BEYOND STABILITY: ATTENTION-TO-MLP INTERACTION IN INFINITE-DEPTH TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard Transformer blocks pass the attention output to the MLP. However, we find that the depth scaling used to stabilize residual accumulation can make this interaction *vanish* as depth increases, even though the architecture remains computationally sequential. We study this effect under Depth-μP and CompleteP residual scalings at initialization and during training (one gradient-descent update). By jointly controlling representation moments and weight-reuse responses, we identify sufficient stability conditions under which the interaction coefficient can remain order one independently of residual scaling. As a result, the accumulated interaction can remain order one as depth increases. By contrast, under the standard tied choice, the accumulated interaction decays as $O(L^{-1/2})$ for Depth-μP and $O(L^{-1})$ for CompleteP, where L is the model depth. Consistent with this analysis, experiments support the predicted depth trends and show that suitable independent coefficients can retain interaction and improve learning over the tied baseline. However, we also observe that the performance gains are sensitive to the interaction coefficient. This leads us to explore input-dependent gates that adapt the interaction strength during training. In our comparisons, these adaptive gates indeed further improve performance, with signed SiLU and linear gates outperforming positive-only sigmoid gates. These findings motivate separating the interaction coefficient from residual scaling, enabling a model design choice between retaining interaction when beneficial for learning and using explicitly parallel blocks for concurrent computation.

1 INTRODUCTION

Recent advances in deep learning have been driven largely by scaling Transformers (Vaswani et al., 2017) along two architectural axes: *width* and *depth*. Studies of width scaling show that neural-network parameterization determines whether training exhibits lazy or rich learning as the width tends to infinity (Chizat et al., 2019; Jacot et al., 2018; Yang & Hu, 2021). The maximal-update parameterization (μP), for example, maintains order-one feature updates in the infinite-width limit for fixed-depth networks and enables hyperparameter (HP) transfer from narrow to wide Transformers (Yang & Hu, 2021; Yang et al., 2021). Rich learning at fixed depth, however, does not guarantee rich learning as depth increases. In particular, some residual scalings introduced for stability can suppress higher-order parameter updates, making learning lazy with respect to a subset of parameters (Bordelon et al., 2024a; Dey et al., 2026). Depth scaling therefore matters not only for stability but also for whether a learning mechanism survives.

Building on this perspective, we study a distinct effect of depth scaling that is particularly relevant to Transformers whose blocks contain separate attention and MLP sublayers, namely its influence on the *within-block attention-to-MLP interaction*. In the standard sequential structure, the attention output is added to the residual representation before entering the MLP (Vaswani et al., 2017). To control accumulation across a depth-L Transformer, Depth-μP and CompleteP use residual multipliers $L^{-1/2}$ and L^{-1}, respectively (Yang et al., 2024; Dey et al., 2026). However, these multipliers also reduce the attention contribution to the MLP input. We show that the resulting within-block interaction vanishes with depth under both scalings. The block then behaves as if attention and MLP were *parallel*, even though the computation remains sequential. This differs from true parallel blocks, which can achieve competitive performance with greater training efficiency (Chowdhery

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.

et al., 2023). The tied sequential block instead pays the sequential computation cost but loses both the interaction benefit and the parallelism benefit. This raises two questions:

> *How can we preserve within-block interaction in the large-depth limit while maintaining stable training? Does preserving it improve learning?*

We address these questions by separating the interaction coefficient from the residual multipliers:

$$X_{\ell+1} = X_\ell + a_L \mathcal{A}_\ell(X_\ell) + b_L \mathcal{M}_\ell(H_\ell), \qquad H_\ell = X_\ell + c_L \mathcal{A}_\ell(X_\ell), \tag{1}$$

where X_ℓ is the residual-stream representation entering block ℓ, and $\mathcal{A}_\ell$ and $\mathcal{M}_\ell$ are its attention and MLP maps. Here a_L and b_L control residual accumulation, while c_L governs how much attention contributes to the MLP input. The standard sequential choice ties these scales together ($c_L = a_L$). Our analysis shows that this tying is not required for stability: both the initialization analysis and the formal one-update analysis, subject to appropriate learning-rate conditions, permit $c_L = \Theta(1)$. Setting $c_L = \Theta(1)$ thereby removes the additional decay factor induced by the tied choice, allowing the accumulated interaction to remain order one under both residual scalings.

We next test whether preserving this interaction improves learning. We compare independent interaction coefficients against the tied baseline and find improved performance across models of varying depth. We also find that these gains are sensitive to the choice of c_L: the best coefficient varies across depth and training. This leads us to explore input-dependent gates that adapt the interaction coefficient during training. In our tests, adapting the interaction coefficient further improves performance. Signed SiLU and linear gates outperform positive-only sigmoid gates, which suggests that both the sign and the magnitude of the interaction matter.

Our contributions are summarized as follows:

- **Stability at initialization and after one update.** We establish sufficient conditions for stability at initialization and, through a formal mean-field analysis, derive additional learning-rate conditions for stability after one parameter update. Both sets of conditions allow $c_L = \Theta(1)$. For causal attention, the initialization condition is also necessary under standard nondegeneracy assumptions on the inputs and activations; see Appendix A for the precise statement.
- **Quantitative bounds on accumulated interaction.** We derive accumulated-interaction bounds at initialization and after one update. Under Depth-μP with $c_L = a_L$, the accumulated interaction decays as $O(L^{-1/2})$ both at initialization and after one update. Under CompleteP, it decays as $O(L^{-3/2})$ at initialization and $O(L^{-1})$ after one update. By contrast, setting $c_L = \Theta(1)$ independently of a_L removes these decay factors, permitting order-one interaction in both cases.
- **Empirical evaluation of within-block interaction.** We test the predicted depth scalings and compare fixed and input-dependent interaction coefficients. Independent coefficients improve over the tied baseline, while adaptive gates improve further. Signed SiLU and linear gates outperform the positive-only sigmoid gate, suggesting that allowing attention to reinforce and oppose the MLP input improves learning. Interestingly, sigmoid coefficients increasingly concentrate near zero in deeper blocks and later in training, whereas signed gates retain both positive and negative values across depth, enabling richer modulation of the attention contribution to the MLP input.

2 Related Work

Width and depth parameterization. Under μP, feature updates remain order-one in the infinite-width limit, and HPs transfer across width (Yang & Hu, 2021; Yang et al., 2021). Depth-μP extends this analysis to depth by scaling residual branches by $L^{-1/2}$, yielding a stable infinite-depth limit with feature diversity and depthwise HP transfer for one-layer residual blocks (Yang et al., 2024). It also identifies limitations when a residual block contains multiple layers, as in Transformers. CompleteP instead uses L^{-1} residual scaling together with corresponding parameter and optimizer scalings, showing that HP transfer alone does not prevent lazy learning with respect to subsets of parameters (Dey et al., 2026). In contrast, our question is distinct from whether parameter updates remain non-lazy. We quantify the accumulated effect of the attention output on the subsequent MLP and show that this interaction can vanish under both depth scalings even when the parameterization supports stable and active feature learning.

Mean-field feature-learning dynamics. Classical mean-field theory provides nonlinear feature-learning limits for two-layer neural networks, with rigorous descriptions of the training dynamics and convergence guarantees under suitable assumptions (Mei et al., 2018; Chizat & Bach, 2018). Extending such feature-learning limits to deep architectures is more subtle. Tensor Programs provide a rigorous framework for infinite-width feature learning in finite-depth networks and form the theoretical basis of μP (Yang & Hu, 2021); its extension to Depth-μP further establishes an infinite-depth limit for one-layer residual networks (Yang et al., 2024). Complementary work based on dynamical mean-field theory (DMFT) studies depthwise scaling in residual networks and various limits of Transformers through covariance kernels and response functions (Bordelon et al., 2024b;a). Our width-limit derivation follows the TP frameworks, whereas the subsequent depth analysis directly controls representation moments and weight-reuse responses for the Transformer interaction studied here. This yields stability conditions and accumulated-interaction bounds without treating the DMFT description itself as the basis of our theoretical guarantees.

Residual pathways and gating. Several works modify residual propagation across layers. ReZero and LayerScale introduce learned residual multipliers (Bachlechner et al., 2021; Touvron et al., 2021), while Hyper-Connections and mHC generalize residual pathways across multiple streams (Zhu et al., 2025; Xie et al., 2026). Attention Residuals replaces fixed residual accumulation with input-dependent aggregation over preceding layers (Team et al., 2026). These methods primarily change how information is propagated across depth. Our focus is instead the within-block attention-to-MLP interaction and how depth scaling affects it. Relatedly, (Qiu et al., 2026) introduce gates within the attention module, whereas our gates control only how much attention enters the subsequent MLP while leaving its residual contribution unchanged. Thus, we separate within-block interaction strength from the residual scaling that governs accumulation across depth.

3 Setup and Interaction Measure

Theoretical setup. To isolate the effect of depth scaling on attention-to-MLP interaction, we analyze a depth-L, width-n Transformer with single-head attention and no normalization. This simplified model retains the sequential attention–MLP structure of each Transformer block, which is the mechanism we study, while removing architectural details that would otherwise obscure the analysis. Our experiments return to standard multi-head Transformers with normalization.

Let $X_\ell \in \mathbb{R}^{T\times n}$ denote the residual-stream representation entering layer ℓ, with fixed sequence length T. The attention and MLP maps in Eq. (1) are

$$\mathcal{A}_\ell(X) = \mathrm{softmax}\big(Q_\ell K_\ell^\top/n\big)\, V_\ell, \qquad \mathcal{M}_\ell(H) = \phi(HW_{1,\ell}^\top)W_{2,\ell}^\top, \tag{2}$$

where $Q_\ell = XW_{Q,\ell}^\top$, $K_\ell = XW_{K,\ell}^\top$, and $V_\ell = XW_{V,\ell}^\top$. The softmax is applied rowwise, and the activation function ϕ is applied elementwise. Following μP, we use $1/n$ attention-logit scaling (Yang & Hu, 2021), initialize all hidden weight entries independently as $\mathcal{N}(0, n^{-1})$. For simplicity, we absorb the attention output projection into the value map and omit all bias terms. We defer details of the embedding and readout parameterizations and the backward adjoints to Sections A.1 and B.1.

Within-block interaction. To isolate the effect of the current attention output on the subsequent MLP, we compare the MLP outputs with and without this contribution while holding the residual input and MLP parameters fixed. We define the *within-block attention-to-MLP interaction* as

$$\mathcal{I}_\ell(X) := \mathcal{M}_\ell(X + c_L\mathcal{A}_\ell(X)) - \mathcal{M}_\ell(X). \tag{3}$$

Thus, $\mathcal{I}_\ell(X)$ measures the change in the MLP output induced by the current attention output.

A small interaction within each block need not vanish after accumulation across depth. To quantify its accumulated contribution, we sum the residual updates and obtain the exact decomposition

$$X_L = X_0 + a_L\sum_{\ell=0}^{L-1}\mathcal{A}_\ell(X_\ell) + b_L\sum_{\ell=0}^{L-1}\mathcal{M}_\ell(X_\ell) + b_L\sum_{\ell=0}^{L-1}\mathcal{I}_\ell(X_\ell). \tag{4}$$

The final term is the *accumulated interaction*, which is the central quantity studied in this paper.[1]

[1]This is a within-block comparison at fixed residual input and parameters, not a comparison between two separately propagated networks.

We first characterize stable depth scaling and accumulated interaction at initialization. We then analyze the first gradient-descent (GD) update as a tractable model for training-time behavior. This is the first point at which learning-induced corrections appear, allowing us to isolate their local and accumulated scaling with depth and learning rate. The main analysis considers GD; Appendix C discusses the corresponding scaling intuition for Adam (Kingma & Ba, 2015) through a signSGD approximation (Bernstein et al., 2018). Finally, the experiments test whether separating c_L from the outer residual multipliers improves learning in standard Transformers.

4 Interaction at Initialization

In this section, we analyze the initialized Transformers. In particular, we first characterize the depth scaling required for stable representations at initialization and then quantify how the same scaling affects the accumulated attention-to-MLP interaction. The key distinction is that stability only requires the interaction coefficient c_L to remain bounded, but does not force it to vanish with depth.

Stable depth scaling. At initialization, each weight matrix is centered, so the attention and MLP residual increments are conditionally centered, and their cross terms vanish in expectation. Stability can therefore be measured by the second moments of the residual stream, which accumulate across L layers at scale $L(a_L^2 + b_L^2)$. Additionally, the MLP input $H_\ell = X_\ell + c_L\mathcal{A}_\ell(X_\ell)$ has bounded second moment only if $c_L\mathcal{A}_\ell(X_\ell)$ does, which requires $c_L = O(1)$. These are the required stability conditions, and the following proposition makes this intuition precise.

Proposition 1 (Initialization stability)**.** *Assume a bounded input* $\sup_{n\geq 1} \frac{1}{nT}\mathbb{E}\|X_0\|_F^2 < \infty$*, and let* ϕ *be Lipschitz continuous. If* $c_L = O(1)$ *and*

$$L(a_L^2 + b_L^2) = O(1), \tag{5}$$

then

$$\sup_{L\geq 1}\sup_{n\geq 1}\sup_{0\leq \ell\leq L}\left(\frac{1}{nT}\mathbb{E}\|X_\ell\|_F^2\right) < \infty. \tag{6}$$

Remark 1. *Proposition 1 shows that a_L and b_L control depthwise accumulation along the residual stream, whereas c_L governs only the attention contribution to the MLP input. Stability does not require $c_L \to 0$; $c_L = \Theta(1)$ remains admissible under stable residual scaling. Standard Transformers instead tie $c_L = a_L$, forcing it to decay with depth. We show below that this tying is what causes the accumulated interaction to vanish, and that $c_L = \Theta(1)$ preserves it while maintaining stability.*

The sufficient condition in Proposition 1 is also necessary under causal attention and mild nondegeneracy assumptions on the input and activation. These assumptions are satisfied by standard language-model inputs and commonly used activations. Therefore, the stable depth scaling identified above is sharp within the bounded-c_L regime and does not require $c_L \to 0$. A detailed discussion of the nondegeneracy assumptions, their verification, and the proof is deferred to Appendix A.

Theorem 1 (Necessary and sufficient initialization stability)**.** *Assume causal self-attention and the input and activation nondegeneracy conditions in Appendix A.5. If $c_L = O(1)$, then the uniform initialization stability in Eq. (6) holds if and only if the depth-scaling condition Eq. (5) holds. Equivalently, for $a_L = L^{-\alpha}$ and $b_L = L^{-\beta}$, stability holds if and only if $\alpha, \beta \geq \frac{1}{2}$.*

Accumulated interaction. Stable representations alone do not determine whether the within-block interaction survives as depth increases. For intuition, a first-order Taylor expansion gives

$$\mathcal{I}_\ell(X) \approx c_L D\mathcal{M}_\ell(X)[\mathcal{A}_\ell(X)], \tag{7}$$

where $D\mathcal{M}_\ell(X)[\cdot]$ denotes the directional derivative of $\mathcal{M}_\ell$ at X. Therefore, the per-layer interaction is proportional to c_L. By Eq. (4), its accumulated contribution is controlled jointly by the residual multiplier b_L and the interaction coefficient c_L. Whether the interaction survives therefore depends on its depth scaling and requires separate analysis.

Theorem 2 (Accumulated interaction at initialization)**.** *Under the assumptions of Proposition 1,*

$$\sup_{n\geq 1}\left(\frac{1}{nT}\mathbb{E}\left\|b_L\sum_{\ell=0}^{L-1}\mathcal{I}_\ell(X_\ell)\right\|_F^2\right)^{1/2} = O\Big(\sqrt{L}\,b_L c_L\Big). \tag{8}$$

Table 1: Accumulated-interaction scaling at initialization and after one gradient update. The effective learning-rate coefficient $\eta_c > 0$ is fixed independently of depth.

		Initialization		After one update	
Scaling	η_L	$c_L = a_L$	$c_L = \Theta(1)$	$c_L = a_L$	$c_L = \Theta(1)$
Depth-μP: $L^{-1/2}$	η_c	$O(L^{-1/2})$	$O(1)$	$O((1+\eta_c)L^{-1/2})$	$O(1+\eta_c)$
CompleteP: L^{-1}	$\eta_c L$	$O(L^{-3/2})$	$O(L^{-1/2})$	$O(L^{-3/2}+\eta_c L^{-1})$	$O(L^{-1/2}+\eta_c)$

The factor $\sqrt{L}$ in Eq. (8) arises from second-moment accumulation across depth. Although $\mathcal{I}_\ell(X_\ell)$ depends on the preceding residual trajectory, the current-layer MLP output weight $W_{2,\ell}$ is independent and centered, so cross-layer interaction terms vanish in expectation. The squared magnitudes therefore add across the L blocks, giving a factor L in the second moment and hence $\sqrt{L}$ in the root-mean-square interaction magnitude. The complete proof is deferred to Appendix A.

For $a_L = b_L = L^{-\alpha}$, Theorem 2 gives an accumulated interaction of $O(L^{\frac{1}{2}-\alpha}c_L)$. Under the standard tied choice $c_L = a_L = L^{-\alpha}$, this reduces to $O(L^{\frac{1}{2}-2\alpha})$, which vanishes for every stable $\alpha \geq \frac{1}{2}$ allowed by Proposition 1. Thus, tying the interaction coefficient to the residual multiplier contributes an additional factor $L^{-\alpha}$, forcing the accumulated interaction to vanish at initialization across all stable power-law residual scalings.

Choosing $c_L = \Theta(1)$ removes this additional decay while remaining compatible with initialization stability. With $L^{-1/2}$ residual scaling, the accumulated interaction can remain $O(1)$; with L^{-1} scaling, it is $O(L^{-1/2})$ and therefore vanishes at initialization. This latter decay arises because the initialized interaction contributions are centered: their second moments accumulate at scale Lb_L^2, yielding a root-mean-square magnitude $\sqrt{L}b_L = L^{-1/2}$. After a parameter update, however, learning introduces response and learning corrections that need not obey the same centered accumulation. We examine these contributions next through the first gradient-descent update. Table 1 summarizes the initialization rates and the corresponding one-update rates derived below.

5 Interaction after One Parameter Update

At initialization, conditional centering controls residual accumulation and within-block interaction. After a parameter update, however, the weights become coupled to the forward features and backward adjoints that generated their gradients, introducing additional response and learning terms. Using Tensor Programs, we derive the infinite-width learning dynamics of these terms and then analyze stability and accumulated interaction after one gradient-descent update as depth increases.

Mean-field learning dynamics. The key step in deriving the mean-field description is to characterize the learned linear maps in the attention and MLP sublayers together with their induced backward fields. These are the only components that directly contain trainable hidden weights; attention mixing, elementwise nonlinearities, and residual additions then couple the resulting fields deterministically. For $\theta \in \{Q, K, V, U, M\}$, write

$$Y^k_{\theta,\ell} = Z^k_{\theta,\ell} W^{k\top}_{\theta,\ell}, \qquad \delta Z^k_{\theta,\ell} = \delta Y^k_{\theta,\ell} W^k_{\theta,\ell}. \tag{9}$$

Here $Z^k_{\theta,\ell}$ is the input to the linear map, while $\delta Y^k_{\theta,\ell}$ and $\delta Z^k_{\theta,\ell}$ are its scaled output and input adjoints. For example, the query, key, and value projections all take $Z^k_{\theta,\ell} = X^k_\ell$. As introduced in Section B.1, we relabel $W_{U,\ell} := W_{1,\ell}$ and $W_{M,\ell} := W_{2,\ell}$ to align the weight names with their associated fields. Thus U_ℓ is the preactivation with input H_ℓ, and M_ℓ is the MLP output with input $\Phi_\ell := \phi(U_\ell)$.

To simplify the presentation, all hidden matrices share the same learning rate η_L with gradient update

$$W^{k+1}_{\theta,\ell} = W^k_{\theta,\ell} - \eta_L \nabla_{W_{\theta,\ell}} \mathcal{L}^k. \tag{10}$$

The embedding, readout, loss, and scaled-adjoint definitions are given in Appendix B.1–B.2.

Substituting the parameter updates into the forward and backward equations yields a Tensor Program, as established in Proposition 6. The TP's Master Theorem then characterizes the infinite-width limit through jointly constructed mean-field variables. We use a bar to denote the mean-field variable associated with a typical width coordinate of the corresponding finite-width quantity. Thus,

$\bar{y}^k_{\theta,\ell} \in \mathbb{R}^T$ corresponds to a typical width coordinate of $Y^k_{\theta,\ell}$, viewed jointly across the T tokens. Applying the TP limiting rules gives the following representation for the learned linear fields.

Proposition 2 (Mean-field learning dynamics)**.** *Fix the depth L and training horizon K. Under the assumptions of Proposition 6, for every $\theta \in \{Q, K, V, U, M\}$, $0 \le \ell < L$, and $0 \le k \le K$,*

$$\bar{y}^k_{\theta,\ell} = \xi^k_{\theta,\ell} + \sum_{s<k} \left[R^{k,s}_{z_\theta,\zeta_\theta,\ell} - \eta_L \rho_{\theta,L} \bar{C}^{k,s}_{z_\theta,\ell} \right] \delta\bar{y}^s_{\theta,\ell}, \tag{11}$$

$$\delta\bar{z}^k_{\theta,\ell} = \zeta^k_{\theta,\ell} + \sum_{s\le k} R^{k,s}_{\delta y_\theta,\xi_\theta,\ell} \bar{z}^s_{\theta,\ell} - \eta_L \rho_{\theta,L} \sum_{s<k} \bar{C}^{k,s}_{\delta y_\theta,\ell} \bar{z}^s_{\theta,\ell}. \tag{12}$$

where $\rho_{\theta,L}$ are the residual multipliers. Here $\bar{C}^{k,s}_{z_\theta,\ell} = \mathbb{E}[\bar{z}^k_{\theta,\ell} \bar{z}^{s\top}_{\theta,\ell}]$ and $\bar{C}^{k,s}_{\delta y_\theta,\ell} = \mathbb{E}[\delta\bar{y}^k_{\theta,\ell} \delta\bar{y}^{s\top}_{\theta,\ell}]$ are uncentered moment kernels. The Gaussian components $(\xi^k_{\theta,\ell}, \zeta^k_{\theta,\ell})$ are jointly centered and satisfy

$$\mathbb{E}[\xi^k_{\theta,\ell} \xi^{s\top}_{\theta,\ell}] = \bar{C}^{k,s}_{z_\theta,\ell}, \qquad \mathbb{E}[\zeta^k_{\theta,\ell} \zeta^{s\top}_{\theta,\ell}] = \bar{C}^{k,s}_{\delta y_\theta,\ell}. \tag{13}$$

Gaussian families associated with distinct initialized matrices or opposite multiplication orientations are independent. The response matrices are defined by

$$R^{k,s}_{z_\theta,\zeta_\theta,\ell} := \mathbb{E}[D_{\zeta^s_{\theta,\ell}} \bar{z}^k_{\theta,\ell}], \quad s < k; \qquad R^{k,s}_{\delta y_\theta,\xi_\theta,\ell} := \mathbb{E}[D_{\xi^s_{\theta,\ell}} \delta\bar{y}^k_{\theta,\ell}], \quad s \le k, \tag{14}$$

and vanish outside the indicated ranges.

Proof sketch. The full derivation is given in Appendix B.3. Proposition 6 first casts the complete forward–backward training computation as a Tensor Program and establishes its joint infinite-width limit via its Master Theorem. Applying the ZHat/ZDot rules to reused initialized weights, together with the empirical-average rule for the accumulated parameter updates, yields the generic linear-field representation of Lemma 6. Specializing that representation to $\eta_\theta = \eta_L$ gives Eq. (11)–Eq. (12). □

The representation in Eq. (11)–Eq. (12) decomposes each learned linear field into three distinct components: a Gaussian term capturing the contribution of the initialized weight matrix, with cross-step correlations preserved via Eq. (13); a response term accounting for the dependence introduced by reusing initialized weight matrices across forward and backward orientations; and a direct learning correction resulting from accumulated parameter updates, governed by the scaling factor $\eta_L \rho_{\theta,L}$. Causally, while the backward response term may depend on the current forward evaluation ($s \le k$), the direct learning correction strictly reflects earlier parameter updates ($s < k$).

The learned linear fields close the forward dynamics through

$$\begin{aligned} &\bar{P}^k_\ell = \mathrm{softmax}\big(\mathbb{E}[\bar{q}^k_\ell \bar{k}^{k\top}_\ell]\big), && \bar{a}^k_\ell = \bar{P}^k_\ell \bar{v}^k_\ell, && \bar{h}^k_\ell = \bar{x}^k_\ell + c_L \bar{a}^k_\ell, \\ &\bar{\phi}^k_\ell = \phi(\bar{u}^k_\ell), && \bar{x}^k_{\ell+1} = \bar{x}^k_\ell + a_L \bar{a}^k_\ell + b_L \bar{m}^k_\ell. \end{aligned} \tag{15}$$

Together with the backward, embedding, and readout relations, these equations define the joint Transformer dynamics; the full system is given in Proposition 7. We next specialize to the first parameter update and derive conditions under which the representation moments and responses remain jointly controlled as depth increases.

Joint stability of representations and responses. Setting $k = 1$ and increasing depth yields the local learning coefficients $\eta_L a_L$ and $\eta_L b_L$ in Equation Eq. (11). Their contributions to the residual stream acquire another residual multiplier, producing the accumulated scales $L\eta_L a_L^2$ and $L\eta_L b_L^2$. Thus, the field moments and responses must be controlled jointly as each enters the other's recursion.

We use the primary response $R^1_\ell := \mathbb{E}[D_{\delta\bar{x}^0_\ell} \bar{x}^1_\ell]$ with the fixed-law derivative defined in Eq. (265). Lemma 9 derives its recursion and expresses the remaining one-update responses through it.

Proposition 3 (Joint mean-field stability after one update)**.** *Assume the setup of Proposition 2, $c_L = O(1)$, and $L(a_L^2 + b_L^2) = O(1)$. Suppose the embedding and readout learning-rate coefficients are bounded independently of depth. If*

$$\eta_L(a_L + b_L) = O(1), \qquad L\eta_L(a_L^2 + b_L^2) = O(1), \tag{16}$$

then

$$\sup_{L\ge 1} \max_{0\le\ell\le L} \left[\frac{1}{T} \mathbb{E}\|\bar{x}^1_\ell\|_2^2 + \|R^1_\ell\|^2_{\mathrm{op}} \right] < \infty. \tag{17}$$

All internal updated forward fields have uniformly bounded second moments as well.

The two conditions in Eq. (16) control the local correction to each learned linear field and its accumulation across depth, respectively. For $a_L = b_L = L^{-\alpha}$, they reduce to

$$\alpha \geq \tfrac{1}{2}, \qquad \eta_L \lesssim \min\{L^{\alpha}, L^{2\alpha-1}\}. \tag{18}$$

Hence, with a depth-independent effective learning-rate coefficient $\eta_c > 0$, the choices $\eta_L = \eta_c$ for $\alpha = 1/2$ and $\eta_L = \eta_c L$ for $\alpha = 1$ are admissible hidden-weight learning rates. These conditions are sufficient for depth stability after one parameter update and, as at initialization in Proposition 1, remain compatible with an independent interaction coefficient $c_L = \Theta(1)$.

