OPINION LEADER DYNAMICS: HOW SPARSE ATTENTION SHAPES TOKEN CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse attention reduces the quadratic cost of global self-attention while retaining strong empirical performance, but how its restricted interactions shape the evolution of token representations remains theoretically underexplored. Modeling tokens as particles on the unit sphere, we introduce **opinion leader dynamics**, a framework that identifies two mechanisms through which token groups converge internally while maintaining distinct limiting directions. In the explicit model, fixed representatives induce a potential that attracts tokens toward distinct local maxima. In the implicit model, disconnected interaction groups evolve toward separate consensus directions. We formulate both models as *reverse Wasserstein gradient flows* and establish exponential convergence under suitable conditions. We further connect these theoretical predictions to token evolution in frontier sparse-attention LLMs that motivate our framework. Across four benchmarks, **Kimi-K3**, **MiniMax-M3**, and **DeepSeek-V4-Flash** consistently exhibit clearer cluster separation and higher clustering scores than the dense-attention model **GLM-4.7-Flash** in projected token representations. These observations support the relevance of the predicted multiple-group structure to trained frontier LLMs, while finite-particle simulations illustrate the theoretical convergence behavior. Together, our results connect restricted token interactions to distinct group-level attractors, providing a dynamical account of how sparse attention can support alignment within groups while preserving separation between them.

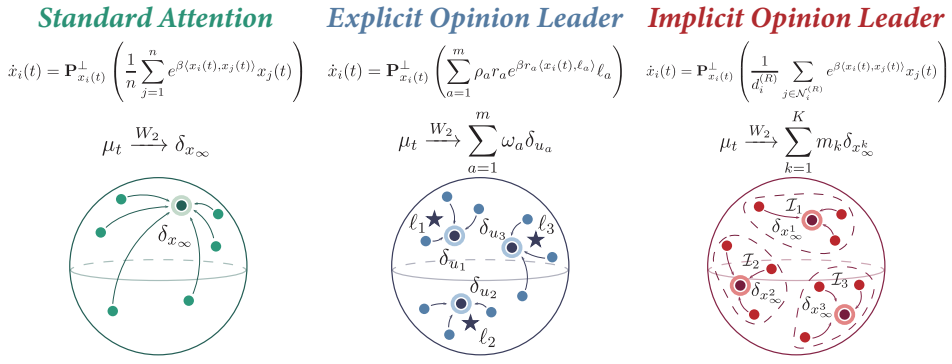

Figure 1: Comparison of three attention dynamics under the respective assumptions of our analysis. ***Left:*** **Standard Attention** drives tokens toward a single consensus x_∞. ***Middle:*** **Explicit opinion leader dynamics** guide token groups toward distinct local maxima u_a of the potential induced by fixed representatives $z_a = r_a \ell_a$. ***Right:*** **Implicit opinion leader dynamics** align tokens within each disconnected component $\mathcal{I}_k$, yielding distinct consensus directions x_∞^k.

1 INTRODUCTION

Transformers (Vaswani et al., 2017) underpin modern large language models (LLMs) (Achiam et al., 2023; Team et al., 2026; Xu et al., 2026; Lai et al., 2026; Zeng et al., 2026), with self-attention governing how tokens exchange and integrate information. Global self-attention allows every token

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

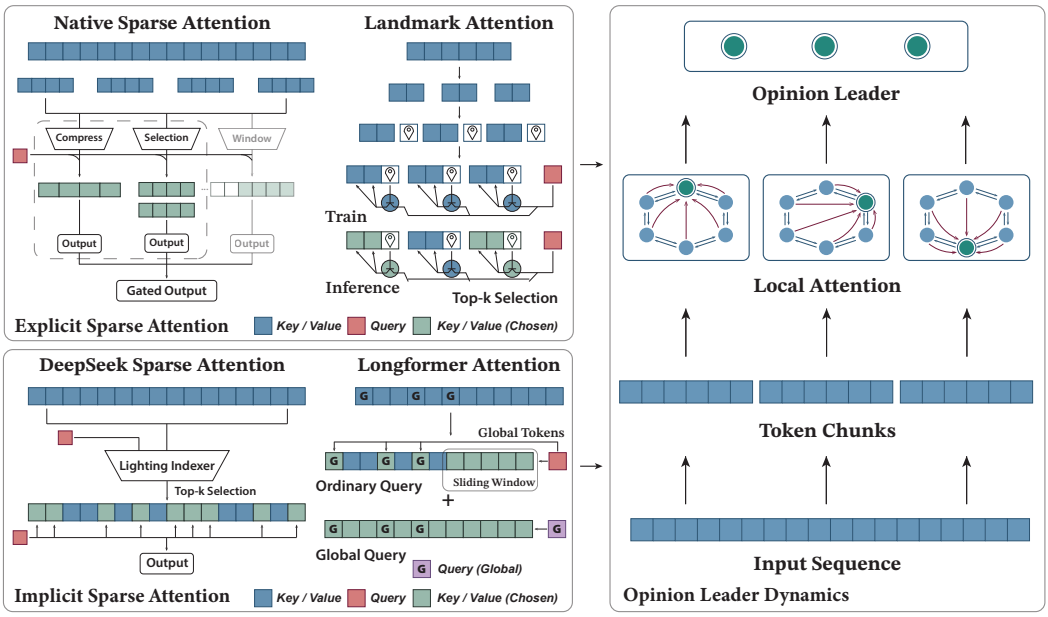

Figure 2: Sparse attention architectures motivating our *opinion leader dynamics* framework. ***Top left:*** Native Sparse Attention (Yuan et al., 2025) and Landmark Attention (Mohtashami & Jaggi, 2023) illustrate mechanisms that use compressed or designated representatives, motivating our *explicit* model. ***Bottom left:*** DeepSeek Sparse Attention (Liu et al., 2025) and Longformer (Beltagy et al., 2020) illustrate mechanisms that restrict interactions through selective or structured attention patterns, motivating our *implicit* model. ***Right:*** A shared schematic of both mechanisms. In the explicit model, fixed representatives induce group-specific opinion leaders; in the implicit model, opinion leaders emerge through groupwise consensus under restricted interactions.

to interact with all other tokens, but its quadratic computational cost limits its efficiency on long sequences. Sparse attention addresses this limitation by restricting interactions or organizing them through compressed representations, often retaining strong empirical performance (Tay et al., 2022; Tang et al., 2024; Lai et al., 2025). These changes affect not only the cost of computation but also the pathways through which tokens influence one another. This raises a fundamental question: *how do restricted interactions shape the clustering and long-term evolution of token representations?*

Recent theoretical work provides a starting point by modeling tokens as interacting particles on the unit sphere and studying their mean-field dynamics (Geshkovski et al., 2023; 2025; Rigollet, 2026). Within this framework, Chen et al. (2025) establish exponential convergence to a Dirac mass at a single consensus direction under suitable conditions. This result characterizes a regime in which global interactions progressively align token representations. Sparse attention changes the underlying interaction structure, raising the possibility that distinct token groups can retain separate limiting directions. Characterizing the conditions and mechanisms supporting such behavior still remains theoretically underexplored.

Existing sparse attention architectures motivate two complementary mechanisms. Methods based on compressed or landmark representations summarize token groups and use these summaries to aggregate information or guide retrieval (Mohtashami & Jaggi, 2023; Yuan et al., 2025; Hu et al., 2025). Such representatives can provide common reference directions toward which tokens evolve. Other methods restrict communication through predefined patterns or content-dependent key selection (Child et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020; Kitaev et al., 2020; Roy et al., 2021; Lu et al., 2025; Liu et al., 2025). These restrictions can favor interaction within token groups, motivating the study of consensus within groups that evolve independently. We isolate these mechanisms through fixed representatives and disconnected interaction components, respectively. These idealizations allow us to characterize two routes to distinct limiting directions and examine their relevance to the token dynamics of frontier sparse-attention LLMs.

Our contributions. We introduce **opinion leader dynamics**, comprising explicit dynamics driven by fixed representatives and implicit dynamics driven by restricted mutual interactions. Both model tokens as particles on the unit sphere, with an *opinion leader* denoting a group's limiting direction (see Fig. 1). Fig. 2 illustrates the motivating architectures and their shared schematic abstraction.

In the *explicit model*, tokens interact with fixed representatives $\{z_a\}_{a=1}^m$, where $z_a = r_a \ell_a$ has magnitude r_a and unit direction ℓ_a. These representatives jointly induce a time-independent potential, whose attracting local maxima serve as explicit opinion leaders. The token dynamics are given by

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\sum_{a=1}^{m} \rho_a r_a e^{\beta r_a \langle x_i(t), \ell_a \rangle} \ell_a \right), \qquad x_i(t) \in \mathbb{S}^{d-1}. \tag{1}$$

Here, ρ_a is the weight of representative a, $\beta > 0$ is the inverse temperature, and $\mathbf{P}^{\perp}_x(y) = y - \langle x, y \rangle x$ projects onto the tangent space $T_x\mathbb{S}^{d-1}$. Because all representatives contribute to the potential, the attracting directions need not coincide with the representative directions themselves.

In the *implicit model*, tokens interact through a fixed sparse graph. Fixed interaction coordinates and a radius R define a symmetric graph with $K \geq 2$ nontrivial connected components $\{\mathcal{I}_k\}_{k=1}^K$. Within each component, tokens evolve according to

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\frac{1}{d_i^{(R)}} \sum_{j \in \mathcal{N}_i^{(R)}} e^{\beta \langle x_i(t), x_j(t) \rangle} x_j(t) \right), \qquad i \in \mathcal{I}_k. \tag{2}$$

Here, $\mathcal{N}_i^{(R)} \subseteq \mathcal{I}_k$ is the interaction neighborhood of token i, and $d_i^{(R)} = |\mathcal{N}_i^{(R)}|$. Each component therefore evolves independently, with its consensus direction determined by the coupled evolution of its tokens. These emergent consensus directions serve as implicit opinion leaders.

Our contributions combine convergence guarantees for both mechanisms with empirical evidence from numerical simulations and frontier sparse-attention LLMs:

1. **Explicit opinion leader dynamics: convergence under fixed representatives.** We formulate token evolution under fixed representatives as a *reverse Wasserstein gradient flow*. We establish convergence of token trajectories to critical points of the induced potential for arbitrary initialization. Under geometric separation and dominance conditions, we further prove exponential convergence of each group to a distinct attracting local maximum.
2. **Implicit opinion leader dynamics: consensus under restricted interactions.** We analyze token evolution on a fixed sparse graph. We establish a *reverse Wasserstein gradient flow* formulation in a degree-weighted product space for each connected component. Under suitable graph and initialization conditions, the components converge exponentially to distinct consensus directions. The convergence-rate bound depends explicitly on the normalized Laplacian spectral gap, linking internal connectivity to the rate of token alignment.
3. **Empirical validation of predicted clustering behavior in frontier LLMs.** We analyze token evolution in **Kimi-K3** (Team et al., 2026), **MiniMax-M3** (Lai et al., 2026), and **DeepSeek-V4-Flash** (Xu et al., 2026), motivating **Explicit**, **Implicit**, and **Hybrid** interpretations, respectively. Across four benchmarks, all three exhibit clearer cluster separation and higher clustering scores in projected representations than the dense-attention model **GLM-4.7-Flash** (Zeng et al., 2025), consistent with the multiple-group structure predicted by our theory. Finite-particle simulations further illustrate the theoretical convergence behavior.

2 Opinion Leader Dynamics

2.1 Explicit Opinion Leader Dynamics

The explicit model describes token evolution under a potential induced by fixed representatives. We first construct the representatives from initial token groups, then characterize the resulting dynamics for arbitrary initialization and under stronger conditions ensuring separated attractors.

Let $\{\mathcal{I}_a\}_{a=1}^m$ be a partition of $\{1,\dots,n\}$ into nonempty groups, with initial token states $x_i(0) \in \mathbb{S}^{d-1}$. For each group, define its size, relative mass, and initial empirical distribution:

$$n_a = |\mathcal{I}_a|, \qquad \omega_a = \frac{n_a}{n}, \qquad \mu_0^a = \frac{1}{n_a}\sum_{i\in\mathcal{I}_a}\delta_{x_i(0)}. \tag{3}$$

The initial token distribution is therefore

$$\mu_0 = \sum_{a=1}^m \omega_a \mu_0^a. \tag{4}$$

A compression map C_θ assigns each group a nonzero representative:

$$z_a = \mathsf{C}_\theta\left((x_i(0))_{i\in\mathcal{I}_a}\right) \neq 0, \qquad r_a = \|z_a\|_2, \qquad \ell_a = \frac{z_a}{r_a} \in \mathbb{S}^{d-1}. \tag{5}$$

The representatives remain fixed throughout the subsequent evolution. More generally, we use the measure-valued formulation in Eq. 4, allowing arbitrary $\mu_0^a \in \mathcal{P}(\mathbb{S}^{d-1})$, with $\omega_a > 0$ and $\sum_{a=1}^m \omega_a = 1$. Each initial group distribution is associated with a fixed nonzero representative z_a.

Define the representative measure

$$\nu = \sum_{a=1}^m \rho_a \delta_{z_a}, \qquad \rho_a > 0, \qquad \sum_{a=1}^m \rho_a = 1. \tag{6}$$

The weights ω_a and ρ_a play different roles: ω_a specifies the mass of token group a, whereas ρ_a controls the contribution of its representative to the interaction field. The representatives are computed at initialization and remain fixed throughout the dynamics. For $\beta > 0$, the time-independent measure ν induces the potential Φ_ν and its spherical gradient field $\mathcal{X}_{\nu,\beta}$:

$$\begin{aligned}\Phi_\nu(x) &= \frac{1}{\beta}\int_{\mathbb{R}^d} e^{\beta\langle x,z\rangle} d\nu(z) = \frac{1}{\beta}\sum_{a=1}^m \rho_a e^{\beta r_a\langle x,\ell_a\rangle},\\ \mathcal{X}_{\nu,\beta}(x) &= \nabla^\circ\Phi_\nu(x) = \sum_{a=1}^m \rho_a r_a e^{\beta r_a\langle x,\ell_a\rangle}\mathbf{P}_x^\perp(\ell_a).\end{aligned} \tag{7}$$

Here, ∇° denotes the spherical gradient. All tokens evolve under this shared field, so the attracting directions are determined jointly by all representatives. The smooth field $\mathcal{X}_{\nu,\beta}$ generates a global flow ϕ_t on $\mathbb{S}^{d-1}$. Define the evolving group distributions and their mixture by

$$\mu_t^a = (\phi_t)_\# \mu_0^a, \qquad \mu_t = \sum_{a=1}^m \omega_a \mu_t^a = (\phi_t)_\#\mu_0. \tag{8}$$

For both empirical and general initial distributions, μ_t satisfies the continuity equation

$$\partial_t \mu_t + \operatorname{div}^\circ(\mu_t \mathcal{X}_{\nu,\beta}) = 0. \tag{9}$$

where div° denotes the spherical divergence. Consider the energy

$$\mathbf{E}_\beta[\mu|\nu] = \int_{\mathbb{S}^{d-1}} \Phi_\nu(x)\, d\mu(x). \tag{10}$$

Along Eq. 9, it satisfies $\frac{d}{dt}\mathbf{E}_\beta[\mu_t|\nu] = \int_{\mathbb{S}^{d-1}} \|\mathcal{X}_{\nu,\beta}(x)\|_2^2\, d\mu_t(x) \geq 0$. Thus, individual tokens follow spherical gradient ascent on Φ_ν, while their distribution evolves as the *reverse Wasserstein gradient flow* of $\mathbf{E}_\beta[\cdot \mid \nu]$. Appendix F.1.1 provides the derivation.

We first establish asymptotic convergence without imposing geometric conditions on the initial groups. Let $\operatorname{Crit}(\Phi_\nu)$ denote the critical set of Φ_ν, and let W_2 be the Wasserstein distance induced by the spherical geodesic distance $d_{\mathbb{S}}$. Additional notation is summarized in Table 2 of Appendix.

Corollary 2.1 (Global asymptotic convergence). *For every $x \in \mathbb{S}^{d-1}$, the limit $\phi_\infty(x) = \lim_{t\to\infty}\phi_t(x)$ exists and belongs to $\operatorname{Crit}(\Phi_\nu)$. Define $\mu_\infty^a = (\phi_\infty)_\#\mu_0^a$ for each a, and $\mu_\infty = \sum_{a=1}^m \omega_a\mu_\infty^a = (\phi_\infty)_\#\mu_0$. Then, as $t\to\infty$,*

$$W_2(\mu_t^a, \mu_\infty^a) \longrightarrow 0 \quad \textit{for every } a, \qquad W_2(\mu_t, \mu_\infty) \longrightarrow 0. \tag{11}$$

Moreover,

$$\operatorname{supp}(\mu_\infty^a) \subseteq \operatorname{Crit}(\Phi_\nu) \quad \textit{for every } a, \qquad \operatorname{supp}(\mu_\infty) \subseteq \operatorname{Crit}(\Phi_\nu). \tag{12}$$

Corollary 2.1 guarantees convergence of every token trajectory, while allowing each group to reach multiple critical points and tokens from different groups to share the same limit. The proof uses the analyticity of Φ_ν and is given in Appendix F.1.2.

To obtain one attracting direction per group and an explicit exponential rate, we next impose concentration, separation, and representative-dominance conditions. Let $M_0^a = \int_{\mathbb{S}^{d-1}} x \, d\mu_0^a(x)$ be the initial mean of group a. We write $S_\theta^+(v) = \{x \in \mathbb{S}^{d-1} : \langle x, v\rangle \geq \cos\theta\}$ for the closed spherical cap of radius θ centered at v.

