NEURAL INFORMATION PROCESSING SYSTEMS

ANCRe: Adaptive Neural Connection Reassignment for Efficient Depth Scaling

Yilang Zhang[*1,2], Bingcong Li[*3], Niao He[3], Georgios B. Giannakis[1]

[1]Department of ECE, University of Minnesota, USA
[2]Machine Learning Research, Morgan Stanley, USA
[3]Department of CS, ETH Zürich, Switzerland

[*]Equal contribution

Paper

Code

Acknowledgements: NSF grants 2212318, 2220292, and 2312547; SNSF Starting Grant

Deep layers are underutilized

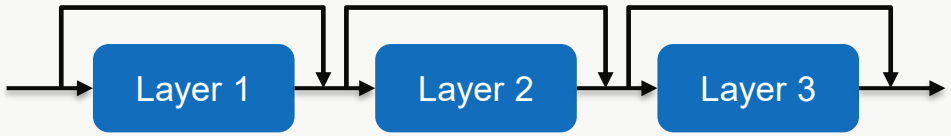
Foundation models are getting deeper

- Llama 3.1: 32, 80, and 126 layers for 8B, 70B, and 405B
- DiT: from 12 (DiT-S) to 28 (DiT-XL) layers

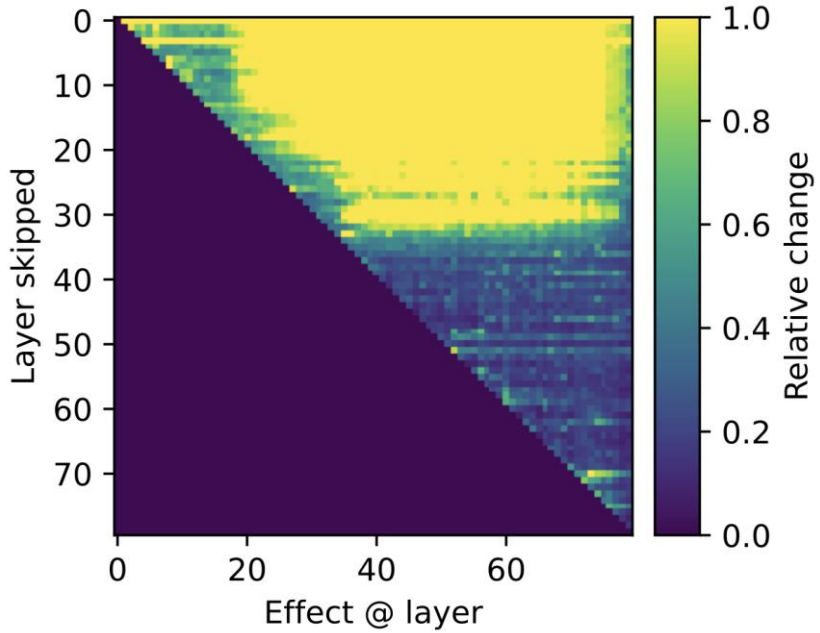
Yet deep layers are often underutilized

- Unlike early layers, deep layers ≈ identity mappings [Csordás et al. ’25]
- Best visual embeddings in intermediate layers [Bolya et al. ’25]

Residual connections: the default for scaling depth



Fixed cascaded layout: layer k outputs $r_k(\mathbf{x}_k) + \mathbf{x}_k$ [He et al. ’16]



Skipping a deep layer has less impact on the outputs of subsequent layers than skipping an early layer
Llama 3.1 70B [Csordás et al. ’25]

This work: where residual connections are placed, i.e., the residual topology, matters for optimization

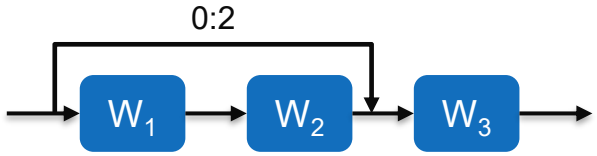
R. Csordás et al., Do language models use their depth efficiently? arXiv 2025. D. Bolya et al., Perception encoder, arXiv 2025. K. He et al., CVPR 2016.

