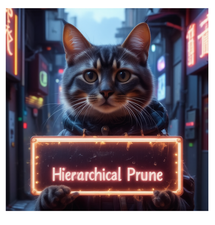
HierarchicalPrune: Position-Aware Compression for Large-Scale Diffusion Models



Young D. Kwon*, Rui Li*, Sijia Li, Da Li, Sourav Bhattacharya, and Stylianos I. Venieris
Samsung AI Center - Cambridge, UK
*Joint First Authors {yd.kwon, rui.li}@samsung.com

Samsung Research
AI Center-Cambridge

Association for the Advancement of Artificial Intelligence

TL;DR: We identify dual hierarchy of SOTA large-scale diffusion models at inter-block and intra-block levels, and design an effective compression framework that significantly reduces resource demands while preserving image generation quality.

Introduction

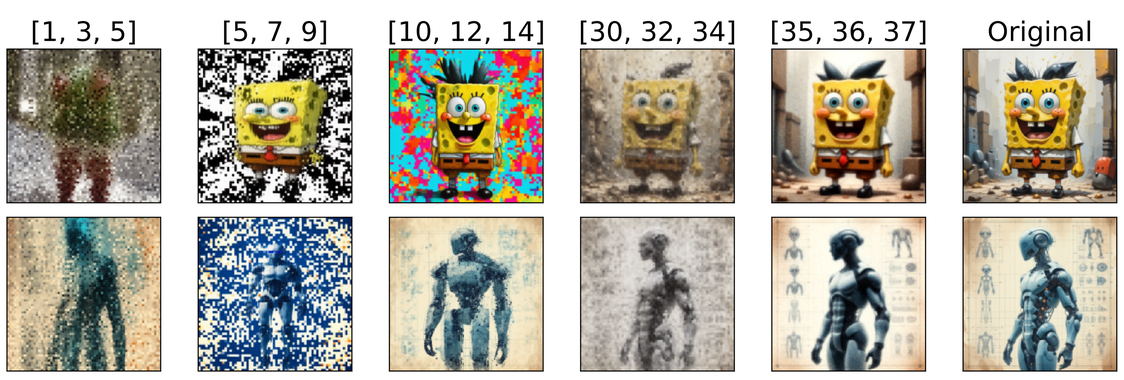
🎨 SOTA diffusion models (DMs), e.g. Stable Diffusion 3.5 [1], FLUX.1 [2], employ multi-modal diffusion Transformer (MMDiT) backbone, with large parameter scale (8-11B), leading to high latency and memory requirements.

✂️ Existing block pruning methods such as BK-SDM [3] and KOALA [4] focus on UNet-based backbones, which do not preserve image quality well when applied to MMDiT-based DMs.

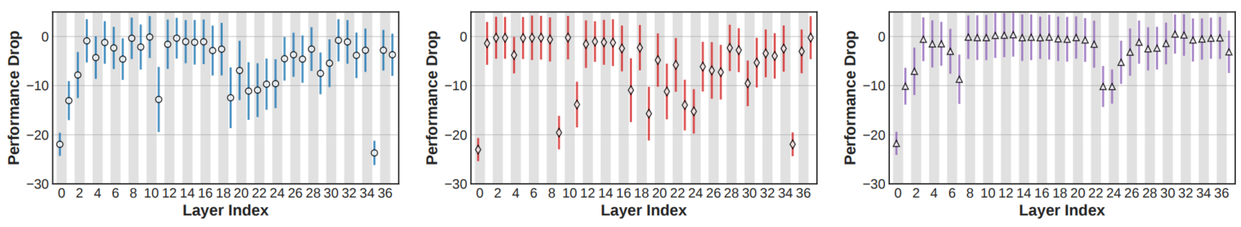
🚀 Based on key observations of the hierarchical nature of MMDiT, we propose a compression method for SOTA DMs that significantly reduces memory footprint and inference latency.

The Two-fold Hierarchy

🔬 Inter-block Hierarchy: Early blocks establish semantic structure. Later blocks handle detailed refinements.



🔬 Intra-block Hierarchy: Not all subcomponents (Attention, MLP) are equal. Their importance varies by position.



