MEAN–COVARIANCE COMPETITION GOVERNS COLLAPSE AND MULTI-CLUSTER STATES IN CAUSAL ATTENTION DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal self-attention can exhibit two distinct regimes: token representations may collapse toward a single dominant direction, forming attention sinks, or they may remain organized into multiple clusters. Despite extensive empirical observations, the mechanism governing the transition between these regimes remains poorly understood. In this work, we develop a dynamical and mean-field framework to explain this transition. We first use a first-order expansion of the causal softmax interaction to decompose the dynamics into two competing terms: a mean-driven alignment term, which depends on the prefix mean of previous tokens, and a covariance-driven routing term, which depends on the prefix covariance and the query–key interaction. We then show that, near an eigendirection of the value matrix, a spectral gap suppresses transverse perturbations and drives collapse toward a single dominant cluster, providing a mechanism for attention-sink formation. Conversely, when the cluster geometry keeps the prefix mean close to zero while maintaining nontrivial covariance, the alignment effect is weakened and covariance-driven routing remains active, allowing multi-cluster structures to persist. A key novelty of our approach is the use of Fourier analysis in the depth variable, which reveals the frequency-dependent structure of causal attention. These results identify causal attention dynamics as a competition between mean-driven spectral alignment and covariance-driven routing. These mechanisms are then examined by $\mathbb{S}^1$ simulations and layerwise analyses of pretrained Llama-3.1-8B model.

1 INTRODUCTION

Transformers Vaswani et al. (2017) have become the dominant architecture for modern sequence modeling, with applications ranging from natural language processing Dehghani et al. (2019); Dai et al. (2019); Al-Rfou et al. (2019); Baevski & Auli (2019); Raffel et al. (2020) to computer vision Dosovitskiy et al. (2021); Ramesh et al. (2021); Radford et al. (2021); Liu et al. (2021), general machine learning Tay et al. (2022); Khan et al. (2022); Lin et al. (2022), and reinforcement learning Janner et al. (2021); Chen et al. (2021). At the core of Transformers lies the attention mechanism, which updates each token representation by aggregating information from other tokens in the sequence. In autoregressive models, this interaction is constrained by a causal mask: each token can attend only to itself and previous tokens. Causal attention is widely used in large-scale generative models, including natural language and image generation systems such as DALL-E Ramesh et al. (2021), VQGAN Esser et al. (2021), and Parti Yu et al. (2022), as well as multimodal foundation models such as hameleon Team (2024), MUSE Chang et al. (2023), and AlphaFold 3 Abramson et al. (2024).

Despite the empirical success of causal self-attention, its internal dynamics remain only partially understood. In particular, causal attention can exhibit two distinct types of behavior. In some regimes, token representations concentrate toward a single dominant direction, producing what is often referred to as an attention sink Abella et al. (2025); Xiao et al. (2023); Geshkovski et al. (2023b). In other regimes, token representations remain separated into multiple clusters, suggesting that causal attention can also preserve or sustain directional diversity Geshkovski et al. (2024); Karagodin et al. (2024). Although these phenomena have been observed empirically and studied from different theoretical

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.

perspectives, the mechanism governing the transition between collapse and multi-cluster structure remains unclear. This raises a fundamental question:

> *What mechanism determines whether causal attention collapses to a single dominant direction or sustains a multi-cluster representation?*

Existing theoretical studies often isolate particular aspects of attention, such as spectral properties, simplified dynamical approximations, or mean-field limits. However, a unified mechanism-level explanation of collapse and multi-cluster formation in causal attention is still lacking. In particular, it remains unclear how the value matrix, the query–key interaction, and the causal prefix structure jointly shape the geometry of token representations.

1.1 Our Main Contributions

In this paper, we provide a unified mechanism-level explanation of how causal attention can transition between collapse toward a single dominant direction and the persistence of multi-cluster representations. Our main contributions are as follows.

- We derive a mean-field transport equation for the empirical distribution of tokens in causal-attention dynamics. We incorporate the token position into each Dirac mass in order to track the causal depth of each token.
- We decompose causal-attention dynamics into a mean-driven alignment term and a covariance-driven routing term. This structure carries from the finite-token system to the position-aware mean-field transport equation and its monokinetic reduction.
- Our main result uses Fourier analysis in logarithmic token position to reveal how the value spectrum controls the decay of linearized modes. The *spectral gap* governs low frequencies, the *leading eigenvalue* governs high frequencies, and Jordan blocks produce polynomial transients. The same linearization holds for exact softmax at any fixed finite inverse temperature.
- We construct stationary clustered measures for the truncated dynamics, including configurations with nonzero prefix covariance and vanishing projected drift.

We illustrate our theoretical findings through controlled simulations on $\mathbb{S}^1$, which exhibit the predicted collapse and multi-cluster regimes, and through layerwise probing of the pretrained Llama-3.1-8B model, which validates the mean-field assumption and confirms that the prefix mean-to-covariance ratio reliably indicates the cluster structure of token embeddings.

Organization. Section 2 derives the decomposition of the causal attention dynamic into mean-driven alignment and covariance-driven routing. Section 3 extends this decomposition to a depth-aware mean-field transport equation and its monokinetic reduction. Section 4 analyzes how these two terms drive attention-sink formation or sustain multi-cluster structure. Section 5 presents simulations and Llama-3.1-8B experiments supporting the theory.

2 Mean- and Covariance-Driven Decomposition of Causal Attention

Following theoretical analyses of attention dynamics Geshkovski et al. (2023a;b); Karagodin et al. (2024); Rigollet (2025), we study a single-head causal attention model with a residual connection and spherical normalization. Continuous time t describes evolution across layers, while normalized token position s will describe variation within the sequence. Our aim is to use this token-depth description to expose the value spectrum in Fourier-mode estimates.

Causal attention. Fix $d \geq 2$ and $L \geq 1$. Let $X = (x_\ell^\top)_{\ell=1}^L \in \mathbb{R}^{L\times d}$, where each $x_\ell \in \mathbb{R}^d$ is a column vector. Given $Q, K, V \in \mathbb{R}^{d\times d}$, causal attention is

$$\mathrm{CauAttn}(X)_\ell = \sum_{j=1}^{\ell} \frac{\exp\big(\beta\langle Qx_\ell, Kx_j\rangle\big)}{\sum_{k=1}^{\ell}\exp\big(\beta\langle Qx_\ell, Kx_k\rangle\big)} Vx_j, \qquad \ell = 1,\dots,L.$$

Here $\beta \geq 0$ is inverse temperature. A dot-product scaling factor such as $d^{-1/2}$ can be included in β or in $T = Q^\top K$. On the unit sphere, $|x_\ell^\top T x_j| \leq \|T\|$, so the expansion below is controlled by $\beta\|T\|$. Choosing $\beta = d^{-1/2}$ alone does not ensure a small correction unless the score scale is also controlled. Small β corresponds to high temperature.

Normalized residual dynamics. Scale normalization, including RMSNorm Zhang & Sennrich (2019), motivates separating token direction from magnitude. We study a spherical idealization with the normalized-residual update and sufficiently small step size $h > 0$:

$$x_\ell^{n+1} = \frac{x_\ell^n + h\ \mathrm{CauAttn}(X^n)_\ell}{\left|x_\ell^n + h\ \mathrm{CauAttn}(X^n)_\ell\right|}. \tag{1}$$

For a unit vector x and fixed increment z,

$$\frac{x + hz}{|x + hz|} = x + h\,\mathsf{P}_x(z) + O(h^2), \qquad \mathsf{P}_x(z) := z - (x \cdot z)x.$$

This first-order expansion motivates the continuous-depth model

$$\partial_t x_\ell(t) = \mathsf{P}_{x_\ell(t)}\left(\sum_{j=1}^{\ell} a_{\ell j}(t) V(t) x_j(t)\right), \qquad \ell = 1, \ldots, L, \tag{2}$$

where

$$a_{\ell j}(t) = \frac{e^{\beta s_{\ell j}(t)}}{\sum_{k=1}^{\ell} e^{\beta s_{\ell k}(t)}}, \qquad s_{\ell j}(t) = x_\ell(t)^\top T(t) x_j(t), \qquad T(t) = Q(t)^\top K(t).$$

The map $\mathsf{P}_x(z)$ projects onto $T_x\mathbb{S}^{d-1}$, so the flow preserves $|x_\ell(t)| = 1$. We assume that V, T are measurable and locally bounded in time; the trajectories are locally absolutely continuous and the evolution equations hold almost everywhere in t. Equation equation 1 specifies the spherical normalization used in this model; it normalizes the whole residual update.

Mean and covariance decomposition. Define the prefix statistics

$$\bar{x}_\ell(t) = \frac{1}{\ell}\sum_{j=1}^{\ell} x_j(t), \qquad \Sigma_\ell(t) = \frac{1}{\ell}\sum_{j=1}^{\ell} \big(x_j(t) - \bar{x}_\ell(t)\big)\big(x_j(t) - \bar{x}_\ell(t)\big)^\top,$$

and $\bar{s}_\ell = \ell^{-1}\sum_{j=1}^{\ell} s_{\ell j}$. For bounded scores, the softmax expansion in Lemma B.1 gives

$$a_{\ell j}(t) = \frac{1}{\ell} + \frac{\beta}{\ell}\big(s_{\ell j}(t) - \bar{s}_\ell(t)\big) + O\left(\frac{\beta^2 M^2}{\ell}\right), \tag{3}$$

where M bounds $\|T(t)\|$ on the time interval considered. For bounded βM, the remainder is uniform in j, ℓ, L. The covariance identity in Lemma B.2 is

$$\frac{1}{\ell}\sum_{j=1}^{\ell} \big(s_{\ell j}(t) - \bar{s}_\ell(t)\big) x_j(t) = \Sigma_\ell(t) T(t)^\top x_\ell(t). \tag{4}$$

Substituting these expressions into equation 2 and retaining first order in β defines the truncated model

$$\begin{aligned} \partial_t x_\ell(t) = {} & \underbrace{\mathsf{P}_{x_\ell(t)}\big(V(t)\bar{x}_\ell(t)\big)}_{\text{mean-driven alignment}} \\ & + \beta\ \underbrace{\mathsf{P}_{x_\ell(t)}\big(V(t)\Sigma_\ell(t)T(t)^\top x_\ell(t)\big)}_{\text{covariance-driven routing}}. \end{aligned} \tag{5}$$

At the same token configuration, the difference between the exact and truncated vector fields is at most $\beta^2\|V(t)\|\|T(t)\|^2$, independently of sequence length (Proposition B.3). This is an instantaneous error bound; the continuum and stationary results below concern equation 5.

The decomposition separates the drift generated by the prefix mean from the first query–key-dependent correction, which is modulated by the prefix covariance. The latter can alter tangential motion, but its direction depends on V, T, and the token configuration. The spectral role of the mean term becomes explicit in the subsequent linearization near a constant eigenprofile.

3 MEAN-FIELD LIMIT FOR CAUSAL ATTENTION DYNAMICS

We pass from the truncated token system equation 5 to a continuum description in normalized token position. Within the measure-based viewpoint Rigollet (2025), causal order is retained by recording token position together with its embedding:

$$\mu_t^L = \frac{1}{L}\sum_{\ell=1}^{L} \delta_{(\ell/L, x_\ell(t))}.$$

This position-aware representation, discussed in the related-work section, keeps the prefix information needed for the analysis.

Assumption 3.1. For solutions of equation 5, assume that the empirical measures converge weakly on $[0,1]\times\mathbb{S}^{d-1}$ for every $t\geq 0$, along the same sequence $L\to\infty$:

$$\mu_t^L \rightharpoonup \mu_t, \qquad \mu_t(ds,dx) = ds\,\mu_{t,s}(dx). \tag{6}$$

Here $\mu_{t,s}$ is a conditional probability law on the sphere, defined for almost every $s\in(0,1]$.

The Lebesgue positional marginal follows from the uniform grid ℓ/L. The disintegration in equation 6 does not assert independence between position and embedding: the law $\mu_{t,s}$ may depend on s. The numerical investigation in Section 5.1 concerns this continuum description; convergence remains an explicit hypothesis of the theorem.

