Association for the Advancement of Artificial Intelligence

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

State-of-the-art Text-to-Image Diffusion Models

Stable Diffusion 3.5
2024 October

Seedream 2.0
2025 April

FLUX 1.0
2024 August

Qwen-Image
2025 August

The Challenge: Scale v.s. Efficiency

8B

Stable Diffusion 3.5
2024 October

Seedream 2.0
2025 April

20B

11B

FLUX 1.0
2024 August

Qwen-Image
2025 August

20B

Common backbone architecture [1][2]: Multi-Modal Diffusion Transformers (MMDiT)

- Significantly outperforms previous generation of models such as SDXL and SD1.5 and smaller models trained from scratch e.g. SANA [3] in real-world evaluations [4];
- **Massive parameter-count;**
- **High Cost:** Requires high-end GPUs (e.g., A100) hence impossible for standard edge deployment.

The Challenge: Scale v.s. Efficiency

Common backbone architecture [1][2]: Multi-Modal Diffusion Transformers (MMDiT)

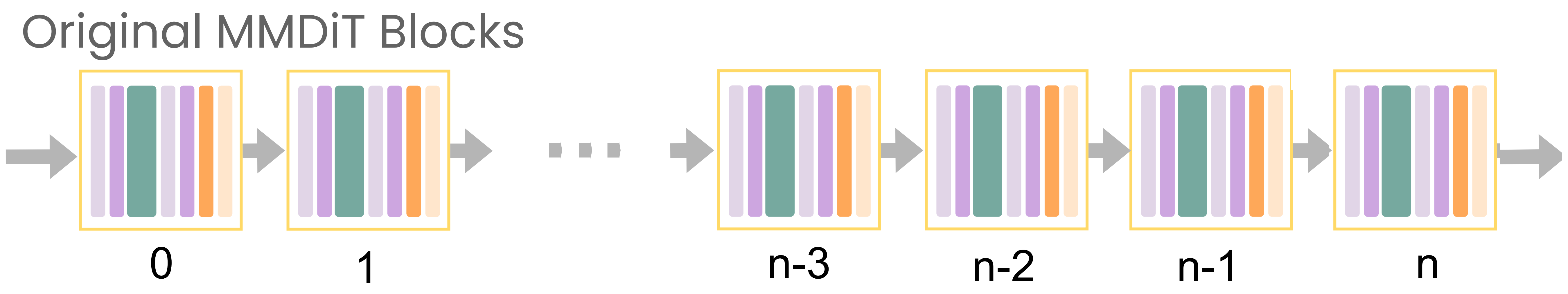
- Significantly outperforms previous generation of models such as SDXL and SD1.5 and smaller models trained from scratch e.g. SANA [3] in real-world evaluations [4];
- **Massive parameter-count;**
- **High Cost:** Requires high-end GPUs (e.g., A100) hence impossible for standard edge deployment.

Previous works [5][6] proved depth-pruning effective for DM compression.

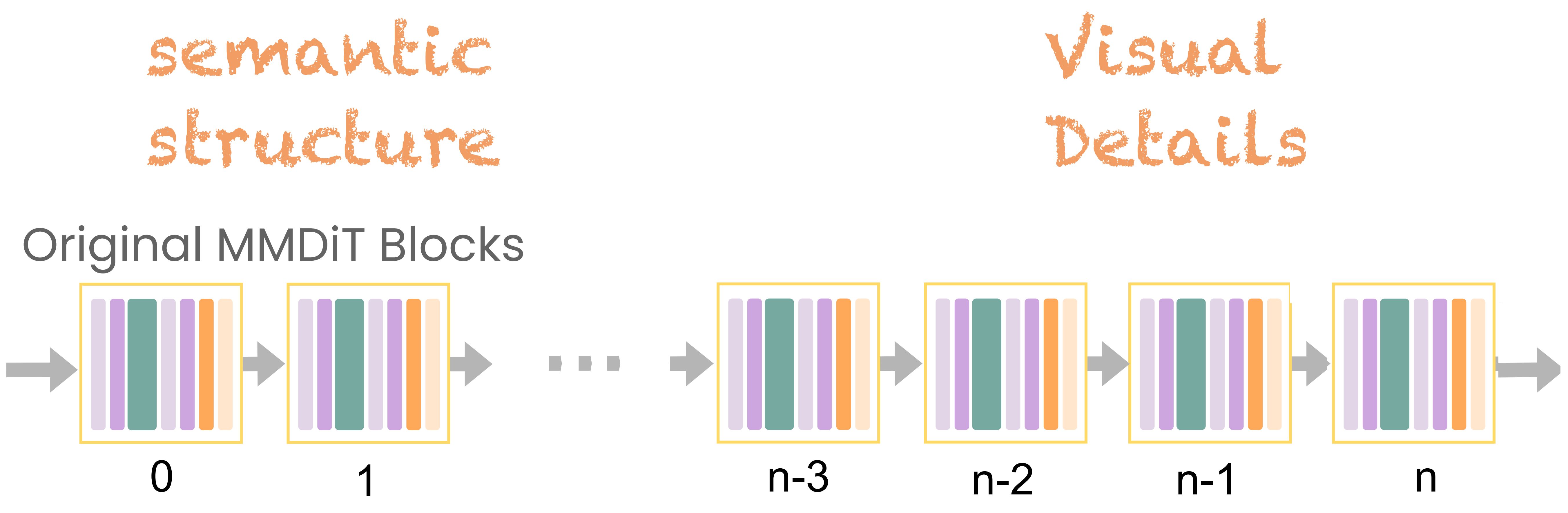
- Targeted UNet-based architecture, as used in Stable Diffusion 1.5, SDXL.
- Fail to generalise to large-scale MMDiTs**:** Significant degradation at >20% compression.

[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] Chen, J. et al. 2025b. SANA-Sprint: One-Step Diffusion with Continuous-Time Consistency Distillation. arXiv:2503.09641.
[4] Artificial Analysis Leaderboard https://artificialanalysis.ai/image/leaderboard/text-to-image
[5] Kim, et al, BK-SDM: A Lightweight, Fast, and Cheap Version of Stable Diffusion. In European Conference on Computer Vision (ECCV'24).
[6] 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).

Hypothesis: The Two-fold Hierarchy

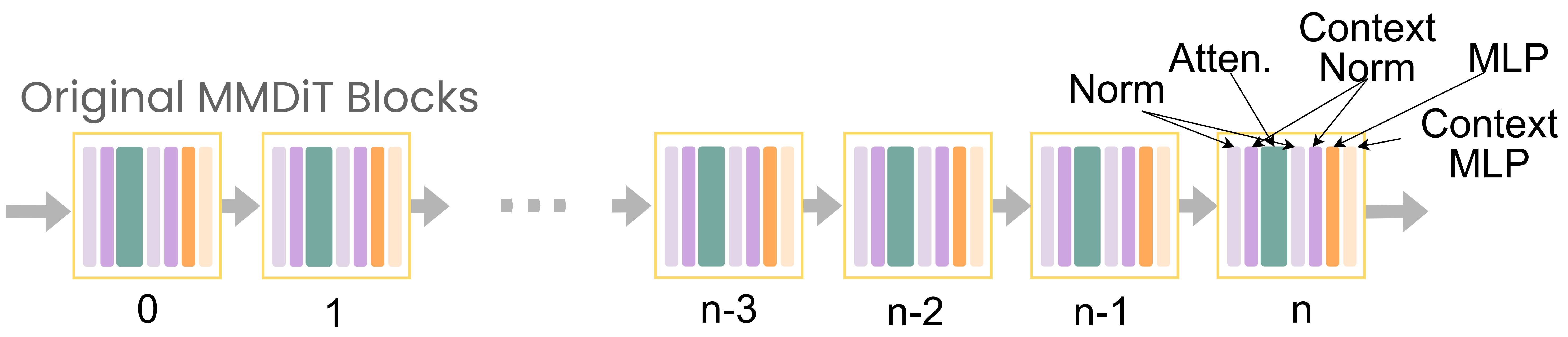


Hypothesis: The Two-fold Hierarchy



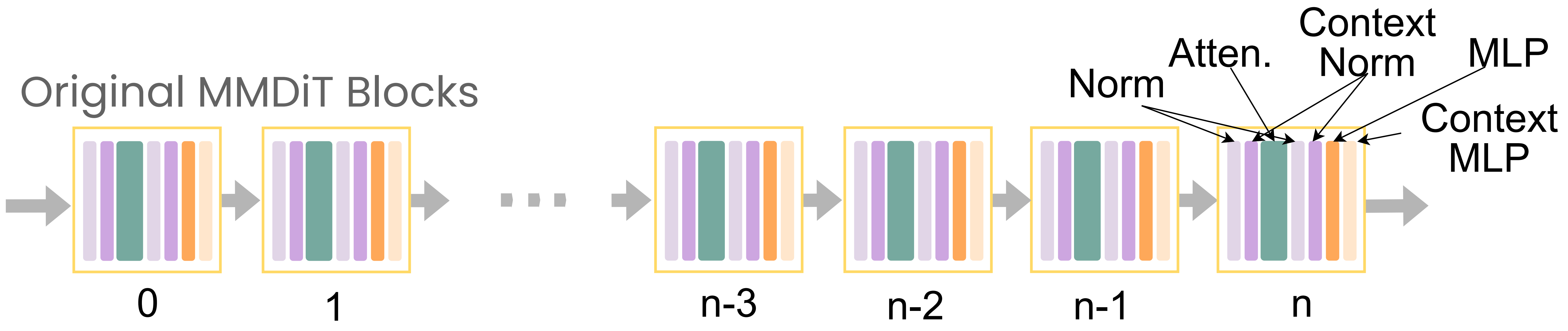
- **Inter-block Hierarchy:** Early blocks establish semantic structure. Later blocks handle detailed refinements.

Hypothesis: 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.

