NeurIPS 2025

Don’t be lazy: CompleteP enables compute-efficient deep transformers

Nolan Dey* Cerebras Systems

Bin Claire Zhang* Cerebras Systems

Lorenzo Noci ETH Zurich, Princeton University

Mufan Li Princeton University

Blake Bordelon Harvard University

Shane Bergsma Cerebras Systems

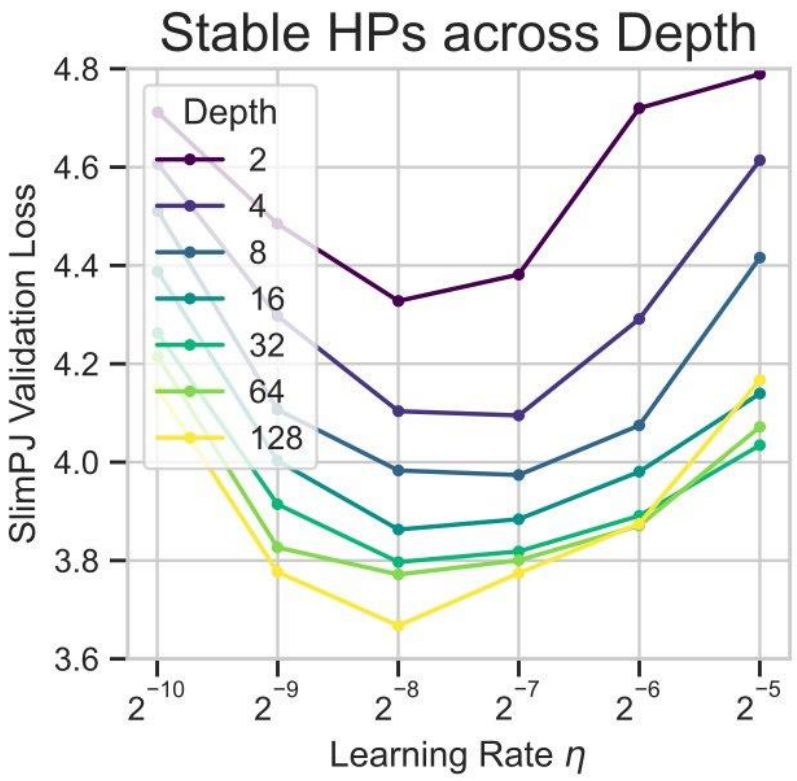
Cengiz Pehlevan Harvard University

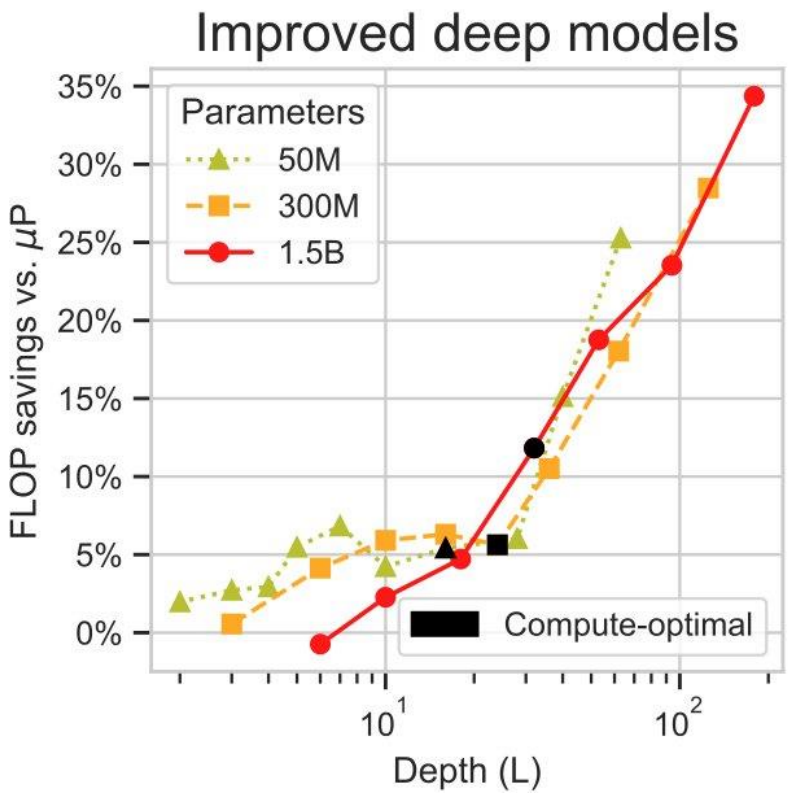
Boris Hanin Princeton University

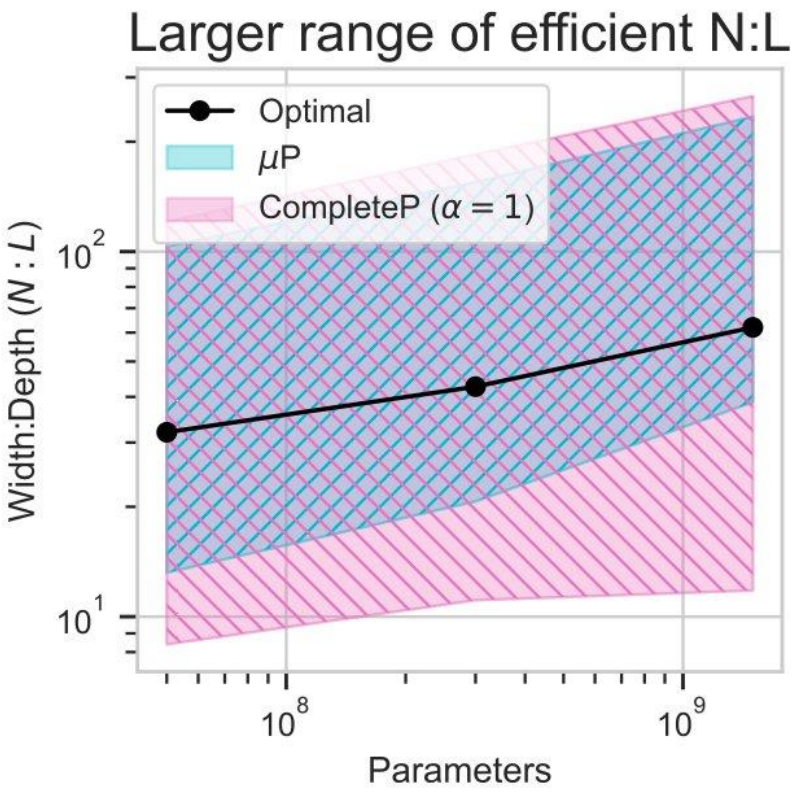
Joel Hestness Cerebras Systems

TLDR

- **TLDR**: We introduce CompleteP, which offers depth-wise hyperparameter (HP) transfer (**Left**), fewer FLOPs to reach the same loss with deeper models (**Middle**), and a larger range of compute-efficient width:depth ratios (**Right**).