Accumulated interaction. Let $\bar{\mathcal{I}}_\ell^1 \in \mathbb{R}^T$ denote the representative interaction from Eq. (3) after one parameter update. To isolate the effect of attention on the MLP, we compare the MLP evaluated at the full input $\bar{h}_\ell^1 = \bar{x}_\ell^1 + c_L \bar{a}_\ell^1$ with the same MLP evaluated after removing the attention contribution, so that its input is just $\bar{x}_\ell^1$. The two evaluations therefore use the same updated MLP weights and residual representation and differ only by $c_L \bar{a}_\ell^1$. Lemma 12 places them in the same Tensor Program and decomposes the resulting interaction into Gaussian, response, and learning components.

Theorem 3 (Accumulated mean-field interaction after one update)**.** *Under the assumptions of Proposition 3, let $a_L = b_L = L^{-\alpha}$ with $\alpha \geq 1/2$. Then*

$$\left(\frac{1}{T} \mathbb{E} \left\| b_L \sum_{\ell=0}^{L-1} \bar{\mathcal{I}}_\ell^1 \right\|_2^2 \right)^{1/2} = O\Big(c_L \left[L^{\frac{1}{2}-\alpha} + \eta_L L^{1-2\alpha} \right] \Big). \tag{19}$$

The two terms in Eq. (19) have different origins. The first, $O(c_L L^{\frac{1}{2}-\alpha})$, comes from the Gaussian component. These interaction terms are orthogonal across layers in the mean-field construction, so their second moments add over depth and give the same $\sqrt{L}$ accumulation as at initialization. The second, $O(c_L \eta_L L^{1-2\alpha})$, arises from the response and learning corrections introduced by the parameter update. Their accumulation is controlled using the joint stability bound in Proposition 3. The complete derivation is given in Appendix B.

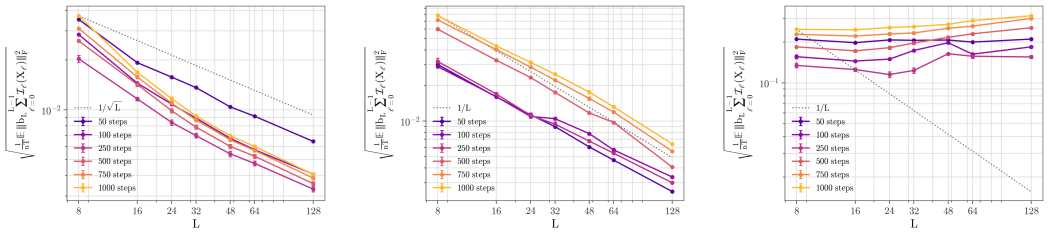
Table 1 compares the accumulated-interaction bounds at initialization and after one parameter update. The tied choice $c_L = a_L$ introduces an additional decay factor and can force the interaction magnitude to vanish. By contrast, an independent $c_L = \Theta(1)$ remains compatible with the sufficient stability conditions and can preserve an $O(1)$ interaction after learning. In particular, under L^{-1} residual scaling, the initialization bound is $O(L^{-1/2})$, whereas after one update it becomes $O(L^{-1/2} + \eta_c)$. Thus, even when the initialized interaction vanishes, training can generate a non-vanishing interaction if its coefficient is decoupled from the residual multiplier. The same conclusion holds under Adam, with only the scaling constants changing; see Appendix C. The experiments next examine whether preserving this interaction improves learning over longer training runs.

6 Empirical Studies

In the previous sections, our theoretical analysis shows that stable residual scaling allows an independent interaction coefficient $c_L = \Theta(1)$ as depth increases, while the standard tied choice causes the within-block interaction to vanish. In this section, we empirically examine whether this vanishing behavior appears in standard Transformers, whether retaining the interaction improves learning, and how different fixed and adaptive choices of the interaction coefficient affect performance.

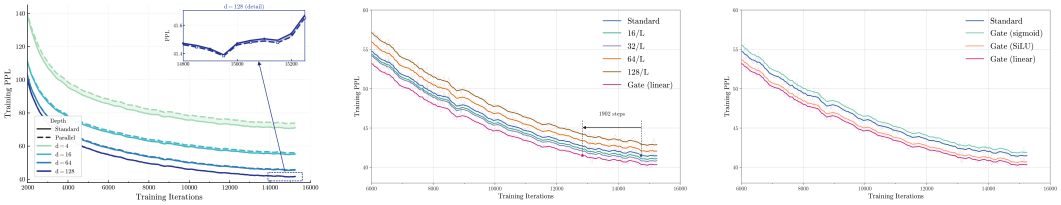
Experimental setup. All experiments use standard dense Transformers with multi-head self-attention and pre-norm, rather than the simplified model used in our theoretical analysis. We primarily follow the experimental setup of (Dey et al., 2026) and compare interaction variants under both Depth-μP and CompleteP residual scalings. We train models with depths $L \in \{4, 16, 64, 128\}$ and width $n = 256$ on the English subset of the open-source C4 corpus (Dodge et al., 2021) using AdamW. Each model is trained on 1B tokens, and we evaluate on the full C4-en validation split. For each depth and residual parameterization, all variants share the same backbone, optimization setup, and data budget. Appendix D provides full experimental details and additional results.

Interaction scaling at finite width. We first test whether the predicted interaction scaling persists in finite-width Transformers. We evaluate the accumulated interaction across model depths and



(a) Depth-μP, tied interaction. (b) CompleteP, tied interaction. (c) Independent $c_L = \Theta(1)$.

Figure 1: **Accumulated interaction across depth and training.** With $c_L = a_L$, interaction decays at the predicted rates under Depth-μP and CompleteP, while $c_L = \Theta(1)$ remains nonvanishing. These empirical trends persist beyond the first parameter update covered by the theory.



(a) Tied vs. parallel (b) Fixed interaction strengths (c) Adaptive interaction

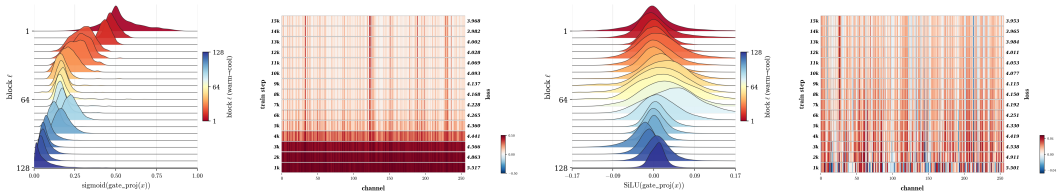
Figure 2: **Effect of attention-to-MLP interaction on learning.** (a) The tied–parallel performance gap narrows with depth. (b) Suitable independent coefficients outperform the tied baseline, with gains depending on interaction strength. (c) Adaptive gates improve further, especially signed gates.

training checkpoints, comparing the tied choice $c_L = a_L$ with $c_L = \Theta(1)$. The effective learning rate η_c is fixed across depths using the corresponding Depth-μP and CompleteP scalings. Figure 1 shows that, once training begins, the tied variants decay approximately as $L^{-1/2}$ and L^{-1} under Depth-μP and CompleteP residual scaling, respectively, consistent with Table 1. In contrast, the independent-coefficient variants maintain substantial interaction across the tested depths. Although tied interactions grow during training, their decay with depth persists—matching predicted scaling rates closely—even though the theory is established only for the first gradient update.

Learning Benefits of Independent Interaction We next examine how interaction strength relates to learning performance. Figure 2(a) compares the tied sequential model with a parallel-block variant ($c_L = 0$): the tied model performs better at shallow depths, but the gap narrows with depth and nearly disappears for the deepest models, consistent with similar observations in Chowdhery et al. (2023). This suggests that the benefit of the sequential structure weakens as its within-block interaction decreases. We therefore test whether preserving this interaction through an independent coefficient improves learning. Figure 2(b) shows that suitable choices outperform the tied baseline, while the gains depend on interaction strength. Preserving interaction can thus improve learning without determining the best coefficient, motivating an adaptive coefficient over a fixed one.

Adaptive and Signed Interaction We replace the fixed coefficient with an input-dependent gate $g_\ell(X_\ell)$, using $H_\ell = X_\ell + g_\ell(X_\ell) \odot \mathcal{A}_\ell(X_\ell)$ while leaving the residual multipliers unchanged. We compare sigmoid, SiLU, and linear gates with the same predictor; sigmoid restricts the coefficients to $(0, 1)$, whereas SiLU and linear gates permit both signs. Gate details are provided in Appendix D. Figure 2(c) shows that adaptive gates further improve over the tied and fixed-coefficient variants, with the signed SiLU and linear gates outperforming the positive-only sigmoid gate. Negative coefficients allow the model to subtract, rather than only add, the attention contribution to the MLP input. While these gates also differ in output range and optimization behavior, the results suggest that restricting the interaction to positive values can be unnecessarily limiting.

Interaction Across Blocks and Training To examine how adaptive interaction evolves across depth and training, we track gate distributions over block positions and checkpoints. Figure 3 shows that sigmoid gates become increasingly concentrated near zero in deeper blocks and later training,



(a) Sigmoid: blocks (b) Sigmoid: training (c) SiLU: blocks (d) SiLU: training

Figure 3: **Adaptive interaction gates across blocks and training.** Sigmoid gates increasingly concentrate near zero in deeper blocks and later in training, reducing the effective interaction, whereas SiLU gates retain substantial positive and negative values across depth and training.

reducing the attention contribution to the MLP input and making these blocks more parallel-like along the interaction pathway. In contrast, SiLU and linear gates retain substantial positive and negative values across depth and training. These models therefore adapt not only the interaction magnitude but also its sign across blocks and training stages.

Overall, the experiments show that separating the interaction coefficient from residual scaling improves learning. Moreover, adaptive gates yield further gains by varying their strength and sign across inputs, blocks, and training steps. These results support treating within-block interaction as an independent architectural choice rather than tying it to residual scaling.

7 Conclusion

In this paper, we studied how depth scaling affects accumulated attention-to-MLP interaction in Transformers. Our initialization analysis and one-update mean-field bounds show that the standard tied choice causes this interaction to vanish under both Depth-μP and CompleteP residual scalings. We identify stability conditions allowing $c_L = \Theta(1)$ independently of a_L and b_L. Experiments with multi-head normalized Transformers show that suitable fixed coefficients retain interaction and improve learning over the tested depths, with further gains from signed adaptive gates. These gains do not imply universal superiority over parallel blocks, which allow concurrent attention and MLP computation. The choice therefore involves a trade-off between learning performance and training efficiency. By contrast, the standard tied design retains the sequential dependency while its explicit within-block interaction becomes negligible as depth increases. Our findings therefore motivate treating residual scaling and within-block interaction as separate design choices.

Limitations. The theory uses single-head attention without normalization and deterministic interaction coefficients. Its quantitative training-time bounds concern the first single-sample gradient update in a width-first mean-field description; they do not establish guarantees for longer training or other optimizers. Although our experiments use multi-head normalized Transformers trained with AdamW, they cover a limited range of model sizes and data budgets rather than industrial-scale training. The observed gains from independent coefficients and signed gates should therefore be interpreted within the tested settings.

Future directions. Extending the joint representation–response analysis to multiple updates, learned gates, and normalized multi-head Transformers trained with adaptive optimizers would connect the theory more directly to practice. Larger-scale, compute-matched experiments would also help determine when retaining interaction justifies its sequential execution cost. More broadly, our findings motivate studying how depth scaling changes feature learning and interactions between sublayers, beyond its role in stabilizing residual accumulation.

AI USE STATEMENT

We used generative AI tools to check and proofread mathematical derivations and proofs and to improve the clarity and organization of the manuscript. The theoretical framework, mathematical claims, primary derivations, and proof arguments were developed by the authors without generative AI. We also used generative AI for brainstorming and literature exploration; all references and related-work claims were verified against the original sources.

For the empirical study, generative AI was used to brainstorm practical experimental choices, such as datasets, model widths, depths, and training scales under our available computational resources, and to assist with debugging code. The research questions, experimental design and execution, and interpretation of results were conducted by the authors. Generative AI was also used to revise descriptions of the results for clarity and structure. All AI-assisted work was reviewed by the authors, with mathematical content independently checked and code changes tested for correctness. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The main text presents the core model setup, theoretical results, and representative experiments needed to follow the paper's central argument. The appendix provides the complete specification and supporting analysis required for reproduction and verification. In particular, it includes the full model parameterization and training setup, including embedding and readout layers, scaled backward variables, gradient updates, and learning-rate conventions; the complete joint mean-field dynamics and proofs underlying the main theoretical results; and detailed experimental configurations, including architectures, datasets, hyperparameters, training procedures, and evaluation protocols. The appendix also reports complementary results and experiments at width 128, which show the same qualitative trends as the width-256 results presented in the main text.

REFERENCES

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in artificial intelligence*, pp. 1352–1361. PMLR, 2021.

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pp. 404–413. PMLR, 2018.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pp. 560–569. PMLR, 2018.

Blake Bordelon, Hamza Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. *Advances in Neural Information Processing Systems*, 2024a.

Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *The Twelfth International Conference on Learning Representations*, 2024b.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 2019.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 2023.

Nolan Dey, Bin Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: Completep enables compute-efficient deep transformers. *Advances in Neural Information Processing Systems*, 2026.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 1286–1305, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

Etai Littwin and Greg Yang. Adaptive optimization in the ∞-width limit. In *The Eleventh International Conference on Learning Representations*, 2023.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *Advances in Neural Information Processing Systems*, 38:100092–100118, 2026.

Kimi Team, Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, Yaoyu Wang, Yucheng Wang, Guanduo Chen, Bohong Yin, et al. Attention residuals. *arXiv preprint arXiv:2603.15031*, 2026.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 32–42. IEEE, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Kuai Yu, et al. mHC: Manifold-constrained hyper-connections. In *Forty-third International Conference on Machine Learning*, 2026.

Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, 2021.

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs VI: Feature learning in infinite depth neural networks. In *International Conference on Learning Representations*, 2024.

Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, and Xun Zhou. Hyper-connections. In *International Conference on Learning Representations*, volume 2025, pp. 97183–97219, 2025.

A Initialization Analysis

This appendix provides the full stability and interaction bounds at initialization, allowing separate residual multipliers and making their finite-width dependence explicit. We first establish the moment identities used in both arguments, then prove the stability and interaction bounds. We subsequently derive the matching necessary condition for causal attention and verify the input and activation assumptions. The final subsection treats full self-attention and Vision Transformer inputs. All variables in this appendix are evaluated at initialization, so we omit training-step superscripts.

A.1 Model and assumptions

We consider a depth-L, width-n Transformer with single-head attention and no normalization. The sequence length T is fixed. Blocks are indexed by $\ell = 0, \dots, L-1$, with residual-stream representations $X_0, \dots, X_L \in \mathbb{R}^{T\times n}$. For $X \in \mathbb{R}^{T\times n}$, define

$$Q_\ell(X) = XW_{Q,\ell}^\top, \qquad K_\ell(X) = XW_{K,\ell}^\top, \qquad V_\ell(X) = XW_{V,\ell}^\top, \tag{20}$$

$$P_\ell(X) = \mathrm{softmax}\left(\frac{Q_\ell(X)K_\ell(X)^\top}{n}\right), \quad \mathcal{A}_\ell(X) = P_\ell(X)V_\ell(X). \tag{21}$$

The softmax is applied rowwise, and a deterministic attention mask may be included. Following μP, we use $1/n$ attention-logit scaling and absorb the attention output projection into the value map.

The MLP map is

$$\mathcal{M}_\ell(H) = \phi(HW_{1,\ell}^\top)W_{2,\ell}^\top, \tag{22}$$

where ϕ is applied entrywise. The residual block is

$$H_\ell = X_\ell + c_L\mathcal{A}_\ell(X_\ell), \tag{23}$$

$$X_{\ell+1} = X_\ell + a_L\mathcal{A}_\ell(X_\ell) + b_L\mathcal{M}_\ell(H_\ell). \tag{24}$$

The coefficients a_L, b_L, c_L are deterministic and may depend on the total depth L. The standard sequential parameterization sets $c_L = a_L$; we keep them separate throughout the analysis.

Following μP, for $\theta \in \{Q, K, V, 1, 2\}$, the entries of $W_{\theta,\ell}$ are initialized independently as $\mathcal{N}(0, n^{-1})$, independently across matrices and blocks. Biases are omitted.

Standing assumptions. We assume that ϕ is C_ϕ-Lipschitz:

$$|\phi(x) - \phi(y)| \le C_\phi|x - y|, \qquad x, y \in \mathbb{R}. \tag{25}$$

Consequently,

$$|\phi(x)|^2 \le 2C_\phi^2|x|^2 + 2|\phi(0)|^2. \tag{26}$$

We further assume that X_0 is independent of the hidden weights and has uniformly bounded normalized second moment:

$$\sup_{L,n\ge 1} \frac{1}{nT}\mathbb{E}\|X_0\|_F^2 < \infty. \tag{27}$$

Throughout, T and C_ϕ are fixed, and asymptotic constants are independent of n, L, and the block index. The input condition covers the standard language-modeling and ViT realizations considered here; see Sections A.6 and A.7.

A.2 Initialized sublayer moments

The proofs use conditional centering of the value weights and the MLP output weights. The value weights suffice for the attention argument because $P_\ell(X)$ does not depend on $W_{V,\ell}$.

Lemma 1 (Initialized sublayer moments)**.** *Let $X \in \mathbb{R}^{T\times n}$ have finite second moment and be independent of the parameters of block ℓ. Then*

$$\mathbb{E}\langle X, \mathcal{A}_\ell(X)\rangle_F = 0, \tag{28}$$

$$\mathbb{E}\|\mathcal{A}_\ell(X)\|_F^2 \le T\mathbb{E}\|X\|_F^2. \tag{29}$$

For any $H \in \mathbb{R}^{T\times n}$ with finite second moment that is independent of $W_{1,\ell}$ and $W_{2,\ell}$,

$$\mathbb{E}\langle H, \mathcal{M}_\ell(H)\rangle_F = 0, \tag{30}$$

$$\mathbb{E}\|\mathcal{M}_\ell(H)\|_F^2 \le 2C_\phi^2\mathbb{E}\|H\|_F^2 + 2nT|\phi(0)|^2. \tag{31}$$

Proof. Condition on $X, W_{Q,\ell}, W_{K,\ell}$. Then $P_\ell(X)X$ is fixed, while $W_{V,\ell}$ is independent and centered. Hence

$$\mathbb{E}[\mathcal{A}_\ell(X) \mid X, W_{Q,\ell}, W_{K,\ell}] = 0, \tag{32}$$

which gives Eq. (28). For each output coordinate j, the jth row of $W_{V,\ell}$ has covariance $n^{-1}I_n$. Therefore,

$$\begin{aligned}\mathbb{E}\big[\|\mathcal{A}_\ell(X)\|_F^2 \mid X, W_{Q,\ell}, W_{K,\ell}\big] &= \sum_{j=1}^{n}\frac{1}{n}\|P_\ell(X)X\|_F^2 \\ &= \|P_\ell(X)X\|_F^2. \end{aligned}\tag{33}$$

Each row of $P_\ell(X)$ is a probability vector, so

$$\|P_\ell(X)\|_{\mathrm{op}}^2 \le \|P_\ell(X)\|_F^2 \le T. \tag{34}$$

Taking expectation in Eq. (33) therefore proves Eq. (29).

For the MLP, condition on $H, W_{1,\ell}$. The output is linear in the independent centered matrix $W_{2,\ell}$, which gives

$$\mathbb{E}[\mathcal{M}_\ell(H) \mid H, W_{1,\ell}] = 0, \tag{35}$$

$$\mathbb{E}\big[\|\mathcal{M}_\ell(H)\|_F^2 \mid H, W_{1,\ell}\big] = \|\phi(HW_{1,\ell}^\top)\|_F^2. \tag{36}$$

The first identity proves Eq. (30). For the second, Eq. (26) implies

$$\begin{aligned}\mathbb{E}\|\phi(HW_{1,\ell}^\top)\|_F^2 &\le 2C_\phi^2\mathbb{E}\|HW_{1,\ell}^\top\|_F^2 + 2nT|\phi(0)|^2 \\ &= 2C_\phi^2\mathbb{E}\|H\|_F^2 + 2nT|\phi(0)|^2. \end{aligned}\tag{37}$$

This proves Eq. (31). □

A.3 Sufficient initialization stability

We first state the finite-depth bound without imposing $c_L = O(1)$. The bounded-coefficient specialization then gives the main-text result.

Proposition 4 (Uniform forward stability at initialization)**.** *Under the activation and input conditions Eq. (25) and Eq. (27), for every $0 \le \ell \le L$,*

$$\frac{1}{nT}\mathbb{E}\|X_\ell\|_F^2 \le \exp\big\{\ell\left[Ta_L^2 + 2C_\phi^2 b_L^2(1+Tc_L^2)\right]\big\} \times \left[\frac{1}{nT}\mathbb{E}\|X_0\|_F^2 + 2\ell b_L^2|\phi(0)|^2\right]. \tag{38}$$

For every $0 \le \ell < L$,

$$\frac{1}{nT}\mathbb{E}\|H_\ell\|_F^2 \le (1+Tc_L^2)\frac{1}{nT}\mathbb{E}\|X_\ell\|_F^2. \tag{39}$$

Consequently, if

$$L\left[a_L^2 + b_L^2(1+c_L^2)\right] = O(1), \tag{40}$$

then

$$\sup_{L,n\ge 1}\max_{0\le\ell\le L}\frac{1}{nT}\mathbb{E}\|X_\ell\|_F^2 < \infty. \tag{41}$$

If, in addition, $c_L = O(1)$, then the residual-stream representation X_ℓ and MLP inputs H_ℓ satisfy the combined bound

$$\sup_{L,n\ge 1}\frac{1}{nT}\left(\max_{0\le\ell\le L}\mathbb{E}\|X_\ell\|_F^2 + \max_{0\le\ell<L}\mathbb{E}\|H_\ell\|_F^2\right) < \infty. \tag{42}$$

Proof. Since X_ℓ depends only on the input and preceding blocks, it is independent of the parameters of block ℓ. Expanding Eq. (23) and applying Eq. (28) gives

$$\begin{aligned}\mathbb{E}\|H_\ell\|_F^2 &= \mathbb{E}\|X_\ell\|_F^2 + c_L^2\mathbb{E}\|\mathcal{A}_\ell(X_\ell)\|_F^2 \\ &\le (1+Tc_L^2)\mathbb{E}\|X_\ell\|_F^2. \end{aligned} \tag{43}$$

This proves Eq. (39).

For the residual update, the cross term between X_ℓ and the attention output vanishes by conditional centering of $W_{V,\ell}$. The cross terms involving the MLP vanish after conditioning on X_ℓ, the current attention parameters, and $W_{1,\ell}$, because $W_{2,\ell}$ is independent and centered. Hence

$$\begin{aligned}\mathbb{E}\|X_{\ell+1}\|_F^2 = \mathbb{E}\|X_\ell\|_F^2 &+ a_L^2\mathbb{E}\|\mathcal{A}_\ell(X_\ell)\|_F^2 \\ &+ b_L^2\mathbb{E}\|\mathcal{M}_\ell(H_\ell)\|_F^2. \end{aligned} \tag{44}$$

Combining Lemma 1 with Eq. (39) yields

$$\frac{1}{nT}\mathbb{E}\|X_{\ell+1}\|_F^2 \le \left[1 + Ta_L^2 + 2C_\phi^2 b_L^2(1+Tc_L^2)\right]\frac{1}{nT}\mathbb{E}\|X_\ell\|_F^2 + 2b_L^2|\phi(0)|^2. \tag{45}$$

Let $r_L = 1 + Ta_L^2 + 2C_\phi^2 b_L^2(1+Tc_L^2) \ge 1$. Iteration gives

$$\begin{aligned}\frac{1}{nT}\mathbb{E}\|X_\ell\|_F^2 &\le r_L^\ell \frac{1}{nT}\mathbb{E}\|X_0\|_F^2 + 2b_L^2|\phi(0)|^2\sum_{j=0}^{\ell-1} r_L^j \\ &\le r_L^\ell\left[\frac{1}{nT}\mathbb{E}\|X_0\|_F^2 + 2\ell b_L^2|\phi(0)|^2\right]. \end{aligned} \tag{46}$$

Using $1 + x \le e^x$ gives Eq. (38). Under Eq. (40), the exponential factor and $Lb_L^2|\phi(0)|^2$ are uniformly bounded. Together with Eq. (27), this gives Eq. (41). If $c_L = O(1)$, Eq. (39) also gives Eq. (42). □

Corollary 1 (Bounded interaction coefficients)**.** *Assume Eq. (25) and Eq. (27), with $c_L = O(1)$. The condition $L(a_L^2 + b_L^2) = O(1)$ is sufficient for the combined uniform bound Eq. (42). In particular, if $a_L = O(L^{-\alpha})$ and $b_L = O(L^{-\beta})$, then $\alpha, \beta \ge 1/2$ are sufficient.*

Proof. For bounded c_L, $a_L^2 + b_L^2 \le a_L^2 + b_L^2(1+c_L^2) \le C(a_L^2 + b_L^2)$ for a depth-independent constant C. The result follows from Proposition 4. □

Remark 2 (Residual scaling and the interaction coefficient)**.** *The sufficient stability condition permits an independent order-one coefficient $c_L = \Theta(1)$ without requiring the standard tying $c_L = a_L$. This statement concerns stability; the accumulated interaction is quantified separately below.*

A.4 Within-block and accumulated interaction

For a fixed residual input X, define

$$\mathcal{I}_\ell(X) := \mathcal{M}_\ell(X + c_L\mathcal{A}_\ell(X)) - \mathcal{M}_\ell(X). \tag{47}$$

Both terms use the same MLP parameters. Summing the residual updates in Eq. (24) gives the exact decomposition

$$X_L = X_0 + a_L\sum_{\ell=0}^{L-1}\mathcal{A}_\ell(X_\ell) + b_L\sum_{\ell=0}^{L-1}\mathcal{M}_\ell(X_\ell) + b_L\sum_{\ell=0}^{L-1}\mathcal{I}_\ell(X_\ell). \tag{48}$$

The last term is the accumulated explicit within-block interaction. The decomposition is evaluated at the same residual inputs X_ℓ; it does not compare two separately propagated networks.

