On the Design of Hyper-Connections: Simplification, Stability, and Scalability

SimpleHC: a research group paper discussion

Xiangdong Zhang, Xiaohan Qin, Sunan Zou, Tuo Dai, Bizhao Shi, Zhuo Xia,
Yebin Yang, Yizhou Chen, Yuliang Liu, Yu Cheng, Junchi Yan

arXiv:2611.05984v1 Submitted 4 Nov 2026

Publication not found as of 5 Nov 2026.
Paper-linked code repository returned 404 on 5 Nov 2026.

Takeaway: shared access to the expanded residual state cuts overhead while keeping mHC-level quality.

The question is which HC interactions justify their cost

Problem

Hyper-Connections add read gates, write gates, and residual-stream mixing. Which interactions improve quality, and which state accesses dominate training overhead?

- Study the three components separately on 15B-A1B MoE models.
- Keep learned read/write gates; replace mixing with identity transitions.
- Share state access across four sublayers, retaining sequential output contributions.

Source: Zhang et al., arXiv:2611.05984v1, Abstract and Sec. 1, PDF pp. 1–2.

HC reads, transforms, and writes across n residual streams

$$u_l = X_l^\top h_l^{\mathrm{pre}}, \qquad o_l = F_l(u_l),$$
$$X_{l+1} = H_l^{\mathrm{res}} X_l + h_l^{\mathrm{post}} o_l^\top.$$

- $X_l \in \mathbb{R}^{n \times C}$: n residual streams of width C; $u_l, o_l \in \mathbb{R}^C$.
- Read/write gates select and distribute information; H_l^{res} mixes old streams.
- mHC conditions gates on the full nC state and approximately constrains mixing to be doubly stochastic using Sinkhorn normalization.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 2, PDF pp. 2–3.

Two interaction groups reach mHC-level quality on 15B

15B-A1B design	Pile loss ↓	5-task mean ↑
mHC: learned gates + Sinkhorn mixing	1.659	55.0
identityHC: learned gates + identity	1.660	55.4
Fixed gates + unconstrained mixing	1.660	56.0
Learned gates + unconstrained mixing	1.657	56.2

Design implication

Learned read/write gating and unconstrained mixing each work as an interaction group. SimpleHC keeps the gating group to avoid repeated residual-matrix products.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 5.2, Table 2, PDF p. 7.

Identity transitions avoid cumulative mixing paths

$$c_{j,k} = (h_j^{\mathrm{pre}})^\top \underbrace{H_{j-1}^{\mathrm{res}} \cdots H_{k+1}^{\mathrm{res}}}_{\Phi_{k+1\to j}} h_k^{\mathrm{post}}.$$

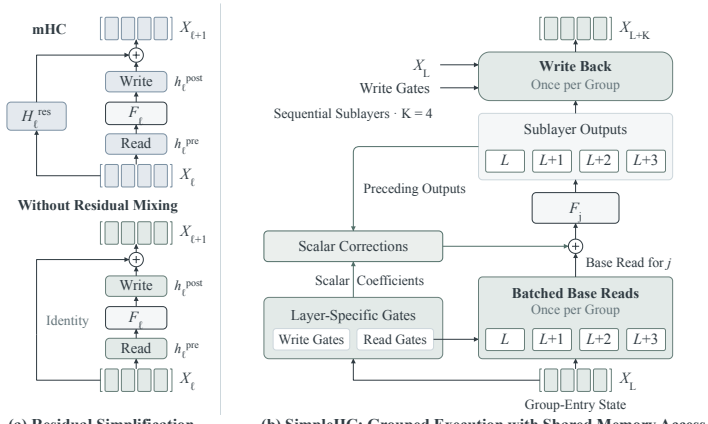
- $c_{j,k}$ weights an earlier output o_k in a later input u_j.
- With $H^{\mathrm{res}} = I$, the path coefficient is just the read/write gate inner product.
- Unconstrained mixing showed larger sampled peak coefficients; Sinkhorn mixing offered little mean-score gain over identity in the 15B study.

Choice

Use full-state, learned read/write gates with identity residual transitions.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 6.1 and App. C–D, PDF pp. 10, 19–20.

SimpleHC simplifies structure and groups execution



Original Fig. 2, PDF p. 3. The right panel keeps sublayers sequential but reads and writes the wide state at group boundaries.