Theorem 3.2 (Conditional mean-field transport equation)**.** *Fix $\beta\geq 0$, and let V, T be measurable and locally bounded in time. Under Assumption 3.1, the limiting laws admit a jointly measurable disintegration and, for almost every $s\in(0,1]$, satisfy*

$$\begin{aligned}\partial_t\mu_{t,s} + \nabla_{\mathbb{S}}\cdot\big(\mu_{t,s}\,\mathsf{P}_x\big(V(t)m(s,t)\big)\big)\\ + \beta\,\nabla_{\mathbb{S}}\cdot\big(\mu_{t,s}\,\mathsf{P}_x\big(V(t)\Sigma(s,t)T(t)^\top x\big)\big) = 0\end{aligned} \tag{7}$$

in the distributional sense in (t,x). The prefix statistics are

$$\begin{aligned}\bar{x}(s,t) &:= \int_{\mathbb{S}^{d-1}} x\,\mu_{t,s}(dx),\\ m(s,t) &:= \frac{1}{s}\int_0^s \bar{x}(u,t)\,du,\\ \Sigma(s,t) &:= \frac{1}{s}\int_0^s\int_{\mathbb{S}^{d-1}} xx^\top\,\mu_{t,u}(dx)\,du - m(s,t)m(s,t)^\top.\end{aligned} \tag{8}$$

The precise weak formulation and proof are given in Theorem B.4; spherical divergence is defined in B.3.

The conditional mean $\bar{x}(s,t)$ and the prefix mean $m(s,t)$ are distinct. The two fluxes in equation 7 are the continuum counterparts of the discrete mean and covariance contributions. The theorem identifies the assumed limit; it does not by itself prove convergence or uniqueness of the nonlinear measure equation.

Monokinetic reduction. A monokinetic conditional law has the form $\mu_{t,s}=\delta_{x(s,t)}$. It describes a single direction at each token position, while allowing that direction to vary with s. It therefore retains positional variation and a possibly nonzero prefix covariance.

Theorem 3.3 (Monokinetic reduction)**.** *Under the hypotheses of Theorem 3.2, assume that $\mu_{t,s}$ is a Dirac mass for almost every (t,s). Then there is a representative $x(s,t)\in\mathbb{S}^{d-1}$ that is locally absolutely continuous in t for almost every s, and satisfies the nonlocal evolution equation*

$$\partial_t x(s,t) = \mathsf{P}_{x(s,t)}\big(V(t)m(s,t) + \beta V(t)\Sigma(s,t)T(t)^\top x(s,t)\big) \tag{9}$$

for almost every $s\in(0,1]$ and almost every t, where

$$m(s,t) = \frac{1}{s}\int_0^s x(u,t)\,du, \qquad \Sigma(s,t) = \frac{1}{s}\int_0^s \big(x(u,t)-m(s,t)\big)\big(x(u,t)-m(s,t)\big)^\top du.$$

No smoothness in token position is required for this formulation. The proof is given in Theorem B.5.

Equation equation 9 is the continuum model used below. Its routing term has magnitude at most $\beta\|V(t)\|\|T(t)\|\|\Sigma(s,t)\|$. A small covariance therefore makes this correction small, but does not alone imply that the mean drift dominates if that drift also vanishes.

4 Effects of Mean-Driven Alignment and Covariance-Driven Routing

We use logarithmic token depth to derive Fourier-mode estimates in which the eigenvalues, spectral gaps, and Jordan structure of V appear explicitly. The analysis concerns linear perturbations near a constant eigenprofile. Stationary clustered measures then provide a complementary description of the truncated model.

4.1 Log-Depth Transformation

Set

$$r = -\log s \in [0, \infty), \qquad X(r,t) = x(e^{-r}, t),$$

and define the exponential prefix operator on the half-line by

$$\mathcal{A}_+[X](r,t) = \int_0^\infty e^{-y} X(r+y,t)\, dy.$$

The substitution $u = e^{-(r+y)}$ gives

$$m(e^{-r},t) = \frac{1}{e^{-r}} \int_0^{e^{-r}} x(u,t)\, du = \mathcal{A}_+[X](r,t).$$

The prefix covariance becomes

$$\Sigma_r[X] = \mathcal{A}_+[XX^\top] - \mathcal{A}_+[X]\mathcal{A}_+[X]^\top.$$

Larger r corresponds to earlier token positions, so the use of $r+y$ follows the original causal prefix order.

Theorem 4.1 (Log-depth form)**.** *Let $X(r,t) = x(e^{-r},t)$, where x solves equation 9. Then, almost everywhere,*

$$\partial_t X = \mathsf{P}_X\big(V(t)\mathcal{A}_+[X] + \beta V(t)\Sigma_r[X]T(t)^\top X\big). \tag{10}$$

The change of variables and its proof are given in Proposition B.6.

4.2 Fourier Modes and the Value Spectrum

Assume that a fixed unit vector $\hat{e}$ satisfies $V(t)\hat{e} = \lambda(t)\hat{e}$ for almost every t. The constant profile $X \equiv \hat{e}$ is stationary. For a bounded tangent perturbation, use the spherical chart

$$X_\varepsilon(r,t) = \frac{\hat{e} + \varepsilon Y(r,t)}{|\hat{e} + \varepsilon Y(r,t)|}, \qquad \hat{e}\cdot Y(r,t) = 0.$$

Its covariance is $O(\varepsilon^2)$, so the routing term has no first variation. Proposition B.7 gives

$$\partial_t Y(r,t) = \mathsf{P}_{\hat{e}}\big(V(t)\mathcal{A}_+[Y](r,t)\big) - \lambda(t)Y(r,t).$$

Direct differentiation of the normalized attention weights gives the same linearization for exact softmax at every fixed finite β (Proposition B.8).

To expose the role of token depth, consider the Fourier harmonic

$$Y_\xi(r,t) = e^{i\xi r}\widehat{Y}(\xi,t), \qquad \xi \in \mathbb{R}, \qquad \hat{e}^\top \widehat{Y}(\xi,t) = 0.$$

Here $\widehat{Y} \in \mathbb{C}^d$ denotes the amplitude of this harmonic, not the integral Fourier transform of an arbitrary half-line profile; taking real parts gives real perturbations. For $\Phi(y) = e^y \mathbf{1}_{\{y\le 0\}}$,

$$\widehat{\Phi}(\xi) = \int_{-\infty}^0 e^y e^{-i\xi y}\, dy = \frac{1}{1-i\xi}.$$

The same multiplier acts directly on the half-line harmonic:

$$\mathcal{A}_+[Y_\xi](r,t) = e^{i\xi r}\widehat{Y}(\xi,t)\int_0^\infty e^{-y}e^{i\xi y}\, dy = \frac{e^{i\xi r}}{1-i\xi}\widehat{Y}(\xi,t).$$

Thus its amplitude satisfies

$$\partial_t \widehat{Y}(\xi,t) = \frac{1}{1-i\xi}\mathsf{P}_{\hat{e}}\big(V(t)\widehat{Y}(\xi,t)\big) - \lambda(t)\widehat{Y}(\xi,t). \tag{11}$$

The projection is extended complex linearly and tangency is preserved. This mode equation combines the causal multiplier with the projected value action. The following estimates hold for every real ξ and $t \geq 0$, with tangent initial amplitudes.

- **Symmetric value matrix (Theorem B.9).** Let V be time-independent and symmetric, with orthonormal eigenvectors satisfying

  $$Ve_k = \lambda_k e_k, \qquad \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d, \qquad \lambda_1 > 0.$$

  For $\hat{e} = \pm e_1$,

  $$|\widehat{Y}(\xi,t)| \leq \exp\left[-\left(\lambda_1 - \frac{\lambda_2}{1+\xi^2}\right)t\right]|\widehat{Y}(\xi,0)|.$$

- **Single Jordan block (Theorem B.10).** Let $V = J_d(\lambda) = \lambda I_d + N_d$, where $\lambda \in \mathbb{R}$ and N_d is the standard nilpotent Jordan shift. With $\hat{e} = \pm e_1$, where e_1 is the first standard basis vector,

  $$|\widehat{Y}(\xi,t)| \leq C_d(1+t^{d-2})\exp\left(-\lambda\frac{\xi^2}{1+\xi^2}t\right)|\widehat{Y}(\xi,0)|.$$

- **Time-dependent value matrix (Theorem B.11).** Assume a fixed real Jordan decomposition

  $$V(t) = S\,\mathrm{diag}\left(\lambda_{\max}(t), J_2(t), \ldots, J_M(t)\right)S^{-1}, \qquad J_b(t) = \lambda_b(t)I_{m_b} + N_b,$$

  where S is constant and invertible, $Se_1 = \hat{e}$ is unit, the N_b are fixed standard nilpotent shifts, and the eigenvalues are real-valued and locally integrable. The leading block has size one. Suppose

  $$\lambda_{\max}(t) - \lambda_b(t) \geq \gamma > 0, \qquad b = 2, \ldots, M,$$

  almost everywhere. Then

  $$|\widehat{Y}(\xi,t)| \leq C(1+t^{m_{\max}-1})\exp\left(-\frac{\xi^2\int_0^t \lambda_{\max}(\tau)\,d\tau + \gamma t}{1+\xi^2}\right)|\widehat{Y}(\xi,0)|,$$

  where $m_{\max} = \max_{b\geq 2} m_b$ and C depends only on d, S.

These estimates are the main spectral consequence of the token-depth formulation. In the symmetric case, the transverse eigencomponent associated with λ_k, $k \geq 2$, has exponent

$$\frac{\lambda_k}{1-i\xi} - \lambda_1 = -\left(\lambda_1 - \frac{\lambda_k}{1+\xi^2}\right) + i\frac{\lambda_k\xi}{1+\xi^2}.$$

The spectral gap controls low-frequency decay, while the dominant eigenvalue controls the high-frequency limit; the imaginary part describes phase evolution. When $\lambda_2 < 0$, high frequencies decay more slowly and the uniform rate is λ_1. The Jordan estimates expose the additional polynomial factors and time-integrated eigenvalues. These are linear mode estimates near a constant eigenprofile; they do not establish global nonlinear collapse or concentration of attention weights on particular tokens.

4.3 Covariance Structure and Stationary Clustered Measures

The covariance term is higher order near a constant profile, but can contribute away from that regime. For a given prefix with approximate cluster directions v_k and weights $\alpha_k > 0$, $\sum_k \alpha_k = 1$,

$$m(s,t) \approx \sum_k \alpha_k v_k, \qquad \Sigma(s,t) \approx \sum_k \alpha_k v_k v_k^\top - m(s,t)m(s,t)^\top.$$

Balancing the prefix mean suppresses $\mathsf{P}_x\big(Vm(s,t)\big)$ without necessarily making the covariance small. Nevertheless, nonzero covariance does not imply a nonzero routing velocity: $V\Sigma T^\top x$ may be radial and disappear under $\mathsf{P}_x(z)$. The following stationary constructions describe this geometry precisely. Throughout these constructions, V, T are fixed, and stationarity refers to the truncated transport equation equation 7.

- **Balanced isotropic mixtures (Theorem B.13).** Assume $VT^\top = I_d$. A position-independent family

$$\mu_s = \sum_{k=1}^{K} \alpha_k \delta_{v_k}, \qquad v_k \in \mathbb{S}^{d-1}, \quad \alpha_k > 0, \quad \sum_k \alpha_k = 1,$$

  is stationary for every $\beta \geq 0$ if

$$\sum_k \alpha_k v_k = 0, \qquad \sum_k \alpha_k v_k v_k^\top = \frac{1}{d} I_d.$$

  Indeed, $m = 0$, $\Sigma = I_d/d$, and $V\Sigma T^\top x = x/d$, so both projected drifts vanish.
- **Antipodal eigenvector mixtures (Theorem B.14).** Assume that $VT^\top$ is symmetric with orthonormal eigenbasis $v_1, \ldots, v_d$, and that $T^\top$ is diagonal in the same basis. For $1 \leq K \leq d$, $\alpha_k > 0$, and $\sum_{k=1}^{K} \alpha_k = 1$, the position-independent family

$$\mu_s = \sum_{k=1}^{K} \alpha_k \frac{\delta_{v_k} + \delta_{-v_k}}{2}$$

  is stationary for every $\beta \geq 0$. Antipodal symmetry gives zero prefix mean, and the shared eigenbasis makes the covariance drift radial at each support point.
- **Measures in an eigenspace at $\beta = 0$ (Corollary B.17).** Let $V_\lambda = \ker(V - \lambda I_d)$ have dimension at least two. Every measurable, time-independent family μ_s supported on $V_\lambda \cap \mathbb{S}^{d-1}$ and antipodally symmetric for almost every s is stationary. This is a special case of the zero-conditional-mean condition: $\int x \, \mu_s(dx) = 0$ for almost every s implies $m(s) = 0$. Eigenspace degeneracy is not needed for that stationary condition.

