From these results, we can infer that pruned T2I models (damaged) need "healing" starting from lower pruning ratios and that combining KD with FT is an essential solution for healing.

Comparison of block importance metrics. We compare our direct quality metric based on Human Preference Score (HPS) for block important score with other similarity-based metrics, cosine similarity [15] and relative magnitude [16] used in LLM-pruning. For this comparison, we also obtain block important scores by computing cosine similarity as [15] and relative magnitude as [11] between the input and output features of each layer. In Fig. 5, we visualize the pruning orders by different metrics. Fig. 4 shows performance comparison at different pruning ratios, and these results are trained by supervised fine-tuning (FT). We can observe that our method using the direct quality metric based on HPS shows better performance across pruning ratios. In addition, when trained with supervised fine-tuning and knowledge-distillation objectives together, our method consistently outperforms other metrics with the different pruning ratio models in Tab. 1. These results suggest that using a direct quality metric based on human preference, which reflects the final image generation quality, is essential for more sensitive T2I models compared to the feature similarity metrics used in LLM pruning [7, 15, 16].

4 Conclusion and Future works

In this work, we empirically studied a simple layer pruning method for the DiT-based T2I model, finding that the direct generation quality metric based on human preference score is essential for identifying redundant layers. However, there is still room for further exploration in pruning large-scale DiT-based T2I models. For future work, we aim to apply our pruning method to state-of-the-art DiT models such as SD3 [5] and Lumina-T2X [6] to evaluate its effectiveness. Additionally, while our approach focuses on block-level pruning, it can be further explored for fine-grained layer pruning [12, 23] within transformer blocks, taking into account the distinct roles of self-attention and MLP components. Furthermore, inspired by techniques [1, 14, 22] in LLM literature, we plan to investigate width-level pruning alongside layer-level pruning (depth).

Acknowledement. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training).

References

1. An, Y., Zhao, X., Yu, T., Tang, M., Wang, J.: Fluctuation-based adaptive structured pruning for large language models. In: AAAI (2024) 1, 4
2. Betzalel, E., Penso, C., Navon, A., Fetaya, E.: A study on the evaluation of generative models. arXiv preprint arXiv:2206.10935 (2022) 3
3. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. arXiv preprint arXiv:2403.04692 (2024) 1, 2, 3, 4, 7, 8, 9
4. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis (2023) 1
5. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024) 1, 4
6. Gao, P., Zhuo, L., Lin, Z., Liu, C., Chen, J., Du, R., Xie, E., Luo, X., Qiu, L., Zhang, Y., et al.: Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. arXiv preprint arXiv:2405.05945 (2024) 1, 4
7. Gromov, A., Tirumala, K., Shapourian, H., Glorioso, P., Roberts, D.A.: The unreasonable ineffectiveness of the deeper layers. arXiv preprint arXiv:2403.17887 (2024) 1, 3, 4
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017) 3
9. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. In: ICCV (2023) 3
10. Lee, Y., Park, K., Cho, Y., Lee, Y.J., Hwang, S.J.: Koala: Empirical lessons toward memory-efficient and fast diffusion models for text-to-image synthesis. arXiv preprint arXiv:2312.04005 (2023) 3, 7
11. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: CVPR (2023) 4
12. Liu, S., Zeng, C., Li, L., Yan, C., Fu, L., Mei, X., Chen, F.: Foldgpt: Simple and effective large language model compression scheme. arXiv preprint arXiv:2407.00928 (2024) 1, 4
13. Luo, Y., Ren, X., Zheng, Z., Jiang, Z., Jiang, X., You, Y.: Came: Confidence-guided adaptive memory efficient optimization. arXiv preprint arXiv:2307.02047 (2023) 7
14. Ma, X., Fang, G., Wang, X.: Llm-pruner: On the structural pruning of large language models. Advances in neural information processing systems **36**, 21702–21720 (2023) 1, 4
15. Men, X., Xu, M., Zhang, Q., Wang, B., Lin, H., Lu, Y., Han, X., Chen, W.: Shortgpt: Layers in large language models are more redundant than you expect. arXiv preprint arXiv:2403.03853 (2024) 1, 3, 4
16. Samragh, M., Farajtabar, M., Mehta, S., Vemulapalli, R., Faghri, F., Naik, D., Tuzel, O., Rastegari, M.: Weight subcloning: direct initialization of transformers using larger pretrained ones. arXiv preprint arXiv:2312.09299 (2023) 1, 4
17. Schuhmann, C., Bevan, P.: Laion pop: 600,000 high-resolution images with detailed descriptions. `https://huggingface.co/datasets/laion/laion-pop` (2023) 3, 7

18. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 3
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS (2017) 1
20. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023) 2, 3
21. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Better aligning text-to-image models with human preference. In: ICCV (2023) 2, 3
22. Xia, M., Gao, T., Zeng, Z., Chen, D.: Sheared llama: Accelerating language model pre-training via structured pruning. arXiv preprint arXiv:2310.06694 (2023) 4
23. Zhong, L., Wan, F., Chen, R., Quan, X., Li, L.: Blockpruner: Fine-grained pruning for large language models. arXiv preprint arXiv:2406.10594 (2024) 1, 3, 4

Appendix

A Implementation details

For finetuning (FT) or FT with knowledge distillation (KD) training objectives, following the original Pixart-Σ training recipe [3], we train models for 50 epochs on a subset (491,567 samples) of LAION-POP dataset[4] [17] with CAME optimizer [13], a batch size of 128 and a learning rate of 2e-5. For rapid verification, we train models for only 50 epochs, so we can expect that the longer training epochs may further improve performance. For KD, we use Pixart-Σ as a teacher model and follow the self-attention-based distillation strategy as in [10]. For inference, we use the same sampler in Pixart-Σ with 20 denoising steps.

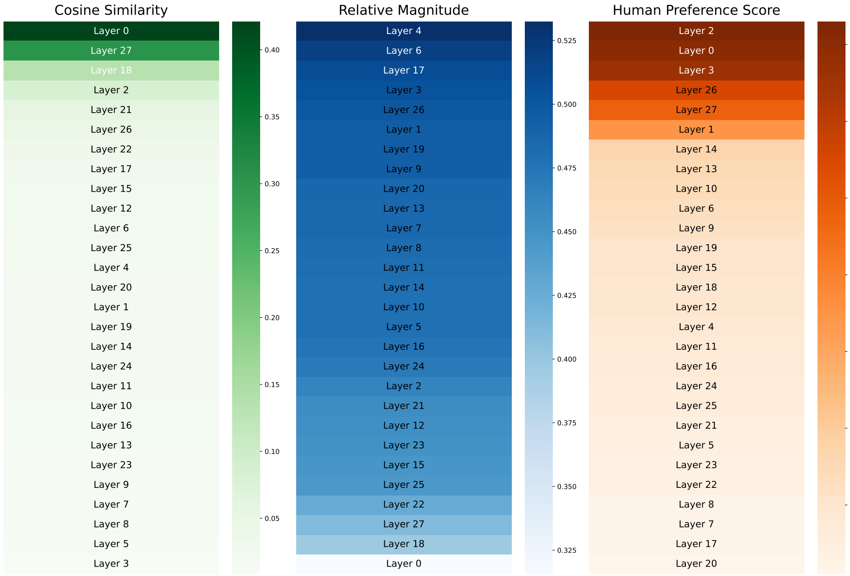

Fig. 5: Comparison for Block Importance score metrics.