Proposition 5 (Local and accumulated interaction bounds)**.** *Under the conditions of Proposition 4, for every $0 \le \ell < L$,*

$$\frac{1}{nT}\mathbb{E}\|\mathcal{I}_\ell(X_\ell)\|_F^2 \le C_\phi^2 Tc_L^2\frac{1}{nT}\mathbb{E}\|X_\ell\|_F^2. \tag{49}$$

For distinct blocks k, ℓ,

$$\mathbb{E}\langle \mathcal{I}_k(X_k), \mathcal{I}_\ell(X_\ell)\rangle_F = 0. \tag{50}$$

Consequently,

$$\frac{1}{nT}\mathbb{E}\left\| b_L \sum_{\ell=0}^{L-1} \mathcal{I}_\ell(X_\ell)\right\|_F^2 = \frac{b_L^2}{nT}\sum_{\ell=0}^{L-1}\mathbb{E}\|\mathcal{I}_\ell(X_\ell)\|_F^2 \le C_\phi^2 T b_L^2 c_L^2 \sum_{\ell=0}^{L-1}\frac{1}{nT}\mathbb{E}\|X_\ell\|_F^2. \tag{51}$$

Proof. By Eq. (22),

$$\mathcal{I}_\ell(X_\ell) = \left[\phi((X_\ell + c_L\mathcal{A}_\ell(X_\ell))W_{1,\ell}^\top) - \phi(X_\ell W_{1,\ell}^\top)\right] W_{2,\ell}^\top. \tag{52}$$

Taking expectation over $W_{2,\ell}$ and using the Lipschitz property of ϕ gives

$$\begin{aligned}\mathbb{E}\|\mathcal{I}_\ell(X_\ell)\|_F^2 &\le C_\phi^2 c_L^2 \mathbb{E}\|\mathcal{A}_\ell(X_\ell)W_{1,\ell}^\top\|_F^2 \\ &= C_\phi^2 c_L^2 \mathbb{E}\|\mathcal{A}_\ell(X_\ell)\|_F^2 \\ &\le C_\phi^2 T c_L^2 \mathbb{E}\|X_\ell\|_F^2. \end{aligned}\tag{53}$$

The equality uses independence of the attention output and $W_{1,\ell}$. Division by nT proves Eq. (49).

For $k < \ell$, let $\mathcal{F}_\ell$ be the sigma-algebra generated by X_0, all parameters in blocks $0, \dots, \ell-1$, and the current-block matrices $W_{Q,\ell}, W_{K,\ell}, W_{V,\ell}, W_{1,\ell}$. Then $\mathcal{I}_k(X_k)$ and the factor multiplying $W_{2,\ell}^\top$ in Eq. (52) are $\mathcal{F}_\ell$-measurable. Since $W_{2,\ell}$ is independent and centered,

$$\mathbb{E}[\mathcal{I}_\ell(X_\ell) \mid \mathcal{F}_\ell] = 0. \tag{54}$$

The tower property gives

$$\mathbb{E}\langle \mathcal{I}_k(X_k), \mathcal{I}_\ell(X_\ell)\rangle_F = \mathbb{E}\langle \mathcal{I}_k(X_k), \mathbb{E}[\mathcal{I}_\ell(X_\ell) \mid \mathcal{F}_\ell]\rangle_F = 0. \tag{55}$$

The case $k > \ell$ follows by symmetry. Expanding the squared norm of the sum and applying Eq. (49) proves Eq. (51). □

Theorem 4 (Accumulated interaction under stable depth scaling)**.** *Assume the conditions Eq. (25) and Eq. (27), with $L[a_L^2 + b_L^2(1 + c_L^2)] = O(1)$. Then, uniformly in width,*

$$\frac{1}{nT}\mathbb{E}\left\| b_L \sum_{\ell=0}^{L-1} \mathcal{I}_\ell(X_\ell)\right\|_F^2 = O(L b_L^2 c_L^2). \tag{56}$$

Proof. Applying Eq. (38) to Eq. (51) gives

$$\begin{aligned}\frac{1}{nT}\mathbb{E}\left\| b_L \sum_{\ell=0}^{L-1} \mathcal{I}_\ell(X_\ell)\right\|_F^2 &\le C_\phi^2 T L b_L^2 c_L^2 \exp\left\{L\left[T a_L^2 + 2C_\phi^2 b_L^2(1 + T c_L^2)\right]\right\} \\ &\quad \times \left[\frac{1}{nT}\mathbb{E}\|X_0\|_F^2 + 2L b_L^2 |\phi(0)|^2\right]. \end{aligned}\tag{57}$$

The exponential and bracketed factors are uniformly bounded under the assumptions. This proves Eq. (56). □

Corollary 2 (Tied and independent interaction coefficients)**.** *Assume the conditions in Theorem 4. Let $a_L = O(L^{-\alpha})$ and $b_L = O(L^{-\beta})$ with $\alpha, \beta \ge 1/2$. For the standard choice $c_L = a_L$,*

$$\left(\frac{1}{nT}\mathbb{E}\left\| b_L \sum_{\ell=0}^{L-1} \mathcal{I}_\ell(X_\ell)\right\|_F^2\right)^{1/2} = O(L^{1/2-\alpha-\beta}). \tag{58}$$

For an independent order-one coefficient $|c_L| = \Theta(1)$,

$$\left(\frac{1}{nT}\mathbb{E}\left\| b_L \sum_{\ell=0}^{L-1} \mathcal{I}_\ell(X_\ell)\right\|_F^2\right)^{1/2} = O(L^{1/2-\beta}). \tag{59}$$

Table 2: Normalized L^2 accumulated-interaction upper bounds at initialization.

$a_L = b_L$	$c_L = a_L$	$\lvert c_L\rvert = \Theta(1)$
$L^{-1/2}$	$O(L^{-1/2})$	$O(1)$
L^{-1}	$O(L^{-3/2})$	$O(L^{-1/2})$

Proof. Both coefficient choices satisfy Eq. (40). Taking the square root of Eq. (56) gives the bound $O(\sqrt{L}\,|b_L c_L|)$; substituting the respective coefficient scalings gives the result. □

As summarized in Table 2, the tied choice yields a vanishing bound under both residual scalings. An independent $c_L = \Theta(1)$ removes the additional decay from tying c_L to a_L, although the initialization bound still vanishes under L^{-1} scaling. The $O(1)$ bound for $L^{-1/2}$ is only an upper bound and does not establish a nonvanishing lower bound.

A.5 Matching necessity for causal attention

We next show that the sufficient depth scaling from Proposition 4 is also necessary under causal attention, given two additional nondegeneracy conditions.

Additional nondegeneracy conditions. Assume first that there exists a constant $c_\phi > 0$, independent of width and depth, such that

$$\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\phi(\sqrt{q}Z)^2\right] \geq c_\phi q, \qquad q \geq 0. \tag{60}$$

We also assume nonvanishing first-token input energy:

$$\inf_{L,n\geq 1} \frac{1}{n}\mathbb{E}\|[X_0]_{1,:}\|_2^2 \geq c_0 > 0. \tag{61}$$

These conditions are verified for the input construction and activations considered in Appendix A.6.

Lemma 2 (Attention lower bound under causal masking)**.** *Assume the initialization setup in Section A.1, causal masking, and Eq. (61). For every* $0 \leq \ell < L$*,*

$$\frac{1}{nT}\mathbb{E}\|\mathcal{A}_\ell(X_\ell)\|_F^2 \geq \frac{c_0}{T}. \tag{62}$$

Proof. Write $P_\ell = P_\ell(X_\ell)$. Under causal masking, $(P_\ell)_{ts} = 0$ for $s > t$, and the first row is $(P_\ell)_{1,:} = (1, 0, \ldots, 0)$. Therefore,

$$\|P_\ell X_\ell\|_F^2 \geq \|[X_\ell]_{1,:}\|_2^2. \tag{63}$$

The same conditional-centering argument as in Eq. (44), applied only to the first row, gives

$$\begin{aligned}\mathbb{E}\|[X_{\ell+1}]_{1,:}\|_2^2 &= \mathbb{E}\|[X_\ell]_{1,:}\|_2^2 + a_L^2\mathbb{E}\|[\mathcal{A}_\ell(X_\ell)]_{1,:}\|_2^2 \\ &\quad + b_L^2\mathbb{E}\|[\mathcal{M}_\ell(H_\ell)]_{1,:}\|_2^2 \\ &\geq \mathbb{E}\|[X_\ell]_{1,:}\|_2^2.\end{aligned} \tag{64}$$

Consequently,

$$\mathbb{E}\|[X_\ell]_{1,:}\|_2^2 \geq \mathbb{E}\|[X_0]_{1,:}\|_2^2 \geq nc_0. \tag{65}$$

Combining Eq. (33), Eq. (63), and Eq. (65) proves the claim. □

Lemma 3 (MLP lower bound)**.** *Under the initialization setup in Section A.1 and Eq. (60),*

$$\frac{1}{nT}\mathbb{E}\|\mathcal{M}_\ell(H_\ell)\|_F^2 \geq c_\phi\frac{1}{nT}\mathbb{E}\|H_\ell\|_F^2. \tag{66}$$

If causal masking and Eq. (61) also hold, then

$$\frac{1}{nT}\mathbb{E}\|\mathcal{M}_\ell(H_\ell)\|_F^2 \geq \frac{c_\phi c_0}{T}(1 + c_L^2). \tag{67}$$

Proof. The MLP input H_ℓ is independent of $W_{1,\ell}$ and $W_{2,\ell}$. Conditional on H_ℓ, each preactivation entry has law

$$[H_\ell W_{1,\ell}^\top]_{tj} \overset{d}{=} \sqrt{\frac{\|[H_\ell]_{t,:}\|_2^2}{n}}\, Z, \qquad Z \sim \mathcal{N}(0,1). \tag{68}$$

First integrating over $W_{2,\ell}$, then over $W_{1,\ell}$, gives

$$\begin{aligned}\frac{1}{nT}\mathbb{E}[\|\mathcal{M}_\ell(H_\ell)\|_F^2 \mid H_\ell] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_Z\left[\phi\left(\sqrt{\frac{\|[H_\ell]_{t,:}\|_2^2}{n}}\, Z\right)^2\right] \\ &\geq \frac{c_\phi}{nT}\|H_\ell\|_F^2. \end{aligned}\tag{69}$$

Taking expectation proves Eq. (66). For the causal case, Eq. (43) and Lemma 2 imply

$$\begin{aligned}\frac{1}{nT}\mathbb{E}\|H_\ell\|_F^2 &= \frac{1}{nT}\mathbb{E}\|X_\ell\|_F^2 + \frac{c_L^2}{nT}\mathbb{E}\|\mathcal{A}_\ell(X_\ell)\|_F^2 \\ &\geq \frac{c_0}{T}(1+c_L^2), \end{aligned}\tag{70}$$

where the first term is bounded below using Eq. (65). Substitution into Eq. (66) gives Eq. (67). □

Theorem 5 (Necessary condition for initialization stability)**.** *Assume the initialization setup in Section A.1, causal masking, Eq. (60), and Eq. (61). If*

$$\sup_{L,n\geq 1}\max_{0\leq\ell\leq L}\frac{1}{nT}\mathbb{E}\|X_\ell\|_F^2 < \infty, \tag{71}$$

then necessarily

$$L[a_L^2 + b_L^2(1+c_L^2)] = O(1). \tag{72}$$

Proof. Sum Eq. (44) over the blocks and apply Lemmas 2 and 3. This gives

$$\frac{1}{nT}\mathbb{E}\|X_L\|_F^2 \geq \frac{1}{nT}\mathbb{E}\|X_0\|_F^2 + \frac{c_0}{T}La_L^2 + \frac{c_\phi c_0}{T}Lb_L^2(1+c_L^2). \tag{73}$$

The left-hand side is uniformly bounded, while $c_0, c_\phi > 0$ and all terms on the right are nonnegative. Thus both La_L^2 and $Lb_L^2(1+c_L^2)$ are bounded, proving the result. □

Corollary 3 (Necessary and sufficient initialization scaling)**.** *Under the assumptions of Theorem 5, including the bounded input moment and Lipschitz activation from Section A.1, Eq. (72) is necessary and sufficient for uniform residual-stream stability at initialization. When $c_L = O(1)$, this condition is equivalent to $L(a_L^2 + b_L^2) = O(1)$ and also ensures the combined bound Eq. (42). In particular, for $a_L = L^{-\alpha}$ and $b_L = L^{-\beta}$ with $c_L = O(1)$, initialization is stable if and only if $\alpha, \beta \geq 1/2$.*

Proof. Combine Proposition 4 with Theorem 5. □

A.6 Verification for language-model inputs and activations

We now verify the input assumptions for token embeddings with absolute sinusoidal positional encoding and the activation lower bound for ReLU, SiLU, and GELU.

Token and positional embeddings. Write $X_0 = SE + P$, where $S \in \{0,1\}^{T\times V}$ is a fixed one-hot token matrix with one nonzero entry in each row, $E \in \mathbb{R}^{V\times n}$ is the effective token embedding matrix, and $P \in \mathbb{R}^{T\times n}$ is the positional encoding. The entries of E are independent, centered, and have variance $\sigma_E^2 = O(1)$; E is independent of the hidden weights. The effective embedding includes any input rescaling used by the implementation. For even n, the sinusoidal encoding is

$$P_{t,2j+1} = \sin\left(t\,10000^{-2j/n}\right), \qquad P_{t,2j+2} = \cos\left(t\,10000^{-2j/n}\right), \qquad 0 \leq j < n/2. \tag{74}$$

The verification below is stated for even widths, using one-based column indices.

Lemma 4 (Input moment verification)**.** *The preceding input construction satisfies*

$$\frac{1}{nT}\mathbb{E}\|X_0\|_F^2 = \frac{1}{n}\mathbb{E}\|[X_0]_{1,:}\|_2^2 = \sigma_E^2 + \frac{1}{2}. \tag{75}$$

Consequently, the bounded input moment and nonvanishing first-token energy assumptions hold. One may take $c_0 = 1/2$ uniformly in width, or $c_0 = \sigma_E^2 + 1/2$ when σ_E^2 is width-independent.

Proof. For every token position t,

$$\|P_{t,:}\|_2^2 = \sum_{j=0}^{n/2-1} \left[\sin^2\left(t\,10000^{-2j/n}\right) + \cos^2\left(t\,10000^{-2j/n}\right)\right] = \frac{n}{2}. \tag{76}$$

Thus $\|P\|_F^2 = nT/2$. Each row of S selects an embedding row, so $\mathbb{E}\|SE\|_F^2 = nT\sigma_E^2$. Repeated tokens do not affect this calculation, which sums coordinate squares. Since the embedding entries are centered and S, P are fixed, $\mathbb{E}\langle SE, P\rangle_F = 0$. Therefore,

$$\frac{1}{nT}\mathbb{E}\|SE + P\|_F^2 = \sigma_E^2 + \frac{1}{2}. \tag{77}$$

The same calculation for the first row gives the other equality in Eq. (75). □

Activation lower bounds. For ReLU and $Z \sim \mathcal{N}(0,1)$,

$$\mathbb{E}[\mathrm{ReLU}(\sqrt{q}Z)^2] = \frac{q}{2}. \tag{78}$$

For $\mathrm{SiLU}(x) = x\sigma(x)$, where σ is the logistic sigmoid, symmetry of Z and $\sigma(-x) = 1 - \sigma(x)$ give

$$\mathbb{E}[\mathrm{SiLU}(\sqrt{q}Z)^2] = \frac{q}{2}\mathbb{E}\left[Z^2\left(\sigma(\sqrt{q}Z)^2 + \sigma(-\sqrt{q}Z)^2\right)\right] \geq \frac{q}{4}, \tag{79}$$

because $s^2 + (1-s)^2 \geq 1/2$. The same argument applies to the exact GELU, $\mathrm{GELU}(x) = x\Phi(x)$, where Φ is the standard normal cumulative distribution function and $\Phi(-x) = 1 - \Phi(x)$. Hence

$$\mathbb{E}[\mathrm{GELU}(\sqrt{q}Z)^2] \geq \frac{q}{4}. \tag{80}$$

Thus Eq. (60) holds with $c_\phi = 1/2$ for ReLU and $c_\phi = 1/4$ for SiLU and GELU.

Theorem 6 (Initialization scaling for causal language Transformers)**.** *Under the initialization of Section A.1, consider causal masking and the token-plus-sinusoidal input construction above, with an activation satisfying Eq. (25) and Eq. (60). Then Eq. (72) is necessary and sufficient for uniform residual-stream stability at initialization. If $c_L = O(1)$, this reduces to $L(a_L^2 + b_L^2) = O(1)$ and also ensures uniform MLP-input moments. Under this stable scaling, the standard $c_L = a_L$ satisfies*

$$\left(\frac{1}{nT}\mathbb{E}\left\|b_L\sum_{\ell=0}^{L-1}\mathcal{I}_\ell(X_\ell)\right\|_F^2\right)^{1/2} = O(L^{-1/2}). \tag{81}$$

For $c_L = \Theta(1)$ and $b_L = \Theta(L^{-1/2})$, the corresponding upper bound is $O(1)$.

Proof. Lemma 4 verifies the input assumptions. Apply Corollary 3 for stability and Theorem 4 for interaction. In the tied case, $L(a_L^2 + b_L^2) = O(1)$ implies $\sqrt{L}\,|a_L b_L| = O(L^{-1/2})$. □

A.7 Full self-attention and Vision Transformer inputs

The sufficient stability condition and accumulated-interaction upper bound derived in Section A.3-A.4 do not require causal masking: their attention estimates use only row stochasticity of $P_\ell(X)$ and independence of the initialized value weights. Hence, for full self-attention without normalization,

$$L\left[a_L^2 + b_L^2(1 + c_L^2)\right] = O(1), \tag{82}$$

$$\frac{1}{nT}\mathbb{E}\left\|b_L\sum_{\ell=0}^{L-1}\mathcal{I}_\ell(X_\ell)\right\|_F^2 = O(Lb_L^2c_L^2) \tag{83}$$

give the same sufficient stability condition and interaction bound. Causal masking enters only the matching necessity argument through the first-token attention lower bound. We next show how a nondegenerate token-Gram matrix can provide an analogous lower bound for full self-attention, and why this condition is natural for Vision Transformer inputs.

A conditional lower bound. The exact identity Eq. (33) can be written as

$$\frac{1}{nT}\mathbb{E}\big[\|\mathcal{A}_\ell(X)\|_F^2 \mid X, W_{Q,\ell}, W_{K,\ell}\big] = \frac{1}{T}\operatorname{tr}\left(P_\ell(X)\frac{XX^\top}{n}P_\ell(X)^\top\right). \tag{84}$$

If $XX^\top/n \succeq \lambda I_T$ for some $\lambda > 0$, then

$$\frac{1}{nT}\mathbb{E}\big[\|\mathcal{A}_\ell(X)\|_F^2 \mid X, W_{Q,\ell}, W_{K,\ell}\big] \geq \frac{\lambda}{T}\|P_\ell(X)\|_F^2 \geq \frac{\lambda}{T}. \tag{85}$$

The last inequality follows because each probability row has squared Euclidean norm at least $1/T$, so $\|P_\ell(X)\|_F^2 \geq 1$. Thus a uniform token-Gram lower bound would provide an alternative to the causal first-token argument.

Positional embeddings at initialization. For Vision Transformer inputs, write $X_0 = X_{\text{patch}} + P$, where X_{patch} contains the projected patches. For illustration, suppose the positional entries are i.i.d., centered, with finite variance $\sigma_P^2 > 0$, independently of the patch projection. For fixed T, the law of large numbers gives

$$\frac{1}{n}PP^\top \xrightarrow[n\to\infty]{\mathbb{P}} \sigma_P^2 I_T. \tag{86}$$

If also

$$\frac{1}{n}X_{\text{patch}}X_{\text{patch}}^\top \xrightarrow[n\to\infty]{\mathbb{P}} \Sigma_{\text{patch}} \succeq 0, \qquad \frac{1}{n}X_{\text{patch}}P^\top \xrightarrow[n\to\infty]{\mathbb{P}} 0, \tag{87}$$

then

$$\frac{1}{n}X_0X_0^\top \xrightarrow[n\to\infty]{\mathbb{P}} \Sigma_{\text{patch}} + \sigma_P^2 I_T \succeq \sigma_P^2 I_T. \tag{88}$$

More generally, the same conclusion follows from a deterministic or random positional representation satisfying $PP^\top/n \to K_P \succeq \lambda_P I_T$, provided the patch Gram matrix converges and the patch–position cross terms vanish. The essential condition in this argument is a nondegenerate limiting positional Gram matrix, together with the stated cross-term control.

Persistence through depth. The token-Gram lower bound can plausibly persist through depth because the new within-block contributions are centered or positive semidefinite. Indeed,

$$\frac{1}{n}H_\ell H_\ell^\top = \frac{1}{n}X_\ell X_\ell^\top + \frac{c_L}{n}\left[X_\ell\mathcal{A}_\ell(X_\ell)^\top + \mathcal{A}_\ell(X_\ell)X_\ell^\top\right] + \frac{c_L^2}{n}\mathcal{A}_\ell(X_\ell)\mathcal{A}_\ell(X_\ell)^\top. \tag{89}$$

Conditional centering of $W_{V,\ell}$ makes the attention cross term zero in expectation. If, for fixed depth,

$$\frac{1}{n}X_\ell\mathcal{A}_\ell(X_\ell)^\top \xrightarrow[n\to\infty]{\mathbb{P}} 0, \tag{90}$$

then, since the final term in Eq. (89) is positive semidefinite,

$$\lambda_{\min}\left(\frac{1}{n}H_\ell H_\ell^\top\right) \geq \lambda_{\min}\left(\frac{1}{n}X_\ell X_\ell^\top\right) - o_{\mathbb{P}}(1). \tag{91}$$

An analogous argument applies to $X_{\ell+1}X_{\ell+1}^\top/n$, with the MLP cross terms centered through $W_{2,\ell}$. This suggests that token-Gram nondegeneracy can propagate across fixed depth.

These observations suggest that the token-Gram nondegeneracy needed for the full-attention lower bound can persist through depth, which would yield the same matching necessity condition as in the causal case. Establishing this uniformly when width and depth grow jointly requires additional concentration estimates and is left to future work.

B TRAINING-TIME ANALYSIS

This appendix develops the training-time analysis. We first present the exact finite-width forward, backward, and parameter-update identities. We then use the Tensor Programs framework and its Master Theorem Yang & Hu (2021) to derive the infinite-width representations of the learned linear fields as $n \to \infty$, with depth and the number of training steps fixed. Building on these representations, we analyze the initialized forward and backward passes and the forward pass after one parameter update to derive sufficient stability conditions. Under these conditions, we then quantify the accumulated attention-to-MLP interaction after one parameter update.

B.1 MODEL, INITIALIZATION, AND NOTATION

Recall from Section 3 that we consider a depth-L, width-n Transformer with single-head attention and no normalization. The sequence length T is fixed throughout the analysis. Let $X_\ell \in \mathbb{R}^{T\times n}$ denote the residual-stream representation entering layer ℓ.

For a fixed input sequence, let $S \in \{0,1\}^{T\times V}$ be its one-hot token matrix, with one unit in each row and vocabulary size V fixed independently of width. The input representation at training step k is

$$X_0^k = SE^k, \tag{92}$$

where $E^k \in \mathbb{R}^{V\times n}$ is the trainable token embedding matrix. Initialize

$$(E^0)_{vj} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_E^2), \tag{93}$$

where $\sigma_E^2 > 0$ is independent of width and depth. The embedding initialization is independent of the hidden and readout initializations. We omit positional encoding in the training-time analysis.

For each layer ℓ, define the query, key, and value matrices by

$$Q_\ell = X_\ell W_{Q,\ell}^\top, \qquad K_\ell = X_\ell W_{K,\ell}^\top, \qquad V_\ell = X_\ell W_{V,\ell}^\top, \tag{94}$$

where $Q_\ell, K_\ell, V_\ell \in \mathbb{R}^{T\times n}$. In particular, the value dimension equals the residual width n, so the value mixture can be added directly to the residual stream. We define the attention logits, attention weights, and attention output by

$$S_\ell := \frac{1}{n} Q_\ell K_\ell^\top \in \mathbb{R}^{T\times T}, \qquad P_\ell := \mathrm{softmax}(S_\ell) \in \mathbb{R}^{T\times T}, \tag{95}$$

$$A_\ell := P_\ell V_\ell \in \mathbb{R}^{T\times n}. \tag{96}$$

Thus, the single-head attention map is

$$\mathcal{A}_\ell(X_\ell) = A_\ell = \mathrm{softmax}\left(\frac{Q_\ell K_\ell^\top}{n}\right) V_\ell. \tag{97}$$

The softmax is applied rowwise. A deterministic attention mask may be included inside the softmax and is omitted from the notation unless needed explicitly. We use the $1/n$ attention-logit scaling from the μP parameterization. We omit a separate attention output projection as a modeling choice. The value matrix is initialized and trained directly; it is not identified with the product of independently initialized value and output matrices.

The input to the MLP is

$$H_\ell = X_\ell + c_L A_\ell. \tag{98}$$

To streamline the mean-field notation, we relabel the MLP weights as

$$W_{U,\ell} := W_{1,\ell}, \qquad W_{M,\ell} := W_{2,\ell}. \tag{99}$$

Thus, the MLP preactivation, postactivation, and output are given by

$$U_\ell := H_\ell W_{U,\ell}^\top, \qquad \Phi_\ell := \phi(U_\ell), \qquad M_\ell := \Phi_\ell W_{M,\ell}^\top, \tag{100}$$

where $H_\ell, U_\ell, \Phi_\ell, M_\ell \in \mathbb{R}^{T\times n}$. Hence,

$$\mathcal{M}_\ell(H_\ell) = M_\ell. \tag{101}$$

The residual block considered in this paper is

$$X_{\ell+1} = X_\ell + a_L A_\ell + b_L M_\ell, \tag{102}$$

The standard sequential parameterization sets $c_L = a_L$. We instead treat the interaction coefficient c_L separately from the outer residual multipliers a_L, b_L. In the analysis below, c_L is a deterministic coefficient, fixed during the update and satisfying $c_L = O(1)$. This includes both the tied choice and an independent order-one coefficient. Input-dependent learned gates require additional derivative terms and are not covered by the scalar-coefficient calculation.

The entries of all weight matrices are initialized independently according to

$$(W^0_{\theta,\ell})_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, n^{-1}), \qquad \theta \in \{Q, K, V, U, M\}, \tag{103}$$

independently across matrices and layers. Biases are omitted. All hidden matrices have shape $n \times n$. The input is independent of the hidden initialization. We use the output-based MLP weight notation $W_{U,\ell} := W_{1,\ell}$ and $W_{M,\ell} := W_{2,\ell}$, matching the two MLP matrices in the main paper. Thus η_U and η_M are the learning rates of $W_{1,\ell}$ and $W_{2,\ell}$, respectively.

B.2 Loss, scaled adjoints, and gradient-descent updates

With the model and initialization fixed, we now specify the loss, scaled adjoints, and gradient-descent dynamics used throughout the training-time analysis.

Given the final residual-stream representation $X_L \in \mathbb{R}^{T\times n}$, define the tokenwise network output by

$$f := X_L W_{\text{out}}^\top \in \mathbb{R}^{T\times C}, \tag{104}$$

where C is fixed as $n \to \infty$. Following the μP readout scaling, initialize

$$(W^0_{\text{out}})_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, n^{-2}). \tag{105}$$

The sample loss is

$$\mathcal{L} := \ell_{\text{loss}}(f, y), \tag{106}$$

where ℓ_{loss} may include any fixed task-dependent token selection, masking, or aggregation applied to f. We assume that the output gradient is Lipschitz:

$$\|\nabla_1 \ell_{\text{loss}}(f, y) - \nabla_1 \ell_{\text{loss}}(f', y)\|_F \le C_{\text{loss}} \|f - f'\|_F. \tag{107}$$

Remark 3. *For softmax cross-entropy,* $\mathrm{CE}(z, y) = -\log(\mathrm{softmax}(z)^\top y)$*, the Hessian is*

$$\nabla_z^2 \,\mathrm{CE} = \mathrm{diag}(p) - pp^\top, \qquad p = \mathrm{softmax}(z), \tag{108}$$

whose operator norm is at most $1/4$*. Hence its gradient is globally* $1/4$*-Lipschitz in the operator norm and* $1/2$*-Lipschitz in the Frobenius norm. For squared loss,*

$$\ell_{\text{loss}}(f, y) = \frac{1}{2}\|f - y\|_F^2, \tag{109}$$

we have $\nabla_f \ell_{\text{loss}} = f - y$ *and may take* $C_{\text{loss}} = 1$*.*

Scaled adjoints. Define the output adjoint by

$$\delta f := \nabla_1 \ell_{\text{loss}}(f, y) \in \mathbb{R}^{T\times C}. \tag{110}$$

For every residual-stream variable X_ℓ, define

$$\delta X_\ell := n \nabla_{X_\ell} \mathcal{L}. \tag{111}$$

Since $f = X_L W_{\text{out}}^\top$, the terminal residual-stream adjoint is

$$\delta X_L = n\, \delta f\, W_{\text{out}}. \tag{112}$$

Under Eq. (105), nW^0_{out} has order-one entries. The corresponding initialization bound for δX_L is established below, and we assume the terminal adjoint remains uniformly bounded over the fixed training horizons considered in the subsequent analysis.