These families establish the existence of clustered equilibria, without asserting their formation or attraction from nearby data. In particular, their nonzero covariance is compatible with a vanishing projected routing field. They complement the Fourier-mode analysis: the latter quantifies the value-spectrum dependence of local linear response, while the stationary constructions describe other geometries allowed by the truncated equation.

5 Experimental Results

In this section, we provide empirical support for the theoretical framework developed above. Our experiments are organized around two objectives. First, we validate Assumption 3.1 by demonstrating that the mean-field limit holds empirically in modern attention-based architectures: as sequence length grows, the empirical distribution of token embeddings converges to a smooth depth-indexed measure. Second, we present representative examples illustrating the role of the prefix mean-to-covariance ratio in the stationary behavior of the system, as analyzed in Section 4, showing that covariance dominance promotes multi-cluster convergence.

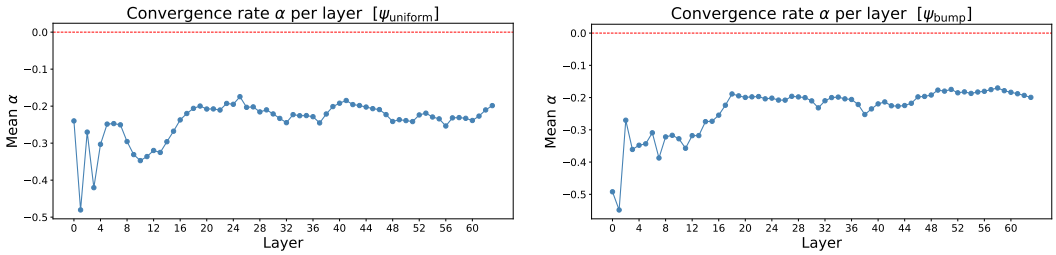
5.1 Empirical Validation of the Mean-Field Limit

We validate Assumption 3.1, which posits that the empirical measure $\mu^L = \frac{1}{L}\sum_{\ell=1}^{L} \delta_{(\ell/L,\, x_\ell)}$ converges weakly to a depth-indexed measure $\mu(ds, dx) = ds \, \mu_s(dx)$ on $[0,1] \times \mathbb{S}^{d-1}$ as $L \to \infty$.

Evaluation method. Since the limit μ is not available in closed form, we assess convergence via a Cauchy criterion: weak convergence implies that, for every bounded continuous test function φ and every depth-weighting function ψ, the gap

$$g(L) \;=\; \left| \langle \mu^L, \psi \otimes \varphi \rangle - \langle \mu^{2L}, \psi \otimes \varphi \rangle \right|$$

vanishes as $L \to \infty$, where $\langle \mu^L, \psi \otimes \varphi \rangle = \frac{1}{L}\sum_{\ell=1}^{L} \psi(\ell/L)\, \varphi(x_\ell)$. A power-law decay $g(L) \sim L^\alpha$ with $\alpha < 0$ provides quantitative evidence of convergence; we estimate α at each layer by linear regression of $\log g(L)$ against $\log L$.



(a) Uniform weighting $\psi_{\text{unif}}(s) = s$.

(b) Bump weighting ψ_{bump} supported on $(0.2,\ 0.8)$.

Figure 1: Mean log–log convergence rate α_j of the Cauchy gap $g(L, j, i)$ as a function of layer j, computed using shared weights from layer 20 of Llama-3.1-8B repeated across 64 layers.

Experimental setup. We use pretrained weights from Llama-3.1-8B, constructing a 64-layer architecture in which each layer reuses the parameters of the 20-th layer to simulate repeated application of a single causal attention block. We probe the limit with two complementary depth-weightings: the uniform weighting $\psi_{\text{unif}}(s) = s$, which captures the global depth profile, and a smooth compactly supported bump $\psi_{\text{bump}}(s) = \exp\big(-\frac{1}{(s-0.2)(0.8-s)}\big)\mathbf{1}_{s\in(0.2,\,0.8)}$, which isolates a strict interior subinterval and rules out boundary-driven convergence. We pair these with four bounded continuous test functions φ_i that probe individual coordinate blocks and their cross-interactions; full definitions and per-sample regression details are deferred to Appendix D.1.

Results. Figure 1 reports the mean of α_j across test functions and corpus samples at each layer, for both depth weightings. Under both ψ_{unif} and ψ_{bump}, we observe $\alpha_j < 0$ at every layer, with values stabilizing in the range $[-0.35, -0.20]$ after a brief transient at shallow layers. The agreement between the two weightings indicates that the convergence is intrinsic to the depth-indexed measure. Together, these results provide empirical support for Assumption 3.1.

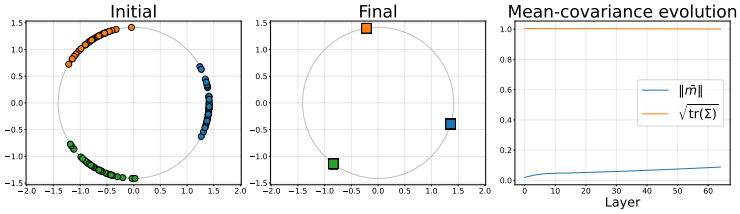
5.2 Mean–Covariance Competition Drives Cluster Structure

The main finding of the analysis in Section 4 is that the interplay between the prefix mean $m(s,t)$ and the prefix covariance $\Sigma(s,t)$ governs the stationary state of the attention dynamics. Specifically, the prefix mean drives collapse toward a *single cluster*, whereas the prefix covariance promotes the formation of *multiple clusters*. To validate this mechanism, we run a 64-layer model with shared weights, where each layer is a causal self-attention module followed by RMSNorm. We set $T = Q^\top K = I$ so that $a_{ij} \propto \beta\langle x_i, x_j\rangle$, and fix $\beta = 0.5$. We consider three initializations on $\mathbb{S}^1$ designed to isolate the roles of $m(s,t)$ and $\Sigma(s,t)$.

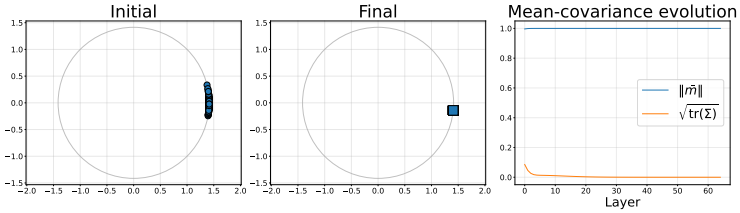
Covariance-dominated (multi-cluster). We place tokens in $K = 3$ balanced clusters at $v_k = (\cos\frac{2\pi k}{3}, \sin\frac{2\pi k}{3})$ with equal mass, so $\sum_k \alpha_k v_k = 0$ and $\Sigma = \mathcal{O}(1)$. Fig. 2a shows tokens converging to exactly three points at the cluster centers, with $\|\bar{m}\| \approx 0$ and $\sqrt{\operatorname{tr}(\bar{\Sigma})} \approx 1$ across all layers. The vanishing prefix mean kills the alignment term $P_x(Vm)$, while the routing term $P_x(V\Sigma T^\top x)$ contracts each cluster onto its center v_k and then vanishes there, leaving the three-cluster state as a fixed point. This behavior matches Theorems B.13–B.14.

Mean-dominated (single-cluster collapse). We initialize all tokens in a tight angular window around $(1, 0)$, an eigendirection of V, so $\|\bar{m}\| \approx 1$ and $\Sigma = \mathcal{O}(\varepsilon^2)$. Fig. 2b shows immediate collapse to a single point: $\sqrt{\operatorname{tr}(\bar{\Sigma})}$ drops to near zero within a few layers while $\|\bar{m}\|$ stays saturated at 1. The covariance term enters at order $\beta\varepsilon^2$ and cannot offset the spectral-gap contraction at order ε predicted by Theorems B.9–B.11, producing the attention-sink regime.

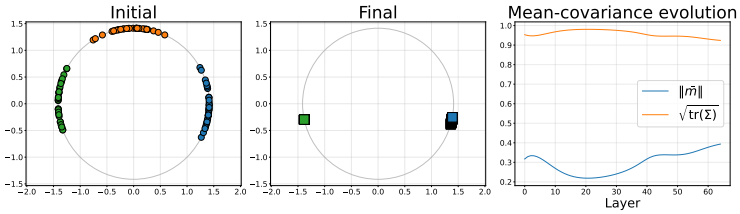
Mixed regime (partial collapse). We use three clusters with non-equispaced angles, so $\sum_k \alpha_k v_k \neq 0$ while Σ stays $\mathcal{O}(1)$. Fig. 2c shows the orange cluster merging into the blue one, leaving exactly two clusters; throughout layer, $\|\bar{m}\| \in [0.22, 0.39]$ and $\sqrt{\operatorname{tr}(\bar{\Sigma})} \approx 0.95$. The non-zero mean activates spectral alignment along the leading eigendirection of V, reducing the cluster count, but the remaining configuration self-organizes into an antipodal pair, preventing full collapse.



(a) Covariance-driven: tokens converge to multiple clusters.



(b) Mean-driven: tokens collapse to a single cluster.



(c) Mixed: partial collapse under competing mean and covariance effects.

Figure 2: Representative regimes of causal attention dynamics on $\mathbb{S}^1$. Each row shows, from left to right, the initial token configuration, the final token configuration after 64 layers, and the layerwise evolution of the prefix-mean norm $\|\bar{m}\|$ and the prefix-covariance trace $\sqrt{\mathrm{tr}(\bar{\Sigma})}$. Colors denote distinct initial clusters.

The three regimes confirm that the stationary geometry of causal attention is set by the competition between $m(s,t)$ and $\Sigma(s,t)$: balanced cluster geometry preserves diversity, mean-aligned configurations collapse, and intermediate cases yield reduced but nontrivial cluster counts. The layerwise ratio $\|\bar{m}\|/\sqrt{\mathrm{tr}(\bar{\Sigma})}$ thus serves as a concrete diagnostic for whether a Transformer layer behaves as a sink-forming alignment layer or a diversity-preserving routing layer.

6 Conclusion and Limitations

We developed a dynamical and mean-field framework for explaining collapse and multi-cluster formation in causal attention dynamics. By expanding the causal softmax interaction to first order, we decomposed the dynamics into a mean-driven alignment term and a covariance-driven routing term. We then derived a depth-aware mean-field transport equation by incorporating token position into the empirical measure. In the monokinetic regime, a log-depth transformation revealed the causal prefix operator as an exponential memory filter, enabling Fourier analysis in the depth variable. Our results show that mean-driven alignment induces spectral-gap contraction near dominant eigendirections of the value matrix, leading to collapse and attention-sink formation. In contrast, covariance-driven routing can sustain multi-cluster structure when the prefix mean is balanced while the prefix covariance remains nontrivial. Overall, causal attention dynamics emerge as a competition between mean-driven spectral alignment and covariance-driven routing.

Limitation. Our analysis focuses on a simplified single-head causal attention flow and relies on a first-order small-β expansion. Extending the framework to full Transformer blocks, higher-order softmax effects, and general distributional regimes remains an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing. We have not used generative AI tools for generating complete proofs or generating entire sections of the manuscript. We have reviewed all AI-assisted work. All AI-assisted text was carefully reviewed and revised by the authors to ensure accuracy, clarity, and consistency with the intended mathematical content. All mathematical arguments, proofs, and claims remain the responsibility of the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Considering the scope and focus of this work, we do not anticipate any adverse societal or ethical consequences.

REPRODUCIBILITY STATEMENT

The source code for all experiments is included in the paper's supplementary materials. Detailed descriptions of our experimental setup can be found in Section 5 and Appendix D. All datasets used are publicly accessible via an anonymous link provided in the README of the supplementary materials.

REFERENCES

Álvaro Rodríguez Abella, João Pedro Silvestre, and Paulo Tabuada. Consensus is all you get: The role of attention in transformers. In *Forty-second International Conference on Machine Learning*, 2025.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3159–3166, 2019.

Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByxZX20qFQ.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? *arXiv preprint arXiv:2312.14820*, 2023.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Jingpu Cheng, Qianxiao Li, Ting Lin, and Zuowei Shen. Interpolation, approximation and controllability of deep neural networks. *arXiv preprint arXiv:2309.06015*, 2023.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive kv cache compression for llms. *arXiv preprint arXiv:2310.01801*, 2023.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=aMjaEkkXJx.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *arXiv preprint arXiv:2312.10794*, 2023b.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36, 2024.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34 (1):014004, 2017.