Layer-specific gates share one group-entry input

$$\bar{h}_j^{\mathrm{pre}} = g_j^{\mathrm{pre}}(X_L), \qquad \bar{h}_j^{\mathrm{post}} = g_j^{\mathrm{post}}(X_L), \quad L \le j < L + K.$$

- Default $K = 4$: two Transformer blocks, four attention/FFN sublayers.
- Each sublayer has independent read/write parameters; all gate mappings see X_L.
- This changes gate conditioning from the current X_j to X_L. It makes all gates available before sequential sublayer execution.

The sole approximation

Shared-input gates replace per-sublayer-state gates; output reconstruction below is exact for the chosen shared-input gates.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 3.2, Eqs. (3)–(4), PDF pp. 4–5.

Scalar corrections preserve earlier sublayer outputs

$$u_j = X_L^\top \bar{h}_j^{\mathrm{pre}} + \sum_{k=L}^{j-1} d_{j,k} o_k,$$
$$d_{j,k} = (\bar{h}_j^{\mathrm{pre}})^\top \bar{h}_k^{\mathrm{post}}.$$

- Batch the K base reads $X_L^\top \bar{h}_j^{\mathrm{pre}}$ in one scan of X_L.
- Each $d_{j,k}$ is a scalar; its product with $o_k \in \mathbb{R}^C$ brings the earlier output into layer j.
- Execute $F_L, \ldots, F_{L+K-1}$ sequentially, then write all updates once:

$$X_{L+K} = X_L + \sum_{k=L}^{L+K-1} \bar{h}_k^{\mathrm{post}} o_k^\top.$$

Source: Zhang et al., arXiv:2611.05984v1, Sec. 3.2, Algorithm 1 and Eqs. (5)–(6), PDF pp. 4–5.

Grouped execution equals sequential execution for fixed shared gates

Induction over sublayers

At group entry, both schedules read the same X_L. If prior outputs agree, expanding sequential residual additions gives exactly the base read plus scalar corrections; the next outputs agree as well.

- The group-boundary write equals the sum of all sequential writes.
- Gradients agree for the same function if the dependence of gates on X_L is retained.
- Floating-point reduction order may change values; this result does not equate shared-input gates with the original $g_j(X_j)$.

Source: Zhang et al., arXiv:2611.05984v1, App. B, PDF pp. 18–19.

Grouping halves logical residual-state traffic at $n = K = 4$

$$Q_{\mathrm{SimpleHC}}(K) = \frac{4nC}{K} + 2C + \frac{C(K-1)(K+4)}{2K}.$$

Residual operations, per token per sublayer	Logical traffic
Fused mHC / identityHC, $n = 4$	$18C$
SimpleHC, $n = K = 4$	$9C$

- Wide gate reads, base reads, and write-back are amortized over K layers.
- The correction term grows with K; $K = 4$ is the lowest among tested $K = 2, 4, 8, 16$.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 4.3, Table 1 and Eqs. (9)–(10), PDF pp. 5–6; logical forward activation traffic only.

Specialized kernels turn fewer accesses into lower step time

Forward

- Jointly project K read/write gates and compute RMS statistics.
- Batch base reads; fuse scalar corrections.
- Write the expanded state once per group.

Backward

- Reuse one wide gradient scan across K outputs.
- Fuse projection and RMS input gradients.
- Pack small gate reductions into tensor-core products.

At $n = K = 4$, one $4C \times 32$ projection supplies eight layer-specific gate vectors.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 4.1–4.2, PDF pp. 5–6; implementations use DeepGEMM and Triton.

Quality and efficiency use distinct experimental setups

Component study

15B-A1B MoE, $n = 4$.
Pile test loss and five downstream tasks.

Quality at scale

30B-A2B MoE, same residual variants.
Pile test loss and 14 downstream tasks.

Training-step efficiency

Separate eight-layer models,
$C = 1024$–6144.
End-to-end step overhead vs. Vanilla.

Stress stability

230B-A6B, 64 layers, $10\times$ reference LR.
Gradient and loss over 28k steps.

Quality backbones use MLA and 8192-token sequences; main training uses Muon/AdamW.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 5.1 and App. A/E, PDF pp. 7, 16–18, 21–22.