Hypothesis: The Two-fold Hierarchy



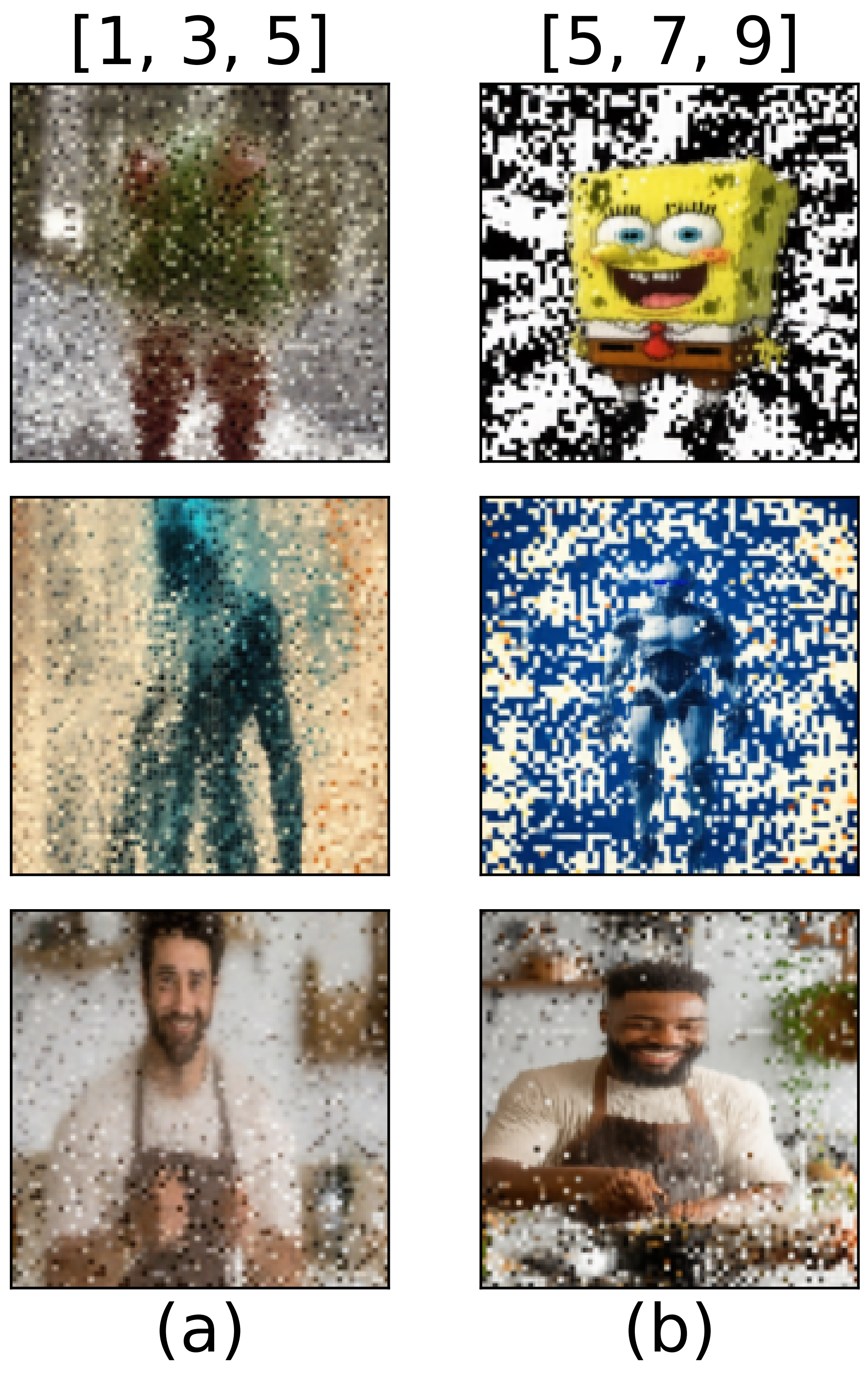
To verify this:

- Data: Samples from HPSv2
- Model: SD3.5 Large Turbo
- Removing 3 non-consecutive blocks at different locations

Core Insights: The Two-fold Hierarchy

🔬 Inter-block Hierarchy

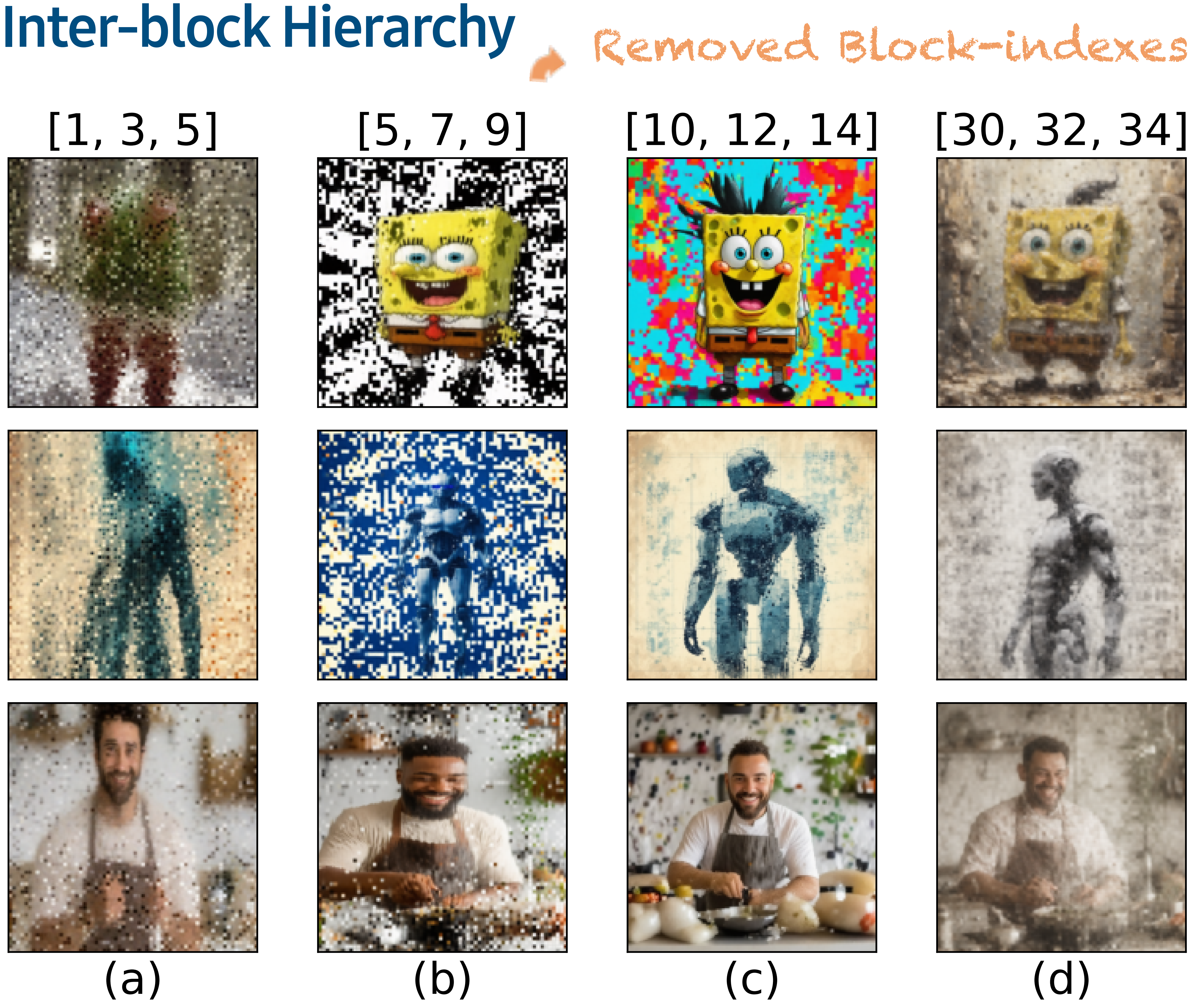
Removed Block-indexes



Early blocks establish semantic structure.
Later blocks handle detailed refinements.

Core Insights: The Two-fold Hierarchy

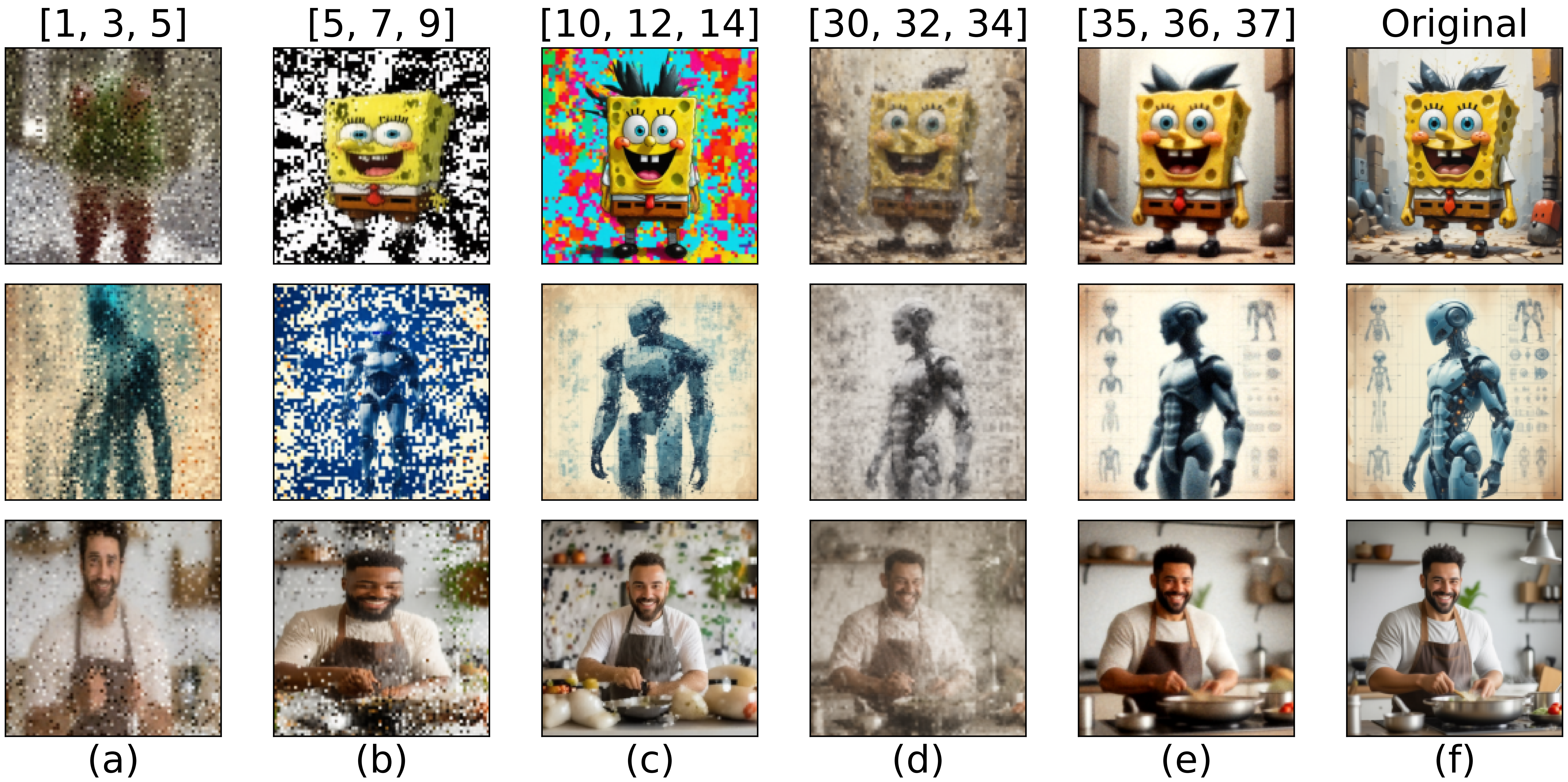
🔬 Inter-block Hierarchy



Early blocks establish semantic structure.
Later blocks handle detailed refinements.