Hakan Inan, Khashayar Khosravi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling. *arXiv preprint arXiv:1611.01462*, 2016.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021.

Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. *Advances in Neural Information Processing Systems*, 37:115652–115681, 2024.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Samory Kpotufe and Abdeslam Boularias. Gradient weights help nonparametric regressors. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Qianxiao Li, Ting Lin, and Zuowei Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 25(5):1671–1709, 2022.

Hongzhou Lin and Stefanie Jegelka. Resnet with one-neuron hidden layers is a universal approximator. *Advances in neural information processing systems*, 31, 2018.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI open*, 3: 111–132, 2022.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJC_jUqxe.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364, 2023.

Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *International conference on machine learning*, pp. 3276–3285. PMLR, 2018.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1412–1421, 2015.

Tan Nguyen, Minh Pham, Tam Nguyen, Khai Nguyen, Stanley Osher, and Nhat Ho. Fourierformer: Transformer meets generalized fourier integral theorem. *Advances in Neural Information Processing Systems*, 35:29319–29335, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Jorge Pérez, Pablo Barceló, and Javier Marinkovic. Attention is turing-complete. *Journal of Machine Learning Research*, 22(75):1–35, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Philippe Rigollet. The mean-field dynamics of transformers. *arXiv preprint arXiv:2512.01868*, 2025.

Domenec Ruiz-Balet and Enrique Zuazua. Neural ode control for classification, approximation, and transport. *SIAM Review*, 65(3):735–773, 2023.

Lars Ruthotto and Eldad Haber. Deep neural networks motivated by partial differential equations. *Journal of Mathematical Imaging and Vision*, 62(3):352–364, 2020.

Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 3515–3530. PMLR, 2022.

Paulo Tabuada and Bahman Gharesifard. Universal approximation power of deep residual neural networks through the lens of control. *IEEE Transactions on Automatic Control*, 68(5):2715–2728, 2022.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

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

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

E Weinan, Yali Duan, Linghua Kong, and Min Guo. A proposal on machine learning via dynamical systems. 2024:08–27, 2017.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Han Zhang, Xi Gao, Jacob Unterman, and Tom Arodz. Approximation capabilities of neural odes and invertible residual networks. In *International Conference on Machine Learning*, pp. 11086–11095. PMLR, 2020.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 2023.

Appendix of "Mean–Covariance Competition Governs Collapse and Multi-Cluster States in Causal Attention Dynamics"

Table of Contents

A Related Works **14**

B Main Theory **15**

B.1 Model and notation 15

B.2 Mean- and covariance-driven decomposition 16

B.3 A conditional mean-field limit 17

B.4 The monokinetic regime 19

B.5 Logarithmic token depth and exact linear response 20

B.6 Fourier modes in logarithmic token depth 22

B.7 Spectral dynamics of logarithmic token-depth modes 23

B.8 Stationary Profiles and Clustered Measures 25

C Conclusion **27**

D Further Experimental Details **27**

D.1 Empirical Validation of the Mean-Field Limit 27

A Related Works

Attention mechanisms and Transformer theory. The Transformer architecture Vaswani et al. (2017) builds on earlier attention mechanisms for neural machine translation and representation learning Bahdanau et al. (2014); Luong et al. (2015); Lin et al. (2017). Theoretical studies have established expressivity and computational properties of Transformers Yun et al. (2019); Pérez et al. (2021), and have examined in-context learning and algorithmic behavior in their forward passes Garg et al. (2022); Akyürek et al. (2022); Von Oswald et al. (2023); Olsson et al. (2022). Our focus is complementary: we study the local dynamics of token representations under a continuous-time model of causal attention, with particular attention to the interaction between token position and the spectrum of the value matrix.

Residual networks and continuous-time models. The propagation of representations through a deep network can be viewed as a discrete dynamical process (LeCun et al., 2015). Connections between neural networks, differential equations, and gradient-based dynamics motivate continuous-time descriptions (Chen et al., 2018; Haber & Ruthotto, 2017; Kpotufe & Boularias, 2012). Residual and continuous-depth architectures provide settings in which to study such descriptions (Lin & Jegelka, 2018; Zhang et al., 2020; Li et al., 2022; Nguyen et al., 2022; Ruiz-Balet & Zuazua, 2023; Cheng et al., 2023). For example, a residual update $z = x + h\, y(x, \theta)$ has the form of an Euler step for an ordinary differential equation. Approximation, interpolation, and controllability questions for these models have received substantial attention (Lin & Jegelka, 2018; Zhang et al., 2020; Li et al., 2022; Tabuada & Gharesifard, 2022; Ruiz-Balet & Zuazua, 2023; Cheng et al., 2023). In the present work, continuous time describes evolution across layers, whereas the separate token-depth variable describes position within the sequence.

Dynamical and mean-field analyses of attention. Continuous-depth formulations connect neural networks and residual architectures to differential equations Weinan et al. (2017); Chen et al. (2018); Lu et al. (2018); Haber & Ruthotto (2017); Ruthotto & Haber (2020). For attention, interacting-particle and mean-field approaches have been used to study clustering, long-time dynamics, and

continuum descriptions Geshkovski et al. (2023b; 2024); Rigollet (2025); Castin et al. (2023); Karagodin et al. (2024). In particular, Castin, Ablin, and Peyré Castin et al. (2023) introduce a mean-field framework for masked attention that augments each token with its position, representing a sequence by a measure of the form $n^{-1}\sum_i \delta_{(s_i,x_i)}$. Their causal attention operator already retains the prefix structure. Rigollet Rigollet (2025) develops a mean-field perspective on Transformer dynamics. We build on this position-aware viewpoint to derive a transport equation for the first-order small-inverse-temperature model and to analyze its prefix operator after a logarithmic change of token depth.

Small-inverse-temperature expansions and simplified models. Structural simplifications have been studied in several related settings, including embedding losses, normalized attention, and dynamical or mean-field models Inan et al. (2016); Sander et al. (2022); Geshkovski et al. (2023b); Castin et al. (2023); Rigollet (2025). Here β denotes inverse temperature, so the small-β expansion is a high-temperature expansion around uniform prefix averaging. Its first-order term decomposes the attention drift into a contribution from the prefix mean and a contribution involving the prefix covariance and query–key matrix. We treat the resulting truncated dynamics as a specified model and distinguish its nonlinear conclusions from statements about exact softmax attention. The linearization about a constant eigenprofile, however, agrees with that of exact softmax for every fixed finite β.

Clustering and attention sinks in causal attention. Clustering and long-time behavior of attention dynamics have been investigated in several settings Geshkovski et al. (2023b; 2024); Karagodin et al. (2024); Rigollet (2025). Karagodin, Polyanskiy, and Rigollet Karagodin et al. (2024) prove almost-everywhere convergence to one cluster for exact finite-token causal attention with $V = I$, arbitrary query and key matrices, and fixed inverse temperature. Under their respective spectral hypotheses, their conjectures distinguish a simple dominant eigendirection, associated with the two antipodal directions, from a dominant eigenspace of dimension at least two, associated with one-cluster convergence. Our estimates concern the linearization near a constant eigenprofile and do not establish these global convergence conjectures. Separately, empirical studies of long-context inference examine attention sinks and the allocation or retention of attention across tokens Xiao et al. (2023); Zhang et al. (2023); Liu et al. (2023; 2024); Ge et al. (2023). Representation alignment and preferential attention to initial tokens are different properties: complete alignment makes the logits within each causal row equal. We therefore interpret our results as statements about representation dynamics, without identifying local alignment with an attention-sink mechanism.

Our contribution relative to prior work. Our central contribution is a Fourier-mode analysis of token-depth dynamics that makes the role of the value spectrum explicit. Starting from the position-aware transport equation, the monokinetic reduction and the change of variables $r = -\log s$ turn causal prefix averaging into an exponential operator with multiplier $(1 - i\xi)^{-1}$. Near a constant eigenprofile, this yields mode estimates in which the eigenvalues and spectral gaps of V determine decay, while its Jordan blocks generate polynomial factors. The covariance term has no first variation there, and direct differentiation shows that exact softmax has the same linearization at every fixed finite inverse temperature. We also construct stationary families for the truncated transport equation; their formation and nonlinear stability are separate questions.

B Main Theory

B.1 Model and notation

We distinguish two variables throughout: $t \geq 0$ is continuous layer depth (the evolution variable), whereas $s \in (0, 1]$ is normalized token position. We occasionally call s token depth, but it is not the evolution variable. Fix an embedding dimension $d \geq 2$ and a sequence length $L \geq 1$. The token representations are $x_\ell(t) \in \mathbb{S}^{d-1}$, $1 \leq \ell \leq L$. All vector norms are Euclidean, and matrix norms are the associated operator norms. Let $Q(t), K(t), V(t) \in \mathbb{R}^{d\times d}$ and set $T(t) = Q(t)^\top K(t)$. Any fixed score normalization is absorbed into T. The nonlinear model assumes that V and T are measurable and locally bounded in t. The later linear evolution results explicitly allow locally integrable coefficients where stated.

For $x \in \mathbb{S}^{d-1}$, define the orthogonal tangent projection by

$$\mathsf{P}_x(z) := z - (x \cdot z)x. \tag{12}$$

For inverse temperature $\beta \geq 0$, the causal scores and attention weights are

$$s_{\ell j} = x_\ell^\top T x_j, \qquad a_{\ell j}(\beta) = \frac{\exp(\beta s_{\ell j})}{\sum_{i=1}^{\ell} \exp(\beta s_{\ell i})}, \qquad 1 \leq j \leq \ell. \tag{13}$$

Time arguments are suppressed when no ambiguity results. We study the continuous-depth, spherical causal-attention model

$$\partial_t x_\ell(t) = \mathsf{P}_{x_\ell(t)} \left(V(t) \sum_{j=1}^{\ell} a_{\ell j}(t) x_j(t) \right). \tag{14}$$

The projection makes the vector field tangent to the sphere and therefore preserves $|x_\ell| = 1$. With the stated coefficient assumptions, this is a Carathéodory ordinary differential equation: its solutions are locally absolutely continuous in t, and the equation holds almost everywhere in t. For each finite L, smooth dependence of the vector field on the token configuration and its local time bounds give a unique global solution for every spherical initial configuration.

B.2 Mean- and covariance-driven decomposition

Define the prefix mean, second moment, and covariance by

$$\bar{x}_\ell = \frac{1}{\ell} \sum_{j=1}^{\ell} x_j, \qquad M_\ell = \frac{1}{\ell} \sum_{j=1}^{\ell} x_j x_j^\top, \qquad \Sigma_\ell = M_\ell - \bar{x}_\ell \bar{x}_\ell^\top. \tag{15}$$

The small-β regime is a high-temperature regime. Its first correction to uniform attention can be expressed exactly in terms of these prefix statistics.

Lemma B.1 (Softmax expansion)**.** *Fix a configuration, a query index ℓ, and real scores $s_{\ell j}$ with $|s_{\ell j}| \leq M$. Set*

$$\bar{s}_\ell = \frac{1}{\ell} \sum_{j=1}^{\ell} s_{\ell j}, \qquad \delta_{\ell j} = s_{\ell j} - \bar{s}_\ell, \qquad v_\ell = \frac{1}{\ell} \sum_{j=1}^{\ell} \delta_{\ell j}^2.$$

For every $b \geq 0$ and $|\beta| \leq b$,

$$a_{\ell j}(\beta) = \frac{1}{\ell} + \frac{\beta}{\ell} \delta_{\ell j} + \frac{\beta^2}{2\ell} (\delta_{\ell j}^2 - v_\ell) + R_{\ell j}(\beta), \qquad |R_{\ell j}(\beta)| \leq \frac{8}{3\ell} e^{2bM} M^3 |\beta|^3. \tag{16}$$

In particular, the remainder is uniform in the query index, the key index, and the sequence length whenever the same score bound applies.

