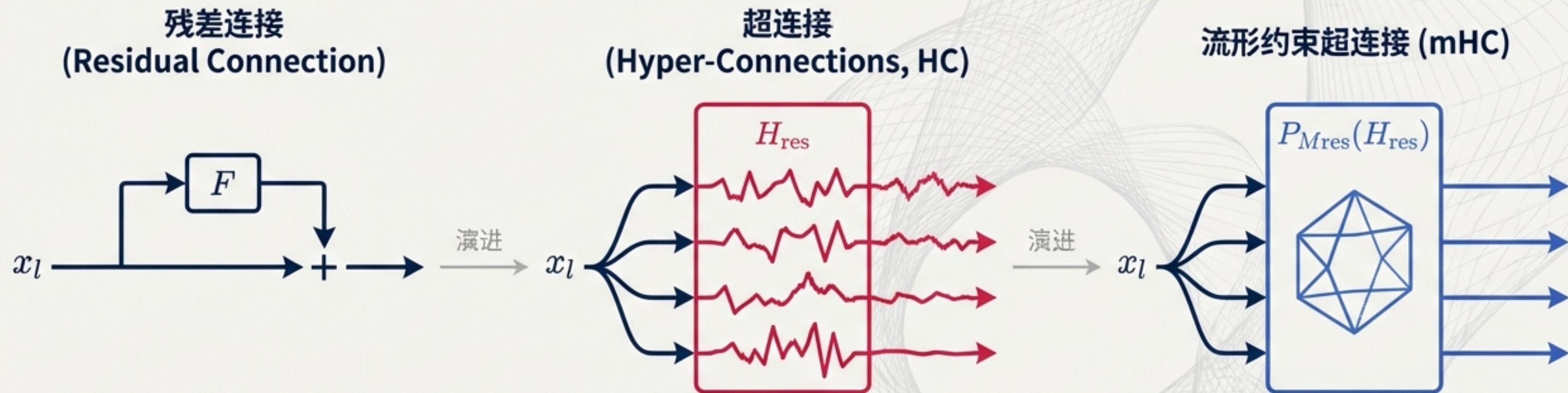


标题：mHC：通过流形约束重塑超连接，实现可扩展的深度学习架构

副标题：DeepSeek-AI 对基础模型拓扑结构设计的深度探索与实践



Zhenda Xie*, Yixuan Wei*, Huanqi Cao*, et al.
DeepSeek-AI
arXiv:2512.24880 [cs.CL]

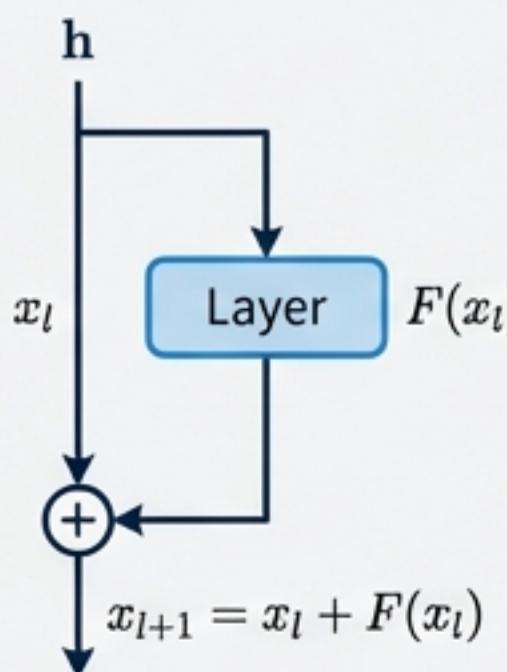
超连接(HC)的出现：一个富有潜力但存在隐患的范式

概念阐述

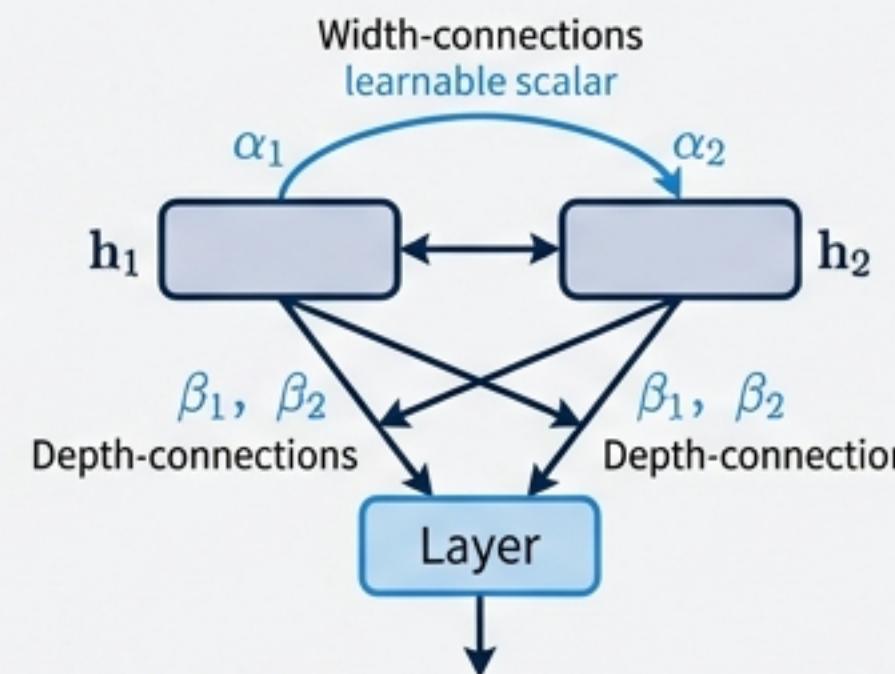
核心思想

传统残差连接 ($x_{l+1} = x_l + F(x_l)$) 只有一个信息流。HC通过将残差流的宽度扩展 n 倍，并引入可学习的连接矩阵，极大地增加了网络拓扑的复杂性和信息容量，而FLOPs开销几乎不变。

$$x_{l+1} = H_{\text{res}} * x_l + H_{\text{post}}^T * F(H_{\text{pre}} * x_l, W_l)$$



(a) Residual Connection

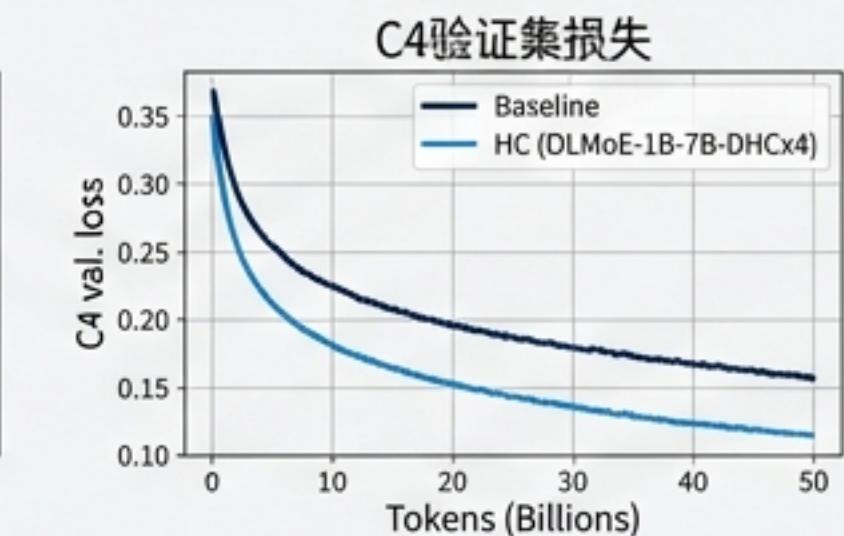
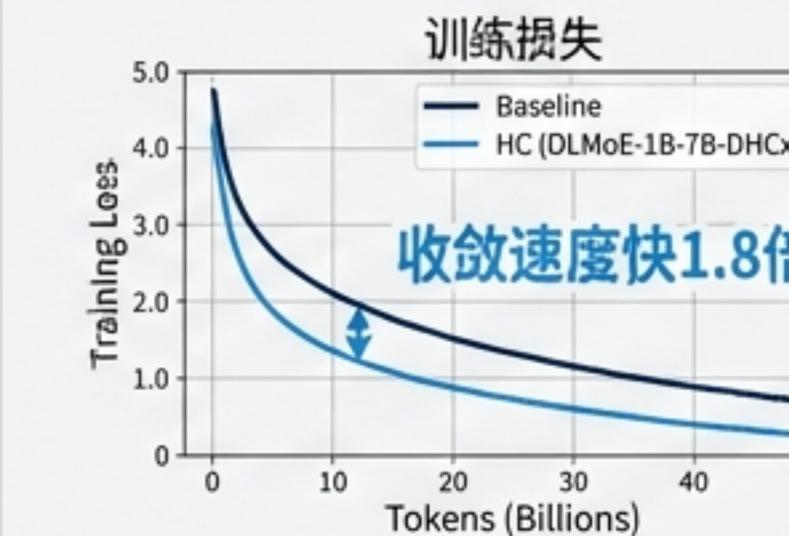