At $K = 4$, independent gates and scalar corrections protect quality

15B-A1B configuration	Pile loss ↓	5-task mean ↑
mHC, per-sublayer gates	1.659	55.0
identityHC, per-sublayer gates	1.660	55.4
SimpleHC, $K = 4$, independent gates	1.660	55.0
Shared input, $K = 8$	1.662	54.3
$K = 4$, one reused gate pair	1.667	53.5
$K = 4$, no scalar corrections	1.705	47.9

Layer-specific gates and scalar corrections both matter for quality.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 5.3, Table 3, PDF p. 8; no-correction variant still writes all outputs at group exit.

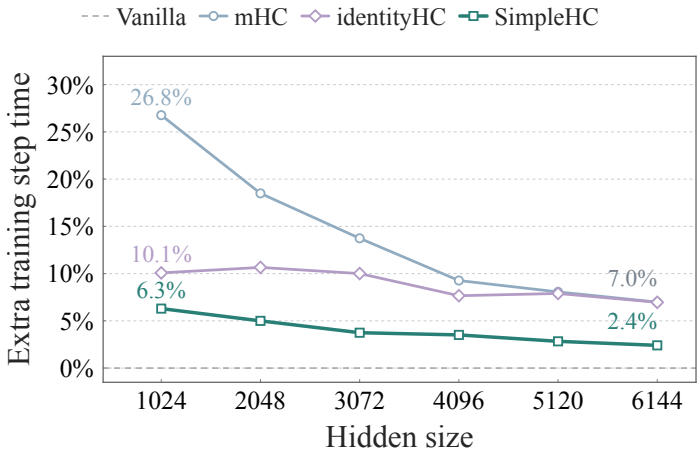
At 30B, SimpleHC matches mHC on aggregate quality

30B-A2B metric	Vanilla	mHC	identityHC	SimpleHC
Pile loss ↓	1.579	1.561	1.561	1.561
14-task mean ↑	59.3	62.9	62.8	63.1
MATH ↑	38.4	38.4	38.9	41.5
HumanEval ↑	39.6	47.6	48.2	45.1
C3 ↑	88.1	89.6	89.6	91.4

- Mean score is comparable; individual tasks have different winners.
- The result supports scaling quality from 15B to the tested 30B MoE setup.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 5.4, Table 4, PDF p. 9.

SimpleHC adds only 2.4% step time at the largest tested width



$C = 6144$

SimpleHC: 2.4%

mHC: about 7.0%

identityHC: about 7.0%

Separate eight-layer benchmark, efficient implementations.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 5.5, original Fig. 3(a), PDF p. 9; relative to Vanilla training step time.

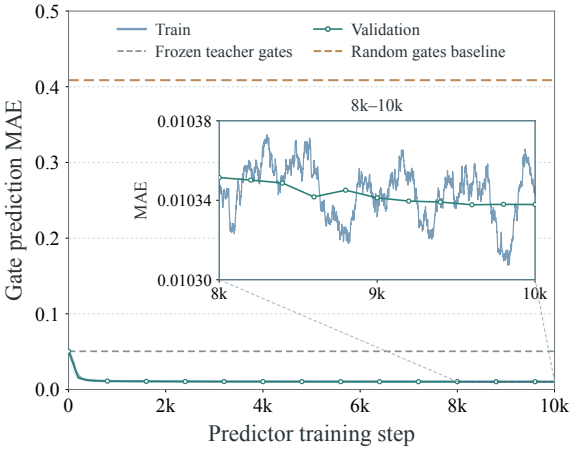
Saved activation storage falls by about 31% in the paper's accounting

Eight layers, $C = 2048$, 8192 tokens	mHC	SimpleHC
Retained storage for backward	3044.5 MiB	2092.875 MiB
Relative reduction	31.26%	

- One group-entry wide state is retained rather than one entry per sublayer.
- Counts sublayer inputs/outputs, residual copies, base reads, and HC auxiliary tensors.
- Excludes attention/FFN internals, optimizer/model states, workspaces, and allocator reservations: this is not peak GPU memory.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 5.5 and App. A.2, Table 7, PDF pp. 9, 17–18; BF16, microbatch 1, no recomputation.