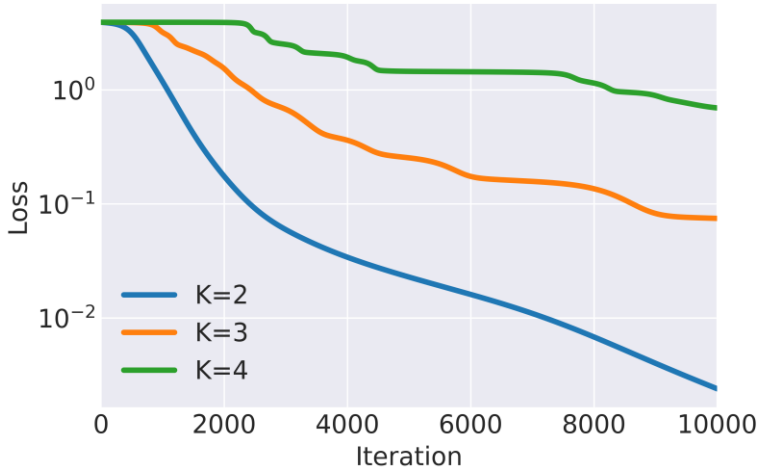
Residual topology matters

Case study: K-layer linear neural network (LNN) $f(\mathbf{x}) = \mathbf{W}_K \cdots \mathbf{W}_2 \mathbf{W}_1 \mathbf{x}$

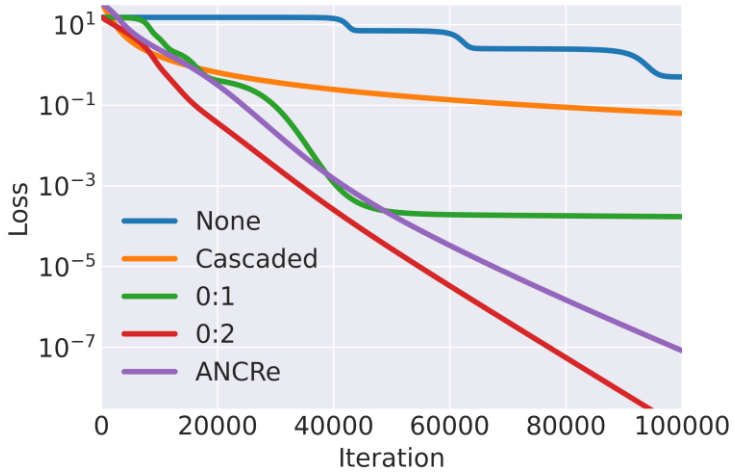
Least-squares regression, trained with gradient descent

Notation: $i{:}\,j$ = shortcut from the output of layer i to the output of layer j; 0 = input

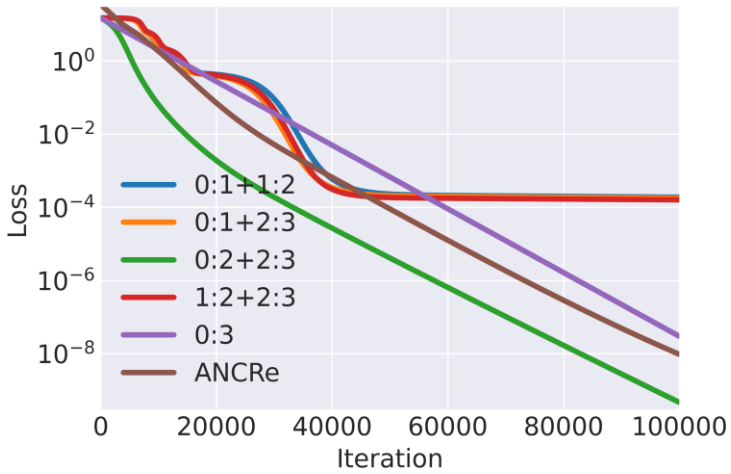




No shortcut, $K = 2,3,4$



$K = 3$, one shortcut



$K = 4$, two shortcuts

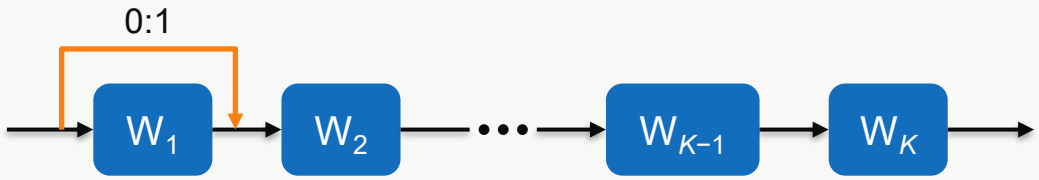
Take-home The conventional cascaded layout is not optimal, and additional residual connections help only when appropriately located.

