Two-Clustering Regime of Token Dynamics in Causal Attention

Trinh T. Nguyen
School of Mathematical Sciences
Shanghai Jiao Tong University, China
q0004349@sjtu.edu.cn

Duy-Tung Pham
Khoury College of Computer Sciences
Northeastern University, USA
pham.duy@northeastern.edu

Hoang-Son Do
Institute of Mathematics
Vietnam Academy of Science and Technology
dhson@math.ac.vn
University of Science and Technology of Hanoi
Vietnam Academy of Science and Technology
do-hoang.son@usth.edu.vn

Tan Minh Nguyen[†]
Department of Mathematics
National University of Singapore
tanmn@nus.edu.sg

Thieu N. Vo[†]
Department of Computer Science
University of Bath, United Kingdom
ntv22@bath.ac.uk

Abstract

We partially resolve Conjecture 1 of *Karagodin et al., Clustering in Causal Attention Masking (NeurIPS 2024)* and extend it to time-dependent parameter regimes. The conjecture predicts that causal attention dynamics converge to two clusters aligned with the eigenvectors associated with the largest eigenvalue $\lambda_{\max}$ of the value transformation. We prove that this behavior holds when $\lambda_{\max} > 0$ and is generally false when $\lambda_{\max} \leq 0$. The resulting two-cluster regime is common in practice and substantially more challenging than the classical single-cluster setting studied previously. A central contribution of our work is the identification of a rigorous dimension reduction phenomenon: despite evolving in a high-dimensional embedding space, token trajectories collapse onto a two-dimensional structure governed by a small number of dominant directions. Building on this insight, we introduce a new analytical framework–instability within the flow–inspired by instability phenomena in physical systems. This framework enables the analysis of steady-regime behavior in non-monotone, non-autonomous attention dynamics. Our theoretical findings are illustrated through simulations and experimental studies using Group Query Attention (GQA) layers.

1 Introduction

Transformer architectures [78] underpin modern artificial intelligence, achieving strong performance in natural language processing [18, 17, 6, 8, 57, 65], computer vision [21, 66, 62, 47, 1], general machine learning [71, 38, 54, 45], and reinforcement learning [34, 12]. Their success is driven by large-scale pretraining, enabling effective transfer to downstream tasks [64, 19, 63, 74]. By replacing

[†]Co-last authors. Correspondence to: q0004349@sjtu.edu.cn & ntv22@bath.ac.uk.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

recurrence and convolution with self-attention, Transformers model global token interactions and capture long-range dependencies, yielding rich contextual representations [46, 79, 16, 73, 52, 55, 60]. As a result, Transformers form the core architecture of large language models. Despite their empirical success, a rigorous theoretical understanding of their internal representations remains limited, an issue that becomes increasingly critical as these models continue to scale.

Transformers can be interpreted through a particle-based perspective, where individual tokens are viewed as interacting entities. Adopting this viewpoint, the works of [28, 29] investigate the continuous-time limit of Transformer models, reformulating their behavior as a dynamical system. Within this framework, the authors demonstrate that token representations evolve toward clustered configurations associated with specific attractors determined by their initial conditions, providing a theoretical explanation for the contextual nature of Transformer representations. Further extending this line of analysis, [27] shows that the self-attention dynamics can exhibit metastability, with tokens remaining confined near multi-cluster configurations over exponentially long time horizons. Complementary results are obtained in [7], where a hardmax, attention-only Transformer is examined in a discrete-time setting using similar analytical techniques. Furthermore, the authors in [61] analyze the flow of tokens across attention layers and investigate how different forms of positional encoding - specifically absolute and rotary - affect these dynamical regimes.

Recent work on token dynamics has been extended to causal attention to better reflect its central role in sequential generation. In this autoregressive setting, each token is allowed to attend only to past tokens, enforcing temporal consistency by masking future information during attention computation. This restriction distinguishes causal attention from full attention, which operates on complete sequences and is therefore better suited to tasks such as machine translation. Because of its ability to support step-by-step generation, causal attention is widely used in large-scale generative models, including natural language and image generation systems such as DALL-E [66], VQGAN [25], and Parti [85], as well as multimodal foundation models like Chameleon [72]. More generally, attention mechanisms that limit interactions to subsets of tokens have enabled recent scaling advances, leading to models such as MUSE [11] and AlphaFold 3 [3]. Unlike full attention, the dynamics induced by causal attention cannot be characterized as a mean-field gradient flow as in [27, 29, 28], necessitating different analytical tools.

Work	**Parameters**	**Initial Data**	**Key Techniques**	**Theoretical Results**
[37]	V constant Q, K general	Arbitrary on $\mathbb{S}^{d-1}$	Explicit $x_1(t)$ formula; Schub genericity theorem	Convergence of $x_1(t)$ to $\pm\xi_{\max}$ Full convergence if $V = I_d$
[2]	$V(t)$ symmetric $Q(t), K(t)$ general	Single hemisphere	Lyapunov function; Input-to-state stability	Global convergence to single cluster $\xi_{\max}$
Ours	$V(t)$ symmetric $Q(t), K(t)$ general	Arbitrary on $\mathbb{S}^{d-1}$	Instability analysis; Hierarchy dim. deduction Error-to-traverse ratio	Global convergence to **two clusters** $\pm\xi_{\max}$; Theoretical justification of **dimensional deduction**

Table 1: Comparison of theoretical analyses for causal-attention dynamics. Our work resolves the two-cluster regime for diagonal $V(t)$ without hemisphere constraints and provides the first theoretical justification for the dimensional deduction phenomenon conjectured from experiments.

1.1 Background – Causal Attention Dynamics and a Two-Clustering Conjecture

A full Transformer layer typically consists of multi-head self-attention, a position-wise feed-forward network (FFN), normalization layers, and residual connections [78]. Our goal is not to reproduce every component of the Transformer architecture, but rather to clarify and formalize the key mathematical structure that governs its behavior. To this end, we simplify the architecture to only the components relevant for our analysis: a single-head self-attention layer with normalization and a residual connection, omitting the FFN block.

Let the input sequence be represented by token embeddings $X = [x_1, \dots, x_L] \in \mathbb{R}^{L\times d}$, where L is the number of tokens and d is the embedding dimension. With parameter matrices $Q, K, V \in \mathbb{R}^{D\times D}$

used to form queries, keys and values, the attention operator is given by

$$\mathrm{Attn}(X)_i = \sum_{j=1}^{L} \left(\frac{e^{\langle Qx_i, Kx_j \rangle}}{\sum_{k=1}^{L} e^{\langle Qx_i, Kx_j \rangle}} \right) Vx_j,$$

for $i = 1, \ldots, L$.

Following [87], we add an RMS normalization layer and a residual connection. Denoting by X^t the activations at layer t, the update to the next layer is

$$X^{t+1} = X^t + \mathrm{RMSNorm}(\mathrm{Attn}(X^t)). \tag{1}$$

The update rule (1) can be interpreted as a finite-difference approximation of a time derivative, with $X^{t+1} - X^t$ playing the role of the increment between consecutive layers, and the RMS normalization playing the role of keeping the token representations lie on the standard unit sphere $\mathbb{S}^{D-1}$. Hence, the sequence $\{X^t\}_t$ can be regarded as a time-discretization of an underlying continuous-time flow which is given by the dynamical system

$$\partial_t x_i(t) = \mathbf{P}_{x_i(t)} \Big(\sum_{j=1}^{L} a_{ij}(t) V(t) x_j(t) \Big),$$

with

$$a_{ij}(t) = \frac{e^{\langle Q(t)x_i(t), K(t)x_j(t) \rangle}}{\sum_{k=1}^{L} e^{\langle Q(t)x_i(t), K(t)x_j(t) \rangle}},$$

and the operator $\mathbf{P}_x(y) := y - \langle x, y \rangle x / \|x\|^2$ is the projector that ensures $x_k(t)$ remains on the sphere. Finally, by imposing a causal masking constraint, so that each token attends only to its predecessors, we obtain the following causal self-attention equation that we analyze in this work:

$$\partial_t x_i(t) = \mathbf{P}_{x_i(t)} \Big(\sum_{j=1}^{i} a_{ij}(t) V(t) x_j(t) \Big), \tag{2}$$

with

$$a_{ij}(t) = \frac{e^{\langle Q(t)x_i(t), K(t)x_j(t) \rangle}}{\sum_{k=1}^{i} e^{\langle Q(t)x_i(t), K(t)x_j(t) \rangle}},$$

for $i, j = 1, \ldots, L$.

The authors in [37] predicted that the eigenvectors of the value matrix plays a central role in determining the qualitative clustering behavior of the causal attention dynamics. In the time-independent setting where Q, K, and V are fixed, the authors in [37] further conjectured the following.

Conjecture 1.1 (Two-clustering configuration of the causal attention dynamics)**.** Assume that Q, K, V are arbitrary time-independent matrices and that V has d distinct positive eigenvalues. Let $\lambda_{\max}$ denote the largest eigenvalue, and let $\xi_{\max}$ be a corresponding unit eigenvector, i.e., $|\xi_{\max}| = 1$. Then, for almost every initial condition

$$X(0) = (x_1(0), \ldots, x_L(0)) \in (\mathbb{S}^{d-1})^L,$$

the causal attention dynamics (2) converges to a two-cluster configuration:

$$\lim_{t \to +\infty} x_i(t) \in \{-\xi_{\max}, \xi_{\max}\}, \qquad \text{for all } i = 1, \ldots, L.$$

This two-cluster behavior has been observed empirically for causal attention in [37] and for full attention in [61]. To the best of our knowledge, however, a rigorous proof of this conjecture is not currently available in the literature.

1.2 Our main contributions

In this paper, we prove Conjecture 1.1 under a mild condition and further generalize it to settings with time-dependent parameters $Q(t), K(t)$ and $V(t)$. Our rigorous theoretical analysis reveals several dynamical mechanisms that govern when and how two-cluster configurations emerge, including hierarchical suppression, intrinsic dimensional reduction, and recurrence to the dominant directions $\pm\xi_{\max}$.

- **Recurrence to the dominant direction $\xi_{\max}$.** Once the first $L-1$ tokens stabilize, the last token is forced to repeatedly re-align with $\xi_{\max}$ along arbitrarily large depths (Theorem 4.2). This recurrent re-alignment creates a "déjà vu" effect: past dominant information is not permanently forgotten, but reappears again and again in the infinite–depth limit.
- **Hierarchical suppression mechanism.** We introduce the *error-to-transverse ratio* $\kappa_\ell(t)$ (see definition in Eq. (5)), a simple but crucial quantity that governs a striking *hierarchical suppression*: depending on its size, the system either collapses entire blocks of coordinates or locks onto exactly one surviving direction (see Theorem 2.2). The error-to-transverse ratio exactly predicts when a token will suppress all but its most favorable direction.
- **Dimensional reduction.** The token dynamics concentrates on a low-dimensional manifold, revealing *dimensionality reduction* within deep sequence models. Under mild conditions, every nontrivial limiting configuration lies in a two-dimensional coordinate plane and satisfies a single scalar algebraic constraint (Theorem 2.3). This rigidity phenomenon appears to be new and, to our knowledge, has not been rigorously analyzed for attention dynamics in the literature.

Existing analyses of the causal-attention ODE can be categorized by their assumptions on the parameter matrices (V, Q, K) and initial data. A summary of the theoretical landscape and the key analytical ideas is provided in Table 1. Our theoretical findings are illustrated through simulations and experimental studies using Group Query Attention (GQA) layers.

Difficulties from previous approaches. The two-cluster regime with time-varying depth is far more challenging than the classical single-cluster setting, which was the primary focus of the analyses in [37] and [2]. The dynamics no longer admit a monotone Lyapunov functional, and thus every existing technique in the literature – including hemisphere-based Lyapunov methods [2] and Schub-type genericity arguments [37] – completely fails in this regime.

Our new approach. Our key contribution is a new mathematical tool that we call the *instability-within-the-flow*, inspired by stability theory in physics. The evolution of a token under a time-dependent value matrix $V(t)$ resembles a small ball rolling on a hill whose summit $r(t)$ itself moves over time. As the "hilltop" $r(t)$ drifts, even slowly, the stable and unstable regions experienced by the token shift accordingly, producing a much richer and more delicate dynamical landscape. A rapidly moving hilltop can trap the flow in an interior region (leading to interior clusters), while a stationary hill forces the trajectory toward one of two extremal states $\{\xi_{\max}, -\xi_{\max}\}$. To our knowledge, this structure has not appeared in previous work, and we believe it offers a promising mathematical tool for understanding directional biases and cluster formation in deep sequence models. The full framework is developed in Sections 3 in the main text and in Section E in appendix.

Organization. Section 2 analyzes hierarchical suppression and high-dimensional dimension-reduction mechanisms. Section 3 visualizes causal attention dynamics as the motion of a small ball on a hill. Section 4 proves the two-cluster behavior of tokens. Section 5 validates the theoretical findings in practical Transformers with GQA layers.

2 Hierarchical Suppression Mechanism and Dimension Reduction

Our first main result is to uncover a high-dimensional reduction mechanism in which the dynamics collapse to exactly two coordinates for arbitrary matrices $Q(t)$ and $K(t)$. Although the dynamics formally reside in a high-dimensional space, the long-time behavior is effectively governed by only a handful of intrinsic degrees of freedom.

The study of the high-dimensional reduction mechanism based on the dynamical system (2) without any additional assumption on the model parameters $Q(t), K(t)$, and $V(t)$ is challenging. To make the proofs feasible while still maintaining the generality practicality of our theoretical findings, we will use the following assumption during this section (see also in [2]).

Assumption 2.1 (Symmetric and time-invariant eigenspace value matrix)**.** *The value matrix $V(t)$ is a symmetric matrix having d different eigenvalues and time-invariant eigenspaces. Furthemore, by a suitable linear change of variables, we can write without loss of generality that:*

$$V(t) = \mathit{diag}\{\lambda_1(t), \lambda_2(t), \ldots, \lambda_d(t)\},$$

so that the eigen-basis coincides with the canonical orthogonal basis $\{\vec{e}_1, \dots, \vec{e}_d\}$ *and*

$$\lambda_1(t) > \lambda_2(t) > \cdots > \lambda_d(t) \quad \textit{for all} \quad t \geq 0.$$

Under Assumption 2.1, the largest eigenvalue $\vec{e}_1 = \xi_{\max}$ will be the leading direction selected by the dynamics. Let us fix a solution $\mathbf{x}(t) = (x_1(t), \dots, x_L(t)) \in C^\infty([0, +\infty), \mathbb{S}^{d-1})^L$ of the dynamical system (2), subject to the initial condition $\mathbf{x}(0) \in (\mathbb{S}^{d-1})^L$. We will derive our theorems by induction on the number L of tokens. In particular, we define the error function

$$\varepsilon_{L-1}(t) = \sum_{j=1}^{L-1} \|x_j(t) - s_j \vec{e}_1\|,$$

where $\|\cdot\|$ denotes the Euclidean norm. Our inductive hypothesis for the first $L-1$ tokens is

$$\varepsilon_{L-1}(t) \;\longrightarrow\; 0 \qquad \text{as } t \to \infty, \tag{3}$$

Here s_j simply records which cluster the token x_j converges to: $+1$ corresponds to the "top" copy of $\vec{e}_1$ and -1 to the "bottom" copy $-\vec{e}_1$. Since we are now focusing on the dynamic of the L-th token, to simplify the notation, we will denote

$$y(t) = (y_1(t), \dots, y_d(t)) := x_L(t). \tag{4}$$

Here, y_1 measures alignment with the dominant direction, and the remaining coordinates $(y_2, \dots, y_d)$ represent transverse components.

Our goal in this section is to identify the leading dynamics of y. We first prove that the dynamic of $y(t)$ induce a hierarchical suppression mechanism as follows:

Theorem 2.2 (hierarchical suppression mechanism – See Theorem C.6)**.** *For each level* $\ell \in \{2, \dots, d\}$*, define*

$$R_\ell(t) = \Big(\sum_{j=\ell}^{d} y_j(t)^2\Big)^{1/2}, \quad \textit{and} \quad \kappa_\ell(t) = \frac{\varepsilon_{L-1}(t)}{R_\ell(t)}. \tag{5}$$

Assume that for some $\ell \geq 2$*,*

$$\kappa_\ell(t) \longrightarrow 0 \qquad \textit{as } t \to \infty. \tag{6}$$

Then the dynamics enforce a hierarchical suppression among the transverse coordinates $\{y_\ell, \dots, y_d\}$ *in the following sense: exactly one of the two alternatives holds:*

(i). (Dominant transverse direction) *There exists an index* $k \in \{\ell, \dots, d\}$ *such that*

$$y_j(t) \longrightarrow 0 \qquad \textit{for all } j \notin \{1, k\},$$

and the trajectory converges to the invariant two-dimensional subspace spanned by $\{e_1, e_k\}$*.*

(ii). (Recursive suppression) *The coordinate* $y_\ell(t) \longrightarrow 0$*, while the remaining transverse vector* $(y_{\ell+1}(t), \dots, y_d(t))$ *remains nontrivial. In this case the same mechanism applies at the next level* $\ell + 1$*, with* $\kappa_{\ell+1}(t)$ *governing the subsequent suppression.*

Iterating this reduction yields a unique surviving transverse coordinate y_k *for some* $k \geq 2$*, and the trajectory collapses onto a low-dimensional invariant subsystem spanned by* $(y_1, y_2, \dots, y_{\ell-1}, y_k)$*.*

Roughly speaking, by the above theorem, we uncover a hierarchical suppression mechanism acting on the transverse coordinates: As the residual error contributed by earlier tokens becomes negligible relative to the remaining transverse energy, the system progressively eliminates all but one transverse direction, yielding an effective reduction in dimensionality.

Our key idea for the proof of this theorem is to separate the direction dynamics from the magnitude dynamics: once we factor out the overall magnitude of a block of coordinates, the remaining dynamics describe a competition over which direction the token points to. Because different coordinates have systematically different growth advantages (encoded by the ordered of the eigenvalues of the value matrix V), this directional dynamics has a built-in drift that continuously sharpens the representation, pushing it toward the boundary where some coordinates vanish. As a result, the system cannot remain

in a persistent mixed state: either a given coordinate loses mass and is suppressed, or it survives and eventually dominates, eliminating all weaker competitors. The only subtlety is noise inherited from earlier tokens, but as long as this residual signal decays faster than the scale of the current competition, it cannot prevent concentration. In effect, the dynamics behave like an expert-selection process: coordinates are tested one by one, and any direction that remains consistently active must ultimately win decisively. This explains why, despite the ambient dimension, the token collapses onto a low-dimensional structure driven by a small number of dominant features. The full detail of the proof of Theorem 2.2 is presented in Section D.2.