(b) Hyper-connections

性能验证

引言

HC在其原始论文中展示了显著的性能优势。



结论：HC的初步成功证明，探索残差连接之外的拓扑结构具有巨大潜力。

标题：HC的致命缺陷：无约束连接破坏了信号传播的稳定性

核心论点：HC虽然性能优越，但其可学习的连接矩阵 H_{res} 是无约束的。这破坏了残差学习赖以成功的核心原则——“恒等映射”（identity mapping）。

左侧：理论分析

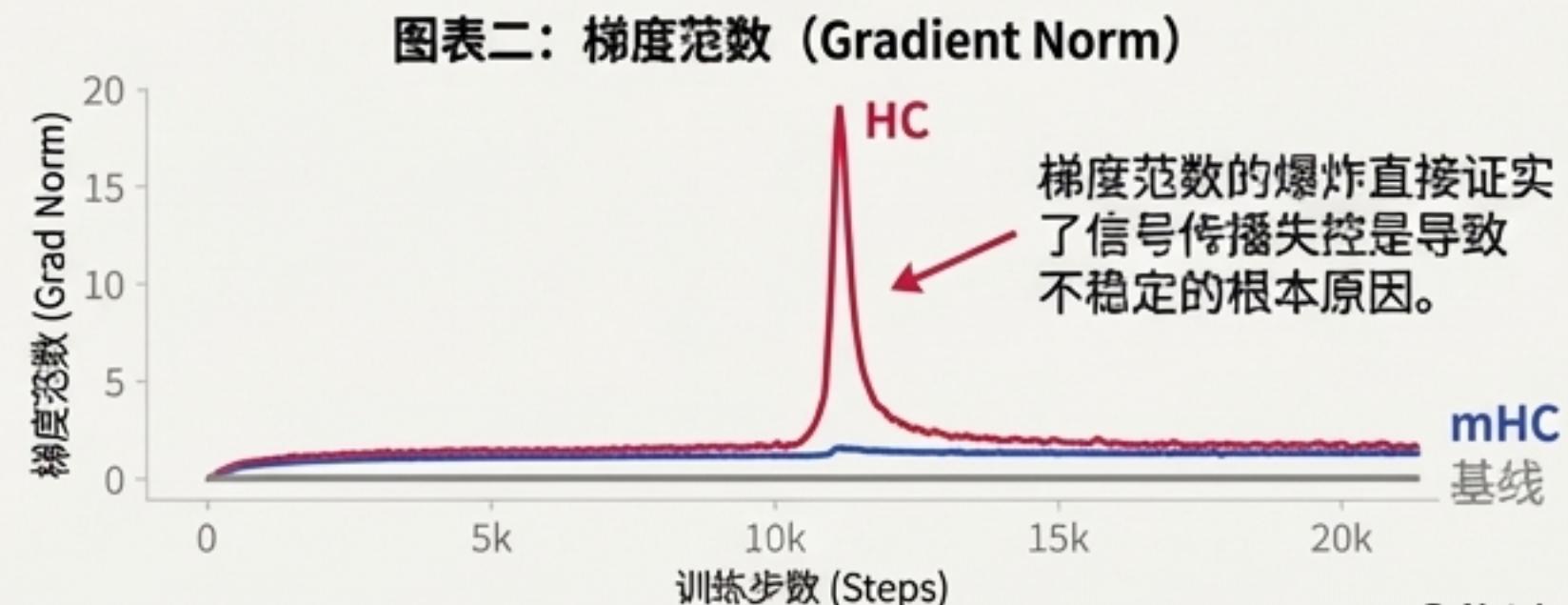
问题根源：在深度网络中，信号从浅层 l 传播到深层 L ，有效路径由复合映射

$$\left(\prod_{i=1}^{L-l} H_{\text{res}, L-i} \right) * x_l.$$

缺陷：由于 H_{res} 无约束，这个复合映射会**偏离单位阵**，导致信号在传播过程中被**无限制地放大或衰减**，最终破坏了训练的稳定性。

“this discrepancy leads to **unbounded signal amplification or attenuation**, resulting in instability during large-scale training.”

右侧：实证证据（27B模型）



标题：信号爆炸的定量诊断：HC复合映射导致增益失控

核心指标解释

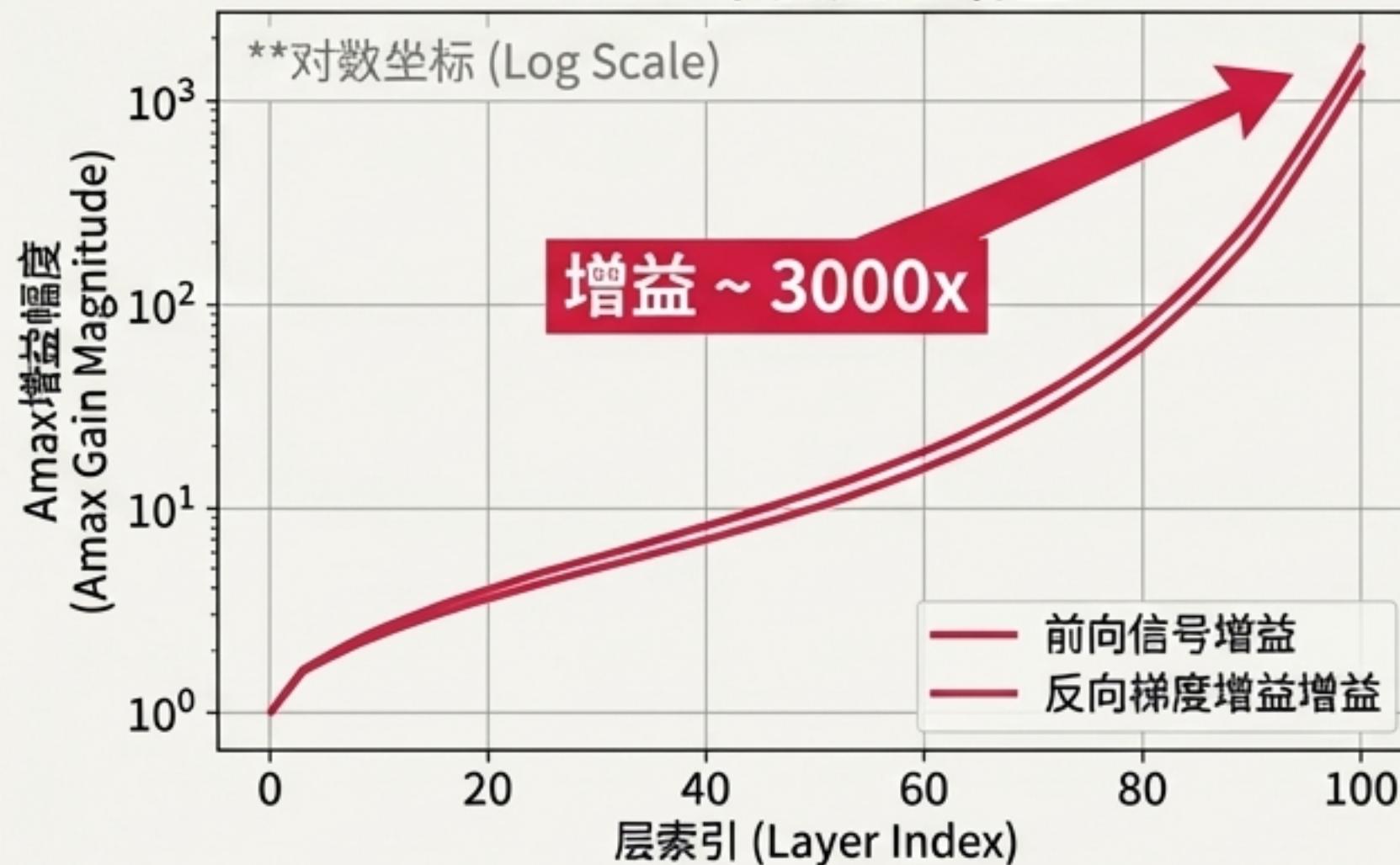
A_{max}增益幅度 (A_{max} Gain Magnitude)：量化信号传播的放大效应。

