PATTERN FORMATION IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

What are the inductive biases of a Transformer architecture? Existing theory on how the forward pass shapes representations either considers whether Transformers escape from *rank collapse* or demonstrates that self-attention drives tokens toward *cluster* patterns. The latter view arises from an elegant dynamical systems perspective, but relies on simplified architectural assumptions, and does not explain the rich structures observed in practice. This leaves a major open question: when a full Transformer escapes rank collapse, how does it structure token representations? Using pattern-formation theory, we show that the dynamical view of Transformers can account for Positional Encoding, Multi-Head Attention, and Output-Value geometry. We demonstrate that a full Transformer architecture imposes an inductive prior by selectively amplifying a rich set of previously unreported patterns, including traveling or rotating waves among others. We characterize the role of each architectural component in controlling which pattern is amplified, which ones stabilize, compete, or coexist. Finally, we show that these structures can act as a controllable *dynamical prior* that facilitates learning. By choosing both task-aligned positional encoding and weight initialization, we demonstrate improved data efficiency and accelerated optimization on controlled sequence tasks and with ConViT on CIFAR-10.

1 INTRODUCTION

Transformers are universal approximators (Yun et al., 2020; Dehghani et al., 2019) and Turing complete (Pérez et al., 2019). These general results establish that Transformers can represent virtually any function *in principle*, but they say very little about the structural biases of the architecture itself. Specifically, what representations does a forward pass through a Transformer architecture naturally promote before learning takes place?

Existing theory that studies this question typically falls into one of two lines of work. The first focuses on *rank collapse* (or oversmoothing), which analyzes how repeated self-attention drives all token representations toward a common vector (Dong et al., 2021; Noci et al., 2022). The second, based on dynamical systems, characterizes the geometry of representations and shows that self-attention drives tokens into discrete *clusters* (Geshkovski et al., 2023; Bruno et al., 2026) on simplified architectures. Yet, these global results are derived under strong simplifications, omitting Positional Encoding (PE), Multi-Head Attention (MHA), and feed-forward nonlinearities (FFN), and do not account for the complex spatial and sequential patterns that Transformers capture in practice. This gap raises a fundamental question: when a full Transformer escapes collapse, what structural priors does the architecture impose on its representations? More practically, how do these patterns affect learning, and can they be controlled at initialization?

In this paper, we show that signals that survive rank collapse in Transformers are neither arbitrary noise, nor clusters alone. Using pattern-formation theory (Cross & Hohenberg, 1993), we demonstrate that around the collapsed state, a randomly initialized Transformer can exhibit complex emerging modes. These joint token–feature modes are characterized by the Transformer's architectural components and are preferentially selected and amplified with depth, revealing a set of spontaneously emerging structures: clusters, standing or traveling waves, and feature rotations. We call the propensity toward these representations a *dynamical prior*: an inductive bias present before learning and dictated by the architecture rather than by data. Recent evidence on the role of architectural biases in Transformers at initialization (Keller & Welling, 2023; Zheng et al., 2025; Li et al., 2026b) motivates the application of our approach to guide the training of these models.

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.

Our contributions are threefold: (1) We derive a matrix-valued dispersion relation (a frequency-resolved gain matrix) for a Transformer block to determine the structure of emerging modes and show how each architectural component contributes to it. We catalog the representations it can amplify, including the known clustered states and previously unreported standing, traveling, feature-rotating and mixed ones (Sec. 4). (2) We derive amplitude equations that characterize how FFN and content-attention nonlinearities control saturation, phase drift and competition (Sec. 5). (3) We validate our theory, show how initializations that amplify task-relevant structure learn faster and are more data-efficient than gain-matched controls on controlled tasks and on CIFAR-10 with ConViT (Sec. 6) [1].

2 Related Work

Signal propagation, rank collapse and oversmoothing. Self-attention in Transformers can rapidly collapse toward token-averaged representations (Dong et al., 2021; Kedia et al., 2024; Saada et al., 2025). This hinders training, as it has been shown to induce vanishing gradients (Noci et al., 2022). This rank collapse phenomenon also plagues graph neural networks (Oono & Suzuki, 2020), BERT and ViTs (Shi et al., 2022; Dovonon et al., 2025), where self-attention imposes a low-pass filter keeping only the homogeneous signal (Wang et al., 2022; Park & Kim, 2022). All of these ask *whether* the signal survives, whereas we consider what *shape it takes*.

Transformers as dynamical systems. Treating tokens as interacting particles (Lu et al., 2020) has enabled further characterization of how clusters (Geshkovski et al., 2023), multiscale dynamics (Bruno et al., 2026), and spectral selection (Tomihari & Karakida, 2026; Kuehn & Yoon, 2026) occur in Transformers. These works provide elegant global results, but typically have to consider simplified dynamics, such as single head, symmetric or identity values, no PE and no FFN. In contrast, our local analysis retains all key architectural components and determines their role in the emergence of structures in Transformers.

Initialization and inductive biases. On the design side, we rely on the observation that initialization-target alignment is necessary for efficient learning (Abbe et al., 2022), and that random networks and Transformers carry systematic architecture-dependent biases (Teney et al., 2024; Li et al., 2026b; Subramaniam et al., 2026). We make one such bias explicit and controllable by showing theoretically how ConViT (d'Ascoli et al., 2021), mimetic (Trockman & Kolter, 2023) initializations and traveling waves (Keller & Welling, 2023; Keller et al., 2024) impose dynamical priors that facilitate the training of these architectures.

3 Setup: Information Flow through a Transformer at Initialization

Transformer Architecture. Let $X = (x_1, \dots, x_N) \in \mathbb{R}^{N\times d}$ denote N token representations $x_i \in \mathbb{R}^d$, with $h \in \{1, \dots, H\}$ indexing attention heads. For head h, let $Q_h, K_h, V_h \in \mathbb{R}^{d_h\times d}$ be the query, key and value matrices and let $b_{ij}^{(h)}$ denote the positional encoding. The attention weights are

$$A_{ij}^{(h)}(X) = \frac{\exp\Big(d_h^{-1/2}\langle Q_h x_i, K_h x_j\rangle + b_{ij}^{(h)}\Big)}{\sum_{m=1}^{N}\exp\Big(d_h^{-1/2}\langle Q_h x_i, K_h x_m\rangle + b_{im}^{(h)}\Big)}. \tag{1}$$

Writing the output projection as $O = [O_1 \cdots O_H]$, define $M_h := O_h V_h \in \mathbb{R}^{d\times d}$, with $\mathrm{rank}(M_h) \le d_h$. The Multi-Head Attention is then expressed as: $\mathrm{MHA}^{(\ell)}(X) = \sum_h \sum_j A_{ij}^{(h)}(X) M_h^{(\ell)} x_j$. Therefore, a Transformer block defines the mapping:

$$[\mathcal{F}^{(\ell)}(X)]_i = \mathcal{H}^{(\ell)}\Big(x_i + \mathrm{MHA}_i^{(\ell)}(X)\Big), \qquad \mathcal{H}^{(\ell)}(y) = y + \phi^{(\ell)}(y), \tag{2}$$

where $\phi^{(\ell)}$ is a Feed Forward Network (FFN). Repeated application of this block gives $X^{(\ell+1)} = \mathcal{F}^{(\ell)}(X^{(\ell)})$. We omit LayerNorm to streamline the analysis; it has been reported as not strictly

[1] The code used to reproduce all experiments will be available upon publication.

necessary to prevent rank collapse (Noci et al., 2022) and can be replaced by point-wise nonlinearities (Zhu et al., 2025). Pre-LN preserves the block structure of the linear theory but changes the base state (Remark A.3, App. A.9).

A result motivating our analysis is that, with small residual scalings, this architecture can be viewed as a Lie–Trotter splitting (Lu et al., 2020) of the form:

$$\frac{dX}{dt} = \underbrace{\mathrm{MHA}(X)}_{\mathcal{P}(X)\text{: propagation across tokens}} + \underbrace{\mathrm{FFN}(X)}_{\mathcal{L}(X)\text{: local pointwise dynamics}} . \tag{3}$$

Equation 3 mirrors the classic *reaction–diffusion* (or *pattern-forming*) systems studied in dynamical systems, mathematical biology (Turing, 1952; Cross & Hohenberg, 1993; Cross & Greenside, 2009), and more recently graph and sequence architectures (Keller & Welling, 2023) (see App. L for a brief primer). In such systems, nonlocal coupling across positions ($\mathcal{P}$) competes with local pointwise nonlinearities ($\mathcal{L}$). A key property of such systems is that when a spatially uniform state becomes unstable, perturbations do not grow arbitrarily; instead, the coupling selectively amplifies specific spatial frequencies, causing structured patterns.

The collapsed state. In a Transformer, the spatially uniform state corresponds to *rank collapse* (Dong et al., 2021; Noci et al., 2022; Shi et al., 2022): the set of homogeneous states $\mathcal{M} = \{\mathbf{1} \otimes a : \ a \in \mathbb{R}^d\}$ in which all N tokens share the exact same feature vector $a \in \mathbb{R}^d$. With uniform attention, without the residual and FFN, $\mathcal{P}$ acts as a pure averaging (diffusion) operator that contracts all tokens onto $\mathcal{M}$. While prior signal-propagation theory asks when $\mathcal{M}$ is attractive or avoided, we use the reaction-diffusion view to study what structures emerge when $\mathcal{M}$ becomes unstable.

4 The Transformer Matrix Dispersion Relation

Our goal in this section is to determine which structures a Transformer architecture preferentially amplifies at initialization. The starting point is to determine what can emerge from the homogeneous manifold $\mathcal{M}$. To do so, consider a small heterogeneous perturbation $U \in \mathbb{R}^{N\times d}$ around collapsed state $X_\star \in \mathcal{M}$. The following simplified Theorem determines the dynamics of this perturbation when it passes a Transformer block.

Theorem 4.1 (Simplified Linear stability analysis)**.** *For a single-head Transformer and with the aforementioned notations, the homogeneous manifold $\mathcal{M}$ is invariant under a Transformer block $\mathcal{F}$: $\mathcal{F}(\mathbf{1}\otimes a) = \mathbf{1}\otimes g(a)$, with $g(a) := \mathcal{H}(a + Va)$.*

Let $X_\star = \mathbf{1}\otimes a_\star \in \mathcal{M}$ with $g(a_\star) = a_\star$, and let $A_\star = A(X_\star)$ be the attention matrix at $X_\star$. In first order in U, token-level perturbations evolve through the Transformer map as:

$$u_i^+ = C_\star\Big[u_i + \sum\nolimits_j (A_\star)_{ij} V u_j\Big], \tag{4}$$

where V is the value/output matrix and $C_\star$ is the FFN Jacobian at $a_\star$.

Further, let $\{\lambda_q | q \in \{0,\dots N-1\}\}$ be the eigenvalues of $A_\star$ with corresponding eigenvectors e_q (the index q becomes a frequency in Sec. 4.1). Then, after a Transformer update, a perturbation $U = e_q \otimes v$ along e_q evolves as:

$$U^+ = e_q \otimes J(q)v, \qquad J(q) = C_\star(I_d + \lambda_q V). \tag{5}$$

The full version of this Theorem is in App. A.1. It states that the eigenvectors (or modes) e_q determine the available token-space structures, while the matrices $J(q)$ determine whether each structure grows ($\rho(J(q)) > 1$) or decays ($\rho(J(q)) < 1$) with depth, and along which feature-space directions (the eigenvectors of $J(q)$). A Transformer can therefore preferentially amplify particular joint token-feature structures even before learning: this is what we consider to be a *dynamical prior*. Note that a Jacobian analysis of weight-tied attention was conducted in Tomihari & Karakida (2026), but without PE and FFN, and in which the token-space structure of $A_\star$ is left unresolved.

Inspired by pattern-formation theory, we call the matrix-valued map $q \to J(q)$ the *Transformer Matrix Dispersion Relation*. We build it progressively, to isolate the roles of PE, OV geometry, MHA, and the FFN, and finally use it to classify the structures promoted at linear order.

4.1 Positional Encoding Lifts Heterogeneous Space Degeneracy

We start with a single attention head without positional encoding and ask: do the linearized dynamics preferentially select a heterogeneous token-space structure?

Proposition 4.2 (No PE: frequency degeneracy)**.** *Without PE ($b_{ij} = 0$), any homogeneous state $X_\star \in \mathcal{M}$ produces uniform attention, i.e. $A_\star = N^{-1}\mathbf{1}\mathbf{1}^\top$. The token-space then decomposes into the two invariant eigenspaces $\mathbb{R}^N = \mathrm{span}\{\mathbf{1}\} \oplus \mathbf{1}^\perp$, corresponding respectively to the homogeneous mode and the $(N-1)$-dimensional space of zero-mean heterogeneous modes. Their associated Jacobian matrices are*

$$J(0) = C_\star(I + V), \qquad J(q) = C_\star \quad (q \neq 0). \tag{6}$$

In linear order, all nonzero token frequencies have identical feature-space amplification.

This result demonstrates that, in linear order, the value matrix V controls the preference between homogeneous and heterogeneous perturbations, consistent with the negative-identity OV shape reported by Trockman & Kolter (2023). However, without positional information, the dispersion relation cannot favor one heterogeneous structure over another [2]. We next show that PE provides precisely the missing mechanism.

Proposition 4.3 (PE lifts frequency degeneracy)**.** *Suppose we have a relative PE, i.e. $b_{ij} = b_r$, where $r = i - j$, with periodic boundary conditions. At any homogeneous state $X_\star \in \mathcal{M}$, the attention matrix is circulant: $[A_\star]_{ij} = A_{i-j}$. Therefore, $\forall k \in \{0, \dots, N-1\}$, with $q = \frac{2\pi k}{N}$, $A_\star$ admits the N Fourier eigenpairs $(\lambda(q), e_q)$, and $J(q) = C_\star(I_d + \lambda(q)V)$ as in Eq. 5:*

$$A_\star e_q = \lambda(q) e_q, \qquad (e_q)_j = \mathrm{e}^{\mathrm{i}qj}, \qquad \lambda(q) = \sum_r A_r \mathrm{e}^{-\mathrm{i}qr}. \tag{7}$$

This result gives positional encoding a direct dynamical interpretation: since $\lambda(q)$ is the Fourier transform of the attention matrix, designing the positional encoding is designing a filter in token-space, while V and $C_\star$ determine how each transmitted mode evolves in feature space. The Transformer can now preferentially amplify specific token-space structures. An extension of this result without periodic boundary conditions is presented in App. A.7.

4.2 Multi-head attention engineers the dispersion relation

Distinguishing modes is not the same as being able to preferentially amplify a specific one. We therefore ask which modes a single attention head can select as the dominant growing mode. For a broad class of PEs, the answer is surprisingly restrictive.

Proposition 4.4 (Single-head PE select spectral edges)**.** *Assume that $C_\star = cI$, $c > 0$, $V = \eta I$, and let the positional encoding be reflection symmetric with real Fourier spectrum $\lambda(q)$ monotonically decreasing on $q \in [0, \pi]$. Then*

$$\rho(J(q)) = c|1 + \eta\lambda(q)|, \qquad \arg\max_{q \in [0,\pi]} \rho(J(q)) \subseteq \{0, \pi\}. \tag{8}$$

Hence, a single head can select either a low-frequency mode ($q_c = 0$) or a high-frequency mode ($q_c = \pi$), but not an isolated intermediate wavelength. This extends to any V (App. A).

This result refines the usual low-pass picture of self-attention (Wang et al., 2022; Park & Kim, 2022): the sign of the OV gain decides whether a single head is low-pass ($\eta > 0$, $q_c = 0$) or high-pass ($-2/(1 + \lambda(\pi)) < \eta < 0$, $q_c = \pi$), but never band-pass. The result covers reflection-symmetric localized positional encodings such as bidirectional ALiBi (Press et al., 2022) and the centered Gaussian kernel of ConViT (d'Ascoli et al., 2021). Isolated intermediate wavelengths therefore require positional encodings outside this class, such as causal ALiBi and off-center ConViT, or with RoPE (Su et al., 2024) that induces a non-symmetric, non-monotone positional profile on $\mathcal{M}$ (more details in App. A.8). We next show that multi-head attention provides another solution: different heads contribute distinct positional spectra and OV geometries, which combine into a matrix-valued filter capable of producing a richer dispersion relation.

[2] We will see in Th. 5.1 that degeneracy is lifted by the nonlinearities, forming clusters of token representations.

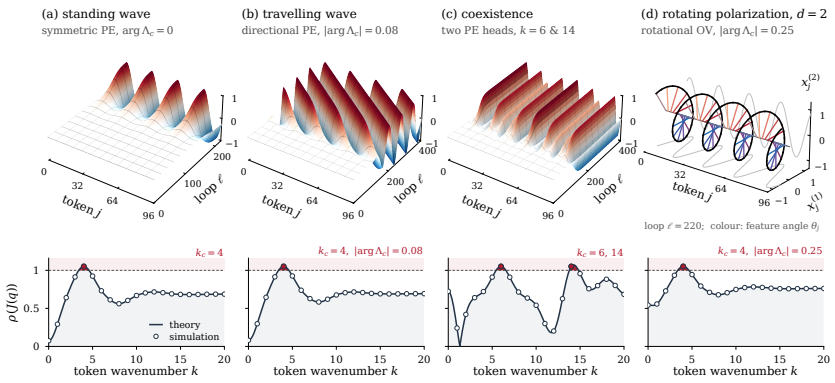


Figure 1: **Taxonomy of patterns in Transformers:** $2D$ visualization of the propagation of the accessible modes, with their corresponding dispersion relation.

Theorem 4.5 (Transformer Matrix Dispersion Relation)**.** *Suppose the PE is translation invariant, i.e.* $b_{i,j} = b_{i-j}$*, let* $X_\star \in \mathcal{M}$*, and let* $M_h := O_h V_h$ *for a concatenated MHA of* H *heads. For a perturbation* $U = e_q \otimes v$ *along Fourier mode* q*, token-wise perturbations evolve as*

$$u_i^+ = \mathrm{e}^{\mathrm{i}qi} J(q) v, \qquad J(q) = C_\star \left[I + \mathcal{D}(q) \right], \qquad \mathcal{D}(q) := \sum\nolimits_{h=1}^{H} \lambda_h(q) M_h. \tag{9}$$

Thus, MHA constructs a matrix-valued spectral filter $\mathcal{D}(q)$*, rather than a simple token–feature coupling.*

Eq. 9 is the central object of this work: at initialization, a Transformer block acts as a matrix-valued spectral filter on token-space modes. For each head, the positional encoding determines the scalar response $\lambda_h(q)$, the OV matrices M_h the feature directions through which that response acts, and the FFN Jacobian $C_\star$ rescales the result.

Scope and weight tying. The results hold on all of $\mathcal{M}$ and along depth, also for layer-dependent blocks (App. A.10). Weight tying is thus not required for the linear theory; we introduce it only in Sec. 5, where depth as repeated iteration of the same map allows us to study the nonlinear stabilization of growing modes, as in Geshkovski et al. (2023); Bruno et al. (2026).

4.3 A CATALOGUE OF JOINT TOKEN–FEATURE PATTERNS

Now that the Transformer Matrix Dispersion Relation is laid out, we can turn its eigenpairs into concrete representations. The following proposition makes the growth of a joint token-feature mode explicit and shows how the phase of Λ_c and the geometry of v_c determine the resulting morphologies.

