ITO-HC: IDENTITY-CENTERED TANH OFF-DIAGONAL HYPER-CONNECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyper-Connections (HC) widen the residual state and learn dynamic read, write, and stream-mixing maps. Manifold-Constrained Hyper-Connections (mHC) stabilize these maps by constraining the residual mixer to the Birkhoff polytope, but repeated nonnegative mixing can suppress stream-difference directions and finite Sinkhorn iterations introduce projection error. We propose *ITO-HC*, an identity-centered, input-independent residual mixer that adds bounded, signed off-diagonal corrections while preserving an exact unit diagonal. ITO-HC retains the mHC-style dynamic pre/post routers and nonlinear Transformer branches, combining an explicit residual highway with controlled cross-stream interactions. We derive a layer-attention formulation that exposes the low-rank dense routing structure of the identity-residual baseline, and we establish conditional perturbation and spectral stability bounds for ITO-HC. Experiments on OpenWebText with GPT-2-style models compare ITO-HC with identity, mHC, mHC-lite, and a single-stream residual baseline across multiple stream widths. The results support the proposed stability–capacity trade-off while highlighting the need for further scale and routing-diagnostic studies.

1 INTRODUCTION

Residual connections provide a short, well-conditioned identity path through depth (He et al., 2016a). Hyper-Connections (HC) generalize this path by maintaining n residual streams and learning how a branch reads from and writes to those streams (Zhu et al., 2024). A layer can be written as

$$X_{l+1} = H_l^{\text{res}} X_l + H_l^{\text{post}} f_l(H_l^{\text{pre}} X_l; W_l), \tag{1}$$

where $X_l \in \mathbb{R}^{n\times C}$, $H_l^{\text{pre}} \in \mathbb{R}^{1\times n}$, $H_l^{\text{post}} \in \mathbb{R}^{n\times 1}$, and $H_l^{\text{res}} \in \mathbb{R}^{n\times n}$. The additional streams create paths for feature reuse across depth, but an unconstrained product of residual mixers can amplify, attenuate, or rotate the residual state. mHC addresses this issue by projecting dynamic residual matrices onto the Birkhoff polytope with Sinkhorn–Knopp iterations (Xie et al., 2025). mHC-lite replaces iterative projection with an exact convex-combination parameterization of permutation matrices (Yang & Gao, 2026). KromHC uses a Kronecker parameterization to reduce the residual-matrix cost (Zhou et al., 2026). More recently, Spectral-Sphere-Constrained Hyper-Connections (sHC) identify limitations of the Birkhoff constraint, including nonnegative convex mixing and parameterization inefficiency, and allow signed mixing under affine and spectral constraints (Liu et al., 2026). We use *iHC* for the identity-residual baseline $H_l^{\text{res}} = I$, which removes only residual-path mixing while retaining the learned pre/post routers and the nonlinear branch (Guo et al., 2026).

We ask whether a fully learned residual mixer is necessary when the pre/post routers and the branch f_l remain learned. We use the following bounded, identity-centered parameterization:

$$H_l^{\text{res}} = I + \gamma_l(\mathbf{1}\mathbf{1}^\top - I) \odot \tanh(S_l), \tag{2}$$

where $S_l = b_l^{\text{res}}$ is a learned layer-specific static residual-logit matrix and $\gamma_l > 0$ controls the maximum off-diagonal magnitude. We call this parameterization ITO-HC. It preserves the self highway, permits signed cross-stream corrections, and avoids a depth-long chain of strictly positive averages.

Our contributions are: we formulate ITO-HC and connect it to the layer-attention view of HC; characterize the rank-n dense channel kernel induced by iHC while explaining why the causal mask

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and dynamic routers prevent it from fully characterizing network expressivity; give perturbation and spectral-norm conditions for the identity-centered residual path and compare its scaling with dynamic mHC and mHC-lite; and evaluate the method on OpenWebText with GPT-2-style models over $n \in \{4, 8, 16, 32, 64\}$.

2 Background and Implementation

2.1 Related work

Identity shortcuts are central to deep residual networks (He et al., 2016a;b), while Highway Networks and DenseNet explore gated or densely reused cross-layer pathways (Srivastava et al., 2015; Huang et al., 2017). In Transformers, recent alternatives enlarge or diversify the residual state through dual residual paths, matrix-valued residual streams, multi-rate mixtures, or parallel latent streams (Xie et al., 2023; Mak & Flanigan, 2025; Bhendawade et al., 2025; Liu et al., 2025).

HC broadens the residual pathway with multiple streams and learned read, write, and residual maps (Zhu et al., 2024). mHC constrains the residual mixer to the Birkhoff polytope of doubly stochastic matrices using Sinkhorn–Knopp projection (Xie et al., 2025); mHC-lite and KromHC replace this iterative projection with exact permutation-mixture and Kronecker parameterizations, respectively (Yang & Gao, 2026; Zhou et al., 2026). The iHC baseline fixes the residual mixer to I while retaining learned pre/post routing. Orthogonal Hyper-Connections (oHC) instead place the mixer on $SO(n)$ to preserve all singular values, and use the term iHC for the identity-mixer baseline (Guo et al., 2026). sHC argues that nonnegativity can limit signed feature interactions and that Birkhoff-constrained mixers may degenerate toward identity (Liu et al., 2026).

2.2 Multi-stream residual computation

At layer l, the branch input and output are

$$z_l = H_l^{\mathrm{pre}} X_l, \qquad y_l = f_l(z_l; W_l), \qquad X_{l+1} = H_l^{\mathrm{res}} X_l + H_l^{\mathrm{post}} y_l. \tag{3}$$

For the generic HC/mHC parameterization, each read, write, and residual logit is the sum of a learned input-independent term and an input-dependent linear projection of the flattened, RMS-normalized stream state, scaled by a small learned gate (Zhang & Sennrich, 2019; Zhu et al., 2024; Xie et al., 2025). ITO-HC retains this dynamic parameterization for the pre/post routers but disables the dynamic residual generator, so $S_l = b_l^{\mathrm{res}}$ is static. The iHC baseline uses the same pre/post routers with $H_l^{\mathrm{res}} = I$. This convention is distinct from the static/static and mixed static/dynamic ablations in Section 5.2. Each branch f_l is one Transformer sublayer: LayerNorm followed by causal self-attention, or LayerNorm followed by a two-layer MLP with GELU activation (Ba et al., 2016; Vaswani et al., 2017; Hendrycks & Gimpel, 2016).

For clarity, the one-fraction, one-input-view implementation can be written in the following compact form. Let $\widehat{x}_l = \mathrm{vec}(X_l) \in \mathbb{R}^{dn}$ and $\bar{x}_l = \mathrm{RMSNorm}(\widehat{x}_l)$. Here b_l^{pre}, b_l^{post}, and b_l^{res} denote the static logits, while $W_l^{\mathrm{pre}}, W_l^{\mathrm{post}}, W_l^{\mathrm{res}}$ are the corresponding dynamic projections; the α terms are learned scalar gates. We use a compact notation in which implementation-specific scalar gates (including the separate post-router gate) and static offsets are absorbed into the displayed α and b symbols. The mHC baseline is

$$\begin{aligned}
H_{l,\mathrm{mHC}}^{\mathrm{pre}} &= \mathrm{sigmoid}(\alpha_l^{\mathrm{pre}} \bar{x}_l W_l^{\mathrm{pre}} + b_l^{\mathrm{pre}})\,, \\
H_{l,\mathrm{mHC}}^{\mathrm{post}} &= 2\ \mathrm{sigmoid}\big(\alpha_l^{\mathrm{post}} \bar{x}_l W_l^{\mathrm{post}} + b_l^{\mathrm{post}}\big)\,, \\
H_{l,\mathrm{mHC}}^{\mathrm{res}} &= \mathrm{SK}_K(\exp(\alpha_l^{\mathrm{res}} \bar{x}_l W_l^{\mathrm{res}} + b_l^{\mathrm{res}}))\,,
\end{aligned} \tag{4}$$

where $W_l^{\mathrm{pre}}, W_l^{\mathrm{post}} \in \mathbb{R}^{dn \times n}$, $W_l^{\mathrm{res}} \in \mathbb{R}^{dn \times n^2}$, and SK_K denotes K Sinkhorn–Knopp iterations followed by reshaping to $n \times n$. In mHC-lite, $H_l^{\mathrm{res}} = \sum_{r=1}^{n!} \pi_{l,r} P_r$, where $\pi_l = \mathrm{softmax}(u_l)$ and P_r ranges over the $n \times n$ permutation matrices; unlike mHC's finite-step Sinkhorn projection, this convex-mixture parameterization is exactly doubly stochastic by construction (Yang & Gao, 2026). KromHC instead sets $H_l^{\mathrm{res}} = \bigotimes_{r=1}^{K} D_{l,r}$, where each $D_{l,r} \in \mathcal{B}_{i_r}$ is a smaller doubly stochastic factor and $n = \prod_{r=1}^{K} i_r$; this avoids the full $n!$ permutation basis and Sinkhorn iterations

Table 1: Per-hyper-connection scaling with stream width n and feature width d.