Core Insights: The Two-fold Hierarchy

🔬 Inter-block Hierarchy

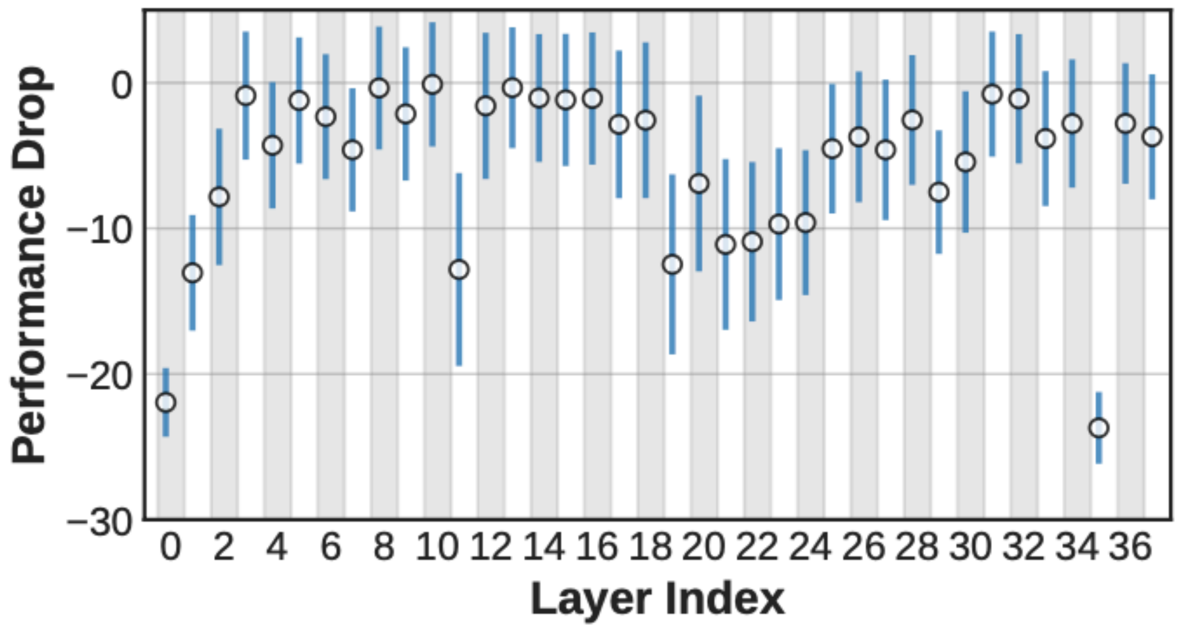


Early blocks establish semantic structure.
Later blocks handle detailed refinements.