Proposition 4.6 (Shape of the emerging patterns)**.** *Let* $a_\star$ *be a fixed point of* g *and* (Λ_c, v_c) *a simple eigenpair of* $J(q_c)$ *with* $|\Lambda_c| > 1$*. The perturbation* $U = e_{q_c} \otimes v_c$ *along mode* q_c *grows exponentially with depth and is of the form*

$$u_j^{(\ell)} \sim B_0 \Lambda_c^\ell v_c \mathrm{e}^{\mathrm{i}q_c j} + \overline{B_0 \Lambda_c^\ell v_c \mathrm{e}^{\mathrm{i}q_c j}}, \tag{10}$$

with $\Lambda_c = \mathrm{e}^{\gamma_c + i\omega_c}$*,* $B_0 = |B_0| \mathrm{e}^{\mathrm{i}\theta_0}$ *the initial perturbation amplitude. Writing* $v_c = v_\Re + \mathrm{i} v_\Im$*, Eq. 10 gives*

$$u_j^{(\ell)} \simeq 2|B_0| \mathrm{e}^{\ell\gamma_c} \left[v_\Re \cos\Theta_{j\ell} - v_\Im \sin\Theta_{j\ell} \right], \qquad \Theta_{j\ell} = q_c j + \ell\omega_c + \theta_0. \tag{11}$$

Eq. 11 gives a direct dictionary between the spectral data of $J(q_c)$ and the resulting representation. The wavenumber q_c sets the spatial scale across tokens, the growth rate γ_c controls amplification with depth, the phase ω_c determines whether the pattern drifts across layers, and the real and imaginary parts of v_c describe its geometry in feature space. Different combinations therefore give rise to different morphologies, simulated in Fig. 1: standing waves ($\omega_c = 0$, $v_c \in \mathbb{R}^d$), traveling waves ($\omega_c \neq 0$, from a directional PE), feature rotations ($v_c \in \mathbb{C}^d$, from non-symmetric OV geometry), and their mixtures. Notably, traveling waves have been *hard*-designed into neural architectures (Keller & Welling, 2023; Keller et al., 2024) for sequence learning. We demonstrate here that they emerge spontaneously in Transformers whenever their PE is directional or their OV non-symmetric.

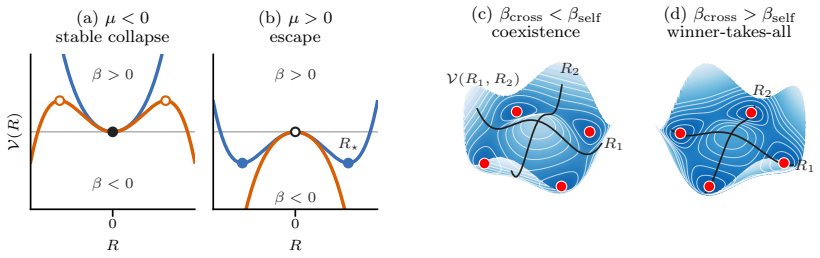


Figure 2: **Depth evolution descends an effective potential.** (a,b) One critical mode, Eq. 12 (blue: $\beta > 0$, orange: $\beta < 0$; disks: stable states, circles: unstable): the dispersion relation sets μ, the nonlinearities set β. (c,d) Two critical modes, Eq. 19, over the signed amplitudes (R_1, R_2): the ratio of cross-suppression β_{cross} to self-saturation β_{self} decides where the wells are.

5 Past Linear Order: Amplitude Equations for Transformers

The dispersion relation in Sec. 4 answers a first question: *which joint token-feature structures does a Transformer initially amplify?* Linear growth, however, cannot be the end of the story. Once the selected structure becomes appreciable, the approximation underlying $J(q)$ breaks down. This leaves three questions that are invisible to the dispersion relation: does a linearly growing mode stabilize at a finite amplitude? When several structures have exactly the same linear gain, which one is eventually realized? When distinct modes grow simultaneously, do they coexist or suppress one another? The non-linear analysis below addresses these questions with the second tool of pattern-formation theory, the *amplitude equation* (Cross & Hohenberg, 1993).

5.1 From linear growth to finite-amplitude patterns

Consider first a single critical mode q_c whose gain is close to the instability threshold, $\rho(J(q_c)) = 1 + \mu$ with $|\mu| \ll 1$. Near this threshold, the selected mode evolves slowly while modes whose gains remain strictly below one decay rapidly, so the token dynamics can be reduced to an analysis of amplitude of the critical modes. For a single stationary pattern with amplitude B, the reduced dynamics has the generic form

$$B^+ - B = \mu B - \beta B^3 = -\mathcal{V}'(B), \qquad \mathcal{V}(B) = -\frac{\mu}{2}B^2 + \frac{\beta}{4}B^4, \qquad \mu = \rho\big(J(q_c)\big) - 1. \quad (12)$$

The two coefficients have different meanings (Fig. 2). The linear coefficient μ is already determined by the dispersion relation of Sec. 4: it sets the curvature of the effective potential around the collapsed state, and therefore when that state loses stability. The nonlinear coefficient β, by contrast, depends on terms discarded by the linearization and determines what replaces it. If $\mu > 0$ and $\beta > 0$, nonlinear feedback counteracts linear instability and the pattern saturates at a finite amplitude $B_\star = \sqrt{\mu/|\beta|}$. If instead $\beta < 0$, the cubic nonlinearity reinforces the instability.

5.2 Amplitude Equations of Clusters and Travelling Waves

We now study the effects of these higher-order terms on two representations: clusters and traveling waves. For clarity, we work with the scalar reduction $d = 1$ of the full Transformer block in Eq. 2 (the general $d > 1$ case reduce to the same form by projection onto the left eigenvector), around $a_\star = 0$ (the extension to $a_\star \neq 0$ is sketched in App. F) with OV gain η, content coupling χ, and translation-invariant PE b_{i-j}:

$$A_{ij}(x) = \frac{e^{\chi x_i x_j + b_{i-j}}}{\sum_m e^{\chi x_i x_m + b_{i-m}}}, \qquad H(x) = x + \eta A(x)x, \qquad F(x) = H(x) + \nu\phi(H(x)), \quad (13)$$

Nonlinearity ϕ is smooth and $\phi(0) = 0$, $C_\star = \alpha_1 := 1 + \nu\phi'(0)$, $\alpha_2 := \nu\phi''(0)/2$ and $\alpha_3 := \nu\phi'''(0)/6$, where ν is the residual scale. H and F are the residual and Transformer mapping of Eq. 2, respectively. Expanding F around $x = 0$ (Lemma B.1, App. B) separates the dispersion relation from the higher-order terms:

$$F(x) = \underbrace{\textstyle\sum_q J(q)\,\hat{x}_q\,e_q}_{\text{dispersion relation}} + \underbrace{\alpha_2(Lx)^{\circ 2} + \alpha_3(Lx)^{\circ 3}}_{\text{FFN nonlinearity}} + \underbrace{\alpha_1\eta\chi\,\mathcal{C}(x)}_{\text{content attention}} + O(\|x\|^4), \quad (14)$$

where $x = \sum_q \hat{x}_q e_q$ and $J(q) = \alpha_1(1 + \eta\lambda_q)$ is the dispersion relation of Theorem 4.1 with $V = \eta$ and λ_q the eigenvalues of $A_\star = A(0)$. The FFN acts on the attention output $Lx := (I + \eta A_\star)x$, and content-dependent attention is $\mathcal{C}(x) := x \circ [A_\star x^{\circ 2} - (A_\star x)^{\circ 2}]$ ($\circ$: component-wise product).

We expose a division of labor between Transformer components. We have shown that PE, OV geometry, and the linearized FFN determine **which mode is initially amplified** through $J(q)$, i.e., μ. We demonstrate here how content-dependent attention and higher order properties of FFN determine, through β, **what happens to that mode once it has grown**.

Theorem 5.1 (Clusters, simplified)**.** *Without PE and with a negative OV gain, $\eta \in (-2, 0)$, all heterogeneous modes share the amplification $J(q \neq 0) = 1 + \mu$ (Proposition 4.2). Near threshold, the separation u of two balanced clusters, $x_i = \pm u$ up to a small common shift, obeys*

$$u^+ - u = \mu u - \beta_{\mathrm{cl}} u^3, \qquad \beta_{\mathrm{cl}} = -\eta\chi - \alpha_3 + \frac{2(1+\eta)}{\eta}\alpha_2^2. \tag{15}$$

If $\mu > 0$ and $\beta_{\mathrm{cl}} > 0$, the clusters saturate at a finite separation $u_\star = \sqrt{\mu/\beta_{\mathrm{cl}}}$, which is stable along the cluster direction. Against intra-cluster perturbations, it is stable only for an odd activation ($\alpha_2 = 0$) with $\alpha_3 < 0$. The full statement and proof are in App. C.

Non-linear effects on clustering. Here the instability is driven by the FFN ($C_\star = \alpha_1 > 1$), with $\eta < 0$ stabilizing the mean, whereas the clusters of Geshkovski et al. (2023) are driven by attention with $V = I$. Which structure emerges from the degenerate sector $\mathbf{1}^\perp$ is decided at cubic order, by ingredients absent from the dispersion relation. Negative-OV content attention, $\eta\chi < 0$, is what holds a cluster: as it tightens, its members attend more to one another. Interestingly, an odd saturating activation ($\alpha_3 < 0$) adds damping, whereas any even part $\alpha_2 \neq 0$ makes one cluster spread, so GELU-type FFNs break balanced clusters that $\tanh$-type FFNs keep. Consequently, activations also bias the emerging structure, which to our knowledge was not noted before. This activation function induced structural bias is also present in message-passing GNNs (Turan et al., 2026).

Theorem 5.2 (Traveling waves, simplified)**.** *With a directional PE, the critical mode has a complex gain $J(q_c) = (1 + \mu)\mathrm{e}^{\mathrm{i}\omega_c}$: at linear order, the wave grows by $1 + \mu$ and advances by a phase ω_c per layer. Near threshold, its amplitude G, measured in the frame moving with the wave, obeys*

$$G^+ - G = \mu G - \beta_{\mathrm{tr}}|G|^2 G, \tag{16}$$

with a complex coefficient β_{tr}. If $\mu > 0$ and $\Re(\beta_{\mathrm{tr}}) > 0$, the wave saturates at a stable amplitude $R_\star = |G_\star|$ and propagates at phase speed Ω per layer:

$$R_\star^2 = \frac{\mu}{\Re(\beta_{\mathrm{tr}})}, \qquad \Omega = \omega_c - \Im(\beta_{\mathrm{tr}})R_\star^2. \tag{17}$$

Thus the real part of β_{tr} controls saturation, while its imaginary part gives the nonlinear correction to the propagation speed. The full statement and proof are in App. E.

As illustrated in Fig. 1(d), a complex eigenvector of $J(q_c)$ adds a feature-plane rotation; the same reduction applies after projection onto that eigenmode.

Scaling laws and design levers. All cases (clusters, traveling waves, and standing waves in App. D) predict amplitude scaling laws invisible at linear order: past threshold, cluster separation and wave amplitude grow as $\sqrt{\mu}$, and the wave speed shifts linearly in μ (Fig. 3a,b). This refines the propagation–collapse dichotomy (Giorlandino & Goldt, 2026; Cowsik et al., 2025) into a family of stable structured representations.

5.3 Competition: winner-takes-all versus coexistence

What happens when two modes become critical at once? The linear theory allows both to grow indistinguishably; the non-linear analysis predicts a competition: stable coexistence or winner-takes-all.

Theorem 5.3 (Two-mode competition, simplified)**.** *Let two standing modes q_1, q_2 cross threshold together with the same gain $1 + \mu$, $\mu > 0$. Their amplitudes obey*

$$B_i^+ - B_i = \big(\mu - \beta_{\mathrm{self}}|B_i|^2 - \beta_{\mathrm{cross}}|B_j|^2\big)B_i, \qquad i \neq j \in \{1, 2\}, \tag{18}$$

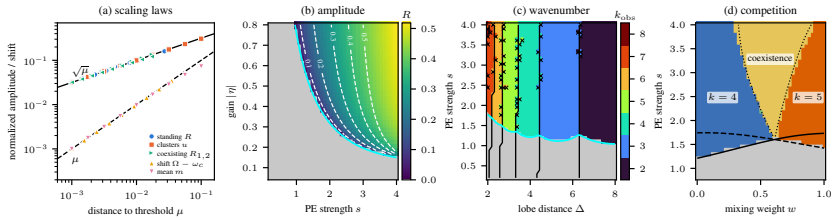


Figure 3: **Prediction of the amplitudes of representations**. Lines: theory, colors: simulation, gray: collapsed state, cyan line: threshold $\max_q |J(q)| = 1$. (a): Amplitude scaling laws for multiple patterns versus μ. (b): Amplitude heatmap of the traveling wave, in the $(|\eta|, s)$ map; white dashed: predicted $R_\star$. (c) Critical mode $k_{\rm obs}$ in the (Δ, s) map; black: theoretically predicted mode transitions ($\times$: nonlinear selection differs); (d): Competition and coexistence of modes 4 (solid), 5 (dashed) in (w, s) with winner-takes-all exchange lines of Thm. 5.3 (dotted).

where $\beta_{\rm self} > 0$ measures how each mode saturates itself and $\beta_{\rm cross}$ how strongly it suppresses the other. Equivalently, the moduli $R_i = |B_i|$ descend the effective potential

$$R_i^+ - R_i = -\frac{\partial \mathcal{V}}{\partial R_i}, \qquad \mathcal{V}(R_1, R_2) = -\frac{\mu}{2}\left(R_1^2 + R_2^2\right) + \frac{\beta_{\rm self}}{4}\left(R_1^4 + R_2^4\right) + \frac{\beta_{\rm cross}}{2} R_1^2 R_2^2, \quad (19)$$

If $\beta_{\rm cross} < \beta_{\rm self}$, the mixed state $R_1^2 = R_2^2 = \mu/(\beta_{\rm self} + \beta_{\rm cross})$ is the stable minimum and the two modes coexist; if $\beta_{\rm cross} > \beta_{\rm self}$, it becomes a saddle, the single-mode states $R_i^2 = \mu/\beta_{\rm self}$ are the stable minima, and the initial condition selects the winner (Fig. 2c,d). The full statement, for unequal gains, and the proof are in App. G.

Kernels and heads that transmit the mixed harmonics $q_1 \pm q_2$ alter the cross-suppression through their stable-mode resolvents, and suppressing those harmonics weakens this route of competition.

6 Experiments

We first validate the theoretical results of Secs. 4 and 5, and then provide practical evidence that accessible representations can induce dynamical priors at initialization on stylized experiments.

6.1 The theory is quantitatively predictive

First, we validate the theory of Sec. 5 by iterating the scalar map in Eq. 13 on a ring of $N = 32$ tokens, with a Transformer using $\phi = \tanh$, OV gain $|\eta|$ and gaussian PE $b_r = sK_\Delta(r)$, where s is the PE strength and Δ the distance between two peaks (details in App. I.2). We measure the dominant Fourier modes, their amplitudes, and, for directional kernels, their phase advance.

The scaling laws of Fig. 3(a) validate our theoretical results on the amplitude growth of multiple representations close to criticality. Panel (b) confirms our predictions in Eq. 17 on how the wave amplitude R increases with gain $|\eta|$ and PE strength s. Panel (c) varies distance Δ between peaks (lobes) of the PE and tests whether the selected wavelength follows the theory $k_{\rm th} = \frac{N}{2\pi} \arg\max_q |J(q)|$. Finally, (d) tests two-mode competition for modes $k = 4$ and $k = 5$, when varying the preference for one mode or the other with mixing w. Our theory clearly distinguishes coexistence from winner-takes-all selection. Together, these results show that the theory predicts not only when collapse is avoided, but also the structure and dynamics of the pattern that replaces it.

6.2 Engineering the prior on controlled tasks

Next, we test whether the dynamical prior we claim to provide can be aligned with a task, and if such an alignment helps the learning process, as stated in (Abbe et al., 2022). Each of the three tasks (Fig. 4, left) targets one of the identified structures and probes its attribute: **T1** reconstructs periodic sequences of mode q_0, favoring $q_c = q_0$. **T2** translates a structured sequence by a prescribed shift, probing the phase θ of the mode. **T3** reconstructs under a prescribed feature-space rotation ψ,

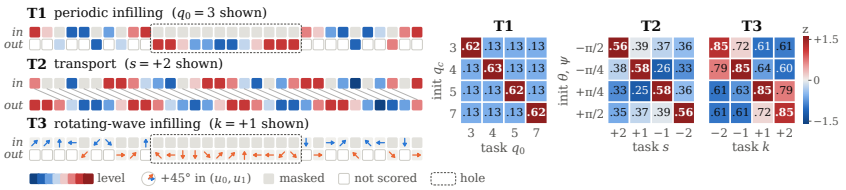
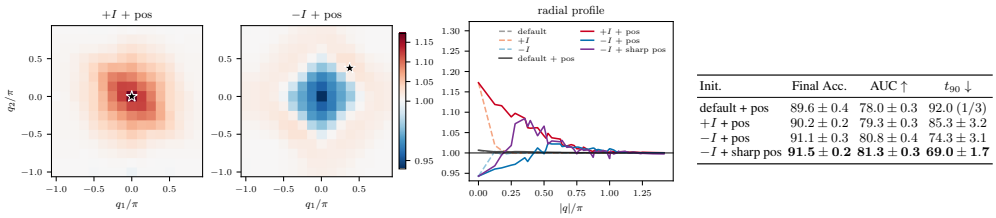


Figure 4: **Task–initialization alignment.** Left: the periodic (T1), transport (T2) and rotation (T3) tasks (perturbed, partially masked targets). Right: AUC for each (task, initialization) pair; the color corresponds to the z-score. Higher scores in the diagonal support the dynamical prior induced by mode criticality.



Init.	Final Acc.	AUC ↑	t_{90} ↓
default + pos	89.6 ± 0.4	78.0 ± 0.3	92.0 (1/3)
$+I$ + pos	90.2 ± 0.2	79.3 ± 0.3	85.3 ± 3.2
$-I$ + pos	91.1 ± 0.3	80.8 ± 0.4	74.3 ± 3.1
$-I$ + sharp pos	$\mathbf{91.5 \pm 0.2}$	$\mathbf{81.3 \pm 0.3}$	$\mathbf{69.0 \pm 1.7}$

(a) (b)

Figure 5: **Dispersion-aware initialization of ConViT on CIFAR-10.** (a) Heatmap of mode gains and radial profile of the dispersion relation for the initializations variants (colours: mode stability; $\star$: leading mode). (b) Final Accuracy, AUC and t_{90}; all variants use localized PE, "sharp": sharpened kernel.

probing the eigenvector geometry of $J(q)$. More details are in App. J. For each task, using identical architectures, parameter counts and gains, we initialize 4 models that differ only in *which* token–feature mode they amplify. We then compute the AUC of their learning curves of each $(\text{task}, \text{init})$ tuple and report their z-scores.

The heatmaps of Fig. 4 (right) clearly indicate that task-aligned initializations benefit from a strong dynamical prior. It originates from engineering the architecture so that the critical mode at initialization is aligned with the problem at hand. Supplementary experiments shown in Fig. 7 show that this practical gain occurs most markedly early in training.