前向信号增益：复合映射矩阵的“最大绝对行和”，代表了信号在前向传播中的最坏情况放大倍数。

反向梯度增益：复合映射矩阵的“最大绝对列和”，对应于梯度在反向传播中的最坏情况放大倍数。

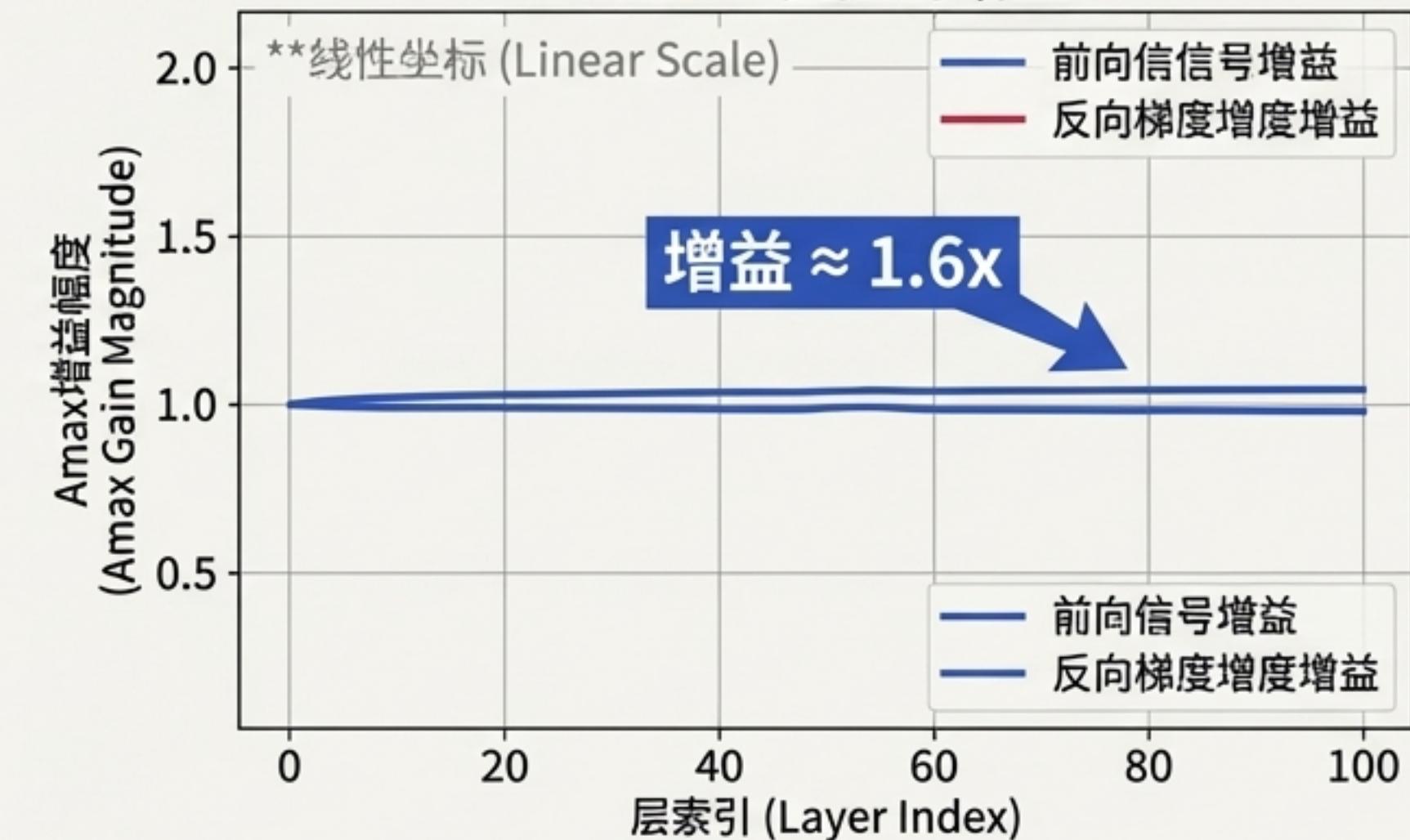
理想值：对于稳定的恒等映射，该值应恒为 1。

HC：不受控的增益



HC中的信号和梯度在跨层传播时经历了**剧烈的、不受控制的放大**，这是训练不稳定的直接原因。

mHC：受控的增益



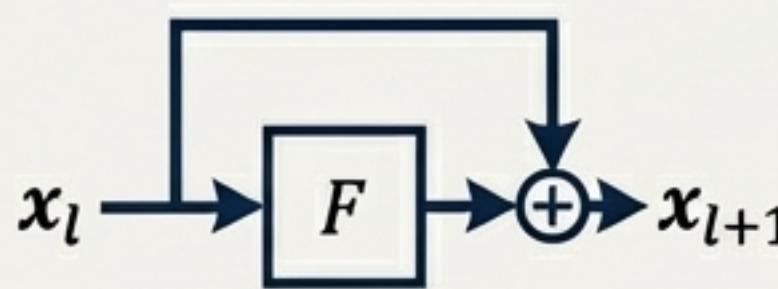
通过流形约束，mHC将信号增益**控制在理想值1附近**，从根本上恢复了传播的稳定性。

标题：解决方案mHC：将连接空间投影至流形以恢复恒等映射

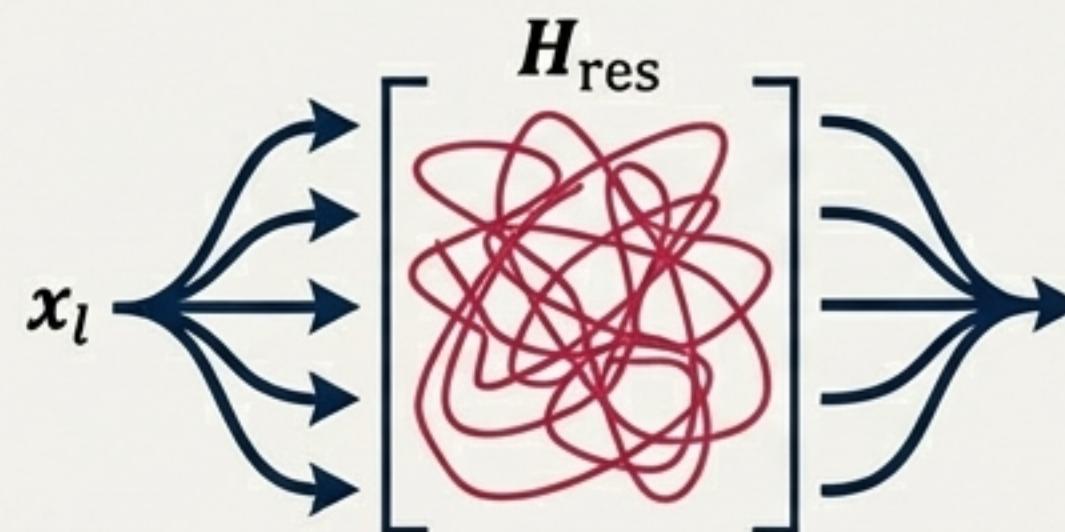
核心思想：mHC的核心创新并非完全抛弃HC，而是通过数学手段“驯服”它。我们不让 H_{res} 在整个实数矩阵空间自由学习，而是将其**投影（Project）**到一个特定的、具有良好性质的**流形（Manifold）**上，从而在保留HC多流信息交互能力的同时，恢复其稳定性。

视觉核心：架构演进图

(a) 残差连接
(Residual Connection)



(b) 超连接 (HC)



$$x_{l+1} = x_l + F(x_l)$$

$$x_{l+1} = H_{\text{res}} * x_l + \dots$$

(c) 流形约束超连接 (mHC)



$$x_{l+1} = P_{M_{\text{res}}}(H_{\text{res}}) * x_l + \dots$$

关键注解：mHC保留了HC的多样化连接模式，同时通过施加数学约束来确保信号传播的稳定性。

标题：理论核心：利用双随机矩阵的性质确保稳定性

核心概念：我们选择的流形是Birkhoff多面体（Birkhoff Polytope），即所有 $n \times n$ 双随机矩阵（Doubly Stochastic Matrices）的集合。

双随机矩阵的定义：

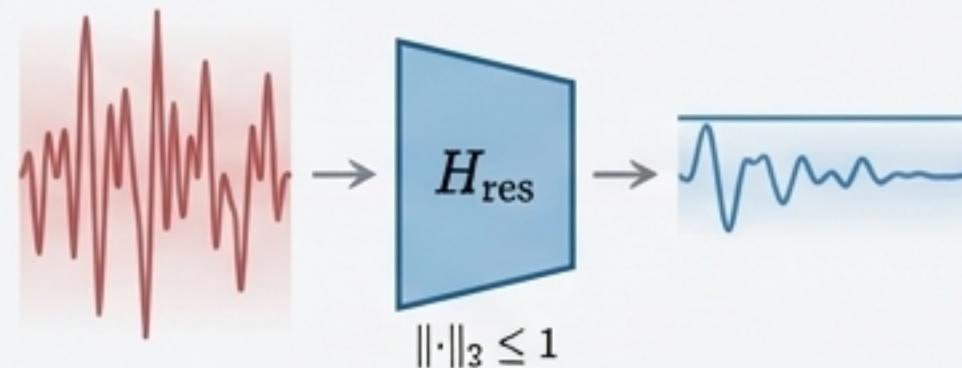
一个非负矩阵，其所有行和与所有列和均为1。

$$H_{\text{res}} \in \mathbb{R}^{n \times n} \text{ 使得 } H_{\text{res}} * 1_n = 1_n, 1_n^T * H_{\text{res}} = 1_n^T, H_{\text{res}} \geq 0.$$

第一列：范数保持 (Norm Preservation)

