DEEPSEEK-AI RESEARCH

Manifold-Constrained Hyper-Connections

Enhancing Stability and Scalability in Neural Architecture Design

mHC
Framework

27B
Model Scale

6.7%
Time Overhead

NAVIGATION

Research Overview

01

Introduction & Motivation

Evolution from residual connections to Hyper-Connections, identifying critical stability and scalability challenges in large-scale training

02

Background: Residual Connection Paradigms

Comparative analysis of Standard Residual Connections, Hyper-Connections, and the proposed Manifold-Constrained approach

03

Problem Analysis: Instability & Overhead

Deep dive into numerical instability caused by unconstrained mappings and system-level memory access bottlenecks

04

Proposed Solution: mHC Framework

Manifold projection onto Birkhoff polytope using Sinkhorn-Knopp algorithm to restore identity mapping properties

05

Methodology & Implementation

Parameterization process, efficient infrastructure design with kernel fusion, recomputing, and DualPipe optimization strategies

06

Experimental Validation

Comprehensive results across model scales (3B/9B/27B), scaling analysis, stability metrics, and downstream benchmark performance

Research Team	Focus Area	Impact
DeepSeek-AI	Macro-Architecture Design	Enhanced Scalability

ARCHITECTURAL EVOLUTION

The Evolution of Residual Connections

Standard Residual

He et al., 2016

Mathematical Formulation

$$x^{(l+1)} = x^l + F(x^l, W^l)$$

Key Properties

- Identity mapping preservation
- Stable signal propagation
- Widely adopted in LLMs

Recursive Extension

$$x^L = x^l + \Sigma F(x^i, W^i)$$

Foundation of Modern Deep Learning

Hyper-Connections

Zhu et al., 2024

Mathematical Formulation

$$x^{(l+1)} = H^{res}_l x^l + H^{post}_l F(H^{pre}_l x^l, W^l)$$

Key Features

- Expands width: C → n×C
- Three learnable mappings
- Preserves FLOPs efficiency
- Enhanced topology

Mappings

$H^{res}_l \in R^{n \times n}$

$H^{pre}_l \in R^{1 \times n}$

$H^{post}_l \in R^{1 \times n}$

Instability at Scale

mHC (Ours)

This Work

Core Innovation

$$H^{res}_l \leftarrow P_{Mres}(H^{res}_l)$$

Key Properties

- Doubly stochastic constraint
- Identity mapping restoration
- Norm preservation (≤1)
- Compositional closure

Benefits

✓ Stable propagation

✓ Scalable training

✓ 6.7% overhead

Best of Both Worlds

Key Insight

HC decouples information capacity from layer input dimension, but unconstrained mappings compromise stability. mHC restores identity mapping via manifold projection.

CRITICAL ANALYSIS

Problems with Hyper-Connections

Numerical Instability

Root Cause

Unconstrained Hˆres_l mappings compromise identity mapping across layers:

$$x^L = (\Pi H^{res}_{L-i}) x^l + \Sigma (\Pi H^{res}_{L-j}) H^{post}_i F(...)$$

Signal Behavior

- Unbounded amplification
- Vanishing gradients
- Loss of feature mean

Empirical Evidence

Loss surge at 12k steps

Gradient norm spikes

Training destabilization

Amax Gain Magnitude **~3000**

Extreme values confirm exploding residual streams (vs. ideal value of 1)

System Overhead

Memory Wall Problem

I/O costs become primary bottleneck in modern architectures:

Memory Access Increase	~n×
Communication Cost	n× Larger

I/O Overhead Breakdown

Calculate Mappings	$2nC + n^2 + 2n$
Residual Merge	$2nC$
Total I/O	$(5n+1)C + n^2 + 2n$

Impact on Training

- Degraded throughput
- Larger pipeline bubbles
- Increased GPU memory

LITERATURE REVIEW

Related Work: Micro vs. Macro Design

Micro Design

Definition

Internal architecture of computational blocks specifying how features are processed across spatial, temporal, and channel dimensions

Convolutional Variants

- Depthwise separable convolutions (Chollet, 2017)
- Grouped convolutions (Xie et al., 2017)
- Parameter sharing & translation invariance

Attention Mechanisms

MQA: Multi-Query Attention (Shazeer, 2019)

GQA: Grouped-Query Attention (Ainslie et al., 2023)

MLA: Multi-Head Latent Attention (Liu et al., 2024)