6.3 Dispersion-aware initialization of ConViT on CIFAR-10

Finally, we train a weight-tied ConViT on CIFAR-10 (more details in App. K) for classification and show that engineering mode amplification facilitates learning in practical architectures. On the left of Fig. 5(a), we plot the dispersion relation for various parameterizations: localized and sharpened PE (d'Ascoli et al., 2021) variants shift amplification toward heterogeneous spatial modes. Reporting classification results in Fig. 5(b) shows that combining $-I$ with localized PE substantially accelerates optimization; sharpening the dispersion structure helps further (91.5% test accuracy). This explains why mimetic initialization's $OV = -I$ (Trockman & Kolter, 2023) and ConViT's localized kernels (d'Ascoli et al., 2021) help: together they move the leading multiplier of $J(\mathbf{q})$ off $\mathbf{q} = 0$. In further experiments we track the dispersion relation during training (Fig. 13) which shows that heterogeneous patterns are favored.

7 Conclusion

In this paper, we show that a randomly initialized Transformer is not structureless once it escapes rank collapse: its forward pass multiplies each token frequency by a matrix-valued gain $J(q)$, set by PE, OV geometry and the FFN, that amplifies specific joint token–feature structures such as traveling and rotative waves. This dynamical prior refines the binary propagate-or-collapse picture into a mode-resolved one. The nonlinear theory says which structures saturate, drift or coexist, and aligning the prior with the task accelerates optimization and improves data efficiency.

AI USE STATEMENT

We used a large language model and coding assistant for two tasks: (i) editing the manuscript, i.e. tightening and restructuring prose and (ii) refining and formatting plots and figures. We have reviewed all AI-assisted work and take full responsibility for the content of this paper.

REFERENCES

Emmanuel Abbe, Elisabetta Cornacchia, Jan Hazla, and Christopher Marquis. An initial alignment between neural network and target is needed for gradient descent to learn. In *International Conference on Machine Learning*, pp. 33–52. PMLR, 2022.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, 2019.

Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. A multiscale analysis of mean-field transformers in the moderate interaction regime. *Advances in Neural Information Processing Systems*, 38:133305–133341, 2026.

Jean-Marc Chomaz. Global instabilities in spatially developing flows: non-normality and nonlinearity. *Annu. Rev. Fluid Mech.*, 37(1):357–392, 2005.

Aditya Cowsik, Tamra Nebabu, Xiaoliang Qi, and Surya Ganguli. Geometric dynamics of signal propagation predict trainability of transformers. *Physical Review E*, 112(5):055301, 2025.

Mark C Cross and Pierre C Hohenberg. Pattern formation outside of equilibrium. *Reviews of modern physics*, 65(3):851, 1993.

Michael Cross and Henry Greenside. *Pattern formation and dynamics in nonequilibrium systems*, volume 1. Cambridge University Press Cambridge, 2009.

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.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, pp. 2793–2803. PMLR, 2021.

Gbetondji Jean-Sebastien Dovonon, Michael M. Bronstein, and Matt J. Kusner. Setting the record straight on transformer oversmoothing. *Transactions on Machine Learning Research*, 2025.

Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International conference on machine learning*, pp. 2286–2296. PMLR, 2021.

Lev Fedorov, Michael Eli Sander, Romuald Elie, Pierre Marion, and Mathieu Lauriere. Clustering in deep stochastic transformers. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=opfSL15Uc4.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *Advances in Neural Information Processing Systems*, 38: 41340–41391, 2026.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36:57026–57037, 2023.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *International Conference on Machine Learning*, pp. 11398–11442. PMLR, 2023.

Alessio Giorlandino and Sebastian Goldt. Two failure modes of deep transformers and how to avoid them: A unified theory of signal propagation at initialisation. In *International Conference on Learning Representations*, 2026.

Hermann Haken. Slaving principle revisited. *Physica D: Nonlinear Phenomena*, 97(1-3):95–103, 10 1996. ISSN 0167-2789. doi: 10.1016/0167-2789(96)00080-2. URL http://dx.doi.org/10.1016/0167-2789(96)00080-2.

Ahmadreza Jeddi, Marco Ciccone, and Babak Taati. LoopFormer: Elastic-depth looped transformers for latent reasoning via shortcut modulation. In *International Conference on Learning Representations*, 2026.

Akhil Kedia, Mohd Abbas Zaidi, Sushil Khyalia, JungHo Jung, Harshith Goka, and Haejun Lee. Transformers get stable: An end-to-end signal propagation theory for language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=30waYPIZUA.

T. Anderson Keller and Max Welling. Neural wave machines: Learning spatiotemporally structured representations with locally coupled oscillatory recurrent neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16168–16189, 2023.

T. Anderson Keller, Lyle Muller, Terrence Sejnowski, and Max Welling. Traveling waves encode the recent past and enhance sequence learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=p4S5Z6Sah4.

Hugo Koubbi, Borjan Geshkovski, and Philippe Rigollet. Homogenized transformers. *arXiv preprint arXiv:2604.01978*, 2026.

Christian Kuehn and Jaeyoung Yoon. Spectral selection in symmetric self-attention dynamics. *arXiv preprint arXiv:2604.26085*, 2026.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

Shuzhen Li, Yifan Zhang, Jiacheng Guo, Quanquan Gu, and Mengdi Wang. Deeploop: Depth scaling for looped transformers. *arXiv preprint arXiv:2607.13491*, 2026a.

Siquan Li, Yao Tong, Haonan Wang, and Tianyang Hu. Transformers are born biased: Structural inductive biases at random initialization and their practical consequences. *arXiv preprint arXiv:2602.05927*, 2026b.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. In *ICLR 2020 Workshop on Deep Differential Equations*, 2020.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1ldO2EFPr.

Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=D78Go4hVcxO.

Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

Jorge Pérez, Javier Marinković, and Pablo Barceló. On the turing completeness of modern neural network architectures. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyGBdo0qFm.

Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=sRKtbGsebH.

Han Shi, JIAHUI GAO, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen M. S. Lee, and James Kwok. Revisiting over-smoothing in BERT from the perspective of graph. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dUV91uaXm3.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Vighnesh Subramaniam, David Mayo, Colin Conwell, Tomaso Poggio, Boris Katz, Brian Cheung, and Andrei Barbu. Training the untrainable: Introducing inductive bias via representational alignment. *Advances in Neural Information Processing Systems*, 38:96753–96791, 2026.

Damien Teney, Armand Mihai Nicolicioiu, Valentin Hartmann, and Ehsan Abbasnejad. Neural redshift: Random networks are not random functions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4786–4796. IEEE, 2024.

Akiyoshi Tomihari and Ryo Karakida. Recurrent self-attention dynamics: An energy-agnostic perspective from jacobians. *Advances in Neural Information Processing Systems*, 38:139928–139961, 2026.

Lloyd N Trefethen. Pseudospectra of linear operators. *SIAM review*, 39(3):383–406, 1997.

Asher Trockman and J Zico Kolter. Mimetic initialization of self-attention layers. In *International Conference on Machine Learning*, pp. 34456–34468. PMLR, 2023.

Erkan Turan, Gaspard Abel, Maysam Behmanesh, Emery Pierson, and Maks Ovsjanikov. Beyond reLU: Bifurcation, oversmoothing, and topological priors. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=AehiGqd7go.

Alan Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society B*, 237:37–72, 1952.

Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=O476oWmiNNp.

Kevin Xu and Issei Sato. On expressive power of looped transformers: Theoretical analysis and enhancement via timestep encoding. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=H4BuhRezCV.

Liu Yang, Kangwook Lee, Robert Nowak, and Dimitris Papailiopoulos. Looped transformers are better at learning learning algorithms. In *International conference on learning representations*, volume 2024, pp. 42195–42214, 2024.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

Tong Zhang, Junhao Hu, Yun Peng, and Tao Xie. When does recurrence become an algorithm? convergence selection in weight-tied looped transformers. *arXiv preprint arXiv:2607.20594*, 2026.

Jianqiao Zheng, Xueqian Li, Hemanth Saratchandran, and Simon Lucey. Structured initialization for vision transformers. In *Advances in Neural Information Processing Systems 38*, NeurIPS 2025, pp. 146840–146865. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025. doi: 10.52202/085713-4413. URL http://dx.doi.org/10.52202/085713-4413.

Jiachen Zhu, Xinlei Chen, Kaiming He, Yann LeCun, and Zhuang Liu. Transformers without normalization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14901–14911. IEEE, June, 2025. doi: 10.1109/cvpr52734.2025.01388. URL http://dx.doi.org/10.1109/cvpr52734.2025.01388.

A Proofs for the matrix-valued linear theory

This section proves the linear results of Sec. 4. Token states are columns $x_i \in \mathbb{R}^d$, and a perturbation is denoted $U = (u_1, \dots, u_N)$. For a token-space vector $s \in \mathbb{C}^N$ and a feature vector $v \in \mathbb{C}^d$, we write $s \otimes v$ for the perturbation whose ith token is $s_i v$.

A.1 Full version Theorem 4.1 and proof

Theorem A.1 (Linear stability and token–feature block decomposition)**.** *Let $\mathcal{M} := \{\mathbf{1} \otimes a : a \in \mathbb{R}^d\}$ denote the homogeneous states. $\mathcal{M}$ is invariant under any Transformer block: $\mathcal{F}(\mathbf{1} \otimes a) = \mathbf{1} \otimes g(a)$ with $g(a) := \mathcal{H}(a + Va)$ for a single head ($V \to \sum_h M_h$ for MHA). Let $X_\star = \mathbf{1} \otimes a_\star \in \mathcal{M}$ and $U^+ := \mathcal{F}(X_\star + U) - \mathcal{F}(X_\star)$ the evolution of perturbation $U \in \mathbb{R}^N \otimes \mathbb{R}^d$ through a Transformer block. Then, to first order in U, the token-wise perturbation u_i^+ evolves as:*

$$u_i^+ = C_\star \left[u_i + \sum_{j=1}^{N} (A_\star)_{ij} V u_j \right], \tag{20}$$

where $A_\star = A(X_\star)$ and $C_\star = D\mathcal{H}(y_\star) = I_d + D\phi(a_\star + Va_\star)$ is the Jacobian of the residual FFN branch at $y_\star = a_\star + Va_\star$ (the $\lambda = 1$ block is $Dg(a_\star) = C_\star(I + V)$). Equivalently, the full linearized Jacobian is $\mathcal{J}_\star = (I_N \otimes C_\star)\left[I_N \otimes I_d + A_\star \otimes V\right]$. If $A_\star = S\Lambda S^{-1}$ is diagonalizable, with $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_N)$, then

$$(S^{-1} \otimes I_d)\mathcal{J}_\star(S \otimes I_d) = \bigoplus_{k=1}^{N} J(\lambda_k), \qquad J(\lambda) = C_\star\left[I_d + \lambda V\right], \tag{21}$$

and $\mathrm{sp}(\mathcal{J}_\star) = \bigcup_k \mathrm{sp}(J(\lambda_k))$ holds even without diagonalizability (Schur form). Hence, the transverse dynamics decompose into independent token-space blocks indexed by the eigenmodes of $A_\star$. If $a_\star$ is a fixed point of g, then $X_\star$ is a linearly stable fixed point of $\mathcal{F}$ if and only if

$$\max_{\lambda \in \mathrm{sp}(A_\star)} \rho(J(\lambda)) < 1. \tag{22}$$

Along an orbit $a_\ell = g^\ell(a_0)$ on $\mathcal{M}$, the transverse dynamics is the cocycle $U \mapsto \mathcal{J}(a_\ell)U$, which splits into the products $\prod_{\ell<L} J_\ell(\lambda_{k,\ell})$, $J_\ell(\lambda) = C(a_\ell)[I_d + \lambda V]$, whenever the $A(\mathbf{1} \otimes a_\ell)$ share an eigenbasis. (Secs. 4.1–4.2).

Its proof is below.

Proof. Write the effective one-head attention output as

$$\mathcal{A}_i(X) := \sum_{j=1}^{N} A_{ij}(X) V x_j, \tag{23}$$

where, in the one-head setting, any output projection can be absorbed into V. Let

$$y_i(X) := x_i + \mathcal{A}_i(X), \qquad [\mathcal{F}(X)]_i = \mathcal{H}(y_i(X)). \tag{24}$$

Every row of $A(X)$ sums to one for every X:

$$A(X)\mathbf{1} = \mathbf{1}. \tag{25}$$

Consequently, at a homogeneous state $X = \mathbf{1} \otimes a$ all value vectors coincide and $\mathcal{A}_i(\mathbf{1} \otimes a) = Va$ for every i, whatever the positional bias or mask; since $\mathcal{H}$ is token-wise, $\mathcal{F}(\mathbf{1} \otimes a) = \mathbf{1} \otimes g(a)$ with $g(a) = \mathcal{H}(a + Va)$, which proves the invariance of $\mathcal{M}$. Now fix $X_\star = \mathbf{1} \otimes a_\star \in \mathcal{M}$, not necessarily a fixed point. Differentiating the row-sum identity at $X_\star$ in an arbitrary direction U gives

$$DA_{X_\star}[U]\mathbf{1} = 0. \tag{26}$$

We now differentiate the attention output. By the product rule,

$$D\mathcal{A}_i(X_\star)[U] = \sum_j \left(DA_{X_\star}[U]\right)_{ij} Va_\star + \sum_j (A_\star)_{ij} Vu_j \tag{27}$$

$$= \left[\sum_j \left(DA_{X_\star}[U]\right)_{ij}\right] Va_\star + \sum_j (A_\star)_{ij} Vu_j \tag{28}$$

$$= \sum_j (A_\star)_{ij} Vu_j, \tag{29}$$

where the first term vanishes by Eq. equation 26. Thus first-order changes in the attention weights do not act on the homogeneous value vector.

Let $y_\star = a_\star + Va_\star$ denote the common value of $y_i(X_\star)$ and set

$$C_\star := D\mathcal{H}(y_\star) = I + D\phi(y_\star). \tag{30}$$

Since $\mathcal{H}$ acts token-wise, the chain rule and Eq. equation 29 yield

$$u_i^+ = C_\star \left[u_i + \sum_{j=1}^{N} (A_\star)_{ij} Vu_j\right], \tag{31}$$

which is the first claim.

Viewing $U \in \mathbb{R}^N \otimes \mathbb{R}^d$, this is equivalently

$$\mathcal{J}_\star = (I_N \otimes C_\star)\left[I_N \otimes I_d + A_\star \otimes V\right]. \tag{32}$$

If $A_\star = S\Lambda S^{-1}$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$, then

$$(S^{-1} \otimes I_d)\mathcal{J}_\star(S \otimes I_d) \tag{33}$$

$$= (I_N \otimes C_\star)\left[I_N \otimes I_d + \Lambda \otimes V\right] = \bigoplus_{k=1}^{N} C_\star(I_d + \lambda_k V). \tag{34}$$

Hence the token eigenmodes of $A_\star$ define invariant d-dimensional feature blocks

$$J(\lambda_k) = C_\star(I_d + \lambda_k V). \tag{35}$$

If $A_\star$ is not diagonalizable, take a Schur decomposition $A_\star = STS^*$ with T upper triangular: $(S^* \otimes I_d)\mathcal{J}_\star(S \otimes I_d) = (I_N \otimes C_\star)[I_N \otimes I_d + T \otimes V]$ is block upper triangular with diagonal blocks $J(T_{kk})$, so $\mathrm{sp}(\mathcal{J}_\star) = \bigcup_k \mathrm{sp}(J(\lambda_k))$ in all cases. Therefore

$$\rho(\mathcal{J}_\star) = \max_{\lambda \in \mathrm{sp}(A_\star)} \rho(J(\lambda)). \tag{36}$$

When $a_\star$ is a fixed point of g, $X_\star$ is a fixed point of $\mathcal{F}$, and for a discrete-time map asymptotic linear stability is equivalent to all multipliers of the Jacobian lying strictly inside the unit circle; the stability criterion is precisely Eq. equation 22. Along an orbit $a_\ell = g^\ell(a_0)$ the same computation at each X_{a_ℓ} gives $U^{(\ell+1)} = \mathcal{J}(a_\ell)U^{(\ell)} + O(\|U^{(\ell)}\|^2)$ with $\mathcal{J}(a_\ell) = (I_N \otimes C(a_\ell))[I_N \otimes I_d + A(a_\ell) \otimes V]$; if the matrices $A(a_\ell)$ share the eigenbasis S, conjugating by $S \otimes I_d$ block-diagonalizes every factor simultaneously and the L-step propagator is $\bigoplus_k \prod_{\ell<L} J_\ell(\lambda_{k,\ell})$, whose growth rate is the transverse Lyapunov exponent stated in the theorem. □

A.2 Proof of Proposition 4.2

Proof. With one head and $b_{ij} = 0$, the content logit is identical for every pair (i, j) at a homogeneous state, because all queries are equal and all keys are equal. Each row of the softmax is therefore uniform:

$$A_\star = \frac{1}{N}\mathbf{1}\mathbf{1}^\top. \tag{37}$$

The vector $\mathbf{1}$ is an eigenvector with eigenvalue 1, whereas every $x \in \mathbf{1}^\perp$ satisfies $A_\star x = 0$. Thus

$$\mathbb{R}^N = \mathrm{span}\{\mathbf{1}\} \oplus \mathbf{1}^\perp \tag{38}$$

is the required invariant decomposition. Substituting the two token-space eigenvalues $\lambda = 1$ and $\lambda = 0$ into Theorem 4.1 gives

$$J(0) = C_\star(I + V), \qquad J(q) = C_\star \quad (q \neq 0), \tag{39}$$

which is Eq. equation 6. Hence every nonzero token frequency has the same feature-space multipliers. □

A.3 Proof of Proposition 4.3

Proof. Assume periodic relative positional logits with $b_{ij} = b_{i-j}$. At a homogeneous state the content part of the one-head logit is independent of i, j, and therefore cancels between numerator and denominator of the row softmax. Consequently

$$(A_\star)_{ij} = A_{i-j}, \qquad A_r = \frac{e^{b_r}}{\sum_s e^{b_s}}, \tag{40}$$

so $A_\star$ is circulant.

For the Fourier vector $(e_q)_j = e^{\mathrm{i}qj}$ and the convention $r = i - j$, we obtain

$$(A_\star e_q)_i = \sum_j A_{i-j} e^{\mathrm{i}qj} \tag{41}$$

$$= e^{\mathrm{i}qi} \sum_r A_r e^{-\mathrm{i}qr}. \tag{42}$$

Thus

$$A_\star e_q = \lambda(q) e_q, \qquad \lambda(q) = \sum_r A_r e^{-\mathrm{i}qr}. \tag{43}$$

On the N-point periodic grid the distinct frequencies are $q = 2\pi k/N$, $k = 0, \ldots, N-1$.