Assumption 2.2 (**Geometric separation and dominance**). Assume $M_0^a \neq 0$ for every a, and define $c_a = M_0^a/\|M_0^a\|_2$. Suppose that $m \geq 2$ and that there exist $\delta, \Delta, \eta > 0$ and $\alpha \in (0, \pi/2)$ such that

$$\begin{aligned} &\operatorname{supp}(\mu_0^a) \subseteq S_\delta^+(c_a), \qquad d_{\mathbb{S}}(\ell_a, c_a) \leq \eta, \qquad a = 1, \dots, m, \\ &d_{\mathbb{S}}(c_a, c_b) \geq \Delta, \qquad\qquad\qquad\qquad 1 \leq a < b \leq m, \end{aligned} \tag{13}$$

with

$$\delta + \eta < \alpha, \qquad 2(\alpha + \eta) < \Delta. \tag{14}$$

Assume further that

$$\mathbf{I}_a(\alpha) > 0, \qquad \mathbf{H}_a(\alpha) > 0, \qquad a = 1, \dots, m, \tag{15}$$

where $\mathbf{I}_a(\alpha)$ and $\mathbf{H}_a(\alpha)$ are defined in Table 2.

The geometric conditions place each initial group inside a cap around its representative and keep these caps separated. The dominance conditions ensure that the field points inward at each cap boundary and that the potential is uniformly strictly concave within the cap. Together, they yield a unique attracting direction for each group.

Theorem 2.3 (**Exponential convergence to separated attractors**). *Under Assumption 2.2, the caps $S_\alpha^+(\ell_a)$ are pairwise disjoint and forward invariant. Each cap contains a unique critical point u_a of Φ_ν, which is an interior nondegenerate attracting local maximizer. If B_a denotes its attraction basin, then*

$$\operatorname{supp}(\mu_0^a) \subseteq S_{\delta+\eta}^+(\ell_a) \Subset \operatorname{int} S_\alpha^+(\ell_a) \subseteq B_a. \tag{16}$$

The attractor locations satisfy

$$d_{\mathbb{S}}(u_a, \ell_a) \leq \alpha, \qquad d_{\mathbb{S}}(u_a, c_a) \leq \alpha + \eta. \tag{17}$$

For $a \neq b$, their separation is bounded below by

$$d_{\mathbb{S}}(u_a, u_b) \geq d_{\mathbb{S}}(c_a, c_b) - 2(\alpha + \eta) \geq \Delta - 2\eta - 2\alpha > 0. \tag{18}$$

For each a, define $D_{0,a}^2 = \int_{\mathbb{S}^{d-1}} d_{\mathbb{S}}(x, u_a)^2 \, d\mu_0^a(x)$. Then

$$\operatorname{supp}(\mu_t^a) \subseteq S_\alpha^+(\ell_a), \qquad W_2(\mu_t^a, \delta_{u_a}) \leq D_{0,a} e^{-\mathbf{H}_a(\alpha)t}, \qquad t \geq 0. \tag{19}$$

Define

$$\mathbf{H}_{\min}(\alpha) = \min_{1 \leq a \leq m} \mathbf{H}_a(\alpha), \qquad \mu_\infty = \sum_{a=1}^m \omega_a \delta_{u_a}, \qquad D_0^2 = \sum_{a=1}^m \omega_a D_{0,a}^2. \tag{20}$$

The limiting distribution has exactly m support points: $\operatorname{supp}(\mu_\infty) = \{u_1, \dots, u_m\}$. *Moreover,*

$$W_2(\mu_t, \mu_\infty) \leq D_0 e^{-\mathbf{H}_{\min}(\alpha)t} \leq (\alpha + \delta + \eta) e^{-\mathbf{H}_{\min}(\alpha)t}, \qquad t \geq 0. \tag{21}$$

Theorem 2.3 shows that each group contracts toward its own *explicit opinion leader* u_a while remaining separated from the other groups. The bound $\mathbf{H}_a(\alpha)$ quantifies the local curvature driving this contraction, and the slowest group determines the guaranteed rate for the full distribution. Each leader is a local maximizer of the potential generated by all representatives, rather than necessarily the representative direction ℓ_a itself. The proof is given in Appendix F.1.3.

2.2 IMPLICIT OPINION LEADER DYNAMICS

The implicit model describes alignment through mutual token interactions on a fixed sparse graph. We first identify the connected components, then establish consensus within each component under suitable initial geometry. The limiting directions emerge from the coupled dynamics of the tokens.

Let $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space and associate each token i with a fixed interaction coordinate $p_i \in \mathcal{Z}$. These coordinates determine which tokens interact, while the token states evolve on $\mathbb{S}^{d-1}$. For an interaction radius $R > 0$, define the neighborhoods and adjacency entries by

$$\mathcal{N}_i^{(R)} = \{j \in \{1, \dots, n\} : d_{\mathcal{Z}}(p_i, p_j) \le R\}, \qquad A_{ij}^{(R)} = \mathbf{1}\left\{j \in \mathcal{N}_i^{(R)}\right\}, \tag{22}$$

Set $d_i^{(R)} = |\mathcal{N}_i^{(R)}|$. Symmetry of the metric and the identity $d_{\mathcal{Z}}(p_i, p_i) = 0$ ensure that $A^{(R)}$ defines an undirected graph G_R with a self-loop at every node.

We write $i \sim_R j$ if tokens i and j are connected by a path in G_R. Equivalently, there exist an integer $s \ge 0$ and a sequence $i_0, \dots, i_s$ such that

$$i_0 = i, \qquad i_s = j, \qquad d_{\mathcal{Z}}(p_{i_{r-1}}, p_{i_r}) \le R, \quad r = 1, \dots, s. \tag{23}$$

We call this relation R-*reachability*. To obtain multiple interaction groups, each containing at least two tokens, we impose the following condition on R.

Assumption 2.4 (Separated interaction scales). Let $\mathbb{T}_n$ denote the set of spanning trees on $\{1, \dots, n\}$, and let $\mathcal{E}(T)$ be the edge set of $T \in \mathbb{T}_n$. Define

$$R_{\mathrm{loc}} = \max_{i \in \{1,\dots,n\}} \min_{j \in \{1,\dots,n\} \setminus \{i\}} d_{\mathcal{Z}}(p_i, p_j), \quad R_{\mathrm{conn}} = \min_{T \in \mathbb{T}_n} \max_{\{i,j\} \in \mathcal{E}(T)} d_{\mathcal{Z}}(p_i, p_j). \tag{24}$$

The interaction radius satisfies $R_{\mathrm{loc}} \le R < R_{\mathrm{conn}}$.

The lower bound R_{loc} ensures that every token has a neighbor other than itself. The upper bound R_{conn} keeps R below the smallest radius at which G_R becomes connected.

For the reachability relation $\sim_R$, the equivalence classes $\{\mathcal{I}_k\}_{k=1}^K$ are precisely the connected components of G_R, giving the partition

$$\{1, \dots, n\} = \bigsqcup_{k=1}^{K} \mathcal{I}_k, \qquad n_k = |\mathcal{I}_k|, \qquad \sum_{k=1}^{K} n_k = n. \tag{25}$$

Assumption 2.4 guarantees $K \ge 2$ and $n_k \ge 2$ for every k. Since $\mathcal{N}_i^{(R)} \subseteq \mathcal{I}_k$ for $i \in \mathcal{I}_k$, Eq. 2 decomposes into K independent particle systems on the connected subgraphs $G_R[\mathcal{I}_k]$. These subgraphs need not be complete: the analysis retains the sparse interactions within each component. Appendix F.2.1 provides the full proof.

For each component, let $\mathbf{A}_k$ and $\mathbf{D}_k$ denote its adjacency and degree matrices. The associated random walk has transition probabilities $\mathsf{P}_{ij}^k = A_{ij}^{(R)}/d_i^{(R)}$ and stationary weights $\pi_i^k = d_i^{(R)}/\operatorname{vol}_k$, where $\operatorname{vol}_k = \sum_{i \in \mathcal{I}_k} d_i^{(R)}$. Symmetry of $\mathbf{A}_k$ gives the detailed-balance identity $\pi_i^k \mathsf{P}_{ij}^k = \pi_j^k \mathsf{P}_{ji}^k$. Connectivity ensures that the normalized Laplacian gap $\lambda_k = \lambda_2(\mathbf{I}_{n_k} - \mathbf{D}_k^{-1/2}\mathbf{A}_k\mathbf{D}_k^{-1/2})$ is positive, where λ_2 denotes the second-smallest eigenvalue. Table 3 summarizes the notation.

To express the dynamics in a Wasserstein framework while preserving the graph structure, we associate each node $i \in \mathcal{I}_k$ with a measure $\nu_{i,t}^k \in \mathcal{P}_2(\mathbb{S}^{d-1})$ and write $\boldsymbol{\nu}_t^k = (\nu_{i,t}^k)_{i \in \mathcal{I}_k}$. The velocity field and continuity equation are

$$\mathcal{X}_{i,\boldsymbol{\nu}^k,\beta}(x) = \sum_{j \in \mathcal{I}_k} \mathsf{P}_{ij}^k \int_{\mathbb{S}^{d-1}} \mathbf{P}_x^{\perp}(y) e^{\beta\langle x,y\rangle}\, d\nu_j^k(y), \; \partial_t \nu_{i,t}^k + \operatorname{div}^{\circ}\left(\nu_{i,t}^k \mathcal{X}_{i,\boldsymbol{\nu}_t^k,\beta}\right) = 0, \; i \in \mathcal{I}_k. \tag{26}$$

Choosing Dirac initial measures yields $\nu_{i,t}^k = \delta_{x_i(t)}$ and recovers the particle dynamics in Eq. 2.

The stationary weights induce the degree-weighted product Wasserstein metric

$$\mathcal{W}_{2,\pi^k}^2\left(\boldsymbol{\nu}^k, \boldsymbol{\eta}^k\right) = \sum_{i \in \mathcal{I}_k} \pi_i^k W_2^2\left(\nu_i^k, \eta_i^k\right). \tag{27}$$

Smoothness of the interaction kernel and compactness of the sphere ensure a unique global characteristic solution for every initial tuple. This solution is locally absolutely continuous in $\mathcal{W}_{2,\pi^k}$, as established in Lemma F.1. Detailed balance yields a graph interaction energy whose *reverse Wasserstein gradient flow* is Eq. 26 in this weighted product space. Appendix F.2.2 gives the energy and its derivation. We use two aggregate measures to relate this formulation to token distributions:

$$\mu_t^k = \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \nu_{i,t}^k, \qquad \widehat{\mu}_t^k = \sum_{i \in \mathcal{I}_k} \pi_i^k \nu_{i,t}^k, \qquad m_k = \frac{n_k}{n}, \qquad \mu_t = \sum_{k=1}^{K} m_k \mu_t^k. \tag{28}$$

The uniform aggregate μ_t^k assigns equal weight to each token, while $\widehat{\mu}_t^k$ uses the degree weights natural to the graph dynamics. We next impose concentration within components and separation between them. Let $\Omega_k = \bigcup_{i \in \mathcal{I}_k} \operatorname{supp}(\nu_{i,0}^k)$ denote the union of initial supports in component k.

Assumption 2.5 (Concentrated and separated initial states). There exist $\alpha \in [0, \pi/4)$ and directions $u_1, \ldots, u_K \in \mathbb{S}^{d-1}$ such that

$$\Omega_k \subseteq S_\alpha^+(u_k), \quad k = 1, \ldots, K, \qquad d_{\mathbb{S}}(u_k, u_\ell) > 2\alpha, \quad 1 \le k < \ell \le K. \tag{29}$$

The directions u_k specify the centers of the initial caps; they are not prescribed consensus directions.

This assumption places each component inside a spherical cap, with a positive gap between caps. The proof shows that these caps remain invariant and that disagreement within each component decays exponentially. The convergence bounds use the initial degree-weighted disagreement V_0^k and the degree-imbalance factor χ_k, which accounts for the difference between uniform and degree-weighted aggregates. Their definitions, together with the constants κ_k, C_k, $\kappa_{\min}$, and C_G used below, are collected in Table 3.

Theorem 2.6 (Componentwise exponential consensus). *Under Assumptions 2.4 and 2.5, each component remains in its initial cap:*

$$\operatorname{supp}(\nu_{i,t}^k) \subseteq S_\alpha^+(u_k), \qquad t \ge 0, \quad k = 1, \ldots, K, \quad i \in \mathcal{I}_k. \tag{30}$$

There exist pairwise distinct directions $x_\infty^k \in S_\alpha^+(u_k)$ such that, with $\boldsymbol{\nu}_\infty^k = (\delta_{x_\infty^k})_{i \in \mathcal{I}_k}$,

$$\mathcal{W}_{2,\pi^k}\left(\boldsymbol{\nu}_t^k, \boldsymbol{\nu}_\infty^k\right) \le C_k (V_0^k)^{1/2} e^{-\kappa_k t}, \qquad t \ge 0. \tag{31}$$

Consequently, the full token distribution converges to a mixture of K distinct point masses:

$$\mu_\infty = \sum_{k=1}^{K} m_k \delta_{x_\infty^k}, \qquad W_2(\mu_t, \mu_\infty) \le C_G e^{-\kappa_{\min} t}, \qquad t \ge 0. \tag{32}$$

Theorem 2.6 establishes internal alignment within each component while preserving separation across components. The limiting directions x_∞^k are the *implicit opinion leaders*, determined by the initial states and interactions within their respective groups. The graph structure enters the convergence bound through $\kappa_k = \lambda_k \cos(2\alpha) e^{\beta \cos(2\alpha)}$. For fixed α and β, this guaranteed rate scales as $\Theta(n_k^{-2})$ for self-looped path graphs and as $\Theta(1)$ for graph families with a uniformly positive normalized Laplacian gap. Thus, the theorem quantifies how connectivity within a group affects its guaranteed rate of consensus. The full proof is provided in Appendix F.2.3.

3 Experiments

3.1 Experimental Setup

Numerical simulations. We simulate Standard Attention dynamics, implicit opinion leader dynamics, and explicit opinion leader dynamics under both general initialization (Corollary 2.1) and initialization satisfying the geometric conditions of Theorem 2.3. Implementation details are provided in Appendix G.1.1. Additional simulations of explicit opinion leader dynamics under the Theorem 2.3's initialization conditions appear in Appendix G.3.

Frontier LLM analysis. We track layerwise evolution of token representations in three frontier sparse-attention LLMs: **Kimi-K3** (Team et al., 2026), **MiniMax-M3** (Lai et al., 2026), and

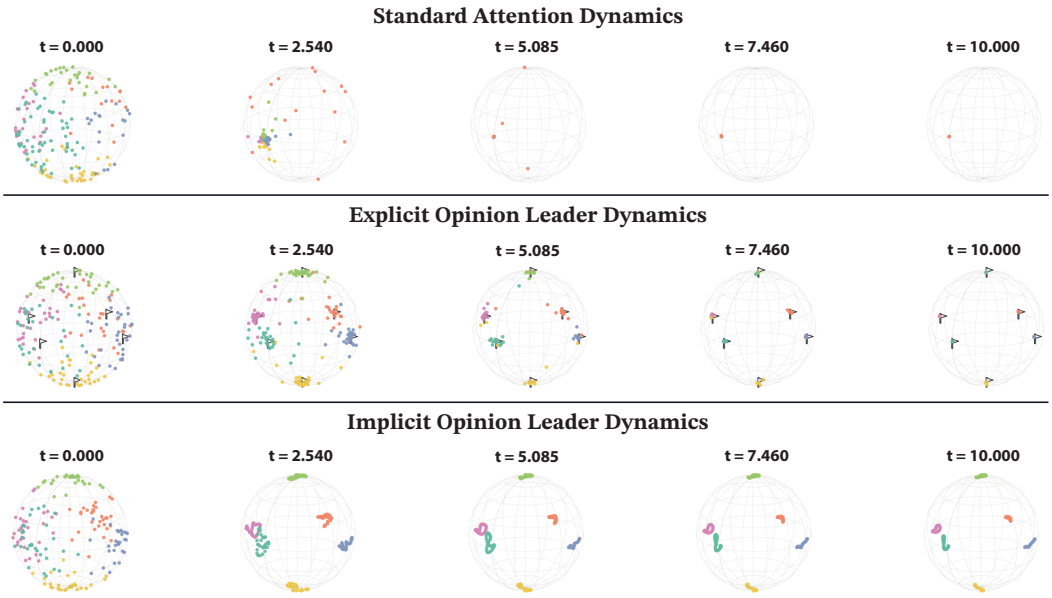

Figure 3: Finite-particle simulations on $\mathbb{S}^2$ of Standard Attention dynamics (*top*), explicit opinion leader dynamics under general initialization (*middle*), and implicit opinion leader dynamics (*bottom*). Columns show five snapshots over $t \in [0, 10]$. Colors are assigned at $t = 0$ and remain fixed.

DeepSeek-V4-Flash (Xu et al., 2026). We also use **GLM-4.7-Flash** (Zeng et al., 2025), which employs standard dense attention for comparison. The analysis covers inputs from four benchmarks: HumanEval (Chen et al., 2021), HellaSwag (Zellers et al., 2019), ARC-Easy (Clark et al., 2018), and MATH (Hendrycks et al., 2021). These benchmarks provide diverse inputs for examining representation dynamics.

We obtain low-dimensional embeddings using UMAP (McInnes et al., 2018), and its spherical variants, with spherical visualizations displaying token representations on $\mathbb{S}^2$. We apply HDBSCAN (McInnes et al., 2017) to the reduced representations to identify token clusters. The resulting cluster labels determine token colors, which remain fixed across layers so that the earlier representations of each final-layer group can be inspected. Extended results are provided in Appendix G.4.

Measuring cluster separation. We use the *cosine silhouette score* (Rousseeuw, 1987) to quantify ***within-cluster cohesion*** and ***between-cluster separation***. For each token, the score compares its average cosine distance to tokens in its own cluster with that to the nearest competing cluster. Scores range in $[-1, 1]$: higher values indicate more cohesive and better-separated clusters, values near zero indicate weak separation, and negative values indicate greater proximity to another cluster. We average the tokenwise scores to obtain an overall measure of clustering quality; the detailed computation is provided in Appendix G.2. All scores are computed in the reduced representation space. For reference, we apply the same analysis to standard Gaussian vectors with matched dimensions.