Method	Additional parameters	Residual construction	Main limitation
mHC	$O(dn^3)$	dynamic + Sinkhorn	cubic generator; iterative projection
mHC-lite	$O(dn\,n!)$	permutation mixture	factorial basis/materialization
KromHC	$O\Big(dn^2 + dn\prod_{r=1}^{K} i_r!\Big)$	Kronecker DS factors	mixed mode factorial term; quadratic only for binary modes
ITO-HC	$O(dn^2)$	static bounded off diagonal residual	no exact DS guarantee

but restricts the mixer to a Kronecker-structured family (Zhou et al., 2026). ITO-HC keeps the same dynamic read/write routers but replaces the Sinkhorn residual with a static identity-centered signed correction:

$$\begin{aligned} H^{\mathrm{pre}}_{l,\mathrm{ITO}} &= \mathrm{sigmoid}(\alpha^{\mathrm{pre}}_l \bar{x}_l W^{\mathrm{pre}}_l + b^{\mathrm{pre}}_l)\,, \\ H^{\mathrm{post}}_{l,\mathrm{ITO}} &= 2\ \mathrm{sigmoid}\big(\alpha^{\mathrm{post}}_l \bar{x}_l W^{\mathrm{post}}_l + b^{\mathrm{post}}_l\big)\,, \\ S_l = b^{\mathrm{res}}_l, &\qquad H^{\mathrm{res}}_{l,\mathrm{ITO}} = I + \gamma_l(\mathbf{1}\mathbf{1}^\top - I) \odot \tanh(S_l). \end{aligned} \tag{5}$$

Thus only the pre/post routers are dynamic, and Eq. 5 is Eq. 10.

The implementation expands the initial stream and reduces the final streams with the repository's 4-mean operator. Thus the language-model head keeps the baseline dimension while the internal router width still scales with n.

With one fraction and one input view, the dynamic pre/post generators each have shape $(dn) \times n$, giving $2dn^2$ parameters per hyper-connection. A fully dynamic mHC residual generator adds an $(dn) \times n^2$ block, or $O(dn^3)$ parameters. ITO-HC disables that generator and retains a static $n \times n$ residual parameterization plus the bounded scale γ_l.

2.3 Scaling with the number of streams

The parameter-count derivations are collected in Appendix A.6; here we summarize only their scaling implications and the main-model comparison. The stream width n is the central scalability variable. Table 1 summarizes the parameter and construction costs: ITO-HC retains dynamic pre/post routers but uses a static, bounded off-diagonal residual correction, yielding quadratic router scaling without a dynamic cubic residual generator. Relative to KromHC, ITO-HC's advantage is primarily simplicity and a smaller constant, not necessarily a better asymptotic order; KromHC requires tensorization and restricts the residual-matrix family.

The routing FLOPs show the same distinction. Dynamic mHC adds $O(dn^3) + O(Kn^2)$ work for coefficient generation and Sinkhorn projection, while materialized mHC-lite adds factorial terms. KromHC has $O(dn^2 + dn\prod_r i_r!)$ parameter-generation cost. ITO-HC uses dynamic pre/post routers but a static residual matrix and elementwise $\tanh$, removing the dynamic cubic residual block, factorial basis, and iterative projection; the attention and feed-forward costs are unchanged.

For the repository's large model, $d = 1024$ and $L = 24$, yielding $2L = 48$ hyper-connections. The shared GPT backbone has $354{,}871{,}296$ trainable parameters, including token and position embeddings. Thus, the full-model count is

$$N^{(m)}_{\mathrm{total}}(n) = 354{,}871{,}296 + 48N_m(n). \tag{6}$$

2.4 Residual-mixer stability analysis

HC parameterizes H^{res}_l as a learnable static matrix, initialized to I, plus a dynamic term: $\tanh$ of a linear projection of the normalized stream state, scaled by a small learnable factor (Zhu et al., 2024). The $\tanh$ bounds only the dynamic term; the static matrix, including its diagonal, is unconstrained.

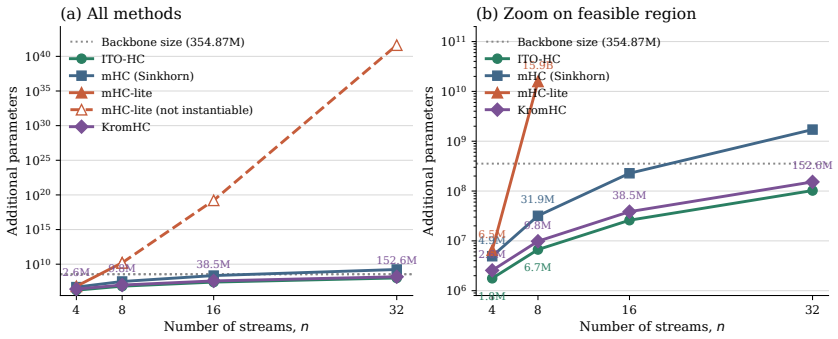

Figure 1: Total trainable parameters for the 24-layer, $d = 1024$ large model, including the shared backbone; both panels use logarithmic axes. Panel (a) compares ITO-HC, mHC, binary-mode KromHC, and mHC-lite; labels above the $n \geq 8$ mHC-lite bars show absolute counts. Panel (b) enlarges the practical range and labels totals in millions. KromHC uses binary factors $i_r = 2$ for the tested widths $n = 2^K$. Non-trainable buffers are excluded.

Xie et al. (2025) report that this compromises the identity-mapping property and permits unbounded amplification or attenuation of the residual signal.

mHC instead constrains H_l^{res} to the Birkhoff polytope of doubly stochastic (DS) matrices,

$$\mathcal{B}_n = \{H \geq 0 : H\mathbf{1} = \mathbf{1},\ \mathbf{1}^\top H = \mathbf{1}^\top\}. \tag{7}$$

mHC approximates the projection onto $\mathcal{B}_n$ with 20 Sinkhorn–Knopp iterations (Sinkhorn & Knopp, 1967; Xie et al., 2025). Xie et al. (2025) motivate the constraint by two properties: every DS matrix has spectral norm at most one, and products of DS matrices remain DS. They also describe repeated DS mixing as "a robust feature fusion mechanism"; the paragraph below shows that, under uniform mixing, the same repetition suppresses stream differences.

mHC-lite removes the finite-iteration error by writing H_l^{res} as a convex combination of the $n!$ permutation matrices, which is exact by the Birkhoff–von Neumann theorem (Birkhoff, 1946; von Neumann, 1953; Yang & Gao, 2026). KromHC reduces this cost with Kronecker products of smaller DS factors (Zhou et al., 2026). Liu et al. (2026) report three limitations of the Birkhoff constraint: learned mixers collapse toward the identity, non-negativity prevents subtractive corrections, and exact parameterizations require either Sinkhorn iterations or a factorial basis.

A Qwen3-scale ablation reported by Romero (2026) sets $H_l^{\text{res}} = I$ while retaining dynamic pre/post routers; the blog refers to this identity-residual baseline as iHC. We use this report as motivation rather than evidence, because its model, data volume, and implementation differ from ours. ITO-HC keeps the exact unit diagonal of iHC and adds signed off-diagonal terms bounded by γ_l. Unlike HC, it bounds the whole deviation from I, not only the dynamic part; unlike mHC, mHC-lite, and KromHC, it does not impose exact double stochasticity.