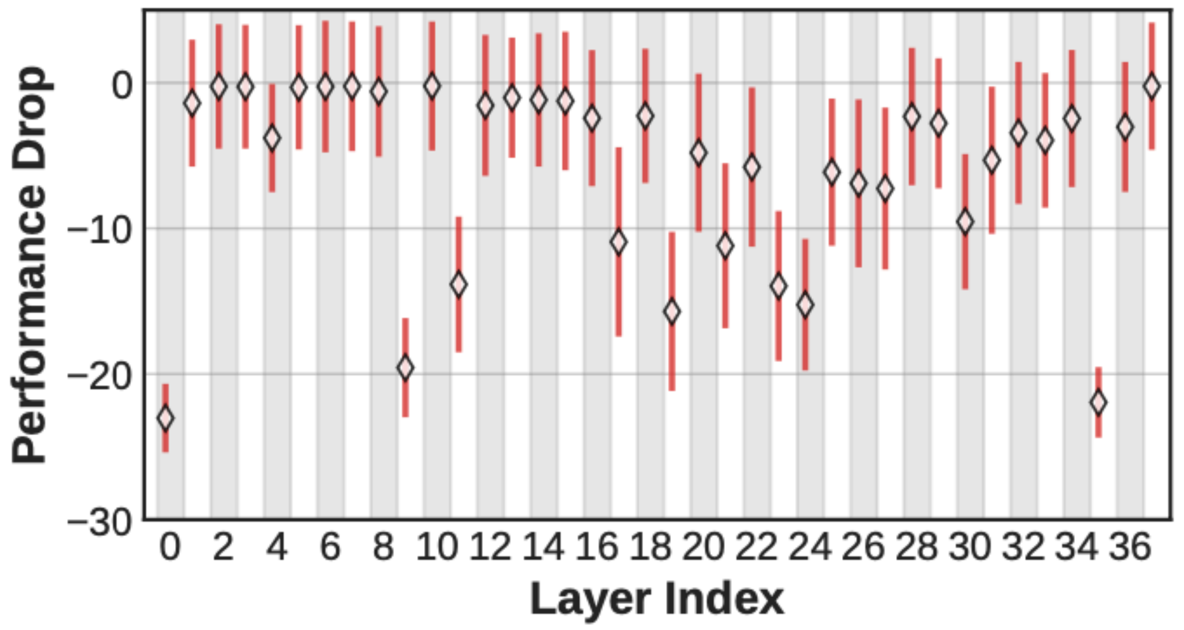
Core Insights: The Two-fold Hierarchy

🔬 Inter-block Hierarchy

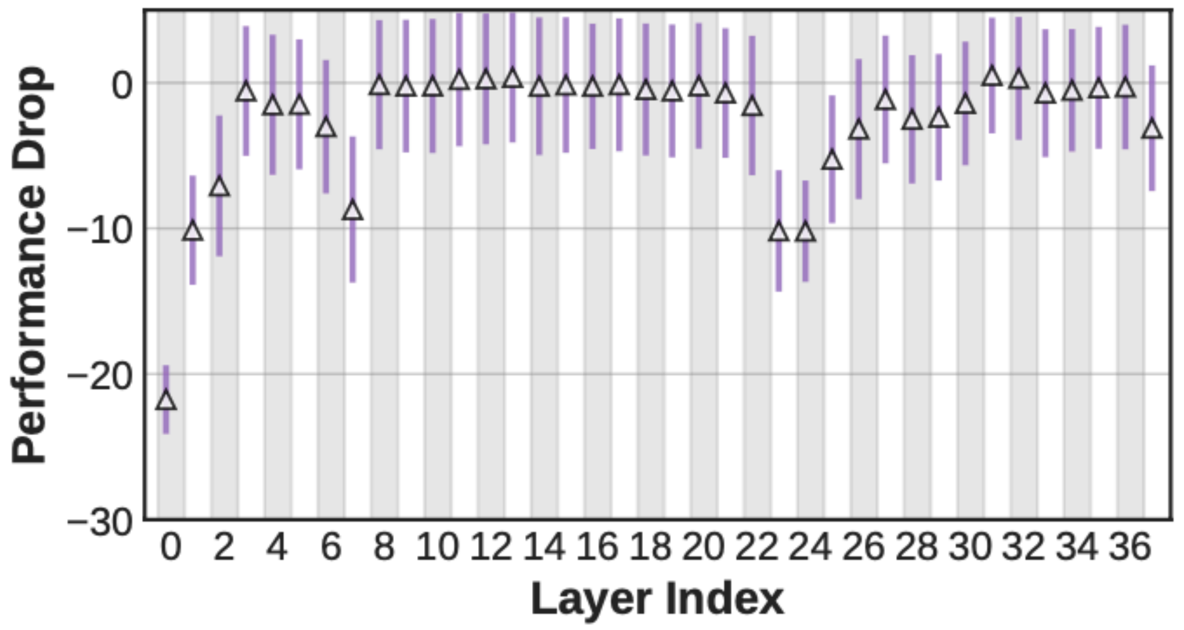
🔬 Intra-block Hierarchy



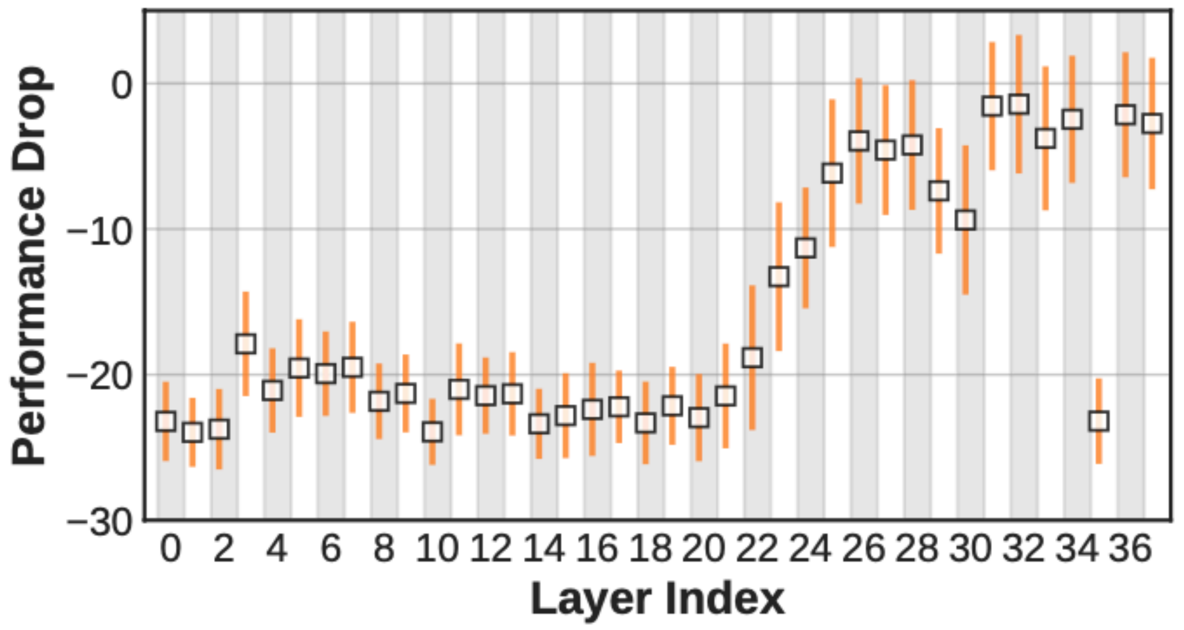
(a) Block Removal



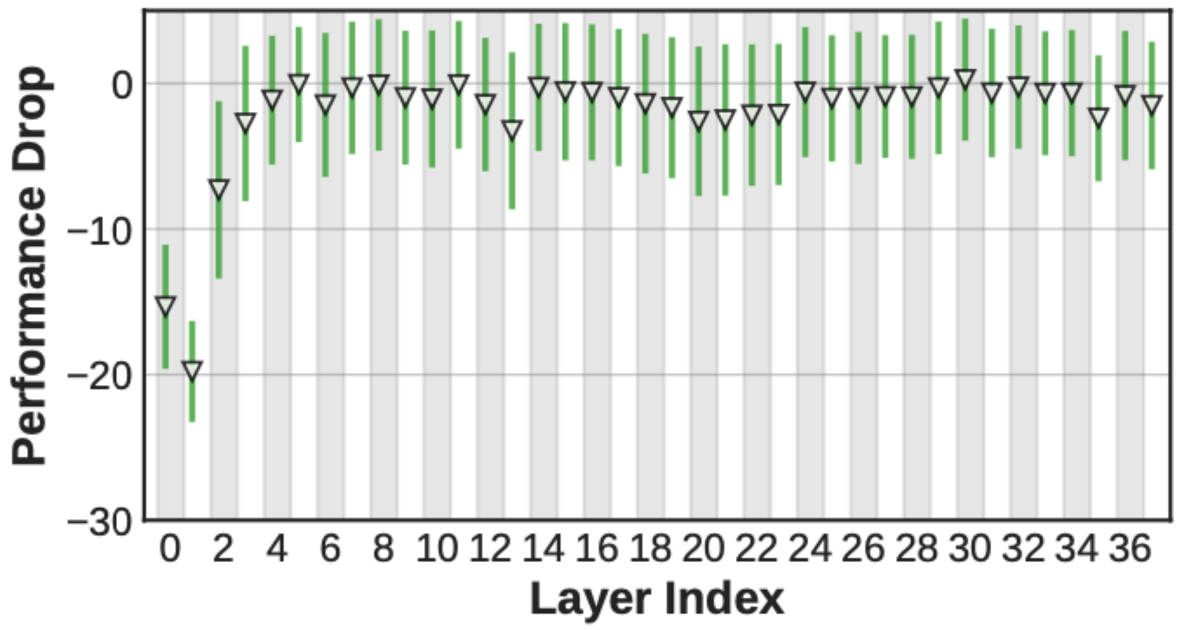
(b) Multi-modal Attention



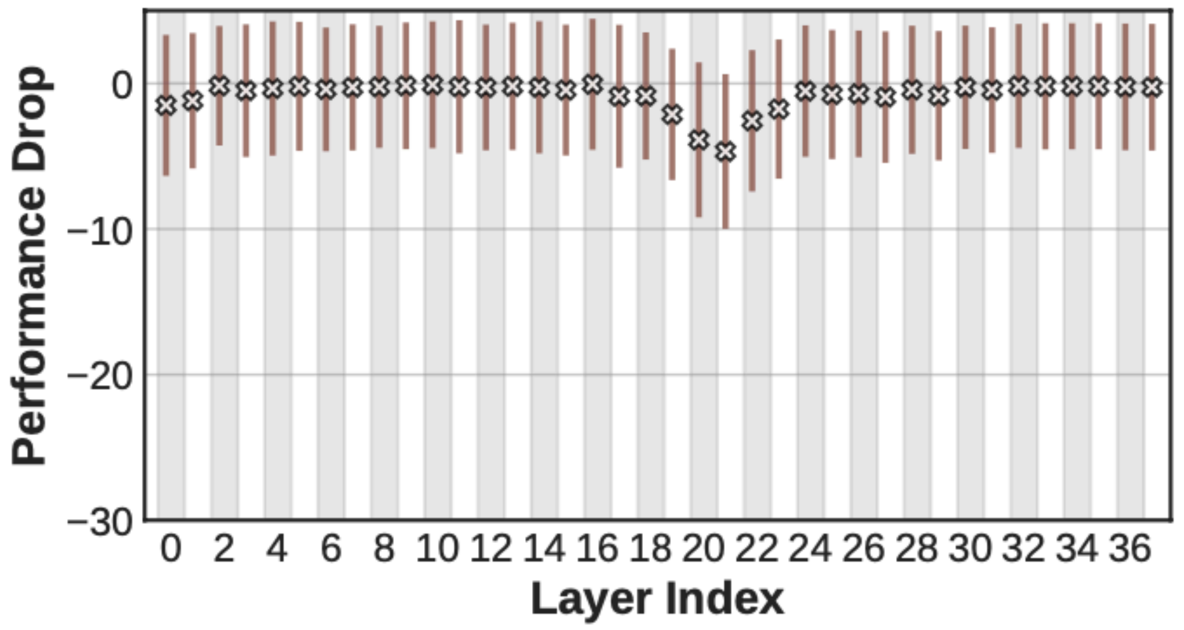
(c) MLP



(d) Norm



(e) Context Norm



(f) Context MLP

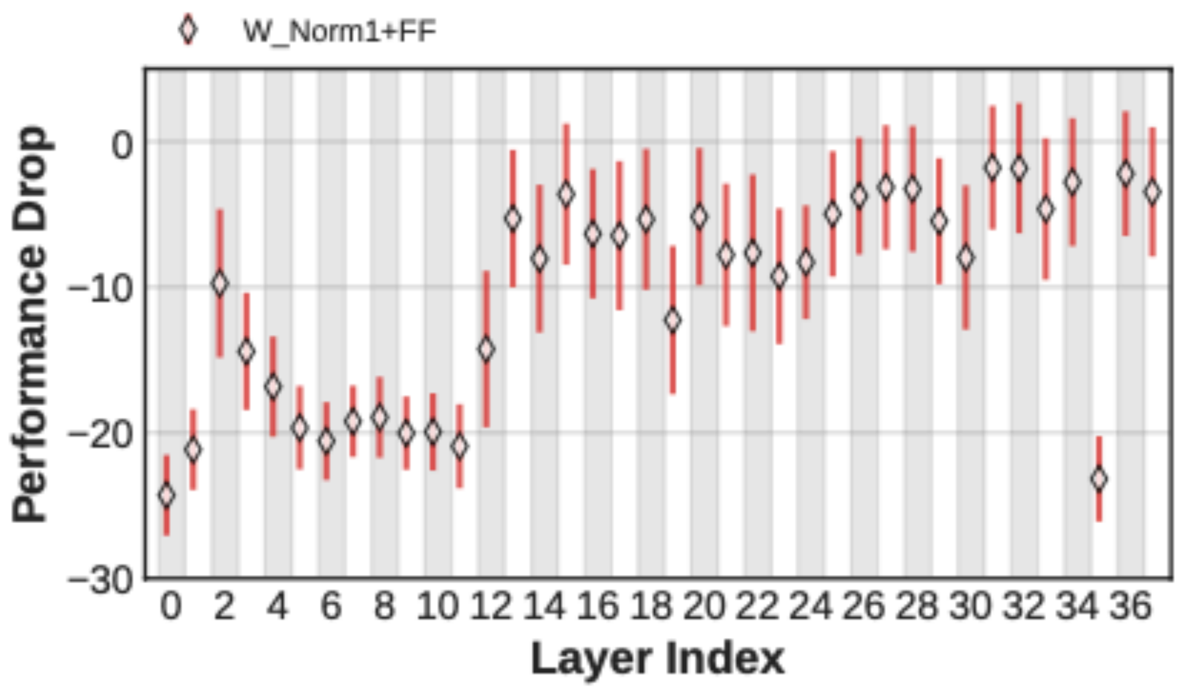
Early blocks establish semantic structure. Later blocks handle detailed refinements.

Not all subcomponents (Attention, MLP) are equal. Their importance varies by position.