3.2 Numerical Simulations

We compare the three dynamics under the initialization settings described below. For Standard Attention, particles are sampled independently from a von Mises–Fisher distribution on $\mathbb{S}^2$. For *explicit opinion leader*, particles are initialized uniformly on $\mathbb{S}^2$ and evolve under $m = 6$ fixed representatives. For *implicit opinion leader*, the $n = 180$ particles are divided equally into $K = 6$ disconnected components, with initial states satisfying our concentration and separation conditions.

Fig. 3 illustrates the resulting contrast. Standard Attention progressively aligns particles toward a single consensus direction. Explicit opinion leader dynamics organize initially dispersed particles into multiple clusters around attractors of the representative-induced potential, whereas implicit opinion leader dynamics align particles within each connected component while maintaining separation. The standard and implicit simulations illustrate the consensus behavior characterized by

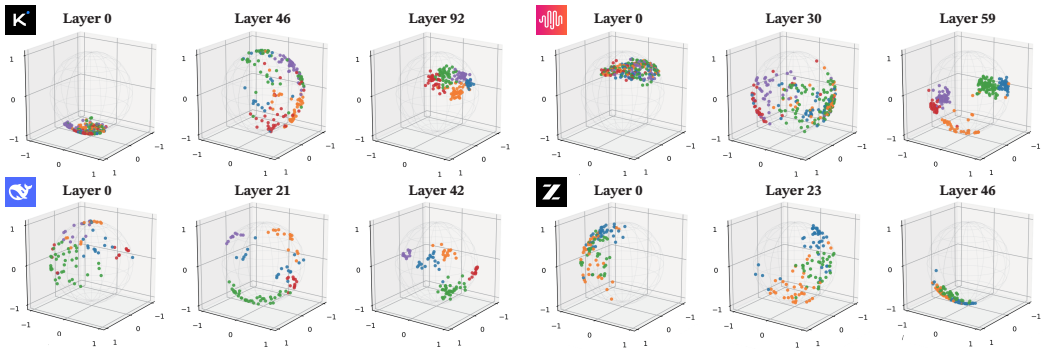

Figure 4: Layerwise evolution of token hidden states for ARC-Easy samples in **Kimi-K3**, **MiniMax-M3**, **DeepSeek-V4-Flash**, and **GLM-4.7-Flash**, visualized on unit sphere using spherical UMAP.

Theorems E.2 and 2.6. The explicit simulation illustrates convergence to critical points as established by Corollary 2.1, with the chosen representatives yielding several attracting clusters.

Table 1: Cosine silhouette scores across LLMs and benchmarks. For each model, the *upper* row reports mean final-layer scores across 100 random samples after projecting hidden states with spherical UMAP, while the *lower* row gives the Gaussian reference score at the same hidden dimension.

Model	HumanEval	ARC-Easy	HellaSwag	MATH
Kimi-K3	**0.77 ± 0.22 (↑ 0.60)**	**0.87 ± 0.15 (↑ 0.68)**	**0.77 ± 0.17 (↑ 0.62)**	**0.86 ± 0.18 (↑ 0.67)**
	0.17 ± 0.04	0.19 ± 0.05	0.15 ± 0.05	0.19 ± 0.04
MiniMax-M3	**0.87 ± 0.07 (↑ 0.71)**	**0.84 ± 0.07 (↑ 0.67)**	**0.85 ± 0.09 (↑ 0.69)**	**0.83 ± 0.11 (↑ 0.66)**
	0.16 ± 0.04	0.17 ± 0.04	0.16 ± 0.04	0.17 ± 0.04
DeepSeek-V4-Flash	**0.71 ± 0.17 (↑ 0.52)**	**0.85 ± 0.12 (↑ 0.63)**	**0.64 ± 0.19 (↑ 0.47)**	**0.84 ± 0.12 (↑ 0.63)**
	0.19 ± 0.06	0.22 ± 0.12	0.17 ± 0.04	0.21 ± 0.13
GLM-4.7-Flash	0.35 ± 0.27 (↑ 0.19)	0.24 ± 0.17 (↑ 0.08)	0.34 ± 0.26 (↑ 0.19)	0.30 ± 0.20 (↑ 0.13)
	0.16 ± 0.06	0.16 ± 0.13	0.15 ± 0.04	0.17 ± 0.12

3.3 Frontier LLM Analysis

Layerwise token dynamics. Fig. 4 visualizes token representations for ARC-Easy samples using spherical UMAP. **Kimi-K3**, **MiniMax-M3**, and **DeepSeek-V4-Flash** exhibit distinct final-layer clusters, whereas **GLM-4.7-Flash** shows weaker intermediate separation followed by a more concentrated final configuration. These observations are qualitatively consistent with the contrast between separated token groups and global alignment studied in our framework.

Quantitative cluster separation. Table 1 reports final-layer cosine silhouette scores over 100 samples per benchmark. Across all four benchmarks, the three sparse-attention LLMs score $\mathbf{0.64}$–$\mathbf{0.87}$, compared with 0.24–0.35 for **GLM-4.7-Flash**. Their gains over matched Gaussian references are also larger: $\mathbf{0.47}$–$\mathbf{0.71}$ versus 0.08–0.19. These consistent differences across models and benchmarks indicate stronger clustering in the representation space, supporting the relevance of the multiple-group structure predicted by our theory to frontier sparse-attention LLMs.

4 Conclusions

We introduce *opinion leader dynamics* to characterize two mechanisms for token clustering under sparse interactions: the **explicit** model where attraction toward representative-induced maxima and the **implicit** model consensus within disconnected interaction groups. Both admit reverse Wasserstein gradient flow formulations and converge exponentially to distinct limiting directions under suitable conditions. Finite-particle simulations illustrate these results. Across four benchmarks, **Kimi-K3**, **MiniMax-M3**, and **DeepSeek-V4-Flash** exhibit stronger clustering behavior than dense **GLM-4.7-Flash**, consistent with the predicted group structure. Together, these findings connect tractable convergence guarantees with representation patterns in frontier sparse-attention LLMs.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code generation and to improve the clarity, grammar, and readability of the manuscript. Generative AI tools did not directly generate or alter the experiment results reported in this paper. The authors reviewed and tested all AI-assisted code and carefully verified all AI-assisted manuscript edits. We take full responsibility for the final content, including all text, claims, results, and associated artifacts.

REFERENCES

Pierre-Antoine Absil, Robert Mahony, and Ben Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005. URL https://epubs.siam.org/doi/abs/10.1137/040605266.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. URL https://arxiv.org/abs/2004.05150.

Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. Emergence of meta-stable clustering in mean-field transformer models. In *ICLR*, 2025a. URL https://arxiv.org/abs/2410.23228.

Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. A multiscale analysis of mean-field transformers in the moderate interaction regime. *NeurIPS*, 2025b. URL https://arxiv.org/abs/2509.25040.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Quantitative clustering in mean-field transformer models. *arXiv preprint arXiv:2504.14697*, 2025. URL https://arxiv.org/abs/2504.14697.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. URL https://arxiv.org/abs/1904.10509.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *NeurIPS*, 2023. URL https://arxiv.org/abs/2305.05465.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 2025. URL https://arxiv.org/abs/2312.10794.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021. URL https://arxiv.org/abs/2103.03874.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *COLM*, 2024. URL https://arxiv.org/abs/2404.06654.

Xiang Hu, Jiaqi Leng, Jun Zhao, Kewei Tu, and Wei Wu. Hardware-aligned hierarchical sparse attention for efficient long-term memory access. *NeurIPS*, 2025. URL https://arxiv.org/abs/2504.16795.

Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. *NeurIPS*, 2024. URL https://arxiv.org/abs/2411.04990.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *ICLR*, 2020. URL https://arxiv.org/abs/2001.04451.

Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, and Xun Zhou. Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference. *ICLR*, 2025. URL https://arxiv.org/abs/2502.20766.

Xunhao Lai, Weiqi Xu, Yufeng Yang, Qiaorui Chen, Yang Xu, Lunbin Zeng, Xiaolong Li, Haohai Sun, Haichao Zhu, Vito Zhang, et al. Minimax sparse attention. *arXiv preprint arXiv:2606.13392*, 2026. URL https://arxiv.org/abs/2606.13392.

Jiaqi Leng, Xiang Hu, Junxiong Wang, Jianguo Li, Wei Wu, and Yucheng Lu. Understanding and improving length generalization in hierarchical sparse attention models. *ICLR*, 2026. URL https://arxiv.org/abs/2510.17196.

Jiaoda Li and Ryan Cotterell. Characterizing the expressivity of local attention in transformers. In *ACL*, 2026. URL https://arxiv.org/abs/2605.00768.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025. URL https://arxiv.org/abs/2512.02556.

Jingkun Liu, Yisong Yue, Max Welling, and Yue Song. Krause synchronization transformers. *ICML*, 2026. URL https://arxiv.org/abs/2602.11534.

Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, et al. Moba: Mixture of block attention for long-context llms. *NeurIPS*, 2025. URL https://arxiv.org/abs/2502.13189.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017. URL https://joss.theoj.org/papers/10.21105/joss.00205.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. URL https://arxiv.org/abs/1802.03426.

Amirkeivan Mohtashami and Martin Jaggi. Random-access infinite context length for transformers. *NeurIPS*, 2023. URL https://arxiv.org/abs/2305.16300.

Philippe Rigollet. The mean-field dynamics of transformers. In *International Congress of Mathematicians*, 2026. URL https://arxiv.org/abs/2512.01868.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 2021. URL https://arxiv.org/abs/2003.05997.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context llm inference. *ICML*, 2024. URL https://arxiv.org/abs/2406.10774.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022. URL https://arxiv.org/abs/2009.06732.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. URL https://arxiv.org/abs/2607.24653.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. URL https://arxiv.org/abs/1706.03762.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. URL https://arxiv.org/abs/2606.19348.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In *ACL*, 2025. URL https://arxiv.org/abs/2502.11089.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *NeurIPS*, 2020. URL https://arxiv.org/abs/2007.14062.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *ACL*, 2019. URL https://arxiv.org/abs/1905.07830.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025. URL https://arxiv.org/abs/2508.06471.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026. URL https://arxiv.org/abs/2602.15763.

APPENDIX CONTENTS

A Discussion and Future Work **14**

B Related Work **14**

C Extended Notation and Definitions **14**

D Interpreting Sparse Attention through Opinion Leader Dynamics **16**

E Standard Attention Dynamics **17**

F Supplementary Proofs **21**
F.1 Explicit Opinion Leader Dynamics 21
F.1.1 Reverse Wasserstein Gradient Flow 21
F.1.2 Proof of Corollary 2.1 22
F.1.3 Proof of Theorem 2.3 23
F.2 Implicit Opinion Leader Dynamics 25
F.2.1 Radius-graph decomposition 25
F.2.2 Graph Structure and Product Wasserstein Flow 25
F.2.3 Proof of Theorem 2.6 27

G Extended Experiment Results **30**
G.1 Implementation Details 30
G.1.1 Numerical Simulations 30
G.1.2 Frontier LLM Analysis 30
G.2 Cosine Silhouette Scores 30
G.3 Explicit Opinion Leader Dynamics under Conditional Initialization 31
G.4 Extended Analysis of Token Dynamics in Frontier LLMs 32
G.4.1 Results across Reduced Embedding Dimensions 32
G.4.2 Results across Dimensionality-Reduction Methods 32
G.4.3 Results on RULER Long-Context Variable Tracking 33

A Discussion and Future Work

Our analysis isolates two mechanisms through which restricted interactions support alignment within token groups while preserving separation between them. To obtain tractable convergence guarantees, we study fixed representatives in the explicit model and disconnected interaction components in the implicit model, together with suitable geometric conditions on initialization. These idealizations allow us to characterize distinct limiting directions and quantify convergence rates.

An important future extension is to incorporate *causal masking* into our analysis. Our current framework does not impose causal constraints, whereas all frontier LLMs examined in our experiments use causal attention. Causal masking makes token interactions directional and can break the symmetry underlying our gradient-flow formulation. We therefore plan to develop a framework for causal opinion leader dynamics, investigating how directed interactions affect the emergence, separation, and stability of token groups.

Despite these differences between theory and practice, the frontier sparse-attention LLMs studied here exhibit the qualitative behavior captured by our framework: their final-layer representations show distinct token groups and consistently stronger cluster separation than the dense-attention comparison model across four benchmarks. These observations support the relevance of the proposed mechanisms beyond the idealized setting, although they do not establish that trained models satisfy every theoretical assumption. Future work will examine which conditions are essential for this correspondence and extend the analysis to evolving representatives, time-dependent interaction patterns, and weak communication between groups.

B Related Work

Attention Dynamics. A growing line of theoretical work models self-attention as an interacting particle system and studies it through mean-field analysis (Geshkovski et al., 2023; Chen et al., 2025; Geshkovski et al., 2025; Rigollet, 2026). Within this framework, Bruno et al. (2025a) show that metastable clusters form and persist near uniform initialization as the number of tokens grows, while Bruno et al. (2025b) identify several regimes at moderate interaction strength, including one in which clusters merge slowly and sequentially. Complementary work considers restricted interaction patterns. Karagodin et al. (2024) analyze finite-token dynamics under causal masking, while Liu et al. (2026) introduce a bounded-confidence sparse attention and study its clustering behavior.

Sparse Architectures. Sparse attention reduces computation by allowing each query to interact with only a subset of keys. Existing methods use local windows or predefined sparse patterns (Child et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020), content-dependent hashing or routing (Kitaev et al., 2020; Roy et al., 2021; Lu et al., 2025), or representative tokens to guide retrieval (Mohtashami & Jaggi, 2023; Hu et al., 2025; Leng et al., 2026). Native Sparse Attention (NSA) combines token compression, blockwise selection, and sliding-window attention (Yuan et al., 2025). DeepSeek Sparse Attention (DSA) instead uses a lightweight learned indexer to select a query-dependent top-k subset of preceding tokens (Liu et al., 2025). Recent theoretical work also shows that combining local and global attention can be more expressive than using either alone (Li & Cotterell, 2026). Motivated by these architectures, we formulate two idealized models within *opinion leader dynamics* framework: one uses fixed representatives, while the other restricts interactions through a sparse graph.

C Extended Notation and Definitions

This section collects supplementary notation for the two models of *opinion leader dynamics*. Table 2 records the notation associated with *explicit opinion leader dynamics*, while Table 3 reports the notation used in *implicit opinion leader dynamics*. The listed notation is used throughout the model formulations, assumptions, convergence results, and proofs developed in the main text and appendix.

Table 2: Extended key notation for *explicit opinion leader dynamics*.

Notation	Definition
$\mathrm{Crit}(\Phi_\nu)$	$\{x \in \mathbb{S}^{d-1} : \nabla^\circ \Phi_\nu(x) = 0\}$
μ_t^a	$(\phi_t)_\# \mu_0^a, \qquad a = 1, \dots, m$
θ_{ab}	$d_{\mathbb{S}}(\ell_a, \ell_b), \qquad a \neq b$
$\gamma_{ab}(\alpha)$	$\cos\big(\theta_{ab} - \alpha\big), \qquad a \neq b$
$\mathbf{I}_a(\alpha)$	$\rho_a r_a \exp\big(\beta r_a \cos\alpha\big) \sin^2\alpha$ $- \sin\alpha \sum_{b \neq a} \rho_b r_b \exp\big(\beta r_b \gamma_{ab}(\alpha)\big) \sin\big(\theta_{ab} - \alpha\big), \qquad a = 1, \dots, m$
$\mathbf{H}_a(\alpha)$	$\rho_a r_a \exp\big(\beta r_a \cos\alpha\big)\big(\cos\alpha - \beta r_a \sin^2\alpha\big)$ $- \sum_{b \neq a} \rho_b r_b \exp\big(\beta r_b \gamma_{ab}(\alpha)\big)\big(1 + \beta r_b\big), \qquad a = 1, \dots, m$

Table 3: Extended key notation for *implicit opinion leader dynamics*.

Notation	Definition
$\mathbf{A}_k$	$\Big(A_{ij}^{(R)}\Big)_{i,j \in \mathcal{I}_k}$
$\mathbf{D}_k$	$\mathrm{diag}\Big(\Big(d_i^{(R)}\Big)_{i \in \mathcal{I}_k}\Big)$
λ_k	$\lambda_2\Big(\mathbf{I}_{n_k} - \mathbf{D}_k^{-1/2} \mathbf{A}_k \mathbf{D}_k^{-1/2}\Big)$
vol_k	$\sum_{i \in \mathcal{I}_k} d_i^{(R)}$
P_{ij}^k	$\frac{A_{ij}^{(R)}}{d_i^{(R)}}$
π_i^k	$\frac{d_i^{(R)}}{\mathrm{vol}_k}$
V_0^k	$\frac{1}{2} \iint_{\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}} \|x - y\|_2^2 \, d\widehat{\mu}_0^k(x) \, d\widehat{\mu}_0^k(y)$
χ_k	$\frac{\mathrm{vol}_k}{n_k \min_{i \in \mathcal{I}_k} d_i^{(R)}}$
κ_k	$\lambda_k \cos(2\alpha) e^{\beta \cos(2\alpha)}$
C_k	$\frac{2e^{\beta(1-\cos(2\alpha))}}{\lambda_k \cos(2\alpha)}$
$\kappa_{\min}$	$\min_{1 \leq k \leq K} \kappa_k$
C_G	$\Big(\sum_{k=1}^K m_k \chi_k C_k^2 V_0^k\Big)^{1/2}$

D Interpreting Sparse Attention through Opinion Leader Dynamics

A wide range of existing sparse attention architectures can be interpreted through the **Explicit** and **Implicit** models of our *opinion leader dynamics*. As summarized in Table 4, Landmark Attention (Mohtashami & Jaggi, 2023), Kimi Delta Attention (Team et al., 2026), and Hierarchical Sparse Attention (Hu et al., 2025) are **Explicit** Sparse Attention for their **representative-based mechanisms**; Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), DeepSeek Sparse Attention (V3.2) (Liu et al., 2025), and MiniMax Sparse Attention (Lai et al., 2026) are **Implicit** Sparse Attention for their **restricted interaction mechanisms**; NSA (Yuan et al., 2025) combines three branches: compressed block representations, top-k block selection, and local sliding-window attention, which together incorporate both **explicit representatives** and a **restricted interaction pattern**. DeepSeek-V4 Attention similarly combines the two mechanisms through its Compressed Sparse Attention (CSA) and Heavily Compressed Attention (HCA) modules. Both compress token groups into key-value entries that serve as **explicit representatives**. Their local sliding-window branches, together with query-dependent selection in CSA, define a **restricted interaction pattern** (Xu et al., 2026). Accordingly, we classify NSA and DeepSeek-V4 Attention as **Hybrid** Sparse Attention.