Repeated DS mixing contracts stream differences. Let $G_L = H_{L-1}^{\text{res}} \cdots H_0^{\text{res}}$ and $J_n = \mathbf{1}\mathbf{1}^\top/n$, where $H_l^{\text{res}} \in \mathcal{B}_n$ for $l = 0, \ldots, L-1$. Every doubly stochastic matrix preserves the mean direction and the stream-difference subspace $\mathbf{1}^\perp$. Define

$$\rho_l = \left\| H_l^{\text{res}}\big|_{\mathbf{1}^\perp} \right\|_2. \tag{8}$$

Then $\|G_L - J_n\|_2 \leq \prod_{l=0}^{L-1} \rho_l$. In particular, if every entry of every mixer is at least $\varepsilon > 0$, then $\rho_l \leq 1 - n\varepsilon$ and

$$\|G_L - J_n\|_2 \leq (1 - n\varepsilon)^L, \qquad \sigma_{\min}(G_L) \leq (1 - n\varepsilon)^L \longrightarrow 0. \tag{9}$$

Thus, under a common positive-mixing lower bound, the product preserves the mean direction but exponentially suppresses stream-difference directions and approaches the rank-one averaging matrix J_n. The invariant-subspace proof is given in Appendix A.7. Double stochasticity alone is not enough: a sequence of identity or permutation matrices exhibits no such decay. The bound describes how the residual path transports an earlier contribution. Each branch still writes new stream

differences through H_l^{post}. Finite Sinkhorn–Knopp iterations introduce a separate issue: the output need not be exactly doubly stochastic, so neither the invariant-subspace argument nor the contraction bound applies exactly. Classical matrix-scaling analyses show that convergence and finite-step error depend on the input matrix, while mHC-lite documents the resulting approximation gap for the fixed iteration budget used by mHC (Linial et al., 1998; Franklin & Lorenz, 1989; Knight, 2008; Chakrabarty & Khanna, 2021; Yang & Gao, 2026).

In the standard mHC construction, exponentiating finite logits gives a strictly positive matrix before Sinkhorn, and Sinkhorn preserves positivity. Under exact double-stochastic constraints, this implies $\rho_l < 1$ at each layer. The implementation uses a finite number of Sinkhorn iterations, however, so the constraints are approximate; positivity alone does not establish the exact invariant-subspace bound for the implemented matrix. A depth-uniform contraction rate in the ideal doubly stochastic setting still requires a uniform lower bound on entries.

This gives the central stability–expressivity trade-off. The mHC constraint prevents residual-path amplification, but repeated positive mixing attenuates stream-difference directions; as G_L approaches J_n, information transported by the residual path becomes increasingly concentrated in the mean direction and difficult to recover. New differences can still be written by $H_l^{\text{post}} y_l$, but previously transported differences are not preserved. ITO-HC gives up the exact doubly stochastic guarantee in exchange for an identity-centered, signed perturbation: under the cumulative bounds in Eq. 22, it avoids an architectural drive toward J_n while retaining a nonzero residual-path singular spectrum. The checkpoint comparison in Section 5.3 illustrates this distinction at finite depth.

3 ITO-HC

3.1 Parameterization

For the proposed ITO-HC configuration, the residual logits are input-independent: $S_l = b_l^{\text{res}} \in \mathbb{R}^{n\times n}$ is a learned matrix specific to layer l. Define $M = \mathbf{1}\mathbf{1}^\top - I$ and

$$E_l = \gamma_l M \odot \tanh(S_l), \qquad H_l^{\text{res}} = I + E_l. \tag{10}$$

The diagonal is exactly one and $|E_{l,ij}| < \gamma_l$ for $i \neq j$. In our implementation, γ_l is represented by an exponential log-scale, initialized to 0.05 in the ITO-HC configuration and learned independently at each layer. There is no input-dependent residual-logit generator in ITO-HC; only H_l^{pre} and H_l^{post} retain the dynamic generator. Pure identity residuals preserve each stream's direct self path but provide no residual-path communication between streams. ITO-HC retains this unit diagonal and adds bounded off-diagonal interactions: positive entries can strengthen a cross-stream contribution, while negative entries permit subtractive or canceling corrections. The $\tanh$ nonlinearity maps arbitrary logits to $(-1, 1)$, so γ_l bounds each off-diagonal deviation from identity; this elementwise bound alone is not a global stability guarantee. The design therefore adds controlled signed communication without restricting the mixer to nonnegative convex averaging, while making no claim that it remains doubly stochastic or preserves the stream mean. As shown in Appendix A.2, with fixed routers these mixers transport the branch-write directions, whose accumulated contribution remains low rank, while the total change from the initial state also includes a separate residual-transport term.

Unlike a Sinkhorn matrix, Eq. 10 is neither necessarily nonnegative nor doubly stochastic. Exact preservation of the mean direction would additionally require $E_l\mathbf{1} = 0$ and $\mathbf{1}^\top E_l = 0$. We use the identity-centered signed perturbation as a stability–expressivity trade-off and measure violations of these constraints rather than assuming them away.

3.2 Layer-attention derivation

We derive the scalar influence of an earlier branch on a later branch input by expanding the residual recursion (the full expansion is given in Appendix A.1). For $i < j$,

$$z_j = \sum_{i<j} a_{ji} y_i, \qquad a_{ji} = H_j^{\text{pre}} \left(\prod_{k=i+1}^{j-1} H_k^{\text{res}} \right) H_i^{\text{post}}. \tag{11}$$

Stacking the branch outputs gives $Z = AY$; under the source/target indexing, A is strictly lower triangular. For iHC,

$$a_{ji} = H_j^{\text{pre}} H_i^{\text{post}}. \tag{12}$$

Let $p_j = (H_j^{\text{pre}})^\top \in \mathbb{R}^n$ and $q_i = H_i^{\text{post}} \in \mathbb{R}^n$, and define $P = [p_1^\top; \ldots; p_T^\top]$ and $Q = [q_1^\top; \ldots; q_T^\top]$. The resulting unmasked channel-strength matrix is

$$C = PQ^\top, \qquad \operatorname{rank}(C) \le n. \tag{13}$$

Equation 13 characterizes only the *dense, unmasked scalar routing kernel*. When $n < T$, it cannot represent a target matrix whose rank exceeds n; this is a useful capacity diagnostic, not a rank theorem for the full Transformer.

The actual causal layer matrix is instead

$$A = L \odot (PQ^\top), \qquad L_{ji} = \mathbb{1}[i < j]. \tag{14}$$

The rank bound does not generally survive the Hadamard mask. With shifted causal indexing, the matrix may even have a nonzero diagonal and be full rank. Thus $n < T$ restricts only the unmasked completion C; it does not prove that A, or the full nonlinear Transformer, is rank deficient.

For ITO-HC, the residual transitions add signed path corrections to the factorized identity kernel. The detailed expansion and the equivalent gauge factorization are deferred to Appendix A.1. That calculation also makes the limitation explicit: whenever the residual transitions are invertible, they can be absorbed into effective read/write factors, so ITO-HC does not claim an unconditional full-rank dense kernel. Its intended benefit is a controlled, signed state transformation before later nonlinear branches and dynamic routers, while retaining the identity highway. With fixed read/write vectors, the residual mixer transports each branch write to a different stream direction without removing the low-rank factorization of the accumulated writes. The total change relative to the initial state has an additional transport term, however; the precise decomposition and its perturbative expansion are given in Appendix A.2.

3.3 Role of dynamic pre/post routing

The pre and post routers act on different factors of the stream–feature representation. Under the identity-residual simplification, let $X_l \in \mathbb{R}^{n \times C}$ and write

$$a_l(X_l) = H_l^{\text{pre}}(X_l) \in \mathbb{R}^{1 \times n}, \qquad b_l(X_l) = H_l^{\text{post}}(X_l) \in \mathbb{R}^{n \times 1}. \tag{15}$$

The branch update is

$$\Phi_l(X_l) = X_l + b_l(X_l) y_l, \qquad y_l = f_l(a_l(X_l) X_l)\,, \tag{16}$$

where $y_l \in \mathbb{R}^{1 \times C}$. Setting $H_l^{\text{res}} = I$ leaves the nonlinear Transformer branch intact; it does not replace f_l by an identity map.

With iHC, the cumulative update over L branches factorizes as

$$\begin{aligned} X_L - X_0 &= B_L(X) Y_L(X), \\ \operatorname{rank}(X_L - X_0) &\le \min\{n, C, L\}. \end{aligned} \tag{17}$$

Here $B_L \in \mathbb{R}^{n \times L}$ collects post directions and $Y_L \in \mathbb{R}^{L \times C}$ collects branch features. This is a fixed-input, per-token stream–feature statement, not a claim about the rank of the full sequence-level attention Jacobian. The detailed matrix-space and differential analysis is deferred to Appendix A.3.

Fixed-input geometry. For a fixed input, the relaxed output envelope is $\mathcal{S}_X = \{B_L(X)Y : Y \in \mathbb{R}^{L \times C}\}$, with dimension $\operatorname{rank}(B_L(X))C \le LC$. Hence the rank ceiling in Eq. 17 concerns one input and one token; the reachable set is further restricted by the nonlinear branches and state recursion. The full matrix-space characterization and determinantal-variety dimension are given in Appendix A.3.

Static post routing fixes the stream-side basis across inputs, whereas dynamic post routing can change it; dynamic pre routing changes the feature-side coefficients even when the post basis is fixed. This distinction is the mechanistic reason to retain both routers, and is tested by the ablation in Section 5.2.