Regarding the dimension reduction mechanism, we further prove another high-dimensional reduction mechanism in which the dynamics collapse to exactly two coordinates, as we see in the following theorem.

Theorem 2.3 (Dimension reduction mechanism – See Theorem C.7)**.** *Let Q and K be arbitrary matrices. Assume that $x_i(t) \to \pm\vec{e}_1$ for all $i \le L-1$, and that the condition* (6) *holds for the final L^{th} token. Then there exists a finite subset $\mathcal{S} \subset \mathbb{S}^{d-1}$ such that, for almost every initial condition, the trajectory $x_L(t)$ converges to some $y_\star \in \mathcal{S}$ as $t \to \infty$.*

Moreover, if the limit $y_\star$ is not equal to $\pm\vec{e}_1$, then $y_\star$ must lie in a two–dimensional coordinate plane and take the form

$$(\alpha_1, 0, \dots, 0, \alpha_\ell, 0, \dots, 0), \qquad \alpha_1^2 + \alpha_\ell^2 = 1,$$

that is, only the first and the ℓ-th coordinates are nonzero.

By Theorem 2.3, we uncover a dimensional reduction mechanism in which the dynamics collapse to exactly two coordinates for arbitrary matrices Q and K. Intuitively speaking, once the parameter matrices Q and K have stabilized, we shows that, asymptotically, the token dynamics are confined to the one-dimensional manifold $\mathbb{S}^1 \times \{0\}^{d-2}$. This residual degree of freedom corresponds to a phase (or orientation) ambiguity in the embedding space. A natural question is whether the token may continue to rotate along this manifold indefinitely, even after the training dynamics appear to have converged. Using tools from complex analysis, we prove that the set of admissible long-time limit points on $\mathbb{S}^1$ is necessarily discrete. Consequently, the token representation must converge to a single fixed configuration (see Lemma G.6). In particular, apparent stabilization of the parameter matrices Q and K implies convergence of the token dynamics. We refer to Theorem C.7 and Section D.3 for the complete argument.

3 Visualization of Tokens' Evolution as Motion of a Small Ball on a Hill

The self-attention mechanism–which updates token representations based on a weighted combination of other tokens–lies at the heart of their expressive power. While recent theoretical work has analyzed the equilibrium states of such systems under static attention patterns, real training dynamics involve rapidly shifting attention, raising a fundamental question: *How do token embeddings evolve when the attention landscape itself changes over time?* In this section, we propose another dynamical perspective that captures this phenomenon intuitively from the motion of a small ball on a hill.

Moving hills and metastable clustering. Consider a token's trajectory under a time-dependent attention matrix $V(t)$ as the motion of a ball on a hill whose summit itself moves. In the static case, the ball inevitably rolls to one of two stable fixed points (analogous to token collapse to $\pm\xi_{\max}$). But if the hilltop moves sufficiently fast, the ball may become trapped near an intermediate position, giving rise to interior clusters–token representations that do not converge to extremal points but instead settle into a metastable or oscillatory state.

To make this intuition concrete, we consider a minimal scalar model:

$$x'(t) = (1 - x^2)(x - r(t)), \qquad x(t) \in (-1, 1).$$

If $r(t) \equiv r_0$ is constant, almost every initial condition converges to ± 1, unless initialized exactly at the unstable equilibrium $x = r_0$. But if $r(t)$ moves rapidly between -1 and 1, the ball cannot roll off either side; it remains trapped near an intermediate value, forming an interior cluster.

A tractable attention model. We now connect this toy analogy to a simplified Transformer setting. Assume $Q^\top K = I_d$ and fix the first token as $x_1(t) \equiv \vec{e}_1$. Let the second token be

$x_2(t) = x(t)\vec{e}_1 + y(t)\vec{e}_2$. Its dynamics reduce (see Theorem C.9) to a scalar equation of the form

$$x'(t) = (1 - x^2)\, h(x)\, g(t, x),$$

where $g(t, x)$ encodes the time-varying attention from the first token, and $h(x) > 0$ is a positive weighting. Crucially, under mild conditions on the attention coefficients $\lambda_1(t)$ and $\lambda_2(t)$, there exists a moving root $r(t) \in (-1, 1)$ satisfying $g(t, r(t)) = 0$, which plays the role of the moving hilltop. If $\lambda_2(t) \to \lambda_{2,\infty} \in (0, 1)$, then $r(t) \to r_\infty \in (-1, 0)$, and the interior cluster stabilizes.

Deneric convergence to extremal or metastable states. Our main technical contribution in this section is a rigorous treatment of this moving-root mechanism in the presence of small perturbations. We prove that, under natural regularity conditions, the dynamics admit at most one exceptional trajectory $x_\star(t)$ that does not converge to ± 1. All other initial conditions are driven to the boundary, except possibly those starting exactly on this special interior path. In particular, we prove that:

Theorem 3.1 (See Theorem E.1)**.** *Consider the perturbed dynamics*

$$x'(t) = (1 - x^2)\, h(x)\, g(t, x) + w(t, x),$$

where $w(t, x)$ is a small, asymptotically vanishing perturbation. Assume that $h(x) > 0$, and $\partial_x g(t, x) \geq c_0 > 0$, as well as there exists a moving root $r(t) \in (-1, 1)$ with $g(t, r(t)) = 0$ and $r'(t) \to 0$. Then:

(i). There exists at most one interior trajectory $x_\star(t)$ that avoids convergence to ± 1.

(ii). If $\sup_x |w(t, x)| \to 0$, then for almost every *initial condition, we have $x(t) \to \pm 1$.*

Intuitively speaking, this result formalizes the hilltop analogy: a rapidly moving but asymptotically slowing root $r(t)$ can sustain at most one persistent interior path; all other trajectories collapse to the poles. The analysis provides a mathematical foundation for understanding how time-varying attention can produce metastable token structures, a phenomenon observed in training dynamics but previously lacking a rigorous dynamical explanation.

Theorem 3.1 is then used to analyze the dominant direction of the dynamic of the leading token $y(t)$ (see equation (4)). Indeed, under suitable block-structure assumptions on $Q(t)$ and $K(t)$, we derive an effective differential equation governing the evolution of the projection $x_L \cdot \xi_{\max} = y_1$, treating all remaining coordinates as perturbations. This reduction is carried out in Theorem D.10. Formally, the dynamics of y_1 take the form

$$\partial_t y_1 = (1 - y_1^2)\, M(t, y_1) + o(1), \tag{7}$$

where $o(1)$ denotes a quantity that vanishes as $t \to \infty$. Since the equation is non-autonomous and the right-hand side is not monotone for the two-cluster regime $\{\xi_{\max}, -\xi_{\max}\}$, we cannot linearize around a stationary solution and apply the stable manifold theorem as in [37, 28], nor use the Lyapunov method as in [7]. Instead, the right-hand side of (7) admits a "moving equilibrium" in the interior of $(-1, 1)$–a phenomenon already visible in the simpler setting (37). This behavior corresponds precisely to the moving-hilltop analogy in the stability theory of physical systems, which we discuss in Section E. Our key insight is that, under suitable time-varying assumptions on the parameter matrices, this unstable equilibrium cannot trap the solution y_1 in the interior of $(-1, 1)$. Consequently, $y(t)$ must converge to one of the two clusters: $\xi_{\max}$ (corresponding to $y_1 = 1$) or $-\xi_{\max}$ (corresponding to $y_1 = -1$). The full analytical framework is developed in Theorem E.1 and Section F. Importantly, our analysis is not limited to establishing convergence to the two-cluster configuration. Rather, it also provides insight into alternative clustering behaviors that may arise in practice and elucidates the instability mechanisms that can unlock such dynamics. In particular, we present an example in which oscillatory drifting of the unstable equilibrium within the interior can trap the L-th token and lead to dynamics distinct from convergence to $\pm\xi_{\max}$. This demonstrates that our assumptions are essentially sharp; see Proposition E.3.

Our work in this section bridges intuitive dynamical analogies with precise asymptotics, offering a new lens through which to study token evolution in Transformers. We believe this perspective will inform both theoretical analysis and the design of stable training procedures.

4 Recurrent Collapse to the Dominant Direction

Transformer training dynamics often exhibit a striking phenomenon: despite the high-dimensional embedding space, token representations tend to converge toward a small number of dominant

directions. In this section, we combine the results of the preceding sections to formalize this behavior by showing that under slowly varying attention patterns, token trajectories ultimately reduce to two-dimensional dynamics, with nearly all solutions recurrently approaching the dominant direction determined by the value matrix.

Assumption 4.1 (Block structure of the query-key matrix)**.** *In addition to Assumption 2.1, we will also use the following assumptions in this section:*

(i) The query-key matrix $\mathcal{T}(t) = Q^\top K(t)$ has the block structure

$$\mathcal{T}(t) = \begin{pmatrix} m_0(t) & 0 \\ 0 & m_1(t) I_{d-1} + \mathcal{B}(t) \end{pmatrix}, \qquad \mathcal{B}(t)^\top = -\mathcal{B}(t), \tag{8}$$

for some functions $m_0(t) > 0$ such that $m_1(t) < m_0(t) + \frac{1}{2\beta}$ for all $t \geq 0$.

(ii) Furthermore, $\lim_{t\to\infty}(V(t), \mathcal{T}(t)) = (V_\infty, \mathcal{T}_\infty)$, and $\lim_{t\to\infty} V'(t) = 0$, where

$$V_\infty = \mathit{diag}\{\lambda_{1,\infty}, \ldots, \lambda_{d,\infty}\}, \qquad \lambda_{1,\infty} > \lambda_{2,\infty} > \cdots > \lambda_{d,\infty}, \qquad \lambda_{1,\infty} > 0.$$

We note that the setting considered under Assumption 4.1 is substantially more general than the weight-tying framework studied in [37]. In particular, our formulation goes beyond weight tying by allowing the model parameters to depend on time. Under this assumption, we prove the following result.

Theorem 4.2 (See Theorem C.1 for a formal statement)**.** *Assume that $x_1(t), \ldots, x_{L-1}(t) \to \pm\xi_{\max}$, where $\xi_{\max}$ is the dominant eigen-direction of the limiting attention structure. Let us consider the evolution of a token $x_L(t) \in \mathbb{S}^{d-1}$ under self-attention with time-dependent weight matrices $V(t)$ and query-key product $\mathcal{T}(t) = Q(t)^\top K(t)$. Then aside from a single exceptional trajectory, every solution* recurrently approaches *the dominant state $\pm\xi_{\max}$.*

$$x_L(t_n) \to \pm\xi_{\max} \quad \textit{along some subsequence} \quad t_n \to \infty.$$

We further prove that the condition is unavoidable $\lambda_{\max} > 0$ in the following theorem.

Theorem 4.3 (Sharpness of the condition $\lambda_1 > 0$)**.** *Suppose that $\lambda_1 < 0$. Then there exist $\beta > 0$ and $\lambda_2 < \lambda_1$ such that, for $L = d = 2$, $Q = K = I_2$, $V = \mathrm{diag}(\lambda_1, \lambda_2)$, and $x_1(t) \equiv \vec{e}_1$, one can find a nonempty open set $U \subset \mathbb{S}^1$ with the property that for all $x_2(0) \in U$, the limit $\lim_{t\to\infty} x_2(t)$ exists and satisfies*

$$\lim_{t\to\infty} x_2(t) \notin \{\pm\vec{e}_1\}.$$

Intuitively speaking, Theorem 4.2 shows that tokens tends to the unit length eigenvectors $\pm\xi_{\max}$ of the value matrix $V(t)$ whenever Assumption 4.1 holds.

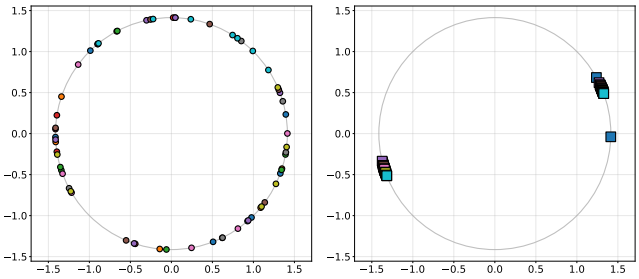
Remark 4.4 (Limitations of Assumption 4.1 in practical Transformers)**.** Assumption 4.1 is unlikely to hold exactly in practical pre-trained Transformers, since the value matrix $V(t)$ need not be symmetric or have time-invariant eigenvectors, and the key-query matrix $\mathcal{T}(t)$ need not possess the prescribed block structure. Nevertheless, this assumption covers the weight-tying setting considered in [37], and therefore yields a proof of Conjecture 1.1 under that setting. Moreover, our formulation goes beyond the weight-tying case by allowing the model parameters to be time-dependent.

5 Empirical Validation

In this section, we validate the emergence of the two-cluster regime through simulations with both single-head attention and in Grouped-Query Attention (GQA) [5], a more general mechanism. We then investigate the role of the value projection matrix, highlighting its importance in shaping token dynamics and model adaptation.

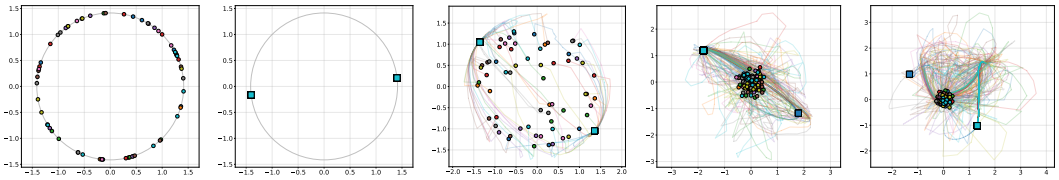
5.1 Validation on Grouped-Query Attention

We begin by validating the condition for the two-cluster regime described in our paper. To this end, we conduct simulations using single-head attention with varying hidden dimensions, including 2D, 3D, 32D, and 128D. Each configuration employs 200 time-dependent layers that satisfy the



(a) Initial positions. (b) Final positions.

Figure 1: Initial (circles) and final (squares) token positions after processing through weight-sharing causal GQA with RMSNorm blocks. Tokens organize into two opposing clusters on the circle.



(a) Initial positions of tokens in 2D. (b) Final positions of tokens in 2D. (c) Token trajectories in 3D. (d) Token trajectories in 32D. (e) Token trajectories in 128D.

Figure 2: Tokens initial (circle) and final (square) positions and trajectories across various choices of hidden dimension.

assumptions outlined in Theorem 4.2. As shown in Figure 2, the results exhibit behavior consistent with the predicted behavior, that is, token representations converge toward one of two opposing points on the sphere.

To demonstrate that this phenomenon also arises in more sophisticated attention mechanisms, we then run a simulation using grouped-query attention (GQA) [5], which shares key-value heads across groups of query heads to reduce KV-cache memory and accelerate inference. This attention mechanism is widely adopted in modern LLMs such as Llama3 [23] and Mistral [35]. We consider a model composed of 5000 weight-sharing blocks containing GQA and RMSNorm, operating in a 2-dimensional space, with 4 attention heads and a single KV head, each of dimension 2. The initial positions of 64 tokens in the sequence (Figure 1a) and all model parameters are randomly chosen. As illustrated in Figure 1b, the final token positions tend to organize into two opposing clusters on the circle.

5.2 Importance of the Value Projection Matrix

Prior studies highlight the role of the value projection matrix V in Transformer models, as it primarily affects information flow and training convergence [69, 76, 10]. This property influences both LLM pre-training [90, 51] and fine-tuning [33, 81]. From our token-dynamic perspective in this work, we offer a novel explanation for this importance.

By Theorem 4.2, the convergence points $\pm\xi_{\max}$ of the token dynamics are the eigenvectors associated with the maximum eigenvalue of V. These points emerge as stable attractors for every token in the sequence, dictating the long-term evolution of token representations. As a result, V plays a critical role in shaping token trajectories and thus significantly influences model behavior.

In line with this observation, ablation studies in [33] demonstrate that fine-tuning combinations including V yields better performance. The experiment adapts GPT-3 175B to the WikiSQL [89] and MultiNLI [82] datasets, examining which subsets of self-attention weight matrices should be adapted. The number of trainable parameters is kept equal across configurations by inversely scaling the fine-tuning rank with the number of weights. We replicate these results in Table 2 and observe that only fine-tuning Q and/or K fails to achieve optimal performance. This provides further evidence of V's importance in fine-tuning strategies.

Table 2: Validation accuracy ($\%, \uparrow$) on WikiSQL and MultiNLI when fine-tuning GPT-3 with LoRA applied to different subsets of self-attention weight matrices. All configurations use 18M trainable parameters. The standard deviation across random seeds is consistent for each dataset and reported in the first column. For the same number of fine-tuned weight types, combinations that include V yield higher accuracy. Results reproduced in [33].

Fine-tuning weight	Q	K	V	Q, K	Q, V
Fine-tuning rank	8	8	8	4	4
WikiSQL ($\pm 0.5\%$)	70.4	70.0	73.0	71.4	73.7
MultiNLI ($\pm 0.1\%$)	91.0	90.8	91.0	91.3	91.3

6 Conclusion

In this paper, we introduce a new mathematical framework that partially resolves the two-cluster conjecture for causal attention with symmetric value matrices under arbitrary initial token configurations. In addition to establishing global convergence to $\pm\xi_{\max}$, our analysis provides the first rigorous explanation of the dimensional reduction phenomenon, which had previously been supported primarily by empirical observations. Central to our approach is a new analytical tool–which we term instability-within-the-flow–motivated by stability theory in physics. This perspective yields a principled route toward the conjecture, and it elucidates how geometric, dynamical, and statistical mechanisms jointly shape token dynamics. Overall, our results offer a theoretically grounded and practically useful characterization of clustering behavior in modern large-scale sequence models.

Acknowledgments and Disclosure of Funding

This research is supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-012-SGIL). This research is also supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2023) (A-8002040-00-00, A-8002039-00-00), the NUS Presidential Young Professorship Award (A-0009807-01-00), the NUS Artificial Intelligence Institute–Seed Funding (A-8003062-00-00), and the Cross Faculty Grant 2025, CFG25-012 (A-8004460-00-00). T. T. Nguyen acknowledges support from the NSFC Excellent Young Scientists Fund (Overseas) and the Research Start-up Fund of Shanghai Jiao Tong University.

References

[1] Laziz U Abdullaev, Maksim Tkachenko, and Tan M Nguyen. Revisiting transformers with insights from image filtering and boosting. *SIAM Journal on Imaging Sciences*, 2025.

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

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

[4] Juan A Acebrón, Luis L Bonilla, Conrad J Pérez Vicente, Félix Ritort, and Renato Spigler. The kuramoto model: A simple paradigm for synchronization phenomena. *Reviews of modern physics*, 77(1):137–185, 2005.