性质：双随机矩阵的谱范数 $\|H_{\text{res}}\|_2 \leq 1$ 。这意味着该变换是非扩张性的。

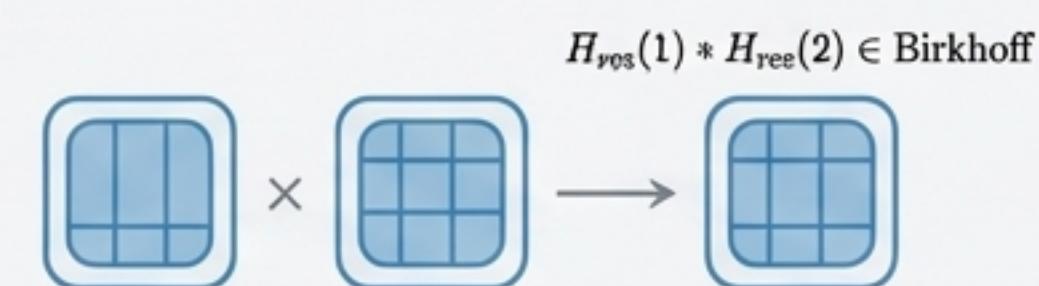
效果：从根本上抑制了信号和梯度的爆炸问题。



第二列：复合闭包性 (Compositional Closure)

性质：两个双随机矩阵的乘积仍然是双随机矩阵。

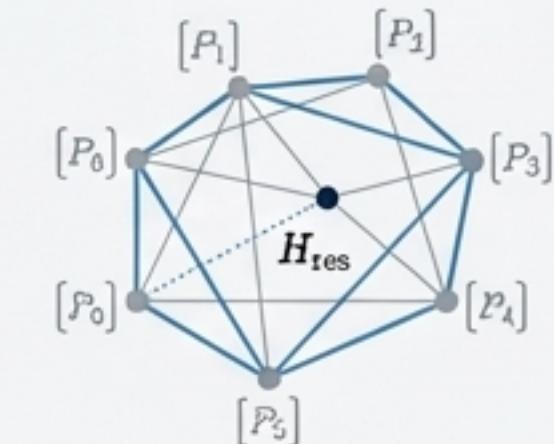
效果：确保了跨越多层的复合映射 ΠH_{res} 依然保持稳定，从而保证了全局的稳定性。



第三列：几何解释 (Geometric Interpretation)

性质：Birkhoff多面体是所有置换矩阵的凸包。

效果：这意味着 H_{res} 的作用可以被理解为对不同信息流进行排列组合的“凸组合”，是一种鲁棒的特征融合机制。



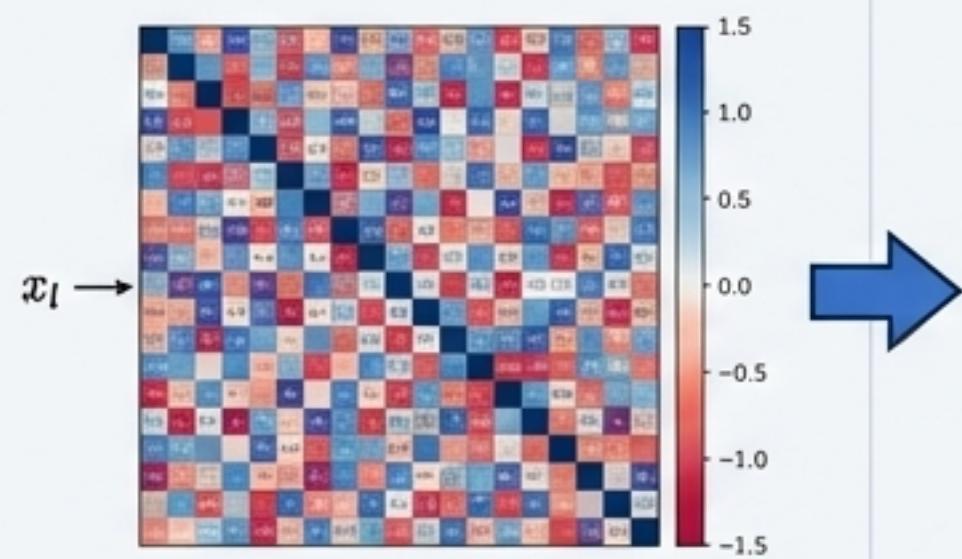
结论：通过将 H_{res} 约束在Birkhoff多面体内，mHC获得了一套“免费的”理论保障，使其在拥有HC灵活性的同时，具备了比标准残差连接更强的稳定性。

标题：实现机制：通过Sinkhorn-Knopp算法实现流形投影

Step 1: 生成无约束矩阵

首先，像HC一样，通过模型的输入 x_l 动态生成一个无约束的参数矩阵 \tilde{H}_{res}

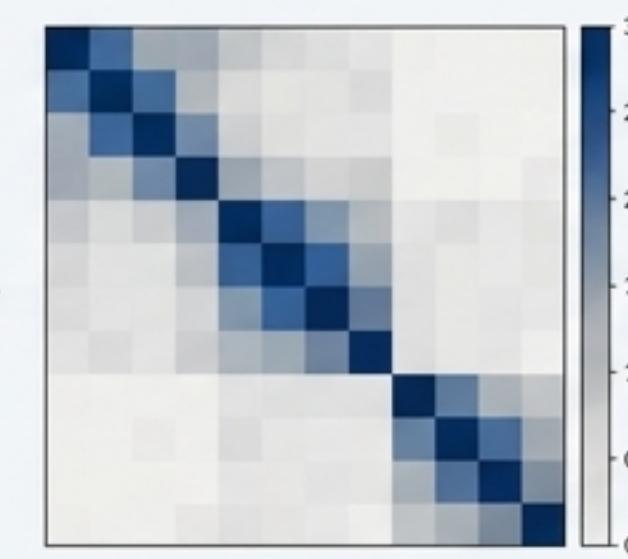
$$\tilde{H}_{\text{res}} = \alpha_{\text{res}} * \text{mat}(\text{vec}(x_l)' * \varphi_{\text{res}}) + b_{\text{res}}$$



Step 2: 确保非负性

对 \tilde{H}_{res} 取指数，得到一个严格为正的矩阵 $M(0)$ 。

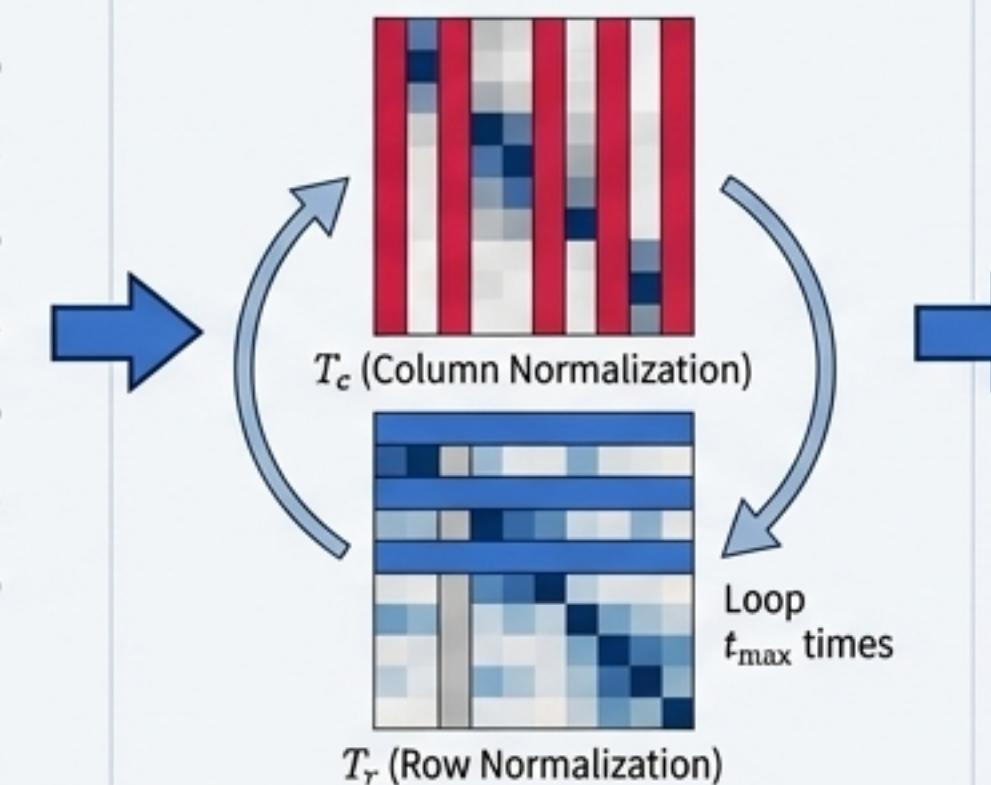
$$M(0) = \exp(\tilde{H}_{\text{res}})$$



Step 3: 迭代归一化 (Sinkhorn-Knopp 迭代)