Table 4: Connections between existing sparse attention mechanisms and our *opinion leader dynamics* framework. For each method, we summarize its main design and identify its corresponding primary dynamical interpretation within our framework: **Explicit**, **Implicit**, and **Hybrid**.

Attention mechanism	Description	Interpretation
Landmark Attention (Mohtashami & Jaggi, 2023)	Partitions the sequence into fixed-size blocks, each followed by a landmark token that serves as an **explicit representative**. Query-landmark attention scores gate access to the associated blocks during training and guide their retrieval at inference.	**Explicit**
Kimi Delta Attention (Team et al., 2026)	Compresses past-token information into a gated recurrent state that serves as an **explicit representative** and is updated at each step to provide query-dependent context.	**Explicit**
Hierarchical Sparse Attention (Hu et al., 2025; Leng et al., 2026)	Encodes each fixed-length chunk as a retrieval representation that can be interpreted as an **explicit representative**. Each query scores these representations, attends within the top-k chunks, and combines their outputs with chunk-level weights.	**Explicit**
Longformer (Beltagy et al., 2020)	Defines a **sparse interaction pattern** that combines local sliding-window attention with full-sequence interactions through designated global tokens.	**Implicit**
BigBird (Zaheer et al., 2020)	Defines a **sparse interaction pattern** that combines local windows, random connections, and global tokens to capture long-range dependencies.	**Implicit**
DeepSeek Sparse Attention (V3.2) (Liu et al., 2025)	Restricts interactions for each query to the top-k preceding tokens selected by a lightweight learned indexer, yielding a query-dependent **sparse interaction pattern**.	**Implicit**
MiniMax Sparse Attention (Lai et al., 2026)	Restricts attention within each query-head group to the top-k key-value blocks while always retaining the local block, thereby defining a blockwise **sparse interaction pattern**.	**Implicit**
Native Sparse Attention (Yuan et al., 2025)	Combines compressed blocks that act as **explicit representatives** with a **sparse interaction pattern** defined by top-k block selection and local sliding-window attention.	**Hybrid**
DeepSeek-V4 Attention (Xu et al., 2026)	Combines Compressed Sparse Attention (CSA) and Heavily Compressed Attention (HCA). Compressed key-value entries in both modules serve as **explicit representatives**, while their sliding-window branches and query-dependent selection in CSA define a **sparse interaction pattern**.	**Hybrid**

E Standard Attention Dynamics

Prior work has modeled Standard Attention as an interacting particle system on the unit sphere (Chen et al., 2025; Geshkovski et al., 2025). We briefly review their derivations here for completeness. The analysis of our framework instead uses fixed-potential gradient analysis for *explicit opinion leader dynamics* and graph decomposition with spectral-gap estimates for *implicit opinion leader dynamics*.

Let $x_i(t) \in \mathbb{S}^{d-1}$ denote the representation of token i at time t. For $\beta > 0$, the tokens evolve according to

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\frac{1}{n} \sum_{j=1}^{n} e^{\beta \langle x_i(t), x_j(t) \rangle} x_j(t) \right). \tag{33}$$

The associated empirical measure $\mu_t = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i(t)}$ satisfies

$$\partial_t \mu_t + \operatorname{div}^{\circ} (\mu_t \mathcal{X}_{\mu_t,\beta}) = 0, \qquad \mathcal{X}_{\mu_t,\beta}(x) = \int_{\mathbb{S}^{d-1}} \mathbf{P}^{\perp}_x(y) e^{\beta \langle x,y \rangle} \, d\mu_t(y), \tag{34}$$

where $\operatorname{div}^{\circ}$ denotes the spherical divergence. More generally, Eq. 34 defines the mean-field dynamics for any probability measure on $\mathbb{S}^{d-1}$, with the associated interaction energy

$$\mathbf{E}_{\beta}[\mu] = \frac{1}{2\beta} \iint_{\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}} e^{\beta \langle x,y \rangle} \, d\mu(x) d\mu(y). \tag{35}$$

Let χ be a signed measure with zero total mass, and suppose that $\mu_\varepsilon = \mu + \varepsilon \chi$ remains a probability measure for sufficiently small $|\varepsilon|$. Differentiating Eq. 35 at $\varepsilon = 0$ and using the symmetry of the kernel gives

$$\frac{d}{d\varepsilon} \mathbf{E}_{\beta}[\mu_\varepsilon] \Big|_{\varepsilon=0} = \int_{\mathbb{S}^{d-1}} \left(\frac{1}{\beta} \int_{\mathbb{S}^{d-1}} e^{\beta \langle x,y \rangle} d\mu(y) \right) d\chi(x). \tag{36}$$

Therefore, the first variation of $\mathbf{E}_{\beta}$ is

$$\frac{\delta \mathbf{E}_{\beta}}{\delta \mu}(x) = \frac{1}{\beta} \int_{\mathbb{S}^{d-1}} e^{\beta \langle x,y \rangle} d\mu(y). \tag{37}$$

Taking the spherical gradient yields

$$\begin{aligned} \nabla^{\circ}_x \frac{\delta \mathbf{E}_{\beta}}{\delta \mu}(x) &= \int_{\mathbb{S}^{d-1}} \mathbf{P}^{\perp}_x(y) e^{\beta \langle x,y \rangle} d\mu(y) \\ &= \mathcal{X}_{\mu,\beta}(x). \end{aligned} \tag{38}$$

Hence, Eq. 34 defines a *reverse Wasserstein gradient flow* of $\mathbf{E}_{\beta}$:

$$\partial_t \mu_t + \operatorname{div}^{\circ} \left(\mu_t \nabla^{\circ} \frac{\delta \mathbf{E}_{\beta}}{\delta \mu_t} \right) = 0. \tag{39}$$

Define the squared Wasserstein slope by

$$I_{\beta}(\mu) = \int_{\mathbb{S}^{d-1}} \|\mathcal{X}_{\mu,\beta}(x)\|_2^2 d\mu(x). \tag{40}$$

Using Eq. 39 and integration by parts on the sphere, we obtain

$$\begin{aligned} \frac{d}{dt} \mathbf{E}_{\beta}[\mu_t] &= \int_{\mathbb{S}^{d-1}} \left\langle \nabla^{\circ}_x \frac{\delta \mathbf{E}_{\beta}}{\delta \mu_t}(x), \mathcal{X}_{\mu_t,\beta}(x) \right\rangle d\mu_t(x) \\ &= \int_{\mathbb{S}^{d-1}} \|\mathcal{X}_{\mu_t,\beta}(x)\|_2^2 d\mu_t(x) \\ &= I_{\beta}(\mu_t) \geq 0. \end{aligned} \tag{41}$$

Thus, $\mathbf{E}_{\beta}[\mu_t]$ is nondecreasing along the flow.

For any $x, y \in \mathbb{S}^{d-1}$, we have $\langle x, y\rangle \leq 1$. Therefore, every probability measure $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ satisfies

$$\begin{aligned}\mathbf{E}_\beta[\mu] &= \frac{1}{2\beta} \iint_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} e^{\beta\langle x,y\rangle} d\mu(x) d\mu(y) \\ &\leq \frac{1}{2\beta} \iint_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} e^{\beta} d\mu(x) d\mu(y) \\ &= \frac{e^\beta}{2\beta} = \mathbf{E}_\beta[\delta_{x_0}], \qquad x_0 \in \mathbb{S}^{d-1}. \end{aligned} \tag{42}$$

Equality holds only if $\langle x, y\rangle = 1$ for $\mu \otimes \mu$-almost every (x, y). Since x and y are unit vectors, this condition implies that μ is a Dirac measure. Hence,

$$\mathbf{E}_\beta^\star = \sup_{\mu\in\mathcal{P}(\mathbb{S}^{d-1})} \mathbf{E}_\beta[\mu] = \frac{e^\beta}{2\beta}, \tag{43}$$

and the global maximizers of $\mathbf{E}_\beta$ are precisely the Dirac measures.

We next recall two convergence results established by Chen et al. (2025, *Theorems 2.3 and 2.5*).

Theorem E.1 (Localized initialization). *Let $d \geq 2$, $\beta > 0$, $u \in \mathbb{S}^{d-1}$, and $\alpha \in [0, \pi/2)$. We define $S_\alpha^+(u) = \left\{x \in \mathbb{S}^{d-1} : \langle x, u\rangle \geq \cos\alpha\right\}$, which means the closed spherical cap of radius α centered at u. Assume that*

$$\operatorname{supp}(\mu_0) \subseteq S_\alpha^+(u), \qquad 10(1+\sqrt{\beta})\tan\alpha \leq 1. \tag{44}$$

Then $S_\alpha^+(u)$ is forward invariant, and there exists $x_\infty \in S_\alpha^+(u)$ such that, for every $t \geq 0$,

$$W_2(\mu_t, \delta_{x_\infty}) \leq 20 e^{-\beta} e^{-e^\beta t/20} I_\beta(\mu_0)^{1/2}, \qquad I_\beta(\mu) = \int_{\mathbb{S}^{d-1}} \|\mathcal{X}_{\mu,\beta}(x)\|_2^2 \, d\mu(x). \tag{45}$$

Moreover, any probability measure μ supported on $S_\alpha^+(u)$ satisfies the Polyak–Łojasiewicz inequality

$$\mathbf{E}_\beta^\star - \mathbf{E}_\beta[\mu] \leq 10 e^{-\beta} I_\beta(\mu), \qquad \mathbf{E}_\beta^\star = \frac{e^\beta}{2\beta}. \tag{46}$$

Proof. We follow the proof of *Theorem 2.3* in Chen et al. (2025), written in our notation. The required localization conditions are

$$\operatorname{supp}(\mu_0) \subseteq S_\alpha^+(u), \qquad 10(1+\sqrt{\beta})\tan\alpha \leq 1. \tag{47}$$

Let $\Phi_{0,t}$ denote the characteristic flow generated by $\mathcal{X}_{\mu_t,\beta}$, so that $\mu_t = (\Phi_{0,t})_\# \mu_0$. Define $m(t) = \min_{z\in\operatorname{supp}(\mu_0)} \langle \Phi_{0,t}(z), u\rangle$. At almost every t, the envelope formula gives

$$m'(t) = \min_{x\in A_t} \langle \mathcal{X}_{\mu_t,\beta}(x), u\rangle, \qquad A_t = \{x \in \operatorname{supp}(\mu_t) : \langle x, u\rangle = m(t)\}. \tag{48}$$

For $x \in A_t$ and $y \in \operatorname{supp}(\mu_t)$, the definition of $m(t)$ gives

$$\langle y, u\rangle \geq m(t) = \langle x, u\rangle. \tag{49}$$

Therefore, whenever $m(t) > 0$,

$$\begin{aligned}\langle \mathcal{X}_{\mu_t,\beta}(x), u\rangle &= \int_{\mathbb{S}^{d-1}} \left(\langle y, u\rangle - \langle x, y\rangle\langle x, u\rangle\right) e^{\beta\langle x,y\rangle} d\mu_t(y) \\ &\geq m(t) \int_{\mathbb{S}^{d-1}} \left(1 - \langle x, y\rangle\right) e^{\beta\langle x,y\rangle} d\mu_t(y) \\ &\geq 0. \end{aligned} \tag{50}$$

Since $m(0) \geq \cos\alpha > 0$ and $m'(t) \geq 0$ whenever $m(t) > 0$, absolute continuity yields

$$m(t) \geq m(0) \geq \cos\alpha, \qquad t \geq 0. \tag{51}$$

It follows that

$$\operatorname{supp}(\mu_t) \subseteq S_\alpha^+(u), \qquad t \geq 0. \tag{52}$$

This forward-invariance argument only requires $\alpha < \pi/2$. The stronger condition in Eq. 47 is used below to obtain the quantitative slope decay.

Set $I_t = I_\beta(\mu_t)$. Under Eq. 47, specializing *Lemma 4.1 and Remark 4.2* of Chen et al. (2025) to $A = \beta I_d$ and $\phi'(r) = e^r$ yields

$$\frac{d}{dt} I_t \le -\frac{e^\beta}{10} I_t. \tag{53}$$

Therefore,

$$I_t \le e^{-e^\beta t/10} I_0. \tag{54}$$

The metric derivative of the continuity equation satisfies $|\dot{\mu}_t|_{W_2} \le I_t^{1/2}$. Hence, for $0 \le s \le t$,

$$\begin{aligned} W_2(\mu_s, \mu_t) &\le \int_s^t I_r^{1/2}\, dr \\ &\le 20 e^{-\beta} e^{-e^\beta s/20} I_0^{1/2}. \end{aligned} \tag{55}$$

Thus, $\{\mu_t\}_{t\ge 0}$ is a Cauchy curve in W_2 and converges to some $\mu_\infty \in \mathcal{P}(\mathbb{S}^{d-1})$. Since $S_\alpha^+(u)$ is closed and $\mu_t\big(S_\alpha^+(u)\big) = 1$ for $t \ge 0$, the Portmanteau theorem gives

$$\operatorname{supp}(\mu_\infty) \subseteq S_\alpha^+(u). \tag{56}$$

Smoothness of the kernel on the compact sphere ensures that I_β is continuous with respect to W_2. Since $\mu_t \to \mu_\infty$ in W_2 and $I_\beta(\mu_t) \to 0$ by Eq. 54, we obtain

$$I_\beta(\mu_\infty) = \lim_{t\to\infty} I_\beta(\mu_t) = 0, \tag{57}$$

which shows that the field $\mathcal{X}_{\mu_\infty,\beta}$ vanishes μ_∞-almost everywhere. Continuity in x extends this identity to the support:

$$\mathcal{X}_{\mu_\infty,\beta}(x) = 0, \qquad x \in \operatorname{supp}(\mu_\infty). \tag{58}$$

Choose $x_0 \in \operatorname{supp}(\mu_\infty)$ minimizing $\langle x, u\rangle$. Taking the inner product of Eq. 58 with u gives

$$0 = \int_{\mathbb{S}^{d-1}} \left(\langle y, u\rangle - \langle x_0, y\rangle\langle x_0, u\rangle\right) e^{\beta\langle x_0, y\rangle} d\mu_\infty(y). \tag{59}$$

By the choice of x_0, the integrand is bounded below by

$$\langle x_0, u\rangle \left(1 - \langle x_0, y\rangle\right) e^{\beta\langle x_0, y\rangle} \ge 0. \tag{60}$$

Moreover, $\langle x_0, u\rangle \ge \cos\alpha > 0$. Hence, Eq. 59 implies $\langle x_0, y\rangle = 1$ for μ_∞-almost every y. Therefore, $\mu_\infty = \delta_{x_\infty}$ for some $x_\infty \in S_\alpha^+(u)$. Letting $t \to \infty$ in Eq. 55 yields

$$W_2(\mu_t, \delta_{x_\infty}) \le 20 e^{-\beta} e^{-e^\beta t/20} I_\beta(\mu_0)^{1/2}. \tag{61}$$

This proves Eq. 45.

The preceding argument applies to any probability measure supported on $S_\alpha^+(u)$. Start the flow from such a measure μ. By Eqs. 41 and 54,

$$\begin{aligned} \mathbf{E}_\beta^\star - \mathbf{E}_\beta[\mu] &= \lim_{T\to\infty} \left(\mathbf{E}_\beta[\mu_T] - \mathbf{E}_\beta[\mu]\right) \\ &= \int_0^\infty I_\beta(\mu_t)\, dt \\ &\le 10 e^{-\beta} I_\beta(\mu). \end{aligned} \tag{62}$$

Because $\mathbf{E}_\beta^\star = \frac{e^\beta}{2\beta}$, Eq. 62 proves Eq. 46 and completes the proof of Theorem E.1. □

Theorem E.1 establishes exponential convergence to a single consensus direction, with rate $e^\beta/20$ uniformly in the number of tokens. Theorem E.2 moves beyond this setting by replacing the localization condition with L^2 regularity and a nonzero initial mean.

Theorem E.2 (L^2 **mean-field initialization**). *Let $d \geq 2$, and let σ denote the normalized uniform measure on $\mathbb{S}^{d-1}$. Assume that*

$$d\mu_0 = f_0\, d\sigma, \qquad f_0 \in L^2(\mathbb{S}^{d-1}, \sigma), \qquad R_0 = \left\| \int_{\mathbb{S}^{d-1}} x\, d\mu_0(x) \right\|_2 > 0. \tag{63}$$

Then, for every $t > 0$, the measure μ_t admits a density $f_t \in L^2(\mathbb{S}^{d-1}, \sigma)$ with respect to σ. Moreover, there exist constants $\beta_0, C_0, T_0 > 0$, depending on μ_0, such that, for every $0 < \beta < \beta_0$, there exists $x_\infty \in \mathbb{S}^{d-1}$ satisfying

$$W_2(\mu_t, \delta_{x_\infty}) \leq C_0 e^{-t/100}, \qquad t > T_0. \tag{64}$$

Proof. We specialize the quantitative estimates of Chen et al. (2025) to Standard Attention dynamics. Take

$$A = I_d, \qquad \varphi'(r) = e^{\beta r}, \qquad r \in [-1, 1]. \tag{65}$$

The parametrization in Eq. 65 yields the same kernel $e^{\beta\langle x,y\rangle}$ as the localized choice $A = \beta I_d$ and $\varphi'(r) = e^r$, while making the dependence on the small parameter β explicit.