Feed-Forward Networks

MoE: Mixture-of-Experts (Fedus et al., 2022)

Enables massive parameter scaling without proportional computational costs

Macro Design

Definition

Inter-block topological structure dictating how feature representations are propagated, routed, and merged across distinct layers

Early Approaches

ResNet	Identity
DenseNet	Dense
FractalNet	Multi-path
DLA	Recursive

Recent Stream Expansion

HC: Hyper-Connections (Zhu et al., 2024)

RMT: Residual Matrix Transformer (Mak & Flanigan, 2025)

MUDDFormer: Multiway Dynamic Dense (Xiao et al., 2025)

CORE METHODOLOGY

mHC Core Innovation: Manifold Projection

Theoretical Foundation

Inspired by identity mapping principle, mHC constrains Hˆres_l onto the **Birkhoff polytope manifold** using doubly stochastic matrices. This preserves stability while enabling information exchange.

Doubly Stochastic Constraint

$$P_{Mres}(H^{res}_l) = \{ H^{res}_l \in R^{(n\times n)} \mid H^{res}_l 1_n = 1_n,\ 1_n^T H^{res}_l = 1_n^T,\ H^{res}_l \geqslant 0 \}$$

1	1	≥0
Row Sum	Col Sum	Entries

Additional Constraints

Non-negativity constraints on input/output mappings:

- $H^{pre}_l \geqslant 0$ — Prevents signal cancellation
- $H^{post}_l \geqslant 0$ — Special manifold projection

Three Key Properties

1 Norm Preservation

Spectral norm ≤ 1, effectively mitigating gradient explosion

$\|H^{res}_l\|_2 \leqslant 1$

2 Compositional Closure

Preserves stability throughout the entire model depth

Doubly stochastic matrices closed under multiplication

3 Geometric Interpretation

Convex combination of permutations for robust feature fusion

Birkhoff polytope = convex hull of permutation matrices

Signal Propagation

Hˆres_l xˆl functions as convex combination: conserves feature mean & regularizes signal norm

IMPLEMENTATION DETAILS

Parameterization & Manifold Projection Process

1 Input Preparation

Flatten hidden matrix:

$x_l \rightarrow \tilde{x}_l = \mathrm{vec}(x_l)$

Preserves full context

RMSNorm: Applied to last dimension for normalization

2 Mapping Generation

$\hat{H}^{pre}_l$
Dynamic + static

$\hat{H}^{post}_l$
Dynamic + static

$\hat{H}^{res}_l$
Dynamic + static

3 Constraint Application

$H^{pre}_l = \sigma(\hat{H}^{pre}_l)$

$H^{post}_l = 2\sigma(\hat{H}^{post}_l)$

$H^{res}_l = \mathrm{Sinkhorn}(\hat{H}^{res}_l)$

Sinkhorn-Knopp Algorithm

Iterative normalization ($t_{max} = 20$):

$M^{(0)} = \exp(\hat{H}^{res}_l)$

$M^{(t)} = T_r(T_c(M^{(t-1)}))$

T_r: Row Norm
Scale rows to sum 1

T_c: Col Norm
Scale cols to sum 1

Efficiency Optimizations

Mathematical Equivalence
Reordering operations maintains correctness while improving efficiency

Mixed-Precision Strategy
TF32, BF16, FP32 allocation for optimal accuracy/speed

Custom Backward Kernel
Recomputes intermediate results on-chip during backprop

Key Implementation Note
TileLang framework streamlines kernel implementation with complex calculations, fully utilizing memory bandwidth with minimal engineering.

SYSTEM OPTIMIZATION

Efficient Infrastructure Design

Kernel Fusion

RMSNorm Optimization

Reorder operations to improve efficiency

Divide-by-norm follows matrix multiplication

Three Specialized Kernels

Unified Scan Kernel
Fuses two scans on $\tilde{x}_l$

Coefficient Fusion
Lightweight ops

Sinkhorn Kernel
Single kernel implementation

Memory Bandwidth

Consolidates operations to reduce I/O bottlenecks

Recomputing

Memory Challenge

n-stream design introduces substantial memory overhead

Solution Strategy

- Discard intermediate activations
- Recompute on-the-fly in backward pass
- Store only first layer input x_{l0}

Optimal Block Size

$$L^*_r = \sqrt{(nL/(n+2))}$$

Minimizes total memory footprint

Stored Activations

Every L_r layers: x_{l0} (nC)