Proof. For this proof, write $p_j(\theta) = a_{\ell j}(\theta)$, $\bar{s}(\theta) = \sum_j p_j(\theta) s_{\ell j}$, and $v(\theta) = \sum_j p_j(\theta)(s_{\ell j} - \bar{s}(\theta))^2$. Differentiating the finite softmax sums gives

$$p_j' = p_j(s_{\ell j} - \bar{s}), \qquad p_j'' = p_j\big((s_{\ell j} - \bar{s})^2 - v\big),$$

and

$$p_j''' = p_j\big((s_{\ell j} - \bar{s})^3 - 3(s_{\ell j} - \bar{s})v - \kappa_3\big), \qquad \kappa_3 = \sum_i p_i (s_{\ell i} - \bar{s})^3.$$

The score bound implies $|s_{\ell j} - \bar{s}| \leq 2M$, $v \leq M^2$, and $|\kappa_3| \leq 2Mv \leq 2M^3$. Also $p_j(\theta) \leq e^{2bM}/\ell$ for $|\theta| \leq b$. Consequently,

$$|p_j'''(\theta)| \leq \frac{16}{\ell} e^{2bM} M^3.$$

Taylor's theorem at $\theta = 0$, where $p_j(0) = 1/\ell$, now gives equation 16 and the stated remainder bound. □

Lemma B.2 (Covariance identity)**.** *For the scores in equation 13,*

$$\frac{1}{\ell} \sum_{j=1}^{\ell} (s_{\ell j} - \bar{s}_\ell) x_j = \Sigma_\ell T^\top x_\ell. \tag{17}$$

Proof. Put $w = T^\top x_\ell$. Then $s_{\ell j} = w \cdot x_j$, $\bar{s}_\ell = w \cdot \bar{x}_\ell$, and hence

$$\frac{1}{\ell}\sum_j (s_{\ell j} - \bar{s}_\ell) x_j = \left(\frac{1}{\ell}\sum_j x_j x_j^\top - \bar{x}_\ell \bar{x}_\ell^\top\right) w = \Sigma_\ell T^\top x_\ell.$$

□

Proposition B.3. *At every spherical token configuration and every time for which V, T are defined, the exact vector field in equation 14 satisfies*

$$\partial_t x_\ell = \mathsf{P}_{x_\ell}(V\bar{x}_\ell) + \beta \mathsf{P}_{x_\ell}(V\Sigma_\ell T^\top x_\ell) + \mathcal{R}_\ell, \qquad |\mathcal{R}_\ell| \le \beta^2 \|V\|\|T\|^2.$$

The bound is independent of ℓ and L.

Proof. Fix the configuration and define $F(\theta) = \sum_{j=1}^{\ell} a_{\ell j}(\theta) x_j$. Let $\mathbb{E}_\theta$ denote expectation over the finite set of keys with these attention weights, and let S denote the corresponding score. Direct differentiation yields

$$F'(0) = \Sigma_\ell T^\top x_\ell, \qquad F''(\theta) = \mathbb{E}_\theta\left[(S - \mathbb{E}_\theta S)^2 (x - F(\theta))\right].$$

Since $|x - F(\theta)| \le 2$ and $|S| \le \|T\|$,

$$|F''(\theta)| \le 2\,\mathrm{Var}_\theta(S) \le 2\|T\|^2.$$

Taylor's formula with integral remainder therefore gives $|F(\beta) - F(0) - \beta F'(0)| \le \beta^2 \|T\|^2$. Apply the linear map $z \mapsto \mathsf{P}_{x_\ell}(Vz)$ and use $|\mathsf{P}_{x_\ell}(z)| \le |z|$. □

We now define the first-order truncated system by

$$\partial_t x_\ell = \mathsf{P}_{x_\ell}\left(V(t)\bar{x}_\ell + \beta V(t)\Sigma_\ell T(t)^\top x_\ell\right). \tag{18}$$

Unless stated otherwise, the continuum results below concern this system. The covariance term is the first query–key-dependent correction to prefix averaging. Its tangential effect depends on V, T, and the token configuration; it need not oppose alignment. Proposition B.3 compares vector fields at the same configuration. A trajectory comparison requires an additional stability estimate, and the proposition does not give a uniform approximation over arbitrarily long layer depth.

B.3 A CONDITIONAL MEAN-FIELD LIMIT

For a solution of equation 18, set $s_\ell = \ell/L$ and define the joint empirical measure

$$\mu_t^L = \frac{1}{L}\sum_{\ell=1}^{L} \delta_{(s_\ell, x_\ell(t))} \quad \text{on } [0,1] \times \mathbb{S}^{d-1}.$$

The inclusion of token position retains the causal prefix information. Its positional marginal is the uniform grid measure and converges to Lebesgue measure on $[0,1]$. Thus any weak probability limit has a disintegration

$$\mu_t^L \rightharpoonup \mu_t, \qquad \mu_t(ds, dx) = ds\, \mu_{t,s}(dx), \tag{19}$$

where $\mu_{t,s}$ is a probability measure on the sphere for almost every token position s. This conditional law describes the distribution obtained by first taking the continuum limit and then shrinking positional neighborhoods around almost every s; no pointwise convergence of individual tokens is assumed.

Spherical weak formulation. For a smooth scalar function ψ on the sphere, $\nabla_{\mathbb{S}}\psi(x) = \mathsf{P}_x\big(\nabla\widetilde{\psi}(x)\big)$, where $\widetilde{\psi}$ is any smooth extension to a neighborhood of the sphere. If $b(x)$ is tangential and ν is a finite measure, define the distributional divergence by

$$\langle \nabla_{\mathbb{S}} \cdot (\nu b), \psi\rangle = -\int_{\mathbb{S}^{d-1}} \nabla_{\mathbb{S}}\psi(x) \cdot b(x)\, \nu(dx).$$

In particular, no density of ν is required.

Theorem B.4. *Fix $\beta \geq 0$ and measurable, locally bounded matrices V, T. For each L, let the spherical trajectories solve equation 18. Assume that μ_t^L converges weakly as a probability measure on $[0,1] \times \mathbb{S}^{d-1}$ for every $t \geq 0$. Then the limiting measures have the form equation 19, with a jointly measurable disintegration. Define, for $s > 0$,*

$$\begin{aligned} \bar{x}(s,t) &= \int_{\mathbb{S}^{d-1}} x \, \mu_{t,s}(dx), \\ m(s,t) &= \frac{1}{s} \int_0^s \bar{x}(u,t) \, du, \\ \Sigma(s,t) &= \frac{1}{s} \int_0^s \int_{\mathbb{S}^{d-1}} xx^\top \, \mu_{t,u}(dx) \, du - m(s,t)m(s,t)^\top. \end{aligned} \tag{20}$$

For almost every $s \in (0,1]$, the conditional laws satisfy

$$\partial_t \mu_{t,s} + \nabla_{\mathbb{S}} \cdot \big(\mu_{t,s} \mathsf{P}_x \big(V(t) m(s,t) + \beta V(t) \Sigma(s,t) T(t)^\top x \big) \big) = 0 \tag{21}$$

in the distributional sense in $(t,x) \in (0,\infty) \times \mathbb{S}^{d-1}$. The conditional laws admit representatives for which, for almost every s, integration against each smooth spherical test function is locally absolutely continuous in t.

More explicitly, for every $\varphi \in C_c^\infty((0,1] \times \mathbb{R}^d)$ and every $t \geq 0$,

$$\begin{aligned} \langle \mu_t, \varphi \rangle - \langle \mu_0, \varphi \rangle = &\int_0^t \int_0^1 \int_{\mathbb{S}^{d-1}} \nabla_x \varphi(s,x) \cdot \mathsf{P}_x \big(V(\tau) m(s,\tau) \big) \, \mu_{\tau,s}(dx) \, ds \, d\tau \\ &+ \beta \int_0^t \int_0^1 \int_{\mathbb{S}^{d-1}} \nabla_x \varphi(s,x) \cdot \mathsf{P}_x \big(V(\tau) \Sigma(s,\tau) T(\tau)^\top x \big) \, \mu_{\tau,s}(dx) \, ds \, d\tau. \end{aligned} \tag{22}$$

Here smoothness on $(0,1]$ means smoothness up to the endpoint 1; compact support keeps the test function away from $s = 0$.

Proof. The positional marginal of μ_t is Lebesgue measure, because this is the weak limit of $L^{-1} \sum_\ell \delta_{\ell/L}$. Disintegration therefore gives equation 19. Every prefix covariance is positive semidefinite and satisfies

$$\operatorname{tr} \Sigma_\ell = 1 - |\bar{x}_\ell|^2 \leq 1.$$

The same identity holds for the covariance of the limiting prefix law. Consequently, both discrete and limiting means have norm at most 1, and both covariances have operator norm at most 1. In particular,

$$|\partial_t x_\ell(t)| \leq c(t) := \|V(t)\| \big(1 + \beta \|T(t)\| \big) \tag{23}$$

almost everywhere, uniformly in ℓ, L. The function c is locally integrable. This bound, first for Lipschitz test functions and then by uniform approximation for continuous test functions, shows that $t \mapsto \mu_t$ is weakly continuous: for every bounded continuous test function g, the map $t \mapsto \langle \mu_t, g \rangle$ is continuous. In particular, $t \mapsto \mu_t$ is measurable, and a jointly measurable conditional probability kernel may be chosen.

We first justify convergence of the prefix moments at a fixed time. For a bounded continuous vector- or matrix-valued function f on the sphere, define

$$G_L^f(s) = \frac{1}{L} \sum_{j/L \leq s} f(x_j(t)), \qquad G^f(s) = \int_0^s \int_{\mathbb{S}^{d-1}} f(x) \, \mu_{t,u}(dx) \, du.$$

The limiting measure gives zero mass to every positional slice. Weak convergence therefore implies $G_L^f(s) \to G^f(s)$ for each $s \in [0,1]$. Moreover, for $0 \leq a \leq b \leq 1$,

$$|G_L^f(b) - G_L^f(a)| \leq \|f\|_\infty \big(b - a + L^{-1} \big), \qquad |G^f(b) - G^f(a)| \leq \|f\|_\infty (b - a).$$

A finite mesh argument upgrades pointwise convergence to uniform convergence on $[0,1]$. Applying this to $f(x) = x$ and $f(x) = xx^\top$ shows that the functions

$$m_L(s,t) = \frac{G_L^x(s)}{s}, \qquad \Sigma_L(s,t) = \frac{G_L^{xx^\top}(s)}{s} - m_L(s,t) m_L(s,t)^\top$$

converge uniformly to $m(s,t), \Sigma(s,t)$ on every interval $[\delta, 1]$, $\delta > 0$. At the grid points $s = s_\ell$ these functions equal $\bar{x}_\ell$ and Σ_ℓ exactly. For each fixed time, the limiting prefix functions are continuous on $(0, 1]$ and locally Lipschitz there.

For a test function φ as in the theorem, the chain rule for absolutely continuous trajectories gives

$$\begin{aligned}\langle \mu_t^L, \varphi\rangle - \langle \mu_0^L, \varphi\rangle = \int_0^t \frac{1}{L}\sum_{\ell=1}^{L} \nabla_x \varphi(s_\ell, x_\ell(\tau))\cdot \\ \mathsf{P}_{x_\ell(\tau)}\left(V(\tau)\bar{x}_\ell(\tau) + \beta V(\tau)\Sigma_\ell(\tau)T(\tau)^\top x_\ell(\tau)\right)\, d\tau.\end{aligned} \tag{24}$$

Choose $\delta > 0$ below the positional support of φ. At each fixed time, uniform convergence of the prefix statistics on $[\delta, 1]$, followed by weak convergence of the empirical measure, gives convergence of the integrand to the flux in equation 22. Here the limiting test integrand is continuous in (s, x) on the support of φ. Its discrete counterpart is bounded by $\|\nabla_x \varphi\|_\infty c(\tau)$, independently of L. Dominated convergence in time proves equation 22 for every t. It also proves local absolute continuity of $t \mapsto \langle \mu_t, \varphi\rangle$.

Finally, apply this identity to separated tests $\varphi(s, x) = \eta(s)\psi(x)$ and then to smooth time tests. Fubini's theorem and the arbitrariness of η give equation 21 for almost every s. For a fixed smooth ψ, its conditional weak time derivative is the corresponding flux, bounded by $\|\nabla_{\mathbb{S}}\psi\|_\infty c(t)$. Thus it has a locally absolutely continuous representative, with initial value $\int \psi\, d\mu_{0,s}$. Apply the conditional identities first at rational times to a countable set of smooth spherical test functions dense in C^1. Removing one null set of positions makes all these identities hold simultaneously. The flux bound extends them to every time. At rational times the conditional laws are probability measures; compactness of the sphere and density of the test set give weakly continuous probability-valued representatives for almost every s. For every smooth separated test, their integrals agree with those of μ_t for almost every t and are continuous in t on both sides. They therefore agree for every t, so these representatives still disintegrate μ_t and satisfy equation 21. This establishes the asserted temporal regularity. □

The theorem identifies every pointwise-in-time weak limit satisfying its hypotheses. It does not assert convergence of the full particle sequence from general initial data, uniqueness of the nonlinear measure equation, or an approximation of the exact softmax dynamics at arbitrarily large t.