Dynamic routing and input-conditioned spaces. With dynamic post routing, the stream-side column space $\mathrm{col}(B_L(X))$ becomes input dependent; with dynamic pre routing, the feature-side coefficients are input dependent. The model therefore realizes a family of low-dimensional subspaces rather than one fixed routing basis. Its full Grassmannian/global-span description is deferred to Appendix A.3; only the per-input rank ceiling in Eq. 17 is claimed here.

Dynamic pre routing changes the feature-side factor through the derivative of the input-dependent read, while dynamic post routing can add stream directions through the derivative of $B_L(X)$. The complete local differential and single-branch Jacobian are provided in Appendix A.3.

The local output differential is the sum of a dynamic-post term and a dynamic-pre term,

$$D(B_L Y_L)\,[U] = (DB_L[U])\,Y_L + B_L\,(DY_L[U])\,, \tag{18}$$

so the two routers provide complementary routes to input-conditioned representations. The complete Jacobian chain rule and cross-layer expression are given in Appendix A.3. Dynamic routing still leaves $\mathrm{rank}(C(X)) \le n$ for every fixed input; it supplies a family of low-rank kernels rather than removing the individual-input bound. The static/dynamic ablation is therefore empirical evidence for these pathways, not a universal superiority theorem.

4 Stability analysis

The identity-centered form permits a direct perturbation analysis. Let $\eta_l = \|E_l\|_2$ denote the spectral size of the residual correction at layer l.

Because $|E_{l,ij}| < \gamma_l$ off diagonal and $E_{l,ii} = 0$, a conservative bound is

$$\eta_l = \|E_l\|_2 \le \|E_l\|_F \le \gamma_l \sqrt{n(n-1)}. \tag{19}$$

Therefore,

$$\|H_l^{\mathrm{res}}\|_2 \le 1 + \gamma_l \sqrt{n(n-1)}. \tag{20}$$

More importantly, Weyl's singular-value perturbation bound gives (Horn & Johnson, 2012)

$$\sigma_{\min}(I + E_l) \ge 1 - \|E_l\|_2 = 1 - \eta_l. \tag{21}$$

Consequently, if $\eta_l < 1$ for every layer, the multiplicativity inequality for the smallest singular value yields

$$\sigma_{\min}\left(\prod_{l=0}^{L-1}(I + E_l)\right) \ge \prod_{l=0}^{L-1}(1 - \eta_l) > 0. \tag{22}$$

If, more strongly, $\eta_l \le \bar{\eta} < 1$ and $\sum_l \eta_l \le B$, then $\log(1-x) \ge -x/(1-\bar{\eta})$ on $[0, \bar{\eta}]$ gives the depth-uniform lower bound

$$\sigma_{\min}\left(\prod_{l=0}^{L-1}(I + E_l)\right) \ge \exp\left(-\frac{B}{1-\bar{\eta}}\right) > 0. \tag{23}$$

The same perturbation budget controls the largest singular value:

$$\left\|\prod_{l=0}^{L-1}(I + E_l)\right\|_2 \le \prod_{l}(1 + \eta_l) \le \exp\left(\sum_l \eta_l\right). \tag{24}$$

Equations 23–24 show that, under a bounded cumulative perturbation budget, every residual-stream direction remains bounded away from zero and infinity. This differs from uniformly positive doubly stochastic mixing: ITO-HC is centered at the full-rank identity, permits signed corrections, and is not constrained to preserve the mean while contracting $\mathbf{1}^{\perp}$. It therefore has no architectural requirement to converge to the rank-one consensus matrix J_n; pure identity ($E_l = 0$) is the limiting case with every singular value exactly one.

This conclusion is a sufficient conditional guarantee, not an unconditional property of a unit diagonal. If η_l stays at a fixed nonzero value as $L \to \infty$, the lower bound $\prod_l(1-\eta_l)$ can itself vanish, and specially aligned nonsymmetric perturbations can still make the product ill-conditioned. Accordingly, ITO-HC avoids the *forced positive-averaging mechanism* behind Eq. 9, but it must keep the learned cumulative perturbation controlled to guarantee non-vanishing singular values. From Eq. 19, a directly checkable sufficient condition is $\gamma_l \sqrt{n(n-1)} < 1$ at every layer together with a bounded sum of these quantities. This motivates monitoring both η_l, the cumulative-product singular spectrum, and the learned γ_l during training.

5 EXPERIMENTS

We pretrain decoder-only GPT-2-style models (Radford et al., 2019) on OpenWebText (Gokaslan & Cohen, 2019) with the nanoGPT-derived repository (Karpathy, 2022). The S/M/L configurations have $(6, 8, 512)$, $(12, 12, 768)$, and $(24, 16, 1024)$ layers/heads/width. Runs use length 1024, micro-batch 2, accumulation 128, bfloat16 AdamW (Loshchilov & Hutter, 2019), weight decay 0.1, gradient clipping 1.0, and 10,000 iterations; learning rates are $10^{-3}, 6 \times 10^{-4}, 3 \times 10^{-4}$ with 200 warm-up iterations and cosine decay to one-tenth.

For identity, ITO-HC, and mHC we sweep $n \in \{4, 8, 16, 32, 64\}$. Identity fixes $H^{\text{res}} = I$; ITO-HC uses Eq. 10 with a static $S_l = b_l^{\text{res}}$, dynamic pre/post routers, and γ_l initialized to 0.05; mHC uses its standard dynamic residual parameterization and Sinkhorn projection. We also include the single-stream residual baseline and mHC-lite at $n = 4$. All multi-stream comparisons use the repository's 4-mean reducer and repeated initial expansion. Table 2 reports the currently available losses; missing entries remain explicit rather than being imputed.

Table 2: Available last-iteration OpenWebText losses after 10,000 iterations.

Dataset		OpenWebText					
Method	Streams n	S		M		L	
		Train	Val	Train	Val	Train	Val
iHC	4	3.4270	3.3833	3.1648	3.1348	3.0473	3.0842
	8	3.3822	3.3755	3.1501	3.1216	3.0493	3.0861
	16	3.3992	3.3753	3.1414	3.1128	3.0344	3.0712
	32	3.4200	3.3754	3.1355	3.1085	3.0290	3.0635
	64	3.3757	3.3607	3.1296	3.1017	3.0265	3.0606
ITO-HC	4	3.3933	3.3765	3.1589	3.1293	3.0461	3.081
	8	3.3699	3.3659	3.1487	3.1190	3.0463	3.0827
	16	3.3855	3.3684	3.1361	3.1085	3.0288	3.0657
	32	3.3774	3.3612	3.1234	3.0982	3.0187	3.0559
	64	3.3711	3.3573	3.1147	3.0884	2.9926	3.0310
mHC	4	3.3964	3.3817	3.1554	3.1472	3.0562	3.0916
mHC-lite	4	3.3931	3.3776	3.1623	3.1338	3.0553	3.0898
KromHC	4	3.3916	3.3785	3.1581	3.1507	3.0486	3.0823
Residual	1	3.4177	3.4598	3.2086	3.2240	3.0934	3.1277

5.1 WHY SCALE THE NUMBER OF STREAMS?

The prior mHC and mHC-lite studies use $n = 4$, while KromHC also tests larger widths (Xie et al., 2025; Yang & Gao, 2026; Zhou et al., 2026). Multiple streams provide up to $\min(n, L)$ fixed-input stream directions; the normalized 4-mean keeps output amplitudes comparable. Details, including the $n > L$ qualification, are in Appendix A.5.

5.2 ABLATING DYNAMIC PRE/POST ROUTING

Small-model static/dynamic ablations and wide static-routing stress tests are in Appendix A.8; they are diagnostic, not parameter- or memory-matched scaling experiments. The complementary residual- mixer product diagnostic appears below.

The S/M/L backbone configurations and their parameter counts are listed in Appendix 3; Figure 1 reports large-model totals including Hyper-Connection parameters.

5.3 RESIDUAL-MIXER PRODUCT DIAGNOSTIC

To test the contraction mechanism in an actual forward pass, we captured the post-Sinkhorn residual mixers of the trained $n = 4$ large mHC checkpoint after combining static and input-dependent logits and applying the same 20 Sinkhorn iterations. We sampled two length-1024 OpenWebText validation sequences and report sample 0, token 1023. The 48-mixer chain and annotated matrices are shown in Figure 2. Every mHC row remains normalized, but product rows become similar

mixtures of the four streams. Its spectrum is

$$\sigma(G_{mHC}) = (1.00876,\ 0.15560,\ 0.04360,\ 0.00427), \qquad \mathrm{cond}(G_{mHC}) = 236.0. \tag{25}$$

The smallest direction retains about 0.4%, consistent with positive-mixing contraction; this is a single-input diagnostic, not a population estimate.