[5] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore, December 2023. Association for Computational Linguistics.

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

[7] Albert Alcalde, Giovanni Fantuzzi, and Enrique Zuazua. Clustering in pure-attention hardmax transformers and its role in sentiment analysis. *arXiv preprint arXiv:2407.01602*, 2024.

[8] Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. In *International Conference on Learning Representations*, 2019.

[9] Han Bao, Ryuichiro Hataya, and Ryo Karakida. Self-attention networks localize when qk-eigenspectrum concentrates. *arXiv preprint arXiv:2402.02098*, 2024.

[10] Etienne Boursier and Claire Boyer. Softmax as linear attention in the large-prompt regime: a measure-based perspective, 2025.

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

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

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

[14] Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12020–12030, 2022.

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

[16] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[17] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics.

[18] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[20] 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*, pages 2793–2803. PMLR, 2021.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[22] Gbetondji JS Dovonon, Michael M Bronstein, and Matt J Kusner. Setting the record straight on transformer oversmoothing. *arXiv preprint arXiv:2401.04301*, 2024.

[23] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[24] Subhabrata Dutta, Tanya Gautam, Soumen Chakrabarti, and Tanmoy Chakraborty. Redesigning the transformer architecture with insights from multi-particle dynamical systems. *Advances in Neural Information Processing Systems*, 34:5531–5544, 2021.

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

[26] Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. *Advances in Neural Information Processing Systems*, 35:33054–33065, 2022.

[27] Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024.

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

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

[30] Xiaojun Guo, Yifei Wang, Tianqi Du, and Yisen Wang. Contranorm: A contrastive learning perspective on oversmoothing and beyond. *arXiv preprint arXiv:2303.06562*, 2023.

[31] Seung-Yeal Ha, Jeongho Kim, Jinyeong Park, and Xiongtao Zhang. Complete cluster predictability of the cucker–smale flocking model on the real line. *Archive for Rational Mechanics and Analysis*, 231:319–365, 2019.

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

[33] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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

[35] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[36] Amir Joudaki, Hadi Daneshmand, and Francis Bach. On the impact of activation and normalization in obtaining isometric embeddings at initialization. *Advances in Neural Information Processing Systems*, 36:39855–39875, 2023.

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

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

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

[40] Ulrich Krause et al. A discrete nonlinear and non-autonomous model of consensus formation. *Communications in difference equations*, 2000:227–236, 2000.

[41] Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics: January 23–29, 1975, Kyoto University, Kyoto/Japan*, pages 420–422. Springer, 1975.

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

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

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

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

[46] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *International Conference on Learning Representations*, 2017.

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

[48] 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. *arXiv preprint arXiv:1906.02762*, 2019.

[49] Sebastien Motsch and Eitan Tadmor. Heterophilious dynamics enhances consensus. *SIAM review*, 56(4):577–621, 2014.

[50] Ho Huu Nghia Nguyen, Tan Nguyen, Huyen Vo, Stanley Osher, and Thieu Vo. Improving neural ordinary differential equations with nesterov's accelerated gradient method. *Advances in Neural Information Processing Systems*, 35:7712–7726, 2022.

[51] Tam Nguyen, Tan Nguyen, and Richard Baraniuk. Mitigating over-smoothing in transformers via regularized nonlocal functionals. *Advances in Neural Information Processing Systems*, 36:80233–80256, 2023.

[52] Tam Minh Nguyen, Tan Minh Nguyen, Dung DD Le, Duy Khuong Nguyen, Viet-Anh Tran, Richard Baraniuk, Nhat Ho, and Stanley Osher. Improving transformers with probabilistic attention keys. In *International Conference on Machine Learning*, pages 16595–16621. PMLR, 2022.

[53] Tam Minh Nguyen, César A Uribe, Tan Minh Nguyen, and Richard Baraniuk. Pidformer: Transformer meets control theory. In *Forty-first International Conference on Machine Learning*, 2024.

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

[55] Tan Minh Nguyen, Tam Minh Nguyen, Hai Ngoc Do, Khai Nguyen, Vishwanath Saragadam, Minh Pham, Nguyen Duy Khuong, Nhat Ho, and Stanley Osher. Improving transformer with an admixture of attention heads. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[56] Tuan Nguyen, Hirotada Honda, Takashi Sano, Vinh Nguyen, Shugo Nakamura, and Tan Minh Nguyen. From coupled oscillators to graph neural networks: Reducing over-smoothing via a kuramoto model-based approach. In *International Conference on Artificial Intelligence and Statistics*, pages 2710–2718. PMLR, 2024.

[57] Stefan K Nielsen, Laziz U Abdullaev, Rachel S Teo, and Tan M Nguyen. Elliptical attention. *Advances in Neural Information Processing Systems*, 37:109748–109789, 2024.

[58] 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.

[59] Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36, 2024.

[60] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics.

[61] Duy-Tung Pham, An The Nguyen, Viet-Hoang Tran, Nhan-Phu Chung, Xin T Tong, Tan M Nguyen, and Thieu N Vo. Dynamical properties of tokens in self-attention and effects of positional encoding. *Advances in Neural Information Processing Systems*, 2025.

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

[63] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[64] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

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

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

[68] Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael T Schaub. Residual connections and normalization can provably prevent oversmoothing in gnns. *arXiv preprint arXiv:2406.02997*, 2024.

[69] Bingqing Song, Boran Han, Shuai Zhang, Jie Ding, and Mingyi Hong. Unraveling the gradient descent dynamics of transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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

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

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

[73] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.

[74] Rachel Teo and Tan Nguyen. Molex: Mixture of layer experts for fine-tuning with sparse upcycling. In *International Conference on Learning Representations*, volume 2025, pages 70264–70289, 2025.

[75] Rachel S Teo and Tan M Nguyen. Momentumsmoe: Integrating momentum into sparse mixture of experts. *Advances in Neural Information Processing Systems*, 37:28965–29000, 2024.

[76] Rachel S Teo and Tan M Nguyen. Unveiling the hidden structure of self-attention via kernel principal component analysis. *Advances in Neural Information Processing Systems*, 37:101393–101427, 2024.

[77] Matthew Thorpe, Tan Minh Nguyen, Hedi Xia, Thomas Strohmer, Andrea Bertozzi, Stanley Osher, and Bao Wang. Grand++: Graph neural diffusion with a source term. In *International Conference on Learning Representations*, 2022.

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

[79] Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In Tal Linzen, Grzegorz Chrupala, Yonatan Belinkov, and Dieuwke Hupkes, editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy, August 2019. Association for Computational Linguistics.

[80] Thieu N Vo, Tung D Pham, Xin T Tong, and Tan Minh Nguyen. Demystifying the token dynamics of deep selective state space models. *International Conference on Learning Representations (ICLR)*, 2025.

[81] Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. LoRA-pro: Are low-rank adapters properly optimized? In *The Thirteenth International Conference on Learning Representations*, 2025.

[82] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, June 2018.

[83] Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[84] Hedi Xia, Vai Suliafu, Hangjie Ji, Tan Nguyen, Andrea Bertozzi, Stanley Osher, and Bao Wang. Heavy ball neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 34:18646–18659, 2021.

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

[86] Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR, 2023.

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

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

[89] Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning, 2017.

[90] Zhanchao Zhou, Tianyi Wu, Zhiyun Jiang, Fares Obeid, and Zhenzhong Lan. Value residual learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28341–28356, Vienna, Austria, July 2025. Association for Computational Linguistics.

Appendix of "Two-Clustering Regime of Token Dynamics in Causal Attention"

Table of Contents

A Related Works **17**

B Setting **18**

C Main Results **20**

D Deriving the Leading Dynamics **23**

D.1 Equations for the Coordinates of the Lth Token 24

D.2 Hierarchical Suppression Mechanism of Transverse Coordinates 25

D.3 Proof of Theorem C.7 31

E Instability within the Flow Mechanisms **31**

F Time-dependent $Q(t), V(t), K(t)$ **36**

G Technical Lemmas **38**

A Related Works

ResNets and Neural ODEs Our work relates to the literature that views deep neural networks (DNNs) through a dynamical-systems perspective. In particular, standard DNNs can be interpreted as discrete-time dynamical processes in which representations propagate sequentially across layers [42]. Residual networks (ResNets) [13, 32, 39] provide a canonical example of this viewpoint, while their continuous-time analogues–neural ordinary differential equations (neural ODEs) [44, 88, 84, 43, 50, 77, 67, 15]–make the limiting transition to continuous time explicit. Under this formulation, the residual update $z = x + y(x, \theta)$ is understood as a discretization of a flow governed by an ordinary differential equation, which maps an initial condition $x(0)$ to a terminal state $x(T)$. A substantial body of work investigates the approximation, interpolation, and controllability properties of these architectures [44, 88, 43, 70, 67, 15, 75, 50].

Oversmoothing and Rank Collapse Transformers are susceptible to two prominent degeneracies: oversmoothing and rank collapse. Oversmoothing drives token representations toward indistinguishability, thereby reducing expressive capacity. In particular, Nguyen et al. [51, 56], Wu et al. [83], Dovonon et al. [22], and Chen et al. [14] show that self-attention alone is insufficient to avert this phenomenon. Proposed mitigations include *ContraNorm*, a normalization scheme motivated by contrastive learning [30], as well as residual connections coupled with normalization, which have been empirically shown to alleviate oversmoothing [68].

Rank collapse, by contrast, refers to the degeneration of hidden states onto a low-dimensional subspace. A growing body of work [20, 26, 58, 36, 86, 59, 53, 9] identifies this effect as a key bottleneck and develops stabilization mechanisms aimed at preserving representation diversity and improving downstream performance.

Token Dynamics and Clustering Effects A complementary line of work models sequence architectures, including Transformers and Mamba, as interacting particle systems. Lu et al. [48] interpret self-attention as a numerical solver for a convection–diffusion equation, whereas Dutta et al. [24] introduce *TransEvolve*, a particle-inspired design that reduces computational complexity. More broadly, this viewpoint places Transformer dynamics within the landscape of nonlinear collective-behavior models, such as flocking [31], synchronization [41, 4], and consensus dynamics [40, 49].

Within this dynamical framework, the asymptotic behavior of token trajectories has been investigated through clustering phenomena. Geshkovski et al. [29] established convergence to limiting objects shaped by the initial conditions, and subsequent studies [28, 27] identified metastable multi-cluster configurations that can persist for extended times before ultimately merging into a single cluster. Alcalde et al. [7] further demonstrated asymptotic convergence to clustered equilibria for hardmax Transformers. Pham et al. [61] investigated how different forms of positional encoding - specifically absolute and rotary - affect these dynamical regimes. These results have been recently extended to causal attention in [37], all type of attentions in [2]. Furthermore, Vo et al. [80] extended the analysis to the dynamical properties of tokens in a deep selective state-space model, and based on that, suggested improvements to the model by excluding negatively impactful scenarios and reordering tokens using importance scores.

B Setting

In this paper, we study the dynamical properties of the system

$$\partial_t x_i = \sum_{j=1}^{i} a_{ij}(t)\, \mathbf{P}_{x_i}(V x_j), \tag{9}$$

where $x_i : [0,\infty) \to \mathbb{S}^{d-1}$ denotes the i-th token evolving on the unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, with $d \geq 2$. The projection operator $\mathbf{P}$ is defined by

$$\mathbf{P}_y(x) = x - (x \cdot y)y, \qquad y \in \mathbb{S}^{d-1}.$$

The matrix $V \in \mathbb{R}^{d\times d}$ is referred to as the *value matrix*. For notational simplicity, we assume throughout that

$$V(t) = \mathrm{diag}\{\lambda_1(t), \lambda_2(t), \ldots, \lambda_d(t)\},$$

so that the eigen-basis coincides with the canonical basis $\{\vec{e}_1, \ldots, \vec{e}_d\}$ and

$$\lambda_1(t) > \lambda_2(t) > \cdots > \lambda_d(t) \quad \text{for all} \quad t \geq 0. \tag{10}$$

We denote by $Q(t), K(t) \in \mathbb{R}^{d\times d}$ the *query* and *key* matrices, and define

$$Z_i(t) = \sum_{j=1}^{i} e^{\beta\langle Q x_i(t), K x_j(t)\rangle},$$

where $\beta > 0$ is a fixed parameter. The coefficients a_{ij} are then given by

$$a_{ij}(t) = a_{ij}(x_i(t), x_j(t)) = \frac{e^{\beta\langle Q x_i, K x_j\rangle}}{Z_i(t)}.$$

Define $Q(t)^\top K(t) = \mathcal{T}(t)$, so that

$$a_{ij}(t) = \frac{e^{\beta\langle x_i, \mathcal{T} x_j\rangle}}{Z_i(t)}, \qquad Z_i(t) = \sum_{j=1}^{i} e^{\beta\langle x_i, \mathcal{T} x_j\rangle}.$$

In particular, the evolution equation for the L-th token reads

$$\partial_t x_L = \sum_{j=1}^{L} a_{Lj}\, \mathbf{P}_{x_L}(V x_j).$$

The dynamical systems perspective—in which tokens are viewed as interacting particles—has attracted considerable interest from both the machine learning and mathematical communities. This framework provides a mathematically rigorous lens for understanding how Transformers internally organize information through token interactions. The synthesis of geometric, dynamical, and statistical insights makes the characterization of token dynamics both theoretically elegant and crucial for interpreting modern large-scale models.

Token Dynamics for Causal Attention. Token dynamics for causal attention have been studied recently in [37] for the time-independent setting and in [2] for the time-dependent setting. Since both works are closely related to our analysis, we briefly review their main results below.

- *Time-Independent & Causal Attention:* Karagodin et al. [37] established the convergence of the first token $x_1(t)$ for arbitrary constant value matrices V and proved global convergence of all tokens to $x_1(0)$ in the special case $V = I_d$. Their analysis leverages explicit closed-form solutions and tools from generic dynamical systems (Schub genericity theorem). Importantly, they also predicted and conjectured a dimensional deduction mechanism based on experimental observations of real Transformer value matrices. The work of Karagodin et al. [37] further conjectured this dimensionality reduction mechanism and highlighted its practical implications:

  > *An important practical consequence of these conjectures is that they describe a dimensionality reduction mechanism induced by the value matrix V. Token trajectories converge to $L \cap \mathbb{S}^{d-1}$ and, once restricted to this set, evolve as if governed by the scalar matrix λI_d on the smaller sphere $\mathbb{S}^{\dim L-1}$. For the pre-trained model of Lan et al. (2020), the spectra of the value matrices are shown in Figure 6, where some heads even exhibit a negative top eigenvalue $\lambda_{\max}$. A natural direction for future work is to analyze value matrices V arising in real-world networks and relate their leading eigenspaces to the semantic roles of layers and tokens.*

- *Fully Time-Dependent Case:* Abella et al. [2] analyzed the causal model with fully time-dependent matrices $V(t), K(t), Q(t)$. Assuming all initial tokens lie within a single hemisphere of $\mathbb{S}^{d-1}$, they employed a Lyapunov argument to prove global convergence to a single cluster $\xi_{\max}$, fully resolving the single-cluster regime. The hemisphere condition is preserved under the dynamics, enabling a clean analytical framework, but restricts admissible initial configurations rarely satisfied in practice. However, their results do not resolve the conjecture established in [37].

In contrast, our work offers a mathematically rigorous lens for explaining how Transformers internally organize information through token interactions. To this end, we introduce a new analytical tool–which we term instability-within-the-flow–inspired by stability theory in physics. This framework enables us to partially resolve the conjecture and clarifies the synthesis of geometric, dynamical, and statistical perspectives underlying token dynamics. As a result, our characterization is both theoretically principled and practically informative for interpreting modern large-scale models.

Structure of the proofs and key ideas . Our first step is to derive the leading-order dynamics of the Lth token, assuming that all preceding tokens have converged to $\pm\xi_{\max}$. This analysis is carried out in Section D and substantially simplifies the dynamics of the final token. Next, for notational convenience, we assume $\xi_{\max} = \vec{e}_1$, which can be achieved without loss of generality by applying a unitary rotation to the solution. We then explicitly derive the coordinate-wise evolution equations for the Lth token, as stated in Lemma D.4. These steps establish a natural and systematic setup for studying the dynamics of the final token.

Our first key result is presented in Theorem C.6. Roughly speaking, we uncover a high-dimensional reduction mechanism in which the dynamics collapse to *exactly two coordinates* for arbitrary matrices Q and K. Our key idea is to *separate the direction dynamics from the magnitude dynamics*: once we factor out the overall magnitude of a block of coordinates, the remaining dynamics describe a competition over which direction the token points to. Because different coordinates have systematically different growth advantages (encoded by the ordered of the eigenvalues (10) of the value matrix V), this directional dynamics has a built-in drift that continuously sharpens the representation, pushing it toward the boundary where some coordinates vanish. As a result, the system cannot remain in a persistent mixed state: either a given coordinate loses mass and is suppressed, or it survives and eventually dominates, eliminating all weaker competitors. The only subtlety is noise inherited from earlier tokens, but as long as this residual signal decays faster than the scale of the current competition, it cannot prevent concentration. In effect, the dynamics behave like an expert-selection process: coordinates are tested one by one, and any direction that remains consistently active must ultimately win decisively. This explains why, despite the ambient dimension, the token collapses onto

a low-dimensional structure driven by a small number of dominant features. The full detail of the proof of Theorem C.6 is presented in Section D.2.