For this choice, the perturbation parameter in Chen et al. (2025) satisfies

$$\begin{aligned} \varepsilon_\beta &= 3\|e^{\beta(\cdot)} - 1\|_{C^1([-1,1])} \\ &\leq 3\left(e^\beta - 1 + \beta e^\beta\right) \longrightarrow 0 \qquad \text{as } \beta \to 0. \end{aligned} \tag{66}$$

Therefore, there exists $\beta_0 = \beta_0(\mu_0) > 0$ such that, for every $0 < \beta < \beta_0$,

$$\varepsilon_\beta \leq \min\left\{ \frac{1}{100}, c_{\mathrm{u}} R_0^6 \right\}. \tag{67}$$

These are the small-perturbation conditions required by *Theorems 3.5 and 3.8* of Chen et al. (2025).

Since the velocity field is smooth on the compact sphere, its characteristic flow is a diffeomorphism for every finite t. Hence, absolute continuity and L^2 regularity are preserved: if $d\mu_0 = f_0\, d\sigma$ with $f_0 \in L^2(\mathbb{S}^{d-1}, \sigma)$, then

$$d\mu_t = f_t\, d\sigma, \qquad f_t \in L^2(\mathbb{S}^{d-1}, \sigma), \qquad t > 0. \tag{68}$$

Define

$$M_t = \int_{\mathbb{S}^{d-1}} x\, d\mu_t(x), \qquad R_t = \|M_t\|_2, \qquad U_t = \frac{M_t}{R_t}. \tag{69}$$

After possibly reducing β_0, *Lemma 6.6* of Chen et al. (2025) yields a constant $\lambda \in (0, 1)$ such that

$$R_t \geq \lambda R_0 > 0, \qquad t \geq 0. \tag{70}$$

Thus, U_t is well defined for all $t \geq 0$.

Fix $\alpha = \pi/100$ and define the mass outside the moving cap by

$$p_t = \mu_t\left(\mathbb{S}^{d-1} \setminus S_\alpha^+(U_t)\right). \tag{71}$$

Theorem 3.5 of Chen et al. (2025) gives

$$\frac{d}{dt} I_\beta(\mu_t) \leq -I_\beta(\mu_t) + 100 p_t. \tag{72}$$

Moreover, *Theorem 3.8* of Chen et al. (2025) gives a time $T_{\mathrm{out}} > 0$ such that

$$p_t \leq \|f_0\|_{L^2(\mathbb{S}^{d-1},\sigma)} \exp\left(-\frac{d-1}{16}(t - T_{\mathrm{out}}) \right), \qquad t \geq T_{\mathrm{out}}. \tag{73}$$

Set $b = \frac{d-1}{16}$. Combining Eqs. 72 and 73 and applying Grönwall's inequality gives

$$\begin{aligned} I_\beta(\mu_t) \leq\ & e^{-(t-T_{\mathrm{out}})} I_\beta(\mu_{T_{\mathrm{out}}}) \\ & + 100\|f_0\|_{L^2(\mathbb{S}^{d-1},\sigma)} \int_{T_{\mathrm{out}}}^t e^{-(t-r)} e^{-b(r-T_{\mathrm{out}})}\, dr, \qquad t \geq T_{\mathrm{out}}. \end{aligned} \tag{74}$$

Since $d \geq 2$, we have $b = (d-1)/16 \geq 1/16$. We may therefore choose $a = 1/50 < \min\{1, b\}$. For $t \geq T_{\text{out}}$, the convolution term satisfies

$$\int_{T_{\text{out}}}^{t} e^{-(t-r)} e^{-b(r-T_{\text{out}})}\, dr \leq C_{a,b} e^{-a(t-T_{\text{out}})}. \tag{75}$$

Combining this estimate with the preceding bound and absorbing T_{out} into the constant, there exists $C_I > 0$ such that

$$I_\beta(\mu_t) \leq C_I e^{-t/50}, \qquad t \geq T_{\text{out}}. \tag{76}$$

Integrating the metric derivative, for all $t \geq s \geq T_{\text{out}}$, gives

$$\begin{aligned} W_2(\mu_s, \mu_t) &\leq \int_s^t I_\beta(\mu_r)^{1/2}\, dr \\ &\leq C_I^{1/2} \int_s^t e^{-r/100}\, dr \\ &\leq 100 C_I^{1/2} e^{-s/100}. \end{aligned} \tag{77}$$

Therefore, $\{\mu_t\}_{t\geq 0}$ is a Cauchy curve in W_2 and converges to some $\mu_\infty \in \mathcal{P}(\mathbb{S}^{d-1})$. Letting $t \to \infty$ in Eq. 77 yields

$$W_2(\mu_s, \mu_\infty) \leq 100 C_I^{1/2} e^{-s/100}, \qquad s \geq T_{\text{out}}. \tag{78}$$

Moreover, the continuity of I_β under W_2 convergence and Eq. 76 imply

$$I_\beta(\mu_\infty) = 0, \tag{79}$$

which means μ_∞ is stationary.

It remains to identify the stationary limit. By compactness of $\mathbb{S}^{d-1}$, there exists a sequence $t_k \to \infty$ and a direction $U_\infty \in \mathbb{S}^{d-1}$ such that

$$U_{t_k} \longrightarrow U_\infty. \tag{80}$$

Define $g(x) = (\cos\alpha - \langle x, U_\infty\rangle)_+$. For every k, Eq. 71 gives

$$\int_{\mathbb{S}^{d-1}} g(x)\, d\mu_{t_k}(x) \leq \|U_{t_k} - U_\infty\|_2 + 2p_{t_k}. \tag{81}$$

By Eq. 73, the right-hand side converges to zero. Passing to the limit yields

$$\operatorname{supp}(\mu_\infty) \subseteq S_\alpha^+(U_\infty). \tag{82}$$

Since μ_∞ is stationary, the minimizer argument used in Eq. 59 applies with $u = U_\infty$. Therefore,

$$\mu_\infty = \delta_{x_\infty} \tag{83}$$

for some $x_\infty \in S_\alpha^+(U_\infty)$.

Combining Eq. 78 with $\mu_\infty = \delta_{x_\infty}$, and setting $C_0 = 100 C_I^{1/2}$, $T_0 = T_{\text{out}}$, we obtain

$$W_2(\mu_t, \delta_{x_\infty}) \leq C_0 e^{-t/100}, \qquad t \geq T_0. \tag{84}$$

This proves Eq. 64 and completes the proof of Theorem E.2. □

Theorem E.2 guarantees exponential convergence after T_0 with decay $e^{-t/100}$. The restriction to small β is essential, as large β can lead to non-consensus limits (Chen et al., 2025).

F Supplementary Proofs

F.1 Explicit Opinion Leader Dynamics

F.1.1 Reverse Wasserstein Gradient Flow

Let χ be a signed measure with zero total mass and set $\mu_\varepsilon = \mu + \varepsilon\chi$. Since ν is fixed, differentiating Eq. 10 at $\varepsilon = 0$ gives

$$\frac{d}{d\varepsilon} \mathbf{E}_\beta[\mu_\varepsilon|\nu]\Big|_{\varepsilon=0} = \int_{\mathbb{S}^{d-1}} \left(\frac{1}{\beta} \int_{\mathbb{R}^d} e^{\beta\langle x,z\rangle} d\nu(z) \right) d\chi(x). \tag{85}$$

Hence,

$$\frac{\delta \mathbf{E}_\beta[\mu|\nu]}{\delta \mu}(x) = \frac{1}{\beta}\int_{\mathbb{R}^d} e^{\beta\langle x,z\rangle} d\nu(z) = \Phi_\nu(x). \tag{86}$$

Taking the spherical gradient yields

$$\begin{aligned}\nabla_x^\circ \frac{\delta \mathbf{E}_\beta[\mu|\nu]}{\delta \mu}(x) &= \int_{\mathbb{R}^d} \mathbf{P}_x^\perp(z) e^{\beta\langle x,z\rangle} d\nu(z) \\ &= \mathcal{X}_{\nu,\beta}(x).\end{aligned} \tag{87}$$

Therefore, Eq. 9 is a *reverse Wasserstein gradient flow* of $\mathbf{E}_\beta[\cdot|\nu]$:

$$\partial_t \mu_t + \mathrm{div}^\circ\left(\mu_t \nabla_x^\circ \frac{\delta \mathbf{E}_\beta[\mu_t|\nu]}{\delta \mu_t}\right) = 0. \tag{88}$$

Define the corresponding squared Wasserstein slope by

$$\mathcal{I}_\beta[\mu|\nu] = \int_{\mathbb{S}^{d-1}} \|\mathcal{X}_{\nu,\beta}(x)\|_2^2 d\mu(x). \tag{89}$$

Using Eq. 88 and integration by parts on the sphere, we obtain

$$\begin{aligned}\frac{d}{dt}\mathbf{E}_\beta[\mu_t|\nu] &= \int_{\mathbb{S}^{d-1}} \left\langle \nabla_x^\circ \frac{\delta \mathbf{E}_\beta[\mu_t|\nu]}{\delta \mu_t}(x), \mathcal{X}_{\nu,\beta}(x)\right\rangle d\mu_t(x) \\ &= \int_{\mathbb{S}^{d-1}} \|\mathcal{X}_{\nu,\beta}(x)\|_2^2 d\mu_t(x) \\ &= \mathcal{I}_\beta[\mu_t|\nu] \geq 0.\end{aligned} \tag{90}$$

Thus, the conditional energy is nondecreasing along the flow. Its derivative vanishes at time t if and only if $\mathcal{X}_{\nu,\beta} = 0$ holds μ_t-almost everywhere.

F.1.2 Proof of Corollary 2.1

Proof. Fix $x \in \mathbb{S}^{d-1}$ and write $x_t = \phi_t(x)$. Since $\dot{x}_t = \nabla^\circ \Phi_\nu(x_t)$,

$$\frac{d}{dt}\Phi_\nu(x_t) = \|\nabla^\circ \Phi_\nu(x_t)\|_2^2 \geq 0. \tag{91}$$

Because $\mathbb{S}^{d-1}$ is compact, $\Phi_\nu(x_t)$ is bounded above and therefore converges. Integrating Eq. 91 gives

$$\int_0^\infty \|\nabla^\circ \Phi_\nu(x_t)\|_2^2 \, dt < \infty. \tag{92}$$

Since Φ_ν is real analytic on $\mathbb{S}^{d-1}$, the Łojasiewicz gradient inequality ensures that every trajectory converges to a single critical point (Absil et al., 2005). Consequently, $\phi_\infty(x) = \lim_{t\to\infty}\phi_t(x) \in \mathrm{Crit}(\Phi_\nu)$ is well defined for every $x \in \mathbb{S}^{d-1}$. Since each ϕ_t is continuous, the pointwise limit ϕ_∞ is Borel measurable.

For each group, the measure $(\phi_t, \phi_\infty)_\# \mu_0^a$ couples $\mu_t^a = (\phi_t)_\# \mu_0^a$ and $\mu_\infty^a = (\phi_\infty)_\# \mu_0^a$. Thus,

$$W_2^2(\mu_t^a, \mu_\infty^a) \leq \int_{\mathbb{S}^{d-1}} d_{\mathbb{S}}\left(\phi_t(x), \phi_\infty(x)\right)^2 \, d\mu_0^a(x). \tag{93}$$

The integrand converges pointwise to zero and is bounded by π^2. Dominated convergence gives $W_2(\mu_t^a, \mu_\infty^a) \to 0$ for every a.

Combining these groupwise couplings yields

$$W_2^2(\mu_t, \mu_\infty) \leq \sum_{a=1}^m \omega_a W_2^2(\mu_t^a, \mu_\infty^a) \longrightarrow 0. \tag{94}$$

Since ϕ_∞ takes values in $\mathrm{Crit}(\Phi_\nu)$, all limiting measures are supported on the critical set. □

F.1.3 Proof of Theorem 2.3

Proof. For $a \neq b$, the triangle inequality gives

$$\theta_{ab} = d_{\mathbb{S}}(\ell_a, \ell_b) \geq d_{\mathbb{S}}(c_a, c_b) - d_{\mathbb{S}}(\ell_a, c_a) - d_{\mathbb{S}}(\ell_b, c_b) \geq \Delta - 2\eta > 2\alpha. \tag{95}$$

Thus, the caps $\mathcal{S}_\alpha^+(\ell_a)$ are pairwise disjoint. Moreover, for $x \in \operatorname{supp}(\mu_0^a)$,

$$d_{\mathbb{S}}(x, \ell_a) \leq d_{\mathbb{S}}(x, c_a) + d_{\mathbb{S}}(c_a, \ell_a) \leq \delta + \eta < \alpha. \tag{96}$$

Hence,

$$\operatorname{supp}(\mu_0^a) \subseteq \mathcal{S}_{\delta+\eta}^+(\ell_a) \Subset \operatorname{int} \mathcal{S}_\alpha^+(\ell_a). \tag{97}$$

For later use, if $x \in \mathcal{S}_\alpha^+(\ell_a)$ and $b \neq a$, then

$$d_{\mathbb{S}}(x, \ell_b) \geq \theta_{ab} - d_{\mathbb{S}}(x, \ell_a) \geq \theta_{ab} - \alpha. \tag{98}$$

Since $0 < \theta_{ab} - \alpha < \pi$ and the cosine is decreasing on $[0, \pi]$,

$$\langle x, \ell_b \rangle \leq \cos(\theta_{ab} - \alpha) = \gamma_{ab}(\alpha). \tag{99}$$

Fix $x \in \partial\mathcal{S}_\alpha^+(\ell_a)$, so that $\langle x, \ell_a \rangle = \cos\alpha$. The derivative of $h_a(x) = \langle x, \ell_a \rangle$ along the vector field is

$$\begin{aligned} Dh_a(x)[\mathcal{X}_{\nu,\beta}(x)] &= \langle \mathcal{X}_{\nu,\beta}(x), \ell_a \rangle \\ &= \sum_{b=1}^m \rho_b r_b e^{\beta r_b \langle x, \ell_b \rangle} \left\langle \mathbf{P}_x^\perp \ell_b, \mathbf{P}_x^\perp \ell_a \right\rangle. \end{aligned} \tag{100}$$

The term with $b = a$ is

$$\rho_a r_a e^{\beta r_a \cos\alpha} \sin^2\alpha. \tag{101}$$

For $b \neq a$, Eq. 99 gives

$$\begin{aligned} \left\langle \mathbf{P}_x^\perp \ell_b, \mathbf{P}_x^\perp \ell_a \right\rangle &= \cos\theta_{ab} - \langle x, \ell_b \rangle \cos\alpha \\ &\geq \cos\theta_{ab} - \cos\alpha\cos(\theta_{ab} - \alpha) \\ &= -\sin\alpha\sin(\theta_{ab} - \alpha). \end{aligned} \tag{102}$$

If the projected inner product is nonnegative, its contribution to Eq. 100 is nonnegative. If it is negative, Eqs. 99 and 102 imply

$$\begin{aligned} &\rho_b r_b e^{\beta r_b \langle x, \ell_b \rangle} \left\langle \mathbf{P}_x^\perp \ell_b, \mathbf{P}_x^\perp \ell_a \right\rangle \\ &\quad \geq -\rho_b r_b e^{\beta r_b \gamma_{ab}(\alpha)} \sin\alpha \sin(\theta_{ab} - \alpha). \end{aligned} \tag{103}$$

Summing over b yields

$$Dh_a(x)[\mathcal{X}_{\nu,\beta}(x)] \geq \mathbf{I}_a(\alpha) > 0. \tag{104}$$

Thus, the vector field points strictly into the cap along its boundary, so $\mathcal{S}_\alpha^+(\ell_a)$ is forward invariant.

Write

$$\Phi_\nu = \sum_{b=1}^m \Phi_b, \qquad \Phi_b(x) = \frac{\rho_b}{\beta} e^{\beta r_b \langle x, \ell_b \rangle}. \tag{105}$$

For a unit vector $v \in T_x\mathbb{S}^{d-1}$,

$$\nabla_{\mathbb{S}}^2 \Phi_b(x)[v, v] = \rho_b r_b e^{\beta r_b \langle x, \ell_b \rangle} \left(\beta r_b \langle v, \ell_b \rangle^2 - \langle x, \ell_b \rangle \right). \tag{106}$$

For $x \in \mathcal{S}_\alpha^+(\ell_a)$,

$$\langle x, \ell_a \rangle \geq \cos\alpha, \qquad |\langle v, \ell_a \rangle| \leq \|\mathbf{P}_x^\perp \ell_a\|_2 \leq \sin\alpha. \tag{107}$$

Since $\mathbf{H}_a(\alpha) > 0$, we have $\cos\alpha - \beta r_a \sin^2\alpha > 0$. Therefore,

$$\nabla_{\mathbb{S}}^2 \Phi_a(x)[v, v] \leq -\rho_a r_a e^{\beta r_a \cos\alpha} \left(\cos\alpha - \beta r_a \sin^2\alpha \right). \tag{108}$$

For $b \neq a$, Eq. 99 and $|\langle x, \ell_b \rangle| \leq 1$ give

$$\nabla_{\mathbb{S}}^2 \Phi_b(x)[v, v] \leq \rho_b r_b e^{\beta r_b \gamma_{ab}(\alpha)} (1 + \beta r_b). \tag{109}$$

Indeed, this bound is immediate when the parenthesis in Eq. 106 is nonpositive; otherwise, the exponential is bounded using Eq. 99. Summing Eqs. 108 and 109 yields

$$\nabla^2_{\mathbb{S}}\Phi_\nu(x)[v,v] \le -\mathbf{H}_a(\alpha) < 0, \qquad x \in \mathcal{S}^+_\alpha(\ell_a). \tag{110}$$

Because $\alpha < \pi/2$, the cap $\mathcal{S}^+_\alpha(\ell_a)$ is geodesically convex. Compactness ensures that Φ_ν attains a maximum on the cap. At a boundary point, Eq. 104 makes $\mathcal{X}_{\nu,\beta}$ a feasible ascent direction, and

$$D\Phi_\nu[\mathcal{X}_{\nu,\beta}] = \|\mathcal{X}_{\nu,\beta}\|_2^2 > 0. \tag{111}$$

Thus, the maximum lies in the interior and is a critical point. The uniform concavity in Eq. 110 makes this point unique; denote it by u_a. The same bound shows that u_a is a nondegenerate local maximizer.