Applying Theorem 4.1 to the token eigenvalue $\lambda(q)$ gives, for $U = e_q \otimes v$,

$$U^+ = e_q \otimes J(q)v, \qquad J(q) = C_\star[I + \lambda(q)V], \tag{44}$$

Since the discrete Fourier transform is invertible, the full Jacobian is similar over $\mathbb{C}$ to $\mathrm{diag}_q\, J(q)$ and therefore

$$\rho(\mathcal{J}_\star) = \max_q \rho(J(q)). \tag{45}$$

This gives the stability condition of Theorem A.1. If $\lambda(q)$ varies with q, then different token frequencies see different feature-space blocks, so the heterogeneous frequency degeneracy is lifted. □

A.4 Proof of Proposition 4.4

Proof. Under the scalar specialization $C_\star = cI$, $V = \eta I$,

$$J(q) = c[1 + \eta\lambda(q)]I, \tag{46}$$

and hence

$$\rho(J(q)) = c|1 + \eta\lambda(q)|. \tag{47}$$

Because $\lambda(q)$ is real and monotone on $[0, \pi]$, its image is the interval with endpoints $\lambda(0)$ and $\lambda(\pi)$. Define

$$g(x) := |1 + \eta x|. \tag{48}$$

The function g is convex. Therefore its maximum over any compact interval is attained at an endpoint, which gives

$$\max_{q\in[0,\pi]} \rho(J(q)) = c\max\left\{|1+\eta\lambda(0)|,\, |1+\eta\lambda(\pi)|\right\}. \tag{49}$$

Thus an isolated interior frequency cannot be a strict maximizer. If $\eta \neq 0$ and $\lambda(q)$ is strictly monotone, no interior point can tie an endpoint maximum, and

$$\arg\max_{q\in[0,\pi]} \rho(J(q)) \subseteq \{0,\pi\}, \tag{50}$$

as stated in the non-degenerate case. For a general V with $C_\star = cI$, the eigenvalues of $J(q)$ are $c(1+\nu\lambda(q))$, $\nu \in \mathrm{sp}(V)$, so $\rho(J(q)) = c\max_\nu |1+\nu\lambda(q)|$ is a maximum of convex functions of $\lambda(q)$, hence convex, and the same endpoint argument applies. □

A.5 Proof of Theorem 4.5

Proof. For head h, write the post-output-projection contribution as

$$\mathcal{A}_i^{(h)}(X) := \sum_{j=1}^{N} A_{ij}^{(h)}(X) M_h x_j, \qquad M_h = O_h V_h. \tag{51}$$

The full pre-FFN residual state is

$$y_i(X) = x_i + \sum_{h=1}^{H} \mathcal{A}_i^{(h)}(X). \tag{52}$$

As in the proof of Theorem 4.1, row-stochasticity gives

$$DA_{X_\star}^{(h)}[U]\mathbf{1} = 0. \tag{53}$$

Since the value vector $M_h a_\star$ is the same at every token, the derivative of the hth head is therefore

$$D\mathcal{A}_i^{(h)}(X_\star)[U] = \sum_{j=1}^{N} (A_{h,\star})_{ij} M_h u_j. \tag{54}$$

With $C_\star = D\mathcal{H}(y_\star)$, the full linearization is

$$u_i^+ = C_\star \left[u_i + \sum_{h=1}^{H}\sum_{j=1}^{N} (A_{h,\star})_{ij} M_h u_j \right]. \tag{55}$$

For translation-invariant relative positional logits, the content logit of each head is constant at $X_\star$ and cancels from the row softmax. Therefore every $A_{h,\star}$ is circulant. Write

$$A_{h,\star} e_q = \lambda_h(q) e_q. \tag{56}$$

For the separated perturbation $U = e_q \otimes v$, Eq. equation 55 becomes

$$u_i^+ = e^{\mathrm{i}qi} C_\star \left[I + \sum_{h=1}^{H} \lambda_h(q) M_h \right] v. \tag{57}$$

Thus the feature-space block in Fourier sector q is

$$J(q) = C_\star \left[I + \sum_{h=1}^{H} \lambda_h(q) M_h \right] = C_\star[I + \mathcal{D}(q)], \tag{58}$$

which is Eq. equation 9. Since the Fourier transform block-diagonalizes all circulant $A_{h,\star}$ simultaneously, the full Jacobian is similar to $\mathrm{diag}_q\, J(q)$ and

$$\rho(\mathcal{J}_\star) = \max_q \rho(J(q)). \tag{59}$$

Hence a homogeneous fixed point is linearly stable exactly when $\max_q \rho(J(q)) < 1$. Along an orbit a_ℓ on $\mathcal{M}$ every $A_h(a_\ell)$ is circulant, since translation invariance holds at each homogeneous state, so the Fourier basis diagonalizes all layers simultaneously and the L-layer propagator of mode q is $\prod_{\ell<L} J_\ell(q)$ with $J_\ell(q) = C(a_\ell)[I + \sum_h \lambda_h(q; a_\ell) M_h]$. □

A.6 Proof of Proposition 4.6

Proof. Let v_c be an eigenvector of the critical block:

$$J(q_c)v_c = \Lambda_c v_c. \tag{60}$$

Then $e_{q_c} \otimes v_c$ is an eigenvector of the full complexified Jacobian. Because the Transformer parameters are real,

$$J(-q_c) = \overline{J(q_c)}, \tag{61}$$

so $e_{-q_c} \otimes \overline{v_c}$ is the conjugate mode with multiplier $\overline{\Lambda_c}$.

A real perturbation in the corresponding two-dimensional real invariant subspace can be written

$$u_j^{(0)} = B_0 v_c e^{\mathrm{i}q_c j} + \overline{B_0 v_c e^{\mathrm{i}q_c j}}. \tag{62}$$

After ℓ applications of the linearized block,

$$u_j^{(\ell)} = B_0 \Lambda_c^\ell v_c e^{\mathrm{i}q_c j} + \overline{B_0 \Lambda_c^\ell v_c e^{\mathrm{i}q_c j}}, \tag{63}$$

which is Eq. equation 10.

Now write

$$\Lambda_c = e^{\gamma_c + \mathrm{i}\omega_c}, \qquad B_0 = |B_0| e^{\mathrm{i}\theta_0}, \qquad v_c = a + \mathrm{i}b. \tag{64}$$

Taking the real part explicitly gives

$$u_j^{(\ell)} = 2|B_0| e^{\ell\gamma_c} \left[a \cos \Theta_{j\ell} - b \sin \Theta_{j\ell} \right], \qquad \Theta_{j\ell} = q_c j + \ell\omega_c + \theta_0, \tag{65}$$

which is Eq. equation 11 with the amplitude written explicitly as $|B_0|$. The standing, traveling, feature-rotating, and mixed cases are obtained by specializing q_c, ω_c, and the real and imaginary parts of v_c. □

A.7 Open boundaries: bulk equivalence

Propositions 4.3–4.6 use periodic boundary conditions. The following statement makes precise the sense in which they describe the bulk of a sequence with open boundaries.

Proposition A.2 (Bulk equivalence)**.** *Let the positional kernel have finite range ξ, i.e. $A_r = 0$ for $|r| > \xi$, and let $\mathcal{F}^{\mathrm{o}}$ and $\mathcal{F}^{\mathrm{p}}$ denote the block with open and periodic boundary conditions on N tokens. Then for every state X and every i with $\xi < i \le N - \xi$, $[\mathcal{F}^{\mathrm{o}}(X)]_i = [\mathcal{F}^{\mathrm{p}}(X)]_i$. Consequently, for every X and every $L \ge 1$,*

$$[(\mathcal{F}^{\mathrm{o}})^L(X)]_i = [(\mathcal{F}^{\mathrm{p}})^L(X)]_i \qquad \text{for all } L\xi < i \le N - L\xi, \tag{66}$$

and the same holds for the linearizations $\mathcal{J}^{\mathrm{o}}$, $\mathcal{J}^{\mathrm{p}}$ at any homogeneous state.

Proof. Row i of the attention involves only the tokens j with $A_{i-j} \neq 0$, i.e. $|i - j| \le \xi$. For $\xi < i \le N - \xi$ these are the same tokens in both settings, with the same content and positional logits and hence the same softmax normalizer, and no wrap-around index occurs. Since the FFN is token-wise, $[\mathcal{F}^{\mathrm{o}}(X)]_i = [\mathcal{F}^{\mathrm{p}}(X)]_i$ for such i. For $L \ge 2$, $[\mathcal{F}^{\mathrm{o}}(X^{\mathrm{o}})]_i$ with $L\xi < i \le N - L\xi$ depends only on the tokens X_j^{o} with $|i - j| \le \xi$, all of which satisfy $(L-1)\xi < j \le N - (L-1)\xi$ and therefore coincide with X_j^{p} by induction. The statement for the linearizations follows by differentiating. □

The dispersion relation is therefore a bulk property, exact on a core of $N - 2L\xi$ tokens. For kernels with exponential tails (ALiBi, Gaussian), truncating at range ξ changes each attention row by at most the tail mass $\delta(\xi)$ in total variation, and the L-step discrepancy is $O\big(L\,\delta(\xi)\max_q \rho(J(q))^L\big)$, negligible once ξ is a few decay lengths. Two remarks. First, for reflection-symmetric kernels the open-boundary attention matrix is $A^{\mathrm{o}} = D^{-1}T$ with T symmetric Toeplitz and D the diagonal of row normalizers; it is similar to the symmetric matrix $D^{-1/2}TD^{-1/2}$, so its spectrum is real and close to the circulant symbol for localized kernels. For directional kernels A^{o} is a non-normal Toeplitz matrix whose eigenvalues do not converge to the circulant symbol (Trefethen, 1997): the periodic dispersion relation governs finite-depth growth but not the $L \to \infty$ spectrum, which is the distinction between convective and absolute instability in open flows (Chomaz, 2005). Second, the argument applies verbatim to causal attention with a local kernel: the rows $i > \xi$ are those of the circulant built from the one-sided kernel $A_r \mathbb{1}_{r \ge 0}$, whose symbol is complex, so causal masking enters the theory as an extreme directional kernel.

A.8 RoPE ON THE HOMOGENEOUS MANIFOLD

Remark A.3 (Homogeneous manifold, RoPE, LayerNorm and untied weights)**.** *(i) Theorems 4.1 and 4.5 hold at every point of $\mathcal{M}$: along an orbit only $C(a_\ell)$ (and, for RoPE, the kernel) varies, and finite-depth growth is governed by $\prod_\ell J_\ell(q)$, which is what Sec. K.3.1 measures. (ii) RoPE is not an additive bias, but at X_a its logit is $a^\top Q^\top R_{j-i} K a$: a translation-invariant, generally non-symmetric kernel of size $O(\|a\|^2)$, so Theorem 4.5 applies with an a-dependent symbol $\lambda(q;a)$, invisible at $a=0$. (iii) LayerNorm is pointwise: with pre-LN the block structure is preserved with $\mathcal{D}(q)\to\mathcal{D}(q)$ $D\mathrm{LN}(a)$ and $C_\star = I + D\phi(\mathrm{LN}(y_\star))\, D\mathrm{LN}(y_\star)$, which requires $a\neq 0$ (App. A.9).*

Rotary encoding replaces the additive bias by position-dependent rotations of queries and keys: the logit is $d_h^{-1/2}\langle R_i Q x_i, R_j K x_j\rangle = d_h^{-1/2} x_i^\top Q^\top R_{j-i} K x_j$, where $R_r = \bigoplus_k R(\theta_k r)$ is block-diagonal with 2×2 rotations at frequencies $\theta_1,\dots,\theta_{d_h/2}$ and $R_i^\top R_j = R_{j-i}$. At a homogeneous state X_a the logit equals

$$b_r(a) := d_h^{-1/2}\, a^\top Q^\top R_r K a = d_h^{-1/2} \sum_k \Big[\langle p_k, k_k\rangle \cos(\theta_k r) + (p_k \times k_k)\sin(\theta_k r)\Big], \qquad r = j-i, \tag{67}$$

where $p_k, k_k \in \mathbb{R}^2$ are the kth 2-blocks of Qa and Ka and $p\times k = p_1 k_2 - p_2 k_1$. Hence $b_r(a)$ is translation invariant, the attention at X_a is circulant (on the ring, for $\theta_k \in \frac{2\pi}{N}\mathbb{Z}$; otherwise up to the boundary effects of App. A.7), and Theorem 4.5 applies with the a-dependent symbol $\lambda(q;a)$, the DFT of $A_r(a) \propto \mathrm{e}^{b_r(a)}$; the proof is unchanged since the values Vx_j are not rotated. Three properties follow. (i) $b_r(a) = O(\|a\|^2)$: at $a=0$ RoPE yields uniform attention and no wavelength selection, which is why the base state must be taken on $\mathcal{M}$ rather than at the origin. (ii) $b_r(a)$ is reflection symmetric only if all $p_k \times k_k$ vanish; generically it is directional, $\lambda(q;a)$ is complex, and RoPE supports traveling modes with a single head. (iii) $b_r(a)$ is not monotone: if one block dominates, $A_r(a) \propto \mathrm{e}^{\rho\cos(\theta_k r - \varphi)}$ is a periodic (von Mises) kernel of period $2\pi/\theta_k$, whose DFT is supported on the multiples of θ_k with $\lambda(\pm\theta_k; a) = \mathrm{e}^{\mp \mathrm{i}\varphi} I_1(\rho)/I_0(\rho)$ dominating the non-zero frequencies (I_n the modified Bessel functions); for instance, for $\eta<0$ and $\varphi=\pi$ the leading mode of $|L_q| = |1+\eta\lambda(q;a)|$ is the intermediate wavelength $q_c = \pm\theta_k$, set jointly by the rotary frequency and the content a. Proposition 4.4 does not apply because its monotonicity hypothesis fails.

A.9 LayerNorm

Consider the pre-LN block $[\mathcal{F}(X)]_i = \mathcal{H}\big(x_i + \sum_h \sum_j A_{ij}^{(h)}(\mathrm{LN}(X)) M_h \mathrm{LN}(x_j)\big)$, $\mathcal{H}(y) = y + \phi(\mathrm{LN}(y))$, with $\mathrm{LN}(x) = \gamma \circ z(x) + \beta$, $z(x) = (x - \mu(x)\mathbf{1}_d)/\sigma(x)$. Since LN is token-wise, $\mathrm{LN}(X_a)$ is homogeneous, so at X_a the values $M_h\mathrm{LN}(a)$ coincide across tokens and the content logits are constant: the two facts used in the proofs of Theorems 4.1 and 4.5. Repeating those proofs with the chain rule gives, for $\sigma(a) > 0$,

$$\begin{aligned} J(q) &= C_\star\Big[I + \sum_h \lambda_h(q) M_h N_1\Big], \qquad N_1 = D\mathrm{LN}(a), \\ C_\star &= I + D\phi(\mathrm{LN}(y_\star))\, D\mathrm{LN}(y_\star), \qquad y_\star = a + \sum_h M_h \mathrm{LN}(a), \end{aligned} \tag{68}$$

with

$$D\mathrm{LN}(x) = \frac{1}{\sigma(x)}\,\mathrm{diag}(\gamma)\Big[P - \frac{z(x)z(x)^\top}{d}\Big], \qquad P = I - \tfrac{1}{d}\mathbf{1}_d\mathbf{1}_d^\top, \tag{69}$$

where $P - zz^\top/d$ is the orthogonal projector onto $\{\mathbf{1}_d, z(x)\}^\perp$. Thus LayerNorm preserves the Fourier block structure and only inserts the rank-$(d-2)$ factor N_1: it removes the mean and scale directions from the feature dynamics and rescales the coupling by $1/\sigma(a)$. Two consequences follow. First, LN is singular at $a=0$, so the base state must be a point $a\neq 0$ of $\mathcal{M}$. Second, with pre-LN the increment $g(a) - a = \sum_h M_h\mathrm{LN}(a) + \phi(\mathrm{LN}(y(a)))$ is bounded and, for large $\|a\|$, depends on a only through the direction of Pa; a fixed point of g is therefore not guaranteed, the residual norm typically grows along depth, and the transverse coupling decays like $1/\sigma(a_\ell)$, which flattens the dispersion relation with depth (consistent with the reduced effect of positional sharpening under pre-LN in Table 1). Post-LN and normalization-free pointwise substitutes such as DyT (Zhu et al., 2025) renormalize the state, so fixed points on $\mathcal{M}$ exist and Sec. 5 applies, with the Taylor coefficients of the normalization contributing to α_2, α_3.

A.10 LAYER-DEPENDENT WEIGHTS

Nothing in the proofs of Theorems 4.1 and 4.5 uses that the block is the same at every layer. With layer-dependent parameters $(Q_h^{(\ell)}, K_h^{(\ell)}, M_h^{(\ell)}, \phi^{(\ell)})$ and translation-invariant PE, each layer is circulant at any homogeneous state, the Fourier basis diagonalizes all layers simultaneously, and the L-layer transverse propagator of mode q is $\prod_{\ell<L} J_\ell(q)$ with $J_\ell(q) = C^{(\ell)}(a_\ell)[I + \sum_h \lambda_h^{(\ell)}(q) M_h^{(\ell)}]$, exactly as along an orbit of the tied block. Weight tying is used only to speak of fixed points in Sec. 5.

Two further remarks connect this to standard, untied Transformers. First, at i.i.d. initialization the blocks $(Q_h^{(\ell)}, K_h^{(\ell)}, M_h^{(\ell)}, \phi^{(\ell)})$ are independent draws from one distribution, so $J_\ell(q)$ fluctuates around a common mean $\bar{J}(q)$ and $\prod_{\ell<L} J_\ell(q)$ is a product of i.i.d. random matrices whose growth rate is governed by $\bar{J}(q)$ up to fluctuation corrections; this is the regime of the homogenized and stochastic limits studied by Koubbi et al. (2026) and Fedorov et al. (2026), who find that resampled weights mitigate but do not remove collapse. Second, the tied setting is itself of independent interest: looped and weight-tied Transformers (Dehghani et al., 2019; Lan et al., 2020; Bai et al., 2019; Giannou et al., 2023; Yang et al., 2024; Xu & Sato, 2025; Geiping et al., 2026; Jeddi et al., 2026; Li et al., 2026a; Zhang et al., 2026) apply one block repeatedly, and their initialization and stability (Li et al., 2026a; Tomihari & Karakida, 2026) is exactly the problem studied here. An untied control experiment on CIFAR-10 is left for future work.

B CUBIC EXPANSION AND MINIMAL NONLINEAR CLOSURE

All scalar nonlinear results in Sec. 5 start from the same local expansion of Eq. equation 13. We derive it once, then use the linear sectors of Sec. 4 to determine which additional modes must be retained in each case. This makes the reduced ansatz a consequence of nonlinear mode generation rather than an independent assumption.

For the remainder estimates below we assume ϕ is C^5 in a neighborhood of the origin; the displayed coefficients depend only on its first three derivatives. Let $A_0 := A(0)$ and

$$L := I + \eta A_0, \qquad J(q) = \alpha_1 L_q, \qquad L_q := 1 + \eta\lambda_q, \tag{70}$$

where the λ_q are the eigenvalues of A_0. For vectors, $x^{\circ m}$ denotes component-wise powers and $u \circ v$ the Hadamard product; $\Pi_q y$ denotes the Fourier coefficient of y in sector q.

Lemma B.1 (Third-order expansion of the scalar Transformer map)**.** *For the scalar map in Eq. 13 around $x = 0$,*

$$F(x) = \alpha_1 Lx + \alpha_2 (Lx)^{\circ 2} + \alpha_3 (Lx)^{\circ 3} + \alpha_1 \eta\chi\, \mathcal{C}(x) + O(\|x\|^4), \tag{71}$$

where, writing $x = \sum_q \hat{x}_q e_q$, the first-order term $\alpha_1 Lx = \sum_q J(q)\, \hat{x}_q\, e_q$ is the linear map of Theorem 4.1, $C_\star(I + A_\star V)$, with $C_\star = \alpha_1$, $V = \eta$ and $A_\star = A_0$; it acts on each Fourier mode through the scalar dispersion relation $J(q)$. The content-attention term is

$$\mathcal{C}(x) := x \circ \left[A_0(x^{\circ 2}) - (A_0 x)^{\circ 2}\right]. \tag{72}$$

In particular, the state dependence of attention first enters at cubic order at the zero state.