The ITO-HC diagnostic uses static residual matrices (dynamic residual generator disabled) and dynamic pre/post routers. Its signed chain is the right panel of Figure 2; its cumulative product has

$$\sigma(G_{\mathrm{ITO}}) = (1.66502,\ 1.29762,\ 0.90939,\ 0.50104), \qquad \mathrm{cond}(G_{\mathrm{ITO}}) = 3.32, \tag{26}$$

The minimum remains substantial, unlike mHC: signed identity-centered corrections avoid the same consensus collapse. This finite- depth comparison illustrates, but does not prove, the conditional result.

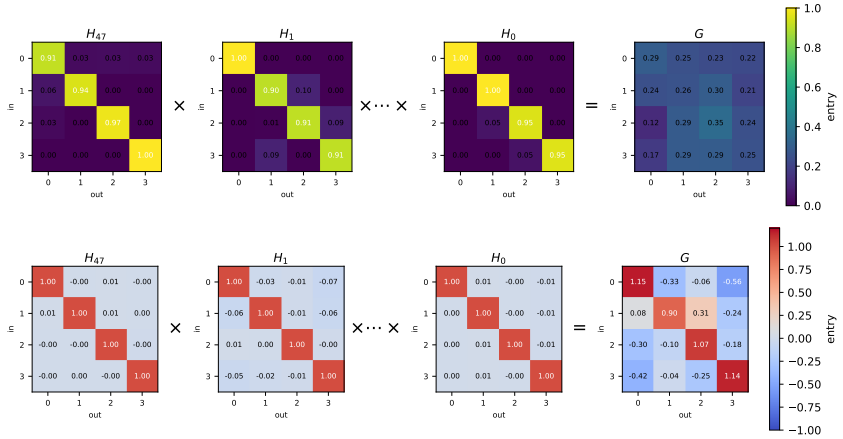

Figure 2: Residual-mixer products for sample 0, token 1023 from the large $n = 4$ checkpoints. Top: mHC after 20 Sinkhorn iterations; bottom: ITO-HC with static residual mixers and dynamic pre/post routers. Each panel shows $H_{47} \times \cdots \times H_1 \times H_0 = G$ across all 48 forward mixers;

6 Discussion and Limitations

The rank bound in Eq. 13 concerns a dense scalar kernel, not the full Transformer; causal masking, input-dependent routers, and branch Jacobians add further behavior. ITO-HC has conditional, perturbation-norm stability rather than exact double stochasticity, and signed corrections may change the stream mean. Larger-scale training, transfer, and broader routing diagnostics remain open.

7 Conclusion

We introduced ITO-HC as an identity-centered alternative to fully dynamic residual mixing. The main conclusions are:

- **Architecture.** ITO-HC preserves an exact unit diagonal while adding bounded, signed off-diagonal corrections. It therefore keeps a direct self highway and permits controlled cross-stream interactions.
- **Analysis.** For ITO-HC, perturbation analysis gives conditional lower and upper bounds on the residual-path singular spectrum under a controlled cumulative perturbation budget.
- **Empirical comparison.** OpenWebText experiments compare ITO-HC with iHC, mHC, mHC-lite, KromHC, and a single-stream residual baseline across multiple stream widths. The results support the proposed training stability of ITO-HC, while larger-scale training and broader routing diagnostics remain important directions for future work.

AI USE STATEMENT

Generative AI tools assisted with developing and checking mathematical claims and derivation exposition, shaping checkpoint diagnostics, interpreting routing measurements, implementing diagnostic and plotting scripts, and editing and restructuring the manuscript. They were not used to generate synthetic data or to replace the reported training runs. The authors reviewed the AI-assisted text, mathematical arguments, code, and numerical results, and take responsibility for the final content and claims.

ETHICS STATEMENT

This study uses the publicly available OpenWebText corpus and involves no human participants, new collection of personal or sensitive information, or human-facing deployment. The corpus may inherit biases or other limitations from its source web data.

REPRODUCIBILITY STATEMENT

Model definitions, training settings, stream-width comparisons, and reported losses are specified in Sections 2–5. Mathematical derivations and routing diagnostics are provided in Sections 4–5 and Appendix A. Training and analysis scripts, configuration details, and plotting utilities will be supplied as anonymous supplementary material.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Nikhil Bhendawade, Mahyar Najibi, Devang Naik, and Irina Belousova. M2R2: Mixture of multi-rate residuals for efficient transformer inference, 2025.

Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Universidad Nacional de Tucuman Revista*, 5:147–151, 1946.

Deeparnab Chakrabarty and Sanjeev Khanna. Better and simpler error analysis of the Sinkhorn–Knopp algorithm for matrix scaling. *Mathematical Programming*, 188(1):395–407, 2021. doi: 10.1007/s10107-020-01503-3.

Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114–115:717–735, 1989. doi: 10.1016/0024-3795(89)90490-4.

Aaron Gokaslan and Vanya Cohen. OpenWebText corpus. Web corpus, 2019. URL `https://skylion007.github.io/OpenWebTextCorpus/`.

Haoqiang Guo, Xuyi Chen, Bo Ke, Yishu Lei, Ziyang Xu, Shikun Feng, Ximen, and Wenhan Luo. oHC: Orthogonal hyper-connections on $SO(4)$ via quaternions, 2026. URL `https://arxiv.org/abs/2609.02672`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016a. doi: 10.1109/CVPR.2016.90.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pp. 630–645. Springer, 2016b.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Andrej Karpathy. nanoGPT. GitHub repository, 2022. URL https://github.com/karpathy/nanoGPT.

Philip A. Knight. The Sinkhorn–Knopp algorithm: Convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008. doi: 10.1137/060659624.

Nathan Linial, Alex Samorodnitsky, and Avi Wigderson. A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, pp. 644–652, 1998. doi: 10.1145/276698.276880.

Houjun Liu, Shikhar Murty, Christopher D. Manning, and Róbert Csordás. Thoughtbubbles: An unsupervised method for parallel thinking in latent space, 2025.

Zhaoyi Liu, Haichuan Zhang, and Ang Li. Beyond the Birkhoff polytope: Spectral-Sphere-Constrained Hyper-Connections, 2026. URL https://arxiv.org/abs/2603.20896.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Brian Mak and Jeffrey Flanigan. Residual matrix transformers: Scaling the size of the residual stream. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

Tyler Romero. Your DeepSeek mHC might not need the "m". Tyler Romero technical blog, English translation of a Zhihu technical blog post, 2026. URL https://www.tylerromero.com/translations/zhihu/your-deepseek-mhc-might-not-need-the-m/. Accessed 2026-09-25.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

John von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. In Harold William Kuhn and Albert William Tucker (eds.), *Contributions to the Theory of Games, Volume II*, pp. 5–12. Princeton University Press, 1953. doi: 10.1515/9781400881970-002.

Shufang Xie, Huishuai Zhang, Junliang Guo, Xu Tan, Jiang Bian, Hany Hassan Awadalla, Arul Menezes, Tao Qin, and Rui Yan. ResiDual: Transformer with dual residual connections, 2023.

Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Kuai Yu, Liang Zhao, Shangyan Zhou, Zhean Xu, Zhengyan Zhang, Wangding Zeng, Shengding Hu, Yuqing Wang, Jingyang Yuan, Lean Wang, and Wenfeng Liang. mHC: Manifold-constrained hyper-connections. *arXiv preprint arXiv:2512.24880*, 2025.

Yongyi Yang and Jianyang Gao. mHC-lite: You don't need 20 Sinkhorn-Knopp iterations. *arXiv preprint arXiv:2601.05732*, 2026.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Wuyang Zhou, Yuxuan Gu, Giorgos Iacovides, and Danilo Mandic. KromHC: Manifold-constrained hyper-connections with kronecker-product residual matrices. *arXiv preprint arXiv:2601.21579*, 2026.

Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, and Xun Zhou. Hyper-Connections. *arXiv preprint arXiv:2409.19606*, 2024.

A Additional derivations and ablations

A.1 Detailed layer-attention derivation

Set $X_0 = H_0^{\text{post}} y_0$ and recursively expand Eq. 3. For $i < j$,

$$X_{j-1} = \sum_{i<j} \left(\prod_{k=i+1}^{j-1} H_k^{\text{res}} \right) H_i^{\text{post}} y_i, \tag{27}$$

which gives Eq. 11. For iHC, the transition product is the identity and Eq. 12 follows. With $p_j = (H_j^{\text{pre}})^\top$ and $q_i = H_i^{\text{post}}$, stacking these vectors gives the dense completion $C = PQ^\top$ in Eq. 13. The causal matrix is instead $A = L \odot (PQ^\top)$, so the rank bound for C does not generally transfer through the Hadamard mask.