B.4 The monokinetic regime

A monokinetic conditional law has the form $\mu_{t,s} = \delta_{x(s,t)}$. This means that the limiting conditional distribution at a given token position is a single mass. The direction may nevertheless vary with s; in particular, this assumption is distinct from a global collapse to one direction. For example, the profile $x(s) = (\cos(\pi s), \sin(\pi s), 0, \ldots, 0)$ is monokinetic and has nontrivial positional variation. Nor does weak convergence to a monokinetic law imply uniform approximation of every individual token by this profile.

Theorem B.5 (Monokinetic reduction)**.** *Under the hypotheses and conclusions of Theorem B.4, assume that $\mu_{t,s}$ is a Dirac mass for almost every (t, s). Then its representing profile admits a version such that, for almost every s, $t \mapsto x(s,t)$ is locally absolutely continuous, takes values in $\mathbb{S}^{d-1}$ for every t, and satisfies*

$$\partial_t x(s,t) = \mathsf{P}_{x(s,t)}\left(V(t)m(s,t) + \beta V(t)\Sigma(s,t)T(t)^\top x(s,t)\right) \tag{25}$$

for almost every t, where

$$\begin{aligned} m(s,t) &= \frac{1}{s}\int_0^s x(u,t)\, du, \\ \Sigma(s,t) &= \frac{1}{s}\int_0^s \big(x(u,t) - m(s,t)\big)\big(x(u,t) - m(s,t)\big)^\top du. \end{aligned} \tag{26}$$

No spatial smoothness of the profile is needed for this almost everywhere formulation.

Proof. For almost every s, the conditional equation tested against each coordinate function shows that

$$q(s,t) := \int_{\mathbb{S}^{d-1}} x\, \mu_{t,s}(dx)$$

has a locally absolutely continuous representative. At almost every time the conditional law is a Dirac mass, so $q(s,t) = x(s,t)$ and $|q(s,t)| = 1$ there. Continuity implies $|q(s,t)| = 1$ for every time. Using the narrowly continuous conditional representative in Theorem B.4, the law is then $\delta_{q(s,t)}$ at every time. One way to see this is that a probability measure on the unit sphere whose mean has norm 1 must be concentrated at that mean. Set $x = q$.

The conditional first and second moments are therefore $x(s,t)$ and $x(s,t)x(s,t)^\top$. Substituting these into equation 20 gives equation 26; expanding its centered integrand yields the equivalent expression

$$\Sigma(s,t) = \frac{1}{s}\int_0^s x(u,t)x(u,t)^\top\,du - m(s,t)m(s,t)^\top.$$

The conditional equation tested against the coordinate functions now gives equation 25 almost everywhere in time, for almost every position. □

Although the conditional law is a Dirac mass, the prefix covariance in equation 26 generally remains nonzero: it measures variation across earlier token positions. It vanishes when the profile is constant almost everywhere on that prefix. Thus the monokinetic reduction retains both the prefix-mean term and the covariance-dependent correction without asserting a direction of their nonlinear competition.

B.5 Logarithmic token depth and exact linear response

Here $s \in (0,1]$ denotes normalized token position and t denotes continuous layer depth. These variables play different roles. We analyze how perturbations vary across token positions while they evolve in t. The main spectral conclusion is particularly explicit when V is time-independent and symmetric, with $\lambda_1 > \lambda_2$ and $\lambda_1 > 0$: the amplitudes of transverse Fourier modes obey

$$|\widehat{Y}(\xi,t)| \le e^{-\kappa(\xi)t}|\widehat{Y}(\xi,0)|, \qquad \kappa(\xi) = \frac{\lambda_1-\lambda_2}{1+\xi^2} + \frac{\lambda_1\xi^2}{1+\xi^2}.$$

This displays how token-depth frequency combines the spectral gap with the dominant value eigenvalue in the decay rate. Theorem B.9 gives the precise statement; the subsequent theorems describe nilpotent transients and time-dependent eigenvalues. The change of variables

$$r = -\log s \in [0,\infty), \qquad X(r,t) = x(e^{-r},t)$$

turns the prefix average into a translation-invariant averaging formula on a half-line. Define

$$\mathcal{A}_+F(r) := \int_0^\infty e^{-y}F(r+y)\,dy, \qquad r \ge 0, \tag{27}$$

and

$$\Sigma_r[X] := \mathcal{A}_+(XX^\top) - (\mathcal{A}_+X)(\mathcal{A}_+X)^\top.$$

The subscript $+$ records that the physical domain is $[0,\infty)$.

Proposition B.6 (Prefix averaging in logarithmic coordinates)**.** *If x satisfies the monokinetic equation equation 25, then $X(r,t) = x(e^{-r},t)$ satisfies*

$$\partial_t X = \mathsf{P}_X\big(V(t)\mathcal{A}_+X + \beta V(t)\Sigma_r[X]T(t)^\top X\big). \tag{28}$$

The change of variables requires no differentiability in token position: it is valid almost everywhere whenever the prefix integrals and the original evolution equation are defined.

Proof. In the prefix mean, substitute $u = se^{-y}$:

$$\frac{1}{s}\int_0^s x(u,t)\,du = \int_0^\infty e^{-y}x(se^{-y},t)\,dy = \mathcal{A}_+X(-\log s,t).$$

The same substitution applied to $x(u,t)x(u,t)^\top$ transforms the prefix second moment into $\mathcal{A}_+(XX^\top)$. Subtracting the outer product of the transformed means gives $\Sigma_r[X]$. Since the change of token coordinate is independent of t, the claimed equation follows. □

Filter orientation. Writing $A = \mathcal{A}_+ F$, differentiation in the distributional sense gives

$$\partial_r A = A - F.$$

Thus $\mathcal{A}_+$ uses larger values of r: it is anti-causal in increasing r. In the coordinate $q = -r = \log s \leq 0$, the same operator is the stable causal exponential filter

$$\widetilde{A}(q) = \int_0^\infty e^{-y} \widetilde{F}(q-y)\, dy, \qquad \partial_q \widetilde{A} = \widetilde{F} - \widetilde{A}.$$

This orientation agrees with the original prefix order in s. The integral formula, rather than an arbitrary initial condition at $r = 0$, selects the appropriate solution of $A_r = A - F$.

Proposition B.7 (Linearization at a constant eigenprofile)**.** *Let $\hat{e} \in \mathbb{S}^{d-1}$ be independent of time and suppose*

$$V(t)\hat{e} = \lambda(t)\hat{e}$$

for almost every t. Then $X(r,t) \equiv \hat{e}$ is a stationary profile for equation 28. For bounded tangent variations $Y(r) \in \hat{e}^\perp$, use the spherical chart

$$X_\varepsilon(r) = \frac{\hat{e} + \varepsilon Y(r)}{|\hat{e} + \varepsilon Y(r)|}.$$

At each time when $V(t)$ and $T(t)$ are finite, the derivative of the vector field in equation 28 at this profile is

$$\partial_t Y = \mathsf{P}_{\hat{e}}\big(V(t)\mathcal{A}_+ Y\big) - \lambda(t) Y. \tag{29}$$

In particular, β and T do not enter this linearized equation. The same assertions hold at $-\hat{e}$, with the same eigenvalue $\lambda(t)$ and the same tangent operator.

Proof. At a constant profile, $\mathcal{A}_+\hat{e} = \hat{e}$ and $\Sigma_r[\hat{e}] = 0$, so the projected drift vanishes. The normalized chart satisfies $X_\varepsilon = \hat{e} + \varepsilon Y + O(\varepsilon^2)$ uniformly in r. Since covariance is unchanged by adding a constant,

$$\Sigma_r[X_\varepsilon] = \varepsilon^2 \left[\mathcal{A}_+(YY^\top) - (\mathcal{A}_+ Y)(\mathcal{A}_+ Y)^\top\right] + O(\varepsilon^3) \tag{30}$$

in the uniform norm. In particular, its first derivative is zero. For fixed z, differentiating $\mathsf{P}_{X_\varepsilon}(z) = z - (X_\varepsilon \cdot z)X_\varepsilon$ at $\varepsilon = 0$ gives

$$\frac{d}{d\varepsilon}\mathsf{P}_{X_\varepsilon}(z)\bigg|_{\varepsilon=0} = -(Y \cdot z)\hat{e} - (\hat{e} \cdot z)Y.$$

Taking $z = V\hat{e} = \lambda\hat{e}$ gives $-\lambda Y$; differentiating the averaged value gives $\mathsf{P}_{\hat{e}}\big(V\mathcal{A}_+ Y\big)$. The covariance term contributes nothing at first order. The argument at $-\hat{e}$ is identical because $\mathsf{P}_{-\hat{e}}(z) = \mathsf{P}_{\hat{e}}(z)$ for every z and $V(-\hat{e}) = \lambda(-\hat{e})$. □

The same linear response holds for exact softmax attention. The preceding cancellation is not limited to a small-β model. For completeness, define the exact continuum prefix average

$$\mathcal{M}_\beta[x](s) := \frac{\displaystyle\int_0^s e^{\beta x(s)^\top T x(u)} x(u)\, du}{\displaystyle\int_0^s e^{\beta x(s)^\top T x(u)}\, du}, \qquad s > 0. \tag{31}$$

Here β is any fixed finite nonnegative number. The exact monokinetic softmax vector field is $\mathsf{P}_{x(s)}\big(V\mathcal{M}_\beta[x](s)\big)$. This formula is considered as an exact continuum model; no additional discrete-to-continuum limit for exact softmax is needed for the derivative calculation below.

Proposition B.8 (Temperature-independent derivative of exact softmax)**.** *At the aligned configuration $x_1 = \cdots = x_L = \hat{e}$, with $V\hat{e} = \lambda\hat{e}$, the exact finite-particle softmax dynamics has tangent linearization*

$$\partial_t y_\ell = \mathsf{P}_{\hat{e}}\left(V\frac{1}{\ell}\sum_{j=1}^{\ell} y_j\right) - \lambda y_\ell, \qquad y_j \in \hat{e}^\perp. \tag{32}$$

For the exact continuum vector field defined by equation 31, its derivative at $x(s) \equiv \hat{e}$, in every bounded measurable tangent direction y, is

$$\partial_t y(s) = \mathsf{P}_{\hat{e}}\left(V \frac{1}{s}\int_0^s y(u)\,du\right) - \lambda y(s). \tag{33}$$

Both derivatives are independent of β and T. For time-dependent matrices and a common eigenvector, these statements hold at almost every time, and equation 33 becomes equation 29 after the logarithmic change of variables.

Proof. At an aligned finite configuration all scores in a prefix agree, hence $a_{\ell j} = 1/\ell$. Let a dot denote differentiation with respect to the perturbation parameter at zero. Normalization implies $\sum_j \dot{a}_{\ell j} = 0$. Consequently,

$$\frac{d}{d\varepsilon}\sum_{j=1}^{\ell} a_{\ell j}(\varepsilon)x_j(\varepsilon)\bigg|_{\varepsilon=0} = \frac{1}{\ell}\sum_{j=1}^{\ell} y_j + \hat{e}\sum_{j=1}^{\ell}\dot{a}_{\ell j} = \frac{1}{\ell}\sum_{j=1}^{\ell} y_j.$$

Differentiating the projection gives the term $-\lambda y_\ell$.

For the continuum calculation, write $\langle f\rangle_s = s^{-1}\int_0^s f(u)\,du$, and use the normalized spherical chart for x_ε. Put

$$c = e^{\beta \hat{e}^\top T \hat{e}}, \qquad h(s,u) = \beta\big(y(s)^\top T\hat{e} + \hat{e}^\top T y(u)\big).$$

The numerator and denominator of equation 31, after dividing each by s, have expansions

$$N_\varepsilon(s) = c\hat{e} + \varepsilon c\big(\langle y\rangle_s + \hat{e}\langle h(s,\cdot)\rangle_s\big) + O(\varepsilon^2),$$
$$D_\varepsilon(s) = c + \varepsilon c\langle h(s,\cdot)\rangle_s + O(\varepsilon^2).$$

The errors are uniform in $s > 0$ for bounded y, fixed finite β, and finite matrices. This follows from the uniform score bound and the fact that $\langle\cdot\rangle_s$ is an average of norm one. The quotient derivative is therefore $D\mathcal{M}_\beta[\hat{e}]y = \langle y\rangle_s$: the two score variation terms cancel. Differentiating the projection completes the proof. This also proves a derivative in the uniform norm, rather than only a pointwise formal expansion. □

B.6 Fourier modes in logarithmic token depth

We analyze the linearized equation equation 29 one Fourier mode at a time. For each $\xi \in \mathbb{R}$, consider

$$Y_\xi(r,t) = e^{i\xi r}\widehat{Y}(\xi,t), \qquad \widehat{Y}(\xi,t) \in \mathbb{C}^d, \qquad \hat{e}^\top \widehat{Y}(\xi,t) = 0. \tag{34}$$

Here $\widehat{Y}$ denotes the amplitude of a prescribed harmonic, rather than the integral Fourier transform of an arbitrary half-line profile. The projection $\mathsf{P}_{\hat{e}}(z)$ is extended complex linearly; taking real parts gives real tangent perturbations.