Assume $a_L, b_L > 0$. For the internal MLP variables, define the coefficient-stripped adjoints by

$$\delta Y := b_L^{-1} n \nabla_Y \mathcal{L}, \qquad Y \in \{M_\ell, \Phi_\ell, U_\ell, H_\ell\}. \tag{113}$$

For the internal attention variables, define

$$\delta Y := a_L^{-1} n \nabla_Y \mathcal{L}, \qquad Y \in \{A_\ell, Q_\ell, K_\ell, V_\ell\}. \tag{114}$$

For the fixed-dimensional attention matrices, define

$$\delta P_\ell := a_L^{-1} \nabla_{P_\ell} \mathcal{L}, \qquad \delta S_\ell := a_L^{-1} \nabla_{S_\ell} \mathcal{L}. \tag{115}$$

The factor n is omitted from δP_ℓ and δS_ℓ because $P_\ell, S_\ell \in \mathbb{R}^{T\times T}$ and T is fixed. These normalizations keep the residual multipliers a_L and b_L explicit in the backward equations. In particular, the attention-output adjoint below contains the additional ratio $b_L c_L / a_L$.

MLP backward pass. From Eq. (102), we have

$$\delta M_\ell = b_L^{-1} n \nabla_{M_\ell} \mathcal{L} = \delta X_{\ell+1}. \tag{116}$$

Backpropagating through the MLP gives

$$\delta \Phi_\ell = \delta M_\ell W_{M,\ell}, \tag{117}$$

$$\delta U_\ell = \delta \Phi_\ell \odot \phi'(U_\ell), \tag{118}$$

$$\delta H_\ell = \delta U_\ell W_{U,\ell}. \tag{119}$$

Equivalently,

$$\delta H_\ell = \left[(\delta X_{\ell+1} W_{M,\ell}) \odot \phi'(U_\ell) \right] W_{U,\ell}. \tag{120}$$

Hence

$$n \nabla_{H_\ell} \mathcal{L} = b_L \delta H_\ell. \tag{121}$$

Adjoint of the attention output. The attention output A_ℓ reaches the loss through the residual update and through the MLP input $H_\ell = X_\ell + c_L A_\ell$. Therefore

$$\begin{aligned} n \nabla_{A_\ell} \mathcal{L} &= a_L \delta X_{\ell+1} + c_L \, n \nabla_{H_\ell} \mathcal{L} \\ &= a_L \delta X_{\ell+1} + b_L c_L \delta H_\ell. \end{aligned} \tag{122}$$

Thus

$$\delta A_\ell = \delta X_{\ell+1} + \frac{b_L c_L}{a_L} \delta H_\ell, \tag{123}$$

Attention backward pass. Since $A_\ell = P_\ell V_\ell$, the value and attention-weight adjoints are

$$\delta V_\ell = P_\ell^\top \delta A_\ell, \tag{124}$$

$$\delta P_\ell = \frac{1}{n} \delta A_\ell V_\ell^\top. \tag{125}$$

Because the softmax is applied rowwise,

$$\delta S_\ell = P_\ell \odot \left(\delta P_\ell - \left[(\delta P_\ell \odot P_\ell) \mathbf{1}_T \right] \mathbf{1}_T^\top \right). \tag{126}$$

Equivalently, for $i, j \in [T]$,

$$(\delta S_\ell)_{ij} = \frac{(P_\ell)_{ij}}{n} \left\langle (\delta A_\ell)_{i,:}, (V_\ell)_{j,:} - (A_\ell)_{i,:} \right\rangle. \tag{127}$$

Using $S_\ell = n^{-1} Q_\ell K_\ell^\top$ then gives

$$\delta Q_\ell = \delta S_\ell K_\ell, \qquad \delta K_\ell = \delta S_\ell^\top Q_\ell. \tag{128}$$

Residual-stream backward recursion. Define the coefficient-stripped backward fields induced by the query, key, and value projections:

$$\delta X_{Q,\ell} := \delta Q_\ell W_{Q,\ell}, \qquad \delta X_{K,\ell} := \delta K_\ell W_{K,\ell}, \qquad \delta X_{V,\ell} := \delta V_\ell W_{V,\ell}. \tag{129}$$

The residual-stream adjoint satisfies

$$\delta X_\ell = \delta X_{\ell+1} + b_L \delta H_\ell + a_L \left(\delta X_{Q,\ell} + \delta X_{K,\ell} + \delta X_{V,\ell} \right). \tag{130}$$

Parameter gradients. The parameter gradients are

$$\nabla_E \mathcal{L} = \frac{1}{n} S^\top \delta X_0, \tag{131}$$

$$\nabla_{W_{M,\ell}} \mathcal{L} = \frac{b_L}{n} \delta M_\ell^\top \Phi_\ell, \tag{132}$$

$$\nabla_{W_{U,\ell}} \mathcal{L} = \frac{b_L}{n} \delta U_\ell^\top H_\ell, \tag{133}$$

$$\nabla_{W_{Q,\ell}} \mathcal{L} = \frac{a_L}{n} \delta Q_\ell^\top X_\ell, \tag{134}$$

$$\nabla_{W_{K,\ell}} \mathcal{L} = \frac{a_L}{n} \delta K_\ell^\top X_\ell, \tag{135}$$

$$\nabla_{W_{V,\ell}} \mathcal{L} = \frac{a_L}{n} \delta V_\ell^\top X_\ell, \tag{136}$$

$$\nabla_{W_{\text{out}}} \mathcal{L} = (\delta f)^\top X_L. \tag{137}$$

The embedding gradient sums the residual-input gradients over occurrences of each token. The MLP and attention gradients carry the outer residual multipliers b_L and a_L, respectively.

Gradient-descent updates. Following the μP scaling, we use learning rates $n\eta_E$ for the embedding, η_θ for the hidden weights $\theta \in \{Q, K, V, U, M\}$, and η_{out}/n for the readout. The coefficients $\eta_E, \eta_\theta, \eta_{\text{out}} \geq 0$ are independent of width and fixed throughout training. All parameters are updated simultaneously. For $k \geq 1$,

$$E^k = E^{k-1} - \eta_E S^\top \delta X_0^{k-1}, \tag{138}$$

$$W_{M,\ell}^k = W_{M,\ell}^{k-1} - \eta_M \frac{b_L}{n} (\delta M_\ell^{k-1})^\top \Phi_\ell^{k-1}, \tag{139}$$

$$W_{U,\ell}^k = W_{U,\ell}^{k-1} - \eta_U \frac{b_L}{n} (\delta U_\ell^{k-1})^\top H_\ell^{k-1}, \tag{140}$$

$$W_{Q,\ell}^k = W_{Q,\ell}^{k-1} - \eta_Q \frac{a_L}{n} (\delta Q_\ell^{k-1})^\top X_\ell^{k-1}, \tag{141}$$

$$W_{K,\ell}^k = W_{K,\ell}^{k-1} - \eta_K \frac{a_L}{n} (\delta K_\ell^{k-1})^\top X_\ell^{k-1}, \tag{142}$$

$$W_{V,\ell}^k = W_{V,\ell}^{k-1} - \eta_V \frac{a_L}{n} (\delta V_\ell^{k-1})^\top X_\ell^{k-1}, \tag{143}$$

$$W_{\text{out}}^k = W_{\text{out}}^{k-1} - \frac{\eta_{\text{out}}}{n} (\delta f^{k-1})^\top X_L^{k-1}. \tag{144}$$

Unrolling from initialization gives

$$E^k = E^0 - \eta_E \sum_{s=0}^{k-1} S^\top \delta X_0^s, \tag{145}$$

$$W_{M,\ell}^k = W_{M,\ell}^0 - \frac{\eta_M b_L}{n} \sum_{s=0}^{k-1} (\delta M_\ell^s)^\top \Phi_\ell^s, \tag{146}$$

$$W_{U,\ell}^k = W_{U,\ell}^0 - \frac{\eta_U b_L}{n} \sum_{s=0}^{k-1} (\delta U_\ell^s)^\top H_\ell^s, \tag{147}$$

$$W_{Q,\ell}^k = W_{Q,\ell}^0 - \frac{\eta_Q a_L}{n} \sum_{s=0}^{k-1} (\delta Q_\ell^s)^\top X_\ell^s, \tag{148}$$

$$W_{K,\ell}^k = W_{K,\ell}^0 - \frac{\eta_K a_L}{n} \sum_{s=0}^{k-1} (\delta K_\ell^s)^\top X_\ell^s, \tag{149}$$

$$W_{V,\ell}^k = W_{V,\ell}^0 - \frac{\eta_V a_L}{n} \sum_{s=0}^{k-1} (\delta V_\ell^s)^\top X_\ell^s, \tag{150}$$

$$W_{\text{out}}^k = W_{\text{out}}^0 - \frac{\eta_{\text{out}}}{n} \sum_{s=0}^{k-1} (\delta f^s)^\top X_L^s. \tag{151}$$

B.3 Tensor-program limit and mean-field feature dynamics

For fixed depth L and training horizon K, substituting the cumulative parameter updates into the forward and backward equations expresses the training dynamics as a Tensor Program (TP). Its Master Theorem then identifies the infinite-width empirical limits with expectations over the corresponding mean-field variables. We then apply its limiting rules to the learned linear fields to obtain their mean-field dynamics, and combine these with the embedding, attention, residual, and readout relations to close the full Transformer training dynamics.

B.3.1 Tensor-program representation and empirical convergence

Fix the depth L, sequence length T, vocabulary size V, output dimension C, and training horizon K, and let $n \to \infty$. We work under the model, initialization, and batch-size-one gradient-descent setup of Sections B.1 and B.2, including the trainable embedding and readout and the stated width scalings of all learning rates. All coefficients other than the width n are held fixed.

A map ψ is pseudo-Lipschitz if there exist $p \geq 0$ and $C_\psi < \infty$ such that

$$|\psi(u) - \psi(v)| \leq C_\psi \big(1 + |u|^p + |v|^p\big)|u - v|. \tag{152}$$

We additionally assume $\phi \in C^2(\mathbb{R})$ with bounded first and second derivatives and retain the loss-gradient regularity condition Eq. (107). That is, ϕ, ϕ', and $\nabla_1 \ell(f, y)$ are Lipschitz continuous.

After substituting the cumulative parameter updates, the resulting computation uses only initialized Gaussian matrix multiplications and their transposes, coordinatewise operations, and empirical averages over width, and hence falls within the TP framework of (Yang & Hu, 2021, Definition G.1).

Proposition 6 (Tensor-program representation and empirical convergence)**.** *Fix the depth L and training horizon K, and let $n \to \infty$. Assume that ϕ, ϕ', and the loss gradient are Lipschitz. Then the Transformer training computation under the μP initialization and updates in Sections B.1–B.2, through step K, admits a finite Tensor-Program representation. Let $z_n^1, \ldots, z_n^r \in \mathbb{R}^n$ be any fixed collection of vectors in this program, and let $\bar{z}^1, \ldots, \bar{z}^r$ be their joint Tensor-Program representatives (Yang & Hu, 2021, Definition G.3). For every pseudo-Lipschitz $\psi : \mathbb{R}^r \to \mathbb{R}$,*

$$\frac{1}{n} \sum_{j=1}^{n} \psi\big((z_n^1)_j, \ldots, (z_n^r)_j\big) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}\left[\psi(\bar{z}^1, \ldots, \bar{z}^r)\right]. \tag{153}$$

Every scalar variable in the program converges a.s. to its deterministic Tensor-Program limit.

Proof. The rows of E^0 and nW_{out}^0 form the initial Gaussian vectors, independently of the hidden Gaussian matrices. The cumulative embedding update Eq. (145) gives

$$E^k = E^0 - \eta_E \sum_{s<k} S^\top \delta X_0^s, \qquad X_0^k = SE^0 - \eta_E \sum_{s<k} SS^\top \delta X_0^s. \tag{154}$$

Since S has fixed dimensions, these are finite linear combinations of initial embedding vectors and previously computed adjoints.

For a hidden map $Y^k = Z^k W^{k\top}$ and its induced backward field $\delta Z^k = \delta Y^k W^k$, let $\rho \in \{a_L, b_L\}$ be its residual multiplier and η its learning-rate coefficient. Substitution of the weight update yields

$$Y^k = Z^k W^{0\top} - \eta\rho \sum_{s<k} \left(\frac{1}{n} Z^k Z^{s\top}\right) \delta Y^s, \tag{155}$$

$$\delta Z^k = \delta Y^k W^0 - \eta\rho \sum_{s<k} \left(\frac{1}{n} \delta Y^k \delta Y^{s\top}\right) Z^s. \tag{156}$$

The first terms are initialized-matrix multiplications. Each entry of a kernel is a coordinate average, and multiplication by that kernel is a finite linear combination of token rows with scalar coefficients.

Attention uses the same operations. In particular,

$$(S_\ell^k)_{tu} = \frac{1}{n} \sum_{j=1}^{n} (Q_\ell^k)_{tj} (K_\ell^k)_{uj}, \qquad (\delta P_\ell^k)_{tu} = \frac{1}{n} \sum_{j=1}^{n} (\delta A_\ell^k)_{tj} (V_\ell^k)_{uj}. \tag{157}$$

Softmax is a continuous function of finitely many score scalars. Attention mixing, activation, residual addition, and their backward operations are coordinatewise functions of vectors and previously computed scalars.

For the readout, Eq. (151) gives

$$f^k = \frac{1}{n} X_L^k (nW_{\text{out}}^0)^\top - \eta_{\text{out}} \sum_{s<k} \left(\frac{1}{n} X_L^k X_L^{s\top}\right) \delta f^s, \tag{158}$$

$$\delta X_L^k = \delta f^k (nW_{\text{out}}^0) - \eta_{\text{out}} \sum_{s<k} \delta f^k (\delta f^s)^\top X_L^s. \tag{159}$$

Thus the output entries are coordinate averages, the output gradient is a continuous function of these scalars, and the terminal adjoint is a finite linear combination of program vectors. The cumulative update also expresses each row of nW_{out}^k as such a linear combination.

For fixed L, T, V, C, K, the construction has finitely many instructions. Functions involving vector coordinates are pseudo-Lipschitz jointly in those coordinates and their scalar arguments; scalar-only functions are continuous. These properties verify Setup G.2 and Assumption L.4 of Yang & Hu (2021). Theorem G.4 of that reference therefore gives Eq. (153) and scalar convergence. □

Throughout this section, bars denote the joint Tensor-Program representatives from Proposition 6. For $Y^k \in \mathbb{R}^{T\times n}$, let $\bar{y}^k \in \mathbb{R}^T$ collect the representative coordinates of its rows. Applying Proposition 6 to coordinate products gives

$$\frac{1}{n} Y^k Z^{s\top} \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[\bar{y}^k \bar{z}^{s\top}]. \tag{160}$$

For same-field moments, we write

$$\bar{C}_{y,\ell}^{k,s} := \mathbb{E}[\bar{y}_\ell^k \bar{y}_\ell^{s\top}], \tag{161}$$

which denotes an uncentered second-moment kernel. Bars on fixed-dimensional scalar arrays denote their deterministic limits, and sums over $s < k$ mean $s = 0, \ldots, k-1$.

B.3.2 Query dynamics as a representative example

We use the query field as a representative example of the Tensor-Program limiting rules. Substituting the cumulative query-weight update separates each forward and backward field into an initialized-weight term and an explicit learning drift. The former is determined by the ZHat/ZDot rules, while the latter follows from empirical-kernel convergence.

Finite-width identities. Fix a layer ℓ and define

$$Q_\ell^{k|0} := X_\ell^k W_{Q,\ell}^{0\top}, \tag{162}$$

$$\delta X_{Q,\ell}^{k|0} := \delta Q_\ell^k W_{Q,\ell}^0, \tag{163}$$

$$C_{X,\ell}^{k,s} := \frac{1}{n} X_\ell^k X_\ell^{s\top}, \tag{164}$$

$$C_{\delta Q,\ell}^{k,s} := \frac{1}{n} \delta Q_\ell^k \delta Q_\ell^{s\top}. \tag{165}$$

The superscript $k \mid 0$ indicates that the step-k input or adjoint is evaluated using the initialized query matrix. Substituting Eq. (148) gives the exact identities

$$Q_\ell^k = Q_\ell^{k|0} - \eta_Q a_L \sum_{s<k} C_{X,\ell}^{k,s} \delta Q_\ell^s, \tag{166}$$

$$\delta X_{Q,\ell}^k = \delta X_{Q,\ell}^{k|0} - \eta_Q a_L \sum_{s<k} C_{\delta Q,\ell}^{k,s} X_\ell^s. \tag{167}$$

Lemma 5 (Query mean-field dynamics)**.** *Under Proposition 6, the initialized-weight fields $Q_\ell^{k|0}$ and $\delta X_{Q,\ell}^{k|0}$ admit the mean-field representations*

$$\bar{q}_\ell^{k|0} = \xi_{Q,\ell}^k + \sum_{s<k} R_{x,\zeta_Q,\ell}^{k,s} \delta \bar{q}_\ell^s, \tag{168}$$

$$\delta \bar{x}_{Q,\ell}^{k|0} = \zeta_{Q,\ell}^k + \sum_{s\le k} R_{\delta q,\xi_Q,\ell}^{k,s} \bar{x}_\ell^s. \tag{169}$$

Here $\xi_{Q,\ell}^k, \zeta_{Q,\ell}^k \in \mathbb{R}^T$ are the centered Gaussian components associated with the forward and backward uses of $W_{Q,\ell}^0$, with covariances

$$\mathbb{E}\left[\xi_{Q,\ell}^k \xi_{Q,\ell}^{s\top}\right] = \bar{C}_{x,\ell}^{k,s}, \qquad \mathbb{E}\left[\zeta_{Q,\ell}^k \zeta_{Q,\ell}^{s\top}\right] = \bar{C}_{\delta q,\ell}^{k,s}. \tag{170}$$

The response matrices are

$$R_{x,\zeta_Q,\ell}^{k,s} := \mathbb{E}\left[D_{\zeta_{Q,\ell}^s} \bar{x}_\ell^k\right], \quad s<k; \qquad R_{\delta q,\xi_Q,\ell}^{k,s} := \mathbb{E}\left[D_{\xi_{Q,\ell}^s} \delta \bar{q}_\ell^k\right], \quad s \le k, \tag{171}$$

and are zero outside the indicated ranges. Consequently, the learned query fields satisfy

$$\bar{q}_\ell^k = \xi_{Q,\ell}^k + \sum_{s<k} \left[R_{x,\zeta_Q,\ell}^{k,s} - \eta_Q a_L \bar{C}_{x,\ell}^{k,s}\right] \delta \bar{q}_\ell^s, \tag{172}$$

$$\delta \bar{x}_{Q,\ell}^k = \zeta_{Q,\ell}^k + \sum_{s\le k} R_{\delta q,\xi_Q,\ell}^{k,s} \bar{x}_\ell^s - \eta_Q a_L \sum_{s<k} \bar{C}_{\delta q,\ell}^{k,s} \bar{x}_\ell^s. \tag{173}$$

The forward and backward Gaussian families are independent of one another and of the Gaussian families associated with distinct initialized matrices.

Proof. We apply the Tensor-Program rules (Yang & Hu, 2021, Definition G.3) directly to the initialized-weight terms in Eq. (166) and Eq. (167).

For the forward field

$$Q_\ell^{k|0} = X_\ell^k W_{Q,\ell}^{0\top},$$

the earlier uses of the same initialized matrix in the opposite orientation are

$$\delta X_{Q,\ell}^{s|0} = \delta Q_\ell^s W_{Q,\ell}^0, \qquad s<k.$$

Their input representatives are $\delta \bar{q}_\ell^s$, and their Gaussian components are $\zeta_{Q,\ell}^s$. The ZHat rule contributes the centered Gaussian field $\xi_{Q,\ell}^k$, while the ZDot rule contributes, for token t,

$$\sum_{s<k} \sum_{u=1}^T \mathbb{E}\left[\frac{\partial(\bar{x}_\ell^k)_t}{\partial(\zeta_{Q,\ell}^s)_u}\right] (\delta \bar{q}_\ell^s)_u. \tag{174}$$

Here the derivative is taken with respect to the corresponding Gaussian argument in the recursively constructed Tensor-Program representation, with the remaining Gaussian arguments and deterministic limiting scalars held fixed (Yang & Hu, 2021, Remark L.1). Collecting the token coordinates gives Eq. (168), with the response defined in Eq. (171).

For the backward field

$$\delta X^{k|0}_{Q,\ell} = \delta Q^k_\ell W^0_{Q,\ell},$$

the opposite-orientation uses are now

$$Q^{s|0}_\ell = X^s_\ell W^{0\top}_{Q,\ell}, \qquad s \le k,$$

because the step-k forward pass precedes its backward pass. Their input representatives are $\bar{x}^s_\ell$, and their Gaussian components are $\xi^s_{Q,\ell}$. Thus ZDot contributes

$$\sum_{s\le k}\sum_{u=1}^{T} \mathbb{E}\left[\frac{\partial(\delta\bar{q}^k_\ell)_t}{\partial(\xi^s_{Q,\ell})_u}\right](\bar{x}^s_\ell)_u, \tag{175}$$

which gives Eq. (169), with the response again defined in Eq. (171). The inclusion of $s = k$ reflects the within-step ordering of the forward and backward computations.

The ZHat rule gives the Gaussian covariances in Eq. (170) and the stated independence structure.

It remains to account for the explicit learning drifts. By Eq. (160),

$$C^{k,s}_{X,\ell} \xrightarrow{\text{a.s.}} \bar{C}^{k,s}_{x,\ell}, \qquad C^{k,s}_{\delta Q,\ell} \xrightarrow{\text{a.s.}} \bar{C}^{k,s}_{\delta q,\ell}. \tag{176}$$

Substituting these limits together with Eq. (168) and Eq. (169) into Eq. (166) and Eq. (167) gives Eq. (172) and Eq. (173). □

The query calculation separates the two effects of training: opposite-orientation reuse of the initialized matrix generates the response terms through ZDot, while the parameter updates generate the explicit learning drifts through the limiting empirical kernels. The same two mechanisms apply to every learned linear field.

B.3.3 Extension to all learned linear fields

The query calculation depends only on the structure of a learned linear map, its transposed use in backpropagation, and its cumulative parameter update. The same Tensor-Program rules therefore apply to every learned linear field in the Transformer.

For each $\theta \in \{Q, K, V, U, M\}$, write

$$Y^k_{\theta,\ell} = Z^k_{\theta,\ell} W^{k\top}_{\theta,\ell}, \qquad \delta Z^k_{\theta,\ell} = \delta Y^k_{\theta,\ell} W^k_{\theta,\ell}, \tag{177}$$

with

θ	$Z_{\theta,\ell}$	$Y_{\theta,\ell}$	$\delta Y_{\theta,\ell}$	$\delta Z_{\theta,\ell}$	$\rho_{\theta,L}$
Q	X_ℓ	Q_ℓ	δQ_ℓ	$\delta X_{Q,\ell}$	a_L
K	X_ℓ	K_ℓ	δK_ℓ	$\delta X_{K,\ell}$	a_L
V	X_ℓ	V_ℓ	δV_ℓ	$\delta X_{V,\ell}$	a_L
U	H_ℓ	U_ℓ	δU_ℓ	δH_ℓ	b_L
M	Φ_ℓ	M_ℓ	δM_ℓ	$\delta \Phi_\ell$	b_L

(178)

The cumulative update has the common form

$$W^k_{\theta,\ell} = W^0_{\theta,\ell} - \frac{\eta_\theta \rho_{\theta,L}}{n} \sum_{s<k} (\delta Y^s_{\theta,\ell})^\top Z^s_{\theta,\ell}. \tag{179}$$

Define the initialized-weight fields

$$Y^{k|0}_{\theta,\ell} := Z^k_{\theta,\ell} W^{0\top}_{\theta,\ell}, \qquad \delta Z^{k|0}_{\theta,\ell} := \delta Y^k_{\theta,\ell} W^0_{\theta,\ell}. \tag{180}$$

Then

$$Y^k_{\theta,\ell} = Y^{k|0}_{\theta,\ell} - \eta_\theta \rho_{\theta,L} \sum_{s<k} C^{k,s}_{Z_\theta,\ell} \delta Y^s_{\theta,\ell}, \tag{181}$$

$$\delta Z^k_{\theta,\ell} = \delta Z^{k|0}_{\theta,\ell} - \eta_\theta \rho_{\theta,L} \sum_{s<k} C^{k,s}_{\delta Y_\theta,\ell} Z^s_{\theta,\ell}, \tag{182}$$

where

$$C^{k,s}_{Z_\theta,\ell} := \frac{1}{n} Z^k_{\theta,\ell} Z^{s\top}_{\theta,\ell}, \qquad C^{k,s}_{\delta Y_\theta,\ell} := \frac{1}{n} \delta Y^k_{\theta,\ell} \delta Y^{s\top}_{\theta,\ell}. \tag{183}$$

Lemma 6 (Mean-field dynamics of the learned linear fields)**.** *Under Proposition 6, for every $\theta \in \{Q, K, V, U, M\}$, the initialized-weight fields admit the representations*

$$\bar{y}^{k|0}_{\theta,\ell} = \xi^k_{\theta,\ell} + \sum_{s<k} R^{k,s}_{z_\theta,\zeta_\theta,\ell} \delta \bar{y}^s_{\theta,\ell}, \tag{184}$$

$$\delta \bar{z}^{k|0}_{\theta,\ell} = \zeta^k_{\theta,\ell} + \sum_{s\le k} R^{k,s}_{\delta y_\theta,\xi_\theta,\ell} \bar{z}^s_{\theta,\ell}. \tag{185}$$

Here $\xi^k_{\theta,\ell}$ and $\zeta^k_{\theta,\ell}$ are centered Gaussian components satisfying

$$\mathbb{E}\left[\xi^k_{\theta,\ell} \xi^{s\top}_{\theta,\ell}\right] = \bar{C}^{k,s}_{z_\theta,\ell}, \qquad \mathbb{E}\left[\zeta^k_{\theta,\ell} \zeta^{s\top}_{\theta,\ell}\right] = \bar{C}^{k,s}_{\delta y_\theta,\ell}, \tag{186}$$

where

$$\bar{C}^{k,s}_{z_\theta,\ell} := \mathbb{E}\left[\bar{z}^k_{\theta,\ell} \bar{z}^{s\top}_{\theta,\ell}\right], \qquad \bar{C}^{k,s}_{\delta y_\theta,\ell} := \mathbb{E}\left[\delta \bar{y}^k_{\theta,\ell} \delta \bar{y}^{s\top}_{\theta,\ell}\right]. \tag{187}$$

The corresponding response matrices are

$$R^{k,s}_{z_\theta,\zeta_\theta,\ell} := \mathbb{E}\left[D_{\zeta^s_{\theta,\ell}} \bar{z}^k_{\theta,\ell}\right], \quad s<k; \qquad R^{k,s}_{\delta y_\theta,\xi_\theta,\ell} := \mathbb{E}\left[D_{\xi^s_{\theta,\ell}} \delta \bar{y}^k_{\theta,\ell}\right], \quad s \le k, \tag{188}$$

and are zero outside the indicated ranges.