交替对矩阵的行和与列和进行归一化，使其趋近于1。

$$M(t) = T_r(T_c(M(t-1)))$$

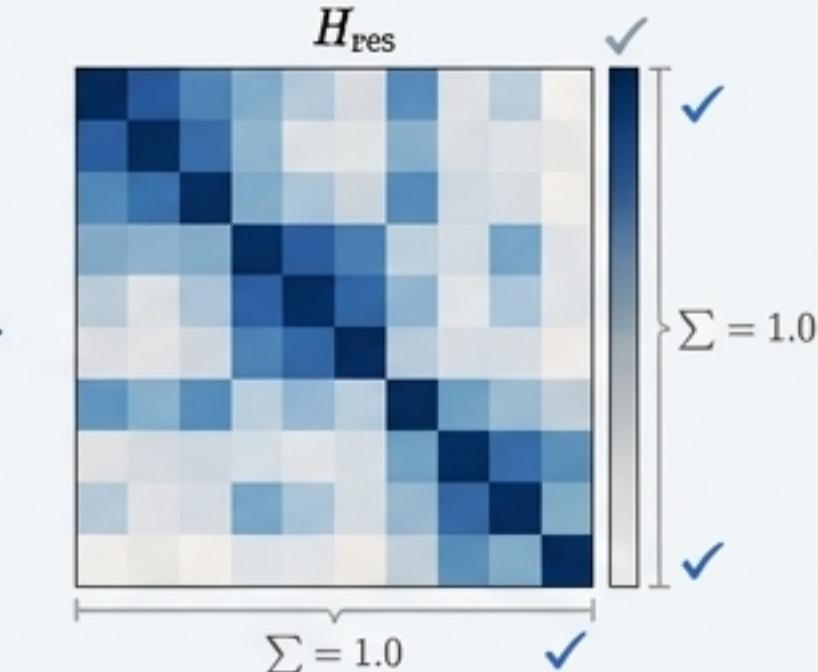


实践：在实验中，我们设置迭代次数 $t_{\max} = 20$ ，在效率和精度之间取得了良好平衡。

Step 4: 获得约束矩阵

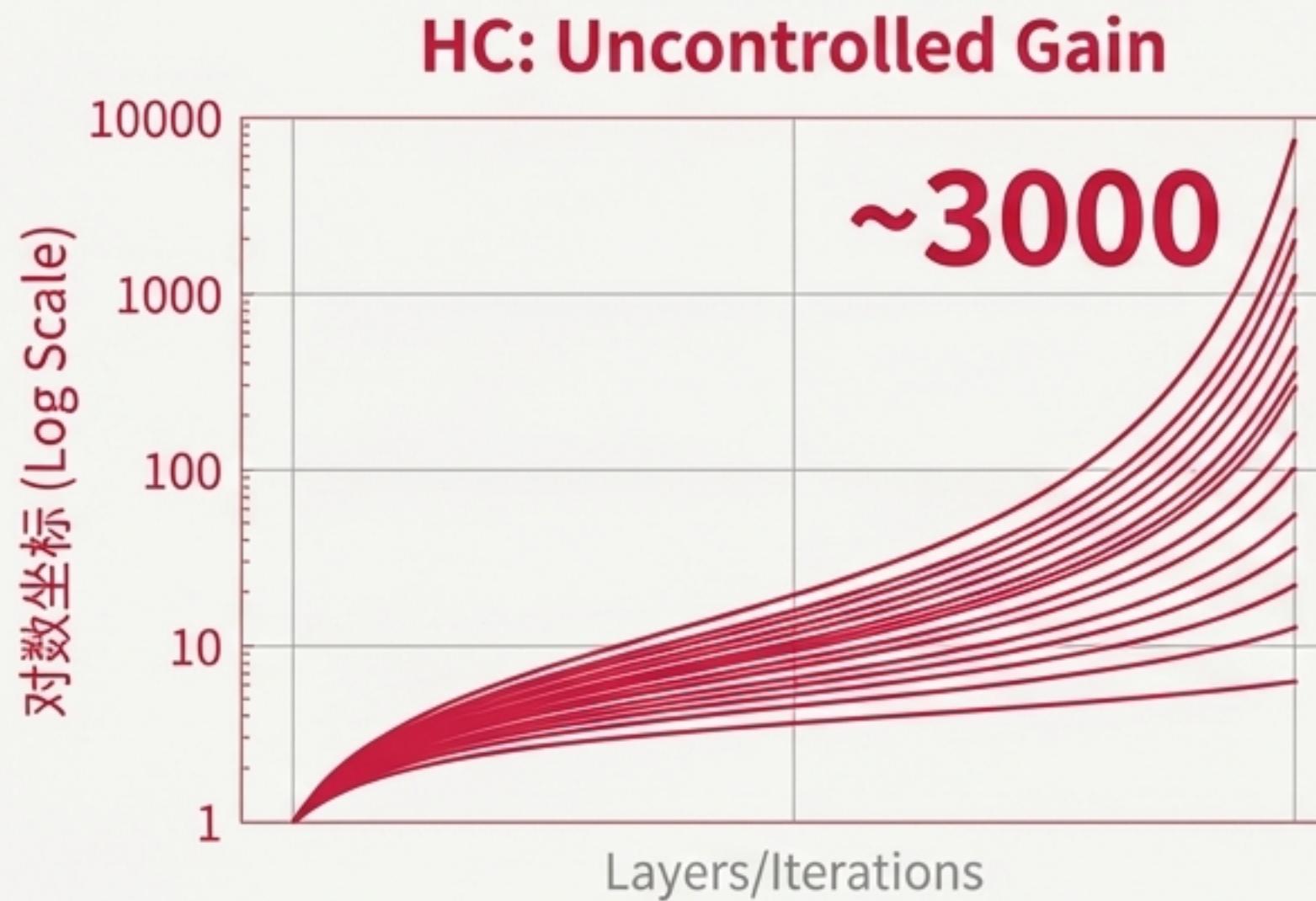
迭代结束后的矩阵即为我们需要的 H_{res} 。

$$H_{\text{res}} = M(t_{\max})$$

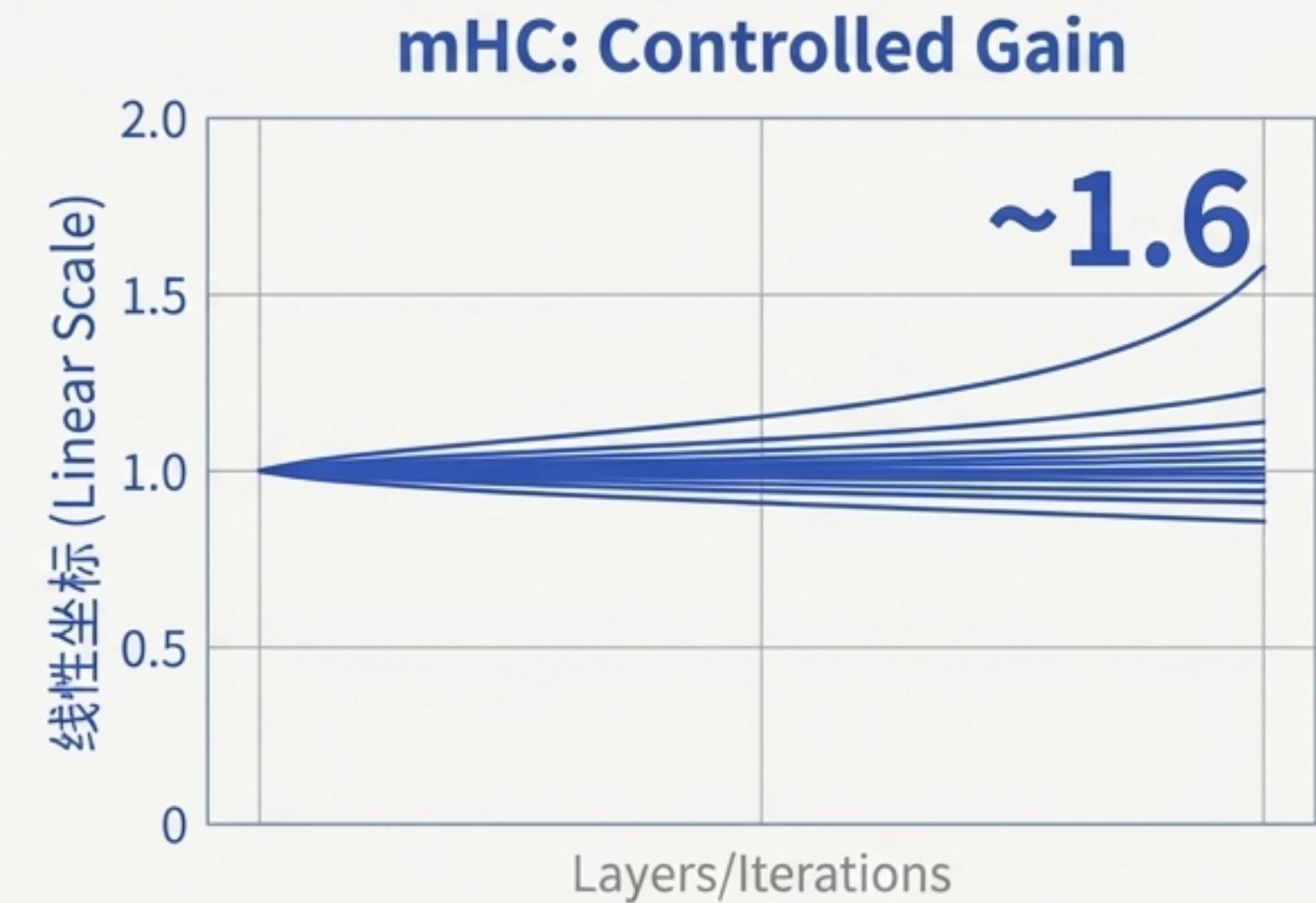


标题：立竿见影的效果：mHC成功将信号增益控制在稳定范围

核心信息：mHC的设计初衷是恢复信号传播的稳定性。Amax增益幅度的分析结果直接证明了我们已经成功实现了这一目标。



无约束的复合映射导致信号增益无法控制，造成数值不稳定。



流形约束将复合映射的增益严格控制在1附近，确保了端到端的传播稳定性。

mHC将信号增益的失控幅度降低了三个数量级，从根本上解决了HC的可扩展性问题。

标题：连接矩阵可视化：从数值层面直观感受HC与mHC的差异

第一行：HC (Hyper-Connections)