Every layer: $F(\ldots)$ (C), $x_l H^{pre}_l$ (C)

DualPipe Overlapping

Pipeline Challenge

n-stream incurs substantial communication latency across stages

Extended Schedule

- Overlaps communication & computation
- High-priority stream for FFN layers
- Prevents blocking communication

Key Optimizations

No persistent kernels: Prevents extended stalls

Preemption support: Flexible scheduling

Decoupled recompute: Local caching

Compute Streams

- Normal compute stream
- Communication stream
- High-priority stream

EMPIRICAL VALIDATION

Experimental Setup & Main Results

Experimental Configuration

Architecture

Base: DeepSeek-V3	Type: MoE
Expansion: n = 4	Sinkhorn Iter: t_max = 20

Model Variants

27B (Primary)	Proportional data
9B	Compute scaling
3B	Compute scaling
3B	1T tokens

Training Stability (27B)

Loss Reduction	Time Overhead
-0.021	6.7%

Key Finding

mHC achieves stable training with minimal overhead.

Downstream Performance

Benchmark Results

Benchmark	Baseline	HC	mHC
BBH	43.8	48.9	**51.0**
DROP	47.0	51.6	**53.9**
GSM8K	46.7	53.2	**53.8**
MMLU	59.0	63.0	**63.4**
PIQA	78.5	79.9	**80.5**

BBH (Reasoning)
+2.1% vs HC

DROP (Reasoning)
+2.3% vs HC

SCALING PROPERTIES

Scaling & Stability Analysis

Compute Scaling

Models: 3B → 9B → 27B parameters

Performance advantage **robustly maintained** with minimal attenuation at higher bud gets

3B	9B	27B
+1.8%	+1.5%	+1.2%

Token Scaling

3B model on fixed 1T token corpus

Within-run dynamics show **consistent advantage** throughout training

Key Finding

mHC maintains stable improvements across all training stages

Propagation Stability

Single-Layer Mapping

Forward Gain	Backward Gain
~1.0	~1.1

Slight deviation due to 20 iterations

Composite Mapping

Forward	Backward
~1.2	~1.6

Deviation increases but remains bounded

Critical Improvement

HC Max Gain ~3000 → mHC Max Gain ~1.6

3 Orders of Magnitude

Reduction in signal amplification

RESEARCH IMPACT

Key Contributions & Impact

1 Problem Identification

Critical Discovery
HC's signal divergence compromises energy conservation.

Root Cause Analysis
Unconstrained Hˆres_l → instability & scalability issues.

Empirical Evidence
Amax gain magnitude of ~3000 validates instability.

2 Theoretical Framework

Manifold Projection
Projects Hˆres_l onto Birkhoff polytope via Sinkhorn-Knopp.

Theoretical Properties
Norm preservation, closure, and geometric interpretation.

Signal Transformation
Becomes convex combination, conserving feature mean.

3 Practical Implementation

Infrastructure Optimizations
Kernel fusion, recomputing, and DualPipe overlapping.

System Efficiency
Minimal 6.7% overhead for n=4, enabling large-scale use.

Engineering Contribution
TileLang-based kernels for complex calculations.

4 Empirical Validation

Comprehensive Evaluation
Validated across multiple scales (3B/9B/27B) & benchmarks.

Scalability & Stability
Robust scaling & 3-order magnitude stability improvement.

Performance Gains
+2.1% on BBH, +2.3% on DROP vs. HC.

CONCLUSIONS & OUTLOOK

Manifold-Constrained Hyper-Connections

Key Achievements

Problem Solved

Successfully addresses HC's instability by projecting residual connections onto doubly stochastic manifold, transforming signal propagation into convex feature combinations.

Identity Mapping Restored

Restores identity mapping property, enabling stable large-scale training with negligible overhead through infrastructure optimization.

Empirical Success

Comprehensive validation across model scales (3B/9B/27B) and benchmarks demonstrates superior scalability, stability, and performance.

Future Directions

Manifold Exploration

Explore diverse manifold constraints tailored to specific learning objectives beyond doubly stochastic matrices.

Topological Impact

Deepen understanding of how topological structures influence optimization and representation learning.

Next-Gen Architectures

Rejuvenate interest in macro-architecture design to illuminate new pathways for foundational model evolution.

mHC	6.7%	3×