Exponential gap in convergence

Setting: K-layer LNN with one shortcut, regression with L-2 loss $\mathcal{L}$ on whitened inputs $\mathbf{X}\mathbf{X}^\top = \mathbf{I}$

Gradient flow $\frac{\mathrm{d}\mathbf{W}_k(t)}{\mathrm{d}t} = -\nabla_{\mathbf{W}_k}\mathcal{L}(t)$ from a small initialization

Shortcut $0{:}\,1$: sublinear rate

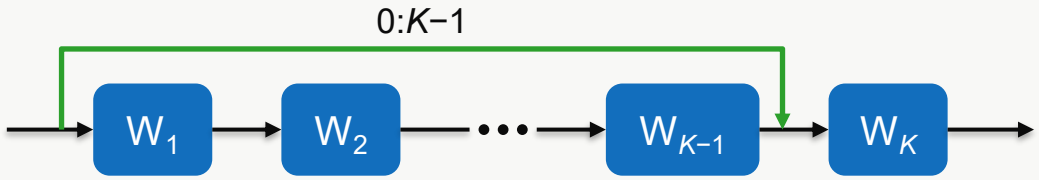


$$\mathcal{L}(t) \geq \Omega(1/t^{\frac{K-1}{K-2}})$$

for some small initialization

Deeper is slower: $\Omega(1/t^2)$ at $K = 3$, toward $\Omega(1/t)$ as $K \to \infty$

Shortcut $0{:}\,K-1$: linear rate



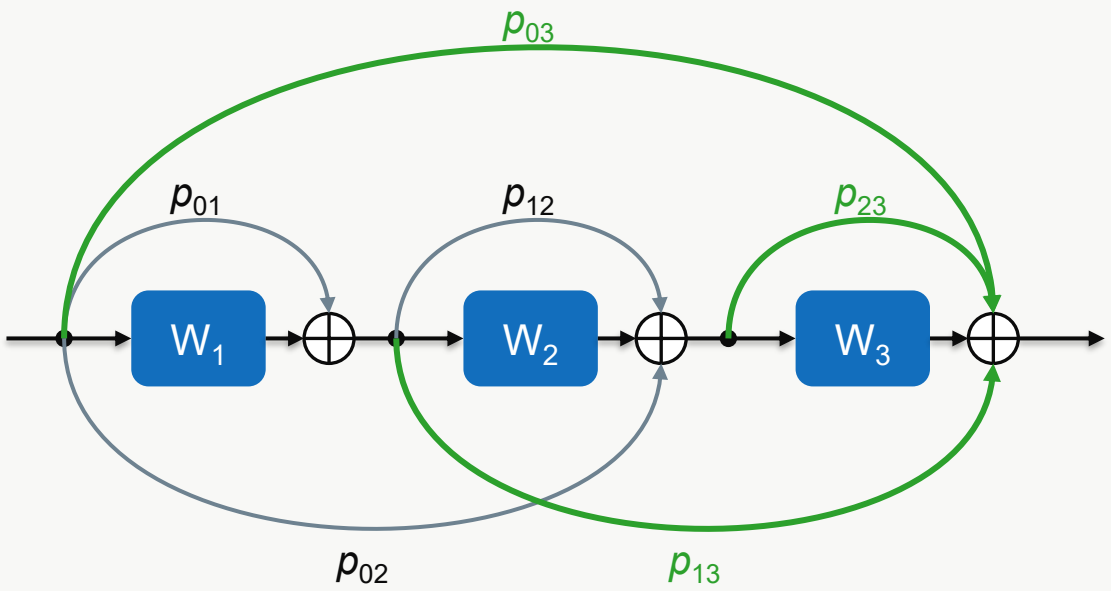
$$\mathcal{L}(t) \leq \mathcal{L}(0)\, e^{-2(1-\lambda)^2 t}$$

for any sufficiently small initialization

$0 < \lambda < 1$ depends on the initialization

Take-home Shortcut topology alone can lead to an exponential gap in convergence rate.