[4] https://huggingface.co/datasets/Ejafa/ye-pop

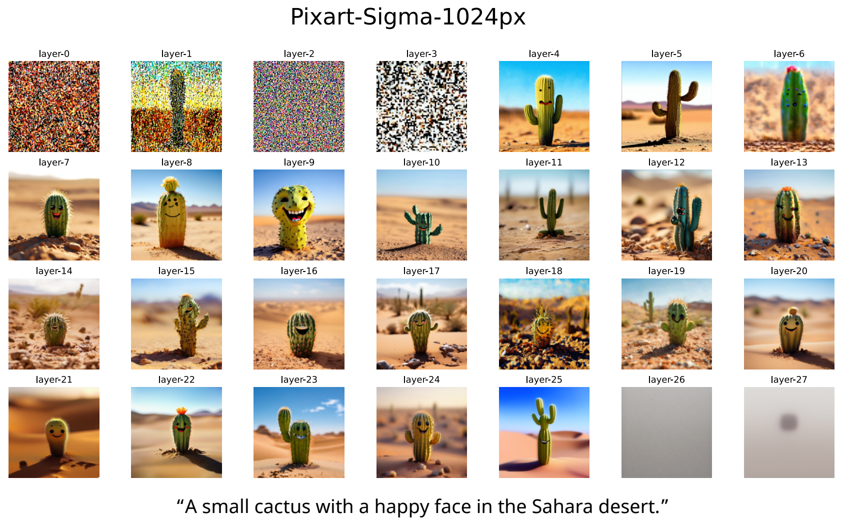

Fig. 6: Generated images after removing each block with Pixart-Σ-1024px [3]

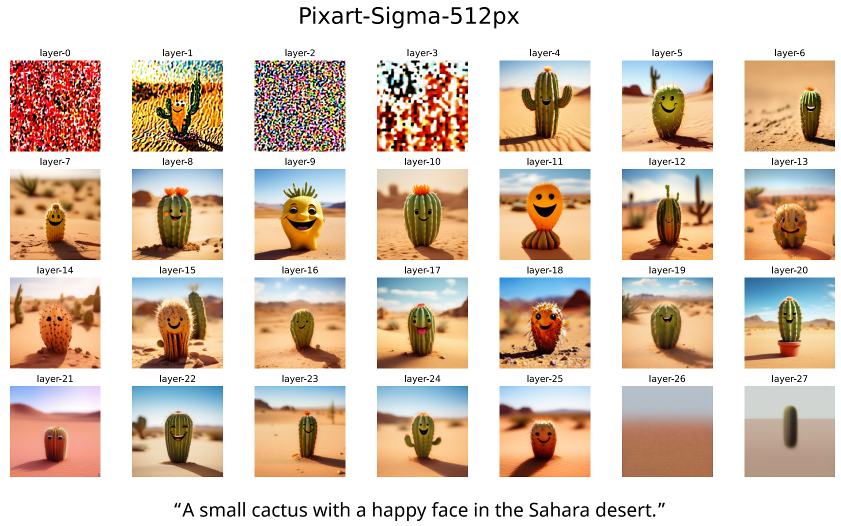

Fig. 7: Generated images after removing each block with Pixart-Σ-512px [3]

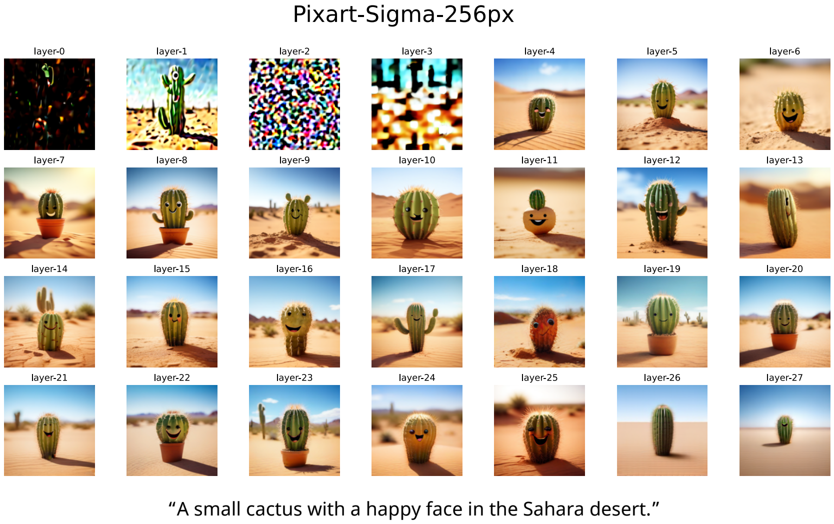

Fig. 8: Generated images after removing each block with Pixart-Σ-256px [3]