Based on Theorem C.6, we develop our theory in two separate directions:

- Structured time-dependent $Q(t), K(t)$ (Theorem C.1). Under suitable block-structure assumptions on $Q(t)$ and $K(t)$, we derive an effective ODE governing the evolution of the projection $x_L \cdot \xi_{\max}$ (later denoted by y_1), treating all remaining coordinates as perturbations. This reduction is carried out in Theorem D.10. Formally, the dynamics of y_1 take the form

$$\partial_t y_1 = (1 - y_1^2)\, M(t, y_1) + o(1), \tag{11}$$

  where $o(1)$ denotes a quantity that vanishes as $t \to \infty$. Since the equation is non-autonomous and the right-hand side is not monotone for the two-cluster regime $\{\xi_{\max}, -\xi_{\max}\}$, we cannot linearize around a stationary solution and apply the stable manifold theorem as in [37, 28], nor use the Lyapunov method as in [7]. Instead, the right-hand side of (11) admits a "moving equilibrium" in the interior of $(-1, 1)$—a phenomenon already visible in the simpler setting (37). This behavior corresponds precisely to the moving-hilltop analogy in the stability theory of physical systems, which we discuss in Section E. Our key insight is that, under suitable time-varying assumptions on the parameter matrices, this unstable equilibrium cannot trap the solution y_1 in the interior of $(-1, 1)$. Consequently, x_L must converge to one of the two clusters: $\xi_{\max}$ (corresponding to $y_1 = 1$) or $-\xi_{\max}$ (corresponding to $y_1 = -1$). The full analytical framework is developed in Theorem E.1 and Section F. Importantly, our analysis is not limited to establishing convergence to the two-cluster configuration. Rather, it also provides insight into alternative clustering behaviors that may arise in practice and elucidates the instability mechanisms that can unlock such dynamics. In particular, we present an example in which oscillatory drifting of the unstable equilibrium within the interior can trap the L-th token and lead to dynamics distinct from convergence to $\pm\xi_{\max}$. This demonstrates that our assumptions are essentially sharp; see Proposition E.3.
- General stationary Q, K (Theorem C.7). Once the parameter matrices Q and K have stabilized, the dimension-reduction result of Theorem D.5 shows that, asymptotically, the token dynamics are confined to the manifold

$$\mathbb{S}^1 \times \{0\}^{d-2}.$$

  This residual degree of freedom corresponds to a phase (or orientation) ambiguity in the embedding space. A natural question is whether the token may continue to rotate along this manifold indefinitely, even after the training dynamics appear to have converged. Using tools from complex analysis, we prove that the set of admissible long-time limit points on $\mathbb{S}^1$ is necessarily discrete. Consequently, the token representation must converge to a single fixed configuration (see Lemma G.6). In particular, apparent stabilization of the parameter matrices Q and K implies convergence of the token dynamics. We refer to Theorem C.7 and Section D.3 for the complete argument.

C Main Results

Theorem C.1. *Assume that $\mathcal{T}(t) = Q^\top K(t)$ has the block structure*

$$\mathcal{T}(t) = \begin{pmatrix} m_0(t) & 0 \\ 0 & m_1(t) I_{d-1} + \mathcal{B}(t) \end{pmatrix}, \qquad \mathcal{B}(t)^\top = -\mathcal{B}(t), \tag{12}$$

for some functions $m_0(t) > 0$ and

$$m_1(t) < m_0(t) + \frac{1}{2\beta} \qquad \textit{for all } t \geq 0. \tag{13}$$

Assume further that

$$\lim_{t\to\infty} (V(t), \mathcal{T}(t)) = (V_\infty, \mathcal{T}_\infty), \qquad \lim_{t\to\infty} V'(t) = 0,$$

where

$$V_\infty = \textit{diag}\{\lambda_{1,\infty}, \ldots, \lambda_{d,\infty}\}, \qquad \lambda_{1,\infty} > \lambda_{2,\infty} > \cdots > \lambda_{d,\infty}, \qquad \lambda_{1,\infty} > 0.$$

Suppose that

$$x_i(t) \longrightarrow \pm\vec{e}_1 \qquad \text{as } t \to \infty, \quad i = 1, \dots, L-1.$$

Then either of the following must hold

- *There exists a sequence $t_n \to \infty$ such that $x_L(t_n) \to \pm\vec{e}_1$*
- *Then there exists an index $\ell \in \{2, \dots, d\}$ such that the following holds. For every solution $x_L(t)$, all transverse components except one decay,*

$$x_L(t) \cdot \vec{e}_j \longrightarrow 0 \qquad \text{as } t \to \infty, \quad \forall\, j \neq 1, \ell, \tag{14}$$

so that the long–time dynamics collapses onto the two–dimensional subspace $\operatorname{span}\{\vec{e}_1, \vec{e}_\ell\}$.

Within this reduced dynamics, a conditional uniqueness property holds: if $x_L^{(1)}(t)$ and $x_L^{(2)}(t)$ are two solutions (possibly with different initial data) whose transverse components coincide,

$$x_L^{(1)}(t) \cdot \vec{e}_j \equiv x_L^{(2)}(t) \cdot \vec{e}_j \qquad \forall\, t \geq 0,\ \forall\, j \neq 1, \ell, \tag{15}$$

then, except possibly for a single exceptional trajectory in this constraint class, the corresponding solution satisfies

$$x_L(t_n) \longrightarrow \pm\vec{e}_1 \qquad \text{along some sequence } t_n \to \infty.$$

In particular, apart from this exceptional orbit, all solutions recurrently return *to the dominant direction in the infinite–time limit.*

Remark C.2. We show that, after stabilization of the earlier tokens, the dynamics of the remaining token $x_L(t)$ admits only two long–time possibilities. Either the trajectory visits the dominant direction $\vec{e}_1$ infinitely many times as $t \to \infty$, or the dynamics collapses onto a two–dimensional invariant subspace. Within this reduced two–dimensional regime, and under a natural conditional uniqueness assumption, all trajectories except possibly a single exceptional one recurrently return to $\vec{e}_1$. In this sense, our theorem provides a principled theoretical explanation for the emergence of $\vec{e}_1$ as the dominant direction, even in the presence of time–varying attention.

Remark C.3 (On the structural assumption for $\mathcal{T}(t) = Q^\top K(t)$)**.** The block structure in (12) ensures that the distinguished direction $\vec{e}_1$, corresponding to the leading eigenvalue of V_∞, remains an invariant axis under the $Q^\top K(t)$–induced interaction. No mixing occurs between $\vec{e}_1$ and the transverse coordinates; all nontrivial coupling is confined to the orthogonal $(d-1)$–dimensional subspace through the skew–symmetric component $\mathcal{B}(t)$. The coefficients $m_0(t)$ and $m_1(t)$ control the relative strength of the interaction along and perpendicular to $\vec{e}_1$, while $\mathcal{B}(t)$ contributes only rotational effects within the transverse space. This decomposition isolates the dominant direction dynamically and is precisely the structure needed for the reduction to a one–dimensional effective equation along y_1.

Remark C.4. The restriction (15) is not artificial. By part (i), all transverse coordinates except $\vec{e}_\ell$ decay asymptotically and are therefore slaved to the dynamics in $\operatorname{span}\{\vec{e}_1, \vec{e}_\ell\}$. The conditional uniqueness statement in part (ii) should thus be understood as a fiber–wise result: for each fixed transverse profile, the reduced two–dimensional dynamics admits at most one non–dominant trajectory.

Remark C.5. The structural assumptions

$$\lambda_1(t) > 0, \qquad m_0(t) > 0, \qquad m_1(t) < m_0(t) + \frac{1}{2\beta}$$

can in fact be replaced by the more general asymptotic condition

$$\lambda_{1,\infty}\, \beta m_{0,\infty} L \;+\; (\lambda_{1,\infty} - \lambda_{k,\infty}) \min\Big\{1,\ 1 - 2\beta(m_{1,\infty} - m_{0,\infty})\Big\} \;>\; 0, \qquad \text{for all } k \geq 2.$$

However, to keep the presentation streamlined and the exposition focused, we will restrict attention to the simpler set of hypotheses stated in the theorem, under which all essential ideas already appear transparently.

Theorem C.6 (Hierarchical Suppression Mechanism)**.** *Let $y(t) \in \mathbb{S}^{d-1}$ be the L-th token, and let $\varepsilon_{L-1}(t)$ denote the residual error contributed by the first $L-1$ tokens. For each $\ell \in \{2, \dots, d\}$, define*

$$R_\ell(t) = \Big(\sum_{j=\ell}^{d} y_j(t)^2\Big)^{1/2}, \qquad \kappa_\ell(t) = \frac{\varepsilon_{L-1}(t)}{R_\ell(t)}.$$

Suppose that for some $\ell \geq 2$,

$$\kappa_\ell(t) \longrightarrow 0, \qquad (t \to \infty). \tag{16}$$

Then exactly one of the following two alternatives occurs:

(A) y_ℓ becomes the dominant transverse direction.

$$y_j(t) \longrightarrow 0, \qquad \textit{for all } j \neq 1, \ell.$$

In this case the suppression process terminates at level ℓ.

(B) y_ℓ is suppressed.

$$y_\ell(t) \longrightarrow 0.$$

Here the transverse vector $(y_{\ell+1}, \ldots, y_d)$ remains nontrivial, and the same mechanism applies at the next level $\ell + 1$ using $\kappa_{\ell+1}(t)$.

Thus, whenever the error-to-transverse ratio becomes negligible at level ℓ, the dynamics eliminate all but one of the coordinates $y_\ell, \ldots, y_d$. Iterating this reduction yields a unique surviving transverse coordinate y_k, and the trajectory collapses onto the (ℓ)-dimensional subsystem spanned by

$$(y_1, y_2, \ldots, y_{\ell-1}, y_k).$$

Theorem C.7. *Let Q and K be arbitrary matrices. Assume that $x_i(t) \to \pm\vec{e}_1$ for all $i \leq L-1$, and that the DDSC condition* (34) *holds for the final L^{th} token. Then there exists a finite subset $\mathcal{S} \subset \mathbb{S}^{d-1}$ such that, for almost every initial condition, the trajectory $x_L(t)$ converges to some $y_\star \in \mathcal{S}$ as $t \to \infty$.*

Moreover, if the limit $y_\star$ is not equal to $\pm\vec{e}_1$, then $y_\star$ must lie in a two–dimensional coordinate plane and take the form

$$(\alpha_1, 0, \ldots, 0, \alpha_\ell, 0, \ldots, 0), \qquad \alpha_1^2 + \alpha_\ell^2 = 1,$$

that is, only the first and the ℓ-th coordinates are nonzero. In this case the limit point $y_\star$ satisfies the scalar algebraic condition

$$G(y_\star) = 0, \qquad \textit{for some } \ell \in \{2, \ldots, d\},$$

where

$$G(y) = e^{\beta\, y\cdot \mathcal{T} y}\,(\lambda_1 - \lambda_\ell) y_1 \;+\; \lambda_1 \big[N_1 e^{\beta\, y\cdot \mathcal{T}\vec{e}_1} - N_2 e^{-\beta\, y\cdot \mathcal{T}\vec{e}_1}\big]\,.$$

Remark C.8. Theorem C.7 reveals an unexpectedly rigid structure in the high-dimensional token dynamics. Although the system is nonlinear and a priori allows infinitely many potential limit configurations, the theorem shows that the long-time behavior is governed by a *finite* set of isolated equilibria. This finiteness holds for *all* choices of Q and K and for *almost every* initial condition. Even more striking is that every nontrivial limit point $y_\star$ is forced to lie on a two-dimensional coordinate plane and is completely characterized by the scalar equation $G(y_\star) = 0$. Thus, a high-dimensional nonlinear problem collapses to solving a single one-dimensional equation and checking the sign of the eigenvalues of a 2×2 linearized system.

Theorem C.9 (Sharpness of the condition $\lambda_1 > 0$)**.** *Suppose $\lambda_1 < 0$. Then there exist $\beta > 0$ and $\lambda_2 < \lambda_1$ such that, for $L = d = 2$, $\mathcal{T} = I_2$, and $x_1(t) \equiv \vec{e}_1$, one can find a nonempty open set $U \subset \mathbb{S}^1$ with the property that for all $x_2(0) \in U$, the limit*

$$\lim_{t\to\infty} x_2(t)$$

exists and satisfies

$$\lim_{t\to\infty} x_2(t) \notin \{\pm\vec{e}_1\}.$$

Proof. We write $x_2(t) = x(t)\vec{e}_1 + y(t)\vec{e}_2$,. From the equation $\partial_t x_2 = a\,\mathbf{P}_{x_2}(Vx_2) + b\lambda_1 \mathbf{P}_{x_2}(\vec{e}_1)$, and the fact that $x_2 \in \mathbb{S}^1$ for all time, we deduce

$$x'(t) = (1 - x^2)\,\frac{g(x)}{e^\beta + e^{\beta x}}, \quad g(x) = e^\beta(\lambda_1 - \lambda_2)x + \lambda_1 e^{\beta x}.$$

Thus, the first coordinate of x_2 satisfies a scalar ODE taking values in $(-1, 1)$. In particular, if g admits two distinct roots $r_1, r_2 \in (-1, 1)$, then by the vector field method we obtain

$$\lim_{t\to\infty} x(t) \in \{r_1, r_2\}$$

whenever the trajectory $x(t)$ remains in the interval (r_1, r_2). We now show that such roots exist when $\lambda_1 < 0$ and β is chosen appropriately. Suppose $r \in (-1, 1)$ is a root of g. Since $\lambda_1 < 0$ and $\lambda_1 - \lambda_2 > 0$, it follows that $r > 0$. The root condition $g(r) = 0$ is equivalent to

$$c(\beta r) = \gamma, \qquad c(x) = xe^{-x}, \qquad \gamma = \frac{|\lambda_1|\beta}{e^{\beta}(\lambda_1 - \lambda_2)} = \frac{|\lambda_1|\beta}{e^{\beta}(|\lambda_2| - |\lambda_1|)}. \tag{17}$$

Since $c'(x) = e^{-x}(1 - x)$, the function c is strictly increasing on $(0, 1)$ and strictly decreasing on $(1, \infty)$. We will show that when $\beta > 1$ and $\gamma \in (\beta e^{-\beta}, e^{-1})$, the equation $c(x) = \gamma$ has two solutions, one solution in $(0, \beta^{-1})$ and one solution in $(\beta^{-1}, 1)$.
As $r \in (0, 1)$, we examine solutions of $c(\bar{r}) = \gamma$ with $\bar{r} \in (0, \beta)$.
Indeed, because $\gamma \in (c(0), c(1)) = (0, e^{-1})$ and c is increasing on $(0, 1)$, there exists a solution $\bar{r}_1 \in (0, 1)$. On $(1, \beta)$, since c is decreasing and $\gamma \in (c(\beta), c(1)) = (\beta e^{-\beta}, e^{-1})$, there exists another solution $\bar{r}_2 \in (1, \beta)$.
Consequently,

$$g(r_1) = g(r_2) = 0, \qquad r_1 = \tfrac{\bar{r}_1}{\beta} \in (0, \beta^{-1}), \qquad r_2 = \tfrac{\bar{r}_2}{\beta} \in (\beta^{-1}, 1).$$

Now if we take any data $x(0) = x_2 \cdot \vec{e}_1 \in (r_1, r_2)$, we have $x(t) \to x_\star \in \{r_1, r_2\}$. This completes the proof. □

D Deriving the Leading Dynamics

We define the error function

$$\varepsilon_{L-1}(t) = \sum_{j=1}^{L-1} \left\| x_j(t) - s_j \vec{e}_1 \right\|$$

Our inductive hypothesis for the first $L - 1$ tokens is

$$\varepsilon_{L-1}(t) \longrightarrow 0 \qquad \text{as } t \to \infty, \tag{18}$$

Here s_j simply records *which cluster* the token x_j converges to: $+1$ corresponds to the "top" copy of $\vec{e}_1$ and -1 to the "bottom" copy $-\vec{e}_1$. To simplify the notation, in this paper we denote

$$y = x_L$$

to be the Lth token. Our goal is to identify the leading dynamics of y. The main results are stated in Theorem C.6. We assume (18). The evolution equation for $y = x_L$ reads

$$\partial_t y \;=\; a_{LL}\mathbf{P}_y(Vy) + \sum_{j=1}^{L-1} a_{Lj}\mathbf{P}_y(Vx_j). \tag{19}$$

Here

$$a_{LL} = \frac{e^{\beta\langle y, \mathcal{T}y\rangle}}{Z_L(t)}, \qquad a_{Lj} = \frac{e^{\beta\langle y, \mathcal{T}x_j\rangle}}{Z_L(t)},$$

with

$$Z_L(t) = e^{\beta\langle y, \mathcal{T}y\rangle} + \sum_{j=1}^{L-1} e^{\beta\langle y, \mathcal{T}x_j\rangle}.$$

To extract the leading dynamics of $y \in \mathbb{S}^{d-1}$, we introduce

$$Z(y) = e^{\beta y \cdot \mathcal{T}y} + \sum_{j=1}^{L-1} e^{s_j \beta y \cdot \mathcal{T}\vec{e}_1},$$

and define

$$a(y) = \frac{e^{\beta y \cdot \mathcal{T}y}}{Z(y)}, \qquad b(y) = \frac{1}{Z(y)} \sum_{j=1}^{L-1} s_j e^{s_j \beta y \cdot \mathcal{T}\vec{e}_1}. \tag{20}$$

Remark D.1. The functions Z, a, b are the functions $Z_L, a_{LL}, \sum_{j=1}^{L-1} a_{Lj}$ evaluated at $x_L = y, x_j = s_j\vec{e}_1$.