Problems with training deep transformers

- Deeper = more unstable
  - Too risky to train a large-scale deep network

- As depth is varied, models do not share the same optimal hyperparameters
  - Expensive to tune hyperparameters
  - More likely to choose suboptimal hyperparameters

- Deeper-narrower models are not competitive with shallower-wider models
  - Everyone trains with width:depth = 100

Infinite depth parameterizations

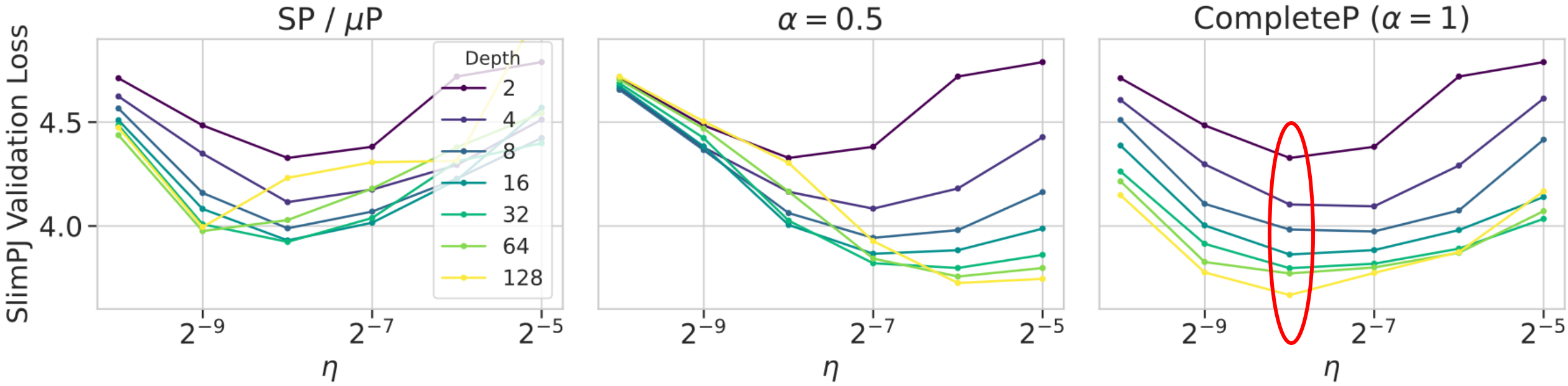
- ResNet:
  - $h^{l+1} = h^l + L^{-\alpha} F_l(h^l), l \in \{1, \ldots, L\}$
  - $\eta_l = \eta_{base} L^{\alpha-1}$

- Yang et al. [1] argue $\alpha = 0.5$ is best in practice, but that **depth-wise HP transfer is not possible** for any α
  - Several works have since adopted this prescription!