Proof. Row by row,

$$A_{ij}(x) = \frac{(A_0)_{ij} \mathrm{e}^{\chi x_i x_j}}{\sum_m (A_0)_{im} \mathrm{e}^{\chi x_i x_m}}. \tag{73}$$

Since $\chi x_i x_j = O(\|x\|^2)$, expansion of the exponential and row normalizer gives

$$A_{ij}(x) = (A_0)_{ij} \left[1 + \chi x_i x_j - \chi x_i \sum_m (A_0)_{im} x_m + O(\|x\|^4)\right]. \tag{74}$$

Multiplying by x_j and summing,

$$A(x)x = A_0 x + \chi \mathcal{C}(x) + O(\|x\|^5). \tag{75}$$

Hence $H(x) = Lx + \eta\chi\mathcal{C}(x) + O(\|x\|^5)$. Taylor expansion of $F(x) = H(x) + \nu\phi(H(x))$ gives Eq. 71. The content logit is quadratic at $x = 0$, so no attention-dependent quadratic term appears. □

B.1 From linear sectors to the minimal nonlinear ansatz

At linear order the Fourier sectors are independent: mode q is multiplied by $J(q)$. Nonlinear products couple them by wave-vector addition. For a simple critical pair $\pm q_c$, write only its linear contribution as

$$x_c = B\, e_{q_c} + \overline{B}\, e_{-q_c}. \tag{76}$$

The quadratic term in Lemma B.1 generates only

$$0, \quad \pm 2q_c, \tag{77}$$

whereas the direct cubic terms generate $\pm q_c$ and $\pm 3q_c$. We call the pair *non-resonant* if $4q_c \not\equiv 0 \pmod{2\pi}$, so that 0, $\pm 2q_c$ and $\pm 3q_c$ are distinct from $\pm q_c$ (this excludes $q_c = \pi$), and if $|J(0)|, |J(2q_c)| < 1$. Under the non-resonance assumptions used below, the mean m and second harmonic are linearly stable. To obtain the equation for B through cubic order, it is therefore sufficient to retain their $O(|B|^2)$ response:

$$x_j = B\mathrm{e}^{\mathrm{i}q_c j} + \overline{B}\mathrm{e}^{-\mathrm{i}q_c j} + m + Z\mathrm{e}^{2\mathrm{i}q_c j} + \overline{Z}\mathrm{e}^{-2\mathrm{i}q_c j} + O(|B|^3), \tag{78}$$

where Z is the quadratic amplitude factor. The third harmonic is indeed generated at $O(|B|^3)$, but its feedback into the critical sector occurs only beyond cubic order, so it does not need to be included in the minimal closure.

C Clusters: proof of Theorem 5.1

We first state the full version of Theorem 5.1.

Theorem C.1 (Clusters, full version of Theorem 5.1)**.** *Suppose there is no PE, N is even and $\eta \in (-2, 0)$, so that the homogeneous direction is stable, while the heterogeneous sector $\mathbf{1}^\perp$ loses stability through $J(q \neq 0) = \alpha_1 = 1 + \mu$. For a balanced two-cluster pattern $x = u\, s + m\mathbf{1}$, $s_i \in \{\pm 1\}$ and $\sum_i s_i = 0$, the only quadratic mode generated by us is the homogeneous one. It is bound[3] to the cluster amplitude as*

$$m = -\frac{\alpha_2}{\eta} u^2 + O(\mu u^2, u^4), \tag{79}$$

and

$$u^+ - u = \mu u - \beta_{\mathrm{cl}} u^3 + O(\mu u^3, u^5), \qquad \beta_{\mathrm{cl}} = -\eta\chi - \alpha_3 + \frac{2(1+\eta)}{\eta}\alpha_2^2. \tag{80}$$

If $\mu > 0$ and $\beta_{\mathrm{cl}} > 0$, then $u_\star^2 = \mu/\beta_{\mathrm{cl}} + O(\mu^2)$ and the branch is stable along the cluster line. Intra-cluster perturbations have multipliers $1 + \mu \pm 2\alpha_2 u_\star + O(\mu)$, so the balanced state is stable near threshold only if $\alpha_2 = 0$; in that case it is stable if $\alpha_3 < 0$ (and marginal at cubic order when $\alpha_3 = 0$).

Without PE,

$$A_0 = \frac{1}{N}\mathbf{1}\mathbf{1}^\top, \qquad L = I + \frac{\eta}{N}\mathbf{1}\mathbf{1}^\top. \tag{81}$$

Thus $Lx = x$ on $\mathbf{1}^\perp$ and $L\mathbf{1} = L_0\mathbf{1}$ with $L_0 = 1 + \eta$. The heterogeneous multiplier is $J(q \neq 0) = \alpha_1 = 1 + \mu$, whereas the homogeneous multiplier $J(0) = \alpha_1 L_0$ remains strictly inside the unit circle. The critical sector is degenerate, so we reduce on the balanced two-cluster direction and then check perturbations transverse to it.

Proof. Generated sector and slaving. Let $s_i \in \{\pm 1\}$ with $\sum_i s_i = 0$. Since $s^{\circ 2} = \mathbf{1}$, the quadratic FFN term generated by the critical direction us has no heterogeneous component: it drives only the homogeneous sector. The minimal closure is therefore

$$x = us + m\mathbf{1}, \qquad m = O(u^2). \tag{82}$$

For $x = us$, balanced clusters give the exact attention weights

$$A_{ij}(us) = \frac{1}{N}\big(1 + \tanh(\chi u^2) s_i s_j\big), \tag{83}$$

[3] Haken (1996) refers to this phenomenon as "slaving".

so $A(us)(us) = \chi u^3 s + O(u^5)$. Since $m = O(u^2)$, its effect on the attention weights enters only beyond cubic order.

Before separating the two sectors, write the whole cubic update on the minimal subspace. Because

$$L(us + m\mathbf{1}) = us + L_0 m\mathbf{1}, \tag{84}$$

we have, through cubic order,

$$(Lx)^{\circ 2} = u^2\mathbf{1} + 2L_0 mu\, s + O(u^4), \tag{85}$$

$$(Lx)^{\circ 3} = u^3 s + O(u^4), \tag{86}$$

$$\mathcal{C}(x) = u^3 s + O(u^4). \tag{87}$$

Hence Lemma B.1 gives the unprojected expression

$$\begin{aligned} F(us + m\mathbf{1}) = &\Big[J(0)m + \alpha_2 u^2\Big]\mathbf{1} \\ &+ \Big[\alpha_1 u + 2\alpha_2 L_0 mu + (\alpha_3 + \alpha_1\eta\chi)u^3\Big]s + O(\mu u^3, u^4). \end{aligned} \tag{88}$$

The two displayed directions are linearly distinct: the first is the stable homogeneous sector and the second lies in the critical heterogeneous sector. Reading the coefficient of $\mathbf{1}$ therefore gives

$$m^+ = J(0)m + \alpha_2 u^2 + O(\mu u^2, u^4). \tag{89}$$

Because $|J(0)| < 1$ while u changes only by $O(\mu u, u^3)$, the mean follows the slowly varying cluster amplitude:

$$m = \frac{\alpha_2}{1 - J(0)} u^2 + O(u^4) = -\frac{\alpha_2}{\eta} u^2 + O(\mu u^2, u^4), \tag{90}$$

where the second equality evaluates the stable resolvent at onset.

Projection onto the cluster amplitude. We now read the coefficient of s in Eq. equation 88. The three cubic contributions are visible before the projection: content attention, the local cubic FFN, and the quadratic FFN evaluated on the slaved mean. Using $\alpha_1 = 1 + \mu$ and absorbing the $O(\mu u^3)$ correction gives

$$\begin{aligned} u^+ - u &= \mu u + \left[\underbrace{\eta\chi}_{\text{content attention}} + \underbrace{\alpha_3}_{\text{cubic FFN}} + \underbrace{2\alpha_2(1+\eta)\frac{m}{u^2}}_{\text{slaved mean}}\right] u^3 + O(\mu u^3, u^5) \\ &= \mu u + \left(\eta\chi + \alpha_3 - \frac{2(1+\eta)}{\eta}\alpha_2^2\right) u^3 + O(\mu u^3, u^5) = \mu u - \beta_{\text{cl}} u^3 + O(\mu u^3, u^5), \end{aligned} \tag{91}$$

which is Eq. 80. For $\mu > 0$ and $\beta_{\text{cl}} > 0$, $u_\star^2 = \mu/\beta_{\text{cl}} + O(\mu^2)$, and the multiplier along the cluster line is $1 - 2\mu + O(\mu^2)$.

Transverse stability. Because the entire zero-mean space is critical without PE, retain all of $\mathbf{1}^\perp$. For a perturbation $\delta \in \{\mathbf{1}, s\}^\perp$ at $x_\star = u_\star s + m_\star \mathbf{1}$, Eq. equation 72 gives $D\mathcal{C}(x_\star)\delta = u_\star^2\delta$, and $L\delta = \delta$. Hence

$$DF(x_\star)\delta = \alpha_1\delta + 2\alpha_2 u_\star\, s \circ \delta + 2\alpha_2 L_0 m_\star\delta + 3\alpha_3 u_\star^2\delta + \alpha_1\eta\chi u_\star^2\delta + O(u_\star^3)\delta. \tag{92}$$

The involution $\delta \mapsto s \circ \delta$ has eigenvalues ± 1 on this transverse space, so

$$\Lambda_\perp^\pm = 1 + \mu \pm 2\alpha_2 u_\star + (2\alpha_3 - \beta_{\text{cl}})u_\star^2 + O(u_\star^3). \tag{93}$$

If $\alpha_2 \neq 0$, one multiplier exceeds one because $u_\star = O(\mu^{1/2}) \gg \mu$. If $\alpha_2 = 0$,

$$\Lambda_\perp = 1 + \frac{2\alpha_3}{\beta_{\text{cl}}}\mu + O(\mu^2), \tag{94}$$

which is below one for $\beta_{\text{cl}} > 0$ exactly when $\alpha_3 < 0$; $\alpha_3 = 0$ is marginal at cubic order. This proves the theorem. □

D STANDING WAVES

The standing-wave case is the cleanest illustration of the general mechanism: one critical Fourier pair quadratically generates a mean and a second harmonic, both of which are linearly stable and therefore slaved.

Proposition D.1 (Standing waves)**.** *Assume a reflection-symmetric scalar positional kernel and a simple non-resonant critical pair $\pm q_c$ with*

$$J(q_c) = 1 + \mu, \qquad |J(q)| < 1 \quad (q \notin \{\pm q_c\}) \tag{95}$$

near $\mu = 0$. Then the critical Fourier coefficient obeys

$$B^+ - B = \mu B - \beta_{\mathrm{st}} |B|^2 B + O(\mu |B|^3, |B|^5), \tag{96}$$

where, with $\lambda_c := \lambda_{q_c}$, $\lambda_2 := \lambda_{2q_c}$ and $L_c := L_{q_c}$,

$$\begin{aligned} -\beta_{\mathrm{st}} = {} & \alpha_1 \eta \chi (2 + \lambda_2 - 3\lambda_c^2) + 3\alpha_3 L_c^3 \\ & + 2\alpha_2^2 L_c^3 \left[\frac{2L_0}{1 - J(0)} + \frac{L_{2q_c}}{1 - J(2q_c)} \right], \end{aligned} \tag{97}$$

all quantities on the right being evaluated at $\mu = 0$. If $\mu > 0$ and $\beta_{\mathrm{st}} > 0$, then $|B_\star|^2 = \mu/\beta_{\mathrm{st}} + O(\mu^2)$; the phase is stationary.

Proof. Reflection symmetry makes all λ_q and $J(q)$ real. Let $L_2 := L_{2q_c}$.

Generated sectors and slaving. The quadratic product of the critical pair generates only the mean and second harmonic. We therefore truncate the invariant graph at quadratic order,

$$x_j^{[2]} = B \mathrm{e}^{\mathrm{i} q_c j} + \overline{B} \mathrm{e}^{-\mathrm{i} q_c j} + m + Z \mathrm{e}^{2\mathrm{i} q_c j} + \overline{Z} \mathrm{e}^{-2\mathrm{i} q_c j}. \tag{98}$$

The omitted cubic stable correction contains, in particular, the generated third harmonic, but it cannot feed back into q_c at cubic order. Here all L_q and λ_q are real. Substituting this truncation into Lemma B.1 and collecting equal Fourier sectors gives the complete cubic update before any sector is discarded:

$$\begin{aligned} F(x^{[2]}) = {} & \Big[J(0) m + 2\alpha_2 L_c^2 |B|^2 \Big] e_0 \\ & + \Big\{ \Big[J(q_c) B + 2\alpha_2 L_c \big(L_0 m B + L_2 Z \overline{B} \big) + 3\alpha_3 L_c^3 |B|^2 B + \alpha_1 \eta \chi (2 + \lambda_2 - 3\lambda_c^2) |B|^2 B \Big] e_{q_c} + \text{c.c.} \Big\} \\ & + \Big\{ \big[J(2q_c) Z + \alpha_2 L_c^2 B^2 \big] e_{2q_c} + \text{c.c.} \Big\} \\ & + \Big\{ \Big[2\alpha_2 L_c L_2 B Z + \big(\alpha_3 L_c^3 + \alpha_1 \eta \chi (\lambda_2 - \lambda_c^2) \big) B^3 \Big] e_{3q_c} + \text{c.c.} \Big\} + O(|B|^4). \end{aligned} \tag{99}$$

This makes the bookkeeping visible. The 0 and $2q_c$ sectors are forced at quadratic order and are stable; q_c receives the cubic feedback that controls saturation; and $3q_c$ is generated at cubic order but cannot return to q_c at the same order.

Reading the 0 and $2q_c$ lines of Eq. equation 99,

$$m^+ = J(0) m + 2\alpha_2 L_c^2 |B|^2 + O(|B|^4), \qquad Z^+ = J(2q_c) Z + \alpha_2 L_c^2 B^2 + O(|B|^4). \tag{100}$$

At stationary onset $J(q_c) = 1$, so

$$m = \frac{2\alpha_2 L_c^2}{1 - J(0)} |B|^2 + O(|B|^4), \qquad Z = \frac{\alpha_2 L_c^2}{1 - J(2q_c)} B^2 + O(|B|^4). \tag{101}$$

The two denominators are exactly the stable linear resolvents of the generated sectors.

Projection onto the critical mode. We now read the q_c line of Eq. equation 99. In projection notation, its three nonlinear pieces are

$$\Pi_{q_c} \mathcal{C}(x_c) = (2 + \lambda_2 - 3\lambda_c^2) |B|^2 B, \tag{102}$$

$$\Pi_{q_c} (L x_c)^{\circ 3} = 3 L_c^3 |B|^2 B, \tag{103}$$

$$\Pi_{q_c} \big[(Lx)^{\circ 2} - (L x_c)^{\circ 2} \big] = 2 L_c \big(L_0 m B + L_2 Z \overline{B} \big) + O(|B|^4). \tag{104}$$

Thus the projections are simply a compact way of reading terms that are already explicit in Eq. equation 99. Substituting Eq. equation 101 into the last line and weighting the three pieces by the coefficients in Lemma B.1 gives

$$-\beta_{\mathrm{st}} = \underbrace{\alpha_1\eta\chi(2+\lambda_2-3\lambda_c^2)}_{\text{content attention}} + \underbrace{3\alpha_3 L_c^3}_{\text{cubic FFN}} + \underbrace{2\alpha_2^2 L_c^3\left[\frac{2L_0}{1-J(0)}+\frac{L_2}{1-J(2q_c)}\right]}_{\text{slaved mean and second harmonic}}, \tag{105}$$

which is Eq. equation 97. Hence Eq. equation 96 follows. All coefficients are real, so the phase is stationary. On the nonzero branch the radial multiplier is $1-2\mu+O(\mu^2)$, while the slaved sectors retain multipliers $J(0)+O(\mu)$ and $J(2q_c)+O(\mu)$ strictly inside the unit circle. □

E TRAVELING WAVES: PROOF OF THEOREM 5.2

We first state the full version of Theorem 5.2.

Theorem E.1 (Traveling waves, full version of Theorem 5.2)**.** *Assume a simple non-resonant critical pair $\pm q_c$ for a directional kernel. Write its scalar dispersion gain as*

$$J(q_c) = (1+\mu)\mathrm{e}^{\mathrm{i}\omega_c}, \tag{106}$$

with $\mu = 0$ at onset and $\omega_c \neq 0$, and no strong resonance: $\mathrm{e}^{\mathrm{i}k\omega_c} \neq 1$ for $k = 1,\dots,4$. If B_ℓ is the critical Fourier coefficient and $B_\ell = \mathrm{e}^{\mathrm{i}\ell\omega_c}G_\ell$, then

$$G^+ - G = \mu G - \beta_{\mathrm{tr}}|G|^2 G + O(\mu|G|^3, |G|^5), \tag{107}$$

where the explicit coefficient $\beta_{\mathrm{tr}} = -\mathrm{e}^{-\mathrm{i}\omega_c}\beta_{q_c}^{\mathrm{tr}}$ is derived below, with $\beta_{q_c}^{\mathrm{tr}}$ in Eq. 117. The second harmonic is bound in a frame rotating at $2\omega_c$. If $\mu > 0$ and $\Re(\beta_{\mathrm{tr}}) > 0$, the amplitude saturates at

$$R_\star^2 = \frac{\mu}{\Re(\beta_{\mathrm{tr}})} + O(\mu^2), \qquad \Omega = \omega_c - \Im(\beta_{\mathrm{tr}})R_\star^2 + O(R_\star^4). \tag{108}$$

Thus the real part of β_{tr} controls saturation, while its imaginary part gives the nonlinear correction to the propagation speed.

For a directional kernel, $\lambda_{-q} = \overline{\lambda_q}$ and $L_{-q} = \overline{L_q}$. At onset let

$$J_c := J(q_c) = \mathrm{e}^{\mathrm{i}\omega_c}, \qquad \lambda_c := \lambda_{q_c}, \quad \lambda_2 := \lambda_{2q_c}, \quad L_c := L_{q_c}, \quad L_2 := L_{2q_c}. \tag{109}$$

The generated sectors are the same as for a standing wave, but the second harmonic now rotates twice as fast as the critical mode.

Proof. The reduction is identical to the standing-wave calculation except for the linear phase accumulated by the generated harmonics.

Generated sectors and slaving. Use the quadratic truncation

$$x_j^{[2]} = B\mathrm{e}^{\mathrm{i}q_c j} + \overline{B}\mathrm{e}^{-\mathrm{i}q_c j} + m + Z\mathrm{e}^{2\mathrm{i}q_c j} + \overline{Z}\mathrm{e}^{-2\mathrm{i}q_c j}. \tag{110}$$

As above, the omitted cubic stable correction cannot feed back into q_c at cubic order. Now $L_{-q_c} = \overline{L}_c$ and $L_{-2q_c} = \overline{L}_2$. The full cubic update of $x^{[2]}$, before projecting onto any Fourier sector, is

$$\begin{aligned}
F(x^{[2]}) = {} & \Big[J(0)m + 2\alpha_2|L_c|^2|B|^2\Big]e_0 \\
& + \Big\{\Big[J_c B + 2\alpha_2\big(L_c L_0 m B + \overline{L}_c L_2 Z\overline{B}\big) + 3\alpha_3|L_c|^2 L_c|B|^2 B \\
& \qquad + \alpha_1\eta\chi\big(2(1-|\lambda_c|^2) + \lambda_2 - \lambda_c^2\big)|B|^2 B\Big]e_{q_c} + \text{c.c.}\Big\} \\
& + \Big\{\big[J(2q_c)Z + \alpha_2 L_c^2 B^2\big]e_{2q_c} + \text{c.c.}\Big\} \\
& + \Big\{\Big[2\alpha_2 L_c L_2 B Z + \big(\alpha_3 L_c^3 + \alpha_1\eta\chi(\lambda_2 - \lambda_c^2)\big)B^3\Big]e_{3q_c} + \text{c.c.}\Big\} + O(|B|^4).
\end{aligned} \tag{111}$$

The generated sectors are therefore exactly the same as in the standing case. What changes is their layer-to-layer phase: the mean is stationary, while the $2q_c$ coefficient forced by B^2 rotates with $J_c^2 = e^{2\mathrm{i}\omega_c}$.