ANCRe: adaptive neural connection reassignment



$$p_{03} + p_{13} + p_{23} = 1$$

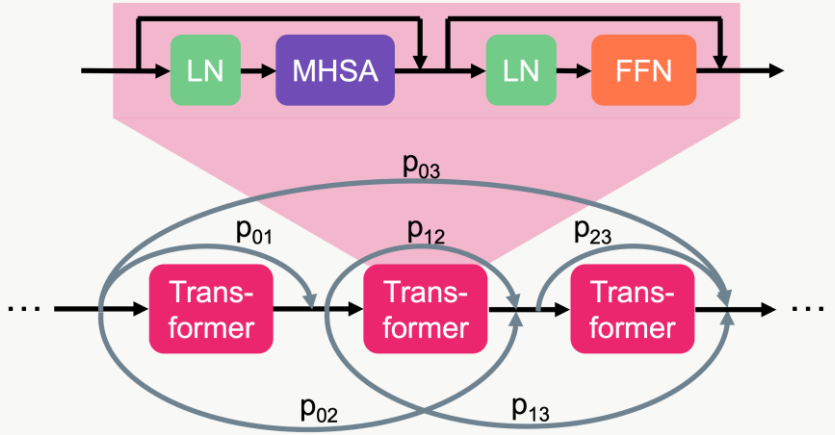
$$p_{ij} = \frac{\exp(c_{ij}/\tau)}{\sum_{k=0}^{j-1} \exp\left(\frac{c_{kj}}{\tau}\right)}$$

Learnable parameters c_{ij}; temperature $\tau > 0$

Ensemble of subnetworks of different depths

Layer j receives $\sum_{i=0}^{j-1} p_{ij} f_i(\mathbf{x})$, where $f_i(\mathbf{x})$ is the output of layer i and p_{ij} a routing coefficient

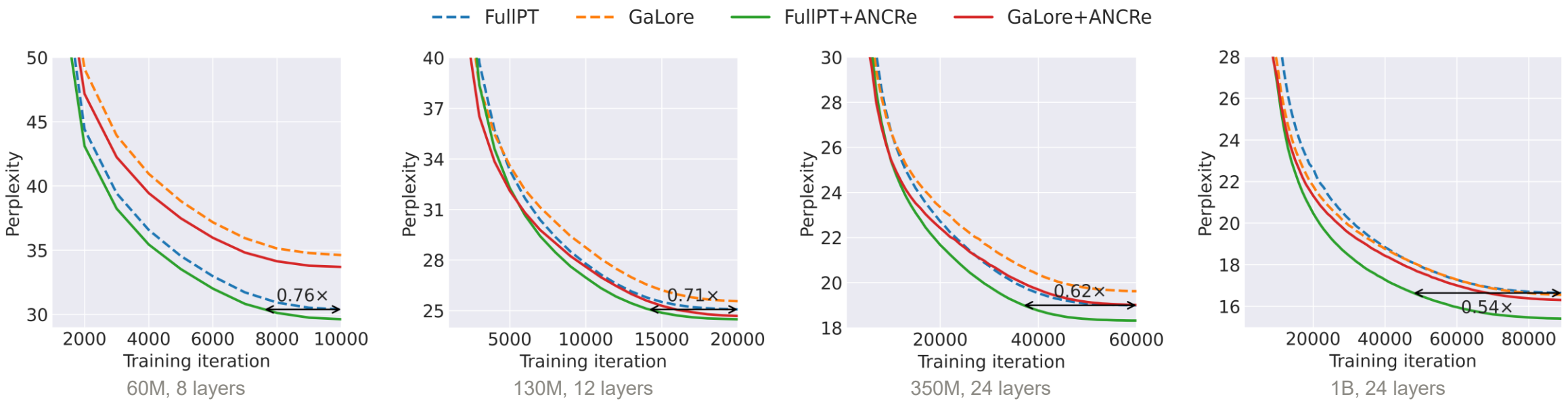
Transformers: shortcuts between block inputs