- Bordelon et al. [3] find $\alpha = 1.0$ allows better learning as $L \to \infty$

- In this work, we study the two promising candidates: $\alpha \in \{0.5, 1\}$

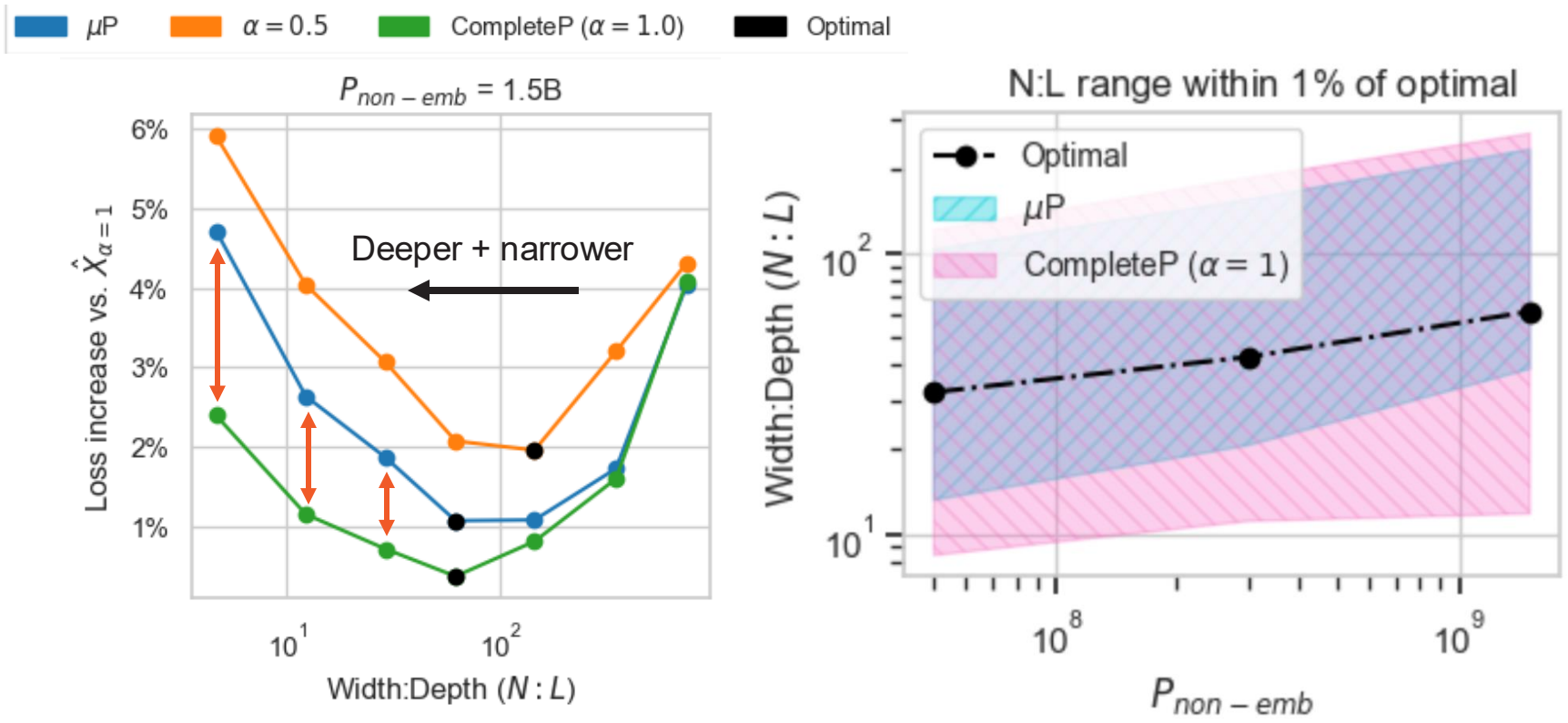
Only $\alpha = 1$ ensures depth-wise HP transfer



Depth-wise HP transfer, with constant tokens, constant batch size

Compute-efficient deep transformers

- Left: CompleteP ($\alpha = 1$) enables more compute-efficient deep transformers than $\{\mu P, \alpha = 0.5\}$
- Right: CompleteP ($\alpha = 1$) enables a larger range of efficient width:depth (N:L) ratios