(a) Block Removal (b) Multi-modal Attention (c) MLP

HierarchicalPrune

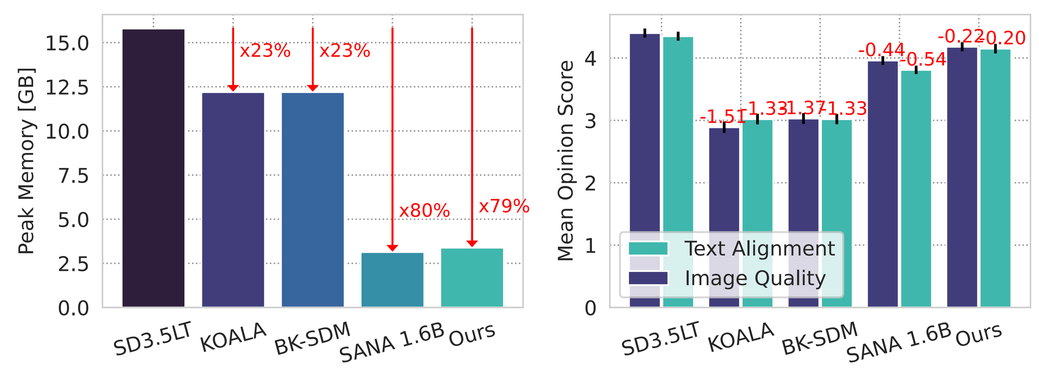
I. Hierarchical Position Pruning (HPP), which identifies and removes less essential later blocks based on position hierarchy.
II. Positional Weight Preservation (PWP), which systematically protects early model portions that are essential for semantic structural integrity.
III. Sensitivity-Guided Distillation (SGDistill), which adjusts knowledge-transfer intensity based on our discovery of block-wise sensitivity variations.

Evaluation

- Baselines: KOALA [3] employs input-output cosine similarity, and BK-SDM [4] uses impact on CLIP Score during block removal as importance scores. Both works target UNet architecture.
- Platforms used for latency & memory measurement: NVIDIA A100 (80GB), A6000 (48GB), and GTX 3090 24GB.

> *"A picture is worth a thousand words."*

> Peak memory reduction (Left) and Mean Opinion Score (Right)



> Image quality scores and the corresponding peak memory measurements in GB and remaining ratios (%) after compression.

Model	Method	Memory (%)	GenEval ⇑	HPSv2 ⇑	Reduction ⇓
Linear DiT	SANA-Sprint	3.14 GB (100%)	0.77	29.61	-
SD3.5 Large Turbo	Original	15.8 GB (100%)	0.71	30.29	-
	KOALA	12.6 GB (79.4%)	0.37	19.99	41.2%
	KOALA (+Quant)	3.56 GB (22.5%)	0.33	18.44	46.4%
	BK-SDM	12.6 GB (79.4%)	0.38	21.21	38.2%
	BK-SDM (+Quant)	3.56 GB (22.5%)	0.34	19.83	43.3%
	Ours (HPP+PWP+Q)	3.56 GB (22.5%)	**0.69**	**28.15**	**4.8%**
	Ours (All)	**3.24 GB (20.5%)**	0.62	26.29	13.3%
FLUX.1 Schnell	Original	22.6 GB (100%)	0.66	29.71	-
	KOALA	15.9 GB (70.5%)	0.38	25.24	28.7%
	BK-SDM	15.9 GB (70.5%)	0.45	27.38	19.8%
	Ours (All)	**4.44 GB (19.6%)**	**0.64**	**28.69**	**3.2%**

> Ablation study of each component and quantisation.

Prunning Ratio	Method	Remaining Memory (%)	GenEval ⇑	HPSv2 ⇑	Reduction ⇓
None (0%)	Original	100%	0.71	30.29	-
Moderate Pruning (20%)	Ours (HPP)	79.4%	0.03	11.08	79.4%
	Ours (+PWP)	79.4%	**0.71**	**28.97**	**2.5%**
	Ours (+Quant)	**22.5%**	0.69	28.15	4.8%
Aggressive Pruning (30%)	Ours (HPP)	71.5%	0.0	7.00	88.4%
	Ours (+PWP)	71.5%	0.46	21.74	31.9%
	Ours (+SGDistill)	71.5%	**0.64**	**27.29**	**10.1%**
	Ours (+Quant)	**20.5%**	0.62	26.29	13.3%

☑ **HierarchicalPrune** achieves reduction rates of **79.5-80.4% in memory & 27.9-38.0% in latency (see paper)** compared to original models.

☑ Only experiences minimum drop of image quality (4.8-5.3%) according to user study and General and HPSv2 benchmarking.

Conclusions

💻 We introduced HierarchicalPrune, a position-aware and fine-grained compression method for billion-parameter scale DMs to friendly to consumer-grade-GPU inferences.

🏋️ HierarchicalPrune delivers significant memory & latency reduction, with minimum drop of image quality.

References
[1] Esser, P. et al. Scaling Rectified Flow Transformers for HighResolution Image Synthesis. In International Conference on Machine Learning (ICML'24).
[2] Black Forest Labs. Flux.1 Model Family. https:// blackforestlabs.ai/announcing-black-forest-labs/
[3] Kim, et al, BK-SDM: A Lightweight, Fast, and Cheap Version of Stable Diffusion. In European Conference on Computer Vision (ECCV'24).
[4] Lee, Y. et al, KOALA: Empirical Lessons toward Memory-Efficient and Fast Diffusion Models for Text-to-Image Synthesis. Advances in Neural Information Processing Systems (NeurIPS'24).