We next derive the contraction rate. Fix $x \in \mathcal{S}^+_\alpha(\ell_a)$ and let $\gamma : [0,1] \to \mathcal{S}^+_\alpha(\ell_a)$ be the constant-speed minimizing geodesic from x to u_a. Set $g(s) = \Phi_\nu(\gamma(s))$. Since $\|\dot\gamma(s)\|_2 = d_{\mathbb{S}}(x,u_a)$, Eq. 110 gives

$$g''(s) \le -\mathbf{H}_a(\alpha) d_{\mathbb{S}}(x,u_a)^2. \tag{112}$$

Moreover, $g'(1) = 0$ because $\nabla^\circ \Phi_\nu(u_a) = 0$. Since $\mathrm{Log}_x(u_a) = \dot\gamma(0)$, integration over $s \in [0,1]$ gives

$$\langle \mathcal{X}_{\nu,\beta}(x), \mathrm{Log}_x(u_a)\rangle = g'(0) \ge \mathbf{H}_a(\alpha) d_{\mathbb{S}}(x,u_a)^2. \tag{113}$$

Let $x_t = \varphi_t(x)$. Forward invariance and the first variation of the squared distance yield

$$\begin{aligned}\frac{1}{2}\frac{d}{dt} d_{\mathbb{S}}(x_t,u_a)^2 &= -\left\langle \mathcal{X}_{\nu,\beta}(x_t), \mathrm{Log}_{x_t}(u_a)\right\rangle \\ &\le -\mathbf{H}_a(\alpha) d_{\mathbb{S}}(x_t,u_a)^2.\end{aligned} \tag{114}$$

Grönwall's inequality therefore gives

$$d_{\mathbb{S}}(\varphi_t(x),u_a) \le e^{-\mathbf{H}_a(\alpha)t} d_{\mathbb{S}}(x,u_a), \qquad t \ge 0. \tag{115}$$

Thus, every point in $\mathcal{S}^+_\alpha(\ell_a)$ converges to u_a, and this cap is contained in the attraction basin B_a of u_a.

Since $u_a \in \mathcal{S}^+_\alpha(\ell_a)$ and $d_{\mathbb{S}}(\ell_a, c_a) \le \eta$,

$$d_{\mathbb{S}}(u_a,\ell_a) \le \alpha, \qquad d_{\mathbb{S}}(u_a,c_a) \le \alpha + \eta. \tag{116}$$

Hence, for $a \ne b$,

$$\begin{aligned} d_{\mathbb{S}}(u_a,u_b) &\ge d_{\mathbb{S}}(c_a,c_b) - d_{\mathbb{S}}(u_a,c_a) - d_{\mathbb{S}}(u_b,c_b) \\ &\ge \Delta - 2\eta - 2\alpha > 0. \end{aligned} \tag{117}$$

Thus, the attractors are pairwise distinct. The boundary and strict concavity arguments above also show that u_a is the only critical point in $\mathcal{S}^+_\alpha(\ell_a)$. Since the caps are disjoint, their union contains exactly the m critical points $u_1,\dots,u_m$.

For each a, forward invariance gives

$$\mathrm{supp}(\mu_t^a) \subseteq \mathcal{S}^+_\alpha(\ell_a), \qquad t \ge 0. \tag{118}$$

Because the second marginal is the Dirac measure δ_{u_a}, the pointwise contraction estimate above yields

$$\begin{aligned} W_2^2(\mu_t^a, \delta_{u_a}) &= \int_{\mathbb{S}^{d-1}} d_{\mathbb{S}}(\varphi_t(x),u_a)^2 d\mu_0^a(x) \\ &\le e^{-2\mathbf{H}_a(\alpha)t} \int_{\mathbb{S}^{d-1}} d_{\mathbb{S}}(x,u_a)^2 d\mu_0^a(x) \\ &= D_{0,a}^2 e^{-2\mathbf{H}_a(\alpha)t}. \end{aligned} \tag{119}$$

This proves the groupwise estimate in Eq. 19.

By linearity of the pushforward,

$$\mu_t = \sum_{a=1}^m \omega_a \mu_t^a. \tag{120}$$

Combining the groupwise couplings gives

$$W_2^2(\mu_t, \mu_\infty) \le \sum_{a=1}^{m} \omega_a W_2^2(\mu_t^a, \delta_{u_a}) \le \sum_{a=1}^{m} \omega_a D_{0,a}^2 e^{-2\mathbf{H}_a(\alpha)t} \le D_0^2 e^{-2\mathbf{H}_{\min}(\alpha)t}. \tag{121}$$

Taking square roots proves the first inequality in Eq. 21. Moreover, for $x \in \mathrm{supp}(\mu_0^a)$,

$$d_{\mathbb{S}}(x, u_a) \le d_{\mathbb{S}}(x, c_a) + d_{\mathbb{S}}(c_a, \ell_a) + d_{\mathbb{S}}(\ell_a, u_a) \le \delta + \eta + \alpha. \tag{122}$$

Therefore, $D_0 \le \alpha + \delta + \eta$. Finally, since every $\omega_a > 0$ and the points $u_1, \dots, u_m$ are pairwise distinct,

$$\mathrm{supp}(\mu_\infty) = \{u_1, \dots, u_m\}, \qquad |\mathrm{supp}(\mu_\infty)| = m, \tag{123}$$

which completes the proof. □

F.2 Implicit Opinion Leader Dynamics

F.2.1 Radius-graph decomposition

We first verify the component structure used in Sec. 2.2. Suppose, for contradiction, that G_R is connected. Then it contains a spanning tree $T_R \in \mathbb{T}_n$. Since every edge of T_R also belongs to G_R, its length is at most R. Hence,

$$R_{\mathrm{conn}} \le \max_{\{i,j\} \in \mathcal{E}(T_R)} d_{\mathcal{Z}}(p_i, p_j) \le R, \tag{124}$$

which contradicts $R < R_{\mathrm{conn}}$. Therefore, G_R is disconnected, and its reachability classes satisfy $K \ge 2$.

For any $i \in \{1, \dots, n\}$, the definition of R_{loc} ensures that there exists $j \ne i$ such that

$$d_{\mathcal{Z}}(p_i, p_j) \le R_{\mathrm{loc}} \le R. \tag{125}$$

Thus, every node has at least one distinct neighbor, so each connected component contains at least two nodes.

Finally, let $i \in \mathcal{I}_k$ and $j \in \mathcal{I}_\ell$ with $k \ne \ell$. Then $d_{\mathcal{Z}}(p_i, p_j) > R$; otherwise, $\{i, j\}$ would be an edge of G_R, implying $i \sim_R j$ and hence $k = \ell$. Therefore, $\mathcal{N}_i^{(R)} \subseteq \mathcal{I}_k$, $i \in \mathcal{I}_k$. After a permutation of the nodes, the adjacency matrix is block diagonal with blocks corresponding to $\{G_R[\mathcal{I}_k]\}_{k=1}^K$. Consequently, the dynamics decompose into K independent graph-coupled systems, one on each connected component.

F.2.2 Graph Structure and Product Wasserstein Flow

Lemma F.1 (Well-posedness of the graph-coupled flow). *For every initial tuple*

$$\boldsymbol{\nu}_0^k = (\nu_{i,0}^k)_{i \in \mathcal{I}_k} \in \mathcal{P}_2(\mathbb{S}^{d-1})^{\mathcal{I}_k}, \tag{126}$$

Eq. 26 admits a unique global characteristic solution

$$\boldsymbol{\nu}^k \in C\left([0, \infty); \mathcal{P}_2(\mathbb{S}^{d-1})^{\mathcal{I}_k}\right). \tag{127}$$

This solution is locally absolutely continuous with respect to $\mathcal{W}_{2,\pi^k}$, and the energy and moment identities used below hold along the flow.

Proof. Define

$$\mathcal{K}_\beta(x, y) = \mathbf{P}_x^\perp(y) e^{\beta\langle x, y\rangle}. \tag{128}$$

The kernel $\mathcal{K}_\beta$ is smooth on the compact set $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, and is therefore uniformly bounded and Lipschitz in both variables. It follows that there exist constants $M_\beta, L_\beta > 0$ such that

$$\sup_{x \in \mathbb{S}^{d-1}} \left\| \mathcal{X}_{i,\boldsymbol{\nu}^k,\beta}(x) \right\|_2 \le M_\beta, \qquad \left\| \mathcal{X}_{i,\boldsymbol{\nu}^k,\beta}(x) - \mathcal{X}_{i,\boldsymbol{\nu}^k,\beta}(x') \right\|_2 \le L_\beta d_{\mathbb{S}}(x, x'). \tag{129}$$

Moreover, for two measure tuples $\boldsymbol{\nu}^k$ and $\boldsymbol{\eta}^k$,

$$\sup_{x \in \mathbb{S}^{d-1}} \left\| \mathcal{X}_{i,\boldsymbol{\nu}^k,\beta}(x) - \mathcal{X}_{i,\boldsymbol{\eta}^k,\beta}(x) \right\|_2 \le L_\beta \sum_{j \in \mathcal{I}_k} \mathsf{P}^k_{ij} W_2(\nu^k_j, \eta^k_j). \tag{130}$$

For $T > 0$, consider the complete metric space

$$\mathcal{C}_T = \{ \boldsymbol{\zeta}^k \in C\left([0,T]; \mathcal{P}_2(\mathbb{S}^{d-1})^{\mathcal{I}_k}\right) : \boldsymbol{\zeta}^k_0 = \boldsymbol{\nu}^k_0 \}, \tag{131}$$

equipped with

$$d_T\left(\boldsymbol{\zeta}^k, \boldsymbol{\eta}^k\right) = \sup_{0 \le t \le T} \mathcal{W}_{2,\pi^k}\left(\boldsymbol{\zeta}^k_t, \boldsymbol{\eta}^k_t\right). \tag{132}$$

For $\boldsymbol{\zeta}^k \in \mathcal{C}_T$, let $\Phi^{\boldsymbol{\zeta}^k}_{i,t}$ denote the characteristic flow generated by $\mathcal{X}_{i,\boldsymbol{\zeta}^k_t,\beta}$, and define

$$\left(\mathcal{T}\boldsymbol{\zeta}^k\right)_{i,t} = \left(\Phi^{\boldsymbol{\zeta}^k}_{i,t}\right)_{\#} \nu^k_{i,0}. \tag{133}$$

Thus, $\mathcal{T}$ maps $\mathcal{C}_T$ into itself. The estimates above, Jensen's inequality, the stationarity identity

$$\sum_{i \in \mathcal{I}_k} \pi^k_i \mathsf{P}^k_{ij} = \pi^k_j, \tag{134}$$

and Grönwall's inequality give

$$d_T\left(\mathcal{T}\boldsymbol{\zeta}^k, \mathcal{T}\boldsymbol{\eta}^k\right) \le L_\beta T e^{L_\beta T} d_T\left(\boldsymbol{\zeta}^k, \boldsymbol{\eta}^k\right). \tag{135}$$

For sufficiently small $T > 0$, the map $\mathcal{T}$ is therefore a contraction. The Banach fixed-point theorem gives local existence and uniqueness. Since M_β and L_β are uniform in time, the same argument can be repeated on consecutive time intervals, yielding a unique global solution.

The characteristic representation also gives

$$\left\| \dot{\boldsymbol{\nu}}^k_t \right\|^2_{\mathcal{W}_{2,\pi^k}} \le \sum_{i \in \mathcal{I}_k} \pi^k_i \int_{\mathbb{S}^{d-1}} \left| \mathcal{X}_{i,\boldsymbol{\nu}^k_t,\beta}(x) \right|^2_2 d\nu^k_{i,t}(x) \le M^2_\beta. \tag{136}$$

Hence, the solution is locally absolutely continuous with respect to $\mathcal{W}_{2,\pi^k}$. Testing the continuity equations against smooth functions and using the smoothness of $\mathcal{K}_\beta$ then yields the energy and moment identities used below. □

Fix a component $\mathcal{I}_k$. Define the reversible edge weights by

$$q^k_{ij} = \pi^k_i \mathsf{P}^k_{ij} = \frac{A^{(R)}_{ij}}{\mathrm{vol}_k} = \pi^k_j \mathsf{P}^k_{ji} = q^k_{ji}. \tag{137}$$

Thus, P^k is reversible with respect to π^k. For $\mathbf{z} = (z_i)_{i \in \mathcal{I}_k}$, let $\overline{z}_{\pi^k} = \sum_{i \in \mathcal{I}_k} \pi^k_i z_i$ and define the graph Dirichlet form

$$\mathcal{D}_k(\mathbf{z}) = \frac{1}{2} \sum_{i,j \in \mathcal{I}_k} \pi^k_i \mathsf{P}^k_{ij} \|z_i - z_j\|^2_2. \tag{138}$$

Since the normalized Laplacian is similar to $\mathbf{I}_{n_k} - \mathsf{P}_k$, the spectral gap λ_k yields the Poincaré inequality

$$\lambda_k \sum_{i \in \mathcal{I}_k} \pi^k_i \|z_i - \overline{z}_{\pi^k}\|^2_2 \le \mathcal{D}_k(\mathbf{z}) \le 2 \sum_{i \in \mathcal{I}_k} \pi^k_i \|z_i - \overline{z}_{\pi^k}\|^2_2. \tag{139}$$

The upper bound follows because the spectrum of the normalized Laplacian is contained in $[0, 2]$. The vector-valued inequality follows by applying the scalar inequality coordinatewise.

Because every node has a self-loop, $\mathsf{P}^k_{ii} = 1/d^{(R)}_i$. Moreover, $d^{(R)}_i \le n_k$, and hence, $\mathrm{tr}(\mathsf{P}_k) = \sum_{i \in \mathcal{I}_k} \frac{1}{d^{(R)}_i} \ge 1$.

Let $1 = \theta_1^k \geq \theta_2^k \geq \cdots \geq \theta_{n_k}^k$ be the eigenvalues of P_k. Since $\sum_{r=2}^{n_k} \theta_r^k \geq 0$, we have $\theta_2^k \geq 0$. Therefore,

$$0 < \lambda_k = 1 - \theta_2^k \leq 1. \tag{140}$$

The graph interaction energy is

$$\mathcal{E}_{\beta,k}^G \left[\boldsymbol{\nu}^k\right] = \frac{1}{2\beta} \sum_{i,j \in \mathcal{I}_k} q_{ij}^k \iint_{\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}} e^{\beta \langle x,y \rangle} \, d\nu_i^k(x) d\nu_j^k(y). \tag{141}$$

We next verify the product Wasserstein gradient structure. By symmetry of q_{ij}^k, the first variation of $\mathcal{E}_{\beta,k}^G$ with respect to the i-th species is

$$\frac{\delta \mathcal{E}_{\beta,k}^G}{\delta \nu_i^k}(x) = \frac{1}{\beta} \sum_{j \in \mathcal{I}_k} q_{ij}^k \int_{\mathbb{S}^{d-1}} e^{\beta \langle x,y \rangle} \, d\nu_j^k(y). \tag{142}$$

Since $q_{ij}^k = \pi_i^k \mathsf{P}_{ij}^k$, taking the spherical gradient gives

$$\begin{aligned} \nabla_x^\circ \left(\frac{1}{\pi_i^k} \frac{\delta \mathcal{E}_{\beta,k}^G}{\delta \nu_i^k}(x) \right) &= \sum_{j \in \mathcal{I}_k} \mathsf{P}_{ij}^k \int_{\mathbb{S}^{d-1}} \mathbf{P}_x^\perp(y) e^{\beta \langle x,y \rangle} \, d\nu_j^k(y) \\ &= \mathcal{X}_{i,\boldsymbol{\nu}^k,\beta}(x). \end{aligned} \tag{143}$$

To verify this identification, let $\xi_i \in L^2(\nu_i^k; T\mathbb{S}^{d-1})$ and consider the perturbation $\nu_{i,\varepsilon}^k = (\exp_x(\varepsilon \xi_i(x)))_{\#} \nu_i^k$. Then

$$\frac{d}{d\varepsilon} \mathcal{E}_{\beta,k}^G \left[\boldsymbol{\nu}_\varepsilon^k\right] \Big|_{\varepsilon=0} = \sum_{i \in \mathcal{I}_k} \pi_i^k \int_{\mathbb{S}^{d-1}} \left\langle \mathcal{X}_{i,\boldsymbol{\nu}^k,\beta}(x), \xi_i(x) \right\rangle \, d\nu_i^k(x). \tag{144}$$

Hence, $\left(\mathcal{X}_{i,\boldsymbol{\nu}^k,\beta}\right)_{i \in \mathcal{I}_k}$ is the *Wasserstein gradient* of $\mathcal{E}_{\beta,k}^G$ under $\mathcal{W}_{2,\pi^k}$.

Define the squared Wasserstein slope by

$$\mathcal{I}_{\beta,k}^G \left[\boldsymbol{\nu}^k\right] = \sum_{i \in \mathcal{I}_k} \pi_i^k \int_{\mathbb{S}^{d-1}} \|\mathcal{X}_{i,\boldsymbol{\nu}^k,\beta}(x)\|_2^2 \, d\nu_i^k(x). \tag{145}$$

Using the continuity equation in Eq. 26 and spherical integration by parts, we obtain

$$\begin{aligned} \frac{d}{dt} \mathcal{E}_{\beta,k}^G \left[\boldsymbol{\nu}_t^k\right] &= \sum_{i \in \mathcal{I}_k} \pi_i^k \int_{\mathbb{S}^{d-1}} \|\mathcal{X}_{i,\boldsymbol{\nu}_t^k,\beta}(x)\|_2^2 \, d\nu_{i,t}^k(x) \\ &= \mathcal{I}_{\beta,k}^G \left[\boldsymbol{\nu}_t^k\right] \geq 0. \end{aligned} \tag{146}$$

Thus, $\boldsymbol{\nu}_t^k$ is a *reverse Wasserstein gradient flow* of $\mathcal{E}_{\beta,k}^G$ in the weighted product space.