How CompleteP works

$\alpha \in \{0.5, 1\}$ stabilize training dynamics

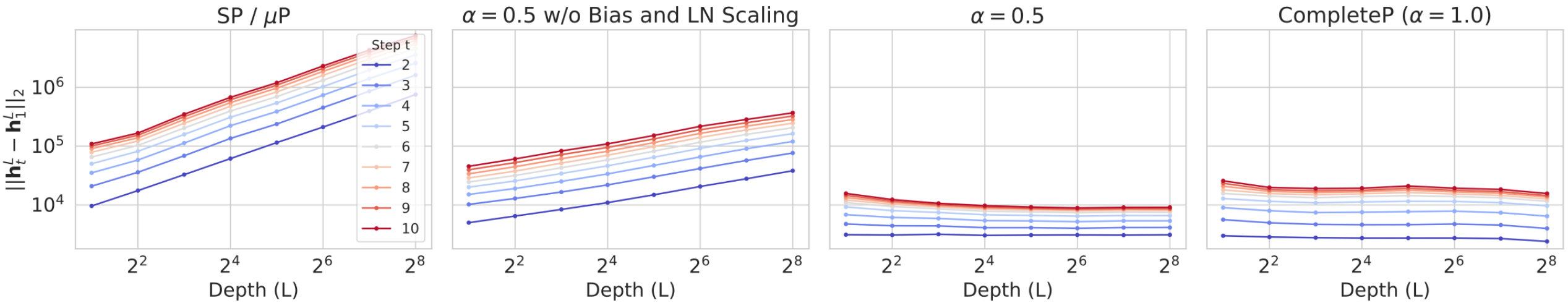
- Setup: For several depths, train models for 10 steps and record activation norms at final residual stream h^L
  - All the points at each depth value comprise a single training run
  - We desire flat lines (desiderata 1 & 2)
- Theory: Any $\alpha \in [0.5, 1]$ will have flat lines (satisfying desiderata 1 and 2)

Exploding

Exploding

Stable

Stable



Only $\alpha = 1$ ensures "Complete Feature Learning"

- **Problem:** How do we theoretically explain the success of α=1?
- **Minimal example:** ResNet with 2-layer linear MLP block: $h^{l+1} = h^l + L^{-\alpha} W^l_{(2)} W^l_{(1)} h^l$
- Consider Δh^{l+1}, the change in h^{l+1} after a weight update $\left(W^l_{(1)}, W^l_{(2)}\right) = \theta^l \to \theta^l + \Delta\theta^l$

$$\begin{aligned}\Delta_{\boldsymbol{\theta}^\ell} h^{\ell+1} &= \langle \nabla_{\boldsymbol{\theta}^\ell} h^{\ell+1}, \Delta\boldsymbol{\theta}^\ell \rangle + \frac{1}{2} \nabla^2_{\boldsymbol{\theta}^\ell} h^{\ell+1} [\Delta\boldsymbol{\theta}^\ell, \Delta\boldsymbol{\theta}^\ell] \\ &= \underbrace{L^{-\alpha} (W^\ell_{(2)} h^\ell \underbrace{\Delta W^\ell_{(1)}}_{L^{\alpha-1}} + W^\ell_{(1)} h^\ell \underbrace{\Delta W^\ell_{(2)}}_{L^{\alpha-1}})}_{L^{-1}} + \underbrace{L^{-\alpha} h^\ell \underbrace{\Delta W^\ell_{(1)} \Delta W^\ell_{(2)}}_{L^{2(\alpha-1)}}}_{L^{\alpha-2}} . \end{aligned} \tag{6}$$

- **Desiderata 3 (Complete Feature Learning):** For all layers, we want all higher-order terms to have the same asymptotic scale as $L \to \infty$
- Only $\alpha = 1$ ensures "Complete Feature Learning", thus we dub it "CompleteP"
- $\alpha = 0.5$ has vanishing higher-order terms ("Lazy") w.r.t. first-order term as $L \to \infty$

CompleteP benefits

- Stable optimal HPs across depths
- Significant FLOPs savings over μP for deep models
- Complete Feature Learning
- Easy to implement (see below)

$$\mathbf{X}^l + L^{-1} \cdot \mathrm{MHA}(\mathrm{LN}(\mathbf{X}^l))$$
$$\mathbf{Z}^l + L^{-1} \cdot \mathrm{MLP}(\mathrm{LN}(\mathbf{Z}^l))$$

References

- [1] Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor Programs VI: Feature Learning in Infinite-Depth Neural Networks. arXiv, 2023.
- [2] Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=KZJehvRKGD.
- [3] Blake Bordelon, Hamza Tahir Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. URL https://openreview.net/forum?id=p0BBKhD5aI.