Consequently, the learned fields satisfy

$$\bar{y}^k_{\theta,\ell} = \xi^k_{\theta,\ell} + \sum_{s<k} \left[R^{k,s}_{z_\theta,\zeta_\theta,\ell} - \eta_\theta \rho_{\theta,L} \bar{C}^{k,s}_{z_\theta,\ell}\right] \delta \bar{y}^s_{\theta,\ell}, \tag{189}$$

$$\delta \bar{z}^k_{\theta,\ell} = \zeta^k_{\theta,\ell} + \sum_{s\le k} R^{k,s}_{\delta y_\theta,\xi_\theta,\ell} \bar{z}^s_{\theta,\ell} - \eta_\theta \rho_{\theta,L} \sum_{s<k} \bar{C}^{k,s}_{\delta y_\theta,\ell} \bar{z}^s_{\theta,\ell}. \tag{190}$$

The Gaussian families associated with distinct initialized matrices and opposite multiplication orientations are independent.

Proof. For each $\theta \in \{Q, K, V, U, M\}$, the initialized-weight products

$$Z^k_{\theta,\ell} W^{0\top}_{\theta,\ell}, \qquad \delta Y^k_{\theta,\ell} W^0_{\theta,\ell}$$

have exactly the same Tensor-Program structure as the query products in Lemma 5.

For the forward product $Z^k_{\theta,\ell} W^{0\top}_{\theta,\ell}$, the previous opposite-orientation uses are $\delta Y^s_{\theta,\ell} W^0_{\theta,\ell}$, $s < k$. Thus the ZHat/ZDot rules give Eq. (184), with response $R^{k,s}_{z_\theta,\zeta_\theta,\ell}$.

For the backward product $\delta Y^k_{\theta,\ell} W^0_{\theta,\ell}$, the previous opposite-orientation uses are $Z^s_{\theta,\ell} W^{0\top}_{\theta,\ell}$, $s \le k$. The same rules give Eq. (185), with response $R^{k,s}_{\delta y_\theta,\xi_\theta,\ell}$, together with the Gaussian covariances in Eq. (186).

Finally, Proposition 6 applied to the coordinate products in Eq. (183) gives

$$C^{k,s}_{Z_\theta,\ell} \xrightarrow{\text{a.s.}} \bar{C}^{k,s}_{z_\theta,\ell}, \qquad C^{k,s}_{\delta Y_\theta,\ell} \xrightarrow{\text{a.s.}} \bar{C}^{k,s}_{\delta y_\theta,\ell}. \tag{191}$$

Substituting these limits and Eq. (184)– Eq. (185) into Eq. (181)– Eq. (182) yields Eq. (189)– Eq. (190). □

Thus, the query, key, value, and two MLP linear maps share the same mean-field structure: opposite-orientation reuse of the initialized matrix generates the response term, while parameter learning generates the explicit kernel drift. The remaining Transformer variables follow by composing these learned linear fields through the attention, activation, residual, embedding, and readout relations.

B.3.4 Closure of the full Transformer dynamics

The learned linear fields in Lemma 6 are coupled through the embedding, attention, activation, residual, and readout operations. We now collect these remaining relations in computational order.

Embedding and input. Let $\bar{e}^0 \in \mathbb{R}^V$ and $\xi_{\text{out}} \in \mathbb{R}^C$ denote the representatives of a column of E^0 and nW^0_{out}, respectively, with

$$\bar{e}^0 \sim \mathcal{N}(0, \sigma_E^2 I_V), \qquad \xi_{\text{out}} \sim \mathcal{N}(0, I_C), \tag{192}$$

independently. The embedding dynamics are

$$\bar{e}^k = \bar{e}^0 - \eta_E \sum_{s<k} S^\top \delta \bar{x}_0^s, \tag{193}$$

$$\bar{x}_0^k = S\bar{e}^k = S\bar{e}^0 - \eta_E \sum_{s<k} SS^\top \delta \bar{x}_0^s. \tag{194}$$

Forward recursion. Starting from $\bar{x}_0^k$, specialize Eq. (189) to $\theta \in \{Q, K, V\}$ and set

$$\bar{S}_\ell^k := \mathbb{E}[\bar{q}_\ell^k \bar{k}_\ell^{k\top}], \qquad \bar{P}_\ell^k := \text{softmax}(\bar{S}_\ell^k), \tag{195}$$

$$\bar{a}_\ell^k := \bar{P}_\ell^k \bar{v}_\ell^k, \qquad \bar{h}_\ell^k := \bar{x}_\ell^k + c_L \bar{a}_\ell^k. \tag{196}$$

Next obtain $\bar{u}_\ell^k$ from Eq. (189) with $\theta = U$, set

$$\bar{\phi}_\ell^k := \phi(\bar{u}_\ell^k), \tag{197}$$

and obtain $\bar{m}_\ell^k$ from the same equation with $\theta = M$. The residual stream evolves as

$$\bar{x}_{\ell+1}^k = \bar{x}_\ell^k + a_L \bar{a}_\ell^k + b_L \bar{m}_\ell^k. \tag{198}$$

Readout. The cumulative readout update gives

$$\bar{f}^k = \mathbb{E}[\bar{x}_L^k \xi_{\text{out}}^\top] - \eta_{\text{out}} \sum_{s<k} \bar{C}_{x,L}^{k,s} \delta \bar{f}^s, \tag{199}$$

$$\delta \bar{f}^k := \nabla_1 \ell_{\text{loss}}(\bar{f}^k, y), \tag{200}$$

$$\delta \bar{x}_L^k = \delta \bar{f}^k \xi_{\text{out}} - \eta_{\text{out}} \sum_{s<k} \delta \bar{f}^k (\delta \bar{f}^s)^\top \bar{x}_L^s. \tag{201}$$

At initialization, $\bar{x}_L^0$ is independent of ξ_{out}, so

$$\bar{f}^0 = 0, \qquad \delta \bar{x}_L^0 = \nabla_1 \ell_{\text{loss}}(0, y) \xi_{\text{out}}. \tag{202}$$

Backward recursion. Starting from $\delta \bar{x}_L^k$, process the layers in reverse order. For the MLP,

$$\delta \bar{m}_\ell^k := \delta \bar{x}_{\ell+1}^k, \qquad \delta \bar{u}_\ell^k := \delta \bar{\phi}_\ell^k \odot \phi'(\bar{u}_\ell^k), \tag{203}$$

where $\delta \bar{\phi}_\ell^k$ and $\delta \bar{h}_\ell^k$ are obtained from Eq. (190) with $\theta = M$ and $\theta = U$, respectively.

For attention,

$$\delta \bar{a}_\ell^k := \delta \bar{x}_{\ell+1}^k + \frac{b_L c_L}{a_L} \delta \bar{h}_\ell^k, \qquad \delta \bar{v}_\ell^k := \bar{P}_\ell^{k\top} \delta \bar{a}_\ell^k, \tag{204}$$

$$\delta \bar{P}_\ell^k := \mathbb{E}[\delta \bar{a}_\ell^k \bar{v}_\ell^{k\top}], \qquad \delta \bar{S}_\ell^k := \bar{P}_\ell^k \odot \left(\delta \bar{P}_\ell^k - [(\delta \bar{P}_\ell^k \odot \bar{P}_\ell^k) \mathbf{1}_T] \mathbf{1}_T^\top \right), \tag{205}$$

$$\delta \bar{q}_\ell^k := \delta \bar{S}_\ell^k \bar{k}_\ell^k, \qquad \delta \bar{k}_\ell^k := \delta \bar{S}_\ell^{k\top} \bar{q}_\ell^k. \tag{206}$$

Specializing Eq. (190) to $\theta \in \{Q, K, V\}$ gives $\delta \bar{x}_{Q,\ell}^k$, $\delta \bar{x}_{K,\ell}^k$, and $\delta \bar{x}_{V,\ell}^k$. Hence

$$\delta \bar{x}_\ell^k = \delta \bar{x}_{\ell+1}^k + b_L \delta \bar{h}_\ell^k + a_L \left(\delta \bar{x}_{Q,\ell}^k + \delta \bar{x}_{K,\ell}^k + \delta \bar{x}_{V,\ell}^k \right). \tag{207}$$

The resulting $\delta \bar{x}_0^k$ completes training step k and enters the embedding and parameter updates at the next step.

Proposition 7 (Joint Transformer mean-field dynamics)**.** *Under Proposition 6, the joint mean-field training dynamics through step K are recursively determined by Lemma 6 together with Eq. (193)–Eq. (207), evaluated in computational order.*

Proof. Lemma 6 gives all learned linear-field representatives. The remaining relations follow directly from the Tensor-Program empirical-average and coordinatewise rules: empirical averages give $\bar{S}_\ell^k$, $\delta\bar{P}_\ell^k$, and the readout kernels, while softmax, activation, attention mixing, residual addition, and their adjoints are fixed-dimensional or coordinatewise operations. The embedding and readout equations follow from their cumulative updates. Hence these relations identify every representative in the Tensor Program of Proposition 6 and close the joint dynamics. □

The following subsections use this fixed-depth mean-field system to study its behavior as L varies, retaining the dependence on a_L, b_L, c_L and the learning rates.

B.4 Initialized forward and backward fields

In this subsection, we specialize the mean-field dynamics of Proposition 7 to initialization, $k = 0$, and derive a sufficient condition for stability as $L \to \infty$. The resulting depth scaling agrees with the finite-width forward-stability condition in Section A. The initialized dynamics and uniform moment bounds provide baseline estimates for the subsequent one-step training analysis.

We first take $n \to \infty$ at each fixed L and then study the resulting mean-field process as $L \to \infty$. Throughout, attention is unmasked, and the sample (S, y), dimensions T, V, C, embedding variance $\sigma_E^2 > 0$, activation, and loss are fixed. We use the assumptions of Proposition 6 and set

$$\gamma_L := a_L^2 + b_L^2, \qquad \sup_{L\geq 1} |c_L| < \infty, \qquad C_\phi := \|\phi'\|_\infty < \infty. \tag{208}$$

For nonnegative quantities, $A \lesssim B$ means $A \leq CB$ with C independent of L and ℓ; C may depend on the fixed problem data and the uniform bound on $|c_L|$, and may change from line to line. The notation $\lesssim_p$ additionally allows dependence on the fixed moment order p.

B.4.1 Initialized forward pass

At $k = 0$, the forward response and learning sums in Eq. (189) are empty. Thus

$$\bar{y}_{\theta,\ell}^0 = \xi_{\theta,\ell}^0, \qquad \theta \in \{Q, K, V, U, M\}. \tag{209}$$

In particular, $\bar{q}_\ell^0, \bar{k}_\ell^0, \bar{v}_\ell^0$ are independent centered Gaussian vectors with covariance $\bar{C}_{x,\ell}^{0,0}$. Consequently,

$$\bar{S}_\ell^0 = \mathbb{E}[\bar{q}_\ell^0 \bar{k}_\ell^{0\top}] = 0, \qquad \bar{P}_\ell^0 = \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^\top, \qquad \|\bar{P}_\ell^0\|_{\mathrm{op}} = 1. \tag{210}$$

The remaining forward relations are

$$\bar{a}_\ell^0 = \bar{P}_\ell^0 \xi_{V,\ell}^0, \qquad \bar{h}_\ell^0 = \bar{x}_\ell^0 + c_L \bar{a}_\ell^0, \tag{211}$$

$$\bar{u}_\ell^0 = \xi_{U,\ell}^0, \qquad \bar{\phi}_\ell^0 = \phi(\bar{u}_\ell^0), \qquad \bar{m}_\ell^0 = \xi_{M,\ell}^0, \tag{212}$$

$$\bar{x}_{\ell+1}^0 = \bar{x}_\ell^0 + a_L \bar{a}_\ell^0 + b_L \bar{m}_\ell^0. \tag{213}$$

The Gaussian components associated with distinct initialized matrices are independent by Lemma 6. In particular, $\bar{a}_\ell^0$ and $\bar{m}_\ell^0$ are independent of $\bar{x}_\ell^0$ and of one another. Their covariances depend on previous fields through deterministic moment kernels:

$$\bar{C}_{a,\ell}^{0,0} = \bar{P}_\ell^0 \bar{C}_{x,\ell}^{0,0} \bar{P}_\ell^{0\top}, \tag{214}$$

$$\bar{C}_{h,\ell}^{0,0} = \bar{C}_{x,\ell}^{0,0} + c_L^2 \bar{C}_{a,\ell}^{0,0}, \tag{215}$$

$$\bar{C}_{u,\ell}^{0,0} = \bar{C}_{h,\ell}^{0,0}, \qquad \bar{C}_{m,\ell}^{0,0} = \bar{C}_{\phi,\ell}^{0,0}. \tag{216}$$

The embedding initialization gives

$$\bar{x}_0^0 = S\bar{e}^0, \qquad \bar{C}_{x,0}^{0,0} = \mathbb{E}[\bar{x}_0^0 \bar{x}_0^{0\top}] = \sigma_E^2 SS^\top, \qquad \mathbb{E}\|\bar{x}_0^0\|_2^2 = T\sigma_E^2. \tag{217}$$

Centering and independence in Eq. (213) give the exact covariance recursion

$$\bar{C}_{x,\ell+1}^{0,0} = \bar{C}_{x,\ell}^{0,0} + a_L^2 \bar{C}_{a,\ell}^{0,0} + b_L^2 \bar{C}_{\phi,\ell}^{0,0}. \tag{218}$$

Proposition 8 (Initialized forward stability)**.** *Under the assumptions of Proposition 6 and the unmasked, bounded-c_L setting above, there is a constant $C > 0$, independent of L and ℓ, such that*

$$T + \mathbb{E}\|\bar{x}_\ell^0\|_2^2 \le e^{C\ell\gamma_L}\left(T + \mathbb{E}\|\bar{x}_0^0\|_2^2\right), \qquad 0 \le \ell \le L. \tag{219}$$

In particular, if

$$\sup_{L\ge 1} L\gamma_L = \sup_{L\ge 1} L(a_L^2 + b_L^2) < \infty, \tag{220}$$

then

$$\sup_{L\ge 1} \max_{0\le \ell\le L} \mathbb{E}\|\bar{x}_\ell^0\|_2^2 < \infty. \tag{221}$$

Proof. Taking traces in Eq. (218) gives

$$\mathbb{E}\|\bar{x}_{\ell+1}^0\|_2^2 = \mathbb{E}\|\bar{x}_\ell^0\|_2^2 + a_L^2 \operatorname{tr} \bar{C}_{a,\ell}^{0,0} + b_L^2 \operatorname{tr} \bar{C}_{\phi,\ell}^{0,0}. \tag{222}$$

By Eq. (214) and $\|\bar{P}_\ell^0\|_{\mathrm{op}} = 1$,

$$\operatorname{tr} \bar{C}_{a,\ell}^{0,0} \le \operatorname{tr} \bar{C}_{x,\ell}^{0,0}. \tag{223}$$

Using Eq. (215)– Eq. (216), $|c_L| \lesssim 1$, and $|\phi(u)| \le |\phi(0)| + C_\phi|u|$, we obtain

$$\begin{aligned} \operatorname{tr} \bar{C}_{\phi,\ell}^{0,0} &\le 2C_\phi^2 \operatorname{tr} \bar{C}_{h,\ell}^{0,0} + 2T|\phi(0)|^2 \\ &\lesssim \operatorname{tr} \bar{C}_{x,\ell}^{0,0} + T. \end{aligned} \tag{224}$$

Substitution into Eq. (222) yields

$$\mathbb{E}\|\bar{x}_{\ell+1}^0\|_2^2 \le (1 + C\gamma_L)\mathbb{E}\|\bar{x}_\ell^0\|_2^2 + CTb_L^2. \tag{225}$$

Since $b_L^2 \le \gamma_L$, the quantity $V_\ell := T + \mathbb{E}\|\bar{x}_\ell^0\|_2^2$ satisfies $V_{\ell+1} \le (1 + C\gamma_L)V_\ell$. Thus

$$V_\ell \le (1 + C\gamma_L)^\ell V_0 \le e^{C\ell\gamma_L} V_0. \tag{226}$$

This proves Eq. (219). The uniform bound follows from Eq. (220) and Eq. (217). □

Corollary 4 (Uniform initialized forward moments)**.** *Under Proposition 8 and Eq. (220), for every $1 \le p < \infty$, the residual fields $\bar{x}_\ell^0$ and the sublayer fields $\bar{q}_\ell^0, \bar{k}_\ell^0, \bar{v}_\ell^0, \bar{a}_\ell^0, \bar{h}_\ell^0, \bar{u}_\ell^0, \bar{\phi}_\ell^0, \bar{m}_\ell^0$ have uniformly bounded L^p norms over L and their valid layer indices.*

Proof. The covariance identities above bound the second moments of all these fields. The residual stream is a sum of the initial Gaussian embedding and independent Gaussian increments. Hence every listed field except $\bar{\phi}_\ell^0$ is centered Gaussian. For $G \sim \mathcal{N}(0, \Sigma)$ in $\mathbb{R}^T$, writing $G \overset{d}{=} \Sigma^{1/2} g$, $g \sim \mathcal{N}(0, I_T)$, gives

$$\|G\|_{L^p} \le (\operatorname{tr} \Sigma)^{1/2} \|g\|_{L^p}. \tag{227}$$

Since T is fixed, the uniform second-moment bounds give the Gaussian L^p bounds. Finally, using $|\phi(u)| \le |\phi(0)| + C_\phi|u|$ gives $\|\bar{\phi}_\ell^0\|_{L^p} \le C_\phi\|\bar{u}_\ell^0\|_{L^p} + \sqrt{T}|\phi(0)|$, which completes the proof. □

B.4.2 Initialized backward pass

Terminal adjoint. Specializing Eq. (201) to $k = 0$ gives

$$\bar{f}^0 = 0, \qquad \delta\bar{f}^0 = \nabla_1 \ell_{\mathrm{loss}}(0, y), \qquad \delta\bar{x}_L^0 = \delta\bar{f}^0 \xi_{\mathrm{out}}. \tag{228}$$

Hence $\delta\bar{x}_L^0$ is centered Gaussian, independent of the initialized hidden forward fields, with

$$\bar{C}_{\delta x,L}^{0,0} = \delta\bar{f}^0 \delta\bar{f}^{0\top}, \qquad \mathbb{E}\|\delta\bar{x}_L^0\|_2^2 = \|\delta\bar{f}^0\|_F^2. \tag{229}$$

The terminal second moment is finite and independent of L because the sample and loss are fixed.

Vanishing responses at initialization. At $k = 0$, the generic backward dynamics Eq. (190) reduce to a backward Gaussian component plus the current-step response. These response terms vanish because the terminal adjoint $\delta\bar{x}_L^0$ is centered and independent of the initialized forward Gaussian fields, and this conditional centering is preserved through the backward recursion.

Lemma 7 (Vanishing initialized backward responses)**.** *Under the assumptions of Proposition 6, for every $\theta \in \{Q, K, V, U, M\}$ and $0 \le \ell < L$,*

$$R^{0,0}_{\delta y_\theta, \xi_\theta, \ell} = 0. \tag{230}$$

The MLP and attention adjoints satisfy

$$\delta\bar{m}_\ell^0 = \delta\bar{x}_{\ell+1}^0, \qquad \delta\bar{\phi}_\ell^0 = \zeta_{M,\ell}^0, \tag{231}$$

$$\delta\bar{u}_\ell^0 = \zeta_{M,\ell}^0 \odot \phi'(\bar{u}_\ell^0), \qquad \delta\bar{h}_\ell^0 = \zeta_{U,\ell}^0, \tag{232}$$

$$\delta\bar{a}_\ell^0 = \delta\bar{x}_{\ell+1}^0 + \frac{b_L c_L}{a_L}\zeta_{U,\ell}^0, \qquad \delta\bar{v}_\ell^0 = \bar{P}_\ell^{0\top}\delta\bar{a}_\ell^0, \tag{233}$$

$$\delta\bar{P}_\ell^0 = \delta\bar{S}_\ell^0 = 0, \qquad \delta\bar{q}_\ell^0 = \delta\bar{k}_\ell^0 = 0. \tag{234}$$

Consequently,

$$\delta\bar{x}_{Q,\ell}^0 = \delta\bar{x}_{K,\ell}^0 = 0, \qquad \delta\bar{x}_{V,\ell}^0 = \zeta_{V,\ell}^0, \tag{235}$$

and the residual adjoint obeys

$$\delta\bar{x}_\ell^0 = \delta\bar{x}_{\ell+1}^0 + b_L\zeta_{U,\ell}^0 + a_L\zeta_{V,\ell}^0. \tag{236}$$

The residual adjoints are jointly centered Gaussian and independent of all initialized forward fields.

Proof. Let $\mathcal{F}_{\mathrm{fwd}}^0$ be generated by the initial embedding and all forward Gaussian components. The ZHat rule (Yang & Hu, 2021, Definition G.3) and the independent readout initialization make the backward Gaussian components and ξ_{out} independent of $\mathcal{F}_{\mathrm{fwd}}^0$. We proceed in reverse layer order, assuming that $\delta\bar{x}_{\ell+1}^0$ is a deterministic linear combination of ξ_{out} and backward Gaussian components from layers $r > \ell$. This holds at $\ell = L - 1$ by Eq. (228).

For the MLP output map, $\delta\bar{m}_\ell^0 = \delta\bar{x}_{\ell+1}^0$ has no $\xi_{M,\ell}^0$ argument in its TP expression. Thus $R^{0,0}_{\delta m, \xi_M, \ell} = 0$, and Eq. (190) gives $\delta\bar{\phi}_\ell^0 = \zeta_{M,\ell}^0$. Hence $\delta\bar{u}_\ell^0 = \zeta_{M,\ell}^0 \odot \phi'(\xi_{U,\ell}^0)$, and its response is

$$R^{0,0}_{\delta u, \xi_U, \ell} = \mathbb{E}\big[\mathrm{diag}\big(\zeta_{M,\ell}^0 \odot \phi''(\xi_{U,\ell}^0)\big)\big] = 0. \tag{237}$$

The equality follows from centering and independence of $\zeta_{M,\ell}^0$. The derivative holds the other Gaussian arguments and deterministic kernels fixed, as in (Yang & Hu, 2021, Remark L.1). Here ϕ'' is the bounded almost-everywhere derivative of the Lipschitz function ϕ'. Equations Eq. (217), Eq. (215)–Eq. (216), and Eq. (218) imply $(\bar{C}_{u,\ell}^{0,0})_{tt} \ge \sigma_E^2 > 0$ for each token t. Thus each component of $\xi_{U,\ell}^0$ has a density, and the displayed derivative is defined almost surely. The generic backward equation therefore gives $\delta\bar{h}_\ell^0 = \zeta_{U,\ell}^0$.

It follows from the attention-output adjoint equation that

$$\delta\bar{a}_\ell^0 = \delta\bar{x}_{\ell+1}^0 + \frac{b_L c_L}{a_L}\zeta_{U,\ell}^0. \tag{238}$$

This is a centered Gaussian vector independent of $\mathcal{F}_{\mathrm{fwd}}^0$. Consequently,

$$\delta\bar{P}_\ell^0 = \mathbb{E}[\delta\bar{a}_\ell^0\bar{v}_\ell^{0\top}] = 0. \tag{239}$$

The softmax adjoint then gives $\delta\bar{S}_\ell^0 = 0$, followed by $\delta\bar{q}_\ell^0 = \delta\bar{k}_\ell^0 = 0$. Their responses and backward Gaussian covariances vanish. Also, $\delta\bar{v}_\ell^0 = \bar{P}_\ell^{0\top}\delta\bar{a}_\ell^0$ has no forward Gaussian argument because $\bar{P}_\ell^0$ is deterministic. Hence $R^{0,0}_{\delta v, \xi_V, \ell} = 0$ and $\delta\bar{x}_{V,\ell}^0 = \zeta_{V,\ell}^0$.

Substituting into Eq. (207) yields Eq. (236). This extends the induction hypothesis to $\delta\bar{x}_\ell^0$ and proves its joint Gaussianity and independence from the initialized forward fields. □

Although the averaged responses vanish, internal adjoints need not be independent of the forward fields; for example, $\delta\bar{u}_\ell^0 = \zeta_{M,\ell}^0 \odot \phi'(\bar{u}_\ell^0)$ retains the forward activation factor.

Backward moment recursion. Define

$$\bar{C}^{0,0}_{\phi',\ell} := \mathbb{E}\left[\phi'(\bar{u}^0_\ell)\phi'(\bar{u}^0_\ell)^\top\right]. \tag{240}$$

By the Gaussian covariance rule Eq. (186) and the independence of the corresponding forward and backward Gaussian components,

$$\bar{C}^{0,0}_{\delta m,\ell} = \bar{C}^{0,0}_{\delta x,\ell+1}, \qquad \bar{C}^{0,0}_{\delta u,\ell} = \bar{C}^{0,0}_{\delta x,\ell+1} \odot \bar{C}^{0,0}_{\phi',\ell}, \tag{241}$$

$$\bar{C}^{0,0}_{\delta a,\ell} = \bar{C}^{0,0}_{\delta x,\ell+1} + \left(\frac{b_L c_L}{a_L}\right)^2 \bar{C}^{0,0}_{\delta u,\ell}, \qquad \bar{C}^{0,0}_{\delta v,\ell} = \bar{P}^{0\top}_\ell \bar{C}^{0,0}_{\delta a,\ell} \bar{P}^0_\ell. \tag{242}$$

By Lemma 7, $\delta\bar{x}^0_{\ell+1}$ uses only the initial readout and deeper-layer backward Gaussian components. It is therefore independent of $\zeta^0_{U,\ell}$ and $\zeta^0_{V,\ell}$, which are also independent of one another. Taking the covariance of Eq. (236) gives

$$\begin{aligned}\bar{C}^{0,0}_{\delta x,\ell} &= \bar{C}^{0,0}_{\delta x,\ell+1} + b_L^2 \bar{C}^{0,0}_{\delta u,\ell} + a_L^2 \bar{C}^{0,0}_{\delta v,\ell} \\ &= \bar{C}^{0,0}_{\delta x,\ell+1} + b_L^2 \bar{C}^{0,0}_{\delta u,\ell} + \bar{P}^{0\top}_\ell \left[a_L^2 \bar{C}^{0,0}_{\delta x,\ell+1} + b_L^2 c_L^2 \bar{C}^{0,0}_{\delta u,\ell}\right] \bar{P}^0_\ell.\end{aligned} \tag{243}$$

Proposition 9 (Initialized backward stability)**.** *Under the assumptions of Proposition 6 and the unmasked, bounded-c_L setting above, there exists a constant $C > 0$, independent of L and ℓ, such that for $0 \le \ell \le L$,*

$$\mathbb{E}\|\delta\bar{x}^0_\ell\|_2^2 \le e^{C(L-\ell)\gamma_L} \|\delta\bar{f}^0\|_F^2. \tag{244}$$

Hence, under the baseline depth condition Eq. (220),

$$\sup_{L\ge 1} \max_{0\le\ell\le L} \mathbb{E}\|\delta\bar{x}^0_\ell\|_2^2 < \infty. \tag{245}$$

Proof. Since $\|\phi'\|_\infty = C_\phi$, Eq. (241) gives

$$\operatorname{tr} \bar{C}^{0,0}_{\delta u,\ell} \le C_\phi^2 \operatorname{tr} \bar{C}^{0,0}_{\delta x,\ell+1}. \tag{246}$$

Taking traces in Eq. (243) and using $\|\bar{P}^0_\ell\|_{\mathrm{op}} = 1$ therefore yields

$$\begin{aligned}\mathbb{E}\|\delta\bar{x}^0_\ell\|_2^2 &\le \left[1 + a_L^2 + C_\phi^2 b_L^2 (1 + c_L^2)\right] \mathbb{E}\|\delta\bar{x}^0_{\ell+1}\|_2^2 \\ &\le (1 + C\gamma_L)\mathbb{E}\|\delta\bar{x}^0_{\ell+1}\|_2^2,\end{aligned} \tag{247}$$

where we used $|c_L| \lesssim 1$ and $\gamma_L = a_L^2 + b_L^2$.