Gates differ by sublayer yet are predictable from X_L



Write-gate MAE

Trained predictor: 0.01034

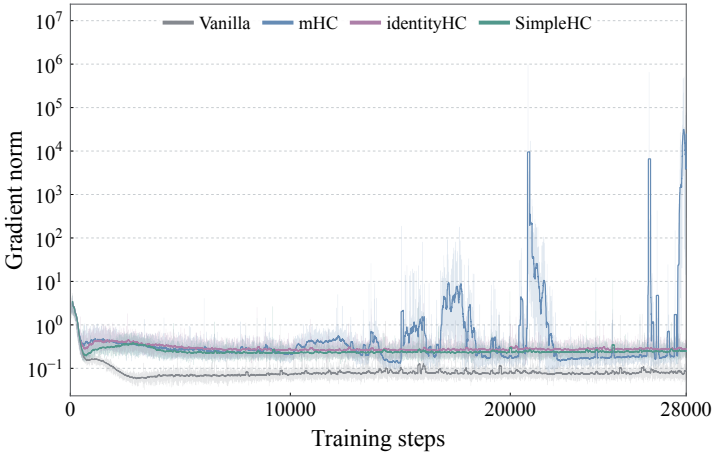
Frozen mapping: 0.05044

Constant gate: 0.06893

Frozen 15B identityHC teacher; only independent, layer-specific predictors are trained.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 6.2, original Fig. 5(b), PDF p. 11; test values from App. D.3, PDF p. 21.

SimpleHC stays stable in a combined 230B-A6B stress test



Combined stress

64 layers

512 routed experts

10× reference LR

mHC spikes reach about 1.9×10^6; SimpleHC and identityHC stay bounded across 28k steps.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 6.3, original Fig. 6(b), PDF p. 11; setting and diagnostics in App. E, PDF pp. 21–23.

The main benefit comes from shared access, with clear scope

1. Learned read/write gates with identity transitions provide mHC-level quality in the tested 15B and 30B MoE models.
2. Shared-input gates plus exact scalar corrections reduce logical residual traffic from $18C$ to $9C$ at $n = K = 4$.
3. Specialized execution yields 2.4% extra step time at $C = 6144$ in an eight-layer benchmark.

Interpret the evidence at its measured scope

The timing benchmark is eight layers; the 31% storage figure is a selected activation accounting; the stress test combines several pressures.

Source: Zhang et al., arXiv:2611.05984v1, Secs. 3–6 and App. A/E, PDF pp. 3–12, 17–18, 21–23.

Paper and status links

- Xiangdong Zhang et al. *On the Design of Hyper-Connections: Simplification, Stability, and Scalability*. arXiv:2611.05984v1, submitted 4 Nov 2026.
- Versioned paper PDF and author TeX source.
- Publication not found as of 5 Nov 2026; the paper-linked code repository returned 404 on that date.

Discussion prompt

Can group-entry conditioning remain accurate for larger K or a different residual architecture without reintroducing frequent wide-state reads?

Backup: fixed gates expose the role of each interaction group

15B-A1B configuration	Pile loss ↓	5-task mean ↑
identityHC	1.660	55.4
Fixed read gate, learned write gate	1.676	53.6
Learned read gate, fixed write gate	1.678	52.8
Fixed gates + Sinkhorn mixing	1.678	52.5
Fixed gates + unconstrained mixing	1.660	56.0

Fixed read weights are $1/n$ and fixed write weights are 1; all comparisons use the table's stated gate inputs and mixing type.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 5.2, Table 2, PDF p. 7.

Backup: the stress observations do not isolate one cause

- In one recorded event, an FFN branch output became very large before write gating; a finite-iteration Sinkhorn matrix later had row sums $(2, 1, 0.5, 0.5)$ and could further amplify it.
- A separate backward spike showed a strong input-dependent mixing-gradient path. The paper leaves mathematical sensitivity versus numerical backward effects unresolved.
- In standard 15B training, maximum post-warmup 1,000-step rolling gradient-norm standard deviation was 0.032 for SimpleHC, versus 0.206 for mHC and 0.248 for identityHC.

Source: Zhang et al., arXiv:2611.05984v1, Sec. 6.3 and App. E.2–E.3, PDF pp. 11, 22–23.