For ITO-HC, substituting $H_k^{\text{res}} = I + E_k$ into the interval product gives the first-order path expansion

$$a_{ji} = H_j^{\text{pre}} H_i^{\text{post}} + \sum_{k=i+1}^{j-1} H_j^{\text{pre}} E_k H_i^{\text{post}} + O(E^2), \tag{28}$$

where $O(E^2)$ collects products containing at least two residual corrections. This expansion shows signed cross-stream paths, but does not by itself prove a larger matrix rank. A sharper algebraic statement follows when every H_k^{res} is invertible. Define

$$G_t = H_t^{\text{res}} H_{t-1}^{\text{res}} \cdots H_1^{\text{res}}, \qquad G_0 = I. \tag{29}$$

Then

$$\prod_{k=i+1}^{j-1} H_k^{\text{res}} = G_{j-1} G_i^{-1}, \tag{30}$$

and hence

$$a_{ji} = \underbrace{H_j^{\text{pre}} G_{j-1}}_{\widehat{p}_j^\top} \underbrace{G_i^{-1} H_i^{\text{post}}}_{\widehat{q}_i}. \tag{31}$$

Consequently, the corresponding unmasked completion remains

$$C_{\text{ITO-HC}} = \widehat{P}\widehat{Q}^\top, qquad \operatorname{rank}(C_{\text{ITO-HC}}) \le n. \tag{32}$$

Thus an invertible residual mixer is a depth-dependent change of stream coordinates in this scalar analysis. Its practical effect comes from changing the states at which nonlinear branches and dynamic routers are evaluated; it is not an unconditional full-rank guarantee.

A.2 Fixed-routing residual transport and perturbations

This subsection isolates the role of the residual mixer by fixing the read and write routers. Let $a_l = H_l^{\text{pre}}$ and $b_l = H_l^{\text{post}}$ be fixed, let $H_l = H_l^{\text{res}}$, and write

$$X_{l+1} = H_l X_l + b_l y_l, \qquad y_l = f_l(a_l X_l). \tag{33}$$

For $0 \le k \le L$, define the ordered product

$$T_{L,k} = H_{L-1} H_{L-2} \cdots H_k, \qquad T_{L,L} = I. \tag{34}$$

Repeated substitution gives

$$X_L = T_{L,0} X_0 + \sum_{i=0}^{L-1} T_{L,i+1} b_i y_i. \tag{35}$$

Define

$$B_L^{(H)} = \begin{bmatrix} T_{L,1} b_0 & T_{L,2} b_1 & \cdots & b_{L-1} \end{bmatrix}, \qquad Y_L = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{L-1} \end{bmatrix}. \tag{36}$$

The accumulated branch write relative to the transported initial state is therefore

$$X_L - T_{L,0}X_0 = B_L^{(H)}Y_L, \qquad \operatorname{rank}(B_L^{(H)}Y_L) \le \min\{n, L, C\}. \tag{37}$$

The residual mixer changes each write direction from b_i to $T_{L,i+1}b_i$, but it does not by itself increase the number of branch write directions. In contrast, the change relative to the original initial state is

$$X_L - X_0 = (T_{L,0} - I)X_0 + B_L^{(H)}Y_L. \tag{38}$$

The first term is transport of the pre-existing state and can have rank up to $\min\{n, C\}$. Thus $X_L - X_0$ has no general rank bound controlled only by L; a useful consequence of subadditivity is

$$\operatorname{rank}(X_L - X_0) \le \min\left\{n, C, \operatorname{rank}(T_{L,0} - I) + L\right\}. \tag{39}$$

For $H_l = I$, $T_{L,0} = I$ and Eq. 37 reduces to the identity-residual factorization used in the main text.

For ITO-HC, write $H_k = I + E_k$. Holding the branch outputs fixed while expanding the transport product gives

$$T_{L,i+1} = I + \sum_{k=i+1}^{L-1} E_k + \sum_{i+1 \le m < k \le L-1} E_k E_m + \cdots. \tag{40}$$

Let $\eta_k = \|E_k\|_2$ and $q_i = \sum_{k=i+1}^{L-1} \eta_k$. By spectral-norm submultiplicativity, the norm of all terms of order at least two is bounded by

$$\prod_{k=i+1}^{L-1} (1 + \eta_k) - 1 - q_i \le e^{q_i} - 1 - q_i. \tag{41}$$

Consequently, the higher-order path terms are $O(q_i^2)$ when the cumulative perturbation q_i is small. The elementwise parameter bound alone does not ensure this regime: from $|E_{k,rs}| < \gamma_k$ off diagonal one obtains only $\eta_k \le \|E_k\|_F < \gamma_k\sqrt{n(n-1)}$.

The preceding path expansion treats y_i as fixed coefficients. In the actual nonlinear recursion, changing H_l also changes later branch inputs. To make this dependence explicit, set $H_l(\epsilon) = I + \epsilon\Delta_l$ and keep a_l, b_l fixed. At $\epsilon = 0$, let $J_l = Df_l(a_lX_l)$ and define $U_l = \frac{dX_l}{d\epsilon}\big|_0$. Differentiating Eq. 33 gives

$$U_{l+1} = U_l + \Delta_l X_l + b_l J_l a_l U_l. \tag{42}$$

The direct term $\Delta_l X_l$ is residual transport, while the last term is its indirect propagation through the branch Jacobian. If $V_l = \frac{d^2X_l}{d\epsilon^2}\Big|_0$ and the branch has a bounded second derivative, then

$$V_{l+1} = V_l + 2\Delta_l U_l + b_l\left[J_l a_l V_l + D^2 f_l(a_l X_l)[a_l U_l, a_l U_l]\right]. \tag{43}$$

Thus second-order effects comprise both multiple residual corrections and the branch Hessian response to the perturbed state. Neglecting higher-order terms in the full network requires control of the cumulative η_l, branch Jacobians and Hessians, and state norms; it does not follow from the elementwise γ_l bound alone.

A.3 Detailed dynamic pre/post routing analysis

For the identity-residual analysis, write $a_l(X_l) = H_l^{\text{pre}}(X_l)$ and $b_l(X_l) = H_l^{\text{post}}(X_l)$. The exact update is Eq. 16, and the local branch expansion, when the second derivatives are bounded, is

$$f_l(\bar{z}_l + \delta z_l) = f_l(\bar{z}_l) + J_{f_l}(\bar{z}_l)\delta z_l + O(\|\delta z_l\|^2). \tag{44}$$

With iHC,

$$X_L = X_0 + \sum_{l=0}^{L-1} b_l(X_l) y_l, \tag{45}$$

and define

$$B_L(X) = \begin{bmatrix} b_0(X_0) & b_1(X_1) & \cdots & b_{L-1}(X_{L-1}) \end{bmatrix},$$
$$Y_L(X) = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{L-1} \end{bmatrix}. \tag{46}$$

Then $X_L - X_0 = B_L(X)Y_L(X)$ and the rank bound in Eq. 17 follows from the rank inequality for a matrix product. For a fixed input, the associated coefficient-relaxed envelope is

$$\mathcal{S}_X := \{B_L(X)Y : Y \in \mathbb{R}^{L\times C}\} = \{M \in \mathbb{R}^{n\times C} : \mathrm{col}(M) \subseteq U_X\}, \quad U_X := \mathrm{col}(B_L(X)), \tag{47}$$

and, with $r_X = \mathrm{rank}(B_L(X))$,

$$\dim \mathcal{S}_X = r_X C \le LC. \tag{48}$$

If $n > L$, each fixed-input update is therefore rank deficient in the stream dimension even when the feature coefficients are unrestricted. If both factors were allowed to vary freely, the largest fixed-rank envelope would be the determinantal variety

$$\mathcal{D}_L := \{M \in \mathbb{R}^{n\times C} : \mathrm{rank}(M) \le L\}, qquad \dim \mathcal{D}_L = L(n + C - L). \tag{49}$$

For static pre and post routers, $b_l(X_l) = b_l^0$ and

$$B_L(X) = B_L^0 = \begin{bmatrix} b_0^0 & \cdots & b_{L-1}^0 \end{bmatrix}, qquad U_X = U_0 := \mathrm{col}(B_L^0), \tag{50}$$

so all relaxed updates use the same space

$$\mathcal{S}_{\mathrm{static}} = \{B_L^0 Y : Y \in \mathbb{R}^{L\times C}\}, qquad \dim \mathcal{S}_{\mathrm{static}} = \mathrm{rank}(B_L^0)C \le LC. \tag{51}$$

Dynamic pre can still change the attainable coefficient matrices, but cannot create new stream-side basis directions when post routing is static.