Iterating from layer L and using Eq. (229),

$$\mathbb{E}\|\delta\bar{x}^0_\ell\|_2^2 \le (1 + C\gamma_L)^{L-\ell} \|\delta\bar{f}^0\|_F^2 \le e^{C(L-\ell)\gamma_L} \|\delta\bar{f}^0\|_F^2. \tag{248}$$

Thus Eq. (244) holds, and the uniform bound follows from $L\gamma_L = O(1)$. □

Corollary 5 (Uniform initialized backward moments)**.** *Under Proposition 9 and Eq. (220), for every $1 \le p < \infty$, the residual and MLP adjoints, and weighted attention adjoints*

$$a_L \delta\bar{a}^0_\ell, \qquad a_L \delta\bar{v}^0_\ell, \qquad a_L \delta\bar{x}^0_{V,\ell} \tag{249}$$

have uniformly bounded L^p norms over L and their valid layer indices, while $\delta\bar{P}^0_\ell, \delta\bar{S}^0_\ell, \delta\bar{q}^0_\ell, \delta\bar{k}^0_\ell, \delta\bar{x}^0_{Q,\ell}, \delta\bar{x}^0_{K,\ell}$ vanish identically. If additionally

$$\sup_{L\ge 1} \left|\frac{b_L c_L}{a_L}\right| < \infty, \tag{250}$$

then $\delta\bar{a}^0_\ell$, $\delta\bar{v}^0_\ell$, and $\delta\bar{x}^0_{V,\ell}$ are uniformly bounded in L^p as well.

Proof. The residual adjoints are Gaussian by Lemma 7. Thus Proposition 9 and Eq. (227) give their uniform L^p bounds. For the internal MLP fields, the covariance rule, Eq. (246), and the same Gaussian moment bound yield

$$\|\delta\bar{\phi}^0_\ell\|_{L^p} + \|\delta\bar{h}^0_\ell\|_{L^p} \lesssim_p \|\delta\bar{x}^0_{\ell+1}\|_{L^2}. \tag{251}$$

Together with $\delta\bar{m}_\ell^0 = \delta\bar{x}_{\ell+1}^0$ and

$$\|\delta\bar{u}_\ell^0\|_{L^p} \le C_\phi \|\zeta_{M,\ell}^0\|_{L^p}, \tag{252}$$

this bounds all MLP adjoints.

For attention, Eq. (233) gives

$$\|a_L \delta\bar{a}_\ell^0\|_{L^p} \le a_L \|\delta\bar{x}_{\ell+1}^0\|_{L^p} + |b_L c_L| \, \|\zeta_{U,\ell}^0\|_{L^p}. \tag{253}$$

The right-hand side is uniformly bounded because Eq. (220) bounds a_L and b_L, and $\sup_L |c_L| < \infty$. Since $\|\bar{P}_\ell^0\|_{\mathrm{op}} = 1$, the same bound holds for $a_L \delta\bar{v}_\ell^0$. Moreover, $\delta\bar{v}_\ell^0$ and $\zeta_{V,\ell}^0$ are centered Gaussian with the same covariance, so

$$\|a_L \delta\bar{x}_{V,\ell}^0\|_{L^p} = \|a_L \zeta_{V,\ell}^0\|_{L^p} = \|a_L \delta\bar{v}_\ell^0\|_{L^p}. \tag{254}$$

The vanishing fields are identified in Eq. (234)– Eq. (235). Under Eq. (250), Eq. (233) also bounds $\delta\bar{a}_\ell^0$ without the factor a_L, and the same reasoning gives the remaining unweighted bounds. □

Thus the same condition $L\gamma_L = L(a_L^2 + b_L^2) = O(1)$ suffices for both initialized passes when $c_L = O(1)$. The subsequent one-step analysis therefore treats this architectural condition as given and studies the additional response and learning terms introduced after the first parameter update.

B.5 One-update forward dynamics and joint stability

We study the forward pass after the first simultaneous parameter update, using the joint mean-field dynamics of Proposition 7. We retain the unmasked setting and fixed data of Section B.4, and assume

$$\sup_{L\ge 1} |c_L| < \infty, \qquad \sup_{L\ge 1} L\gamma_L < \infty, \qquad \gamma_L = a_L^2 + b_L^2. \tag{255}$$

The initialized moment bounds are therefore available. Learning-rate coefficients may depend on L but remain independent of width. All depth estimates concern the mean-field process obtained by first taking $n \to \infty$ at fixed L.

Updated input and forward fields. The embedding update gives

$$\bar{x}_0^1 = \bar{x}_0^0 - \eta_E S S^\top \delta\bar{x}_0^0. \tag{256}$$

The remaining fields follow by specializing Lemma 6 to $k = 1$. The query and key equations simplify because their initialized adjoints vanish.

Lemma 8 (One-update forward representations)**.** *Under Proposition 6 and the initialization setting above, for $0 \le \ell < L$,*

$$\bar{q}_\ell^1 = \xi_{Q,\ell}^1, \qquad \bar{k}_\ell^1 = \xi_{K,\ell}^1, \qquad \bar{S}_\ell^1 = 0, \qquad \bar{P}_\ell^1 = \bar{P}_\ell^0 = \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^\top. \tag{257}$$

The value and MLP fields satisfy

$$\bar{v}_\ell^1 = \xi_{V,\ell}^1 + \left[R_{x,\zeta_V,\ell}^{1,0} - \eta_V a_L \bar{C}_{x,\ell}^{1,0}\right] \delta\bar{v}_\ell^0, \tag{258}$$

$$\bar{a}_\ell^1 = \bar{P}_\ell^0 \bar{v}_\ell^1, \qquad \bar{h}_\ell^1 = \bar{x}_\ell^1 + c_L \bar{a}_\ell^1, \tag{259}$$

$$\bar{u}_\ell^1 = \xi_{U,\ell}^1 + \left[R_{h,\zeta_U,\ell}^{1,0} - \eta_U b_L \bar{C}_{h,\ell}^{1,0}\right] \delta\bar{u}_\ell^0, \tag{260}$$

$$\bar{\phi}_\ell^1 = \phi(\bar{u}_\ell^1), \tag{261}$$

$$\bar{m}_\ell^1 = \xi_{M,\ell}^1 + \left[R_{\phi,\zeta_M,\ell}^{1,0} - \eta_M b_L \bar{C}_{\phi,\ell}^{1,0}\right] \delta\bar{m}_\ell^0. \tag{262}$$

The kernels are the joint moments $\bar{C}_{z,\ell}^{1,0} = \mathbb{E}[\bar{z}_\ell^1 \bar{z}_\ell^{0\top}]$, $z \in \{x, h, \phi\}$. The Gaussian components retain the joint law of Lemma 6; in particular,

$$\mathbb{E}[\xi_{\theta,\ell}^1 \xi_{\theta,\ell}^{s\top}] = \bar{C}_{z_\theta,\ell}^{1,s}, \qquad s \in \{0, 1\}, \tag{263}$$

where $z_Q = z_K = z_V = x$, $z_U = h$, and $z_M = \phi$.

Proof. Apply Eq. (189) at $k = 1$. Equation Eq. (234) gives $\delta\bar{q}^0_\ell = \delta\bar{k}^0_\ell = 0$, eliminating the query and key corrections. Their remaining Gaussian components are independent and centered, so $\bar{S}^1_\ell = \mathbb{E}[\xi^1_{Q,\ell}\xi^{1\top}_{K,\ell}] = 0$. Unmasked softmax gives Eq. (257). The other identities are the specializations $\theta = V, U, M$ and the deterministic attention and activation relations. The Gaussian covariance identity is Eq. (186) at these training steps. □

Thus the first updated attention weights remain uniform, although the query and key representations can change through their updated inputs. Their learning-rate coefficients do not enter this forward pass. Substitution into the residual update gives

$$\begin{aligned}\bar{x}^1_{\ell+1} = \bar{x}^1_\ell &+ a_L\bar{P}^0_\ell\xi^1_{V,\ell} + b_L\xi^1_{M,\ell} \\ &+ a_L\bar{P}^0_\ell\left[R^{1,0}_{x,\zeta_V,\ell} - \eta_V a_L\bar{C}^{1,0}_{x,\ell}\right]\delta\bar{v}^0_\ell \\ &+ b_L\left[R^{1,0}_{\phi,\zeta_M,\ell} - \eta_M b_L\bar{C}^{1,0}_{\phi,\ell}\right]\delta\bar{m}^0_\ell. \end{aligned} \tag{264}$$

We next identify the responses in this recursion from the symbolic Tensor-Program expressions.

Response identities and recursion. We close the responses in Eq. (264) through the primary residual-stream response

$$R^1_\ell := \mathbb{E}\Big[D_{\delta\bar{x}^0_\ell}\bar{x}^1_\ell\Big], \qquad 0 \le \ell \le L. \tag{265}$$

To define this derivative, express the earlier initialized residual adjoints through $\delta\bar{x}^0_\ell$ using

$$\delta\bar{x}^0_j = \delta\bar{x}^0_\ell + \sum_{r=j}^{\ell-1}\left(b_L\zeta^0_{U,r} + a_L\zeta^0_{V,r}\right), \qquad 0 \le j \le \ell, \tag{266}$$

which follows from Eq. (236). Substitute this identity into the earlier adjoints appearing in the updated residual computation, and differentiate with respect to the remaining argument $\delta\bar{x}^0_\ell$. All other Gaussian arguments, moment kernels, attention matrices, and response coefficients are held fixed. Thus R^1_ℓ measures how the updated residual responds to the initialized adjoint at the same layer. The derivatives with respect to Gaussian components retain the Tensor-Program convention (Yang & Hu, 2021, Remark L.1).

Lemma 9 (One-update response identities)**.** *Under Lemma 8 and the initialized backward identities, the primary response starts from*

$$R^1_0 = -\eta_E S S^\top. \tag{267}$$

For $0 \le \ell < L$, the active responses satisfy

$$R^{1,0}_{x,\zeta_V,\ell} = a_L R^1_\ell, \tag{268}$$

$$R^{1,0}_{h,\zeta_U,\ell} = b_L R^1_\ell + b_L c_L^2\bar{P}^0_\ell\left(R^1_\ell - \eta_V\bar{C}^{1,0}_{x,\ell}\right)\bar{P}^{0\top}_\ell, \tag{269}$$

$$R^{1,0}_{\phi,\zeta_M,\ell} = \mathbb{E}\Big[\mathrm{diag}\big(\phi'(\bar{u}^1_\ell)\big)\left(R^{1,0}_{h,\zeta_U,\ell} - \eta_U b_L\bar{C}^{1,0}_{h,\ell}\right)\mathrm{diag}\big(\phi'(\bar{u}^0_\ell)\big)\Big]. \tag{270}$$

The primary response obeys

$$R^1_{\ell+1} = R^1_\ell + a_L^2\bar{P}^0_\ell\left(R^1_\ell - \eta_V\bar{C}^{1,0}_{x,\ell}\right)\bar{P}^{0\top}_\ell + b_L R^{1,0}_{\phi,\zeta_M,\ell} - \eta_M b_L^2\bar{C}^{1,0}_{\phi,\ell}. \tag{271}$$

Proof. All derivatives are taken in the representative expressions at the fixed joint Tensor-Program law. Moment kernels, attention matrices, and response coefficients are therefore held fixed, together with the other Gaussian arguments (Yang & Hu, 2021, Definition G.3 and Remark L.1). For differentiation with respect to $\delta\bar{x}^0_\ell$, we use the expansion Eq. (266). We apply the chain rule to these expressions and then take expectations, as in Eq. (265). The embedding update Eq. (256) gives

$$R^1_0 = \mathbb{E}\Big[D_{\delta\bar{x}^0_0}\bar{x}^1_0\Big] = -\eta_E S S^\top, \tag{272}$$

which proves Eq. (267).

Residual-input responses. In the residual computation through layer $\ell - 1$, the initialized backward fields enter through the embedding update and the adjoints $\delta\bar{v}^0_j, \delta\bar{m}^0_j$, $j < \ell$. By Eq. (231) and

Eq. (233), these adjoints are linear in the initialized residual adjoints and $\zeta^0_{U,j}$. After applying Eq. (266), the current-layer Gaussian arguments $\zeta^0_{U,\ell}$ and $\zeta^0_{V,\ell}$ enter $\bar{x}^1_\ell$ only through $\delta\bar{x}^0_\ell$. The initialized backward recursion

$$\delta\bar{x}^0_\ell = \delta\bar{x}^0_{\ell+1} + b_L\zeta^0_{U,\ell} + a_L\zeta^0_{V,\ell} \tag{273}$$

therefore gives the realization-wise chain-rule identities

$$\begin{aligned} D_{\zeta^0_{V,\ell}}\bar{x}^1_\ell &= a_L D_{\delta\bar{x}^0_\ell}\bar{x}^1_\ell, \\ D_{\zeta^0_{U,\ell}}\bar{x}^1_\ell &= b_L D_{\delta\bar{x}^0_\ell}\bar{x}^1_\ell, \\ D_{\delta\bar{x}^0_{\ell+1}}\bar{x}^1_\ell &= D_{\delta\bar{x}^0_\ell}\bar{x}^1_\ell. \end{aligned} \tag{274}$$

Taking expectations and using Eq. (265) proves Eq. (268) and yields

$$\mathbb{E}\Big[D_{\zeta^0_{U,\ell}}\bar{x}^1_\ell\Big] = b_L R^1_\ell, \qquad \mathbb{E}\Big[D_{\delta\bar{x}^0_{\ell+1}}\bar{x}^1_\ell\Big] = R^1_\ell. \tag{275}$$

MLP-input response. Substituting Eq. (268) into Eq. (258) gives

$$\bar{v}^1_\ell = \xi^1_{V,\ell} + a_L\left(R^1_\ell - \eta_V\bar{C}^{1,0}_{x,\ell}\right)\delta\bar{v}^0_\ell. \tag{276}$$

The initialized value adjoint is

$$\delta\bar{v}^0_\ell = \bar{P}^{0\top}_\ell\left(\delta\bar{x}^0_{\ell+1} + \frac{b_L c_L}{a_L}\zeta^0_{U,\ell}\right), \tag{277}$$

so $D_{\zeta^0_{U,\ell}}\delta\bar{v}^0_\ell = (b_L c_L/a_L)\bar{P}^{0\top}_\ell$. Differentiating $\bar{h}^1_\ell = \bar{x}^1_\ell + c_L\bar{P}^0_\ell\bar{v}^1_\ell$ and taking expectation gives

$$\begin{aligned} R^{1,0}_{h,\zeta_U,\ell} &= \mathbb{E}\Big[D_{\zeta^0_{U,\ell}}\bar{x}^1_\ell\Big] + c_L\bar{P}^0_\ell a_L\left(R^1_\ell - \eta_V\bar{C}^{1,0}_{x,\ell}\right)\frac{b_L c_L}{a_L}\bar{P}^{0\top}_\ell \\ &= b_L R^1_\ell + b_L c_L^2\bar{P}^0_\ell\left(R^1_\ell - \eta_V\bar{C}^{1,0}_{x,\ell}\right)\bar{P}^{0\top}_\ell, \end{aligned} \tag{278}$$

where the last equality uses Eq. (275). This proves Eq. (269).

Activation response. In Eq. (260), the Gaussian argument $\zeta^0_{M,\ell}$ enters through $\delta\bar{u}^0_\ell = \zeta^0_{M,\ell}\odot\phi'(\bar{u}^0_\ell)$. Holding the other Gaussian arguments and the law-dependent coefficients fixed, the chain rule gives

$$D_{\zeta^0_{M,\ell}}\bar{\phi}^1_\ell = \mathrm{diag}\big(\phi'(\bar{u}^1_\ell)\big)\left(R^{1,0}_{h,\zeta_U,\ell} - \eta_U b_L\bar{C}^{1,0}_{h,\ell}\right)\mathrm{diag}\big(\phi'(\bar{u}^0_\ell)\big). \tag{279}$$

Taking expectation gives Eq. (270). Boundedness of ϕ' ensures integrability; both activation-derivative factors remain inside the joint expectation.

Primary response recursion. Equations Eq. (277) and Eq. (231) give

$$D_{\delta\bar{x}^0_{\ell+1}}\delta\bar{v}^0_\ell = \bar{P}^{0\top}_\ell, \qquad D_{\delta\bar{x}^0_{\ell+1}}\delta\bar{m}^0_\ell = I_T. \tag{280}$$

Differentiate Eq. (264) with its forward Gaussian arguments and deterministic coefficients held fixed, and take expectation. By Eq. (265),

$$\begin{aligned} R^1_{\ell+1} = \mathbb{E}\Big[D_{\delta\bar{x}^0_{\ell+1}}\bar{x}^1_\ell\Big] &+ a_L\bar{P}^0_\ell\left(R^{1,0}_{x,\zeta_V,\ell} - \eta_V a_L\bar{C}^{1,0}_{x,\ell}\right)\bar{P}^{0\top}_\ell \\ &+ b_L\left(R^{1,0}_{\phi,\zeta_M,\ell} - \eta_M b_L\bar{C}^{1,0}_{\phi,\ell}\right). \end{aligned} \tag{281}$$

Substituting Eq. (275) and Eq. (268) proves Eq. (271). □

The residual recursion Eq. (264) and the response recursion Eq. (271) now evolve the updated residual and its primary response together. The remaining responses are given by Eq. (268)– Eq. (270).

Local field and response bounds. To retain the dependence on the attention-to-MLP coefficient, define

$$\nu_L := (a_L^2 + b_L^2 c_L^2)^{1/2}, \qquad \lambda_L := \eta_V \nu_L + b_L(\eta_U + \eta_M). \tag{282}$$

We use the initialized attention bounds with their residual factor a_L retained. Define the joint energy

$$\mathcal{E}_\ell := 1 + \|\bar{x}_\ell^1\|_{L^2}^2 + \|R_\ell^1\|_{\mathrm{op}}^2. \tag{283}$$

The additive constant accommodates $\phi(0) \neq 0$.

Lemma 10 (Local one-update estimates)**.** *Under Eq. (255), assume* $\sup_L \lambda_L < \infty$. *Then, uniformly for* $0 \le \ell < L$,

$$\|\bar{q}_\ell^1\|_{L^2} + \|\bar{k}_\ell^1\|_{L^2} + \|\bar{v}_\ell^1\|_{L^2} + \|\bar{a}_\ell^1\|_{L^2} + \|\bar{h}_\ell^1\|_{L^2} + \|\bar{u}_\ell^1\|_{L^2} + \|\bar{\phi}_\ell^1\|_{L^2} + \|\bar{m}_\ell^1\|_{L^2} \lesssim \mathcal{E}_\ell^{1/2}, \tag{284}$$

$$\|\bar{C}_{x,\ell}^{1,0}\|_{\mathrm{op}} + \|\bar{C}_{h,\ell}^{1,0}\|_{\mathrm{op}} + \|\bar{C}_{\phi,\ell}^{1,0}\|_{\mathrm{op}} \lesssim \mathcal{E}_\ell^{1/2}. \tag{285}$$

The active responses satisfy

$$\begin{aligned} \|R_{x,\zeta_V,\ell}^{1,0}\|_{\mathrm{op}} &\le a_L \mathcal{E}_\ell^{1/2}, \\ \|R_{h,\zeta_U,\ell}^{1,0}\|_{\mathrm{op}} &\lesssim b_L(1 + c_L^2 \eta_V)\mathcal{E}_\ell^{1/2}, \\ \|R_{\phi,\zeta_M,\ell}^{1,0}\|_{\mathrm{op}} &\lesssim b_L(1 + c_L^2 \eta_V + \eta_U)\mathcal{E}_\ell^{1/2}. \end{aligned} \tag{286}$$

The constants are independent of learning rates or depth.

Proof. By the initialized backward moment bounds and Eq. (242),

$$\|a_L \delta\bar{v}_\ell^0\|_{L^2} \lesssim \nu_L, \qquad \|\delta\bar{u}_\ell^0\|_{L^2} + \|\delta\bar{m}_\ell^0\|_{L^2} \lesssim 1. \tag{287}$$

Cauchy–Schwarz and the initialized forward moment bounds give

$$\|\bar{C}_{z,\ell}^{1,0}\|_{\mathrm{op}} \le \|\bar{z}_\ell^1\|_{L^2}\|\bar{z}_\ell^0\|_{L^2} \lesssim \|\bar{z}_\ell^1\|_{L^2}, \qquad z \in \{x, h, \phi\}. \tag{288}$$

For brevity write $F_\ell = \mathcal{E}_\ell^{1/2}$ in this proof. Equation Eq. (257) and the Gaussian covariance rule imply $\|\bar{q}_\ell^1\|_{L^2} = \|\bar{k}_\ell^1\|_{L^2} = \|\bar{x}_\ell^1\|_{L^2}$. Using Eq. (276) and Eq. (287),

$$\begin{aligned} \|\bar{v}_\ell^1\|_{L^2} &\le \|\bar{x}_\ell^1\|_{L^2} + \|R_\ell^1 - \eta_V \bar{C}_{x,\ell}^{1,0}\|_{\mathrm{op}}\|a_L \delta\bar{v}_\ell^0\|_{L^2} \\ &\lesssim (1 + \eta_V \nu_L)\|\bar{x}_\ell^1\|_{L^2} + \nu_L \|R_\ell^1\|_{\mathrm{op}} \lesssim F_\ell. \end{aligned} \tag{289}$$

Here $\nu_L^2 \lesssim \gamma_L$, so ν_L is uniformly bounded. Since $\|\bar{P}_\ell^0\|_{\mathrm{op}} = 1$ and c_L is bounded, $\|\bar{a}_\ell^1\|_{L^2} + \|\bar{h}_\ell^1\|_{L^2} \lesssim F_\ell$. Equation Eq. (269) then gives the stated bound on $R_{h,\zeta_U,\ell}^{1,0}$ and

$$\|\bar{u}_\ell^1\|_{L^2} \lesssim \left[1 + b_L(1 + c_L^2 \eta_V + \eta_U)\right] F_\ell \lesssim F_\ell. \tag{290}$$

The last inequality uses $b_L c_L^2 \le |c_L|\nu_L$ and the bound on λ_L. Lipschitzness of ϕ gives

$$\|\bar{\phi}_\ell^1\|_{L^2} \le C_\phi \|\bar{u}_\ell^1\|_{L^2} + \sqrt{T}|\phi(0)| \lesssim F_\ell. \tag{291}$$

In Eq. (270), both diagonal matrices have operator norm at most C_ϕ. Therefore

$$\begin{aligned} \|R_{\phi,\zeta_M,\ell}^{1,0}\|_{\mathrm{op}} &\le C_\phi^2 \left\| R_{h,\zeta_U,\ell}^{1,0} - \eta_U b_L \bar{C}_{h,\ell}^{1,0} \right\|_{\mathrm{op}} \\ &\lesssim b_L(1 + c_L^2 \eta_V + \eta_U) F_\ell. \end{aligned} \tag{292}$$

Finally, the Gaussian covariance rule and Eq. (262) imply

$$\|\bar{m}_\ell^1\|_{L^2} \lesssim \left[1 + b_L(1 + c_L^2 \eta_V + \eta_U + \eta_M)\right] F_\ell \lesssim F_\ell. \tag{293}$$

These estimates and Eq. (288) prove the field and kernel bounds. The value response follows directly from Eq. (268). □

Residual increments and Gaussian cross terms. Separate the residual increment as

$$\bar{x}^1_{\ell+1} = \bar{x}^1_\ell + G^1_\ell + D^1_\ell, \tag{294}$$

where

$$G^1_\ell := a_L \bar{P}^0_\ell \xi^1_{V,\ell} + b_L \xi^1_{M,\ell}, \tag{295}$$

$$D^1_\ell := a_L^2 \bar{P}^0_\ell \left(R^1_\ell - \eta_V \bar{C}^{1,0}_{x,\ell}\right) \delta \bar{v}^0_\ell + b_L \left(R^{1,0}_{\phi,\zeta_M,\ell} - \eta_M b_L \bar{C}^{1,0}_{\phi,\ell}\right) \delta \bar{x}^0_{\ell+1}. \tag{296}$$

Lemma 11 (Gaussian cross-term cancellation)**.** *Under the preceding assumptions, G^1_ℓ is a centered Gaussian and independent of the pair $(\bar{x}^1_\ell, D^1_\ell)$. Consequently,*

$$\mathbb{E}\langle \bar{x}^1_\ell, G^1_\ell \rangle = \mathbb{E}\langle D^1_\ell, G^1_\ell \rangle = 0, \tag{297}$$

and

$$\|\bar{x}^1_{\ell+1}\|^2_{L^2} = \|\bar{x}^1_\ell\|^2_{L^2} + \|G^1_\ell\|^2_{L^2} + 2\mathbb{E}\langle \bar{x}^1_\ell, D^1_\ell \rangle + \|D^1_\ell\|^2_{L^2}. \tag{298}$$

Proof. We first make the random factors in D^1_ℓ explicit. Substituting Eq. (277) and Eq. (231) into Eq. (296) and using Eq. (271) gives

$$D^1_\ell = (R^1_{\ell+1} - R^1_\ell)\delta \bar{x}^0_{\ell+1} + a_L b_L c_L \bar{P}^0_\ell \left(R^1_\ell - \eta_V \bar{C}^{1,0}_{x,\ell}\right) \bar{P}^{0\top}_\ell \zeta^0_{U,\ell}. \tag{299}$$

All matrix coefficients in this expression are deterministic: the moment kernels and response matrices are expectations under the fixed joint mean-field law.