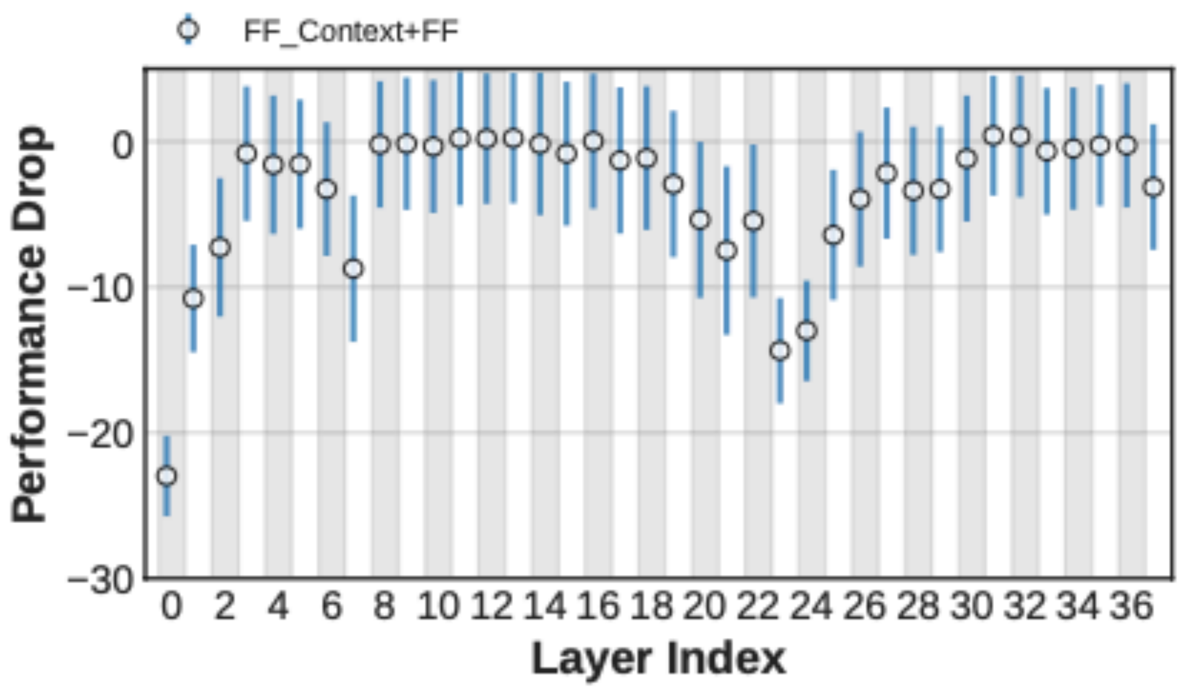
Core Insights: The Two-fold Hierarchy

🔬 Inter-block Hierarchy

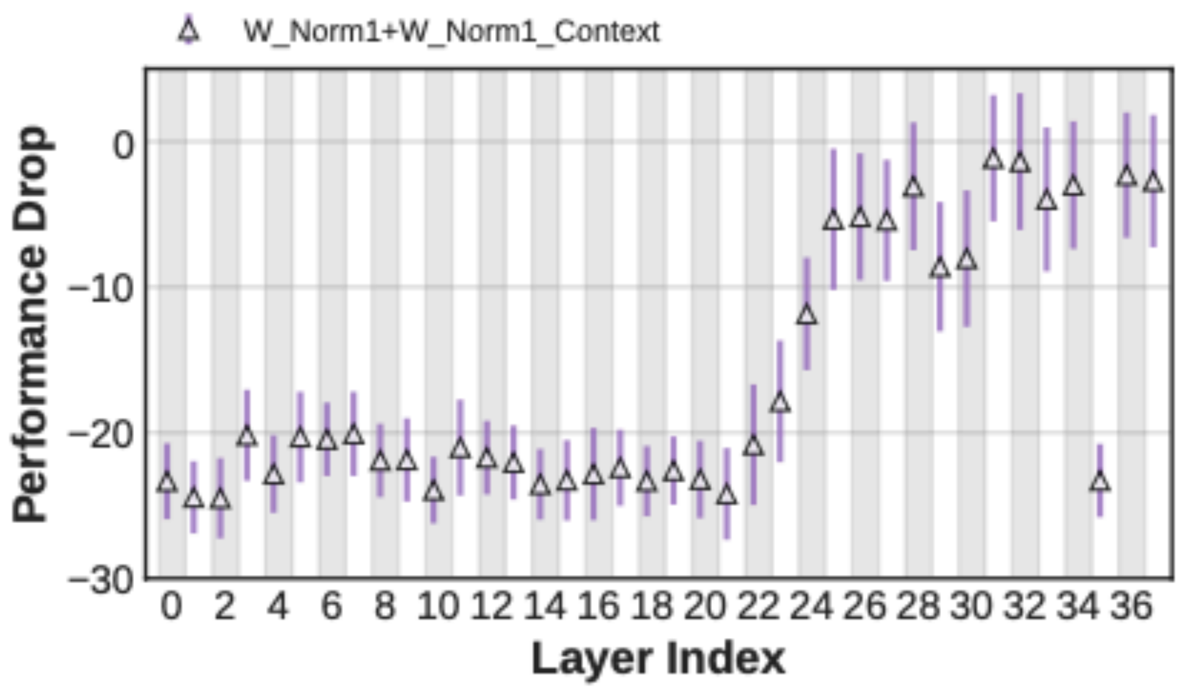
🔬 Intra-block Hierarchy



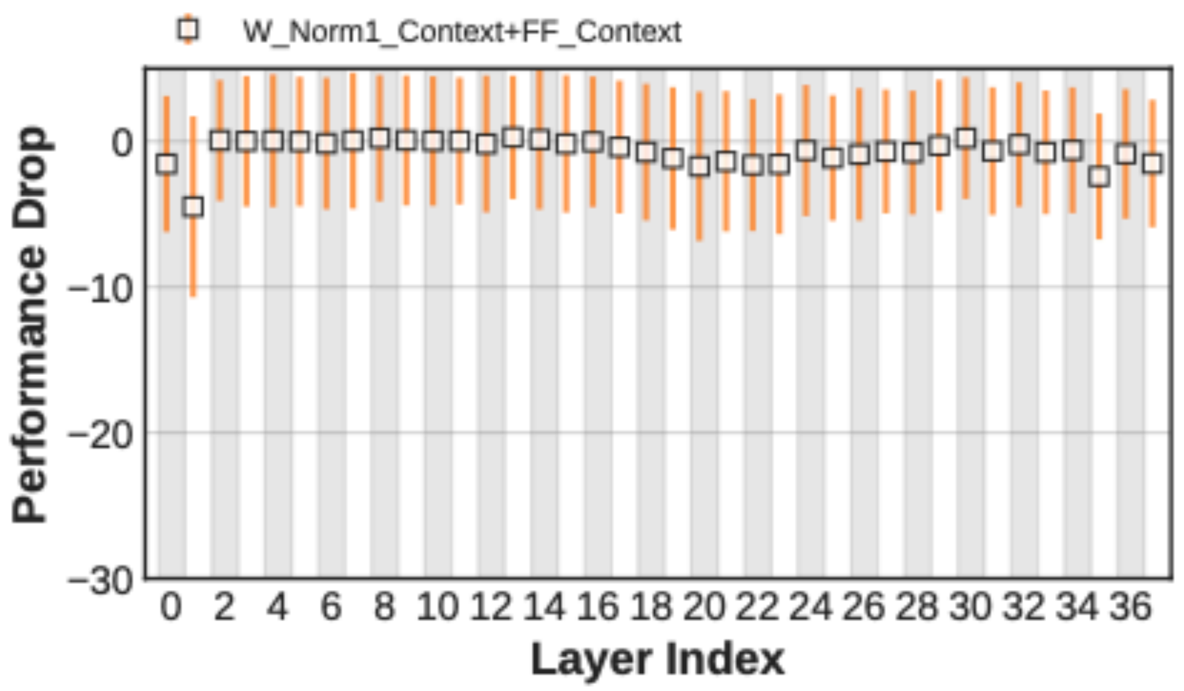
(a) Norm+MLP



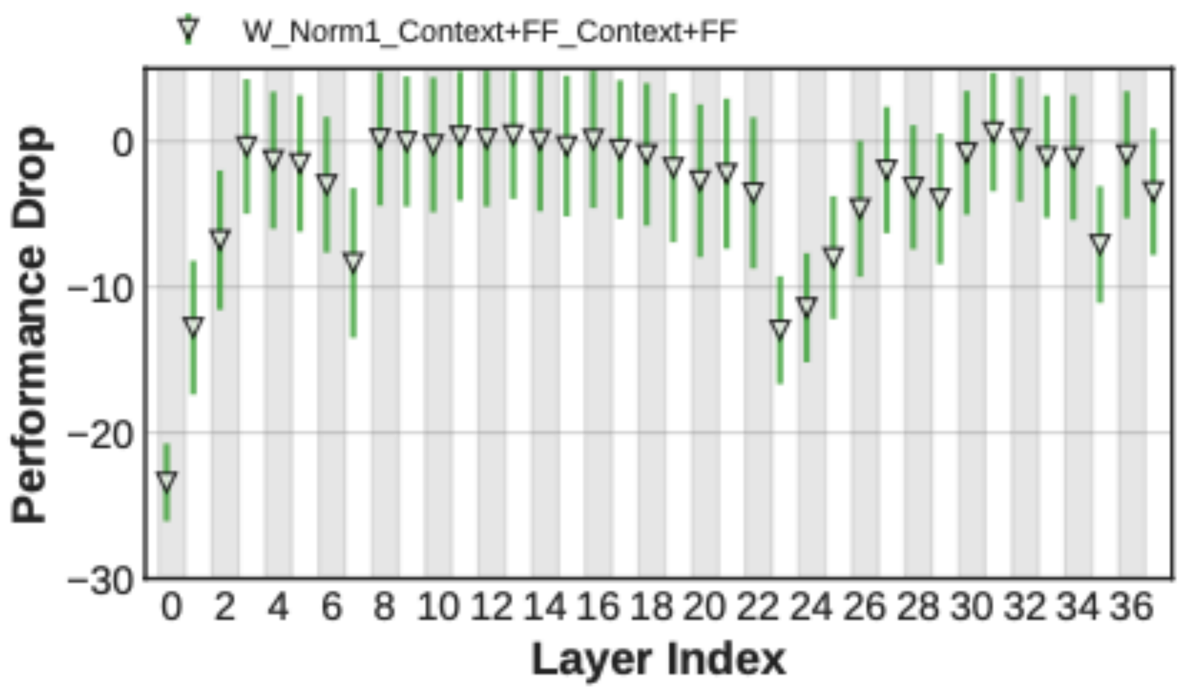
(b) MLP & Context MLP



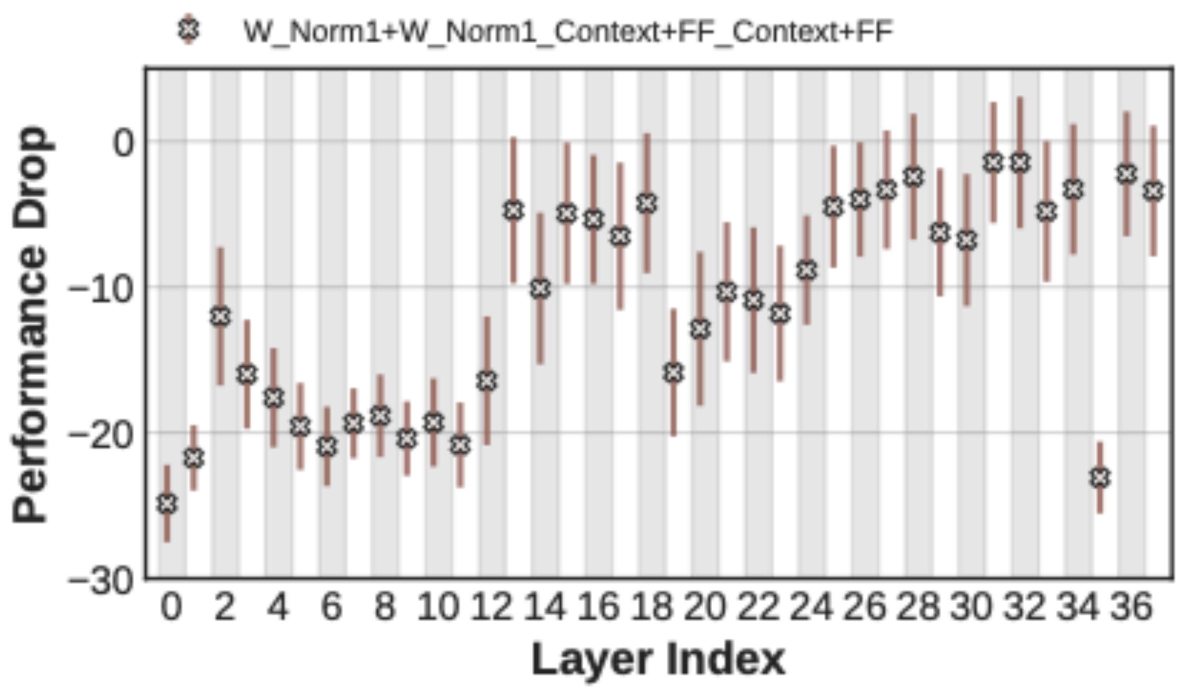
(c) Norm & Context Norm



(d) Context Norm & Context MLP



(e) Context Norm & All MLP



(f) All Norm & All MLP

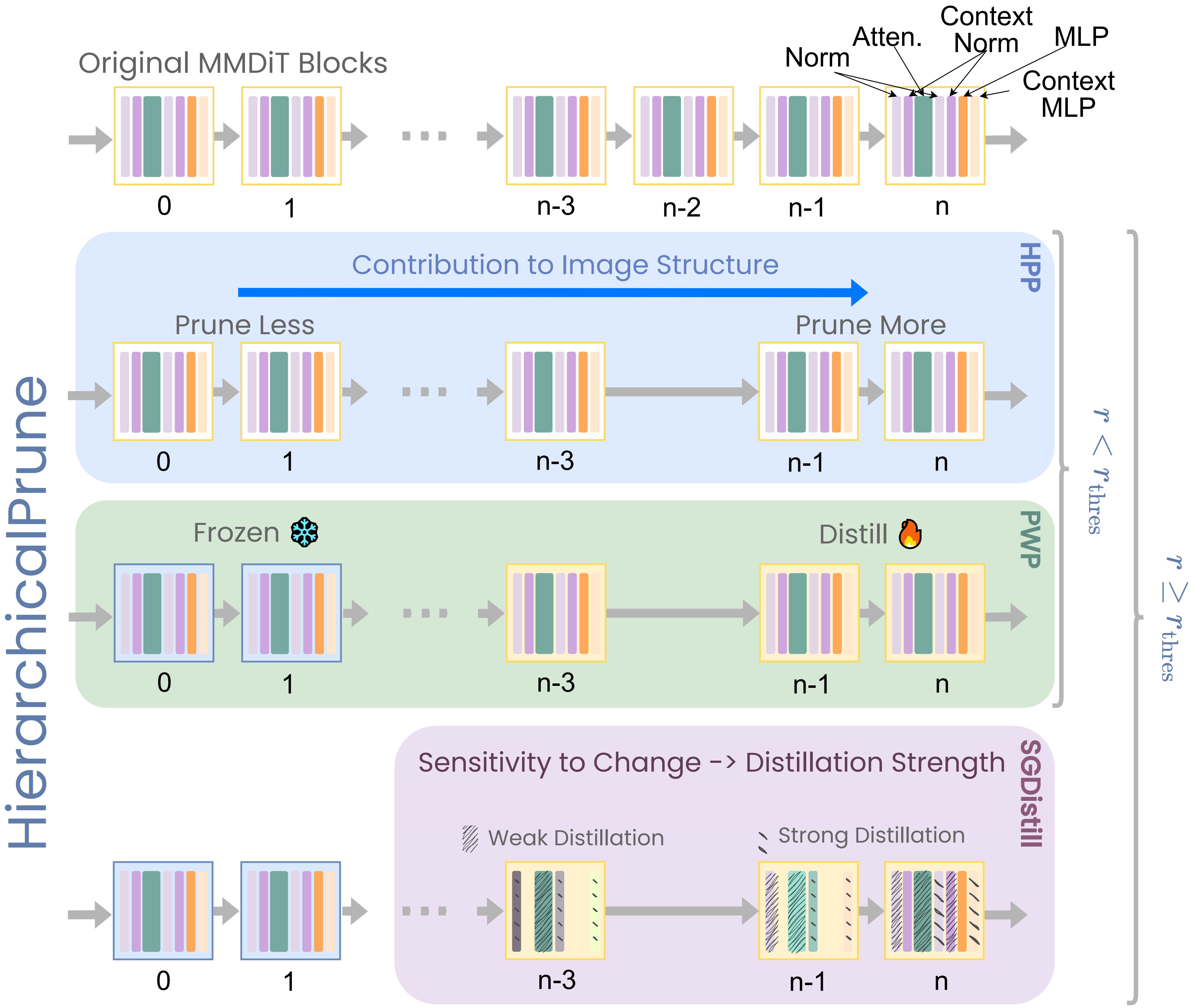