Reading the first and third lines of Eq. equation 111 gives

$$m^+ = J(0)m + 2\alpha_2|L_c|^2|B|^2 + O(|B|^4), \qquad Z^+ = J(2q_c)Z + \alpha_2 L_c^2 B^2 + O(|B|^4). \tag{112}$$

Hence

$$m = \frac{2\alpha_2|L_c|^2}{1-J(0)}|B|^2 + O(|B|^4), \qquad Z = \frac{\alpha_2 L_c^2}{\mathrm{e}^{2\mathrm{i}\omega_c} - J(2q_c)}B^2 + O(|B|^4). \tag{113}$$

Relative to Eq. equation 101, only the second-harmonic resolvent changes: it is evaluated in the frame rotating at twice the critical phase.

Projection onto the critical mode. We now read the q_c line of Eq. equation 111. The same three mechanisms contribute, and in projection notation they are

$$\Pi_{q_c}\mathcal{C}(x_c) = \left[2(1-|\lambda_c|^2) + \lambda_2 - \lambda_c^2\right]|B|^2 B, \tag{114}$$

$$\Pi_{q_c}(Lx_c)^{\circ 3} = 3|L_c|^2 L_c |B|^2 B, \tag{115}$$

$$\Pi_{q_c}\left[(Lx)^{\circ 2} - (Lx_c)^{\circ 2}\right] = 2\left(L_c L_0 m B + \overline{L}_c L_2 Z\overline{B}\right) + O(|B|^4). \tag{116}$$

The complex conjugate on L_c in the second-harmonic feedback is essential: it is the $(-q_c)+(2q_c) = q_c$ interaction. Using Eq. equation 113 therefore gives, at onset,

$$\begin{aligned}\beta_{q_c}^{\mathrm{tr}} = {} & \underbrace{\alpha_1\eta\chi\left[2(1-|\lambda_c|^2) + \lambda_2 - \lambda_c^2\right]}_{\text{content attention}} + \underbrace{3\alpha_3|L_c|^2 L_c}_{\text{cubic FFN}} \\ & + \underbrace{2\alpha_2^2 L_c|L_c|^2\left[\frac{2L_0}{1-J(0)} + \frac{L_2}{\mathrm{e}^{2\mathrm{i}\omega_c} - J(2q_c)}\right]}_{\text{slaved mean and second harmonic}}.\end{aligned} \tag{117}$$

Thus

$$B^+ = J(q_c)B + \beta_{q_c}^{\mathrm{tr}}|B|^2 B + O(\mu|B|^3, |B|^5). \tag{118}$$

Writing $J(q_c) = (1+\mu)\mathrm{e}^{\mathrm{i}\omega_c}$ and $B_\ell = \mathrm{e}^{\mathrm{i}\ell\omega_c}G_\ell$ removes the linear rotation:

$$G^+ - G = \mu G + \beta_{\mathrm{tr}}|G|^2 G + O(\mu|G|^3, |G|^5), \qquad \beta_{\mathrm{tr}} := \mathrm{e}^{-\mathrm{i}\omega_c}\beta_{q_c}^{\mathrm{tr}}. \tag{119}$$

Writing $G = R\mathrm{e}^{\mathrm{i}\theta}$ then gives

$$R^+ - R = \mu R - \Re(\beta_{\mathrm{tr}})R^3 + O(\mu R^3, R^5), \qquad \theta^+ - \theta = -\Im(\beta_{\mathrm{tr}})R^2 + O(R^4), \tag{120}$$

from which Eq. 108 follows. For a joint token–feature mode, projection onto the corresponding left eigenvector gives the same co-rotating normal form; a complex right eigenvector produces the feature-plane rotation described by Eq. equation 11. □

F Fixed points away from the origin

The previous analysis assumes $a_\star = 0$. The same reduction applies around a nonzero homogeneous fixed point $a_\star \neq 0$, as may arise with biases or LayerNorm. The only qualitative difference is that content attention then contributes already at quadratic order.

Write

$$x = a_\star\mathbf{1} + u, \qquad y_\star = (1+\eta)a_\star, \qquad A_0 := A(a_\star\mathbf{1}), \qquad L := I + \eta A_0,$$

with

$$\alpha_k := \delta_{k1} + \frac{r}{k!}\phi^{(k)}(y_\star).$$

For each attention row define

$$\kappa_{2,i}(u) := (A_0 u^{\circ 2})_i - (A_0 u)_i^2,$$

and

$$\kappa_{3,i}(u) := (A_0 u^{\circ 3})_i - 3(A_0 u^{\circ 2})_i (A_0 u)_i + 2(A_0 u)_i^3.$$

Proof sketch. For a fixed row i,

$$\chi x_i x_j = \underbrace{\chi a_\star (a_\star + u_i)}_{\text{independent of } j} + \chi(a_\star + u_i) u_j,$$

so the first term cancels from the row softmax. The perturbed attention row is therefore an exponential tilt of A_0. Expanding its mean in cumulants gives

$$[A(x)x]_i = a_\star + (A_0 u)_i + \chi a_\star \kappa_{2,i}(u) + \chi u_i \kappa_{2,i}(u) + \frac{1}{2}(\chi a_\star)^2 \kappa_{3,i}(u) + O(\|u\|^4).$$

Expanding the FFN around $y_\star$ then yields

$$u^+ = \alpha_1 L u + \mathcal{N}_2(u) + \mathcal{N}_3(u) + O(\|u\|^4),$$

with

$$\mathcal{N}_2(u) = \alpha_2 (Lu)^{\circ 2} + \alpha_1 \eta \chi a_\star \kappa_2(u),$$

and

$$\mathcal{N}_3(u) = \alpha_3 (Lu)^{\circ 3} + \alpha_1 \eta \chi\, u \circ \kappa_2(u) + \frac{1}{2}\alpha_1 \eta (\chi a_\star)^2 \kappa_3(u) + 2\alpha_2 \eta \chi a_\star (Lu) \circ \kappa_2(u).$$

Hence the center-manifold reduction is unchanged: the linear dispersion relation still determines the critical modes, while nonlinear interactions generate stable harmonics that slave to them. Compared with $a_\star = 0$, the new term

$$\alpha_1 \eta \chi a_\star \kappa_2(u)$$

adds an attention-mediated quadratic forcing of the slaved modes. For example, for a standing critical pair $\pm q_c$, the mean and second-harmonic drives become

$$p_0 = \alpha_2 L_c^2 + \alpha_1 \eta \chi a_\star (1 - \lambda_c^2), \qquad p_2 = \alpha_2 L_c^2 + \alpha_1 \eta \chi a_\star (\lambda_2 - \lambda_c^2).$$

The same slaving and projection steps therefore give

$$B^+ - B = \mu B + \beta_{q_c}^{\mathrm{st}} |B|^2 B + \cdots$$

with modified coefficients. Setting $a_\star = 0$ recovers the previous analysis.

G Proof of Theorem 5.3

We first state the full version of Theorem 5.3.

Theorem G.1 (Two-mode competition, full version of Theorem 5.3)**.** *In the reflection-symmetric scalar reduction, let two distinct standing modes q_1, q_2 be near critical, with all other sectors generated up to cubic order strictly stable and with no resonances $n_1q_1 + n_2q_2 \equiv 0 \pmod N$ for $0 < |n_1| + |n_2| \le 4$. Their amplitudes obey*

$$B_1^+ - B_1 = (\mu_1 + \beta_{11}|B_1|^2 + \beta_{12}|B_2|^2)B_1 + \cdots,$$
$$B_2^+ - B_2 = (\mu_2 + \beta_{22}|B_2|^2 + \beta_{21}|B_1|^2)B_2 + \cdots, \tag{121}$$

where $\mu_i = J(q_i) - 1$, and the self-coupling β_{ii} and cross-coupling β_{ij}, negative when saturating and suppressing respectively, are computed below. At simultaneous onset the cross-couplings are symmetric to cubic order, $\beta_{12} = \beta_{21}$, so the moduli $R_i = |B_i|$ obey

$$R_i^+ - R_i = -\frac{\partial \mathcal{V}}{\partial R_i}, \qquad \mathcal{V}(R_1, R_2) = -\sum_{i=1,2} \frac{\mu_i}{2} R_i^2 - \sum_{i=1,2} \frac{\beta_{ii}}{4} R_i^4 - \frac{\beta_{12}}{2} R_1^2 R_2^2. \tag{122}$$

In the symmetric case $\mu_i = \mu > 0$, $\beta_{ii} = -\beta_{\mathrm{self}}$ and $\beta_{ij} = -\beta_{\mathrm{cross}}$ with $\beta_{\mathrm{self}} > 0$: if $\beta_{\mathrm{cross}} < \beta_{\mathrm{self}}$, the mixed state $R_1^2 = R_2^2 = \mu/(\beta_{\mathrm{self}} + \beta_{\mathrm{cross}})$ is stable and the two modes coexist; if $\beta_{\mathrm{cross}} > \beta_{\mathrm{self}}$, the two single-mode states $R_i^2 = \mu/\beta_{\mathrm{self}}$ are stable and the initial condition selects the winner.

Assume a reflection-symmetric scalar kernel and two distinct standing modes q_1, q_2 simultaneously near onset, all other sectors being strictly stable. Write

$$x_j = B_1 \mathrm{e}^{\mathrm{i}q_1 j} + \overline{B}_1 \mathrm{e}^{-\mathrm{i}q_1 j} + B_2 \mathrm{e}^{\mathrm{i}q_2 j} + \overline{B}_2 \mathrm{e}^{-\mathrm{i}q_2 j} + O(|B|^2). \tag{123}$$

The non-resonance condition of Theorem G.1 ensures that the quadratic products generate only stable sectors

$$0, \qquad \pm 2q_1, \qquad \pm 2q_2, \qquad \pm(q_1 + q_2), \qquad \pm(q_1 - q_2), \tag{124}$$

and that no other cubic wave-vector identity folds into $\pm q_1$ or $\pm q_2$.

Proof. Generated sectors and slaving. Define the critical part

$$x_c := B_1 e_{q_1} + \overline{B}_1 e_{-q_1} + B_2 e_{q_2} + \overline{B}_2 e_{-q_2}. \tag{125}$$

Let $L_i := L_{q_i}$ and write the $O(|B|^2)$ stable correction explicitly as

$$\begin{aligned} x_s = m &+ Z_1 e_{2q_1} + \overline{Z}_1 e_{-2q_1} + Z_2 e_{2q_2} + \overline{Z}_2 e_{-2q_2} \\ &+ W_+ e_{q_1+q_2} + \overline{W}_+ e_{-(q_1+q_2)} + W_- e_{q_1-q_2} + \overline{W}_- e_{-(q_1-q_2)}. \end{aligned} \tag{126}$$

Before projecting anything, the quadratic product of the two critical pairs is

$$\begin{aligned} (Lx_c)^{\circ 2} = {} & 2\big(L_1^2|B_1|^2 + L_2^2|B_2|^2\big)e_0 \\ & + \Big[L_1^2 B_1^2 e_{2q_1} + L_2^2 B_2^2 e_{2q_2} + 2L_1L_2B_1B_2 e_{q_1+q_2} + 2L_1L_2B_1\overline{B}_2 e_{q_1-q_2} + \text{c.c.}\Big]. \end{aligned} \tag{127}$$

Thus the list of stable sectors in the theorem is read directly from the unprojected quadratic term. Through cubic order the full map is

$$\begin{aligned} F(x_c + x_s) = {} & \alpha_1 L(x_c + x_s) + \alpha_2 (Lx_c)^{\circ 2} + 2\alpha_2 (Lx_c) \circ (Lx_s) \\ & + \alpha_3 (Lx_c)^{\circ 3} + \alpha_1 \eta\chi\, \mathcal{C}(x_c) + O(|B|^4). \end{aligned} \tag{128}$$

The two direct cubic terms are supported on

$$\pm q_1,\ \pm q_2,\ \pm 3q_1,\ \pm 3q_2,\ \pm(2q_1 \pm q_2),\ \pm(q_1 \pm 2q_2). \tag{129}$$

The non-resonance assumption ensures that none of the noncritical entries in this list folds back onto $\pm q_1$ or $\pm q_2$. Therefore the quadratically generated modes in Eq. equation 126 are the only stable modes that can feed back into a critical equation at cubic order.

At simultaneous standing onset $J(q_i) = 1$. Reading the coefficients of the quadratically generated sectors in Eq. equation 128 gives

$$m = \frac{2\alpha_2 \left(L_1^2|B_1|^2 + L_2^2|B_2|^2\right)}{1 - J(0)}, \tag{130}$$

$$Z_i = \frac{\alpha_2 L_i^2}{1 - J(2q_i)} B_i^2, \tag{131}$$

$$W_+ = \frac{2\alpha_2 L_1 L_2}{1 - J(q_1 + q_2)} B_1 B_2, \qquad W_- = \frac{2\alpha_2 L_1 L_2}{1 - J(q_1 - q_2)} B_1 \overline{B}_2. \tag{132}$$

The self-interaction of mode q_i is exactly the standing-wave calculation of Proposition D.1; its cubic coefficient is $\beta_{ii} = -\beta_{\mathrm{st}}(q_i)$, so that $\beta_{\mathrm{self}} = \beta_{\mathrm{st}}$.

Projection onto one critical mode. Fix $i \neq j$. To make the projection transparent, first display the terms of each unprojected polynomial in Eq. equation 128 that lie in sector q_i:

$$2(Lx_c) \circ (Lx_s) \supset 2\Big[L_i L_0 m B_i + L_i L_{2q_i} Z_i \overline{B}_i + L_j L_{q_i+q_j} W_+ \overline{B}_j + L_j L_{q_i-q_j} W_- B_j\Big] e_{q_i}, \tag{133}$$

$$(Lx_c)^{\circ 3} \supset \Big[3L_i^3 |B_i|^2 B_i + 6L_i L_j^2 |B_j|^2 B_i\Big] e_{q_i}, \tag{134}$$

$$\mathcal{C}(x_c) \supset \Big[(2 + \lambda_{2q_i} - 3\lambda_{q_i}^2)|B_i|^2 B_i + 2\big(1 - \lambda_{q_j}^2 + \lambda_{q_i+q_j} + \lambda_{q_i-q_j} - 2\lambda_{q_i}\lambda_{q_j}\big)|B_j|^2 B_i\Big] e_{q_i}. \tag{135}$$

All omitted terms lie in the noncritical cubic sectors listed above. Thus Π_{q_i} simply extracts the displayed coefficients. In particular, the direct cross pieces are

$$\Pi_{q_i}(Lx_c)^{\circ 3} \supset 6L_i L_j^2 |B_j|^2 B_i, \tag{136}$$

$$\Pi_{q_i}\mathcal{C}(x_c) \supset 2\left[1 - \lambda_{q_j}^2 + \lambda_{q_i+q_j} + \lambda_{q_i-q_j} - 2\lambda_{q_i}\lambda_{q_j}\right] |B_j|^2 B_i. \tag{137}$$

The mean and mixed harmonics in Eq. equation 132 feed back through the quadratic FFN. Their contribution is

$$\beta_{ij}^{\mathrm{slave}} = 4\alpha_2^2 L_i L_j^2 \left[\frac{L_0}{1 - J(0)} + \frac{L_{q_i+q_j}}{1 - J(q_i+q_j)} + \frac{L_{q_i-q_j}}{1 - J(q_i-q_j)}\right]. \tag{138}$$

Therefore the cross coefficient is the sum of direct attention, direct cubic FFN, and slaved feedback:

$$\beta_{ij} = \underbrace{2\alpha_1\eta\chi\left[1 - \lambda_{q_j}^2 + \lambda_{q_i+q_j} + \lambda_{q_i-q_j} - 2\lambda_{q_i}\lambda_{q_j}\right]}_{\text{content attention}} + \underbrace{6\alpha_3 L_j^2 L_i}_{\text{cubic FFN}} + \underbrace{4\alpha_2^2 L_i L_j^2 \left[\frac{L_0}{1 - J(0)} + \frac{L_{q_i+q_j}}{1 - J(q_i+q_j)} + \frac{L_{q_i-q_j}}{1 - J(q_i-q_j)}\right]}_{\text{slaved mean and mixed harmonics}}. \tag{139}$$

This gives Eq. 121.

Coexistence versus exclusion. At simultaneous onset $J(q_1) = J(q_2) = 1$ implies $L_1 = L_2$; because the symbol is even, Eq. equation 139 is symmetric under $i \leftrightarrow j$, so $\beta_{12} = \beta_{21}$ to cubic order. The radial dynamics is therefore the gradient map of Eq. equation 19. In the symmetric case $\mu_i = \mu > 0$, $\beta_{ii} = -\beta_{\mathrm{self}}$ and $\beta_{ij} = -\beta_{\mathrm{cross}}$ with $\beta_{\mathrm{self}} > 0$, the mixed state has $R_1^2 = R_2^2 = \mu/(\beta_{\mathrm{self}} + \beta_{\mathrm{cross}})$. Linearizing the radial increment there gives

$$-2\mu, \qquad -2\mu\,\frac{\beta_{\mathrm{self}} - \beta_{\mathrm{cross}}}{\beta_{\mathrm{self}} + \beta_{\mathrm{cross}}}, \tag{140}$$

so coexistence is stable for $\beta_{\mathrm{cross}} < \beta_{\mathrm{self}}$. A single-mode state, say $R_1^2 = \mu/\beta_{\mathrm{self}}$, has transverse multiplier

$$1 + \mu\left(1 - \frac{\beta_{\mathrm{cross}}}{\beta_{\mathrm{self}}}\right) + O(\mu^2), \tag{141}$$

so it is stable against invasion by mode 2 exactly when $\beta_{\mathrm{cross}} > \beta_{\mathrm{self}}$. By symmetry the same holds for the other single-mode state. □

H Effective potentials

The potential representation used in the main text follows directly from the cubic amplitude equations.

One mode. For a single critical mode, the amplitude equation in the co-rotating frame has the form

$$G^+ - G = \mu G - \Gamma |G|^2 G + \cdots . \tag{142}$$

Writing $G = R\mathrm{e}^{\mathrm{i}\theta}$ gives, to cubic order,

$$R^+ - R = \mu R - \Re(\Gamma) R^3 + \cdots , \tag{143}$$

while $\Im(\Gamma)$ produces the nonlinear phase drift. Defining

$$\beta := \Re(\Gamma), \tag{144}$$

the radial dynamics becomes

$$R^+ - R = -\frac{d\mathcal{V}}{dR}, \tag{145}$$

with

$$\mathcal{V}(R) = -\frac{\mu}{2} R^2 + \frac{\beta}{4} R^4. \tag{146}$$

For standing waves and clusters the cubic coefficient is real. For traveling waves, the potential describes only the radial dynamics, while the imaginary part of Γ controls the phase velocity.