Moreover, by Eq. (228) and Eq. (236), the initialized terminal adjoint and backward recursion give

$$\delta \bar{x}^0_{\ell+1} = \delta \bar{f}^0 \xi_{\text{out}} + \sum_{r=\ell+1}^{L-1} \left(b_L \zeta^0_{U,r} + a_L \zeta^0_{V,r}\right). \tag{300}$$

Thus D^1_ℓ depends only on the readout and initialized backward Gaussian components. In contrast,

$$G^1_\ell = a_L \bar{P}^0_\ell \xi^1_{V,\ell} + b_L \xi^1_{M,\ell} \tag{301}$$

uses only the current-layer forward Gaussian components. The ZHat independence rule (Yang & Hu, 2021, Definition G.3) therefore gives independence of G^1_ℓ and D^1_ℓ.

To obtain independence from the pair $(\bar{x}^1_\ell, D^1_\ell)$, unroll the residual recursion:

$$\bar{x}^1_\ell = \bar{x}^0_0 - \eta_E S S^\top \delta \bar{x}^0_0 + \sum_{j<\ell} (G^1_j + D^1_j). \tag{302}$$

The initial input uses the embedding Gaussian, each G^1_j uses forward Gaussian components from layer $j < \ell$, and each D^1_j uses only the readout and initialized backward Gaussian components, as shown above. The current-layer forward pair $(\xi^1_{V,\ell}, \xi^1_{M,\ell})$ is jointly independent of this entire collection. Hence G^1_ℓ is independent of $(\bar{x}^1_\ell, D^1_\ell)$.

Since $\mathbb{E}[G^1_\ell] = 0$, both cross moments in Eq. (297) vanish. Expanding $\|\bar{x}^1_\ell + G^1_\ell + D^1_\ell\|^2_{L^2}$ then gives Eq. (298). □

The temporal correlations in Eq. (263) are retained. The cancellation above uses the particular variables entering the updated residual, rather than independence from the entire initialized forward history.

Joint stability through depth. Define the accumulated-learning and total recursion scales

$$\rho_L := \eta_V \nu_L^2 + b_L^2(\eta_U + \eta_M) = \eta_V(a_L^2 + b_L^2 c_L^2) + b_L^2(\eta_U + \eta_M), \qquad \kappa_L := \gamma_L + \rho_L. \tag{303}$$

The term $\eta_V a_L^2$ comes from the value correction in the residual stream. The term $\eta_V b_L^2 c_L^2$ comes from the same update entering the MLP response through Eq. (269).

Proposition 10 (Joint second-forward and response stability)**.** *Under Proposition 6 and Eq. (255), assume* $\sup_L \lambda_L < \infty$*. There is a constant* $C > 0$*, independent of* L *and* ℓ*, such that*

$$\mathcal{E}_\ell \le \exp\{C\ell(\kappa_L + \kappa_L^2)\}\,\mathcal{E}_0, \qquad 0 \le \ell \le L, \tag{304}$$

with $\mathcal{E}_0 \lesssim 1 + \eta_E^2$*. In particular, if*

$$\sup_L \eta_E < \infty, \qquad \sup_L L\rho_L < \infty, \tag{305}$$

then

$$\sup_{L\ge 1} \max_{0\le \ell\le L} \left(\|\bar{x}_\ell^1\|_{L^2}^2 + \|R_\ell^1\|_{\mathrm{op}}^2\right) < \infty. \tag{306}$$

All internal second-forward fields have uniformly bounded second moments, and the active responses satisfy Eq. (286) uniformly through depth.

Proof. We bound the two components of $\mathcal{E}_\ell$ using Lemma 10. First, Eq. (271) gives

$$\begin{aligned}\|R_{\ell+1}^1 - R_\ell^1\|_{\mathrm{op}} &\lesssim a_L^2\|R_\ell^1\|_{\mathrm{op}} + \left[\eta_V a_L^2 + b_L^2(1 + c_L^2\eta_V + \eta_U + \eta_M)\right]\mathcal{E}_\ell^{1/2} \\ &\lesssim \kappa_L \mathcal{E}_\ell^{1/2}.\end{aligned} \tag{307}$$

The operator-norm triangle inequality therefore yields

$$\|R_{\ell+1}^1\|_{\mathrm{op}}^2 \le \|R_\ell^1\|_{\mathrm{op}}^2 + C(\kappa_L + \kappa_L^2)\mathcal{E}_\ell. \tag{308}$$

For the Gaussian increment, covariance and independence give

$$\begin{aligned}\|G_\ell^1\|_{L^2}^2 &= a_L^2\,\mathrm{tr}\Big(\bar{P}_\ell^0 \bar{C}_{x,\ell}^{1,1} \bar{P}_\ell^{0\top}\Big) + b_L^2\,\mathrm{tr}\,\bar{C}_{\phi,\ell}^{1,1} \\ &\lesssim \gamma_L \mathcal{E}_\ell.\end{aligned} \tag{309}$$

For the response and learning increment, Eq. (287) gives

$$\begin{aligned}\|D_\ell^1\|_{L^2} &\lesssim a_L\nu_L\left(\|R_\ell^1\|_{\mathrm{op}} + \eta_V\|\bar{C}_{x,\ell}^{1,0}\|_{\mathrm{op}}\right) \\ &\quad + b_L\|R_{\phi,\zeta_M,\ell}^{1,0}\|_{\mathrm{op}} + \eta_M b_L^2\|\bar{C}_{\phi,\ell}^{1,0}\|_{\mathrm{op}} \\ &\lesssim \kappa_L \mathcal{E}_\ell^{1/2},\end{aligned} \tag{310}$$

where $a_L\nu_L \le \nu_L^2 \lesssim \gamma_L$. By Cauchy–Schwarz in Eq. (298),

$$\|\bar{x}_{\ell+1}^1\|_{L^2}^2 \le \|\bar{x}_\ell^1\|_{L^2}^2 + C(\kappa_L + \kappa_L^2)\mathcal{E}_\ell, \tag{311}$$

since $\gamma_L \le \kappa_L$. Combining this with Eq. (308) gives

$$\mathcal{E}_{\ell+1} \le \left[1 + C(\kappa_L + \kappa_L^2)\right]\mathcal{E}_\ell. \tag{312}$$

Iteration proves Eq. (304).

For the boundary, $\bar{x}_0^0$ and $\delta\bar{x}_0^0$ are independent and centered by the initialization analysis. Hence

$$\begin{aligned}\mathcal{E}_0 &= 1 + T\sigma_E^2 + \eta_E^2\mathbb{E}\|SS^\top\delta\bar{x}_0^0\|_2^2 + \eta_E^2\|SS^\top\|_{\mathrm{op}}^2 \\ &\lesssim 1 + \eta_E^2.\end{aligned} \tag{313}$$

Under Eq. (305), $L\kappa_L = L\gamma_L + L\rho_L$ is uniformly bounded, as is $L\kappa_L^2 = (L\kappa_L)^2/L$. The depth bound and boundary estimate therefore give Eq. (306). The remaining conclusions follow from Lemma 10. □

Higher moments and the updated readout. The residual recursion also gives higher-moment bounds, while its primary response identifies the dependence on the initialized readout.

Corollary 6 (Forward moments and readout after one update)**.** *Under the uniform-stability conditions of Proposition 10, every hidden forward field and the updated embedding have uniformly bounded* L^p *norms for each fixed* $1 \le p < \infty$*. The output is*

$$\bar{f}^1 = \left(R_L^1 - \eta_{\mathrm{out}}\bar{C}_{x,L}^{1,0}\right)\delta\bar{f}^0. \tag{314}$$

If $\sup_L \eta_{\mathrm{out}} < \infty$*, then* $\bar{f}^1$ *and* $\delta\bar{f}^1 = \nabla_1\ell_{\mathrm{loss}}(\bar{f}^1, y)$ *are uniformly bounded as well, and the representatives of the columns of* nW_{out}^1 *have uniformly bounded* L^p *norms.*

Proof. The residual recursion expresses $\bar{x}_\ell^1$ as a deterministic linear combination of jointly Gaussian fields, as established in the proof of Lemma 11. The query, key, value, attention-output, MLP-input, and MLP-output fields are also deterministic linear combinations of jointly Gaussian variables, by Eq. (257)– Eq. (262) and the initialized adjoint representations. Their second-moment bounds therefore imply uniform L^p bounds by Eq. (227). For the MLP preactivation,

$$\begin{aligned}\|\bar{u}_\ell^1\|_{L^p} &\le \|\xi_{U,\ell}^1\|_{L^p} + \left\| R_{h,\zeta_U,\ell}^{1,0} - \eta_U b_L \bar{C}_{h,\ell}^{1,0} \right\|_{\mathrm{op}} \|\delta\bar{u}_\ell^0\|_{L^p} \\ &\lesssim_p 1. \end{aligned} \tag{315}$$

This uses the initialized L^p bounds, the local-response estimates, and $b_L c_L^2 \eta_V + b_L \eta_U \lesssim \lambda_L$. Lipschitzness of ϕ then bounds $\bar{\phi}_\ell^1$.

At $\ell = L$, the response identity and $\delta\bar{x}_L^0 = \delta\bar{f}^0 \xi_{\mathrm{out}}$ give

$$D_{\xi_{\mathrm{out}}} \bar{x}_L^1 = R_L^1 \delta\bar{f}^0. \tag{316}$$

The readout enters the updated residual only through the initialized adjoints, while the other Gaussian arguments are independent of it. Thus, using Gaussian integration by parts on ξ_{out}, yields

$$\mathbb{E}[\bar{x}_L^1 \xi_{\mathrm{out}}^\top] = \mathbb{E}[D_{\xi_{\mathrm{out}}} \bar{x}_L^1] = R_L^1 \delta\bar{f}^0. \tag{317}$$

Substitution into Eq. (199) at $k = 1$ proves Eq. (314). Its boundedness follows from the bounds on R_L^1 and $\bar{C}_{x,L}^{1,0}$. The Lipschitz loss-gradient assumption then gives

$$\|\delta\bar{f}^1\|_F \le \|\delta\bar{f}^0\|_F + C_{\mathrm{loss}} \|\bar{f}^1\|_F. \tag{318}$$

Finally, the updated boundary representatives are

$$\bar{e}^1 = \bar{e}^0 - \eta_E S^\top \delta\bar{x}_0^0, \qquad \xi_{\mathrm{out}} - \eta_{\mathrm{out}} (\delta\bar{f}^0)^\top \bar{x}_L^0. \tag{319}$$

Their L^p bounds follow from the initialized moment bounds and the bounded boundary learning-rate coefficients. □

Sufficient depth scalings. The conditions obtained above are

$$\begin{aligned} &L(a_L^2 + b_L^2) = O(1), \qquad c_L = O(1), \\ &\eta_V \sqrt{a_L^2 + b_L^2 c_L^2} + b_L(\eta_U + \eta_M) = O(1), \\ &L\left[\eta_V (a_L^2 + b_L^2 c_L^2) + b_L^2 (\eta_U + \eta_M)\right] = O(1), \\ &\eta_E = O(1), \qquad \eta_{\mathrm{out}} = O(1). \end{aligned} \tag{320}$$

The first line is inherited from initialization; the second controls local field corrections; the third controls their accumulation through depth. The final line controls the embedding and readout boundaries. The readout coefficient is needed for the updated output bound, rather than the hidden forward recursion, because the parameter updates are simultaneous.

These estimates use the initialized attention adjoints multiplied by a_L, so they apply without an additional bound on $b_L c_L / a_L$. Under the balanced-ratio condition Eq. (250), $\nu_L \asymp a_L$, simplifying the local condition to $\eta_V a_L + b_L(\eta_U + \eta_M) = O(1)$. All conditions are sufficient for the first updated forward pass; subsequent training steps require the corresponding later-time responses.

Corollary 7 (Common power-law scaling)**.** *Suppose $a_L = b_L = L^{-\alpha}$, $c_L = O(1)$, and $\eta_V = \eta_U = \eta_M = \eta_L \ge 0$, with bounded embedding and readout learning-rate coefficients. A sufficient condition for the uniform bounds above is*

$$\alpha \ge \frac{1}{2}, \qquad \eta_L \lesssim \min\{L^\alpha, L^{2\alpha - 1}\}. \tag{321}$$

For $\alpha \in [1/2, 1]$, this reduces to $\eta_L \lesssim L^{2\alpha-1}$.

Proof. Here $\gamma_L = 2L^{-2\alpha}$, $\nu_L \asymp L^{-\alpha}$, $\lambda_L \asymp \eta_L L^{-\alpha}$, and $\rho_L \asymp \eta_L L^{-2\alpha}$. The claims follow by substituting these expressions into Eq. (320), together with Eq. (257). □

B.6 ACCUMULATED INTERACTION AFTER ONE UPDATE

We now bound the accumulated attention-to-MLP interaction using the one-update field and response estimates. We retain the unmasked setting and the uniform-stability assumptions of Proposition 10. For the updated network, define

$$\mathcal{I}_\ell^1(X) := \mathcal{M}_\ell^1\big(X + c_L\mathcal{A}_\ell^1(X)\big) - \mathcal{M}_\ell^1(X). \tag{322}$$

Both evaluations use the same updated MLP parameters. Only the attention contribution to the MLP input is removed, and both terms are evaluated at the residual input of the full updated network. The residual recursion therefore gives the exact decomposition

$$X_L^1 = X_0^1 + a_L\sum_{\ell=0}^{L-1}\mathcal{A}_\ell^1(X_\ell^1) + b_L\sum_{\ell=0}^{L-1}\mathcal{M}_\ell^1(X_\ell^1) + b_L\sum_{\ell=0}^{L-1}\mathcal{I}_\ell^1(X_\ell^1). \tag{323}$$

The last sum is the accumulated explicit within-block interaction.

Joint full and attention-ablated evaluations. At each layer, write

$$U_{\ell,x}^1 := X_\ell^1 W_{U,\ell}^{1\top}, \qquad \Phi_{\ell,x}^1 := \phi(U_{\ell,x}^1), \qquad M_{\ell,x}^1 := \Phi_{\ell,x}^1 W_{M,\ell}^{1\top}. \tag{324}$$

Their differences from the full evaluations satisfy

$$\Delta U_\ell^1 := U_\ell^1 - U_{\ell,x}^1 = c_L A_\ell^1 W_{U,\ell}^{1\top}, \tag{325}$$

$$\Delta\Phi_\ell^1 := \Phi_\ell^1 - \Phi_{\ell,x}^1, \qquad \Delta M_\ell^1 := M_\ell^1 - M_{\ell,x}^1 = \Delta\Phi_\ell^1 W_{M,\ell}^{1\top}. \tag{326}$$

These additional forward evaluations use the trained matrices without altering the training updates. In particular, their cumulative-update terms retain the original training inputs H_ℓ^0 and Φ_ℓ^0:

$$\Delta U_\ell^1 = c_L A_\ell^1 W_{U,\ell}^{0\top} - \eta_U b_L c_L\left(\frac{1}{n}A_\ell^1 H_\ell^{0\top}\right)\delta U_\ell^0, \tag{327}$$

$$\Delta M_\ell^1 = \Delta\Phi_\ell^1 W_{M,\ell}^{0\top} - \eta_M b_L\left(\frac{1}{n}\Delta\Phi_\ell^1\Phi_\ell^{0\top}\right)\delta M_\ell^0. \tag{328}$$

For their joint representatives, set

$$\Delta\bar{u}_\ell^1 := \bar{u}_\ell^1 - \bar{u}_{x,\ell}^1, \quad \Delta\bar{\phi}_\ell^1 := \phi(\bar{u}_\ell^1) - \phi(\bar{u}_{x,\ell}^1), \quad \bar{\mathcal{I}}_\ell^1 := \bar{m}_\ell^1 - \bar{m}_{x,\ell}^1. \tag{329}$$

Lemma 12 (Joint interaction representation)**.** *Under Proposition 6, the additional full and attention-ablated evaluations have joint representatives satisfying*

$$\Delta\bar{u}_\ell^1 = c_L\xi_{U,a,\ell}^1 + c_L\left(R_{a,\zeta_U,\ell}^{1,0} - \eta_U b_L\bar{C}_{a,h,\ell}^{1,0}\right)\delta\bar{u}_\ell^0, \tag{330}$$

$$\bar{\mathcal{I}}_\ell^1 = \xi_{M,\Delta\phi,\ell}^1 + \left(R_{\Delta\phi,\zeta_M,\ell}^{1,0} - \eta_M b_L\bar{C}_{\Delta\phi,\phi,\ell}^{1,0}\right)\delta\bar{m}_\ell^0. \tag{331}$$

Here

$$\bar{C}_{a,h,\ell}^{1,0} := \mathbb{E}[\bar{a}_\ell^1\bar{h}_\ell^{0\top}], \qquad \bar{C}_{\Delta\phi,\phi,\ell}^{1,0} := \mathbb{E}[\Delta\bar{\phi}_\ell^1\bar{\phi}_\ell^{0\top}], \tag{332}$$

$$R_{a,\zeta_U,\ell}^{1,0} := \mathbb{E}[D_{\zeta_{U,\ell}^0}\bar{a}_\ell^1], \qquad R_{\Delta\phi,\zeta_M,\ell}^{1,0} := \mathbb{E}[D_{\zeta_{M,\ell}^0}\Delta\bar{\phi}_\ell^1]. \tag{333}$$

The centered Gaussian components have covariances

$$\mathbb{E}[\xi_{U,a,\ell}^1\xi_{U,a,\ell}^{1\top}] = \mathbb{E}[\bar{a}_\ell^1\bar{a}_\ell^{1\top}], \qquad \mathbb{E}[\xi_{M,\Delta\phi,\ell}^1\xi_{M,\Delta\phi,\ell}^{1\top}] = \mathbb{E}[\Delta\bar{\phi}_\ell^1\Delta\bar{\phi}_\ell^{1\top}]. \tag{334}$$

They belong to the same joint Gaussian families as the corresponding full evaluations. The components $\xi_{M,\Delta\phi,\ell}^1$ are independent across layers and jointly independent of the initial readout and initialized backward Gaussian components.

Proof. Insert the additional evaluations into the step-1 forward pass, without changing any training field or update. After the expansions Eq. (327)– Eq. (328), they use the same Gaussian matrices, coordinatewise operations, and empirical averages as in Proposition 6. The augmented computation therefore satisfies the same Tensor-Program hypotheses.

The preceding opposite-orientation evaluation of $W^0_{U,\ell}$ has input δU^0_ℓ and Gaussian component $\zeta^0_{U,\ell}$. Applying ZHat/ZDot to $A^1_\ell W^{0\top}_{U,\ell}$ gives its Gaussian component and the response $R^{1,0}_{a,\zeta_U,\ell}\delta\bar{u}^0_\ell$. The empirical average in Eq. (327) converges to $\bar{C}^{1,0}_{a,h,\ell}$, proving Eq. (330). The same argument for $W^0_{M,\ell}$, with input $\Delta\Phi^1_\ell$ and preceding backward input δM^0_ℓ, gives Eq. (331). These are the rules of (Yang & Hu, 2021, Definition G.3), with joint convergence supplied by its Theorem G.4.

The Gaussian covariance rule couples evaluations using the same matrix through their input cross moments. If $\xi^1_{U,x,\ell}$ and $\xi^1_{M,x,\ell}$ denote the ablated Gaussian components, linearity of the inputs gives

$$\xi^1_{U,\ell} - \xi^1_{U,x,\ell} = c_L\xi^1_{U,a,\ell}, \qquad \xi^1_{M,\ell} - \xi^1_{M,x,\ell} = \xi^1_{M,\Delta\phi,\ell} \tag{335}$$

almost surely in the joint construction. The covariance rule gives Eq. (334). Independence across distinct output matrices $W^0_{M,\ell}$ and from backward Gaussian families and initial vectors follows from ZHat. □

The added evaluations also give, at each fixed depth,

$$\frac{1}{nT}\left\| b_L \sum_{\ell=0}^{L-1} \mathcal{I}^1_\ell(X^1_\ell)\right\|_F^2 \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{T}\mathbb{E}\left\| b_L \sum_{\ell=0}^{L-1} \bar{\mathcal{I}}^1_\ell\right\|_2^2. \tag{336}$$

This is the empirical-convergence result applied to the squared norm of the finite sum of interaction fields.

Accumulation through depth. We next bound the Gaussian component and response correction in Eq. (331). The response difference is obtained by differentiating the full and ablated preactivations at the same fixed joint law.

Proposition 11 (Accumulated interaction after one update)**.** *Assume the uniform-stability hypotheses of Proposition 10, with ϕ' Lipschitz. Then*

$$\frac{1}{\sqrt{T}}\left\| b_L \sum_{\ell=0}^{L-1} \bar{\mathcal{I}}^1_\ell\right\|_{L^2} \lesssim |c_L|\left[\sqrt{L}\, b_L + Lb_L^2\big(1 + |c_L|\eta_V + \eta_U + \eta_M\big)\right]. \tag{337}$$

The constant is independent of L and the individual learning rates; it may depend on the fixed problem data and the uniform-stability bounds. Since $|c_L| \lesssim 1$, the same estimate holds with $|c_L|\eta_V$ replaced by η_V.

Proof. The preceding stability results give uniform L^2 bounds for the initialized and updated forward fields, the initialized MLP adjoints, and R^1_ℓ. We use the local learning-rate condition

$$\eta_V\nu_L + b_L(\eta_U + \eta_M) \lesssim 1, \qquad \nu_L := \sqrt{a_L^2 + b_L^2 c_L^2}, \tag{338}$$

from Eq. (282). All moment kernels and response coefficients below are deterministic, and derivatives hold them and the other Gaussian arguments fixed.

Preactivation and activation differences. Differentiating $\bar{a}^1_\ell = \bar{P}^0_\ell\bar{v}^1_\ell$ using Eq. (276) and Eq. (277), then taking expectation, gives

$$R^{1,0}_{a,\zeta_U,\ell} = b_L c_L \bar{P}^0_\ell\left(R^1_\ell - \eta_V\bar{C}^{1,0}_{x,\ell}\right)\bar{P}^{0\top}_\ell. \tag{339}$$

By Cauchy–Schwarz, $\|\bar{C}^{1,0}_{a,h,\ell}\|_{\mathrm{op}} \le \|\bar{a}^1_\ell\|_{L^2}\|\bar{h}^0_\ell\|_{L^2} \lesssim 1$. The Gaussian covariance in Eq. (334) likewise gives $\|\xi^1_{U,a,\ell}\|_{L^2} = \|\bar{a}^1_\ell\|_{L^2} \lesssim 1$. Consequently, Eq. (330) yields

$$\|\Delta\bar{u}^1_\ell\|_{L^2} \lesssim |c_L|\,[1 + b_L|c_L|(1+\eta_V) + b_L\eta_U] \lesssim |c_L|. \tag{340}$$

The last step uses $b_L|c_L| \le \nu_L$, the boundedness of b_L, c_L, and Eq. (338). Lipschitzness of ϕ and Cauchy–Schwarz then give

$$\|\Delta\bar{\phi}^1_\ell\|_{L^2} \le C_\phi\|\Delta\bar{u}^1_\ell\|_{L^2} \lesssim |c_L|, \tag{341}$$

$$\|\xi^1_{M,\Delta\phi,\ell}\|_{L^2} = \|\Delta\bar{\phi}^1_\ell\|_{L^2} \lesssim |c_L|, \qquad \|\bar{C}^{1,0}_{\Delta\phi,\phi,\ell}\|_{\mathrm{op}} \lesssim |c_L|. \tag{342}$$

Response difference. We denote the coefficients of $\delta\bar{u}_\ell^0$ in the full and ablated preactivations by

$$B_\ell := R^{1,0}_{h,\zeta_U,\ell} - \eta_U b_L \bar{C}^{1,0}_{h,\ell},$$
$$B_{x,\ell} := b_L R^1_\ell - \eta_U b_L \mathbb{E}[\bar{x}^1_\ell \bar{h}^{0\top}_\ell]. \tag{343}$$

Indeed, the ablated initialized-matrix evaluation has response $R^{1,0}_{x,\zeta_U,\ell} = b_L R^1_\ell$ by Eq. (275), while its weight update still uses H^0_ℓ. Thus

$$\bar{u}^1_\ell = \xi^1_{U,\ell} + B_\ell \delta\bar{u}^0_\ell, \qquad \bar{u}^1_{x,\ell} = \xi^1_{U,x,\ell} + B_{x,\ell}\delta\bar{u}^0_\ell, \tag{344}$$

where $\xi^1_{U,x,\ell}$ is the Gaussian component for $X^1_\ell W^{0\top}_{U,\ell}$ in the same joint construction. By linearity of the responses and kernels,

$$B_\ell - B_{x,\ell} = c_L R^{1,0}_{a,\zeta_U,\ell} - \eta_U b_L c_L \bar{C}^{1,0}_{a,h,\ell}. \tag{345}$$

It follows that

$$\|B_{x,\ell}\|_{\mathrm{op}} \lesssim b_L(1+\eta_U), \qquad \|B_\ell - B_{x,\ell}\|_{\mathrm{op}} \lesssim |c_L| b_L \big[|c_L|(1+\eta_V) + \eta_U\big]. \tag{346}$$

For either preactivation, the explicit dependence on $\zeta^0_{M,\ell}$ is through $\delta\bar{u}^0_\ell = \zeta^0_{M,\ell} \odot \phi'(\bar{u}^0_\ell)$. Applying the chain rule and taking expectation therefore gives

$$\begin{aligned} R^{1,0}_{\Delta\phi,\zeta_M,\ell} = \mathbb{E}\big[\mathrm{diag}\big(\phi'(\bar{u}^1_\ell) - \phi'(\bar{u}^1_{x,\ell})\big)\, B_{x,\ell}\, \mathrm{diag}(\phi'(\bar{u}^0_\ell))\big] \\ + \mathbb{E}\big[\mathrm{diag}(\phi'(\bar{u}^1_\ell))(B_\ell - B_{x,\ell})\, \mathrm{diag}(\phi'(\bar{u}^0_\ell))\big]. \end{aligned} \tag{347}$$

The first term records the change in the activation derivative; the second records the change in its preactivation coefficient. Boundedness and Lipschitzness of ϕ' give

$$\begin{aligned} \|R^{1,0}_{\Delta\phi,\zeta_M,\ell}\|_{\mathrm{op}} &\lesssim \|B_{x,\ell}\|_{\mathrm{op}} \|\Delta\bar{u}^1_\ell\|_{L^2} + \|B_\ell - B_{x,\ell}\|_{\mathrm{op}} \\ &\lesssim |c_L| b_L (1 + |c_L|\eta_V + \eta_U). \end{aligned} \tag{348}$$

Using Eq. (342) and the initialized bound on $\delta\bar{m}^0_\ell$ in Eq. (331) then yields

$$\big\|\bar{\mathcal{I}}^1_\ell - \xi^1_{M,\Delta\phi,\ell}\big\|_{L^2} \lesssim |c_L| b_L (1 + |c_L|\eta_V + \eta_U + \eta_M). \tag{349}$$

Accumulated interaction. By Lemma 12, the components $\xi^1_{M,\Delta\phi,\ell}$ are independent and centered across layers. Thus

$$\left\| b_L \sum_{\ell=0}^{L-1} \xi^1_{M,\Delta\phi,\ell} \right\|^2_{L^2} = b_L^2 \sum_{\ell=0}^{L-1} \|\xi^1_{M,\Delta\phi,\ell}\|^2_{L^2} \lesssim c_L^2 L b_L^2. \tag{350}$$

The remainder in each layer is a deterministic matrix times $\delta\bar{m}^0_\ell = \delta\bar{x}^0_{\ell+1}$, whose randomness is explicitly

$$\delta\bar{m}^0_\ell = \delta\bar{f}^0 \xi_{\mathrm{out}} + \sum_{r=\ell+1}^{L-1} (b_L \zeta^0_{U,r} + a_L \zeta^0_{V,r}). \tag{351}$$

Hence the entire Gaussian sum is independent of the entire remainder sum. The remainders may be correlated across layers through this shared backward history. Minkowski's inequality and Eq. (349) give

$$\left\| b_L \sum_{\ell=0}^{L-1} \big(\bar{\mathcal{I}}^1_\ell - \xi^1_{M,\Delta\phi,\ell}\big) \right\|_{L^2} \lesssim |c_L| L b_L^2 (1 + |c_L|\eta_V + \eta_U + \eta_M). \tag{352}$$

Taking square roots in Eq. (350) and combining with Eq. (352) by the triangle inequality proves Eq. (337), with the fixed normalization $T^{-1/2}$ absorbed into the constant. □

Corollary 8 (Interaction under common power-law scaling)**.** *Under Proposition 11, suppose $a_L = b_L = L^{-\alpha}$ and $\eta_V = \eta_U = \eta_M = \eta_L$, with $\alpha \geq 1/2$. Then*

$$\frac{1}{\sqrt{T}} \left\| b_L \sum_{\ell=0}^{L-1} \bar{\mathcal{I}}^1_\ell \right\|_{L^2} \lesssim |c_L| \left[L^{1/2-\alpha} + \eta_L L^{1-2\alpha} \right]. \tag{353}$$

For $c_L = a_L = L^{-\alpha}$, this becomes

$$\frac{1}{\sqrt{T}} \left\| b_L \sum_{\ell=0}^{L-1} \bar{\mathcal{I}}^1_\ell \right\|_{L^2} \lesssim L^{1/2-2\alpha} + \eta_L L^{1-3\alpha}. \tag{354}$$

Table 3: Normalized L^2 accumulated-interaction upper bounds after one update, in the sequential width-then-depth regime. The learning-rate coefficient η_c is fixed independently of depth.

