Exploring Diffusion Transformer Designs via Grafting

Keshigeyan Chandrasegaran (*), Michael Poli (*), Daniel Y. Fu, Dongjun Kim, Lea M. Hadzic, Manling Li,
Agrim Gupta, Stefano Massaroli, Azalia Mirhoseini, Juan Carlos Niebles, Stefano Ermon, Li Fei-Fei

Stanford University · Liquid · DeepMind · together.ai · UC San Diego · RIKEN · Northwestern University · salesforce

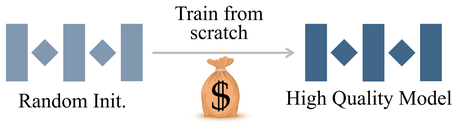
grafting.stanford.edu

Motivation

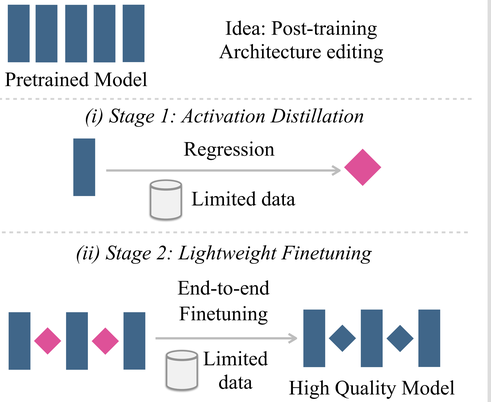
- **Model architecture design** plays a central role in machine learning.
- However, **studying new model architectures** is **expensive** due to pretraining costs.
- **Research Question:** Can we **materialize** new model architectures under **small compute budgets?**

Grafting Diffusion Transformers

Model Architecture Design via Grafting



Grafting Procedure



Grafting is a simple two-stage approach to model architecture editing

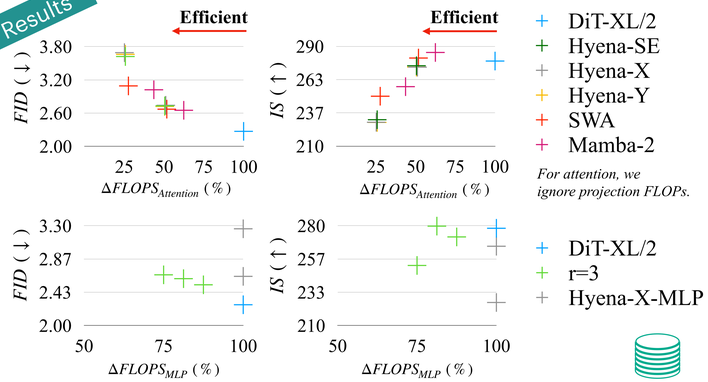
Efficient Hybrid Architectures for Class-conditional Image Generation

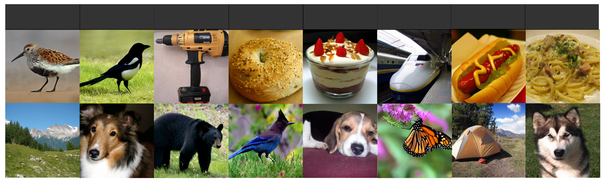
We construct efficient hybrid architectures with good generative quality under small compute budgets.

Testbed

Operator Type: *Which operator types are we replacing?*		
Multi-Head Attention (MHA) / Multi-Layer Perceptron (MLP)		
Efficient Alternative: *What do we replace it with?*		
MHA	Convolutions: Hyena-SE, Hyena-X/ Hyena-Y (Ours)	*K=4*, causal
	Local Attention: Sliding Window Attention (SWA)	*w=4*, bidirectional
	Linear Attention: Mamba-2	*ds=64, E=2*
MLP	Variable expansion ratio	r=3,6
	Hyena-X (Ours)	r=2, K=4, causal
Layer Selection and Replacement Ratio		
Details in the Paper (Sec. 4)		

Results



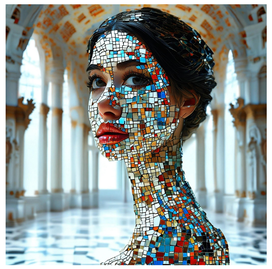


10%

<2%

<24 hours

Hybrid Architectures for Class-conditional Image Generation

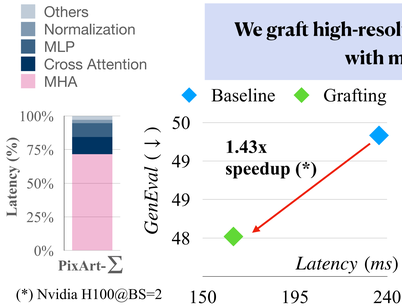


Faster Text-to-Image Generation

Restructure model architectures (Depth2Width)

Grafting Text-to-Image Diffusion Transformers

We graft high-resolution text-to-image DiTs, constructing hybrid architectures with meaningful speedups and minimal quality drop.



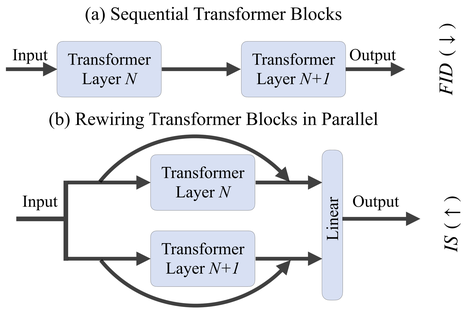
12k synthetic data

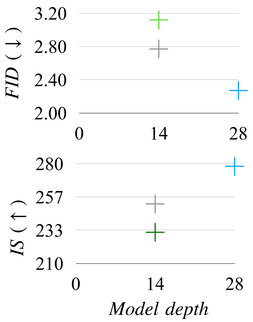
16k sequence length

GenEval	Ratio	Obj(1)	Obj(2)	Count	Colors	Pos	Color Attr.	Overall
Baseline	-	81.45	61.62	46.25	77.13	10.75	21.50	49.75
Grafting (Hyena-X)	50%	80.00	57.07	48.13	70.74	11.25	19.50	47.78

Converting Model Depth to Width via Grafting

Grafting enables architectural restructuring at the transformer block level, allowing model depth to be traded for width.





25%

<5%

<48 hours

References

Peebles, William, and Saining Xie. "Scalable diffusion models with transformers." *ICCV* 2023.
Chen, Junsong, et al. "Pixart-Sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation." *ECCV* 2024.
Poli, Michael, et al. "Hyena hierarchy: Towards larger convolutional language models." *ICML 2023.*
Dao, Tri, and Albert Gu. "Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality." *ICML 2024.*