H_{res} (单层)

18.73	-2.41	5.89	-3.51
-15.29	9.12	-6.04	7.88
4.25	-1.18	12.36	-4.92
-8.63	6.54	-7.21	10.05

列和: [-0.9, 12.0, 5.0, 9.5]

行和: [18.7]

行和: -4.3

行和: 10.5

行和: 0.7]

ΠH_{res} (复合)

-259.2	102.4	35.7	89.1
509.1	-180.5	67.3	-210.4
84.5	76.2	-98.7	115.3
-165.8	-95.4	134.9	-123.6

列和: [168.6, -97.3, 139.2, -129.6]

注解：数值范围大，正负交错，行/列和严重偏离1，表明信号在每个维度上都经历了剧烈的、不可预测的变换。

第二行：mHC (Manifold-Constrained Hyper-Connections)

$P_{Mres}(H_{res})$ (单层)

0.35	0.20	0.25	0.20
0.15	0.40	0.20	0.25
0.25	0.20	0.30	0.25
0.25	0.20	0.25	0.30

列和: [1.00, 1.00, 1.00, 1.00]

行和: [1.00]

行和: 1.00]

行和: 1.00]

行和: 1.00]

$\Pi P_{Mres}(H)$ 复合

$\Pi P_{Mres}(H_{res})$ (复合)

0.24	0.25	0.26	0.25
0.25	0.25	0.24	0.26
0.25	0.26	0.25	0.24
0.26	0.24	0.25	0.25

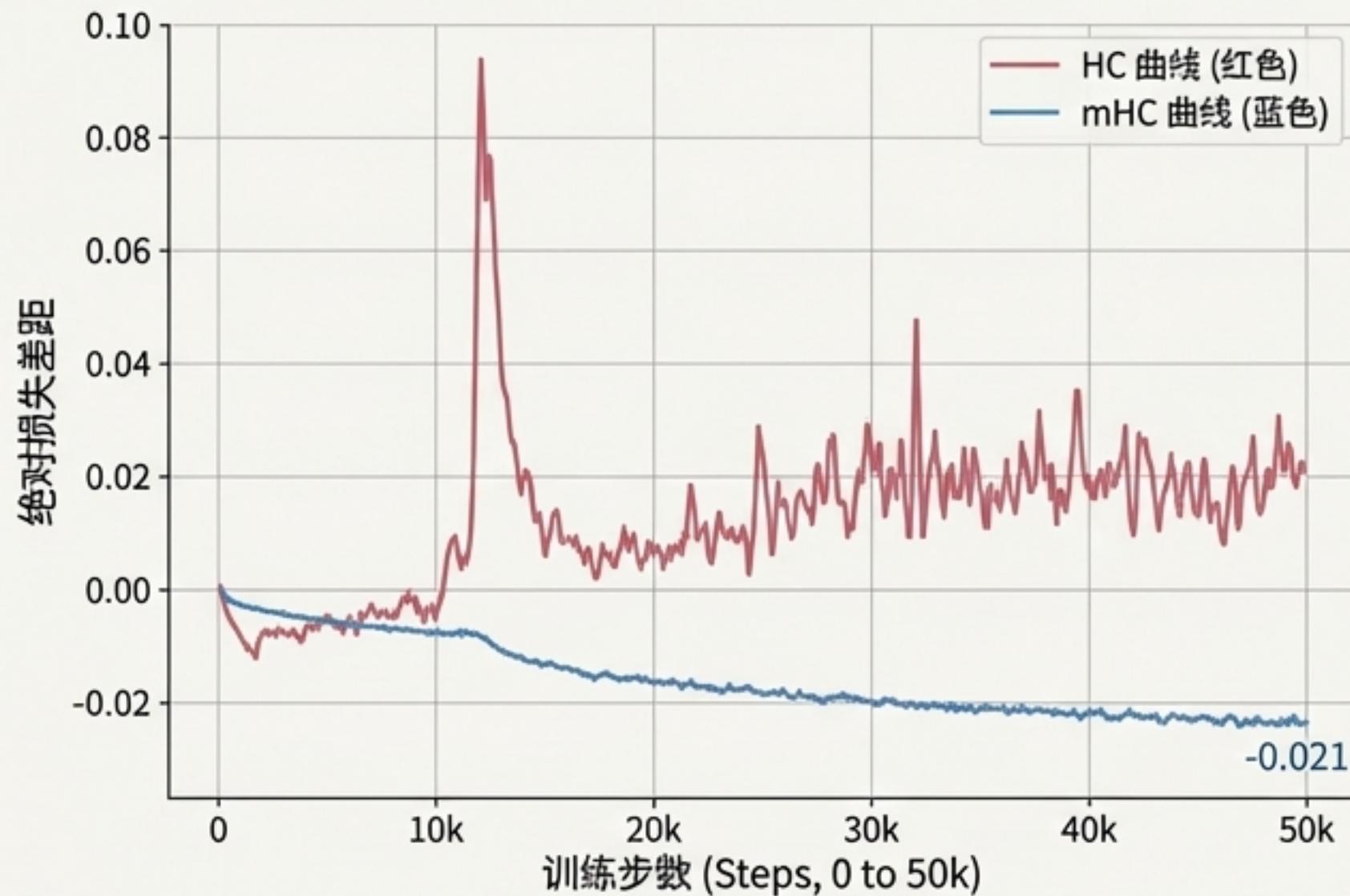
列和: [1.00, 1.00, 1.00, 1.00]