With dynamic post routing,

$$B_L(X) = \begin{bmatrix} b_0(X_0) & \cdots & b_{L-1}(X_{L-1}) \end{bmatrix}, qquad U_X = \mathrm{col}(B_L(X)). \tag{52}$$

If its rank is locally constant, U_X is a point of $\mathrm{Gr}(r, n)$; the global stream span is

$$U_{\mathrm{global}} := \mathrm{span}_{X\in\mathcal{X}} U_X, qquad r_{\mathrm{global}} := \dim U_{\mathrm{global}} \le n, \tag{53}$$

and the span of the relaxed envelopes has dimension $r_{\mathrm{global}}C$:

$$\mathrm{span}_{X\in\mathcal{X}} \mathcal{S}_X = U_{\mathrm{global}} \otimes \mathbb{R}^C, qquad \dim(\mathrm{span}_X \mathcal{S}_X) = r_{\mathrm{global}}C. \tag{54}$$

The actual reachable outputs are generally a smaller nonlinear subset.

Dynamic pre affects the feature side through

$$Dz_l[U] = a_l(X_l)U + (Da_l(X_l)[U])X_l, \tag{55}$$

and therefore

$$\delta y_l \simeq J_{f_l}(\bar{z}_l)\left[a_l(\bar{X}_l)\delta X_l + (Da_l(\bar{X}_l)[\delta X_l])\bar{X}_l\right]. \tag{56}$$

At the cumulative level,

$$D(B_L Y_L)[U] = (DB_L[U])\, Y_L + B_L\,(DY_L[U]), \tag{57}$$

and the single-branch chain rule is

$$D\Phi_l(X_l)[U] = U + (Db_l(X_l)[U])f_l(z_l) \tag{58}$$

$$+ b_l(X_l)Df_l(z_l)[a_l(X_l)U + (Da_l(X_l)[U])X_l]. \tag{59}$$

For self-attention, Df_l is the full token-sequence Jacobian. Finally, the cross-layer read is

$$z_l = a_l(X_l)X_0 + \sum_{i<l} a_l(X_l)b_i(X_i)y_i, \tag{60}$$

and the input-conditioned channel family obeys

$$C(X) = P(X)Q(X)^\top, \mathrm{rank}(C(X)) \le n. \tag{61}$$

A.4 REPOSITORY MODEL CONFIGURATIONS

Table 3: Repository model configurations. The labels S/M/L refer to the configurations used in this paper and are not the vanilla nanoGPT/GPT-2 model-type names. Counts exclude method-specific Hyper-Connection parameters.

Configuration	Layers	Heads	Width d	Backbone parameters
S	6	8	512	45.20M
M	12	12	768	124.48M
L	24	16	1024	354.87M

A.5 STREAM-WIDTH CAPACITY AND OUTPUT-SCALE ANALYSIS

The original mHC and mHC-lite studies use $n = 4$ for their main comparisons, while KromHC additionally tests selected larger widths (Xie et al., 2025; Yang & Gao, 2026; Zhou et al., 2026). Thus $n = 4$ is a practical baseline, but controlled width scaling changes both intermediate capacity and output scale.

The capacity benefit of multiple streams is most transparent before the final stream reducer. Under the identity-residual abstraction used in Section 3.3, the cumulative branch update is

$$\Delta_L(X) := X_L - X_0 = B_L(X)Y_L(X), \qquad B_L(X) \in \mathbb{R}^{n\times L}, \quad Y_L(X) \in \mathbb{R}^{L\times C}. \tag{62}$$

For a fixed input, let $U_X = \mathrm{col}(B_L(X))$ and $r_X = \mathrm{rank}(B_L(X))$. If the feature-side coefficients are relaxed to range over $Y \in \mathbb{R}^{L\times C}$, the corresponding linear envelope is

$$\mathcal{S}_X = \{B_L(X)Y : Y \in \mathbb{R}^{L\times C}\}, \qquad \dim \mathcal{S}_X = r_X C \le \min(n, L)C. \tag{63}$$

The inequality follows because $B_L(X)$ has only L columns. Thus n controls the number of independent stream-side directions that can store and recombine branch features, while L controls the number of branch contributions available in the cumulative factorization.

The single-stream residual baseline corresponds to $n = 1$. In that case $r_X \le 1$ and, whenever the update is nonzero, $\dim \mathcal{S}_X = C$: all branch outputs are accumulated in one residual vector. For multiple streams with sufficiently independent post factors, r_X can grow to $\min(n, L)$, giving up to $\min(n, L)C$ coefficient degrees of freedom and the per-input rank bound

$$\mathrm{rank}(\Delta_L(X)) \le \min(n, C, L). \tag{64}$$

For example, two independent write directions can preserve two different branch features in two stream rows, whereas a single stream can only store their feature-space combination in one row. This is the sense in which multi-stream routing provides parallel latent memory rather than merely duplicating the final output.

The same effect appears in the layer-attention view. With $p_j = (H_j^{\mathrm{pre}})^\top$ and $q_i = H_i^{\mathrm{post}}$, the dense scalar routing kernel has the factorization $C = PQ^\top$ and satisfies $\mathrm{rank}(C) \le n$ (Eq. 13). Increasing n therefore permits more independent source–target routing patterns in the unmasked completion, while dynamic pre/post routers can select different factors for different inputs. This rank statement is a capacity diagnostic rather than a theorem about the full causal Transformer, whose mask, branch Jacobians, and state recursion add further structure.

The regime $n > L$ requires a separate qualification. In that regime, $\mathrm{rank}(\Delta_L(X)) \le L < n$, so increasing n beyond L does not continue to raise the rank ceiling of one fixed-input cumulative update. It can nevertheless enlarge the ambient residual state, the family of input-conditioned stream subspaces $\mathrm{col}(B_L(X))$, and the state trajectories seen by later nonlinear branches. Consequently, improvements for $n > L$ should be interpreted as benefits from state-space geometry and conditional routing, not as evidence that the per-input low-rank bound has disappeared.

Finally, the n-stream advantage is not contradicted by the final reducer. If the output head receives $r^\top X_L$ for a fixed reducer vector r, then the final merged update is a $1 \times C$ vector. The relevant capacity gain occurs before this merge: subsequent attention and MLP branches can read

stream differences through H_l^{pre}, transform them nonlinearly, and write new combinations through H_l^{post}. A reducer applied only after all branches therefore preserves intermediate stream separation, whereas merging after every branch would largely remove it. The normalized 4-mean controls output amplitude across widths; it does not remove intermediate-state routing degrees of freedom.

There is also an output-scale issue when comparing stream widths. The original mHC paper specifies the per-layer pre/post/residual maps, but does not explicitly specify how the final n streams are merged before the language-model head (Xie et al., 2025). In contrast, the mHC-lite implementation used for comparison merges streams by summation ("sum-up") (Yang & Gao, 2026),

$$x_{\text{out}} = \sum_{s=1}^{n} x_s, \tag{65}$$

which we refer to as the sum-up reducer. The repository's generic reducer also defaults to this unnormalized sum. If stream states have comparable scale and weakly correlated signs, the norm of this sum can grow with n (linearly in the worst aligned case), confounding capacity with output rescaling. We therefore use the normalized reducer

$$x_{\text{out}} = \frac{4}{n} \sum_{s=1}^{n} x_s, \tag{66}$$

where the factor 4 gives the $n = 4$ setting the historical sum gain while preventing output gain from growing with n. This convention is used for identity, ITO-HC, and the mHC stream-width sweep; the single-stream residual baseline uses its ordinary identity output. Identity and ITO-HC add $O(dn^2)$ router parameters per hyper-connection, whereas fully dynamic mHC adds an $O(dn^3)$ residual generator.

A.6 Parameter-count derivations

With one fraction and one input view, the exact trainable-parameter count of the main ITO-HC configuration is

$$N_{\text{ITO-HC}} = 2dn^2 + dn + n^2 + 2n + 3 = O(dn^2), \tag{67}$$

where $2dn^2$ is the dynamic pre/post-router cost, dn is the RMSNorm scale, n^2 is the static residual-logit matrix, and the remaining terms are static pre/post logits and scalar gates. The count assumes a trainable γ_l and excludes the intentionally frozen dynamic-residual scale. The dynamic H^{res} generator is disabled in the main ITO-HC configuration. A minimal pure-identity implementation has the same router cost but can omit the residual-logit matrix.