F.2.3 Proof of Theorem 2.6

Proof. By Assumption 2.4 and the radius-graph decomposition in Appendix F.2.1, the sets $\{\mathcal{I}_k\}_{k=1}^K$ are the nontrivial connected components of G_R. Assumption 2.5 further provides pairwise disjoint initial caps. Fix a component $\mathcal{I}_k$ and set $c_\alpha = \cos(2\alpha) > 0$. We first prove forward invariance. The case $\alpha = 0$ is immediate, so suppose that $\alpha > 0$. Let $x \in \partial S_\alpha^+(u_k)$ and $y \in S_\alpha^+(u_k)$. Write

$$x = \cos\alpha \, u_k + \sin\alpha \, v, \qquad y = \cos\theta \, u_k + \sin\theta \, w, \tag{147}$$

where $\theta \leq \alpha$ and $v, w \in u_k^\perp$ are unit vectors. Then

$$\begin{aligned} \left\langle \mathbf{P}_x^\perp(y), u_k \right\rangle &= \langle y, u_k \rangle - \langle x, y \rangle \langle x, u_k \rangle \\ &= \sin\alpha \left(\sin\alpha \cos\theta - \cos\alpha \sin\theta \langle v, w \rangle \right) \\ &\geq \sin\alpha \sin(\alpha - \theta) \geq 0. \end{aligned} \tag{148}$$

Therefore,

$$\left\langle \mathcal{X}_{i,\boldsymbol{\nu}_t^k,\beta}(x), u_k \right\rangle \geq 0 \tag{149}$$

for every boundary point x whenever all nodewise measures are supported in the cap. The characteristic flow associated with Eq. 26 consequently preserves the cap:

$$\operatorname{supp}\left(\nu_{i,t}^k\right) \subseteq S_\alpha^+(u_k), \qquad t \geq 0, \quad i \in \mathcal{I}_k. \tag{150}$$

Define the degree-weighted mean by

$$\overline{x}_t^k = \sum_{i \in \mathcal{I}_k} \pi_i^k \int_{\mathbb{S}^{d-1}} x \, d\nu_{i,t}^k(x) = \int_{\mathbb{S}^{d-1}} x \, d\widehat{\mu}_t^k(x), \tag{151}$$

and define the degree-weighted disagreement by

$$\begin{aligned} V_t^k &= 1 - \|\overline{x}_t^k\|_2^2 \\ &= \int_{\mathbb{S}^{d-1}} \|x - \overline{x}_t^k\|_2^2 \, d\widehat{\mu}_t^k(x) \\ &= \frac{1}{2} \iint_{\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}} \|x - y\|_2^2 \, d\widehat{\mu}_t^k(x) d\widehat{\mu}_t^k(y). \end{aligned} \tag{152}$$

Define the edge-weighted disagreement by

$$\mathcal{D}_k\left[\boldsymbol{\nu}_t^k\right] = \frac{1}{2} \sum_{i,j \in \mathcal{I}_k} q_{ij}^k \iint_{\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}} \|x - y\|_2^2 \, d\nu_{i,t}^k(x) d\nu_{j,t}^k(y). \tag{153}$$

Differentiating the weighted mean through the continuity equations gives

$$\frac{d}{dt}\overline{x}_t^k = \sum_{i \in \mathcal{I}_k} \pi_i^k \int_{\mathbb{S}^{d-1}} \mathcal{X}_{i,\boldsymbol{\nu}_t^k,\beta}(x) \, d\nu_{i,t}^k(x). \tag{154}$$

Using $q_{ij}^k = q_{ji}^k$ and symmetrizing in (i, x) and (j, y), we obtain

$$\begin{aligned} \frac{d}{dt} V_t^k &= -2 \left\langle \overline{x}_t^k, \frac{d}{dt}\overline{x}_t^k \right\rangle \\ &= -\sum_{i,j \in \mathcal{I}_k} q_{ij}^k \iint_{\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}} e^{\beta\langle x,y\rangle} \left(1 - \langle x, y\rangle\right) \left\langle \overline{x}_t^k, x + y \right\rangle \, d\nu_{i,t}^k(x) d\nu_{j,t}^k(y). \end{aligned} \tag{155}$$

Forward invariance implies

$$\langle x, y\rangle \geq c_\alpha, \qquad \left\langle \overline{x}_t^k, x + y \right\rangle \geq 2c_\alpha \tag{156}$$

on the supports of the nodewise measures. Hence,

$$\frac{d}{dt} V_t^k \leq -2c_\alpha e^{\beta c_\alpha} \mathcal{D}_k\left[\boldsymbol{\nu}_t^k\right]. \tag{157}$$

To apply the graph Poincaré inequality, define

$$\mathcal{M}_{i,t}^k = \int_{\mathbb{S}^{d-1}} x \, d\nu_{i,t}^k(x), \qquad \sigma_{i,t}^{k\,2} = \int_{\mathbb{S}^{d-1}} \|x - \mathcal{M}_{i,t}^k\|_2^2 \, d\nu_{i,t}^k(x). \tag{158}$$

Then

$$V_t^k = \sum_{i \in \mathcal{I}_k} \pi_i^k \sigma_{i,t}^{k\,2} + \sum_{i \in \mathcal{I}_k} \pi_i^k \|\mathcal{M}_{i,t}^k - \overline{x}_t^k\|_2^2, \tag{159}$$

while

$$\begin{aligned} \mathcal{D}_k\left[\boldsymbol{\nu}_t^k\right] &= \sum_{i \in \mathcal{I}_k} \pi_i^k \sigma_{i,t}^{k\,2} \\ &\quad + \frac{1}{2} \sum_{i,j \in \mathcal{I}_k} q_{ij}^k \|\mathcal{M}_{i,t}^k - \mathcal{M}_{j,t}^k\|_2^2. \end{aligned} \tag{160}$$

Applying the graph Poincaré inequality to $(\mathcal{M}_{i,t}^k)_{i\in\mathcal{I}_k}$ and using $\lambda_k \le 1$ gives

$$\mathcal{D}_k\left[\boldsymbol{\nu}_t^k\right] \ge \lambda_k V_t^k. \tag{161}$$

Therefore,

$$\frac{d}{dt}V_t^k \le -2\lambda_k \cos(2\alpha) e^{\beta\cos(2\alpha)} V_t^k = -2\kappa_k V_t^k. \tag{162}$$

Gronwall's inequality yields

$$V_t^k \le V_0^k e^{-2\kappa_k t}. \tag{163}$$

We next estimate the metric speed of the product Wasserstein flow. Using Jensen's inequality, $\langle x, y\rangle \le 1$, and $\|\mathbf{P}_x^{\perp}(y)\|_2 \le \|x-y\|_2$, we obtain

$$\begin{aligned}\mathcal{I}_{\beta,k}^G\left[\boldsymbol{\nu}_t^k\right] &\le e^{2\beta}\sum_{i,j\in\mathcal{I}_k} q_{ij}^k \iint_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \|x-y\|_2^2\, d\nu_{i,t}^k(x) d\nu_{j,t}^k(y) \\ &= 2e^{2\beta}\mathcal{D}_k\left[\boldsymbol{\nu}_t^k\right].\end{aligned} \tag{164}$$

The upper graph Poincaré bound and Eq. 160 give

$$\mathcal{D}_k\left[\boldsymbol{\nu}_t^k\right] \le 2V_t^k. \tag{165}$$

Consequently,

$$\left|\dot{\boldsymbol{\nu}}_t^k\right|_{\mathcal{W}_{2,\pi^k}} \le \left(\mathcal{I}_{\beta,k}^G\left[\boldsymbol{\nu}_t^k\right]\right)^{1/2} \le 2e^{\beta}\left(V_0^k\right)^{1/2} e^{-\kappa_k t}. \tag{166}$$

Therefore, for $s > t$,

$$\begin{aligned}\mathcal{W}_{2,\pi^k}\left(\boldsymbol{\nu}_t^k, \boldsymbol{\nu}_s^k\right) &\le \int_t^s \left|\dot{\boldsymbol{\nu}}_r^k\right|_{\mathcal{W}_{2,\pi^k}} dr \\ &\le \frac{2e^{\beta}}{\kappa_k}\left(V_0^k\right)^{1/2} e^{-\kappa_k t}.\end{aligned} \tag{167}$$

The product Wasserstein space is complete, so $\boldsymbol{\nu}_t^k$ converges to a limit $\boldsymbol{\nu}_\infty^k$. Since $V_t^k \to 0$, the aggregate limit has zero variance and must be a Dirac mass. Because every $\pi_i^k > 0$, there exists $x_\infty^k \in S_\alpha^+(u_k)$ such that

$$\boldsymbol{\nu}_\infty^k = \left(\delta_{x_\infty^k}\right)_{i\in\mathcal{I}_k}. \tag{168}$$

Letting $s \to \infty$ in Eq. 167 and using $C_k = 2e^{\beta}/\kappa_k$, we obtain

$$\begin{aligned}\mathcal{W}_{2,\pi^k}\left(\boldsymbol{\nu}_t^k, \boldsymbol{\nu}_\infty^k\right) &= W_2\left(\widehat{\mu}_t^k, \delta_{x_\infty^k}\right) \\ &\le C_k\left(V_0^k\right)^{1/2} e^{-\kappa_k t}.\end{aligned} \tag{169}$$

Since $1/n_k \le \chi_k \pi_i^k$ for every $i \in \mathcal{I}_k$, we also have

$$W_2\left(\mu_t^k, \delta_{x_\infty^k}\right) \le \sqrt{\chi_k}\, C_k\left(V_0^k\right)^{1/2} e^{-\kappa_k t}. \tag{170}$$

For $k \ne \ell$, cap separation gives $d_{\mathbb{S}}(x_\infty^k, x_\infty^\ell) \ge d_{\mathbb{S}}(u_k, u_\ell) - 2\alpha > 0$. Hence, the limiting directions are pairwise distinct.

Finally, combining the componentwise couplings gives

$$\begin{aligned}W_2^2\left(\mu_t, \mu_\infty\right) &\le \sum_{k=1}^K m_k W_2^2\left(\mu_t^k, \delta_{x_\infty^k}\right) \\ &\le \sum_{k=1}^K m_k \chi_k C_k^2 V_0^k e^{-2\kappa_k t} \\ &\le C_G^2 e^{-2\kappa_{\min} t}.\end{aligned} \tag{171}$$

Taking square roots completes the proof. □

G Extended Experiment Results

G.1 Implementation Details

G.1.1 Numerical Simulations

Table 5 summarizes the implementation details for all four attention dynamics, including the shared numerical setup, initialization schemes, representative configurations for the *explicit opinion leader dynamics*, and graph configurations for the *implicit opinion leader dynamics*.

Table 5: Implementation details of the numerical simulations.

Hyperparameters	**Attention dynamics**			
	Standard	**Explicit (General)**	**Explicit (Conditional)**	**Implicit**
Number of particles n	180	180	180	180
Inverse temperature β	1.0	1.0	1.0	1.0
Integrator	RK4	RK4	RK4	RK4
Simulation Time T	10	10	0.05	10
Time step Δt	0.05	0.05	0.01	0.05
Initialization	vMF$(e_1, \kappa = 1)$	Uniform on $\mathbb{S}^2$	Grouped spherical caps	Grouped spherical caps
vMF concentration κ	1.0	✗	✗	✗
Number of representatives m	✗	6	6	✗
Representative directions	✗	$\{\pm e_1, \pm e_2, \pm e_3\}$	$\{\pm e_1, \pm e_2, \pm e_3\}$	✗
Component-center directions	✗	✗	✗	$\{\pm e_1, \pm e_2, \pm e_3\}$
Representative radii r_a	✗	1.85 ($\times 6$)	`auto` ($r_a \approx 6.0468$)	✗
Number of groups or components	✗	✗	6	6
Group or component sizes	✗	✗	30 ($\times 6$)	30 ($\times 6$)
Initial cap radius	✗	✗	$\delta = 0.25$	$\alpha = 0.75$
Explicit invariant-cap radius α_{exp}	✗	✗	$\alpha_{\text{exp}} = 0.28$	✗
Interaction radius R	✗	✗	✗	1.5
Coordinate spacing	✗	✗	✗	1.0
Intercomponent gap	✗	✗	✗	6.0

G.1.2 Frontier LLM Analysis

Table 6 reports the implementation details of the hidden-state analysis for the four frontier LLMs, including GPU allocations, benchmark sampling, dimensionality reduction, and final-layer HDBSCAN clustering. All models follow the same analysis protocol.

G.2 Cosine Silhouette Scores

We give the formal definition of the *cosine silhouette score* $S_{\cos}$ used in our analysis. Let y_i denote the reduced representation of token i, and let $\mathcal{C}(i)$ be its assigned cluster. For nonzero vectors, the cosine distance is

$$d_{\cos}(y_i, y_j) = 1 - \frac{\langle y_i, y_j \rangle}{\|y_i\|_2 \|y_j\|_2}. \tag{172}$$

This distance measures differences in direction rather than magnitude, making it natural for the spherical representations considered here. For unit vectors, it simplifies to $1 - \langle y_i, y_j \rangle$.

For a token in a cluster containing at least two members, define

$$\begin{aligned} a_i &= \frac{1}{|\mathcal{C}(i)| - 1} \sum_{\substack{j \in \mathcal{C}(i) \\ j \neq i}} d_{\cos}(y_i, y_j), \\ b_i &= \min_{\mathcal{C} \neq \mathcal{C}(i)} \frac{1}{|\mathcal{C}|} \sum_{j \in \mathcal{C}} d_{\cos}(y_i, y_j). \end{aligned} \tag{173}$$

Here, a_i measures the average distance from token i to its own cluster, while b_i measures its average distance to the nearest competing cluster. The tokenwise silhouette and its average over the N evaluated tokens are

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}, \qquad S_{\cos} = \frac{1}{N} \sum_{i=1}^{N} s_i. \tag{174}$$

Table 6: Implementation details of the hidden-state analysis across four frontier LLMs.

Hyperparameters	**Models**			
	GLM-4.7-Flash	**DeepSeek-V4-Flash**	**Kimi-K3**	**MiniMax-M3**
GPUs	2 × H100	4 × H100	32 × H100	16 × H100
Benchmarks	HumanEval, ARC-Easy, HellaSwag, and MATH			
Samples per benchmark	100	100	100	100
Hidden-state capture	Prefill	Prefill	Prefill	Prefill
Maximum captured tokens	2048	2048	2048	2048
Maximum sequence length	2048	2048	2048	2048
Observed Transformer layers	12	12	12	12
Hidden-state normalization	ℓ_2	ℓ_2	ℓ_2	ℓ_2
Reduction methods	UMAP, spherical UMAP			
UMAP clustering dimension	32	32	32	32
UMAP neighbors	15	15	15	15
UMAP minimum distance	0.1	0.1	0.1	0.1
Reducer input metric	cosine	cosine	cosine	cosine
Clustering algorithm	HDBSCAN	HDBSCAN	HDBSCAN	HDBSCAN
Clustering distance	cosine	cosine	cosine	cosine
Minimum cluster fraction	0.03	0.03	0.03	0.03
Minimum samples fraction	0.01	0.01	0.01	0.01
Cluster selection method	EOM	EOM	EOM	EOM
HDBSCAN α_{HDB}	1.0	1.0	1.0	1.0
Cluster selection ϵ	0.0	0.0	0.0	0.0
Noise-token assignment	Nearest cluster	Nearest cluster	Nearest cluster	Nearest cluster
Cases with undefined $S_{\cos}$	Report N/A	Report N/A	Report N/A	Report N/A

The score lies in $[-1, 1]$. Values near 1 indicate that a token is much closer to its own cluster than to any competing cluster; values near 0 indicate comparable distances to its own and a competing cluster; negative values indicate that another cluster is closer on average. A higher $S_{\cos}$ therefore indicates more compact and better-separated clusters. Because the score compares distances relatively, it measures cluster separability rather than absolute distance to consensus.

The silhouette score is defined only when the clustering contains at least two clusters. Accordingly, we compute $S_{\cos}$ only when $2 \leq K_{\mathrm{cluster}} \leq N-1$, where K_{cluster} denotes the number of non-noise clusters identified by HDBSCAN. When HDBSCAN identifies fewer than two non-noise clusters, corresponding to either a single non-noise cluster or an all-noise result, we report $S_{\cos}$ as *N/A*.

G.3 Explicit Opinion Leader Dynamics under Conditional Initialization



Figure 5: Finite-particle simulation on $\mathbb{S}^2$ of explicit opinion leader dynamics under conditional initialization. Colors assigned at $t = 0$ indicate membership in six groups defined by representatives and are retained throughout the simulation; flags mark the fixed representatives. Unlike general initialization, the conditional setting satisfies stronger concentration and separation conditions, under which the groups remain separated and converge exponentially to distinct *explicit opinion leaders*.

This experiment complements the general initialization simulation in the main text by considering stronger assumptions. Under conditional initialization, the $n = 180$ particles are divided into six equal groups, each initialized with antithetically paired samples drawn from a spherical cap of radius $\delta = 0.25$ centered at its associated direction. We set the invariant-cap radius to $\alpha = 0.28$ and choose a common representative magnitude such that Assumption 2.2 holds for every group. Fig. 5 shows

that the groups remain separated and converge to distinct limiting directions, consistent with the exponential convergence established in Theorem 2.3.

G.4 Extended Analysis of Token Dynamics in Frontier LLMs

G.4.1 Results across Reduced Embedding Dimensions

Table 7: Cosine silhouette scores for **DeepSeek-V4-Flash** on ARC-Easy across spherical UMAP embedding dimensions. The *upper* row reports mean final-layer scores across 100 random samples after projecting hidden states with spherical UMAP, while the *lower* row gives the Gaussian reference score at the same hidden dimension.