注解：所有元素非负，行/列和严格为1。变换过程可被理解为稳定的信息凸组合与融合，即使在多层复合后依然保持此特性。

标题：大规模训练验证：mHC在27B模型上实现稳定且高效的收敛

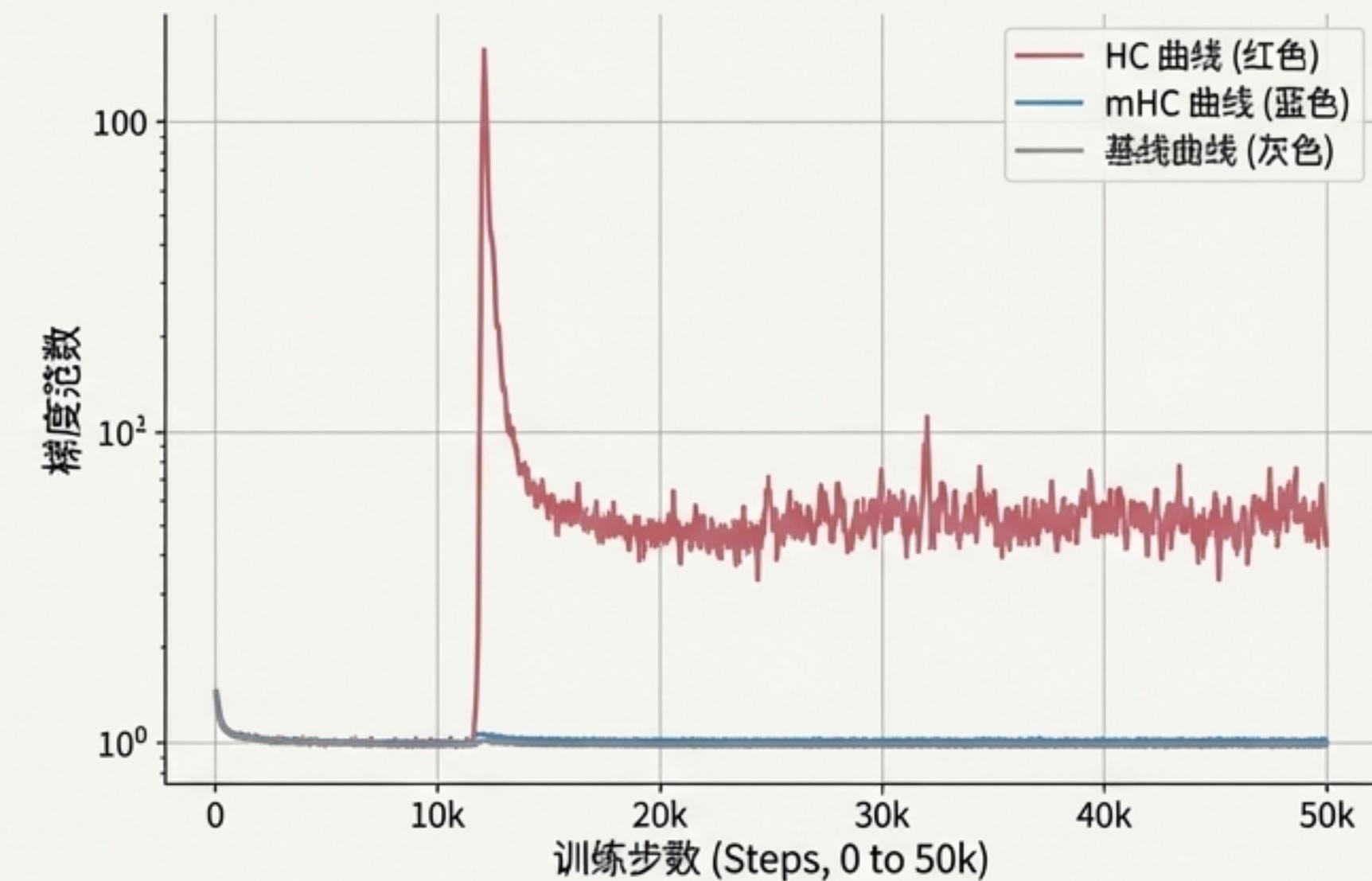
核心信息：理论上的稳定性优势直接转化为实际训练中的性能提升和过程平稳。

左侧：绝对训练损失差距 (Absolute Training Loss Gap vs. Baseline)



注解：mHC避免了HC的训练崩溃，并实现了持续、稳定的性能改进，最终收敛到比基线更优的状态。

右侧：梯度范数 (Gradient Norm)



注解：mHC的梯度行为与健康的基线模型一致，证明其反向传播过程稳定可控。

在27B参数规模上，mHC不仅解决了稳定性问题，还带来了0.021的最终训练损失降低，证明了其在大规模场景下的有效性。

标题：下游任务全面超越：mHC在关键基准测试中取得领先

核心信息：mHC带来的模型质量提升在多种下游任务中得到了验证，尤其在复杂的推理任务上表现突出。

Benchmark	Metric	#Shots	27B Baseline	27B w/ HC	27B w/ mHC
BBH	Acc	3	46.7	48.9	51.0
DROP	F1	3	50.3	51.6	53.9
GSM8K	Acc	8	41.3	42.0	44.1
HellaSwag	Acc	10	79.0	78.5	79.8
MATH	Acc	4	10.6	11.2	11.8
MMLU	Acc	5	59.0	58.2	59.5
PIQA	Acc	0	81.4	81.0	82.1
TriviaQA	Acc	64	63.2	62.8	64.0

综合优势：

mHC在所有八项基准测试中均超越了基线模型，并在大多数任务上优于不稳定的HC。

推理能力增强：

- 在复杂推理任务BBH上，mHC比HC提升了2.1% (51.0 vs 48.9)。
- 在需要阅读理解和数值推理的DROP任务上，mHC比HC提升了2.3% (53.9 vs 51.6)。