Early blocks establish semantic structure. Later blocks handle detailed refinements.

Not all subcomponents (Attention, MLP) are equal. Their importance varies by position.

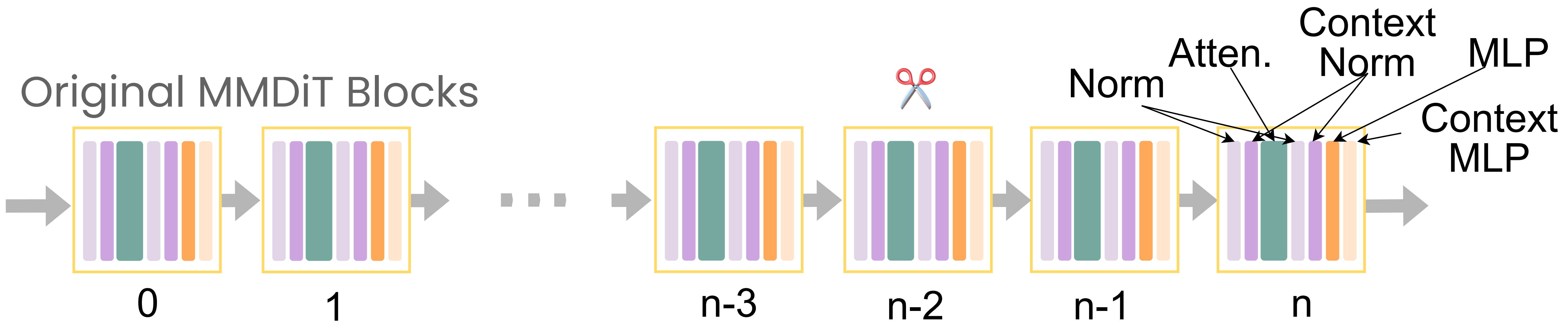
Core Insights: The Two-fold Hierarchy

🔬 Inter-block Hierarchy

🔬 Intra-block Hierarchy

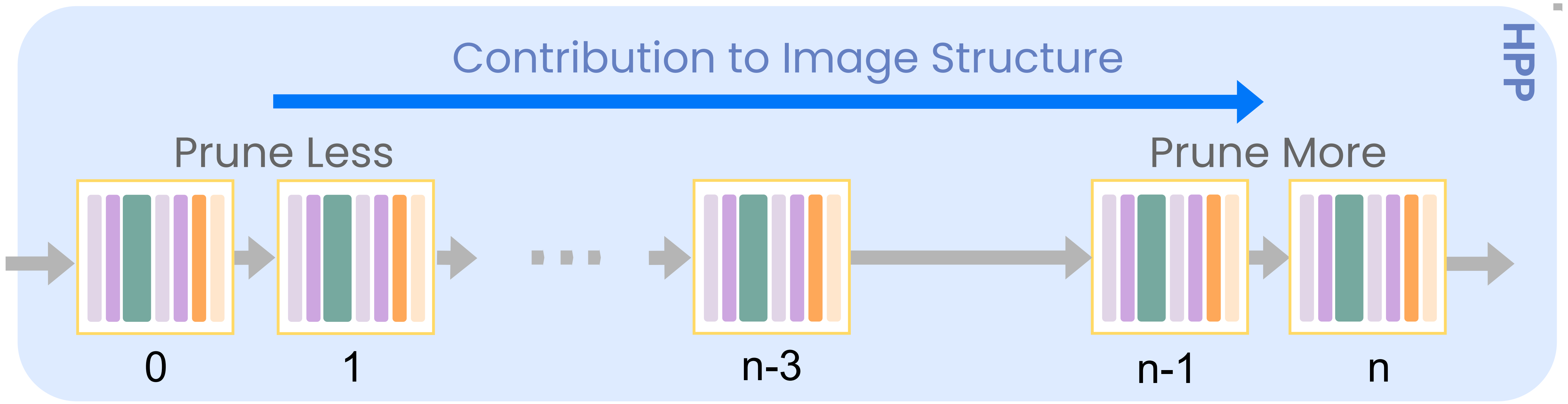


Hierarchical Prune



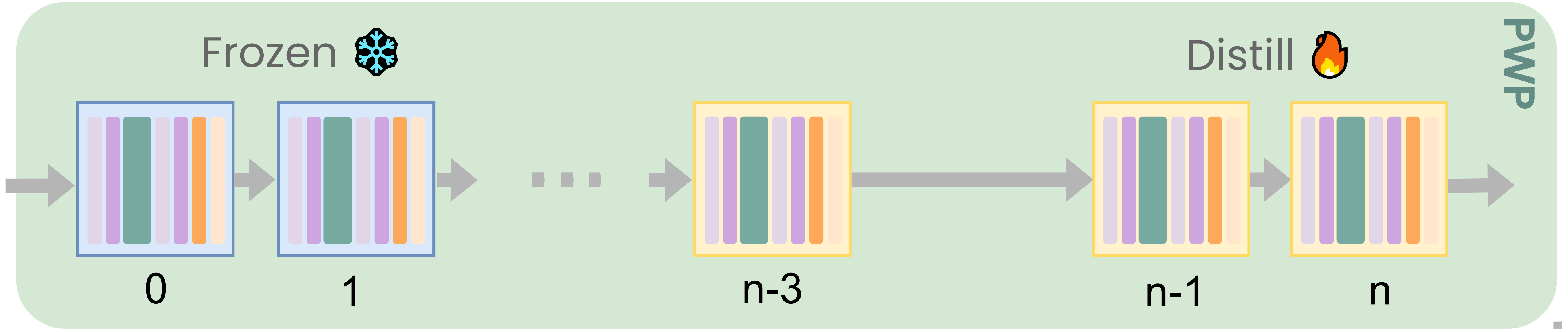
Hierarchical Prune

Hierarchical Position Pruning (HPP)