Model	Spherical UMAP embedding dimension			
	16	32	48	64
DeepSeek-V4-Flash	**0.86** $\pm$ **0.11** ($\uparrow$ **0.66**) 0.20 $\pm$ 0.13	**0.85** $\pm$ **0.12** ($\uparrow$ **0.63**) 0.22 $\pm$ 0.12	**0.85** $\pm$ **0.11** ($\uparrow$ **0.64**) 0.21 $\pm$ 0.12	**0.85** $\pm$ **0.13** ($\uparrow$ **0.63**) 0.22 $\pm$ 0.11

We examine the sensitivity of $S_{\cos}$ to the dimension of the reduced embedding. Using **DeepSeek-V4-Flash** on ARC-Easy, we repeat the quantitative analysis with spherical UMAP embedding dimensions of 16, 32, 48, and 64. As shown in Table 7, the mean cosine silhouette scores remain nearly unchanged at $\mathbf{0.85}$–$\mathbf{0.86}$, while the matched Gaussian references range from 0.20 to 0.22. The resulting gaps are similarly stable at $\mathbf{0.63}$–$\mathbf{0.66}$. The consistent results across embedding dimensions support the robustness of the observed clustering behavior across projection settings.

G.4.2 Results across Dimensionality-Reduction Methods

In addition to the spherical UMAP visualizations for ARC-Easy presented in the main text, we extend the same layerwise analysis of token hidden states to HellaSwag, HumanEval, and MATH for all four LLMs. As shown in Fig. 6, **Kimi-K3**, **MiniMax-M3**, and **DeepSeek-V4-Flash** consistently exhibit distinct final-layer clusters. By contrast, **GLM-4.7-Flash** exhibits weaker group separation and ends with a more concentrated final configuration. These results further demonstrate that the same qualitative pattern holds across benchmarks and aligns with the distinction in our framework between separated token groups and global alignment.

Table 8: Cosine silhouette scores across LLMs and benchmarks. For each LLM, the *upper* row reports mean final-layer scores across 100 random samples after projecting hidden states with UMAP, while the *lower* row gives the Gaussian reference score at the same hidden dimension.

Model	HumanEval	ARC-Easy	HellaSwag	MATH
Kimi-K3	**0.76** $\pm$ **0.27** ($\uparrow$ **0.58**) 0.18 $\pm$ 0.04	**0.89** $\pm$ **0.14** ($\uparrow$ **0.71**) 0.18 $\pm$ 0.06	**0.76** $\pm$ **0.18** ($\uparrow$ **0.61**) 0.15 $\pm$ 0.04	**0.90** $\pm$ **0.14** ($\uparrow$ **0.71**) 0.19 $\pm$ 0.05
MiniMax-M3	**0.87** $\pm$ **0.08** ($\uparrow$ **0.71**) 0.16 $\pm$ 0.03	**0.83** $\pm$ **0.08** ($\uparrow$ **0.66**) 0.17 $\pm$ 0.04	**0.86** $\pm$ **0.08** ($\uparrow$ **0.70**) 0.16 $\pm$ 0.03	**0.84** $\pm$ **0.10** ($\uparrow$ **0.66**) 0.18 $\pm$ 0.03
DeepSeek-V4-Flash	**0.73** $\pm$ **0.16** ($\uparrow$ **0.54**) 0.19 $\pm$ 0.07	**0.85** $\pm$ **0.12** ($\uparrow$ **0.63**) 0.22 $\pm$ 0.11	**0.64** $\pm$ **0.19** ($\uparrow$ **0.46**) 0.18 $\pm$ 0.04	**0.85** $\pm$ **0.11** ($\uparrow$ **0.64**) 0.21 $\pm$ 0.13
GLM-4.7-Flash	0.34 $\pm$ 0.30 ($\uparrow$ 0.19) 0.15 $\pm$ 0.07	0.22 $\pm$ 0.16 ($\uparrow$ 0.05) 0.17 $\pm$ 0.12	0.32 $\pm$ 0.25 ($\uparrow$ 0.17) 0.15 $\pm$ 0.04	0.30 $\pm$ 0.16 ($\uparrow$ 0.13) 0.17 $\pm$ 0.13

To examine whether such observed patterns depend on the choice of dimensionality-reduction methods, we repeat the analysis with UMAP in place of spherical UMAP. As reported in Table 8, the three sparse-attention LLMs score $\mathbf{0.64}$–$\mathbf{0.90}$ across all benchmarks, exceeding their matched Gaussian references by $\mathbf{0.46}$–$\mathbf{0.71}$. By comparison, **GLM-4.7-Flash** scores 0.22–0.34, with smaller differences of 0.05–0.19. The agreement between the results obtained with UMAP and spherical UMAP indicates that the stronger group separation observed in sparse-attention LLMs is robust across dimensionality-reduction methods, providing further empirical support for the clustering behavior predicted by our theory.

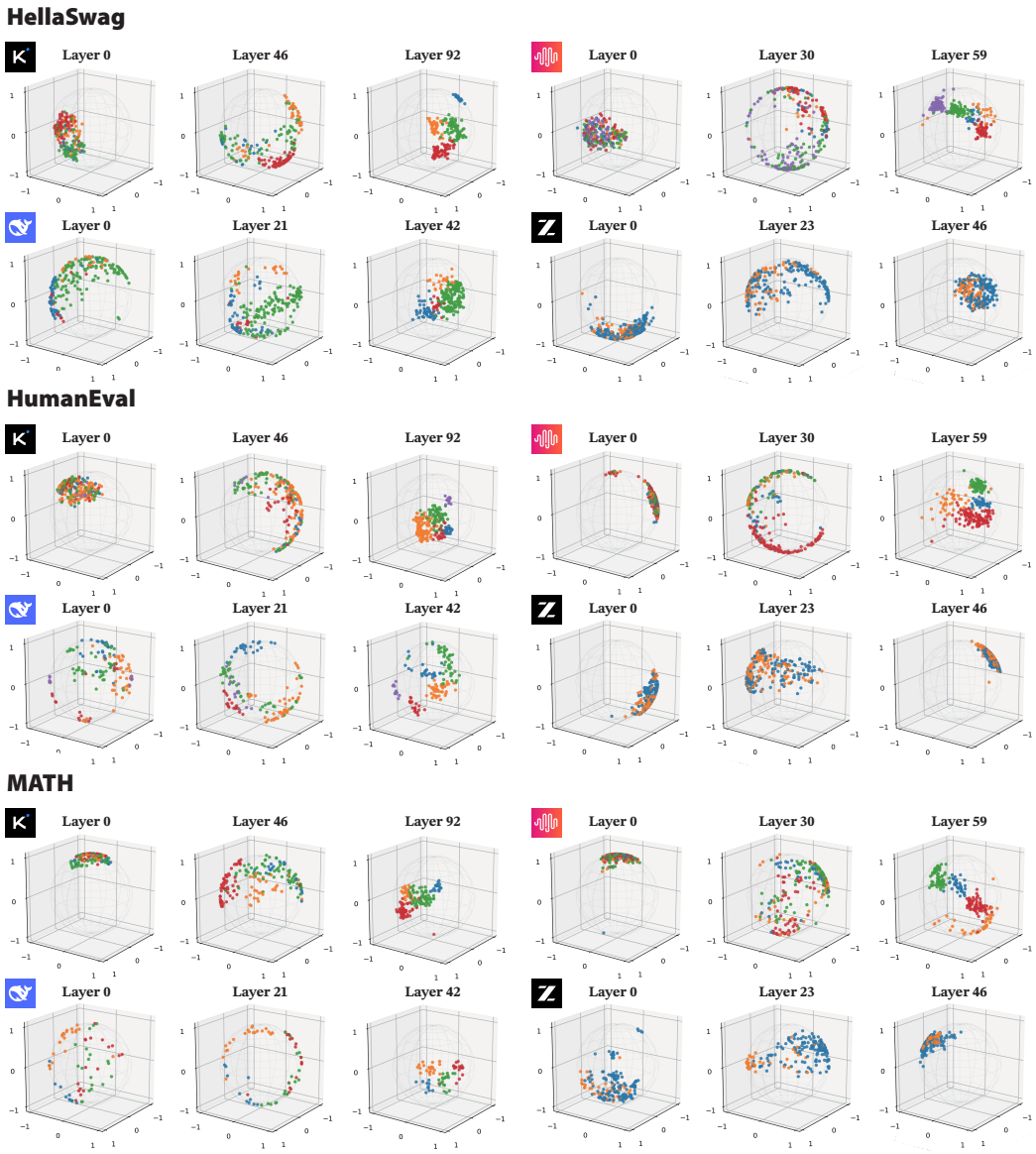

Figure 6: Layerwise evolution of token hidden states for HellaSwag, HumanEval, and MATH samples in **Kimi-K3**, **MiniMax-M3**, **DeepSeek-V4-Flash**, and **GLM-4.7-Flash**, visualized on unit sphere using spherical UMAP.

Fig. 7 visualizes the layerwise evolution of token hidden states after projection with UMAP. Across all four benchmarks, the sparse-attention LLMs exhibit clearly separated final-layer token groups, whereas **GLM-4.7-Flash** shows less distinct group structure, with its final-layer token representations concentrated in a smaller region. This pattern agrees with the quantitative results in Table 8 and closely matches the spherical UMAP findings presented in the main text, indicating that the observed contrast is consistent across dimensionality-reduction methods.

G.4.3 Results on RULER Long-Context Variable Tracking

To further examine token dynamics under longer contexts, we analyze **4K**-token inputs from the RULER (Hsieh et al., 2024) variable-tracking (VT) task with **MiniMax-M3** and **DeepSeek-V4-Flash**. Under a fixed selection budget, longer prompts reduce the fraction of available tokens or blocks retained by sparse attention, producing a more restrictive interaction pattern. Following the protocol in our main analysis, we visualize the layerwise evolution of token hidden states with

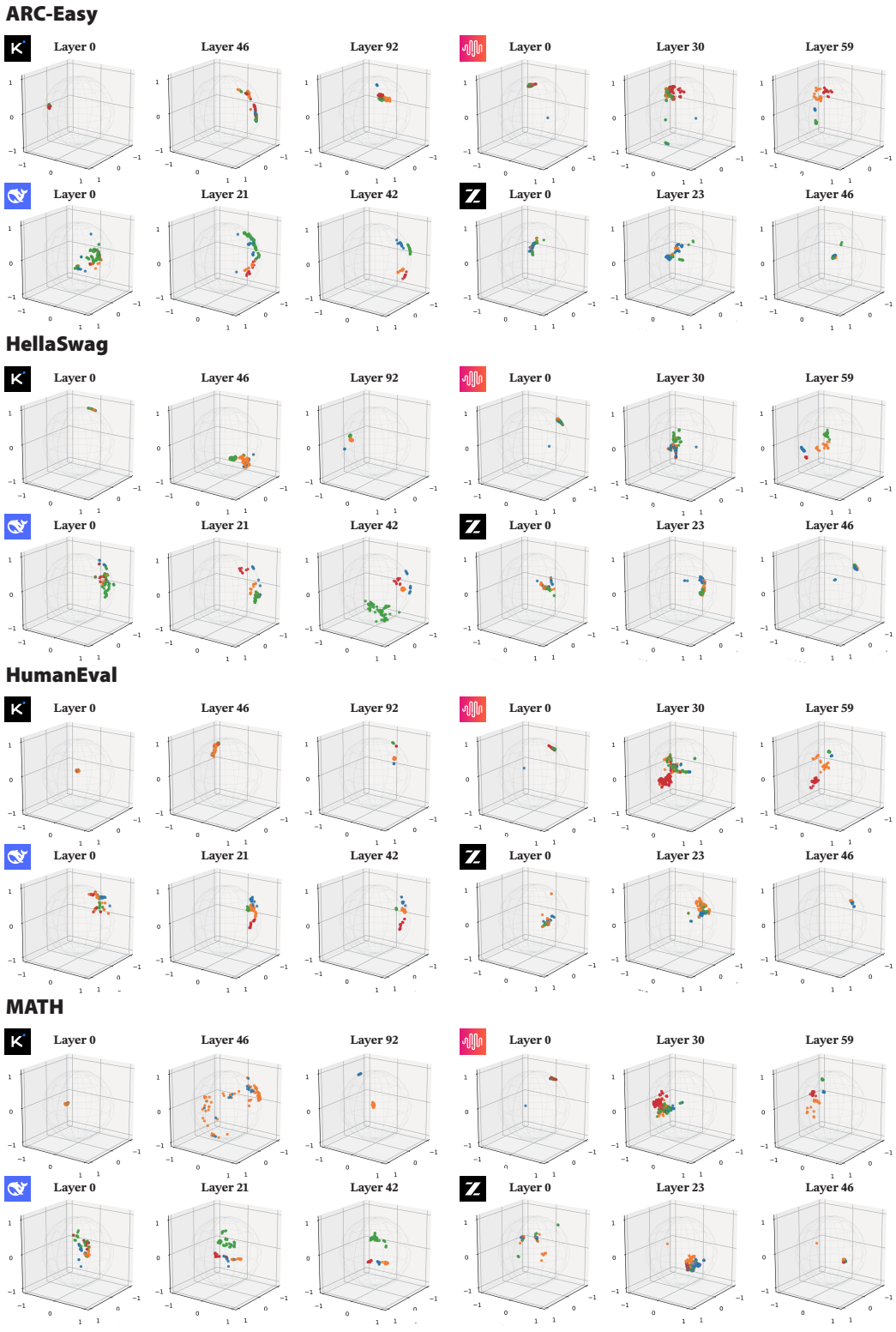

Figure 7: Layerwise evolution of token hidden states for ARC-Easy, HellaSwag, HumanEval, and MATH samples in **Kimi-K3**, **MiniMax-M3**, **DeepSeek-V4-Flash**, and **GLM-4.7-Flash**, visualized on unit sphere using UMAP.

spherical UMAP and report the mean final-layer $S_{\cos}$ over 100 samples, together with a Gaussian reference matched to each hidden dimension.

Table 9: Cosine silhouette scores for **DeepSeek-V4-Flash** and **MiniMax-M3** on RULER VT task (**4K**-token length). The *upper* row reports mean final-layer scores across 100 random samples after projecting hidden states with spherical UMAP, while the *lower* row gives the Gaussian reference score at the same hidden dimension.

Model	RULER VT
DeepSeek-V4-Flash	**0.58 ± 0.19** N/A (< 2 non-noise clusters)
MiniMax-M3	**0.82 ± 0.05** N/A (< 2 non-noise clusters)

Table 9 reports mean final-layer $S_{\cos}$ scores of $\mathbf{0.82 \pm 0.05}$ for **MiniMax-M3** and $\mathbf{0.58 \pm 0.19}$ for **DeepSeek-V4-Flash**. For both LLMs, HDBSCAN identifies fewer than two non-noise clusters in every matched Gaussian reference sample. Since the cosine silhouette score requires at least two clusters, we report $S_{\cos}$ as *N/A*. The findings provide quantitative evidence of multiple-group organization in the reduced token representation space of both sparse-attention LLMs. The layerwise visualizations in Fig. 8 further illustrate how this structure develops across layers: token groups that largely overlap at the input layer reorganize and become more distinct and separated towards the final layers. Together, these results suggest that the multiple-group pattern observed in the main analysis persists under longer-context inputs.

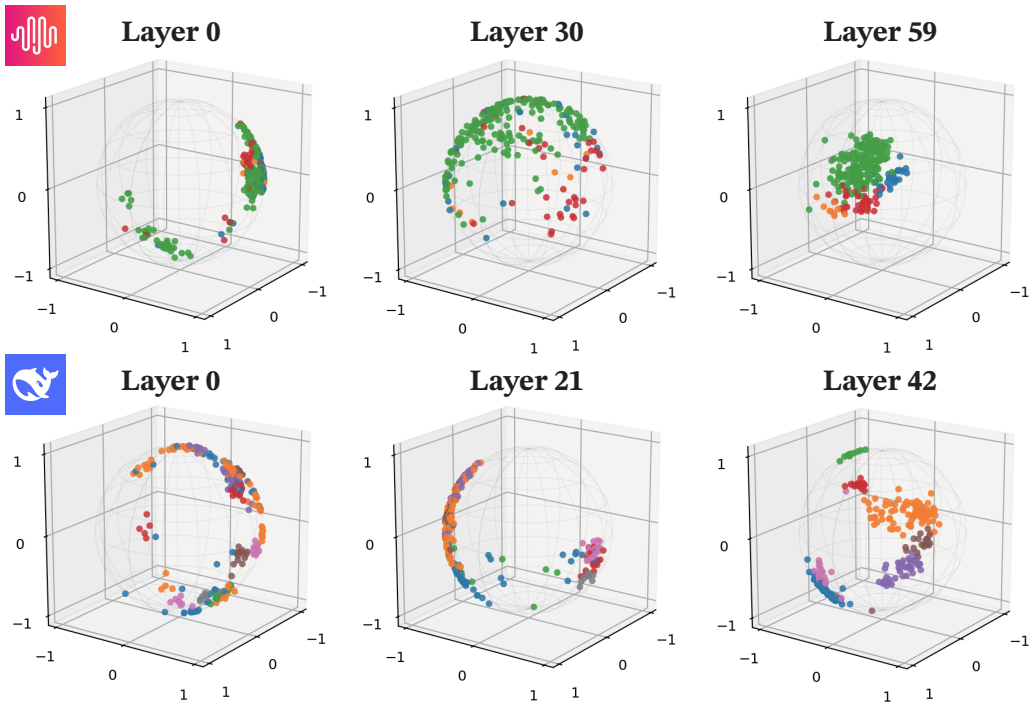

Figure 8: Layerwise evolution of token hidden states for RULER VT samples in **MiniMax-M3** and **DeepSeek-V4-Flash**, visualized on unit sphere using spherical UMAP.