Two competing modes. For two non-resonant standing modes, the amplitude equations are

$$R_1^+ - R_1 = \left(\mu_1 + \beta_{11} R_1^2 + \beta_{12} R_2^2\right) R_1 + \cdots , \tag{147}$$

$$R_2^+ - R_2 = \left(\mu_2 + \beta_{22} R_2^2 + \beta_{21} R_1^2\right) R_2 + \cdots . \tag{148}$$

At onset, in the reflection-symmetric setting considered here, the two critical modes have the same linear multiplier and the cross-coupling is symmetric,

$$\beta_{12} = \beta_{21}. \tag{149}$$

The radial dynamics can therefore be written as

$$R_i^+ - R_i = -\frac{\partial \mathcal{V}}{\partial R_i}, \tag{150}$$

with

$$\mathcal{V}(R_1, R_2) = -\frac{\mu_1}{2} R_1^2 - \frac{\mu_2}{2} R_2^2 - \frac{\beta_{11}}{4} R_1^4 - \frac{\beta_{22}}{4} R_2^4 - \frac{\beta_{12}}{2} R_1^2 R_2^2. \tag{151}$$

Now consider the symmetric saturating case

$$\mu_1 = \mu_2 = \mu > 0, \qquad \beta_{11} = \beta_{22} = -\beta_{\mathrm{self}}, \qquad \beta_{12} = -\beta_{\mathrm{cross}}, \tag{152}$$

with $\beta_{\mathrm{self}} > 0$. Then

$$\mathcal{V}(R_1, R_2) = -\frac{\mu}{2}(R_1^2 + R_2^2) + \frac{\beta_{\mathrm{self}}}{4}(R_1^4 + R_2^4) + \frac{\beta_{\mathrm{cross}}}{2} R_1^2 R_2^2. \tag{153}$$

Its stationary points satisfy

$$R_1 \left(-\mu + \beta_{\mathrm{self}} R_1^2 + \beta_{\mathrm{cross}} R_2^2\right) = 0, \tag{154}$$

$$R_2 \left(-\mu + \beta_{\mathrm{self}} R_2^2 + \beta_{\mathrm{cross}} R_1^2\right) = 0. \tag{155}$$

Besides the origin, there are two single-mode states,

$$(R_1^2, R_2^2) = \left(\frac{\mu}{\beta_{\mathrm{self}}}, 0\right), \qquad \left(0, \frac{\mu}{\beta_{\mathrm{self}}}\right), \tag{156}$$

and a mixed state

$$R_1^2 = R_2^2 = \frac{\mu}{\beta_{\mathrm{self}} + \beta_{\mathrm{cross}}}, \tag{157}$$

provided $\beta_{\mathrm{self}} + \beta_{\mathrm{cross}} > 0$.

At either single-mode state, the Hessian of $\mathcal{V}$ has eigenvalues

$$2\mu, \qquad \mu\left(\frac{\beta_{\mathrm{cross}}}{\beta_{\mathrm{self}}} - 1\right). \tag{158}$$

Therefore the single-mode states are minima when

$$\beta_{\text{cross}} > \beta_{\text{self}}. \tag{159}$$

In this regime, cross-suppression is stronger than self-saturation, so coexistence is unstable and either mode can suppress the other: the winner-takes-all regime.

At the mixed state, the Hessian eigenvalues are

$$2\mu, \qquad 2\mu\frac{\beta_{\text{self}} - \beta_{\text{cross}}}{\beta_{\text{self}} + \beta_{\text{cross}}}. \tag{160}$$

Hence the mixed state is a minimum when

$$\beta_{\text{cross}} < \beta_{\text{self}}, \tag{161}$$

corresponding to stable coexistence of the two modes.

The transition occurs at

$$\beta_{\text{cross}} = \beta_{\text{self}}, \tag{162}$$

where the cubic potential becomes degenerate along

$$R_1^2 + R_2^2 = \frac{\mu}{\beta_{\text{self}}}. \tag{163}$$

Finally, near threshold these minima coincide with the stable fixed points of the cubic amplitude map. Indeed, linearizing

$$R^+ - R = -\nabla\mathcal{V}(R) \tag{164}$$

around a stationary point $R_\star$ gives

$$\delta R^+ = \left[I - D^2\mathcal{V}(R_\star)\right]\delta R. \tag{165}$$

Since the Hessian eigenvalues are $O(\mu)$, for sufficiently small $\mu > 0$, a local minimum of $\mathcal{V}$ has all multipliers inside the unit circle. Thus the potential picture directly reproduces the coexistence and winner-takes-all regimes of the two-mode amplitude equations.

I Experiments

I.1 Effective-potential figure (Figure 2)

Figure 2 plots the effective potentials introduced in Sec. 5. For one critical mode, panels (a) and (b) show

$$\mathcal{V}(B) = -\frac{\mu}{2}B^2 + \frac{\beta}{4}B^4, \qquad B^+ - B = -\mathcal{V}'(B),$$

for $\mu = -1$ and $\mu = 1$, respectively, with $\beta = -1$ and $\beta = 1$ overlaid. The marked extrema illustrate stable and unstable amplitudes below and above the linear threshold. For two critical modes, panels (c) and (d) show schematic surfaces of

$$\mathcal{V}(R_1, R_2) = -\sum_{i=1}^{2}\frac{\mu_i}{2}R_i^2 + \sum_{i=1}^{2}\frac{\beta_{\text{self}}}{4}R_i^4 + \frac{\beta_{\text{cross}}}{2}R_1^2R_2^2,$$

with $\mu_1 = \mu_2 = \beta_{\text{self}} = 1$. When $\beta_{\text{cross}}/\beta_{\text{self}} = 0.4$, the wells lie between the coordinate axes and both modes coexist; when the ratio is 2.5, the wells lie on the axes and the dynamics is winner-takes-all. These schematic surfaces use signed amplitudes (R_1, R_2) to display the symmetry of the normal form.

I.2 Scalar-model phase diagrams (Figure 3)

Figure 3 is generated by iterating the scalar map in Eq. equation 13 on a ring of $N = 32$ tokens. Except for the cluster series in panel (a), the feed-forward branch is $r\tanh(H)$ with $r = -0.1$, so that

$$\alpha_1 = C_\star = 0.9, \qquad \alpha_2 = 0, \qquad \alpha_3 = -\frac{r}{3} = \frac{1}{30} > 0. \tag{166}$$

Thus the cubic FFN term is destabilizing, and saturation is supplied by the state dependence of attention through $\eta\chi$, with $\chi = 1.5$. We use $\eta < 0$ and plot its magnitude $|\eta|$. The cluster series has no PE, $\eta = -0.6$ and the cubic branch $\phi(y) = \mu y + \alpha_2 y^2 - y^3/3$ with $\alpha_2 = 0.3$ (and $r = 1$), so that $\alpha_1 = 1 + \mu$, $\alpha_2 = 0.3$ and $\alpha_3 = -1/3$: the heterogeneous sector is critical as in Theorem 5.1, the even part makes the slaved mean $m = O(\mu)$, and it is started from a balanced two-cluster seed.

The positional logits have the form

$$b_r = sK_\Delta(r), \tag{167}$$

where s is the peak positional logit, or PE strength. Panel (a) compares the numerical amplitudes and the square-root and linear growth in μ prediction obtained in Sec.5 for various representations: wave amplitude R, cluster size u, the coexisting modes $R_{1,2}$, the angular speed of the traveling wave $\Omega - w_c$ and finally cluster mean m. Panels (b) and (c) use symmetric two-lobe kernels $K_\Delta(r) = \mathrm{e}^{-(r-\Delta)^2/2} + \mathrm{e}^{-(r+\Delta)^2/2}$, with $\Delta = 4$ in panel (b), and $\eta = -0.6$ in panel (c). In panel (a), each series sweeps η at a fixed kernel so as to set μ: the standing wave uses $2K_4$ (critical mode k_c the minimizer of λ); the traveling wave uses the one-sided skewed lobe $\mathrm{e}^{-(r-4)^2/(2\zeta_\pm^2)}$, with $\zeta_- = 0.7$ for $r \le 4$ and $\zeta_+ = 2$ for $r > 4$, at $s = 1$, whose Fourier symbol is complex and therefore supports traveling modes; and the coexisting modes $k = 4, 10$ use $4[(1 - w_\star)K_2 + w_\star K_{4.5}]$ with lobe widths 0.7 and $w_\star$ tuned so that $\lambda_4 = \lambda_{10}$. Panel (d) uses the mixture $(1 - w)K_{\Delta=4} + wK_{\Delta=3}$, where the second lobe has width 0.7 and $|\eta| = 0.6$.

At each grid point, we iterate the map for 5000 layers. The initial state is a small random perturbation of amplitude 10^{-3}. In panel (d), we use two single-mode seeds

$$x_j^{(0)} = 0.05\cos(q_k j) + \text{noise}, \qquad k \in \{4, 5\}. \tag{168}$$

After iteration, we Fourier analyse the state. Here R is the modulus of the critical Fourier coefficient, Ω is its phase advance over the final layer, and k_{obs} is the dominant observed wavenumber. In panel (d), the competition outcome is determined from the final amplitudes of modes 4 and 5 for the two initial seeds; bistability means that each seed retains its own mode. Gray regions indicate parameters for which the homogeneous state is linearly stable.

The overlays compare these measurements with the predictions of the amplitude theory. The linear threshold is

$$\max_q |\Lambda(q)| = 1, \qquad \Lambda(q) = \alpha_1\big(1 + \eta\lambda(q)\big), \tag{169}$$

where $\lambda(q)$ is the Fourier symbol of the positional attention kernel. For a selected standing mode, the predicted saturated amplitude is

$$R_\star^2 = \frac{|\Lambda(q_c)| - 1}{\beta_{q_c}^{\mathrm{st}}}. \tag{170}$$

For the traveling-wave series of panel (a), the predicted phase advance and wavenumber is

$$\Omega = \omega_c - \Im(\beta_{\mathrm{tr}})R_\star^2, \qquad k_c = \frac{N}{2\pi}\arg\max_q |\Lambda(q)|. \tag{171}$$

For panel (d), the cubic two-mode theory predicts the single-mode states

$$R_i^2 = -\frac{\mu_i}{\beta_{ii}}, \qquad \mu_j < \frac{\beta_{ji}\mu_i}{\beta_{ii}} \tag{172}$$

as the condition for the i-mode state to suppress mode j.

Near onset, where $|\Lambda(q_c)| < 1.05$, the simulated amplitude agrees with $R_\star$ to within 3% in median error; over the full unstable region, the median error increases to 14%, as expected from a cubic truncation. The observed wavenumber matches the linear prediction at 96% of unstable grid points. The exceptions occur near transition lines and in the resonant sector $k_c = 8$, where $4q_c \equiv 0 \pmod{2\pi}$. Finally, the predicted competition phase agrees with the simulated outcome at 97% of the grid points.

J CONTROLLED EXPERIMENTS: EXPLOITING STRUCTURED REPRESENTATIONS

In this section we give the full construction and evaluation protocol of the controlled experiments described in Section 6.2. The objective is to test consequences of the derived matrix dispersion relation as well as our realized taxonomy. The question we seek to answer is the following: if a task requires a privileged representation, does initializing to amplify the relevant joint token-feature structure provide an optimization advantage? To answer this question, we design three task families, and vary a theoretically identifiable property of the critical mode while holding the architecture, parameter count and maximal linear gain fixed.

J.1 ARCHITECTURE AND GAIN-MATCHED DISPERSION ENGINEERING

All controlled experiments use the following setup. Periodic sequences of $N = 32$ tokens, feature width $d = 32$, $H = 4$ attention heads and $L = 8$ looped interactions of the single weight-tied Transformer block detailed in Section 3. We use relative positional encodings, the standard content-dependent query-key attention as presented, a residual MLP with $\tanh$ activation, and no normalization. We initialize the feedforward weights so that the point-wise residual Jacobian of the homogeneous state satisfies:

$$C_* = \xi I \tag{173}$$

with $\xi = 0.9$. For each engineered "morphology" or structured-representation we calibrate OV matrices and relative positional encodings so that:

$$\max_q \rho(J(q)) = \rho_{target} = 1.10 \tag{174}$$

and

$$\rho(J(0)) = 0.95\rho_{target} \tag{175}$$

This ensures both that all selected morphologies have the same gain, and also that the homogeneous mode is not preferred at initialization.

By engineering this dispersion relation, we can controllably select among three morphologies:

- **Standing-wave**: we place the critical mode at $q_c \in \{3, 4, 5, 7\}$
- **Traveling-wave**: in this case all morphologies share the same $q_c = 4$ but the leading multiplier is assigned an additional token phase $\theta \in \{-\frac{\pi}{2}, -\frac{\pi}{4}, \frac{\pi}{4}, +\frac{\pi}{2}\}$
- **Feature-rotating**: here we also initialize with $q_c = 4$ but engineer feature operator for a critical mode inducing a feature-rotation $\psi \in \{-\frac{\pi}{2}, -\frac{\pi}{4}, \frac{\pi}{4}, +\frac{\pi}{2}\}$

In addition to these engineered initializations we include the following baselines: a gain-matched random morphology with a seed-dependent random critical mode, a standard Transformer initialization with learned relative positional bias, and a standard initialization with sinusoidal absolute positional encoding.

J.2 TASK FAMILIES

Here we describe the 3 task families designed to target a specified morphology.

T1: periodic infilling (wavelength) : this task consists in infilling a scalar periodic signal whose dominant mode is $q_c \in \{3, 4, 5, 7\}$. The signal is perturbed by random phases and amplitude, together with weak nuisance $q = 11$ and $q = 13$ components. The signal is quantized into 8 levels, giving 8 class labels. In addition to the signal perturbation, a block of 12 tokens is removed, and an additional 25 % of the token are randomly masked. The masked block is longer than one period for all carrier frequencies, which prevents the task from being solved by mere copying.

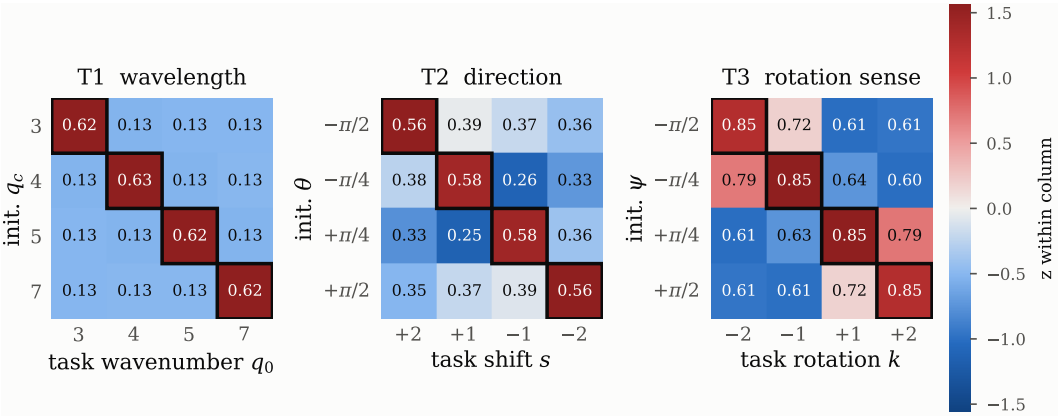


Figure 6: Extended version of Fig. 4. Heatmap of AUC of learning curves for each (task, initialization) pair; color gives the z-scores. Higher scores in the diagonal support the dynamical prior induced by mode criticality.

T2: shift (propagation direction) : input contains a dominant carrier frequency $q_0 = 4$ and weak nuisance modes $q = 3, 5$. These signals are again quantized into 8 classes. The output is the input sequence shifted by $s \in \{-2, -1, +1, +2\}$, 25% of the tokens are masked. A solution to the task would be an accumulated phase shift of

$$\theta_{task} = -\frac{2\pi q_0 s}{N} \tag{176}$$

T3: rotating-wave infilling (feature rotation) : at token j, the input lies in a 2D feature-plane with angle: $\frac{2\pi q_0 j}{N} + \phi_0$, with $q_0 = 4$, and is again quantized into 8 angular states giving 8 labels. We add phase perturbation, mask a block of 12 tokens and randomly mask 25 % of the tokens. The target simply shifts the feature vectors by k 8th turn with $k \in \{-2, -1, +1, +2\}$ which would correspond to a rotation in feature space of:

$$\psi_{task} = \frac{2\pi k}{8} \tag{177}$$

The model must then recreate the carrier across token space while ensuring the correct rotation in feature space.

J.3 Training protocol and Metrics

Optimization uses AdamW, with learning rate 2×10^{-3}, weight decay 10^{-4}, gradient clipping at norm 1 and batch size of 64. Each run lasts 300 gradient steps. For the training-efficiency analysis we employ training data of size:

$$n_{train} \in \{16, 32, 64, 128, 256, 512\} \tag{178}$$

J.4 Results

Figure 6 reports the complete engineered-arm/task grids. The predicted diagonal is visible on all three axes. Wavelength matching produces an especially sharp dissociation: the matched standing wave arms achieve mean early AUC between 0.619 and 0.626, whereas the frozen strongest mismatched arms are between 0.132 and 0.134. Propagation direction produces a smaller but consistent phase specificity: matched arms obtain early AUC 0.559–0.582, compared with 0.362–0.391 for the frozen mismatches. Feature-rotation sense is also resolved: matched arms obtain early AUC approximately 0.853–0.854, with matched-mismatched differences between 0.061 and 0.132 depending on rotation magnitude. Across all 12 tasks, every paired matched–mismatched comparison has the same-sign ordering for all 12 final seeds.

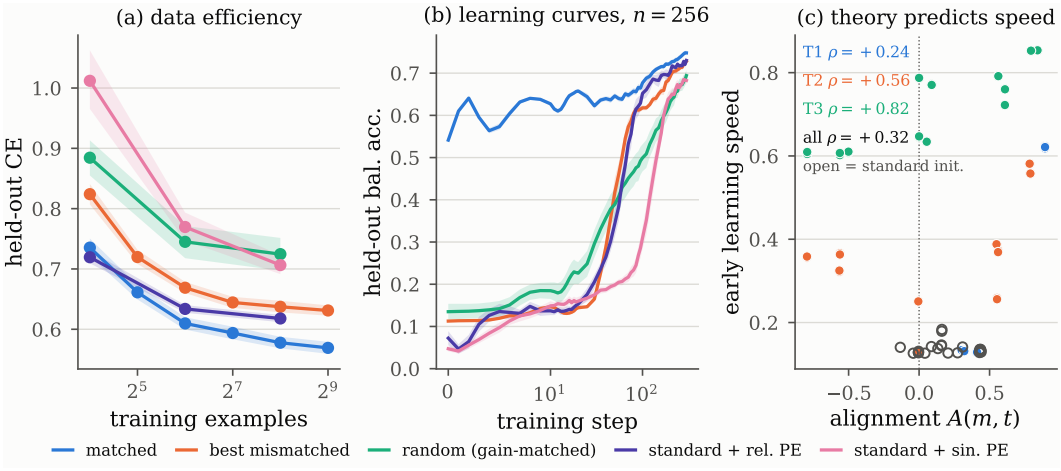


Figure 7: Learning speed and data efficiency on T1: held-out cross-entropy versus training-set size (left) and learning curves (right) for the matched, mismatched and gain-matched random initializations.

J.5 Learning speed and data efficiency

Figure 7 gives complementary views of the effect. For T1, the matched morphology improves held-out cross-entropy relative to the gain-matched random and mismatched structured controls across the training-set-size sweep, with the largest differences in the low-data regime. The corresponding learning curves show that most of the separation occurs early, motivating AUC_{early} as the primary statistic. The standard relative-PE baseline is an intentionally less controlled practitioner baseline, it is not gain matched and does not satisfy the fixed, C_* construction.