For $\Phi(y) = e^y \mathbf{1}_{\{y\le 0\}}$, its Fourier multiplier is

$$\widehat{\Phi}(\xi) = \int_{-\infty}^{0} e^y e^{-i\xi y}\,dy = \frac{1}{1-i\xi}. \tag{35}$$

The same multiplier acts directly on each harmonic on the physical domain $r \ge 0$:

$$\mathcal{A}_+ Y_\xi(r,t) = e^{i\xi r}\widehat{Y}(\xi,t)\int_0^\infty e^{-y}e^{i\xi y}\,dy = \frac{e^{i\xi r}}{1-i\xi}\widehat{Y}(\xi,t).$$

The integral converges absolutely for every real ξ. Substitution into equation 29 therefore gives the exact mode equation

$$\partial_t \widehat{Y}(\xi,t) = \frac{1}{1-i\xi}\mathsf{P}_{\hat{e}}\big(V(t)\widehat{Y}(\xi,t)\big) - \lambda(t)\widehat{Y}(\xi,t). \tag{36}$$

For locally integrable coefficients, this finite-dimensional equation has a unique locally absolutely continuous solution for every initial amplitude. Tangency is preserved because $\partial_t(\hat{e}^\top\widehat{Y}) = -\lambda(t)\hat{e}^\top\widehat{Y}$. Thus Fourier-mode estimates follow directly from equation 36, without a reconstruction of general token profiles. In the original token coordinate, the harmonic is $e^{i\xi r} = s^{-i\xi}$, so ξ measures oscillation across logarithmic token depth.

B.7 SPECTRAL DYNAMICS OF LOGARITHMIC TOKEN-DEPTH MODES

In all three theorems below, $\widehat{Y}(\xi, t)$ solves equation 36 with $\hat{e}^\top \widehat{Y}(\xi, 0) = 0$. The estimates hold for every $\xi \in \mathbb{R}$ and every $t \geq 0$, and all vector norms on $\mathbb{C}^d$ are Euclidean.

Theorem B.9 (Symmetric value matrix). *Let $V \in \mathbb{R}^{d\times d}$ be constant and symmetric, with orthonormal eigenvectors $e_1, \dots, e_d$ and eigenvalues*

$$Ve_k = \lambda_k e_k, \qquad \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d, \qquad \lambda_1 > 0.$$

Choose $\hat{e} = e_1$ or $\hat{e} = -e_1$. Then

$$|\widehat{Y}(\xi, t)| \leq \exp\left[-\left(\lambda_1 - \frac{\lambda_2}{1+\xi^2}\right) t\right] |\widehat{Y}(\xi, 0)|. \tag{37}$$

Consequently, with $\kappa = \min\{\lambda_1, \lambda_1 - \lambda_2\} > 0$,

$$|\widehat{Y}(\xi, t)| \leq e^{-\kappa t} |\widehat{Y}(\xi, 0)|. \tag{38}$$

If $\lambda_2 \geq 0$, then $\kappa = \lambda_1 - \lambda_2$.

Proof. Expand $\widehat{Y} = \sum_{k=2}^{d} z_k e_k$. Each component solves

$$\partial_t z_k = \left(\frac{\lambda_k}{1 - i\xi} - \lambda_1\right) z_k, \qquad z_k(\xi, t) = e^{(\lambda_k/(1-i\xi)-\lambda_1)t} z_k(\xi, 0).$$

Its real growth exponent is $\lambda_k/(1+\xi^2) - \lambda_1$, which is largest when $\lambda_k = \lambda_2$. Orthogonality gives equation 37. Moreover,

$$\lambda_1 - \frac{\lambda_2}{1+\xi^2} = \frac{\lambda_1 - \lambda_2}{1+\xi^2} + \frac{\xi^2 \lambda_1}{1+\xi^2} \geq \kappa.$$

This proves the uniform mode estimate equation 38. The projection and tangent space are the same at the two poles. □

Frequency-dependent contraction and phase. For an eigencomponent, the complex growth exponent is

$$\frac{\lambda_k}{1 - i\xi} - \lambda_1 = -\left(\lambda_1 - \frac{\lambda_k}{1+\xi^2}\right) + i\frac{\lambda_k \xi}{1+\xi^2}.$$

Thus long wavelengths in logarithmic token position are controlled by $\lambda_1 - \lambda_k$, while the damping rate at large $|\xi|$ tends to λ_1. The imaginary part describes the phase evolution of each mode. For the slowest transverse component,

$$\kappa(\xi) = \frac{\lambda_1 - \lambda_2}{1+\xi^2} + \frac{\lambda_1 \xi^2}{1+\xi^2}.$$

If $\lambda_2 > 0$, the damping increases with $|\xi|$; if $\lambda_2 < 0$, it decreases with $|\xi|$, so high frequencies are slower and the uniform rate is λ_1. Thus a low-pass averaging multiplier does not imply that the linearized dynamics always damps high-frequency perturbations faster. The positivity condition $\lambda_1 > 0$ is essential for a uniform positive damping rate over all frequencies, even when the spectral gap is positive.

Theorem B.10 (Single Jordan block). *Let $d \geq 2$ and*

$$V = J_d(\lambda) = \lambda I_d + N_d, \qquad \lambda \in \mathbb{R},$$

where $(N_d)_{k,k+1} = 1$ and all other entries are zero. Let e_1 be the first standard basis vector and set $\hat{e} = e_1$ or $-e_1$. There is a constant C_d, depending only on d, such that

$$|\widehat{Y}(\xi, t)| \leq C_d(1 + t^{d-2}) e^{-\lambda\xi^2 t/(1+\xi^2)} |\widehat{Y}(\xi, 0)|. \tag{39}$$

Proof. Write $\widehat{Y} = (0, z)$ with $z \in \mathbb{C}^{d-1}$. Removing the first component of $V\widehat{Y}$ gives the tangent block $\lambda I_{d-1} + N_{d-1}$, where $N_1 = 0$. Hence

$$z(\xi, t) = e^{(\lambda/(1-i\xi)-\lambda)t} \sum_{j=0}^{d-2} \frac{t^j N_{d-1}^j}{j!(1-i\xi)^j} z(\xi, 0).$$

Since $\|N_{d-1}^j\| \le 1$, $|1-i\xi|^{-1} \le 1$, and

$$\operatorname{Re}\left(\frac{\lambda}{1-i\xi} - \lambda\right) = -\lambda\frac{\xi^2}{1+\xi^2}, \qquad \sum_{j=0}^{d-2}\frac{t^j}{j!} \le C_d(1+t^{d-2}),$$

the asserted mode bound follows. □

Role of the Jordan block. For $\lambda > 0$, each fixed nonzero frequency decays exponentially up to the displayed polynomial factor. The rate tends to zero as $\xi \to 0$. At $\xi = 0$, the amplitude evolves under N_{d-1}: the constant harmonic is neutral when $d = 2$ and may grow polynomially when $d \ge 3$. There is therefore no positive exponential decay rate uniform over all frequencies in this case. For $\lambda < 0$, the scalar exponential grows at nonzero frequencies.

Theorem B.11 (Time-dependent matrix with a fixed Jordan basis)**.** *Suppose*

$$V(t) = S\,\operatorname{diag}\big(\lambda_{\max}(t), J_2(t), \dots, J_M(t)\big)S^{-1}, \qquad J_b(t) = \lambda_b(t)I_{m_b} + N_b, \tag{40}$$

where $M \ge 2$, S is a fixed real invertible matrix, $Se_1 = \hat{e} \in \mathbb{S}^{d-1}$, each N_b is the fixed standard nilpotent Jordan shift of size m_b, and $1 + \sum_{b=2}^{M} m_b = d$. The eigenvalue functions are real-valued and locally integrable. Assume there is $\gamma > 0$ such that

$$\lambda_{\max}(t) - \lambda_b(t) \ge \gamma, \qquad b = 2, \dots, M, \tag{41}$$

for almost every t. Write

$$m_{\max} = \max_{b\ge 2} m_b, \qquad \mathcal{L}(t) = \int_0^t \lambda_{\max}(\tau)\,d\tau, \qquad \Lambda(\xi,t) = \frac{\xi^2\mathcal{L}(t) + \gamma t}{1+\xi^2}.$$

Then a constant C, depending only on d and S, satisfies

$$|\widehat{Y}(\xi,t)| \le C(1+t^{m_{\max}-1})e^{-\Lambda(\xi,t)}|\widehat{Y}(\xi,0)|. \tag{42}$$

If also $\lambda_{\max}(t) \ge \lambda_* > 0$ almost everywhere, then for every $0 < \eta < \min\{\lambda_*, \gamma\}$ there is a constant C_η such that

$$|\widehat{Y}(\xi,t)| \le C_\eta e^{-\eta t}|\widehat{Y}(\xi,0)|. \tag{43}$$

These statements also hold for the profile $-\hat{e}$.

Proof. Let $z = S^{-1}\widehat{Y}$, decomposed into the same blocks as equation 40. For any vector v,

$$S^{-1}\mathsf{P}_{\hat{e}}(v) = S^{-1}v - e_1(\hat{e}^\top v).$$

Thus the projection changes only the first Jordan coordinate. Every block $b \ge 2$ solves the closed equation

$$\partial_t z^{(b)} = \left[\left(\frac{\lambda_b(t)}{1-i\xi} - \lambda_{\max}(t)\right)I_{m_b} + \frac{N_b}{1-i\xi}\right]z^{(b)}.$$

The time-dependent scalar part commutes with the fixed nilpotent matrix, so

$$z^{(b)}(\xi,t) = \exp\left[\int_0^t\left(\frac{\lambda_b(\tau)}{1-i\xi} - \lambda_{\max}(\tau)\right)d\tau\right]\sum_{j=0}^{m_b-1}\frac{t^j N_b^j}{j!(1-i\xi)^j}z^{(b)}(\xi,0).$$

The real scalar exponent obeys

$$\operatorname{Re}\left(\frac{\lambda_b(t)}{1-i\xi} - \lambda_{\max}(t)\right) = -\frac{\xi^2\lambda_{\max}(t)}{1+\xi^2} - \frac{\lambda_{\max}(t) - \lambda_b(t)}{1+\xi^2}.$$

The gap hypothesis and the same nilpotent estimate as above therefore give

$$|z^{(b)}(\xi,t)| \le C_d(1+t^{m_{\max}-1})e^{-\Lambda(\xi,t)}|z^{(b)}(\xi,0)|.$$

Let $w = (z^{(2)}, \dots, z^{(M)})$. Although the first coordinate of z need not vanish, tangency gives

$$\widehat{Y} = \mathsf{P}_{\hat{e}}(Sz) = \mathsf{P}_{\hat{e}}\big(S(0,w)\big), \qquad |\widehat{Y}| \le \|S\|\,|w|, \qquad |w(\xi,0)| \le \|S^{-1}\|\,|\widehat{Y}(\xi,0)|.$$

Squaring the block estimates, summing, and using these inequalities proves equation 42 with $C = C_d\|S\|\|S^{-1}\|$.

Under the additional positivity assumption, $\Lambda(\xi,t) \ge \min\{\lambda_*, \gamma\}t$. For every indicated η, the polynomial is absorbed by the remaining exponential decay, giving equation 43. The argument at the opposite pole uses the same projection and tangent operator. Equivalently, one may replace the first column of S by its negative without changing any norm constants. □

Spectral mechanism. The factor $(1 - i\xi)^{-1}$ comes from causal prefix averaging in logarithmic token depth. In equation 36, it multiplies the projected value action, making the eigenvalues, spectral gaps, and Jordan structure of V explicit in the mode estimates. These are linear response results near a constant eigenprofile, valid also for the exact-softmax linearization. The time-dependent result assumes a fixed Jordan basis; nonlinear cluster formation requires further analysis.