Lemma D.2. *Let N_1 denote the number of tokens among $\{x_1, x_2, \dots, x_{L-1}\}$ converging to $\vec{e}_1$, and N_2 the number converging to $-\vec{e}_1$. Then*

$$\begin{aligned} Z(y) &= e^{\beta y\cdot \mathcal{T} y} + N_1 e^{\beta y\cdot \mathcal{T}\vec{e}_1} + N_2 e^{-\beta y\cdot \mathcal{T}\vec{e}_1}, \\ a(y) &= \frac{e^{\beta y\cdot \mathcal{T} y}}{Z(y)}, \qquad b(y) = \frac{1}{Z(y)}\big(N_1 e^{\beta y\cdot \mathcal{T}\vec{e}_1} - N_2 e^{-\beta y\cdot \mathcal{T}\vec{e}_1}\big). \end{aligned} \tag{21}$$

Proof. This follows directly from the definitions of a, b, and Z. □

D.1 Equations for the Coordinates of the Lth Token

We decompose

$$y = \sum_{i=1}^{d} y_i \vec{e}_i, \qquad y_i(t) = y(t)\cdot \vec{e}_i. \tag{22}$$

Proposition D.3. *The dynamics* (19) *can be rewritten in the form*

$$\partial_t y \;=\; a(y)\,\mathbf{P}_y(Vy) \;+\; b(y)\lambda_1\,\mathbf{P}_y(\vec{e}_1) \;+\; \vec{g}(t), \tag{23}$$

where $\vec{g}(t)$ is an error term satisfying

$$|\vec{g}(t)| \lesssim \varepsilon_{L-1}(t).$$

Proof. Starting from the equation (19) and applying the inductive hypothesis (18), we obtain the representation (23), where the error term $\vec{g}(t)$ is given explicitly by

$$\vec{g}(t) = \;(a_{LL} - a(y))\mathbf{P}_y(Vy) + \sum_{j=1}^{L-1}\Big[a_{Lj} - \frac{s_j e^{s_j \beta y\cdot \mathcal{T}\vec{e}_1}}{Z(y)}\Big]\mathbf{P}_y(Vx_j) + \sum_{j=1}^{L-1} a_{Lj}\,\mathbf{P}_y\big(V(x_j - s_j\vec{e}_1)\big).$$

Since $Z(y)$ and Z_L are uniformly bounded below by a positive constant, it remains only to show that

$$\big|e^{\beta y\cdot \mathcal{T} x_j} - e^{s_j \beta\, y\cdot \mathcal{T}\vec{e}_1}\big| \;\longrightarrow\; 0, \qquad t\to\infty.$$

For any $X, Y \in [-M, M]$ we have the elementary bound

$$|e^X - e^Y| \;\le\; e^M\,|X - Y|.$$

Applying this with

$$X = \beta\, y\cdot \mathcal{T} x_j, \qquad Y = s_j \beta\, y\cdot \mathcal{T}\vec{e}_1,$$

gives

$$|e^{\beta y\cdot \mathcal{T} x_j} - e^{s_j \beta y\cdot \mathcal{T}\vec{e}_1}| \;\le\; e^{\beta|\mathcal{T}\vec{e}_1|}\,|\mathcal{T}|\,\beta\,|y|\,|x_j - s_j\vec{e}_1|.$$

This completes the proof. □

We now decompose $\vec{g}(t)$ as

$$\vec{g}(t) = \sum_{i=1}^{d} g_i(t)\vec{e}_i, \quad g_i(t) = \vec{g}(t)\cdot\vec{e}_i$$

Lemma D.4. *Let y be a solution of* (23)*. Then the first coordinate evolves according to*

$$y_1'(t) \;=\; a\big(\lambda_1 - Vy\cdot y\big)y_1 \;+\; \lambda_1 b(1 - y_1^2) \;+\; g_1(t), \tag{24}$$

while for each $2 \le i \le d$, the coordinate y_i satisfies

$$y_i'(t) \;=\; a\big(\lambda_i - Vy\cdot y\big)y_i \;-\; (\lambda_1 b\, y_1)\, y_i \;+\; g_i(t). \tag{25}$$

Proof. The identities (24) and (25) then follow directly by taking the inner product of (23) with $\vec{e}_i$. Here we use the identities

$$\mathbf{P}_y(Vy)\cdot\vec{e}_i = (\lambda_i - Vy\cdot y)y_i \quad (i\ge 1), \quad \mathbf{P}_y(\vec{e}_1)\cdot\vec{e}_1 = 1 - y_1^2, \quad \mathbf{P}_y(\vec{e}_1)\cdot\vec{e}_i = -y_1 y_i \quad (i\ge 2).$$

□

D.2 Hierarchical Suppression Mechanism of Transverse Coordinates

In this section, we reveal a striking dimensionality–reduction phenomenon in the evolution of the final token $y(t) \in \mathbb{R}^d$. Although the dynamics formally reside in a high-dimensional space, the long-time behavior is effectively governed by only *a handful of intrinsic degrees of freedom*. Put differently, despite the ambient dimension, the token behaves as though it were navigating a much lower-dimensional manifold. To formalize this structure, for each $\ell \in \{2, \dots, d\}$ we define

$$R_\ell(t) = \left(\sum_{j=\ell}^{d} y_j^2(t) \right)^{1/2}, \qquad y_{(\geq \ell)}(t) = (y_\ell(t), \dots, y_d(t)).$$

Thus $R_\ell(t)$ measures the magnitude of the final $d - \ell + 1$ coordinates of the token:

$$y(t) = \Big(y_1(t), \dots, y_{\ell-1}(t), \underbrace{y_\ell(t), \dots, y_d(t)}_{y_{(\geq \ell)}(t)} \Big).$$

We further introduce the *error-to-transverse ratio*

$$\kappa_\ell(t) = \frac{\varepsilon_{L-1}(t)}{R_\ell(t)}, \tag{26}$$

which compares the upstream residual error—generated by the first $L-1$ tokens—to the transverse radius of the current token. One can easily see that

$$\kappa_2(t) \leq \kappa_3(t) \leq \cdots \leq \kappa_d(t)$$

for all $t \geq 0$. We recall Theorem C.6:

Theorem D.5 (Hierarchical Suppression Mechanism)**.** *Let $y(t) \in \mathbb{S}^{d-1}$ be the L-th token, and let $\varepsilon_{L-1}(t)$ denote the residual error contributed by the first $L-1$ tokens. For $\ell \in \{2, \dots, d\}$, suppose that,*

$$\kappa_\ell(t) \longrightarrow 0, \qquad (t \to \infty). \tag{27}$$

Then exactly one of the following two alternatives occurs:

(A) y_ℓ becomes the dominant transverse direction.

$$y_j(t) \longrightarrow 0, \qquad \textit{for all } j \neq 1, \ell.$$

In this case the suppression process terminates at level ℓ.

(B) y_ℓ is suppressed.

$$y_\ell(t) \longrightarrow 0.$$

Key Ideas. To convey the core idea, let us begin with the simplest nontrivial example: $L = 2$ and $d = 3$. In this case the first token is fixed, $x_1(t) \equiv \vec{e}_1$, and we focus on the evolution of the second token $y = x_2(t)$. From Lemma D.4, the coordinates y_2 and y_3 satisfy

$$y_2' = a(\lambda_2 - Vy \cdot y)y_2 - (\lambda_1 b\, y_1)y_2, \qquad y_3' = a(\lambda_3 - Vy \cdot y)y_3 - (\lambda_1 b\, y_1)y_3. \tag{28}$$

Introducing polar coordinates $y_2 = R\cos\theta$, $y_3 = R\sin\theta$, we compute

$$\partial_t \begin{pmatrix} y_2 \\ y_3 \end{pmatrix} = R'(t) \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} + R\,\theta'(t) \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}.$$

Taking the dot product with the perpendicular vector $(-y_3, y_2)$ eliminates the $R'(t)$ term and isolates the angular dynamics:

$$R^2\theta' = \partial_t \begin{pmatrix} y_2 \\ y_3 \end{pmatrix} \cdot \begin{pmatrix} -y_3 \\ y_2 \end{pmatrix} = a(\lambda_3 - \lambda_2)y_2 y_3,$$

and hence

$$\theta'(t) = a(\lambda_3 - \lambda_2)\cos\theta \sin\theta.$$

Thus the two-dimensional dynamics collapse to a single scalar ODE for $\theta(t)$. Because $\lambda_3 < \lambda_2$, the right-hand side has a fixed sign away from $\theta \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$, which forces $\theta(t)$to move monotonically toward the boundary points. Consequently,

$$(y_2(t), y_3(t)) \longrightarrow \{(\pm R, 0),\ (0, \pm R)\},$$

showing that one of the two coordinates must be suppressed in the long-time limit.

This illustrates the key *suppression mechanism*: even in higher dimensions, the dynamics collapse to a low-dimensional angular system in which certain coordinates are monotonically driven to zero, leaving only one dominant direction.

Remark D.6. A major analytical difficulty arises when the transverse radius $R_\ell(t)$ becomes small. In this regime the angular variables develop a $1/R_\ell$ singularity, making it impossible to close the argument using the raw coordinates alone. The quantity $\kappa_\ell(t)$ is introduced precisely to measure this *competition*: it compares the strength of the inherited error $\varepsilon_{L-1}(t)$ to the geometric scale $R_\ell(t)$ on which the angular flow evolves. The condition $\kappa_\ell(t) \to 0$ ensures that the geometric suppression mechanism remains effective even when the transverse radius becomes small, allowing the reduced dynamics to overpower any residual perturbation.

Our theorem shows if a coordinate y_ℓ continues to occupy a nontrivial portion of the transverse space—meaning $R_\ell(t)$ does not decay faster than the upstream error $\varepsilon_{L-1}(t)$—then the dynamics force this coordinate to eventually suppress all deeper coordinates $y_{\ell+1}, \ldots, y_d$. Conversely, if y_ℓ decays to zero, the mechanism automatically shifts to the next level, where the same dichotomy applies. Thus no direction can maintain a marginal presence forever: any coordinate that remains consistently filled (up to an $o(1)$ perturbation from earlier tokens) must ultimately dominate and eliminate all others. To the best of our knowledge, this hierarchical concentration–versus–suppression phenomenon has not been previously justified rigorously in the analysis of token dynamics or related high-dimensional flows.

Intuition behind the mathematical techniques. In the simplified example above, the key idea is to select a vector that is perpendicular to (y_2, y_3) so that, when we take a dot product with the evolution equation, the term $-(\lambda_1 b\, y_1) y_i$ in the equation (28) cancels out. In two dimensions this perpendicular direction is simply $(-y_3, y_2)$. In three dimensions, the coordinates (y_2, y_3, y_4) are expressed in spherical variables, and the appropriate perpendicular directions are given by the standard orthonormal basis associated with those coordinates. More precisely, we write the vector (y_2, y_3, y_4) in spherical coordinates as

$$(y_2, y_3, y_4)^\top = R_2(t)\, \mathcal{M},$$

where

$$\mathcal{M} = \begin{pmatrix} \cos\varphi_1 \\ \sin\varphi_1\ \cos\varphi_2 \\ \sin\varphi_1\ \sin\varphi_2 \end{pmatrix}, \qquad R_2(t) = \sqrt{y_2^2 + y_3^2 + y_4^2}.$$

A natural perpendicular direction is then

$$\partial_{\varphi_1} \mathcal{M} = \begin{pmatrix} -\sin\varphi_1 \\ \cos\varphi_1\ \cos\varphi_2 \\ \cos\varphi_1\ \sin\varphi_2 \end{pmatrix},$$

and indeed $\mathcal{M} \cdot \partial_{\varphi_1} \mathcal{M} = 0$. This same principle extends to general dimension: we choose a direction orthogonal to $(y_2, \ldots, y_d)$ so that the undesirable linear term is eliminated, allowing us to isolate the key dynamics in the remaining coordinates.

Proof of Theorem C.6. To simplify the notation, and without loss of generality, we may assume $\ell = 2$. Our goal is to show that if $\kappa_2(t) \to 0$ as $t \to \infty$, then the dynamics reduce to the two-dimensional subsystem spanned by (y_1, y_k). Since all nontrivial behavior occurs only when $d \geq 3$, we henceforth

also assume $d \geq 3$. We introduce standard angular coordinates $\varphi_1, \ldots, \varphi_{d-2}$ so that

$$\begin{aligned} y_2 &= R_2(t)\, \cos\varphi_1, \\ y_3 &= R_2(t)\, \sin\varphi_1 \cos\varphi_2, \\ y_4 &= R_2(t)\, \sin\varphi_1 \sin\varphi_2 \cos\varphi_3, \\ &\vdots \\ y_{d-1} &= R_2(t)\, \sin\varphi_1 \cdots \sin\varphi_{d-3} \cos\varphi_{d-2}, \\ y_d &= R_2(t)\, \sin\varphi_1 \cdots \sin\varphi_{d-3} \sin\varphi_{d-2}. \end{aligned}$$

We write $\vec{y}_{\geq 2}(t) = \mathcal{M}(R_2(t), \varphi)$, where $\varphi = (\varphi_1, \ldots, \varphi_{d-2}) \in [0, \pi]^{\,d-3} \times [0, 2\pi)$. It is classical that the spherical coordinate map satisfies

$$\partial_{\varphi_i}\mathcal{M} \;\cdot\; \partial_{\varphi_j}\mathcal{M} = 0 \quad (i \neq j), \qquad \mathcal{M} \cdot \partial_{\varphi_i}\mathcal{M} = 0.$$

By the chain rule, the time derivative of $\mathcal{M}$ is

$$\partial_t \mathcal{M} = R_2'(t)\, \frac{\mathcal{M}}{R_2} \;+\; A\, \partial_t \varphi, \tag{29}$$

where the matrix $A \in \mathbb{R}^{(d-1)\times(d-2)}$ has columns

$$A = \big(\partial_{\varphi_1}\mathcal{M},\; \partial_{\varphi_2}\mathcal{M},\; \ldots,\; \partial_{\varphi_{d-2}}\mathcal{M}\big).$$

Recalling (25), we have

$$\partial_t \mathcal{M} = a(y)\, \mathrm{diag}\{\lambda_2, \lambda_3, \ldots, \lambda_d\}\, \mathcal{M} - a(y)\, Vy \cdot y\, \mathcal{M} - \lambda_1 b(y)\, y_1\, \mathcal{M} + \vec{g}_{\geq 2}(t),$$

where $\vec{g}_{\geq 2}(t) = (g_2(t), \ldots, g_d(t))$. Since $\mathcal{M} \cdot \partial_{\varphi_1}\mathcal{M} = 0$, we obtain

$$\partial_t \mathcal{M} \cdot \partial_{\varphi_1}\mathcal{M} = a(y)\, \big[\mathrm{diag}\{\lambda_2, \lambda_3, \ldots, \lambda_d\}\big]\mathcal{M} \cdot \partial_{\varphi_1}\mathcal{M} + \vec{g}_{\geq 2}(t) \cdot \partial_{\varphi_1}\mathcal{M}. \tag{30}$$

Combining (29) and (30), we get

$$A\, \partial_t \varphi \;\cdot\; \partial_{\varphi_1}\mathcal{M} = a(y)\, \big[\mathrm{diag}\{\lambda_2, \lambda_3, \ldots, \lambda_d\}\big]\mathcal{M} \cdot \partial_{\varphi_1}\mathcal{M} + \vec{g}_{\geq 2}(t) \cdot \partial_{\varphi_1}\mathcal{M}.$$

Hence,

$$A\, \partial_t \varphi \;\cdot\; \partial_{\varphi_1}\mathcal{M} = a(y)\, \big[\mathrm{diag}\{\lambda_2, \lambda_3, \ldots, \lambda_d\}\big]\mathcal{M} \cdot \partial_{\varphi_1}\mathcal{M} + \vec{g}_{\geq 2}(t) \cdot \partial_{\varphi_1}\mathcal{M}.$$

Here we used

$$A\, \partial_t \varphi \;\cdot\; \partial_{\varphi_1}\mathcal{M} = \partial_t \varphi \cdot A^{\top} \partial_{\varphi_1}\mathcal{M} = R_2^2\, \partial_t \varphi_1.$$

To compute the right-hand side, consider

$$\sum_{i=2}^{d} \lambda_i\, y_i\, \partial_{\varphi_1} y_i.$$

A direct calculation gives

$$\sum_{i=2}^{d} \lambda_i\, y_i\, \partial_{\varphi_1} y_i = R_2(t)^2\, \sin\varphi_1\, \cos\varphi_1 \Big(-\lambda_2 + \sum_{i=3}^{d} \lambda_i\, \alpha_i^2 \Big), \tag{31}$$

where the coefficients α_i are given by the standard angular factors where

$$\begin{cases} \alpha_3 = \cos\varphi_2, \\ \alpha_4 = \sin\varphi_2\, \cos\varphi_3, \\ \alpha_5 = \sin\varphi_2\, \sin\varphi_3\, \cos\varphi_4, \\ \vdots \\ \alpha_{d-1} = \Big(\prod_{k=2}^{d-3} \sin\varphi_k\Big) \cos\varphi_{d-2}, \\ \alpha_d = \prod_{k=2}^{d-2} \sin\varphi_k. \end{cases}$$

Since the coefficients $\alpha_3^2, \dots, \alpha_d^2$ are nonnegative and satisfy $\sum_{i=3}^d \alpha_i^2 = 1$, the bracket in (31) is estimated as

$$-\lambda_2 + \sum_{i=3}^{d} \lambda_i \alpha_i^2 \le -(\lambda_2 - \lambda_3) < 0$$

whenever $\lambda_1 > \lambda_2 > \cdots > \lambda_d$. This implies

$$\sum_{i=2}^{d} \lambda_i \, y_i \, \partial_{\varphi_1} y_i = -c(t) R_2^2 \cos\varphi_1 \sin\varphi_1.$$

where

$$c(t) \ge \lambda_2 - \lambda_3 > 0 \tag{32}$$

This implies

$$\partial_t \varphi_1 = -a(y)\, c(t)\, \sin\varphi_1 \, \cos\varphi_1 + \frac{\vec{g}_{\ge 2}(t) \, \cdot \, \partial_{\varphi_1} \mathcal{M}}{R_2(t)^2}, \tag{33}$$

where c is given in (32). By Proposition D.3 $|\vec{g}_{\ge 2}(t)| \lesssim \varepsilon_{L-1}(t)$ and the fact that $|\partial_{\varphi_1}\mathcal{M}| \lesssim R_2(t)$, we obtain

$$\frac{\vec{g}_{\ge 2}(t) \, \cdot \, \partial_{\varphi_1} \mathcal{M}}{R_2(t)^2} \;\lesssim\; \kappa_2(t).$$

Since $\kappa_2(t) \to 0$ as $t \to \infty$, the forcing term in the equation (33) converges to 0 as $t \to \infty$. We now show that one of the following two alternatives holds:

- $(\cos\varphi_1(t), \sin\varphi_1(t)) \to (0, \pm 1)$ as $t \to \infty$,
- $(\cos\varphi_1(t), \sin\varphi_1(t)) \to (\pm 1, 0)$ as $t \to \infty$.

Claim 1. If there exists a sequence $t_n \to \infty$ such that

$$(\cos\varphi_1(t_n), \sin\varphi_1(t_n)) \to (\pm 1, 0),$$

then in fact

$$(\cos\varphi_1(t), \sin\varphi_1(t)) \to (\pm 1, 0) \quad \text{as } t \to \infty.$$

Proof of Claim 1. Since $\varphi_1(t_n) \in [0, \pi]$, we may assume $\varphi_1(t_n) \to \bar{\theta} \in [0, \pi]$. Lemma 45 gives $\sin\bar{\theta}\, \cos\bar{\theta} = 0$, hence $\bar{\theta} \in \{0, \frac{\pi}{2}, \pi\}$. Because $\sin\varphi_1(t_n) \to 0$, we further obtain $\bar{\theta} \in \{0, \pi\}$.

Choose $\delta_0 > 0$ sufficiently small so that

$$\cos(\bar{\theta} + \delta)\, \sin(\bar{\theta} + \delta) \ge c_{\delta_0} > 0, \qquad 0 < \delta < \delta_0.$$

We show that $\varphi_1(t) \to \bar{\theta}$. Fix $\delta \in (0, \delta_0)$ and define

$$z(t) = \bar{\theta} + \delta - \varphi_1(t).$$

Since $\varphi_1(t_n) \to \bar{\theta}$, we have $z(t_n) > 0$ for all large n. As $\frac{\vec{g}_{\ge 2}(t) \cdot \partial_{\varphi_1}\mathcal{M}}{R_2^2} \to 0$, we further assume

$$\Big| \frac{\vec{g}_{\ge 2}(t) \cdot \partial_{\varphi_1} \mathcal{M}}{R_2^2} \Big| \le 0.1 \inf_{y \in \mathbb{S}^d} a(y)(\lambda_2 - \lambda_3) c_{\delta_0}, \quad t \ge t_{n_\delta}.$$

Then for every

$$T \ge t_{n_\delta} \quad \text{with} \quad z(T) = 0 \quad (\Leftrightarrow \; \varphi_1(T) = \bar{\theta} + \delta),$$

we have $z'(T) = -\varphi_1'(T) > 0$. Lemma G.1 implies $\varphi_1(t) < \bar{\theta} + \delta$ for all $t \ge t_{n_\delta}$. A symmetric argument yields $\varphi_1(t) > \bar{\theta} - \delta$ for all large t. Thus $\varphi_1(t) \to \bar{\theta}$. This proves the claim.