Negligible overhead

< 1% throughput drop, < 0.01 GB extra memory

Pre-training LLaMA on C4

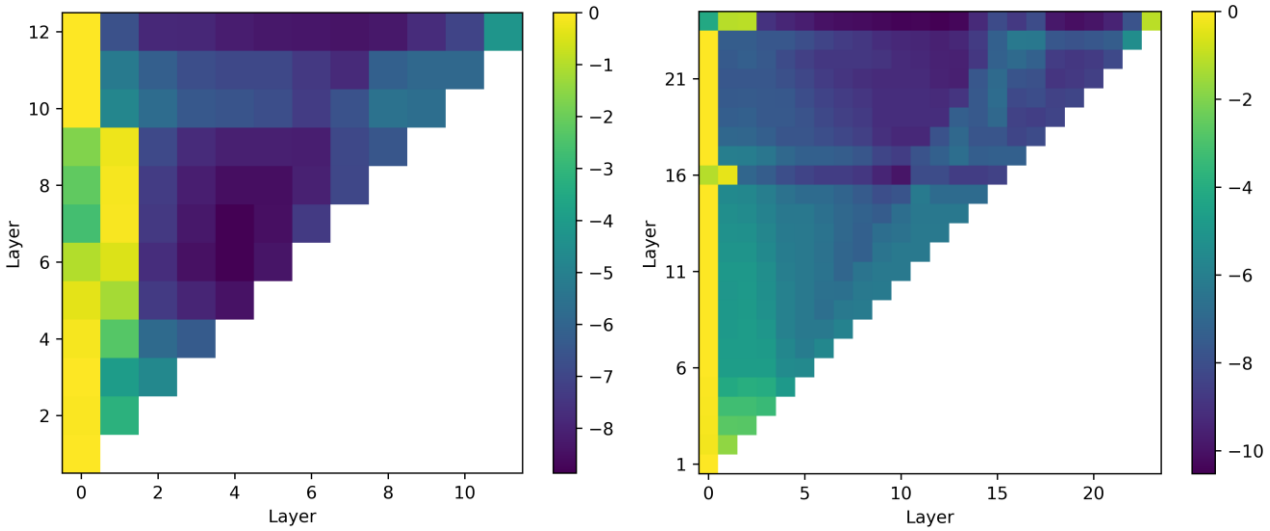


Perplexity ↓	60M	130M	350M	1B
FullPT	30.39	25.07	19.00	16.64
FullPT + HC	29.86	24.51	18.58	16.38
FullPT + mHC	30.21	24.62	18.93	16.14
FullPT + ANCRe	**29.62**	**24.48**	**18.32**	**15.41**
GaLore	34.61	25.56	19.62	16.55
GaLore + ANCRe	**33.69**	**24.68**	**19.01**	**16.45**

- Lowest perplexity in all 8 settings
- Better than HC and mHC at every size
- Up to 1.85× speedup to reach the baseline (1B)
- Gain grows with depth: 24% → 46% fewer iterations

C4 [Raffel et al. '20], LLaMA [Touvron et al. '23], GaLore [Zhao et al. '24], HC [Zhu et al. '25], mHC [Xie et al. '25]

Learned topology



130M, 12 layers

1B, 24 layers

$\log p_{ij}$ learned by ANCRe

Perplexity ↓	130M	350M
ANCRe	**24.48**	**18.32**
Fixed: first and previous layers	25.44	19.94
Retrained with learned topology frozen	24.48	18.49

Learned pattern

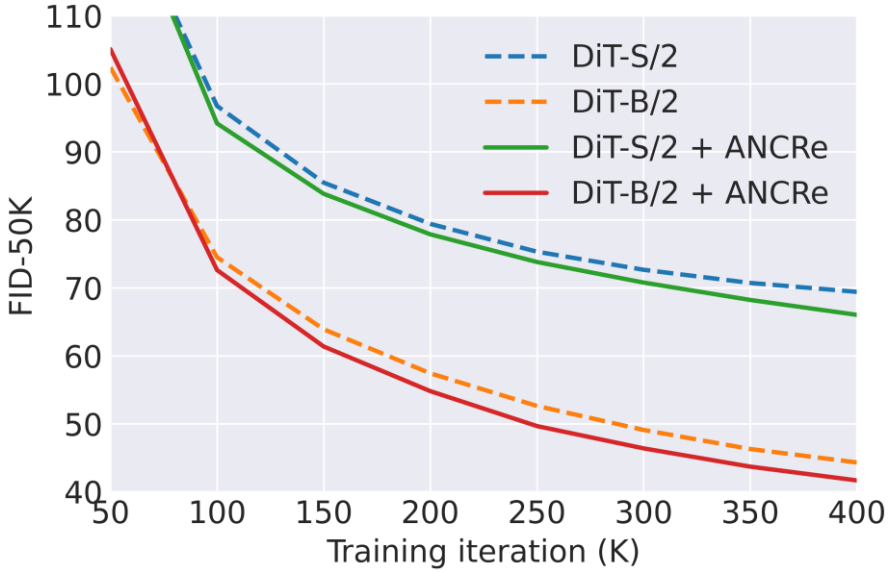
- Deep layers ensemble subnetworks of different depths
- Most weight on shortcuts from the first two layers
- Denser connectivity in larger models