$a_L = b_L$	η_L	$c_L = a_L$	$\lvert c_L \rvert = \Theta(1)$
$L^{-1/2}$	η_c	$O((1+\eta_c)L^{-1/2})$	$O(1+\eta_c)$
L^{-1}	$\eta_c L$	$O(L^{-3/2}+\eta_c L^{-1})$	$O(L^{-1/2}+\eta_c)$

Proof. Substitute the common scaling into Eq. (337). Since $|c_L| \lesssim 1$ and $L^{1-2\alpha} \leq L^{1/2-\alpha}$ for $\alpha \geq 1/2$, the learning-rate-independent remainder is absorbed by the Gaussian term. The tied case follows by setting $|c_L| = L^{-\alpha}$. □

Remark 4 (Endpoint interaction bounds)**.** *The sufficient learning-rate conditions permit $\eta_L = \eta_c$ at $\alpha = 1/2$ and $\eta_L = \eta_c L$ at $\alpha = 1$, with fixed $\eta_c \geq 0$ and bounded embedding and readout coefficients. Table 3 records the resulting normalized L^2 bounds. With $c_L = a_L$, the accumulated interaction vanishes at both endpoints. For order-one c_L, these estimates give an $O(1)$ upper bound when $\eta_c > 0$, but do not establish a nonvanishing interaction. At $\alpha = 1$, the two terms have orders $L^{-1/2}$ and η_c, respectively.*

C EXTENSION TO ADAM THROUGH SIGNSGD

We extend the preceding depth-scaling analysis using signSGD as a scale-invariant surrogate for Adam. We retain the model, initialization, and simultaneous parameter updates of Sections B.1–B.2. The optimizer identities below are exact; the subsequent depth and interaction estimates are scaling predictions based on the SGD analysis, rather than additional convergence theorems.

Adam's first update. Let g^0 denote a parameter gradient at initialization. With zero moments and bias-corrected estimates, Adam has $\widehat{m}_1 = g^0$ and $\widehat{v}_1 = (g^0)^2$, as its first update is

$$\Delta w = -\eta \frac{g^0}{|g^0| + \varepsilon}. \tag{355}$$

For $\varepsilon = 0$, this reduces exactly to $-\eta \operatorname{sign}(g^0)$ for nonzero gradient entries, with $\operatorname{sign}(0) = 0$ (Kingma & Ba, 2015). Thus, at the first update, Adam removes the gradient magnitude and retains primarily its sign, motivating signSGD as a convenient model for analyzing the corresponding depth and width scaling. This equivalence is specific to the first step: at later steps, Adam also incorporates gradient history through its moment estimates (Bernstein et al., 2018; Balles & Hennig, 2018).

Width and depth factors in the update. For a hidden linear map $Y_{\theta,\ell} = Z_{\theta,\ell} W_{\theta,\ell}^\top$, the initialization gradient has the form

$$\nabla_{W_{\theta,\ell}} \mathcal{L}^0 = \frac{\rho_{\theta,L}}{n} \mathsf{G}^0_{\theta,\ell}, \qquad \mathsf{G}^0_{\theta,\ell} := (\delta Y^0_{\theta,\ell})^\top Z^0_{\theta,\ell}, \tag{356}$$

where $\rho_{\theta,L} = a_L$ for $\theta \in \{Q, K, V\}$ and $\rho_{\theta,L} = b_L$ for $\theta \in \{U, M\}$. Since $\rho_{\theta,L} > 0$, sign normalization removes both $\rho_{\theta,L}$ and $1/n$ from the gradient magnitude. The required μP width scaling is therefore supplied by the learning rate:

$$W^1_{\theta,\ell} = W^0_{\theta,\ell} - \frac{\eta_{\theta,L}}{n} \operatorname{sign}(\mathsf{G}^0_{\theta,\ell}), \tag{357}$$

where $\eta_{\theta,L} \geq 0$ is independent of width but may depend on depth. Thus the actual hidden-weight learning rate is $\eta_{\theta,L}/n$, consistent with the Adam width scaling of Yang et al. (2021, Table 3).

Substituting Eq. (357) into the next forward pass gives

$$Y^1_{\theta,\ell} = Z^1_{\theta,\ell} W^{0\top}_{\theta,\ell} - \frac{\eta_{\theta,L}}{n} Z^1_{\theta,\ell} \operatorname{sign}(\mathsf{G}^0_{\theta,\ell})^\top. \tag{358}$$

Since an $n \times n$ sign matrix has operator norm at most n,

$$\|\Delta W_{\theta,\ell}\|_{\mathrm{op}} \leq \eta_{\theta,L}, \qquad \frac{\|Z^1_{\theta,\ell} \Delta W^\top_{\theta,\ell}\|_F}{\sqrt{n}} \leq \eta_{\theta,L} \frac{\|Z^1_{\theta,\ell}\|_F}{\sqrt{n}}. \tag{359}$$

Hence the direct local feature correction is controlled at scale $\eta_{\theta,L}$, rather than $\eta_\theta \rho_{\theta,L}$ as under SGD. The factor $1/n$ therefore remains necessary in the learning rate even though sign normalization removes it from the gradient.

For the boundary layers, the corresponding updates are

$$E^1 = E^0 - \eta_{E,L} \operatorname{sign}(S^\top \delta X^0_0), \tag{360}$$

$$W^1_{\mathrm{out}} = W^0_{\mathrm{out}} - \frac{\eta_{\mathrm{out},L}}{n} \operatorname{sign}\big((\delta f^0)^\top X^0_L\big). \tag{361}$$

Thus, bounded $\eta_{E,L}$ and $\eta_{\mathrm{out},L}$ yield controlled boundary-layer increments, whereas the $n\eta_E$ SGD embedding rate causes width divergence under signSGD.

Initialization and one-update depth scaling. The optimizer does not affect the initialized forward or backward pass, so the sufficient stability condition from Section B.4 remains

$$L(a_L^2 + b_L^2) = O(1), \qquad c_L = O(1). \tag{362}$$

After one sign-normalized update, the gradient factor ρ_L is removed, while the outer residual multiplier remains. Hence a branch with residual multiplier ρ_L has local feature-update scale $\eta_{\theta,L}$

and residual-update scale $\rho_L \eta_{\theta,L}$; under SGD the corresponding scales are $\rho_L \eta_\theta$ and $\rho_L^2 \eta_\theta$. Since updates across layers may share backward information, we conservatively allow their learning contributions to accumulate at order L rather than $\sqrt{L}$.

We assume that, after extracting these explicit scales, the normalized-gradient contractions and their averaged sensitivities introduce no additional powers of depth. This is an assumption on the normalized gradient law rather than a consequence of μP alone. It can be formulated using a smooth sign surrogate at fixed normalized-gradient scale; the ideal sign limit additionally requires control near zero. The nonlinear-outer-product framework of Littwin & Yang (2023) provides a suitable setting for such adaptive updates, so the SGD Tensor-Program kernels cannot simply be reused unchanged.

For separate hidden learning-rate coefficients, define

$$\eta_{\mathrm{att},L} := \eta_{Q,L} + \eta_{K,L} + \eta_{V,L}, \qquad \eta_{\mathrm{mlp},L} := \eta_{U,L} + \eta_{M,L}. \tag{363}$$

A conservative sufficient criterion is

$$\begin{aligned} &\eta_{\mathrm{att},L} + \eta_{\mathrm{mlp},L} = O(1), \\ &L\left[(a_L + b_L|c_L|)\eta_{\mathrm{att},L} + b_L \eta_{\mathrm{mlp},L}\right] = O(1), \qquad \eta_{E,L}, \eta_{\mathrm{out},L} = O(1). \end{aligned} \tag{364}$$

The first line controls local field changes. In the second, attention updates enter the residual directly through a_L and indirectly through the MLP input via $b_L c_L$, while MLP updates enter through b_L.

For $a_L = b_L = L^{-\alpha}$ and a common hidden coefficient $\eta_{\theta,L} = \eta_L$, the conditions reduce to

$$\eta_L = O(1), \qquad L^{1-\alpha}\eta_L = O(1), \qquad \eta_L \lesssim \min\{1, L^{\alpha-1}\}, \quad \alpha \geq \tfrac{1}{2}. \tag{365}$$

By comparison, SGD requires $\eta_L^{\mathrm{SGD}} \lesssim \min\{L^\alpha, L^{2\alpha-1}\}$. For $\alpha \in [1/2, 1]$, comparable depth scales are obtained with

$$\eta_L = \eta_c L^{\alpha-1}, \qquad \eta_L^{\mathrm{SGD}} = \eta_c L^{2\alpha-1} \quad \Longrightarrow \quad \eta_L = L^{-\alpha}\eta_L^{\mathrm{SGD}}. \tag{366}$$

Thus, the effective sign-normalized coefficient is $\eta_c/\sqrt{L}$ for $\alpha = 1/2$ and η_c for $\alpha = 1$, while the actual hidden parameter learning rates include an additional factor $1/n$. These agree with the hidden-weight Adam scalings of Depth-μP and CompleteP, respectively.

Accumulated attention-to-MLP interaction. We retain the interaction definition Eq. (322), evaluating both terms with the same sign-updated weights. Their MLP preactivation difference is

$$\Delta U_\ell^1 = c_L A_\ell^1 W_{U,\ell}^{1\top}, \qquad \mathcal{I}_\ell^1(X_\ell^1) = [\phi(U_\ell^1) - \phi(U_{\ell,x}^1)]W_{M,\ell}^{1\top}. \tag{367}$$

Sign normalization acts on the training gradient and therefore does not remove c_L from this evaluation difference. At finite width,

$$\|\mathcal{I}_\ell^1(X_\ell^1)\|_F \leq |c_L|\kappa_\phi \|W_{U,\ell}^1\|_{\mathrm{op}} \|W_{M,\ell}^1\|_{\mathrm{op}} \|A_\ell^1\|_F. \tag{368}$$

Moreover, the direct contribution from the MLP output-weight update satisfies

$$\frac{\|\Delta\Phi_\ell^1 \Delta W_{M,\ell}^\top\|_F}{\sqrt{nT}} \leq \eta_{M,L} \frac{\|\Delta\Phi_\ell^1\|_F}{\sqrt{nT}}, \tag{369}$$

and hence has scale $|c_L|\eta_{M,L}$ when the local interaction difference has scale $|c_L|$.

To estimate accumulation across depth, we use the same centered-versus-response decomposition as in Section B.6. We assume that the centered component retains its cross-layer cancellation and that, after extracting the explicit scales, normalized-update contractions introduce no additional powers of depth. For $a_L = b_L = L^{-\alpha}$ and a common hidden coefficient $\eta_{\theta,L} = \eta_L$, this gives

$$\begin{aligned} \frac{1}{\sqrt{T}} \left\| b_L \sum_{\ell=0}^{L-1} \bar{\mathcal{I}}_\ell^1 \right\|_{L^2} &\lesssim |c_L| \left[L^{1/2-\alpha} + L^{1-2\alpha} + \eta_L L^{1-\alpha} \right] \\ &\lesssim |c_L| \left[L^{1/2-\alpha} + \eta_L L^{1-\alpha} \right], \qquad \alpha \geq \tfrac{1}{2}. \end{aligned} \tag{370}$$

The second line uses initialization stability to absorb $L^{1-2\alpha}$ into $L^{1/2-\alpha}$. Relative to SGD, sign normalization changes the learning-induced scale from $\eta_L^{\mathrm{SGD}} L^{1-2\alpha}$ to $\eta_L L^{1-\alpha}$.

Table 4: One-update interaction scaling under SGD and sign-normalized updates. The optimizer-specific effective hidden learning-rate coefficients differ, while the resulting interaction bounds coincide. For sign-normalized updates, the actual hidden parameter learning rate is η_L/n.

$a_L = b_L$	η_L^{SGD}	η_L	$c_L = a_L$	$\lvert c_L\rvert = \Theta(1)$
$L^{-1/2}$	η_c	$\eta_c L^{-1/2}$	$O((1+\eta_c)L^{-1/2})$	$O(1+\eta_c)$
L^{-1}	$\eta_c L$	η_c	$O(L^{-3/2}+\eta_c L^{-1})$	$O(L^{-1/2}+\eta_c)$

Using the depth-stable choice $\eta_L = \eta_c L^{\alpha-1}$ for $\alpha \in [1/2, 1]$ gives

$$\frac{1}{\sqrt{T}}\left\| b_L \sum_{\ell=0}^{L-1} \bar{\mathcal{I}}_\ell^1 \right\|_{L^2} \lesssim \begin{cases} L^{1/2-2\alpha} + \eta_c L^{-\alpha}, & c_L = a_L = L^{-\alpha}, \\ L^{1/2-\alpha} + \eta_c, & |c_L| = \Theta(1). \end{cases} \tag{371}$$

The resulting endpoint rates are summarized in Table 4. Although sign-normalized updates require different depth-dependent learning rates from SGD, the accumulated-interaction bounds coincide with those in Table 3: the tied choice $c_L = a_L$ vanishes under both residual scalings, whereas $|c_L| = \Theta(1)$ removes this forced decay and allows an $O(1)$ learning contribution. As before, these are upper bounds and do not imply a positive lower bound.

The sharper tied rate relies on cancellation across layers beyond what uniform local bounds alone provide. Indeed, Eq. (368) gives only the triangle-inequality scale $O(|c_L|Lb_L)$, which vanishes when $c_L = a_L = b_L = L^{-1}$ but not under $L^{-1/2}$ scaling. For the latter, we assume an orthogonality structure analogous to the Gaussian cross-term cancellation established for the SGD one-update dynamics in Lemma 11, leading to second-moment rather than linear accumulation across depth. The resulting scaling agrees with the empirical decay in Section 6 and Figure 1. Proving the corresponding cancellation for sign-normalized or Adam updates remains an open direction.

D Experimental Details and Additional Results

This appendix provides the complete experimental setup and additional results complementing Section 6. The main text uses width $n = 256$ as the primary setting and reports the representative training results. Here we additionally report the corresponding validation results and repeat the main experiments at width $n = 128$ to examine whether the observed interaction scaling and performance trends depend on the reported metric or model width.

D.1 Experimental Setup

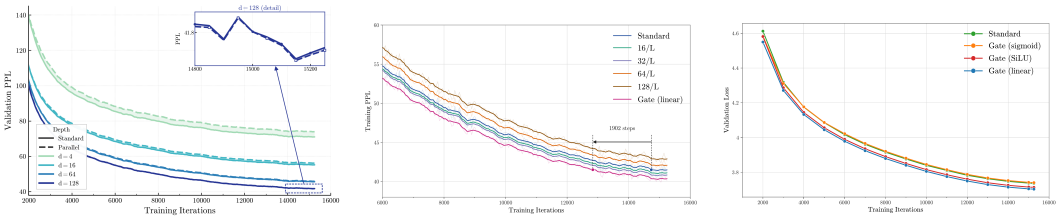
Models and data. All experiments use standard dense Transformers with multi-head self-attention and normalization, rather than the simplified architecture used in the theory. We primarily follow (Dey et al., 2026) and consider both Depth-μP and CompleteP residual parameterizations. Models are trained with AdamW on the English subset of C4. The primary experiments use width $n = 256$ and depths $L \in \{4, 16, 64, 128\}$, with a 1B-token training budget and the full C4-en validation split for evaluation. The additional experiments below repeat the relevant comparisons at width $n = 128$. Within each depth, width, and residual parameterization, all interaction variants use the same backbone, optimization setup, and data budget.

Interaction variants. Recall that, for sequential blocks, the MLP input is $H_\ell = X_\ell + c_L \mathcal{A}_\ell(X_\ell)$. We compare the tied choice $c_L = a_L$, explicitly parallel blocks ($c_L = 0$), and independent fixed coefficients $c_L = \Theta(1)$. For adaptive interaction, we replace c_L with an input-dependent gate, $H_\ell = X_\ell + g_\ell(X_\ell) \odot \mathcal{A}_\ell(X_\ell)$, while leaving the residual multipliers unchanged. Sigmoid, SiLU, and linear gates use the same predictor and differ only in their output activation: sigmoid restricts the coefficients to $(0, 1)$, whereas SiLU and linear gates permit both signs.

Training configuration. The shared training configuration is summarized in Table 5. The effective learning-rate coefficient is held fixed across widths and depths, with the actual depth dependence determined by the corresponding Depth-μP or CompleteP parameterization. The width-128 experiments use the same configuration as the width-256 experiments, changing only the model width.

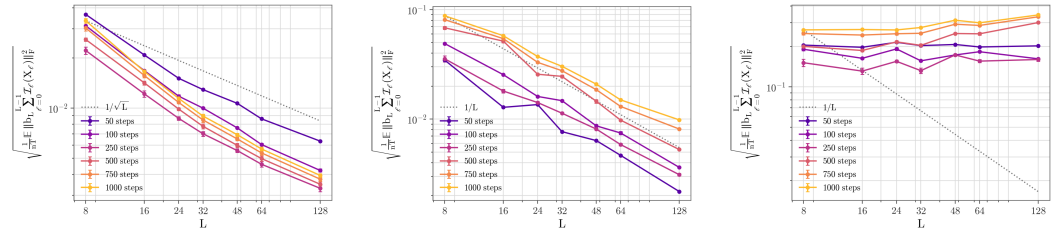
Table 5: Training hyperparameters shared across experiments, primarily following (Dey et al., 2026) for both CompleteP and Depth-μP scaling.

Hyperparameter	Value
Optimizer	AdamW
Peak learning rate	6×10^{-4}
Min. learning rate	6×10^{-5}
LR schedule	linear warmup + cosine decay
Warmup	152 steps
Adam (β_1, β_2)	(0.9, 0.95)
Adam ε	10^{-12}
Weight decay	0.1 (matrices only)
Gradient clip	1.0
Dropout	0
Bias	none
Context length	512
Per-GPU batch size	16
Number of GPUs	8
Gradient accumulation	1
Tokens per step	65,536
Token budget	1B
Training steps	15,259
Precision	bfloat16
Base width / base depth	256 / 1
Init. std	0.02
Seed	1337



(a) Tied vs. parallel (b) Fixed interaction strengths (c) Adaptive interaction

Figure 4: **Validation performance at width $n = 256$.** Validation results show the same trends as the training results in Figure 2: the tied–parallel gap narrows with depth, suitable independent coefficients improve over the tied baseline, and signed adaptive gates perform best.



(a) Depth-μP, tied interaction (b) CompleteP, tied interaction (c) Independent $c_L = \Theta(1)$

Figure 5: **Interaction scaling at width $n = 128$.** Tied interaction exhibits the predicted depth-dependent decay under Depth-μP and CompleteP, while $c_L = \Theta(1)$ remains nonvanishing, reproducing the width-256 behavior in Figure 1.

Interaction measurement. We use the normalized accumulated-interaction measure defined in Section 3. Unless otherwise stated, interaction statistics are evaluated at the indicated training checkpoints using the same models as the corresponding performance experiments.

D.2 Additional Results at Width $n = 256$

Validation performance. The main text reports training perplexity; here we evaluate the same models using validation perplexity. Figure 4 shows the corresponding results. The qualitative conclusions are unchanged: the sequential–parallel gap narrows with depth, suitable independent interaction coefficients improve over the tied baseline, and signed adaptive gates provide the strongest gains. The differences are smaller on validation but preserve the same ordering.

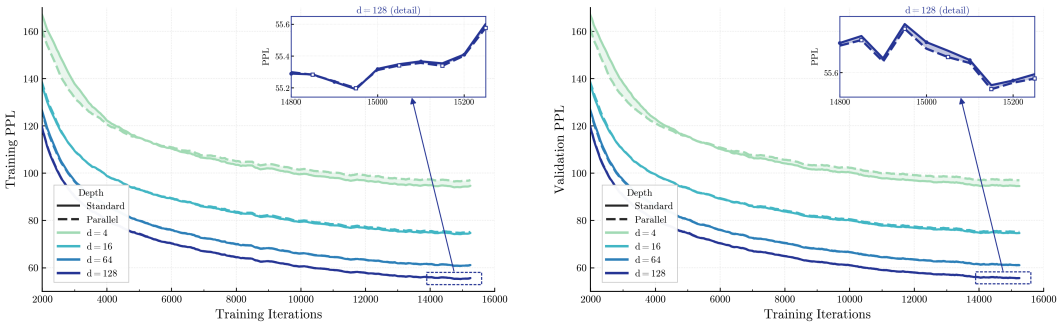
D.3 Width Robustness at $n = 128$

We next repeat the main empirical comparisons at width $n = 128$. These experiments test whether the interaction-scaling and performance trends reported at width 256 persist at a second model width.

Interaction scaling. Figure 5 repeats the accumulated-interaction measurement across depth and training checkpoints. As at width 256, the tied interaction decreases approximately as $L^{-1/2}$ under Depth-μP and L^{-1} under CompleteP, whereas an independent $c_L = \Theta(1)$ remains nonvanishing across depths. The predicted depth-scaling behavior is therefore consistent across both tested widths.

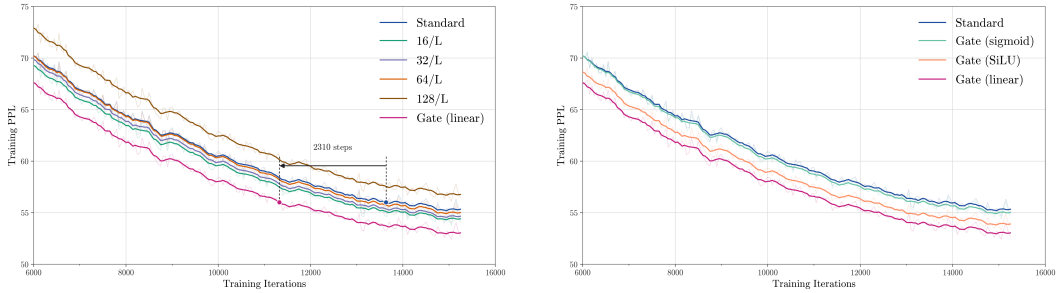
Sequential versus parallel blocks. We first repeat the sequential–parallel comparison. Figure 6 shows both training and validation perplexity. At shallow depths the sequential model retains a visible advantage, while the difference decreases with depth and becomes small for the deepest models. The same trend therefore appears at width 128 and in both training and validation performance.

Fixed and adaptive interaction. Figure 7 examines the deepest model, $L = 128$. Suitable fixed interaction coefficients improve over the tied baseline, while overly strong interaction can degrade performance, confirming that preserving interaction alone does not determine its optimal strength.



(a) Training perplexity (b) Validation perplexity

Figure 6: **Sequential versus parallel blocks at width $n = 128$.** The performance gap decreases with depth in both training and validation perplexity, consistent with the width-256 results.

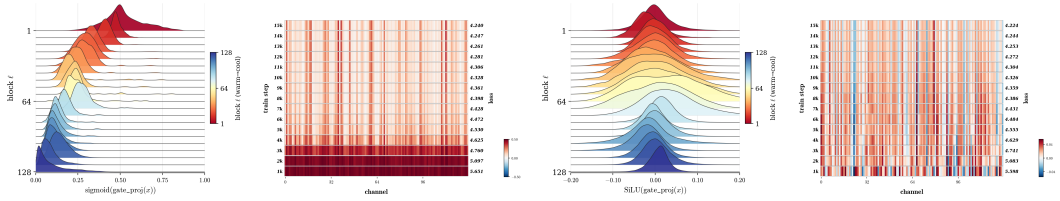


(a) Fixed interaction strengths (b) Adaptive interaction

Figure 7: **Fixed and adaptive interaction at width $n = 128$ and depth $L = 128$.** Suitable fixed coefficients improve over the tied baseline, while adaptive gates improve further, particularly the signed SiLU and linear gates.

Adaptive gates improve further. Among the tested activations, the signed SiLU and linear gates outperform the positive-only sigmoid gate, reproducing the ordering observed at width 256.

Adaptive interaction across blocks and training. Finally, we test whether the learned gate dynamics are robust to width. Figure 8 shows that sigmoid gates again shift toward zero in deeper blocks and later training, reducing the attention contribution to the MLP input. In contrast, SiLU gates retain substantial positive and negative values across depth and training. Thus, the width-128 results reproduce the gate behavior observed at width 256: positive-only gates tend to suppress the interaction, whereas signed gates adapt both its magnitude and direction.



(a) Sigmoid: blocks (b) Sigmoid: training (c) SiLU: blocks (d) SiLU: training

Figure 8: **Adaptive gate dynamics at width $n = 128$.** Sigmoid gates increasingly concentrate near zero in deeper blocks and later training, whereas SiLU gates retain substantial positive and negative values across blocks and checkpoints.

Summary. Across the two tested widths, we observe the same qualitative conclusions: the tied interaction decays with depth at the predicted rates, independent coefficients prevent this forced decay, suitable fixed interaction can improve learning, and signed adaptive gates provide further

gains. The width-256 validation results likewise preserve the performance trends seen in training. Together, these results indicate that the main empirical conclusions are not specific to the displayed training metric or to the width-256 setting.