For comparison, we derive the corresponding counts from the standard mHC, mHC-lite, and KromHC parameterizations (Xie et al., 2025; Yang & Gao, 2026; Zhou et al., 2026):

$$N_{\text{mHC}} = dn^3 + (2d+1)n^2 + dn + 2n + 3, \tag{68}$$

$$N_{\text{mHC-lite}} = (dn+1)n! + 2dn^2 + dn + 2n + 3, \tag{69}$$

$$N_{\text{KromHC}} = 2dn^2 + dn + (dn+1)\prod_{r=1}^{K} i_r! + 2n + 3, \tag{70}$$

where $n = \prod_{r=1}^{K} i_r$ for KromHC. For binary modes, $\prod_r i_r! = n$; this simplification does not hold for arbitrary factors. Counts are per hyper-connection, so a depth-L Transformer with two branches per block multiplies them by $2L$.

A.7 Proof of the doubly stochastic contraction bound

Let $\mathcal{B}_n$ denote the set of doubly stochastic matrices, and use the shorthand

$$H_l := H_l^{\text{res}}, \qquad J_n := \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \qquad P := I - J_n. \tag{71}$$

The vector $\mathbf{1}$ is not normalized, whereas $u := \mathbf{1}/\sqrt{n}$ is a unit vector. Define the mean and difference subspaces

$$U := \text{span}\{\mathbf{1}\} = \text{span}\{u\}, \qquad V := \mathbf{1}^\perp = \{v \in \mathbb{R}^n : \mathbf{1}^\top v = 0\}. \tag{72}$$

Then J_n is the orthogonal projector onto U, and P is the orthogonal projector onto V.

Fix a layer l and suppose $H_l \in \mathcal{B}_n$. The two doubly stochastic identities are

$$H_l\mathbf{1} = \mathbf{1}, \qquad \mathbf{1}^\top H_l = \mathbf{1}^\top. \tag{73}$$

The first identity implies $H_l u = u$, so U is invariant. If $v \in V$, then

$$\mathbf{1}^\top(H_l v) = (\mathbf{1}^\top H_l)v = \mathbf{1}^\top v = 0,$$

so $H_l v \in V$ and V is invariant as well. Moreover,

$$H_l J_n = J_n, \qquad J_n H_l = J_n, \qquad P H_l P = H_l - J_n. \tag{74}$$

The last equality follows by expanding $(I - J_n)H_l(I - J_n)$ and using the first two identities in Eq. 74.

To make the restriction to V explicit, choose a matrix $Q \in \mathbb{R}^{n\times(n-1)}$ whose columns form an orthonormal basis of V. Thus $Q^\top Q = I_{n-1}$, $Q^\top\mathbf{1} = 0$, and $QQ^\top = P$. In the orthogonal basis $[u\ Q]$, the invariance relations give the block decomposition

$$[u\ Q]^\top H_l [u\ Q] = \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{H}_l \end{bmatrix}, \qquad \widetilde{H}_l := Q^\top H_l Q \in \mathbb{R}^{(n-1)\times(n-1)}. \tag{75}$$

The off-diagonal blocks vanish because $H_l u = u$ and $\mathbf{1}^\top H_l = \mathbf{1}^\top$. The contraction factor in the main text is therefore

$$\rho_l := \|H_l|_V\|_2 = \|\widetilde{H}_l\|_2 = \sigma_{\max}(Q^\top H_l Q). \tag{76}$$

This is the quantity that can be computed from a checkpoint: construct any orthonormal basis Q of $\mathbf{1}^\perp$, form $Q^\top H_l Q$, and take its largest singular value. Since H_l need not be symmetric, an eigenvalue calculation is not a substitute for this singular-value calculation.

Now define the depth-L residual transition

$$G_L := H_{L-1}H_{L-2}\cdots H_0. \tag{77}$$

Because U and V are invariant for every layer, the same orthogonal basis gives

$$[u\ Q]^\top G_L [u\ Q] = \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{H}_{L-1}\widetilde{H}_{L-2}\cdots\widetilde{H}_0 \end{bmatrix}. \tag{78}$$

In the same basis, J_n is $\mathrm{diag}(1, 0)$, and hence

$$\|G_L - J_n\|_2 = \left\|\widetilde{H}_{L-1}\cdots\widetilde{H}_0\right\|_2 \le \prod_{l=0}^{L-1}\|\widetilde{H}_l\|_2 = \prod_{l=0}^{L-1}\rho_l, \tag{79}$$

where the inequality is the submultiplicativity of the spectral norm. Equivalently, using Eq. 74 and the fact that a product of doubly stochastic matrices is doubly stochastic,

$$G_L - J_n = P G_L P. \tag{80}$$

This projection identity is useful for checkpoint diagnostics: for exact doubly stochastic matrices, $\|G_L - J_n\|_2$ and the norm of the product on V are the same quantity.

We next derive a uniform bound from strict entrywise positivity. Assume that every entry of every H_l satisfies $H_{l,ij} \ge \varepsilon > 0$ and that $n\varepsilon < 1$. Define

$$K_l := \frac{H_l - \varepsilon\mathbf{1}\mathbf{1}^\top}{1 - n\varepsilon}. \tag{81}$$

The numerator is entrywise nonnegative. Its row sums and column sums are both $1-n\varepsilon$, so $K_l \in \mathcal{B}_n$. Thus

$$H_l = \varepsilon\mathbf{1}\mathbf{1}^\top + (1 - n\varepsilon)K_l. \tag{82}$$

For every $v \in V$, the first term vanishes because $\mathbf{1}\mathbf{1}^\top v = 0$. Consequently,

$$H_l|_V = (1 - n\varepsilon)K_l|_V. \tag{83}$$

Since K_l is nonnegative and doubly stochastic,

$$\|K_l\|_2 \le \sqrt{\|K_l\|_1 \|K_l\|_\infty} = 1,$$

and therefore $\|K_l|_V\|_2 \le 1$. It follows that

$$\rho_l \le 1 - n\varepsilon, \qquad \|G_L - J_n\|_2 \le (1 - n\varepsilon)^L. \tag{84}$$

Finally, the block form in Eq. 78 shows that the singular values of G_L consist of 1 together with the singular values of $\widetilde{H}_{L-1} \cdots \widetilde{H}_0$. For $n \ge 2$, the smallest singular value of the restricted product is no larger than its largest singular value. Therefore

$$\sigma_{\min}(G_L) \le \|G_L - J_n\|_2 \le (1 - n\varepsilon)^L \xrightarrow[L\to\infty]{} 0. \tag{85}$$

This is a conditional result. Exact double stochasticity is needed for the invariant-subspace decomposition; finite-step Sinkhorn iterations generally produce only an approximately doubly stochastic matrix, and ITO-HC does not impose double stochasticity at all. Consequently, the same product bound cannot be applied to those matrices without separately controlling their mean/difference-subspace leakage. Also, strict positivity by itself gives only $\rho_l < 1$ layer by layer; a depth-uniform exponential rate requires one common lower bound ε for all layers.

A.8 Static/dynamic pre/post routing ablations

The four-way small-model ablation at $n = 4$ compares dynamic/dynamic, static/static, static/dynamic, and dynamic/static routing, where the first entry denotes H^{pre} and the second denotes H^{post}. These labels refer only to the pre/post routers: every ITO-HC row uses the proposed static residual logits and has the dynamic H^{res} generator disabled. Static logits remain trainable; only the corresponding state-dependent pre/post generator is removed. Other training settings are fixed. The completed ITO-HC static/static, static/dynamic, and dynamic/static runs are reported alongside the identity ablation; the standalone ITO-HC dynamic/dynamic run remains reported in the main sweep.

The $n = 70$ static run is a wider comparison, not a parameter- or memory-matched comparison: its measured throughput is 4.61×10^5 tokens/s versus 4.85×10^5 for dynamic identity at $n = 4$. The $n = 1024$ run is an extreme width-stress test for whether input-independent state can compensate for missing dynamic directions; it is not an equal-resource comparison. Its results should be read as small-model diagnostics rather than as a general scaling law.

Table 4: Small-model ablation of dynamic pre/post routing for identity and ITO-HC. "Pre/post" specifies whether each router is dynamic or static. The $n = 1024$ row is a width stress test.

Configuration	Streams n	Pre/post	Final train loss	Final val loss
iHC	4	dynamic/dynamic	3.4270	3.3833
	4	static/static	3.4289	3.4150
	4	static/dynamic	3.4110	3.3927
	4	dynamic/static	3.4185	3.4010
	70	static/static	3.4303	3.4136
	1024	static/static	3.4345	3.4168
ITO-HC	4	dynamic/dynamic	3.3933	3.3765
	4	static/static	3.4263	3.4094
	4	static/dynamic	3.4043	3.3875
	4	dynamic/static	3.4084	3.3925
	1024	static/static	3.4154	3.4075