Learning the topology matters

- Fixing the dominant shortcuts is clearly worse
- Retraining with the learned topology frozen keeps nearly all the gain

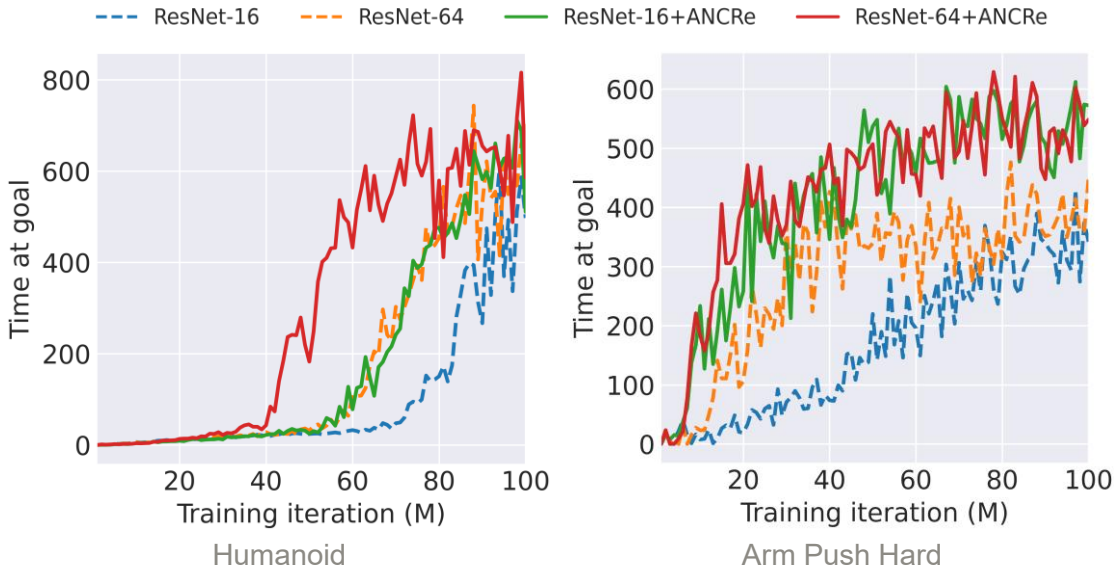
Diffusion models and reinforcement learning

Diffusion: DiT on ImageNet 256×256



- Lower FID throughout training for both model sizes
- All 5 metrics improve, ≈ 6% on average

Goal-conditioned RL with deep ResNets



- ResNet-16 + ANCRe matches or beats ResNet-64 (4× deeper)
- Faster convergence on all 4 tasks (2 shown)

DiT [Peebles & Xie '23], contrastive RL [Eysenbach et al. '22], deep ResNets for RL [Wang et al. '25]

Ablations and overhead

Perplexity ↓	130M	350M
ANCRe	**24.48**	**18.32**
Cascaded (FullPT)	25.07	19.00
All shortcuts, fixed uniform coefficients	diverge	diverge
Cascaded shortcuts, learnable coefficients	27.18	21.72

Model	Method	Params	Mem (GB)	Throughput (tok/s)
350M	Cascaded	0	26.60	99.9k
	HC	172k	46.76	43.6k
	mHC	172k	55.55	33.6k
	ANCRe	**300**	**26.60**	**99.2k**
1B	Cascaded	0	30.38	17.3k
	HC	345k	50.52	12.0k
	mHC	345k	58.50	10.3k
	ANCRe	**300**	**30.38**	**17.3k**

Shortcuts and coefficients

- All shortcuts alone diverge; learned coefficients alone are worse than cascaded
- The gains come from learning the topology, not from adding shortcuts

Overhead

- ANCRe: no extra memory, throughput within 1% of cascaded
- HC, mHC: +20–29 GB, 31–66% lower throughput

HC [Zhu et al. '25], mHC [Xie et al. '25]

Conclusion

Topology matters

Shortcut topology alone can lead to an exponential gap in convergence rate

ANCRe learns it

Softmax-normalized coefficients on all shortcuts, with < 1% overhead

Better use of depth

Faster convergence and better results for LLMs, diffusion models, and deep RL

Limitations	Theory covers linear networks only; gains are largest for deep networks
Outlook	Analysis beyond linear networks; larger-scale pre-training