K Vision experiments

We next test whether the same-dispersion relation based design principle applies in a less controlled image-classification setting. The goal is to alter the initialization of the 2-dimensional dispersion landscape of a ConViT-style attention block and test whether favoring heterogeneous spatial modes changes optimization on CIFAR-10.

K.1 Weight-tied ConViT architecture

To stay in line with our theory, we use a weight-tied ConViT, in which one Transformer block is applied repeatedly across depth. A 2×2 patch embedding maps a 32×32 image to a 16×16 token grid ($N = 256$). The embedding dimension is $d = 216$, we employ $H = 9$ heads and have depth $L = 12$. Unless otherwise stated, we use a gated-positional self-attention (GPSA) as described in (d'Ascoli et al., 2021), and a periodic relative encoding to stay firmly in line with our theory. The architecture has no class token and no absolute positional embedding. Classification uses mean pooling followed by LayerNorm and a linear layer.

K.2 Initialization

We compare seven initializations.

- **default**: uses standard small random weights and disables structured positional locality.
- **+I, -I**: these variants use a Mimetic-style (Trockman & Kolter, 2023) initialization for QK and initialize the full OV operator, from a random matrix plus a positive or negative identity-biased component. It has the form $M_{OV} = 0.4Z \pm 0.4I$

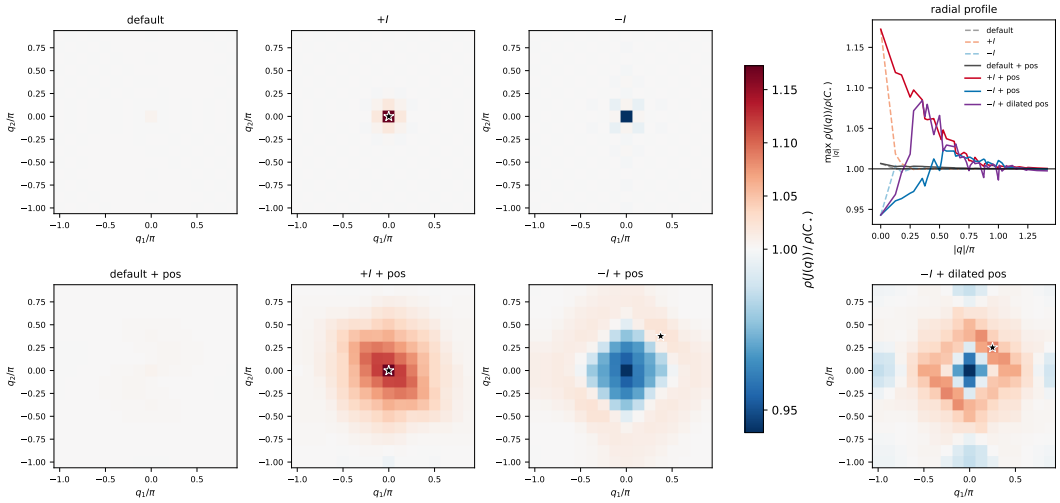


Figure 8: Visualizations of the 2D dispersion relation for various combinations of positional structure and OV geometry

- **+pos**: these variants additionally initialize the GPSA positional logits with a localized ConViT positional encoding. Variants such as $-I + pos$ combine negative mimetic with a localized positional kernel.

These initialization schemes allow us to separate the three ingredients suggested by the theory, namely the interplay of OV geometry and attention positional structure in selecting relevant heterogeneous structures.

K.3 Two-dimensional dispersion relation

For a homogeneous embedded token state x^*, our architecture allows exact computation of the matrix dispersion relation. For each positional frequency $q = (q_1, q_2)$, we have:

$$J(\mathbf{q}) = C_*[I + \epsilon_a \sum_{h=1}^{H} \lambda_h(\mathbf{q}) M_h] \tag{179}$$

where ϵ_a is the learned attention residual scale. In this particular case, and given our initialization this is analytically tractable. Figure 5 visualizes this 2d dispersion for various initialization-positional structure combination. As predicted by the theory, without positional encoding the dispersion is nearly flat or dominated by the homogeneous mode. Local positional encoding makes the dispersion relation frequency dependent. We see the interplay with OV geometry, negative-identity biased OV amplifies heterogeneous segments while positive-identity biased OV amplifies the homogeneous mode. We believe this helps rationalize the empirical observations that led to this initialization in the original work Trockman & Kolter (2023)

K.3.1 Finite-depth validation of the dispersion prediction

Although the dispersion relation is a one-step local quantity, the tied network applies the same block $L = 12$ times. We therefore test whether it predicts finite-depth perturbation growth. For each Fourier sector $\mathbf{q}$, we perturb the homogeneous state by 10^{-6} along the leading eigenvector of $J_0(\mathbf{q})$ and measure the perturbation norm after 12 tied steps. We compare it with the exact linear prediction

$$w_L(\mathbf{q}) = J^{(L-1)}(\mathbf{q}) \cdots J^{(0)}(\mathbf{q}) v_0(\mathbf{q}), \tag{180}$$

and the initialization-time proxy $\rho(J^{(0)}(\mathbf{q}))^L$.

As shown in Fig. 9, the measured growth agrees with the exact linear evolution to below 5×10^{-9} relative error for $+I$+pos, $-I$+pos, and engineered initialization. Moreover, $\rho(J_0(\mathbf{q}))^L$ closely

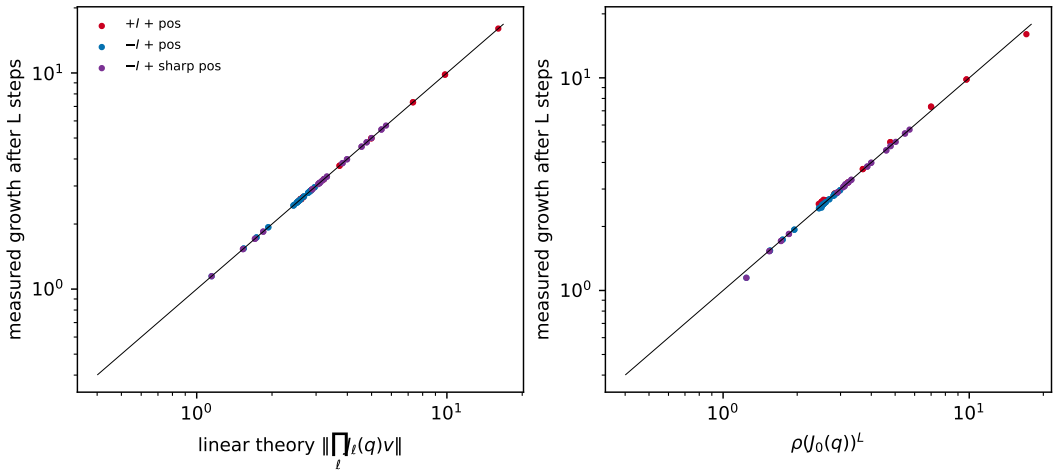


Figure 9: Comparison of linear theory growth prediction vs actual deviation from homogeneous state from various positional encoding and OV geometry combinations

predicts the relative growth across Fourier sectors, with Pearson correlations 0.9984, 0.9993, and 0.9999, respectively. Thus, the initialization-time dispersion map accurately predicts which spatial modes are preferentially amplified through the full tied depth in the local regime.

K.4 CIFAR-10 TRAINING PROTOCOL

We use CIFAR-10 with a fixed class-balanced 45k/5k train/validation split and the standard 10k test set. Training uses random resized crops, horizontal flips, RandAugment, color jitter, random erasing, and standard normalization; full-data results are averaged over three seeds.

Models are trained for 100 full-data-equivalent epochs with AdamW (batch size 512, learning rate 3×10^{-3}, weight decay 10^{-2}), five epochs of warmup followed by cosine decay, and gradient clipping at norm 1. Biases, positional parameters, GPSA gates, and residual scales are exempt from weight decay. For data-efficiency experiments, we keep the number of optimization steps fixed to the full-data schedule across dataset fractions.

We report final test accuracy, test-accuracy AUC, and the first epoch t_{90} reaching 90% test accuracy; seeds that do not reach the threshold are excluded from t_{90} and the number of successful seeds is reported.

K.4.1 OPTIMIZATION RESULTS

Figure 10 summarizes the full-data results. OV sign alone has little effect without positional structure, whereas combining localized attention with negative OV geometry substantially accelerates optimization. Sharpening the positional kernel further improves learning speed, with a smaller improvement in final accuracy.

K.4.2 DATA EFFICIENCY

We repeat the comparison using 5%–100% of the training set while keeping the optimization-step budget fixed. As shown in Fig. 11, the engineered initialization has the highest mean final accuracy and AUC at every evaluated fraction, with the largest gains over the unstructured default in the low- and intermediate-data regimes.

K.4.3 TRAINING DIAGNOSTICS AND DISPERSION EVOLUTION

Figure 12 shows that the positional variants maintain heterogeneous token representations while both the attention residual scale and GPSA positional gate adapt during training. Thus, the engineered initialization biases the initial dynamics without freezing a prescribed attention operator.

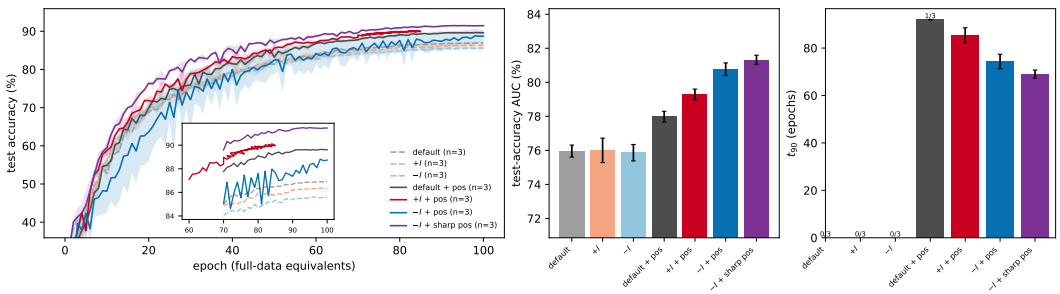


Figure 10: **CIFAR-10 optimization under dispersion-aware initialization.** Test accuracy (left), test-accuracy AUC (middle), and epoch reaching 90% test accuracy (right), over three seeds. Localized positional attention combined with negative OV geometry accelerates training, with the engineered $-I+$ sharp-pos initialization performing best overall.

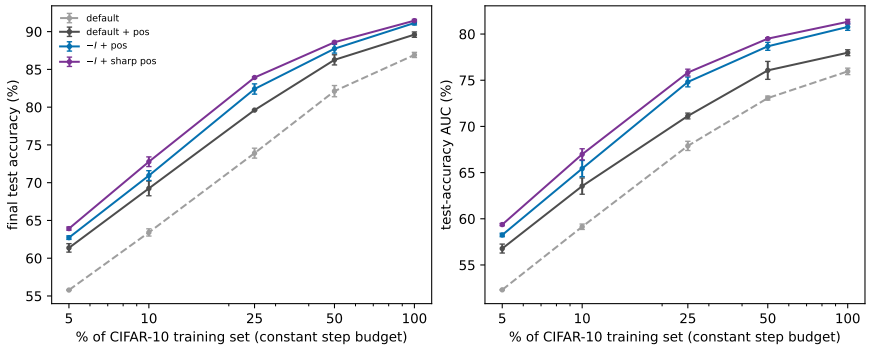


Figure 11: **Data efficiency on CIFAR-10.** Final test accuracy (left) and test-accuracy AUC (right) for different training-set fractions. Error bars show one standard deviation across three seeds, except the 5% default baseline ($n = 2$). The engineered initialization remains strongest across the sweep.

We also recompute the two-dimensional dispersion relation from checkpoints (Fig. 13). Holding negative OV geometry fixed, the sharper positional kernel produces a stronger initial heterogeneous-mode bias, but both spectra subsequently reorganize. Dispersion engineering therefore acts as an initial preference rather than a persistent spectral constraint.

K.4.4 Robustness and scope

Table 1 summarizes two architectural ablations. The benefit of combining positional locality with negative OV geometry persists with pre-LN, although the additional gain from positional sharpening does not. In contrast, injecting the positional bias inside the content softmax (PSA) is markedly less stable at the same initialization scale, with several runs diverging. Thus, the effect is robust to normalization but depends on how positional structure enters attention.

L A pattern-formation primer: the Swift–Hohenberg model

This appendix illustrates, on the canonical pattern-forming Swift–Hohenberg model (Cross & Greenside, 2009), the two objects used in the main text: the dispersion relation and the amplitude equation. As in Eq. 3, the model couples a spatial propagation operator to local pointwise dynamics:

$$\frac{\partial X}{\partial t} = \underbrace{-\left(q_0^2 + \Delta\right)^2 X}_{\mathcal{P}(X)\text{: propagation}} + \underbrace{\mu X + \phi(X)}_{\mathcal{L}(X)\text{: local dynamics}}, \tag{181}$$

where $\phi(0) = 0$ and $\phi(X) = \alpha_0 X + \alpha_1 X^2 + \alpha_2 X^3 + \mathcal{O}(X^4)$. The homogeneous state $X = 0$ is reminiscent of the collapsed Transformer state. To probe its stability, consider a perturbation

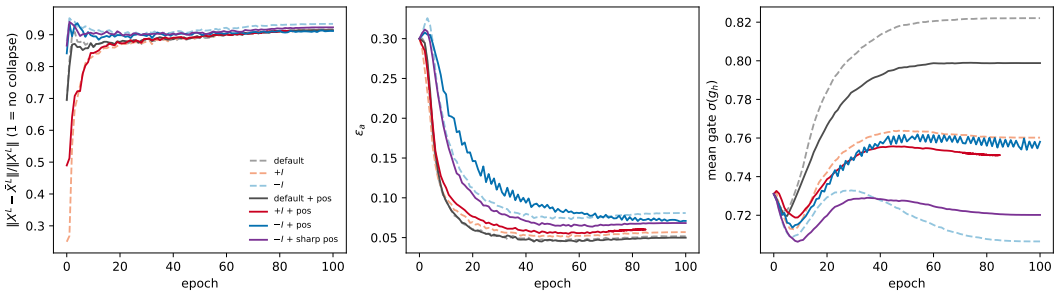
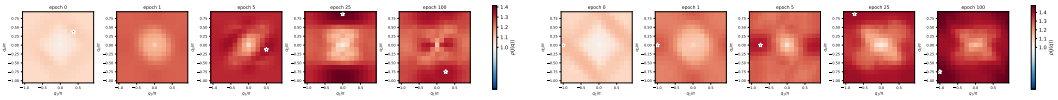


Figure 12: **Training diagnostics for the tied ConViT.** Final-depth token heterogeneity (left), learned attention residual scale ϵ_a (middle), and mean GPSA positional gate (right). Representations remain non-collapsed while the attention parameters adapt throughout training.



$-I$+pos engineered $-I$+sharp pos

Figure 13: **Evolution of the spatial dispersion relation.** $\rho(J(\mathbf{q}))$ at epochs $0, 1, 5, 25$, and 100 for one seed; stars denote the maximally amplified Fourier sector. Both initial spectra evolve substantially during training.

$U \propto \mathrm{e}^{\sigma(k)t+\mathrm{i}k\cdot x}$, where k encodes spatial scale and $\sigma(k)$ its growth rate. Linearizing Eq. 181 gives the dispersion relation

$$\sigma(k) = \mu + \alpha_0 - \left(q_0^2 - |k|^2\right)^2 . \tag{182}$$

Its sign determines whether a perturbation grows or decays, while its shape determines which mode is preferentially amplified: here the critical mode is $q_c = q_0$. Thus, the dispersion relation simultaneously characterizes the loss of stability of the homogeneous state and the structure that emerges from it, as illustrated in Fig. 14.

Linear analysis, however, does not determine the fate of the emerging mode. Near onset, the solution is dominated by $U \approx B(t)\mathrm{e}^{\mathrm{i}q_c x}$, and nonlinear effects govern whether its amplitude diverges or saturates. Expanding Eq. 181 to the third order and projecting onto the critical mode yields reduced amplitude dynamics of the form

$$\frac{dB}{dt} = -\frac{\partial \mathcal{V}}{\partial B}, \qquad \mathcal{V}(B) = -\frac{\sigma}{2}B^2 + \frac{g(\alpha_1, \alpha_2)}{4}B^4. \tag{183}$$

The dispersion relation controls the stability of the homogeneous state, while the nonlinear coefficient $g(\alpha_1, \alpha_2)$ controls the stabilization of the emerging structure; for $g > 0$, the growing mode saturates at finite amplitude. We see in Fig. 14(b) the impact of g, which can be viewed as engineering the stability landscape of the emergent pattern.

This example contains the two ingredients used in the main text: the dispersion relation determines when the homogeneous state loses stability and which structures become accessible, while nonlinearities determine their subsequent fate. Crucially, in this example the propagation operator $\mathcal{P}(X)$ diagonalizes along Fourier modes, reducing the stability analysis to independent spatial scales. In Transformers, positional encoding and multi-head attention play the analogous role (Theorem 4.5), giving rise to a matrix-valued dispersion relation.

Initialization	Final acc. (%)	AUC (%)	t_{90}
GPSA, no normalization			
default + pos	89.6 ± 0.4	78.0 ± 0.3	92.0 (1/3)
$-I$ + pos	91.1 ± 0.3	80.8 ± 0.4	74.3 ± 3.1
$-I$ + sharp pos	$\mathbf{91.5 \pm 0.2}$	$\mathbf{81.3 \pm 0.3}$	$\mathbf{69.0 \pm 1.7}$
GPSA, pre-LN			
default + pos	88.5 ± 0.4	77.0 ± 0.4	–
$-I$ + pos	$\mathbf{89.8 \pm 0.3}$	$\mathbf{79.5 \pm 0.4}$	88.0 (1/3)
$-I$ + sharp pos	89.5 ± 0.2	78.5 ± 0.2	–
PSA, no normalization			
$-I$ + pos	87.3 ± 0.1 (2/3)	75.2 ± 0.2 (2/3)	–
$-I$ + sharp pos	86.8 (1/3)	73.4 (1/3)	–

Table 1: **Normalization and attention-form ablations.** Parentheses indicate the number of contributing seeds when fewer than three runs are available; for t_{90}, they indicate the number of seeds reaching 90% accuracy within 100 epochs.

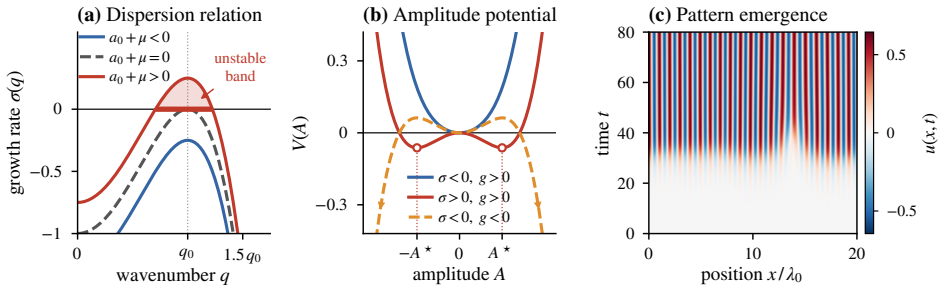


Figure 14: The Swift-Hohenberg illustration: the dispersion relation (a) defines the emergent perturbations, the amplitude potential in (b) dictates their stability, leading to the apparition of patterns in (c).