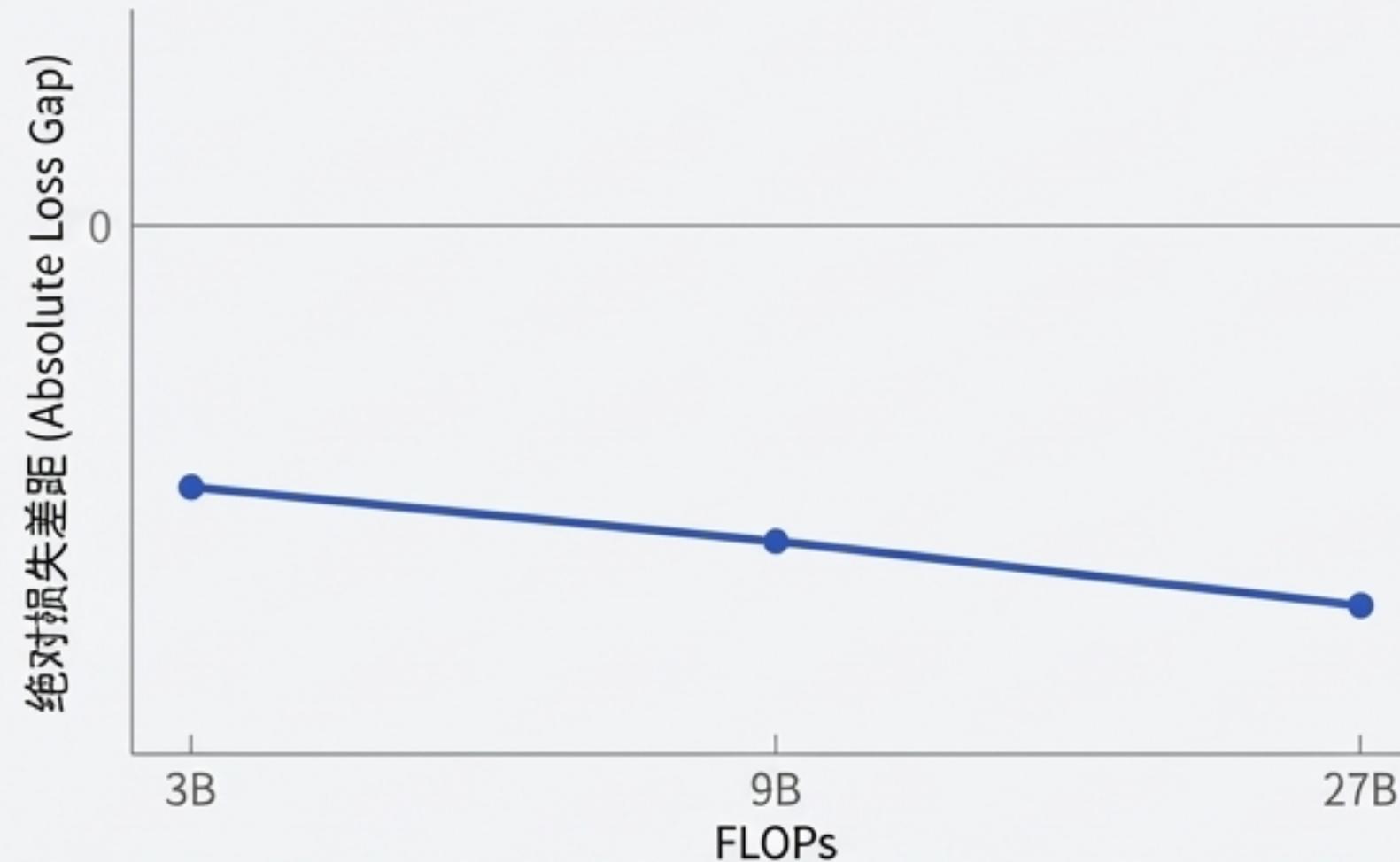
结论：

这些结果表明，mHC的稳定性不仅避免了训练失败，还使得模型能够学习到更强大、更泛化的能力。

标题：可扩展性证明：mHC的性能优势随计算和数据规模持续保持

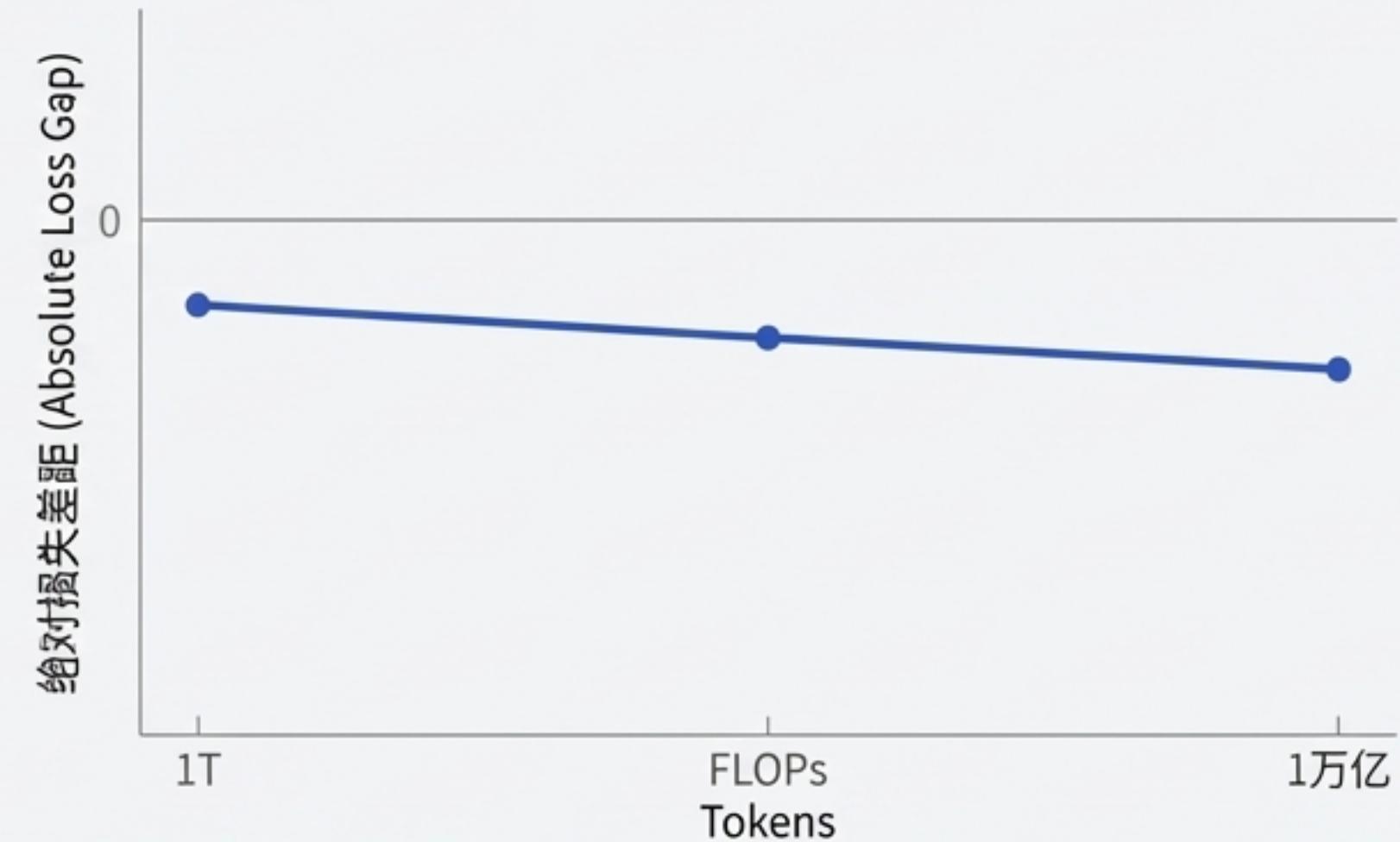
核心信息：mHC的优势并非局限于特定规模，而是在模型和数据两个关键维度上都表现出强大的可扩展性。

左侧：计算扩展曲线（Compute Scaling Curve）



模型规模扩展：从3B扩展到27B，mHC相较于基线的性能优势始终存在，仅有轻微衰减，证明其架构优势在大模型上依然稳固。

右侧：数据扩展曲线（Token Scaling Curve）



训练数据扩展：在长达1万亿Tokens的训练过程中，mHC的优势并未随着训练的深入而减弱，证明了其在长时间训练中的有效性。

总结（页面底部）：无论是在更大的模型上，还是在更长的训练中，mHC都展现出稳定、持续的性能增益。

标题：为效率而设计：通过底层优化将额外训练开销降至6.7%

核心挑战

- n 倍宽度的残差流带来了巨大的**内存访问 (I/O) 开销**和**显存占用**。
- 额外的计算和通信可能成为大规模并行训练的瓶颈。

核函数融合 (Kernel Fusion)



将多个连续的、共享内存访问的操作（如 RMSNorm, 线性投影, Sigmoid, Sinkhorn-Knopp迭代）融合成一个或少数几个定制的 CUDA核。

效果

大幅减少内存带宽瓶颈和核函数启动开销。我们使用TileLang高效实现了复杂计算的融合。

重计算 (Recomputing)



在前向传播后丢弃mHC引入的中间激活值，在反向传播时按需重新计算。通过优化重计算的块大小 L_r^- 来最小化总显存占用。

效果

显著降低了训练过程中的峰值显存占用，使得更大规模的训练成为可能。

通信-计算重叠 (Overlapping Communication)



扩展DualPipe调度策略，为mHC引入的额外计算设置高优先级计算流，确保流水线并行中的通信气泡最小化。

效果

有效隐藏了跨节点通信和重计算带来的延迟，保持了高训练吞吐量。

通过系统性的软硬件协同优化，mHC ($n=4$)
在大规模训练中仅引入了 6.7% 的额外时间开销。

标题：结论与展望：mHC为基础模型的拓扑结构设计开辟了新方向

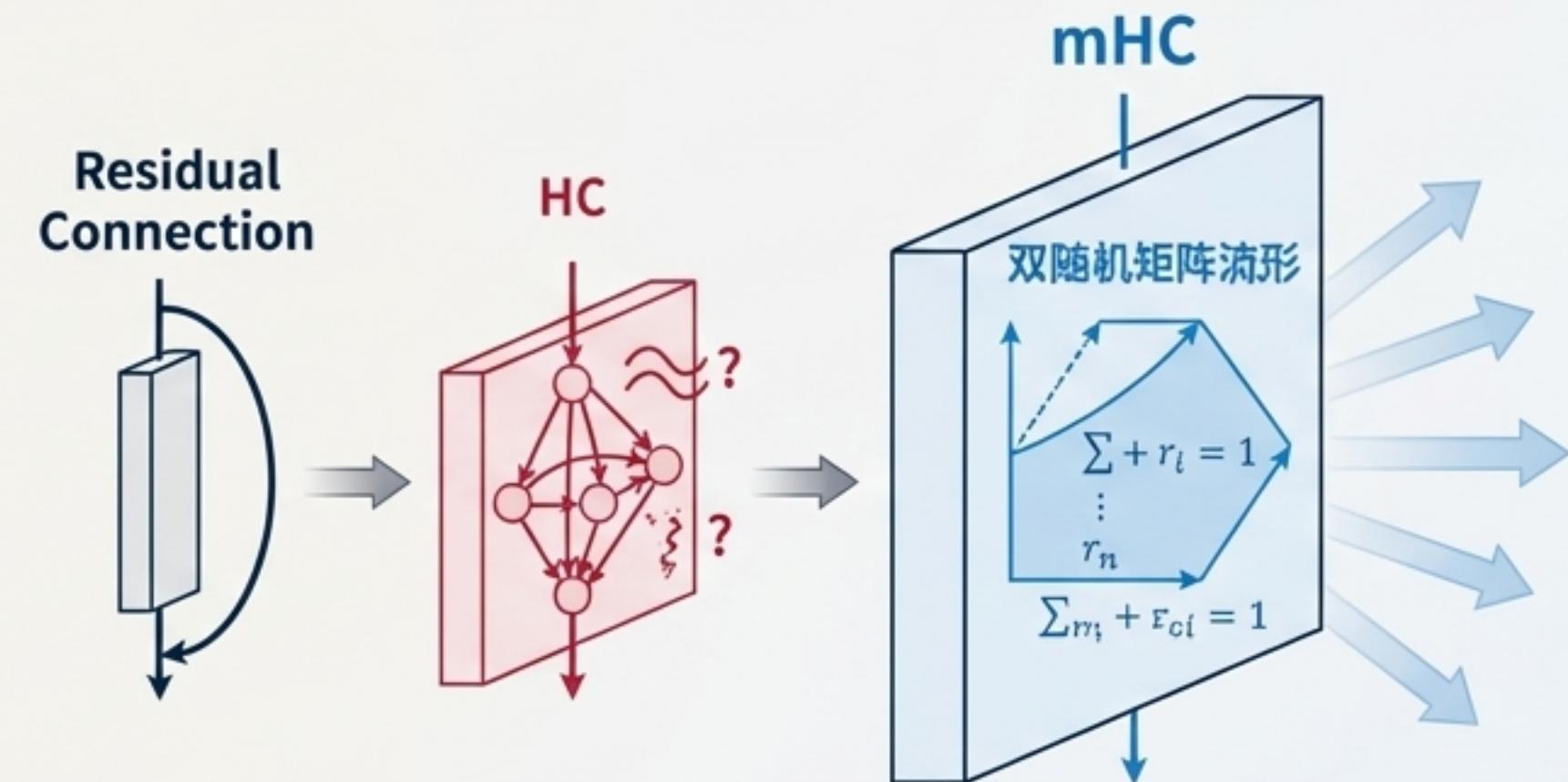
问题与诊断 (The Problem)

- 超连接 (HC) 虽有潜力，但其无约束的设计破坏了恒等映射，导致了大规模训练中的信号传播不稳定问题。

解决方案与验证 (The Solution)

- mHC 通过将连接矩阵投影到双随机矩阵流形 (Birkhoff 多面体)，从理论上恢复了信号传播的稳定性。
- 实验证明，mHC不仅解决了不稳定性，还在大规模训练、下游任务和可扩展性方面展现出全面优势。
- 通过底层系统优化，mHC的额外开销被控制在极低水平 (6.7%)，具备高度实用性。

Residual Connection -> HC -> mHC



展望与贡献 (The Outlook)

- mHC证明了在不显著增加计算复杂度 (FLOPs) 的前提下，通过优化网络拓扑结构是提升模型性能的有效途径。通过优化网络拓扑结构，提升模型稳定性。
- 它为“残差流宽度”这一新的模型缩放维度提供了稳定且可行的实现方案。
- 我们相信，mHC所代表的“通过数学约束引导架构设计”的思想，将为未来更强大、更高效的基础模型的演进提供重要启示。