Claim 2. Exactly one of the following holds:

- $(\cos\varphi_1(t), \sin\varphi_1(t)) \to (0, \pm 1)$,
- $(\cos\varphi_1(t), \sin\varphi_1(t)) \to (\pm 1, 0)$.

Proof of Claim 2. Let

$$\zeta_0 = \liminf_{t\to\infty} |\cos\varphi_1(t)|, \qquad \zeta_1 = \limsup_{t\to\infty} |\cos\varphi_1(t)|.$$

If $\zeta_0 = \zeta_1$, then $\lim_{t\to\infty} |\cos\varphi_1(t)|$ exists. Using Lemma G.2 applied to (33) and the intermediate value theorem, the limit must lie in $\{-1, 0, 1\}$.

Now assume $0 \le \zeta_0 < \zeta_1 \le 1$. Choose $t_n \to \infty$ such that

$$|\cos\varphi_1(t_n)| \to \zeta_0, \qquad \varphi_1(t_n) \to \bar\theta \in [0, \pi].$$

Lemma G.2 again implies $\bar\theta \in \{0, \frac{\pi}{2}, \pi\}$, and $\zeta_0 = |\cos\bar\theta|$.

If $\bar\theta \in \{0, \pi\}$, Claim 1 yields $\varphi_1(t) \to \bar\theta$, so $\zeta_0 = \lim |\cos\varphi_1(t)| = 1$, contradicting $\zeta_0 < \zeta_1$. Thus $\bar\theta = \frac{\pi}{2}$, giving $\zeta_0 = 0$.

Next choose $t_n \to \infty$ such that $|\cos\varphi_1(t_n)| \to \zeta_1 > 0$. Lemma G.2 gives $\varphi_1(t_n) \to \bar\theta \in \{0, \frac{\pi}{2}, \pi\}$ with $\zeta_1 = |\cos\bar\theta|$.

If $\bar\theta \in \{0, \pi\}$, Claim 1 implies $\varphi_1(t) \to \bar\theta$, hence $\zeta_1 = \lim |\cos\varphi_1(t)| = 1 = \zeta_0$, a contradiction. If $\bar\theta = \frac{\pi}{2}$, then $\zeta_1 = 0$, another contradiction. This proves the claim. From the arguments above, we have

$$y_2(t) = R_2(t)\,\cos\varphi_1(t), \qquad \cos\varphi_1(t) \longrightarrow \pm 1 \text{ or } 0 \quad (t \to \infty).$$

If $\cos\varphi_1(t) \to 0$, then $y_2(t) \to 0$, and we apply the same polar–coordinate analysis to the remaining block $(y_3, \dots, y_d)$, now using R_3 and the next angular variable.

If instead $\cos\varphi_1(t) \to \pm 1$, then $y_2(t)$ converges to a nonzero limit while all other transverse coordinates satisfy $y_j(t) \to 0$ for $j \ge 3$. In this case y_2 is the unique surviving transverse direction, and we set $k = 2$. The proof is complete. □

Corollary D.7 (Condition for exactly two surviving coordinates)**.** *Under the hierarchical suppression mechanism, the long-time dynamics satisfy*

$$y_j(t) \to 0 \qquad (j \ne 1, \ell)$$

for some unique $\ell \ge 2$ if and only if there exists an index $\ell \in \{2, \dots, d\}$ such that

$$\kappa_\ell(t) \to 0 \quad \textit{and} \quad y_\ell(t) \not\to 0.$$

In this case the algorithm terminates at level ℓ, and the only nonvanishing coordinates in the limit are y_1 and y_ℓ.

Definition D.8 (Dominant–Direction Selection Condition)**.** For a given level $\ell \in \{2, \dots, d\}$, we say that the *Dominant–Direction Selection Condition* holds if there exists $\ell \in [2, d]$ such that

$$\kappa_\ell(t) \to 0 \qquad \text{and} \qquad y_\ell(t) \not\to 0. \tag{34}$$

Equivalently, this means that

$$\varepsilon_{L-1}(t) \;\ll\; R_\ell(t),$$

while the ℓ-th coordinate itself retains a nonvanishing amplitude. When this condition holds, level ℓ is the first surviving transverse direction, and all deeper coordinates $y_{\ell+1}, \dots, y_d$ are suppressed by the hierarchical mechanism.

Proposition D.9. *If there exist $\delta > 0$ and $T_\delta > 0$ such that*

$$1 - y_1(t)^2 \ge \delta \qquad \textit{for all } t \ge T_\delta,$$

then the condition (34) *holds.*

Proof. Since $y(t) \in \mathbb{S}^{d-1}$, we have $R_2(t)^2 + y_1(t)^2 = 1$. Hence,

$$\kappa_2(t) = \frac{\varepsilon_{L-1}(t)}{\sqrt{1 - y_1(t)^2}}.$$

For all $t \ge T_\delta$, the assumption $1 - y_1(t)^2 \ge \delta$ implies

$$\kappa_2(t) \le \delta^{-1/2}\,\varepsilon_{L-1}(t) \to 0 \qquad \text{as } t \to \infty.$$

If $y_2(t) \not\to 0$, then we choose $\ell = 2$. In that case, Theorem C.6 implies that $y_j(t) \to 0$ for every $j \geq 3$. If instead $y_2(t) \to 0$, we proceed to κ_3. Since $y_2(t) \to 0$, by enlarging T_δ if necessary we have

$$R_3(t)^2 = 1 - y_1(t)^2 - y_2(t)^2 \geq \tfrac{1}{4}\delta \qquad \text{for all } t \geq T_\delta.$$

Therefore,

$$\kappa_3(t) \leq 2\,\delta^{-1/2}\,\varepsilon_{L-1}(t) \to 0.$$

Continuing this argument inductively, we obtain the desired conclusion. □

Theorem D.10. *Recall the definition of N_1, N_2 in Lemma D.2. Under the condition* (34)*, the equation* (24) *can be written as*

$$\partial_t y_1 = (1 - y_1^2)F(y) + U(t),$$

where $U(t) \to 0$ as $t \to \infty$, and

$$\begin{aligned} F(t,y) &= \frac{G(t,y)}{Z(t,y)}, \\ G(t,y) &= e^{\beta\, y\cdot\mathcal{T}(t)y}\,(\lambda_1(t) - \lambda_\ell(t))\,y_1 + \lambda_1(t)\Big[N_1 e^{\beta\, y\cdot\mathcal{T}(t)\vec{e}_1} - N_2 e^{-\beta\, y\cdot\mathcal{T}(t)\vec{e}_1}\Big], \\ Z(t,y) &= e^{\beta\, y\cdot\mathcal{T}(t)y} + N_1 e^{\beta\, y\cdot\mathcal{T}(t)\vec{e}_1} + N_2 e^{-\beta\, y\cdot\mathcal{T}(t)\vec{e}_1}. \end{aligned} \tag{35}$$

When the matrix $\mathcal{T} = Q^\top K$ has the form (12)*, we obtain*

$$\begin{aligned} F(t,y_1) &= \frac{G(t,y_1)}{Z(t,y_1)}, \\ G(t,y_1) &= e^{\beta\,(m_1(t)+(m_0(t)-m_1(t))y_1^2)}\,(\lambda_1(t) - \lambda_\ell(t))\,y_1 + \lambda_1(t)\Big[N_1 e^{\beta m_0(t) y_1} - N_2 e^{-\beta m_0(t) y_1}\Big], \\ Z(t,y_1) &= e^{\beta\,(m_1(t)+(m_0(t)-m_1(t))y_1^2)} + N_1 e^{\beta m_0(t) y_1} + N_2 e^{-\beta m_0(t) y_1}. \end{aligned} \tag{36}$$

Proof. From Corollary D.7, there exists $\ell \in [2, d]$ such that $y_j(t) \to 0$ for all $j \neq 1, \ell$. We recall (24). Observe first that

$$\begin{aligned} \lambda_1 - Vy\cdot y = \lambda_1 \sum_{j=1}^d y_j^2 - \sum_{j=1}^d \lambda_j y_j^2 &= \sum_{j=2}^d (\lambda_1 - \lambda_j) y_j^2 \\ &= (\lambda_1 - \lambda_\ell) y_\ell^2 + \sum_{j\neq\ell} (\lambda_1 - \lambda_j) y_j^2 \\ &= (\lambda_1 - \lambda_\ell)\Big[1 - y_1^2 - \sum_{j\neq\ell} y_j^2\Big] + \sum_{j\neq\ell} (\lambda_1 - \lambda_j) y_j^2 \\ &= (\lambda_1 - \lambda_\ell)(1 - y_1^2) + \sum_{j\neq\ell} (\lambda_\ell - \lambda_j) y_j^2. \end{aligned}$$

Hence the equation for y_1 becomes

$$\partial_t y_1 = (1 - y_1^2)\Big[a(\lambda_1 - \lambda_\ell) y_1 + \lambda_1 b\Big] + g_1(t) + a y_1 \sum_{j\neq\ell} (\lambda_\ell - \lambda_j) y_j^2.$$

By Lemma D.2, we have

$$a(\lambda_1 - \lambda_\ell) y_1 + \lambda_1 b = F(y).$$

Define

$$U(t) = g_1(t) + a y_1 \sum_{j\neq 1,\ell} (\lambda_\ell - \lambda_j) y_j^2,$$

and the desired representation follows. Finally, note that

$$y\cdot\mathcal{T}y = m_1 + (m_0 - m_1) y_1^2, \qquad y\cdot\mathcal{T}\vec{e}_1 = m_0 y_1,$$

which yields the stated formulas for G and Z.

□

D.3 Proof of Theorem C.7

As in the convention of this paper, we write $y = x_L$. By Corollary D.7, there exists some $\ell \in \{2, \dots, d\}$ such that

$$y_j(t) \to 0 \qquad \text{for all } j \neq 1, \ell.$$

Since $y(t)$ remains in the compact set $\mathbb{S}^{d-1}$, we may extract a subsequence $t_n \to \infty$ such that $y(t_n) \to y_\star$ for some limit point $y_\star$. We now show that every such limit point must be an equilibrium.

From the above property, the only possibly nonzero components of $y_\star$ are the first and the ℓ-th. Thus $y_\star$ must be of the form

$$(\alpha_1, 0, \dots, 0, \alpha_\ell, 0, \dots, 0), \qquad \alpha_1^2 + \alpha_\ell^2 = 1.$$

Moreover, since

$$y_\ell^2(t) = 1 - y_1(t)^2 - \sum_{j \neq 1,\ell} y_j(t)^2 \geq \delta - \sum_{j \neq 1,\ell} y_j(t)^2,$$

and $y_j(t) \to 0$, we may increase T_δ so that

$$y_\ell^2(t) \geq \tfrac{1}{4}\,\delta \qquad \forall t \geq T_\delta.$$

Hence $y_\ell(t)$ eventually lies entirely in one of the intervals $(-\infty, -\frac{1}{2}\sqrt{\delta})$ or $(\frac{1}{2}\sqrt{\delta}, \infty)$. Without loss of generality, assume $y_\ell(t) > 0$ for large t. Then we may write

$$y_\ell(t) = \sqrt{1 - y_1(t)^2 - \sum_{j \neq 1,\ell} y_j(t)^2} = \sqrt{1 - y_1(t)^2} + y_{\ell,R}(t),$$

where

$$y_{\ell,R}(t) = \big(1 - y_1(t)^2 - \sum_{j \neq 1,\ell} y_j(t)^2\big)^{1/2} - \sqrt{1 - y_1(t)^2}\,.$$

Clearly $y_{\ell,R}(t) \to 0$.

Using Theorem D.10 and the fact that $y_j(t) \to 0$ for all $j \neq 1, \ell$, the equation for y_1 may be written as

$$\partial_t y_1 = P\big(y_1, \sqrt{1 - y_1^2}\big) + \widetilde{U}(t),$$

where $\widetilde{U}(t) \to 0$ as $t \to \infty$, and P is a complex-analytic function in both variables. Using Lemma G.2 and the fact that $y_1(t_n) \to \alpha_1$, we obtain

$$P(\alpha_1, \alpha_\ell) = 0,$$

so $y_\star$ is an equilibrium point.

Since $\alpha_1^2 + \alpha_\ell^2 = 1$, we may write

$$\alpha_1 = \cos\theta, \qquad \alpha_\ell = \sin\theta.$$

Then $P(\cos\theta, \sin\theta) = 0$. Because P is complex analytic in both variables, the zeros of this equation are isolated. Hence the set of limit points $y_\star$ is isolated. Finally, Lemma G.6 implies that $y(t) \to y_\star$ as $t \to \infty$. This completes the proof.

E Instability within the Flow Mechanisms

The evolution of tokens under a time-dependent $V(t)$ can be intuitively visualized as the motion of a small ball on a hill whose summit itself moves with time, producing much richer mechanism compared to the static case. To illustrate the idea, let's look at the toy model equation $x(t) \in (-1, 1)$ solving the simple ODE

$$x'(t) = (1 - x^2)(x - r(t))$$

We see that if $r(t) = r_0 \in (-1, 1)$ is a constant, then no matter where $x(0)$ in $(-1, 1)$, we will have $x(t) \to \pm 1$, except at the equilibrium $x = r_0$. We see that $r(t)$ acts like an unstable equilibrium. The physical analogy is as follows:

- **Rapidly moving hilltop.** If the hilltop $x = r(t)$ moves quickly from left -1 to right 1, the ball (initial position $x(0)$) cannot roll off either side; so $x(t)$ remains trapped near an intermediate orientation, forming an *interior cluster*.
- **Fixed hill.** If the hill $r(t) = r_0$ is fixed, the ball eventually rolls down one side ± 1.

This "moving-hill" analogy thus captures the physical mechanism behind the emergence of metastable or oscillatory interior clusters on the sphere.

A toy model. To be more precise, we now assume simple scenarios first. First we assume $Q^\top K = I_d$ for simplicity. To make things even more simple, we assume $x_1(t) \equiv \vec{e}_1$. We have

$$x_2(t) = x(t)\vec{e}_1 + y(t)\vec{e}_2,$$

so that $(x(t), y(t))$ solves

$$\partial_t x_2 = a\,\mathbf{P}_{x_2}(Vx_2) + b\lambda_1 \mathbf{P}_{x_2}(\vec{e}_1),$$

As in the proof of Theorem C.9, the x–equation reduces to

$$x'(t) = \frac{(1-x^2)}{e^\beta + e^{\beta x}} \left\{ e^\beta(\lambda_1(t) - \lambda_2(t))x + \lambda_1(t)e^{\beta x} \right\}. \tag{37}$$

We specify

$$\lambda_1(t) = 1$$

for all $t \geq 0$. Let

$$g(t,x) = e^\beta(1-\lambda_2(t))x + e^{\beta x}, \quad h(x) = \frac{1}{e^\beta + e^{\beta x}}.$$

We are left to study

$$x'(t) = (1-x^2)h(x)g(t,x)$$

with $\lambda_2(t) \in (0,1)$. We note that $g(t,1) = e^\beta(1-\lambda_2) + e^\beta > 0$ and $g(t,-1) = e^{-\beta} - e^\beta(1-\lambda_2)$. Hence, as long as

$$1 - \lambda_2(t) > e^{-2\beta}$$

for all $t \geq 0$, we will have $g(t,-1) < 0$. This garantee that there exists $r(t) \in (-1,1)$ such that

$$g(t, r(t)) = 0 \quad \forall t \geq 0$$

At the same time, since

$$\partial_x g(t,x) = e^\beta(1-\lambda_2) + \beta e^{\beta x} \geq e^{-\beta} > 0,$$

the map $t \to r(t)$ is well-defined. We can also see that from $g(t,r) = 0$, we have $r(t) \in (-1,0)$. Moreover, we note that from $r(t)$, we can recover $\lambda_2(t)$ as follows

$$|r(t)|e^{\beta|r(t)|} = \frac{e^{-\beta}}{1-\lambda_2(t)}$$

If $\lambda_2(t) \to \lambda_{2,\infty} \in (0,1)$, we see that

$$r(t) \to r_\infty \in (-1,0)$$

where r_∞ solves

$$|r_\infty|\, e^{\beta|r_\infty|} = \frac{e^{-\beta}}{1-\lambda_{2,\infty}}.$$

This observation motivates our main Theorem E.1, which is formulated precisely to handle models of the form (37) that arise naturally from token dynamics.

Theorem E.1. *Let $x : [0,\infty) \to (-1,1)$ be a solution of the differential equation*

$$x'(t) = (1-x(t)^2)\, h(x(t))\, g(t,x(t)) + w(t,x(t)).$$

Assume the following:

- *The perturbation $w(t,x)$ satisfies*

$$\sup_{x\in[a,b]} |w(t,x)| \longrightarrow 0 \quad \text{as } t\to\infty, \tag{38}$$

for any $a,b\in(-1,1)$, and

$$|w(t,x_1)-w(t,x_2)| \le W(t)\,|x_1-x_2|,$$

where $W(t)\to 0$ as $t\to\infty$.

- *There exists a constant $c_0>0$ such that*

$$h(x)\ge c_0, \qquad \inf_{t\ge 0}\partial_x g(t,x)\ge c_0,$$

for all $x\in\mathbb{R}$.

- *The function $z\mapsto h(z)^{-1}$ is entire.*

- *For each $t\ge 0$, there exists a root $r(t)\in(-1,1)$ of*

$$g(t,r(t))=0,$$

satisfying

$$r(t)\in[\alpha,\beta] \quad \text{for all } t\ge 0, \qquad \lim_{t\to\infty} r'(t)=0,$$

for some $\alpha,\beta\in(-1,1)$.