B.8 Stationary Profiles and Clustered Measures

The linearized estimates describe perturbations near a constant eigenprofile. They are compatible with other stationary configurations, including profiles that occupy both antipodal directions and measures supported on several directions. We first record these configurations and then distinguish their existence from their stability. Except in the first proposition, the matrices V and T in this subsection are fixed, and all measure-valued statements refer to the truncated transport equation equation 21.

Proposition B.12 (Signed eigenprofiles)**.** *Let $\beta \geq 0$. Suppose that $V(t)$ and $T(t)$ are measurable and locally bounded, and that there is a unit vector $\hat{e}$ satisfying*

$$V(t)\hat{e} = \lambda(t)\hat{e} \qquad \text{for almost every } t.$$

For any measurable function $\sigma : [0, \infty) \to \{-1, 1\}$, the profile

$$X_*(r) = \sigma(r)\hat{e}$$

is a stationary solution of equation 28. A constant unit profile $X_*(r) = u$ is stationary if and only if $V(t)u$ is parallel to u for almost every t.

Proof. Set

$$a(r) = \int_0^\infty e^{-y}\sigma(r + y)\, dy.$$

The prefix mean and covariance are

$$\mathcal{A}_+[X_*](r) = a(r)\hat{e}, \qquad \Sigma_r[X_*] = (1 - a(r)^2)\hat{e}\hat{e}^\top.$$

Consequently,

$$V(t)\mathcal{A}_+[X_*](r) = \lambda(t)a(r)\hat{e},$$
$$V(t)\Sigma_r[X_*]T(t)^\top X_*(r) = \lambda(t)(1 - a(r)^2)\sigma(r)\big(\hat{e}^\top T(t)^\top \hat{e}\big)\hat{e}.$$

Both vectors are parallel to $X_*(r)$, so their tangential projections vanish. For a constant profile u, the prefix covariance is zero and the stationary condition reduces to $\mathsf{P}_u(V(t)u) = 0$, which is equivalent to the stated condition. □

This proposition also holds for the exact softmax equation: any weighted average of vectors in $\{\hat{e}, -\hat{e}\}$ lies in their common line, and $V(t)$ preserves that line. In particular, a nonzero prefix sign imbalance alone does not move an exactly antipodal profile toward one pole.

Theorem B.13 (Centered isotropic stationary measures)**.** *Assume $VT^\top = I_d$. Let $v_1, \ldots, v_K \in \mathbb{S}^{d-1}$ and $\alpha_1, \ldots, \alpha_K > 0$ satisfy*

$$\sum_{k=1}^K \alpha_k = 1, \qquad \sum_{k=1}^K \alpha_k v_k = 0, \qquad \sum_{k=1}^K \alpha_k v_k v_k^\top = \frac{1}{d} I_d.$$

Then, for every $\beta \geq 0$, the position-independent family

$$\mu_s = \sum_{k=1}^K \alpha_k \delta_{v_k}, \qquad 0 < s \leq 1,$$

is stationary for equation 21.

Proof. Since μ_s is independent of s, its conditional and prefix means both vanish, and its prefix covariance is $\Sigma(s) = I_d/d$. At every support point v_k,

$$Vm(s) + \beta V\Sigma(s)T^\top v_k = \frac{\beta}{d} VT^\top v_k = \frac{\beta}{d} v_k.$$

The tangential velocity therefore vanishes at each atom, which proves the stationary weak transport equation. The covariance normalization is forced by the unit-sphere constraint: taking traces gives $d\sigma^2 = 1$ whenever the second moment has the form $\sigma^2 I_d$. □

Theorem B.14 (Antipodal mixtures in a shared eigenbasis)**.** *Assume that $VT^\top$ is symmetric with an orthonormal eigenbasis $v_1, \dots, v_d$, and that $T^\top$ is diagonal in the same basis:*

$$VT^\top v_j = \lambda_j v_j, \qquad T^\top v_j = \tau_j v_j.$$

For $1 \le K \le d$ and positive weights with $\sum_{j=1}^K \alpha_j = 1$, the position-independent family

$$\mu_s = \sum_{j=1}^K \alpha_j \, \frac{\delta_{v_j} + \delta_{-v_j}}{2}$$

is stationary for equation 21 for every $\beta \ge 0$.

Proof. Antipodal symmetry gives $m(s) = 0$, and position independence gives

$$\Sigma(s) = \sum_{j=1}^K \alpha_j v_j v_j^\top.$$

For a support point $x = \pm v_j$, we have $\Sigma(s) T^\top x = \alpha_j T^\top x$. Thus

$$V\Sigma(s) T^\top x = \alpha_j V T^\top x = \alpha_j \lambda_j x.$$

Its tangential projection is zero. This argument also covers $\tau_j = 0$ and does not require dividing by an eigenvalue. □

Theorem B.15 (Zero conditional means at $\beta = 0$)**.** *Let $\beta = 0$ and let $s \mapsto \mu_s$ be a measurable, time-independent family of probability measures on $\mathbb{S}^{d-1}$. If*

$$\bar{x}(s) := \int_{\mathbb{S}^{d-1}} x \, \mu_s(dx) = 0 \qquad \textit{for almost every } s \in (0, 1],$$

then $\{\mu_s\}$ is stationary for equation 21, for every fixed matrix V.

Proof. The prefix mean, which is distinct from the conditional mean, satisfies

$$m(s) = \frac{1}{s} \int_0^s \bar{x}(u) \, du = 0.$$

When $\beta = 0$, the velocity in equation 21 is $\mathsf{P}_x(Vm(s)) = 0$. Hence the entire transport flux vanishes. □

Corollary B.16 (Antipodal symmetry at $\beta = 0$)**.** *If $\beta = 0$ and $\mu_s(A) = \mu_s(-A)$ for every Borel set $A \subset \mathbb{S}^{d-1}$ and almost every s, then the time-independent family $\{\mu_s\}$ is stationary for equation 21.*

Proof. Antipodal symmetry implies $\bar{x}(s) = -\bar{x}(s)$, hence $\bar{x}(s) = 0$. Apply Theorem B.15. □

Corollary B.17 (Zero-mean measures in an eigenspace)**.** *Let $\beta = 0$ and let $V_\lambda = \ker(V - \lambda I_d)$ be an eigenspace of V. Every measurable, time-independent family of probability measures supported on $V_\lambda \cap \mathbb{S}^{d-1}$ and satisfying $\int x \, \mu_s(dx) = 0$ for almost every s is stationary for equation 21. In particular, this holds for any antipodally symmetric such family when $\dim V_\lambda \ge 2$.*

Proof. This is another special case of Theorem B.15. The eigenspace support and dimension conditions are not needed for stationarity; they describe a particular geometry within the larger zero-mean class. □

Scope of the equilibrium results. Theorems B.13 and B.14 are sufficient constructions, not a classification of stationary measures. In these constructions the projected covariance drift vanishes at the support points, despite the nonzero covariance. In particular, these examples should not be interpreted as evidence that a nonzero covariance necessarily produces tangential motion. The centered isotropic construction uses only the first two moments and is specific to the truncated equation; higher-order softmax terms need not vanish. Its stationarity consequently does not imply stationarity of the exact softmax model. More generally, finite-time approximation by the truncated dynamics does not alone justify passing its equilibrium or long-time conclusions to exact attention.

C Conclusion

Fourier analysis of logarithmic token depth reveals how the value matrix governs the linear response of causal attention. The prefix operator contributes the multiplier $(1 - i\xi)^{-1}$, leading to the mode equation

$$\partial_t \widehat{Y} = \frac{1}{1 - i\xi} \mathsf{P}_{\hat{e}}\big(V(t)\widehat{Y}\big) - \lambda(t)\widehat{Y}.$$

For a constant symmetric V with $\lambda_1 > \lambda_2$ and $\lambda_1 > 0$, the resulting estimate is

$$|\widehat{Y}(\xi, t)| \le e^{-\kappa(\xi)t} |\widehat{Y}(\xi, 0)|, \qquad \kappa(\xi) = \frac{\lambda_1 - \lambda_2}{1 + \xi^2} + \frac{\lambda_1 \xi^2}{1 + \xi^2}.$$

The spectral gap controls low-frequency decay, whereas the dominant eigenvalue controls the high-frequency limit. The Jordan estimates further expose polynomial transients and the dependence on integrated time-varying eigenvalues. This explicit connection between token-depth frequency and the spectrum of V is the main theoretical contribution.

The same mode equation follows from exact softmax at every fixed finite inverse temperature. The estimates describe linear perturbations of a constant eigenprofile; the stationary measures give complementary examples of clustered configurations in the truncated model. Nonlinear attraction and cluster formation remain beyond these results.

D Further Experimental Details

D.1 Empirical Validation of the Mean-Field Limit

We empirically validate Assumption 3.1, which posits that the empirical measure

$$\mu^L \;=\; \frac{1}{L} \sum_{\ell=1}^{L} \delta_{\left(\frac{\ell}{L},\, x_\ell\right)}$$

converges weakly to a limiting measure $\mu(ds, dx) = ds\, \mu_s(dx)$ as $L \to \infty$, where μ_s is a depth-indexed probability law on the unit sphere $\mathbb{S}^{d-1}$.

Evaluation method. Since the limiting measure μ is not available in closed form, we assess convergence via a Cauchy criterion: if μ^L converges weakly, then for every bounded continuous test function φ and every depth-weighting function ψ,

$$\big|\langle \mu^L, \psi \otimes \varphi\rangle - \langle \mu^{2L}, \psi \otimes \varphi\rangle\big| \;\longrightarrow\; 0 \quad \text{as } L \to \infty,$$

where the paired observable is defined as

$$\langle \mu^L, \psi \otimes \varphi\rangle \;=\; \frac{1}{L} \sum_{\ell=1}^{L} \psi\big(\tfrac{\ell}{L}\big)\, \varphi(x_\ell).$$

A power-law decay rate of this gap as L grows provides quantitative evidence of convergence.

Experimental setup. We conduct the experiment using pretrained weights from Llama-3.1-8B. To approximate the layerwise dynamics studied in our theory, we construct a 64-layer architecture in which each layer reuses the parameters of the 20-th layer of the original model. We probe the limit measure with two complementary depth-weighting functions:

$$\psi_{\text{unif}}(s) \;=\; s, \qquad \psi_{\text{bump}}(s) \;=\; \exp\!\Big(-\tfrac{1}{(s-0.2)(0.8-s)}\Big) \cdot \mathbf{1}_{s \in (0.2,\, 0.8)}.$$

The uniform weighting ψ_{uniform} probes the global depth profile, while the smooth compactly supported bump ψ_{bump} isolates a strict interior subinterval, ensuring that convergence is not driven solely by boundary effects. We use the four bounded continuous test functions

$$\varphi_1(x) = \|x_{[\,:\,d/4]}\|^2, \quad \varphi_2(x) = \|x_{[d/4\,:\,d/2]}\|^2, \quad \varphi_3(x) = \|x_{[d/2\,:\,]}\|^2, \quad \varphi_4(x) = \|x_{[\,:\,d/4]}\|^2 \cdot \|x_{[d/2\,:\,]}\|^2,$$

which probe both individual coordinate blocks and their cross-interactions.

Convergence-rate estimation. For each Cauchy pair $(L, 2L)$, layer index j, test function φ_i, and depth weighting ψ, we compute the absolute gap

$$g(L, j, i) \;=\; \left|\langle \mu_j^L, \psi \otimes \varphi_i \rangle - \langle \mu_j^{2L}, \psi \otimes \varphi_i \rangle\right|$$

on a per-sample basis across the corpus. We then estimate the log–log convergence rate α_j by linear regression of $\log g(L, j, i)$ against $\log L$. A negative value of α_j indicates power-law decay of the Cauchy gap, and hence weak convergence of the empirical measure at layer j.

Results. Figure 1 reports the mean of α_j across test functions and corpus samples at each layer, for both depth weightings. Under both ψ_{unif} and ψ_{bump}, we observe $\alpha_j < 0$ at every layer, with values stabilizing in the range $[-0.35, -0.20]$ after a brief transient at shallow layers. The agreement between the two weightings, despite their different supports, indicates that the convergence is intrinsic to the depth-indexed measure. Together, these results provide empirical support for Assumption 3.1.