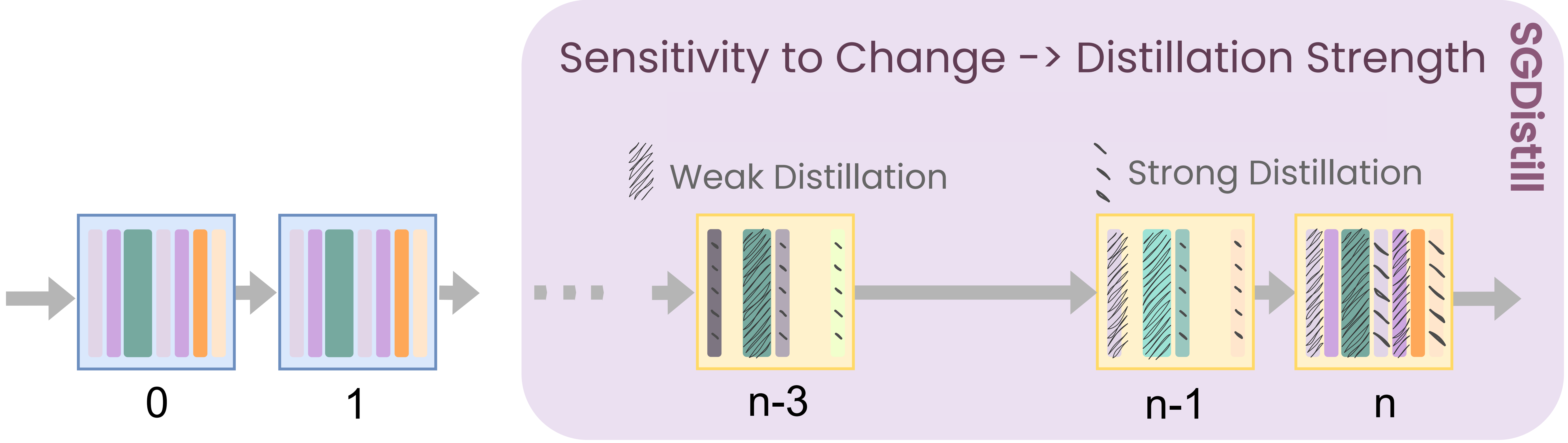
Hierarchical Prune

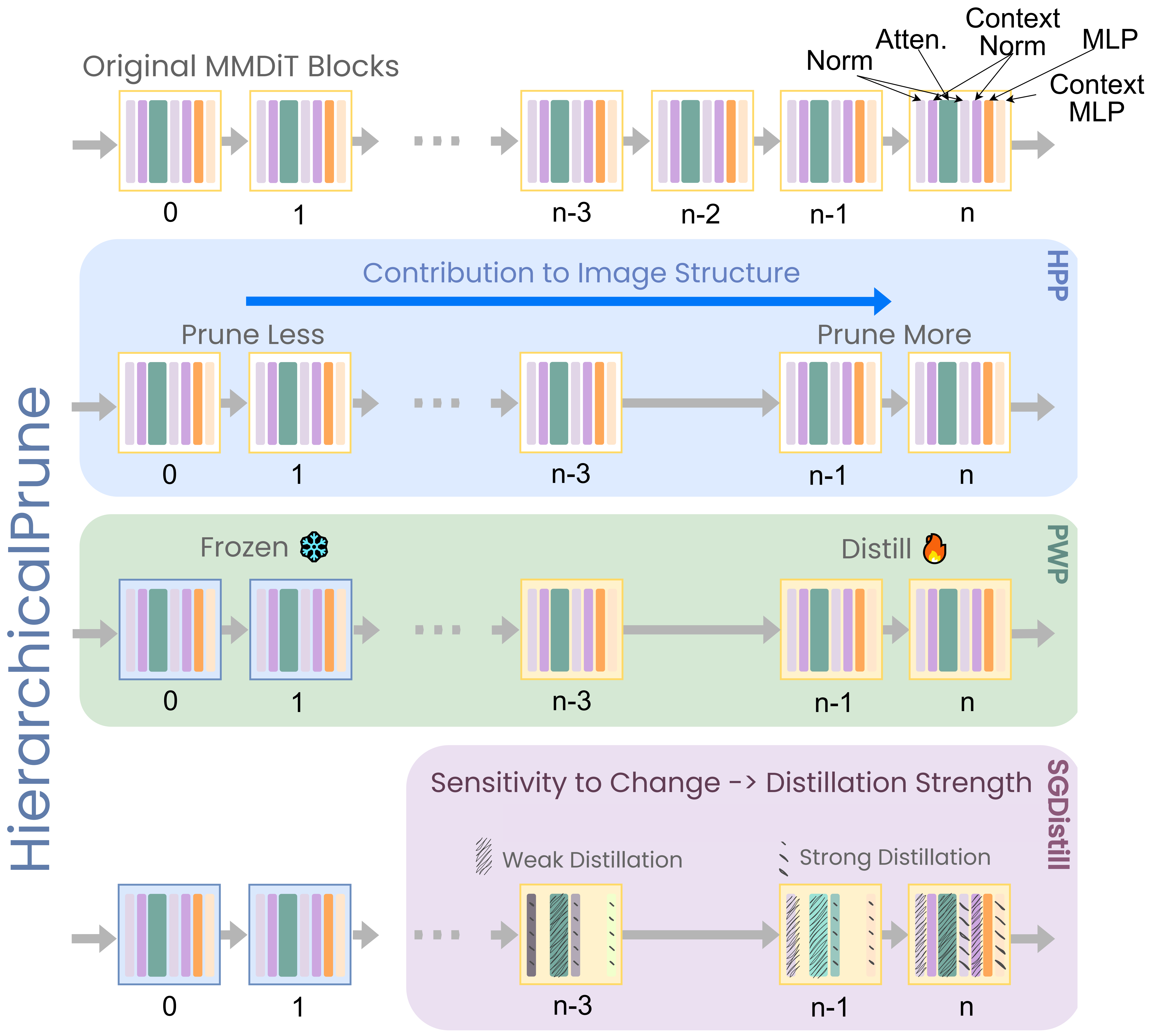
Positional Weight Preservation (PWP)



Hierarchical Prune

Sensitivity-Guided Distillation (SGDistill)





Original MMDiT Blocks
Norm
Atten.
Context Norm
MLP
Context MLP
0
1
n-3
n-2
n-1
n
HierarchicalPrune
HPP
Contribution to Image Structure
Prune Less
Prune More
0
1
n-3
n-1
n
PWP
Frozen
Distill
0
1
n-3
n-1
n
SGDistill
Sensitivity to Change -> Distillation Strength
Weak Distillation
Strong Distillation
0
1
n-3
n-1
n

Evaluation

- SD3.5 Large Turbo (8B) and FLUX.1-Schnell (12B),
- YE-POP dataset

Chen, J. et al. 2025b. SANA-Sprint: One-Step Diffusion with Continuous-Time Consistency Distillation. arXiv:2503.09641.
Kim, et al, BK-SDM: A Lightweight, Fast, and Cheap Version of Stable Diffusion. In European Conference on Computer Vision (ECCV'24).
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).

Evaluation

- SD3.5 Large Turbo (8B) and FLUX.1-Schnell (12B),
- YE-POP dataset
- Baselines:
  - I) BK-SDM (Kim et al. 2024a): proposed block pruning of U-Net-based models using the CLIP score+ distillation of the pruned model
  - ii) KOALA (Lee et al. 2024): each block’s input-output cosine similarity
  - SOTA small-scale DM, SANA (Chen et al. 2025b)

Chen, J. et al. 2025b. SANA-Sprint: One-Step Diffusion with Continuous-Time Consistency Distillation. arXiv:2503.09641.

Kim, et al, BK-SDM: A Lightweight, Fast, and Cheap Version of Stable Diffusion. In European Conference on Computer Vision (ECCV’24).

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).

Evaluation

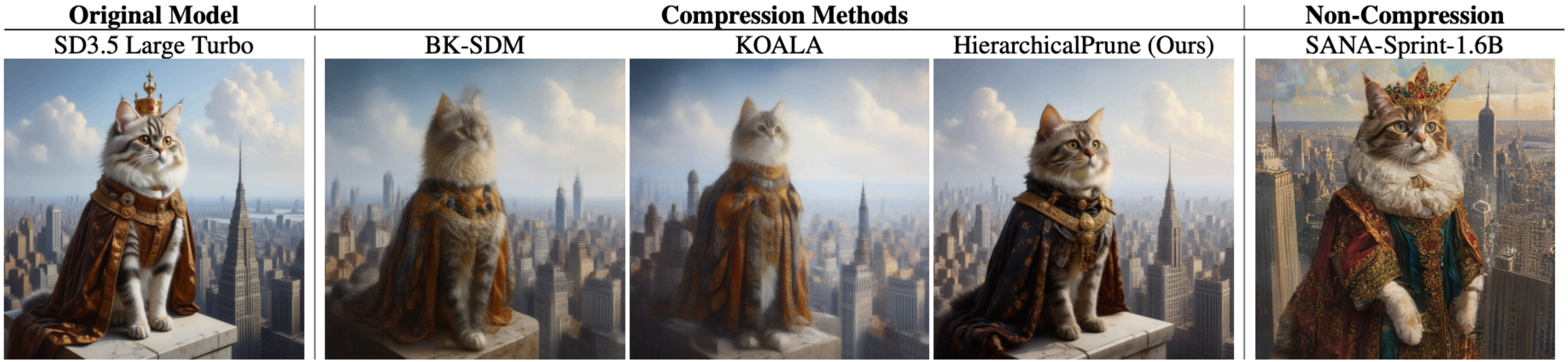
- SD3.5 Large Turbo (8B) and FLUX.1-Schnell (12B),
- YE-POP dataset
- Baselines:
  - I) BK-SDM (Kim et al. 2024a): proposed block pruning of U-Net-based models using the CLIP score+ distillation of the pruned model
  - ii) KOALA (Lee et al. 2024): each block's input-output cosine similarity
  - SOTA small-scale DM, SANA (Chen et al. 2025b)

Chen, J. et al. 2025b. SANA-Sprint: One-Step Diffusion with Continuous-Time Consistency Distillation. arXiv:2503.09641.

Kim, et al, BK-SDM: A Lightweight, Fast, and Cheap Version of Stable Diffusion. In European Conference on Computer Vision (ECCV'24).

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).