Then there exists at most one *trajectory $x_\star(t)\in(-1,1)$ that does not have ±1 as a limit point as $t\to\infty$. Moreover, if the condition* (38) *is replaced by*

$$\sup_{x\in[-1,1]} |w(t,x)| \to 0 \quad \text{as} \quad t\to\infty \tag{39}$$

then for any initial condition $x(0)\neq x_\star(0)$, the corresponding solution satisfies

$$\lim_{t\to\infty} x(t)\in\{-1,1\}.$$

Remark E.2. Theorem E.1 provides a general dynamical principle for one-dimensional token flows of the form (37). Such equations arise naturally after projecting the high-dimensional token dynamics *onto the dominant direction*, where the remaining degrees of freedom contribute through a slowly varying "effective root" $r(t)$. Informally, Theorem E.1 states that once all the remaining (non-dominant) directions of the token dynamics are controlled inside a bounded interval with only mild fluctuations, the evolution along the *dominant direction* becomes essentially one-dimensional and highly rigid. Under these mild conditions, the reduced flow collapses almost completely into one of the two extremal states $\{-1,1\}$: there is at most *one* exceptional trajectory that does not converge to an endpoint, while every other initialization converges decisively to either $+1$ or -1.

Proof. If $x(t)$ does not admit ±1 as a limit point as $t\to\infty$, then there exists $a,b\in(-1,1)$ and $T>0$ such that $x(t)\in[a,b]$ for all $t\ge T$.
We will show that there cannot be two distinct flows x_1,x_2 satisfying the above property. We define $u(t)=\Phi(x(t))$, where

$$\Phi(x)=\int_0^x \frac{1}{(1-w^2)h(w)}\,dw.$$

The function $\Phi:(-1,1)\to\mathbb{R}$ has the following properties:

- $\Phi'(x)=\frac{1}{(1-x^2)h(x)}$ for all $x\in(-1,1)$.

- Φ is an increasing bijection from $(-1,1)$ to $\mathbb{R}$, with $\Phi(x)\to-\infty$ as $x\to(-1)^+$ and $\Phi(x)\to\infty$ as $x\to 1^-$, and $\Phi'(x)>0$.

Now for $u = \Phi(x)$, we can write

$$x = f(u), \quad f = \Phi^{-1}, \quad f : \mathbb{R} \to (-1, 1).$$

The equation for u reads

$$u'(t) = g(t, f(u)) + \frac{w(t, f(u))}{(1 - x^2)h(x)}.$$

Now assume that there are two distinct flows $x_1(t), x_2(t)$ such that

$$x_1(t), x_2(t) \in [a, b] \quad \forall t \geq T.$$

Define $u_1 = \Phi(x_1)$, $u_2 = \Phi(x_2)$, and

$$H(x) = \frac{1}{(1 - x^2)h(x)}.$$

We see that there exists $C_{a,b} > 0$ such that

$$0 < H(x) \leq C_{a,b} \quad \forall x \in [a, b].$$

We have

$$\begin{aligned}
\frac{d}{dt}[u_1(t) - u_2(t)] &= g(t, f(u_1)) - g(t, f(u_2)) + w(t, x_1)\{H(x_1) - H(x_2)\} + H(x_2)(w(t, x_1) - w(t, x_2)) \\
&= \partial_x g(t, f(\xi_u))[f(u_1) - f(u_2)] + w(t, x_1)\{H(x_1) - H(x_2)\} + H(x_2)(w(t, x_1) - w(t, x_2)) \\
&= \partial_x g(t, f(\xi_u)) f'(\bar{\xi}_u)(u_1 - u_2) + w(t, x_1)\{H(x_1) - H(x_2)\} + H(x_2)(w(t, x_1) - w(t, x_2)),
\end{aligned}$$

for some $\xi_u, \bar{\xi}_u \in (u_1(t), u_2(t))$. Now since $f = \Phi^{-1}$, we have

$$f'(u) = \frac{1}{\Phi'(f(u))} = [1 - f(u)^2]h(f(u)).$$

Hence we get

$$\begin{aligned}
\frac{d}{dt}[u_1(t) - u_2(t)] = \partial_x g(t, f(\xi_u))[1 - f(\bar{\xi}_u)^2]h(f(\bar{\xi}_u))[u_1(t) - u_2(t)] \\
+ w(t, x_1)\{H(x_1) - H(x_2)\} + H(x_2)(w(t, x_1) - w(t, x_2)).
\end{aligned}$$

Letting $\xi_x(t) = f(\xi_u)$ and $\bar{\xi}_x = f(\bar{\xi}_u)$, we have $\xi_x, \bar{\xi}_x \in (x_1(t), x_2(t))$, and

$$\begin{aligned}
\frac{d}{dt}[u_1(t) - u_2(t)] = \partial_x g(t, \xi_x)[1 - \bar{\xi}_x^2]h(\bar{\xi}_x)[u_1(t) - u_2(t)] \\
+ w(t, x_1)\{H(x_1) - H(x_2)\} + H(x_2)(w(t, x_1) - w(t, x_2)).
\end{aligned}$$

Since $x_1(t), x_2(t) \in [a, b]$ for $t \geq T$, there exists $c_0 > 0$ such that

$$\begin{cases} 1 - \bar{\xi}_x(t)^2 \geq c_0, \\ |H(x_1) - H(x_2)| \leq c_0 |u_1 - u_2| \end{cases}$$

We define

$$z(t) = (u_1(t) - u_2(t))^2.$$

Then, for T large and $t \geq T$, we get

$$z'(t) \geq \left(c_0 - |w(t)| - \sup_{x_2 \in [a,b]} |H(x_2)| W(t) \right) z(t) \geq 0.1 c_0 z(t),$$

for some constant c_0. Since u_1, u_2 are bounded for $t \geq T$, this is a contradiction. This implies that there is at most one solution $x_\star$ that does not have ± 1 as a limit point. This completes the proof. Now we show that under assumption (39), if ± 1 is a limit point of $\{x(t)\}_{t>0}$, then $x(t) \to \pm 1$ as $t \to \infty$. It suffices to consider the case where $+1$ is a limit point, as the case -1 is treated analogously.

Assume therefore that there exists a sequence $t_n \to \infty$ such that $x(t_n) \to 1$. We claim that $x(t) \to 1$ as $t \to \infty$. By compactness, passing to a subsequence if necessary, we may assume

$$r(t_n) \to r_\infty \in [\alpha, \beta] \subset (-1, 1).$$

Fix $\varepsilon \in (0, 1-\beta)$ and define

$$y(t) := x(t) - r(t) - \varepsilon.$$

Since $x(t_n) \to 1$ and $r(t_n) \to r_\infty$, we have

$$\lim_{n\to\infty} y(t_n) = 1 - r_\infty - \varepsilon \geq 1 - \beta - \varepsilon > 0.$$

Hence there exists n_ε such that

$$y(t_n) > 0 \quad \text{for all } n \geq n_\varepsilon.$$

Using (39) and $r'(t) \to 0$ as $t \to \infty$, we may increase n_ε so that

$$\sup_{x\in[-1,1]} |w(t,x)| + |r'(t)| \leq 0.8\, c_0^2\, \varepsilon \inf_{r\in[\alpha,\beta]} \big(1 - (r+\varepsilon)^2\big), \quad \forall\, t \geq t_{n_\varepsilon}. \tag{40}$$

We now show that

$$y'(T) > 0 \quad \text{for all } T \geq t_{n_\varepsilon} \text{ such that } y(T) = 0.$$

Indeed, if $y(T) = 0$, then $x(T) = r(T) + \varepsilon$, and thus

$$\begin{aligned} y'(T) &= (1-x^2)h(x)g(T,x) + w(T,x(T)) - r'(T) \\ &= \left[1 - (r(T)+\varepsilon)^2\right] h(x(T))\, g(T, r(T)+\varepsilon) + w(T,x(T)) - r'(T) \\ &\geq c_0 \left[1 - (r(T)+\varepsilon)^2\right] g(T, r(T)+\varepsilon) - |w(T,x(T))| - |r'(T)|. \end{aligned}$$

Using $g(T, r(T)) = 0$ and the mean value theorem, we obtain

$$g(T, r(T)+\varepsilon) = \partial_x g(T,\xi)\, \varepsilon \geq c_0 \varepsilon,$$

for some $\xi \in (r(T), r(T)+\varepsilon)$. Therefore,

$$\begin{aligned} y'(T) &\geq c_0^2\, \varepsilon \left[1 - (r(T)+\varepsilon)^2\right] - \sup_{x\in[-1,1]} |w(T,x)| - |r'(T)| \\ &\geq c_0^2\, \varepsilon \inf_{r\in[\alpha,\beta]} \big(1 - (r+\varepsilon)^2\big) - \sup_{x\in[-1,1]} |w(T)| - |r'(T)|. \end{aligned}$$

By (40), we conclude that $y'(T) > 0$.

By Lemma G.1, it follows that

$$y(t) > 0 \quad \text{for all } t \geq t_{n_\varepsilon},$$

which implies

$$x(t) - r(t) > \varepsilon \quad \text{for all } t \geq t_{n_\varepsilon}. \tag{41}$$

Next, we write

$$\begin{aligned} x'(t) &= (1-x^2)h(x)g(t,x) + w(t) \\ &= (1-x^2)h(x)\big[g(t,r(t)) + \partial_x g(t,\xi)(x - r(t))\big] + w(t) \\ &= (1-x^2)h(x)\, \partial_x g(t,\xi)\, (x - r(t)) + w(t), \end{aligned}$$

for some $\xi(t) \in (r(t), x(t))$. Combining this with (41), we obtain for all $t \geq t_{n_\varepsilon}$,

$$x'(t) \geq c_0^2\, \varepsilon\, (1-x^2) - \sup_{x\in[-1,1]} |w(t,x)|.$$

Consequently,

$$\begin{aligned} \frac{d}{dt}(1-x) &= -x'(t) \\ &\leq -c_0^2\, \varepsilon\, (1+x)(1-x) + \sup_{x\in[-1,1]} |w(t,x)| \\ &\leq -c_0^2\, \varepsilon\, (1 + r(t) + \varepsilon)(1-x) + \sup_{x\in[-1,1]} |w(t,x)| \\ &\leq -c_0^2\, \varepsilon\, (1 + \alpha + \varepsilon)(1-x) + \sup_{x\in[-1,1]} |w(t,x)|. \end{aligned}$$

By Lemma G.3, this yields $x(t) \to 1$ as $t \to \infty$.

□

In the next proposition, we show that our theorem is sharp and cannot be strengthened to assert that *all* flows converge to $\pm\vec{e}_1$. We provide an explicit example in the simplified setting $L = d = 2$ with the first token equal to $\vec{e}_1$, and exhibit a trajectory $x_\star$ whose limit is neither $+\vec{e}_1$ nor $-\vec{e}_1$.

Proposition E.3. *Take $\alpha \in (-1, 0)$. Then there exists an ODE of the form* (37) *with suitable β, with $\lambda_1(t) \equiv 1 > \lambda_2(t)$ and $\lambda_2(t) \to \lambda_{2,\infty} \in (0, 1)$ at an exponential rate, such that $x(0) = \alpha$ and*

$$\lim_{t\to\infty} x(t) = \alpha.$$

In particular, the corresponding trajectory $x(t)$ converges to the interior point $\alpha \notin \{-1, 1\}$.

Proof. We take $\lambda_1(t) = 1$. Fix a small $\varepsilon_0 \in (0, 1)$ and set

$$x(t) = \alpha + \varepsilon_0 e^{-t} \sin t.$$

Then $x(t) \to \alpha$ as $t \to \infty$, and for all $t \geq 0$,

$$x(t) \in [\alpha - \varepsilon_0, \alpha + \varepsilon_0] \subset (-1, 0).$$

Next we determine $\lambda_2(t)$. From the equation

$$x'(t) = h(x)\big[e^{\beta}(1 - \lambda_2(t))x + e^{\beta x}\big], \qquad h(x) = \frac{1 - x^2}{e^{\beta} + e^{\beta x}}, \tag{42}$$

we obtain

$$e^{\beta}(1 - \lambda_2)x + e^{\beta x} = \frac{e^{\beta} + e^{\beta x}}{1 - x^2}\, x'(t).$$

Solving for λ_2 gives

$$\lambda_2 = 1 + \frac{e^{\beta(x-1)}}{x} - \frac{(1 + e^{\beta(x-1)})\, x'}{x(1 - x^2)}.$$

Substituting the explicit form of $x(t)$ yields

$$\lambda_2(t) = 1 + \frac{e^{\beta(\alpha+\varepsilon_0 e^{-t}\sin t - 1)}}{\alpha + \varepsilon_0 e^{-t}\sin t} - \frac{\big(1 + e^{\beta(\alpha+\varepsilon_0 e^{-t}\sin t-1)}\big)\, \varepsilon_0 e^{-t}(\cos t - \sin t)}{(\alpha + \varepsilon_0 e^{-t}\sin t)\,\big(1 - (\alpha + \varepsilon_0 e^{-t}\sin t)^2\big)}.$$

As $t \to \infty$, we have $x(t) \to \alpha$ and $x'(t) \to 0$, hence

$$\lambda_2(t) \to \lambda_{2,\infty} = 1 + \frac{e^{\beta(\alpha-1)}}{\alpha} \in (0, 1), \tag{43}$$

provided that $\alpha \in (-1, 0)$ and

$$\beta > \frac{\ln(-\alpha)}{\alpha - 1}.$$

This completes the proof. □

Remark E.4. We have taken the first token to be $\vec{e}_1$, and here $\alpha = x_2(0) \cdot \vec{e}_1 < 0$. Thus the initial position of the second token lies on the opposite hemisphere from $\vec{e}_1$ (see Table 1).

Remark E.5. By Theorem E.1, the solution of (42) with $\lambda_2(t)$ given by (43) converges to ± 1 for every initial condition $x(0) \neq \alpha$. The function $\lambda_2(t)$ itself satisfies the asymptotic behavior

$$\lambda_2(t) = \lambda_{2,\infty} + O(e^{-t}\cos t) + O(e^{-t}\sin t),$$

so $\lambda_2(t)$ approaches its limit exponentially with a decaying oscillation.

F Time-dependent $Q(t), V(t), K(t)$

In this section, we give the proof of Theorem C.1. If $y(t_n) \to \pm\vec{e}_1$ for some sequence $t_n \to \infty$, then the theorem follows immediately. If this does not occur, then there exist $\delta_0 > 0$ and $T > 0$ such that

$$1 - y_1(t)^2 \geq \delta_0 \qquad \forall\, t \geq T.$$

By Theorem D.10 and Proposition D.9, the dynamics for $y_1 = y \cdot \vec{e}_1$ reduce to

$$\partial_t y_1 = (1 - y_1^2)\, F(t, y_1) + U(t, y_1),$$

where $U(t) \to 0$ as $t \to \infty$, and $F(t, y_1)$ is defined in (36):

$$\begin{cases} F(t,y_1) = \dfrac{G(t,y_1)}{Z(t,y_1)}, \\ G(t,y_1) = e^{\beta(m_1(t)+(m_0(t)-m_1(t))y_1^2)}(\lambda_1-\lambda_k)\, y_1 + \lambda_1\left[N_1 e^{\beta m_0(t) y_1} - N_2 e^{-\beta m_0(t) y_1}\right], \\ Z(t,y_1) = e^{\beta(m_1(t)+(m_0(t)-m_1(t))y_1^2)} + N_1 e^{\beta m_0(t) y_1} + N_2 e^{-\beta m_0(t) y_1} \\ U(t,y_1) = g_1(t) + a(t,y_1) y_1 \sum_{j\neq 1,\ell}(\lambda_\ell - \lambda_j) y_j^2 \end{cases}$$

By a direct calculation, we have

$$\partial_{y_1} G(t,y_1) = e^{\beta y\cdot \mathcal{T} y}(\lambda_1-\lambda_k)\Big[1 + 2\beta(m_0-m_1)y_1^2\Big] + \beta m_0\Big[\lambda_1 N_1 e^{\beta m_0 y_1} + \lambda_1 N_2 e^{-\beta m_0 y_1}\Big],$$

First, we will show that

$$G'(y_1) \geq \lambda_1 \beta m_0 L + (\lambda_1-\lambda_k)\min\Big\{1,\, 1-2\beta(m_1-m_0)\Big\}. \tag{44}$$

If $m_0 \geq m_1$, then clearly

$$G'(y_1) \geq \lambda_1\beta m_0 L + \lambda_1 - \lambda_k > 0.$$

If $m_0 < m_1$, then

$$1 + 2\beta(m_0-m_1)y_1^2 = 1 - 2\beta(m_1-m_0)y_1^2 \geq 1 - 2\beta(m_1-m_0),$$

completing the proof of (44). Fix any $t \geq 0$, we see that the map $y_1 \mapsto G(t, y_1)$ is strictly increasing on $\mathbb{R}$, thanks to (13) and the bound (44). Define the limiting function

$$G_\infty(y_1) = \lim_{t\to\infty} G(t, y_1).$$

Our strategy is to apply Theorem E.1. We split the analysis into two cases:

- $G_\infty(-1) > 0$ or $G_\infty(1) < 0$ (monotone case),
- $G_\infty(-1) < 0 < G_\infty(1)$ (sign-changing case).

Case 1: $G_\infty(-1) > 0$ or $G_\infty(1) < 0$. Assume first that $G_\infty(-1) > 0$ (the case $G_\infty(1) < 0$ is symmetric). For all sufficiently large t,

$$G(t, y_1) \geq G(t, -1) \geq 0.9\, G_\infty(-1) > 0.$$

Thus, for $t \geq T$ with T large enough,

$$y_1'(t) \geq (1-y_1(t)^2)\, 0.9\, G_\infty(-1) - |U(t)| \geq 0.9\,\delta_0\, G_\infty(-1) - |U(t)|.$$

Since $U(t) \to 0$, we eventually have $y_1'(t) > 0$. Hence $y_1(t)$ is strictly increasing for large t, and as a bounded monotone function on $(-1, 1)$ it must converge:

$$y_1(t) \to \alpha \in [-1, 1].$$

Taking the limit in the equation yields

$$(1-\alpha^2)\, G_\infty(\alpha) = 0,$$

so $\alpha \in \{-1, 1\}$. Thus $y_1(t) \to \pm 1$, completing Case 1.

Case 2: $G_\infty(-1) < 0 < G_\infty(1)$. In this regime, $G(t, -1) < 0 < G(t, 1)$ for all sufficiently large t. Fix $T > 0$ so that this holds for all $t \geq T$. By strict monotonicity of $G(t, \cdot)$, for each $t \geq T$ there exists a unique $r(t) \in (-1, 1)$ such that

$$G(t, r(t)) = 0.$$

We show that $r(t) \to r_\infty$ for some $r_\infty \in (-1, 1)$ with $G_\infty(r_\infty) = 0$, and that $r'(t) \to 0$.

Step 1: Convergence of $r(t)$. Let $t_n \to \infty$ be arbitrary. Since $r(t_n)$ is bounded, extract a subsequence (not relabeled) such that $r(t_n) \to \alpha$. Passing to the limit in $G(t_n, r(t_n)) = 0$ yields

$$G_\infty(\alpha) = 0.$$

By strict monotonicity of G_∞, the root is unique, hence $\alpha = r_\infty$. Therefore $r(t) \to r_\infty$.

Step 2: Decay of $r'(t)$. Differentiating $G(t, r(t)) = 0$ gives

$$\partial_t G(t, r(t)) + \partial_{y_1} G(t, r(t))\, r'(t) = 0,$$

so

$$r'(t) = -\frac{\partial_t G(t, r(t))}{\partial_{y_1} G(t, r(t))}.$$

From (44), $\partial_{y_1} G(t, r) \geq \lambda_1 \beta m_0 > 0$. Thus

$$|r'(t)| \lesssim |\partial_t G(t, r(t))| \lesssim |\lambda_1'(t)| + |\lambda_k'(t)| \to 0.$$

Step 3: Verification of the monotonicity condition in Theorem E.1. We must show that

$$\inf_{t \geq 0} \frac{\mathrm{d}}{\mathrm{d}y_1} \left[\frac{Z_\infty(y_1)}{Z(t, y_1)}\, G(t, y_1) \right] > 0.$$

Compute

$$\frac{\mathrm{d}}{\mathrm{d}y_1} \left(\frac{Z_\infty}{Z} G \right) = \frac{\mathrm{d}}{\mathrm{d}y_1} \left(\frac{Z_\infty}{Z} \right) G + \frac{Z_\infty}{Z}\, \partial_{y_1} G \geq c_0 \left(\partial_{y_1} G - G \left| \frac{\mathrm{d}}{\mathrm{d}y_1} \left(\frac{Z_\infty}{Z} \right) \right| \right).$$

Since $\partial_{y_1} G$ is uniformly bounded below, it suffices to show

$$\frac{\mathrm{d}}{\mathrm{d}y_1} \left(\frac{Z_\infty}{Z} \right) \to 0 \quad \text{uniformly for } y_1 \in [-1, 1].$$

Indeed,

$$\begin{aligned} \left| \partial_{y_1} \left(\frac{Z_\infty}{Z} \right) \right| &\lesssim |\partial_{y_1} Z_\infty|\, |Z - Z_\infty| + |Z_\infty|\, |\partial_{y_1} (Z - Z_\infty)| \\ &\lesssim |Z - Z_\infty| + |\partial_{y_1} (Z - Z_\infty)| \\ &\lesssim |m_0(t) - m_{0,\infty}| + |m_1(t) - m_{1,\infty}|. \end{aligned}$$

Since $m_0(t) \to m_{0,\infty}$ and $m_1(t) \to m_{1,\infty}$, this quantity tends to zero, completing the verification.

The proof is complete.

G Technical Lemmas

In this subsection we collect several technical lemmas that will be used throughout the paper. These results serve as analytic tools inspired by classical methods in mathematical analysis. Readers interested only in the high-level flow of the argument may safely skip this section on a first pass and return to it as needed.

Lemma G.1. *Let $g : [t_0, \infty) \to \mathbb{R}$ be a differentiable function with a continuous derivative g'. Suppose $g(t_0) > \delta$ for some $\delta \in \mathbb{R}$. If $g'(T) > 0$ for every T in the set*

$$\{T \geq t_0 : \quad g(T) = \delta\}$$

then

$$g(t) > \delta \qquad \text{for all } t \geq t_0.$$

Proof. Suppose, for contradiction, that there exists $t > t_0$ with $g(t) \leq \delta$. Since $g(t_0) > \delta$, there exists a time $T > t_0$ which is the smallest time such that $g(T) = \delta$. We then have $g'(T) > 0$. By continuity of g', there exists $\varepsilon_0 > 0$ such that

$$g'(t) > 0 \qquad \text{for all } t \in (T - \varepsilon_0, T + \varepsilon_0).$$

In particular, for $t \in (T - \varepsilon_0, T)$ we have $g(t) < g(T) = \delta$. By continuity of g, this implies there exists $T_1 \in (t_0, T)$ with $g(T_1) = \delta$, contradicting the minimality of T. Hence no such T can exist, and we conclude

$$g(t) > \delta \qquad \text{for all } t \geq t_0.$$

□

Lemma G.2 (Limit-point principle for scalar residual dynamics)**.** *Let $c : [0, \infty) \to \mathbb{R}_+$ be a smooth function satisfying the uniform bounds*

$$0 < c_\infty \ \le \ c(t) \ \le \ C_\infty < \infty \qquad \textit{for all } t \ge 0.$$

Consider the scalar evolution

$$x'(t) = c(t)\,\mathcal{A}(x(t)) + u(t), \tag{45}$$

where $u(t) \to 0$ as $t \to \infty$ and $\mathcal{A} : \mathbb{R} \to \mathbb{R}$ is continuous. Then any accumulation point of $x(t)$ must be a zero of the drift: if $x(t_n) \to \alpha$ along some sequence $t_n \to \infty$, we must have

$$\mathcal{A}(\alpha) = 0.$$

Proof. Assume that $\mathcal{A}(\alpha) \neq 0$, then we have $\mathcal{A}(\alpha) > 0$ or $\mathcal{A}(\alpha) < 0$.
Case 1: $\mathcal{A}(\alpha) > 0$
Since $\mathcal{A}$ is continuous, we can find $\delta > 0$ such that $\mathcal{A}(x) > 0$ for all $x \in [\alpha - 9\delta, \alpha + 9\delta]$. We will first show that there exists $T_\delta > 0$ such that

$$x(t) > \alpha - \delta \quad \text{for all} \quad t \ge T_\delta. \tag{46}$$

To this end, we consider

$$z(t) = x(t) - \alpha + \delta.$$

Since $x(t_n) \to \alpha$ as $n \to \infty$, we can find $n_0 > 0$ such that $z(t_n) > 0$ for all $n \ge n_0$. Moreover, since $u(t) \to 0$ as $t \to \infty$, we may further assume

$$|u(t)| \le 0.1 c_\infty \mathcal{A}(\alpha - \delta) \quad \text{for all} \quad t \ge t_{n_0}. \tag{47}$$

Now as soon as $z(T) = 0$ for some $T \ge t_{n_0}$, we have $x(T) = \alpha - \delta$, and thus

$$\begin{aligned} z'(T) = x'(T) &= c(T)\mathcal{A}(\alpha - \delta) + u(T) \\ &\ge c_\infty \mathcal{A}(\alpha - \delta) - 0.1 c_\infty \mathcal{A}(\alpha - \delta) > 0 \end{aligned}$$

thanks to (47). By Lemma G.1, we obtain (46). Next we show that for all $\delta > 0$ sufficiently small and for all $T > 0$ large, there exists $t = t(\delta, T) > 0$ such that $|x(t) - \alpha| < \delta$ and $|x'(t)| < \delta$. If this is true, then we can find a sequence $\tau_n \to \infty$ such that $x(\tau_n) \to \alpha$ and $x'(\tau_n) \to 0$. Passing $\tau_n \to \infty$ in (45), we have $\mathcal{A}(\alpha) = 0$. To this end, we consider two cases:
Subcase 1: There exists $T > T_\delta$ such that $x(t) \le \alpha + \delta$ for all $t \ge T$. In this case, we have $t \in [T, T + 2]$ such that $x(T + 2) - x(T) = 2x'(t)$. Hence

$$|x'(t)| = \frac{|x(T + 2) - x(T)|}{2} \le \frac{(\alpha + \delta) - (\alpha - \delta)}{2} = \delta.$$

Subcase 2: For all $T > T_\delta$, there exists $t > T$ such that $x(t) > \alpha + \delta$. In this case, we can find $s_1 > T_\delta$ such that

$$x(s_1) > \alpha + \delta.$$

Since $x(t_n) \to \alpha$ as $n \to \infty$, we can find $t_n > s_1$ such that

$$|x(t_n) - \alpha| < \delta.$$

Next we choose $s_2 > t_n$ such that $x(s_2) > \alpha + \delta$. We have $t_0 \in [s_1, s_2]$ such that

$$x(t_0) = \min_{s \in [s_1, s_2]} x(s).$$

Since $t_n \in (s_1, s_2)$ and $x(t_n) < \alpha + \delta < \min\{x(s_1), x(s_2)\}$, we see that $t_0 \in (s_1, s_2)$. This implies

$$x'(t_0) = 0,$$

which implies $|x'(t_0)| < \delta$. At the same time, we have $x(t_0) - \alpha \le x(t_n) - \alpha < \delta$ and $x(t_0) > \alpha - \delta$ since $t_0 > T$. The proof is complete.
Case 2: $\mathcal{A}(\alpha) < 0$.
Since $\mathcal{A}$ is continuous, we can find $\delta > 0$ such that $\mathcal{A}(x) < 0$ for all $x \in [\alpha - 9\delta, \alpha + 9\delta]$. By considering the function $z(t) = \alpha + \delta - x(t)$ and applying Lemma G.1, we have $T_\delta > 0$ such that $x(t) < \alpha + \delta$ for all $t \ge T_\delta$. Now if there exists $T > T_\delta$ such that $x(t) \ge \alpha - \delta$ for all $t \ge T$. In this case, we have $t \in [T, T + 2]$ such that $x(T + 2) - x(T) = 2x'(t)$ and thus $|x'(t)| \le \delta$. If for

all $T > T_\delta$, there exists $t > T$ such that $x(t) < \alpha - \delta$. In this case, we can find $s_1 > T_\delta$ such that $x(s_1) < \alpha - \delta$. Since $x(t_n) \to \alpha$ as $n \to \infty$, we can find $t_n > s_1$ such that

$$|x(t_n) - \alpha| < \delta.$$

Next we choose $s_2 > t_n$ such that $x(s_2) < \alpha - \delta$. We have $t_0 \in [s_1, s_2]$ such that

$$x(t_0) = \max_{s \in [s_1, s_2]} x(s).$$

Since $t_n \in (s_1, s_2)$ and $x(t_n) > \alpha - \delta > \max\{x(s_1), x(s_2)\}$, we see that $t_0 \in (s_1, s_2)$. This implies

$$x'(t_0) = 0,$$

which implies $|x'(t_0)| < \delta$. At the same time, we have $x(t_0) \geq x(t_n) > \alpha - \delta$ and $x(t_0) < \alpha + \delta$ since $t_0 > T$. The proof is complete. □

Lemma G.3. *Let $x : [0, \infty) \to [0, \infty)$ be a continuous function and $w : [0, \infty) \to \mathbb{R}$ such that $w(t) \to 0$ as $t \to \infty$. Assume that for any $\varepsilon > 0$ sufficiently small, there exist positive constants $A_\varepsilon, T_\varepsilon$ such that*

$$x'(t) \leq -A_\varepsilon x(t) + w(t) \quad \text{for all } t \geq T_\varepsilon.$$

Then

$$\lim_{t \to \infty} x(t) = 0.$$

Proof. Since $w(t) \to 0$ as $t \to \infty$, we can increase $T_\varepsilon > 0$ such that

$$|w(t)| \leq \varepsilon A_\varepsilon \quad \text{for all } t \geq T_\varepsilon.$$

This implies that for $t \geq T_\varepsilon$,

$$x'(t) \leq -A_\varepsilon x(t) + \varepsilon A_\varepsilon.$$

By Gronwall inequality, we have

$$x(t) \leq x(T_\varepsilon) e^{-A_\varepsilon (t - T_\varepsilon)} + \varepsilon \big(1 - e^{-A_\varepsilon (t - T_\varepsilon)}\big).$$

Hence, if

$$t \geq T_\varepsilon + \frac{1}{A_\varepsilon} \ln\left(\frac{1 + x(T_\varepsilon)}{\varepsilon}\right),$$

we get

$$x(t) \leq 2\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary small, we conclude $\lim_{t \to \infty} x(t) = 0$. □

Lemma G.4. *Let $x : [0, \infty) \to [0, \infty)$ be a continuous function and $w : [0, \infty) \to \mathbb{R}$ satisfy $w(t) \to 0$ as $t \to \infty$. Assume also that there exists $M > 0$ such that*

$$0 \leq x(t) \leq M \qquad \text{for all } t \geq 0.$$

Suppose that for every sufficiently small $\varepsilon > 0$, there exist positive constants A_ε and T_ε such that

$$x'(t) \geq A_\varepsilon x(t) + w(t) \qquad \text{for all } t \geq T_\varepsilon.$$

Then

$$\lim_{t \to \infty} x(t) = 0.$$

Proof. Fix any $\varepsilon > 0$. By assumption, there exists $T_\varepsilon > 0$ such that

$$x'(t) \geq A_\varepsilon x(t) + w(t) \qquad \text{for all } t \geq T_\varepsilon.$$

Since $w(t) \to 0$ as $t \to \infty$, we may increase T_ε if necessary so that

$$|w(t)| \leq \varepsilon A_\varepsilon \qquad \text{for all } t \geq T_\varepsilon.$$

Thus, for all $t \geq T_\varepsilon$,

$$x'(t) \geq A_\varepsilon x(t) - \varepsilon A_\varepsilon.$$

Applying Grönwall's inequality, for any $t_1 > t_2 \geq T_\varepsilon$ we obtain

$$x(t_1) \geq \big[x(t_2) - \varepsilon\big] e^{A_\varepsilon (t_1 - t_2)} + \varepsilon.$$

Fix $t_2 \geq T_\varepsilon$ and let $t_1 \to \infty$. Since $x(t)$ is uniformly bounded above on $[0, \infty)$, the right-hand side remains bounded only if

$$x(t_2) - \varepsilon \leq 0.$$

As this holds for every $t_2 \geq T_\varepsilon$, we conclude that $x(t) \to 0$ as $t \to \infty$. □

Lemma G.5. *Consider the triangular dynamical system*

$$\begin{cases} \partial_t x_1 = F_1(x_1), \\ \partial_t x_2 = F_2(x_1, x_2), \\ \quad \vdots \\ \partial_t x_L = F_L(x_1, x_2, \ldots, x_L), \end{cases}$$

where each $F_i : (\mathbb{R}^d)^i \to \mathbb{R}^d$ is continuously differentiable. Let $x^\star = (x_1^\star, \ldots, x_L^\star) \in (\mathbb{S}^{d-1})^L$ be a steady state of this system. The Jacobian matrix of the linearized system at $x^\star$ is block upper triangular. Its eigenvalues are determined solely by the diagonal blocks $D_1F_1(x^\star), D_2F_2(x^\star), \ldots, D_LF_L(x^\star)$ and are given by the union of their spectra.

Proof. The Jacobian at the steady state $x^\star$ is given by

$$J(x^\star) = \begin{bmatrix} D_1F_1(x^\star) & 0 & \cdots & 0 \\ D_1F_2(x^\star) & D_2F_2(x^\star) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ D_1F_L(x^\star) & D_2F_L(x^\star) & \cdots & D_LF_L(x^\star) \end{bmatrix}.$$

This matrix is block upper triangular. A standard result in linear algebra confirms that the eigenvalues of $J(x^\star)$ are simply the eigenvalues of its diagonal blocks $D_iF_i(x^\star)$ for $1 \le i \le L$. □

Lemma G.6. *Let $\gamma : [0, \infty) \to \mathbb{R}^d$ be a continuous map, and define the ω-limit set*

$$K \;=\; \{\, x \in \mathbb{R}^d \;:\; \exists\, t_n \to \infty \text{ with } \gamma(t_n) \to x \,\}.$$

If every point of K is isolated in K, then K consists of a single element.

Proof. Fix $x_0 \in K$. We claim that $\gamma(t) \to x_0$ as $t \to \infty$. Suppose to the contrary that this convergence fails. Then there exist a sequence $s_n \to \infty$ and a number $\delta > 0$ such that

$$|\gamma(s_n) - x_0| \ge \delta \qquad \text{for all } n. \tag{48}$$

Since $x_0 \in K$, for each n we may choose $t_n > s_n$ such that

$$|\gamma(t_n) - x_0| < \frac{\delta}{2^n}. \tag{49}$$

Fix $n_0 \in \mathbb{N}$. By continuity of γ, for every $n > n_0$ there exists a point $w_n \in (s_n, t_n)$ satisfying

$$|\gamma(w_n) - x_0| = \frac{\delta}{2^{\,n_0}}.$$

Since the sphere $\partial B\big(x_0, \frac{\delta}{2^{\,n_0}}\big)$ is compact, we may pass to a subsequence (not relabeled) such that $\gamma(w_n) \to \alpha$ for some

$$\alpha \in \partial B\left(x_0, \frac{\delta}{2^{\,n_0}}\right).$$

Hence

$$K \cap \partial B\left(x_0, \frac{\delta}{2^{\,n_0}}\right) \neq \varnothing.$$

Since n_0 was arbitrary, it follows that K contains points on the spheres $\partial B(x_0, \delta/2^{\,n_0})$ for infinitely many radii. Thus x_0 cannot be an isolated point of K, contradicting the hypothesis.

Therefore no such sequence (s_n) can exist, and we conclude that $\gamma(t) \to x_0$ as $t \to \infty$. This shows that $K = \{x_0\}$. □

NeurIPS Paper Checklist

(i). **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The claims made in the abstract and introduction accurately reflect the paper's theoretical contributions and scope. They are supported by the main theorems (Theorems 2.2, 2.3, 3.1, 4.2, and 4.3) and are stated under explicit assumptions.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

(ii). **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Limitations of this work is discussed in Remark 4.4.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

(iii). **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We do provide the full set of assumptions and complete proofs for the main theorems.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

(iv). **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and conclusions of the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

(v). **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide open access to the data and code in Supplementary Material, with sufficient instructions to faithfully reproduce the main experimental results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

(vi). **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results in Section 5.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

(vii). **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports suitably and correctly defined error bars.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

(viii). **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

(ix). **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

(x). **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This work has no societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

(xi). **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release any data, pre-trained models, image generators, or scraped datasets. Therefore, safeguards for responsible release are not applicable.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

(xii). **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper is primarily theoretical and does not rely on external datasets or pre-trained models. All prior works and relevant external resources used in the paper are properly cited, and no restricted assets are used.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

(xiii). **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce new assets such as datasets, trained models, or software packages. Therefore, asset documentation is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

(xiv). **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. Therefore, participant instructions, screenshots, and compensation details are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

(xv). **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve study participants or human-subject research. Therefore, participant risks, risk disclosure, and IRB approval are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

(xvi). **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of the core methodology. Any use of LLMs is limited to writing, editing, or formatting assistance and does not affect the scientific content, rigor, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.