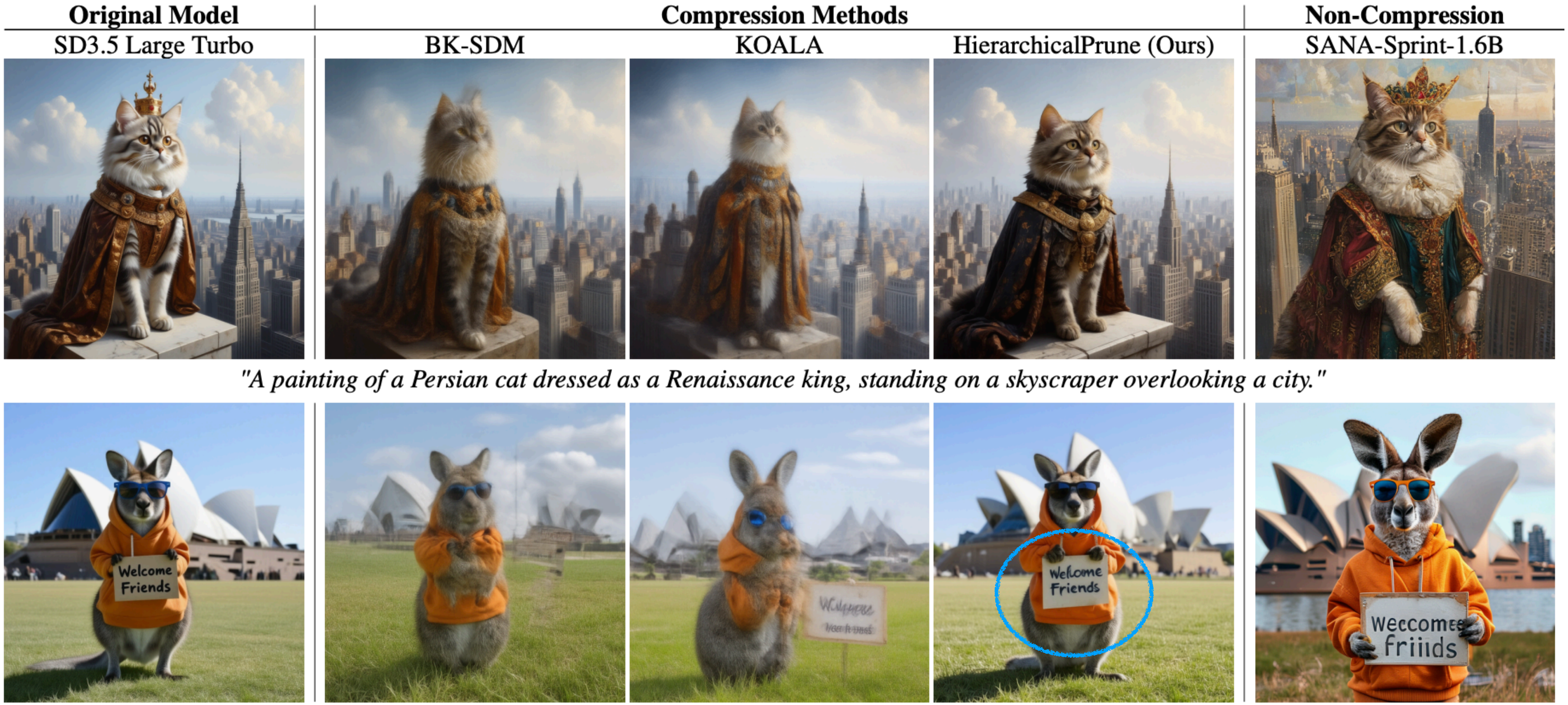
Evaluation



"A painting of a Persian cat dressed as a Renaissance king, standing on a skyscraper overlooking a city."

"A kangaroo in an orange hoodie and blue sunglasses stands on the grass in front of the Sydney Opera House holding a 'Welcome Friends' sign."

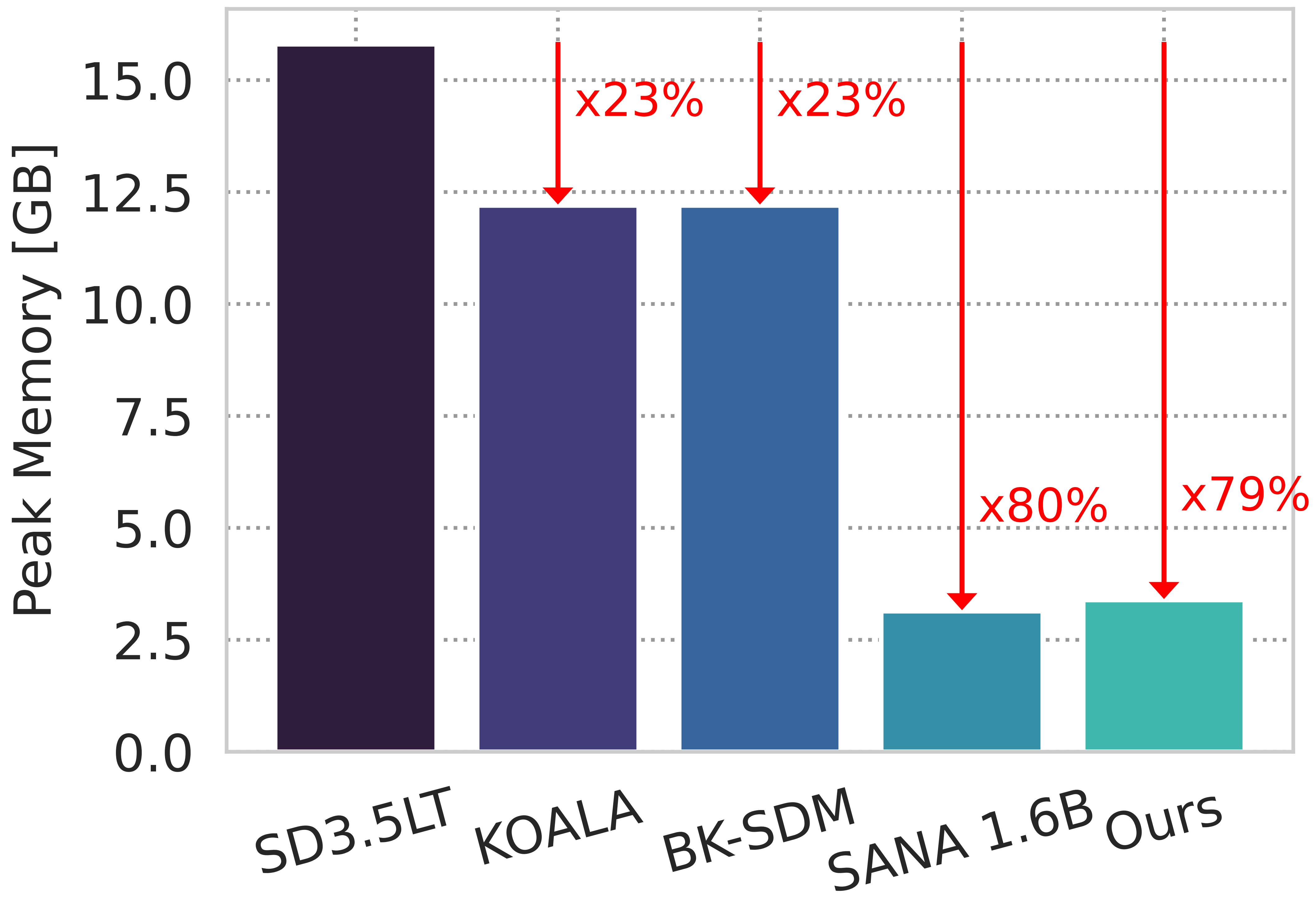
Evaluation



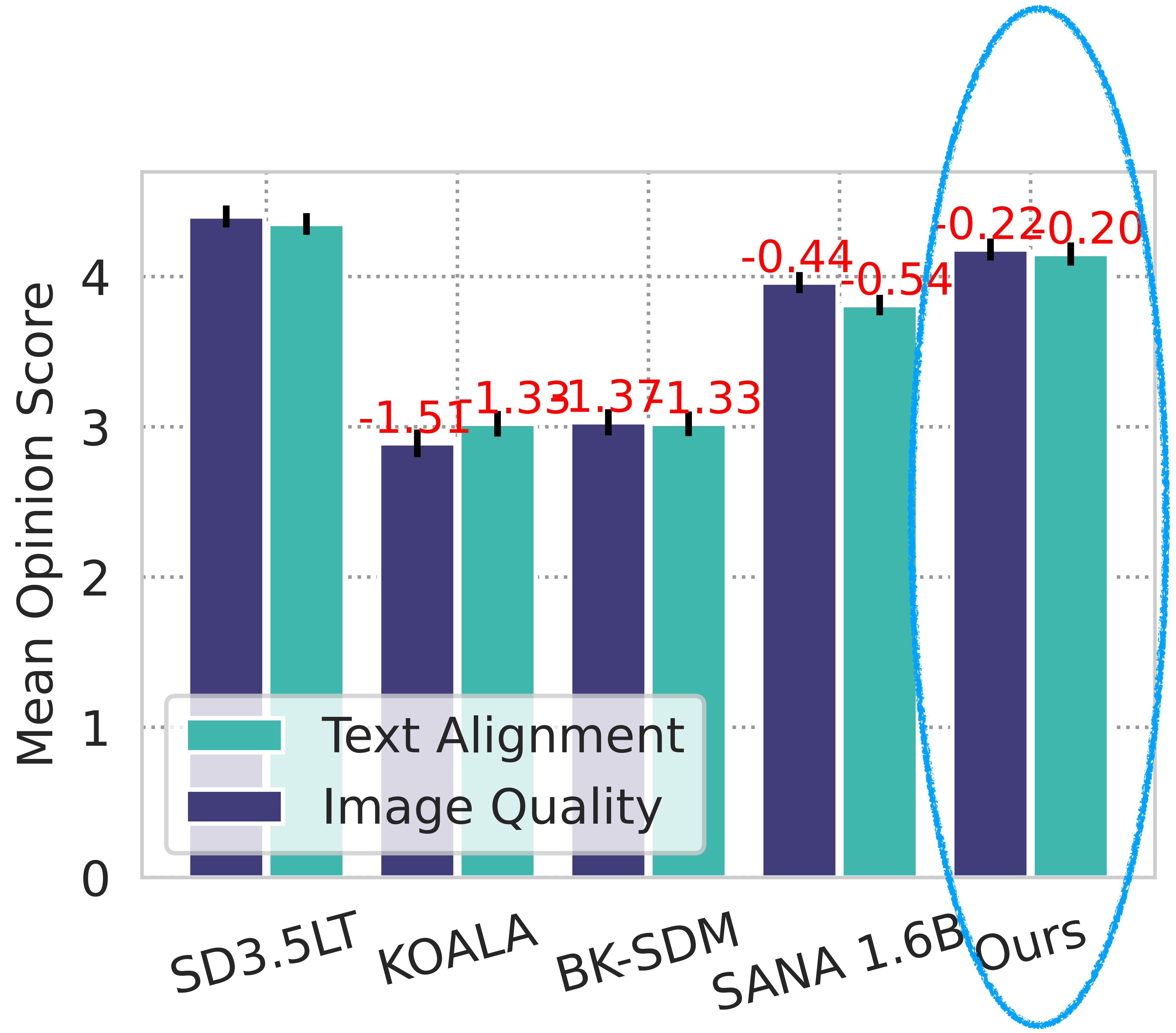
"A painting of a Persian cat dressed as a Renaissance king, standing on a skyscraper overlooking a city."

"A kangaroo in an orange hoodie and blue sunglasses stands on the grass in front of the Sydney Opera House holding a 'Welcome Friends' sign."

Measurements on Consumer Grade GPUs



User Study



Mean Opinion Score
0
1
2
3
4
-1.51
-1.33
-1.37
-1.33
-0.44
-0.54
-0.22
-0.20
Text Alignment
Image Quality
SD3.5LT
KOALA
BK-SDM
SANA 1.6B
Ours

Image Quality by GenEval and HPSv2

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%**

Image Quality by GenEval and HPSv2

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%**

Takeaway

🔬We identify a dual hierarchical structure in MMDiT-based DMs: an inter-block hierarchy and an intra-block hierarchy;

📝 We introduce HierarchicalPrune, establishing a comprehensive, position-aware pruning and distillation framework for large-scale DMs;

🏋️ Through extensive evaluation, we demonstrate that HierarchicalPrune is able to achieve significant memory reduction with minimal quality loss.

Thank you!

{yd.kwon, rui.li}@samsung.com