

Thermodynamic Binding: Freezing *Chimeric* States in Multi-Modal Associative Memories



Code & Paper

Elena Agliari^{1,3,4,5}, Adriano Barra^{1,2,4,5}, Andrea Ladiana^{1,2*}, Andrea Lepre^{1,3}

¹GNFM, Istituto Nazionale d'Alta Matematica, Rome · ²Dept. Basic and Applied Sciences for Engineering, Sapienza Università di Roma · ³Dept. Mathematics, Sapienza Università di Roma · ⁴INFN, Sezione di Lecce · ⁵CNR, Sezione di Lecce · * andrea.ladiana@uniroma1.it

1 · THE PROBLEM

Chimeric States in Multi-Modal Attention

In standard cross-attention, each modality computes its own softmax distribution $\mathbf{p}^{(a)}$ over a shared bank of K prototypes. With L modalities, this admits K^L cross-modal assignments, of which only K are coherent (all modalities agree on the same prototype). The remaining configurations are **chimeric states**: internally contradictory retrievals in which different modalities lock onto incompatible prototypes. This is not a training failure but an **architectural pathology**.

K^L - K chimeric \gg K coherent

e.g., $L = 3, K = 12$: 1716 chimeric vs 12 coherent (143 : 1)

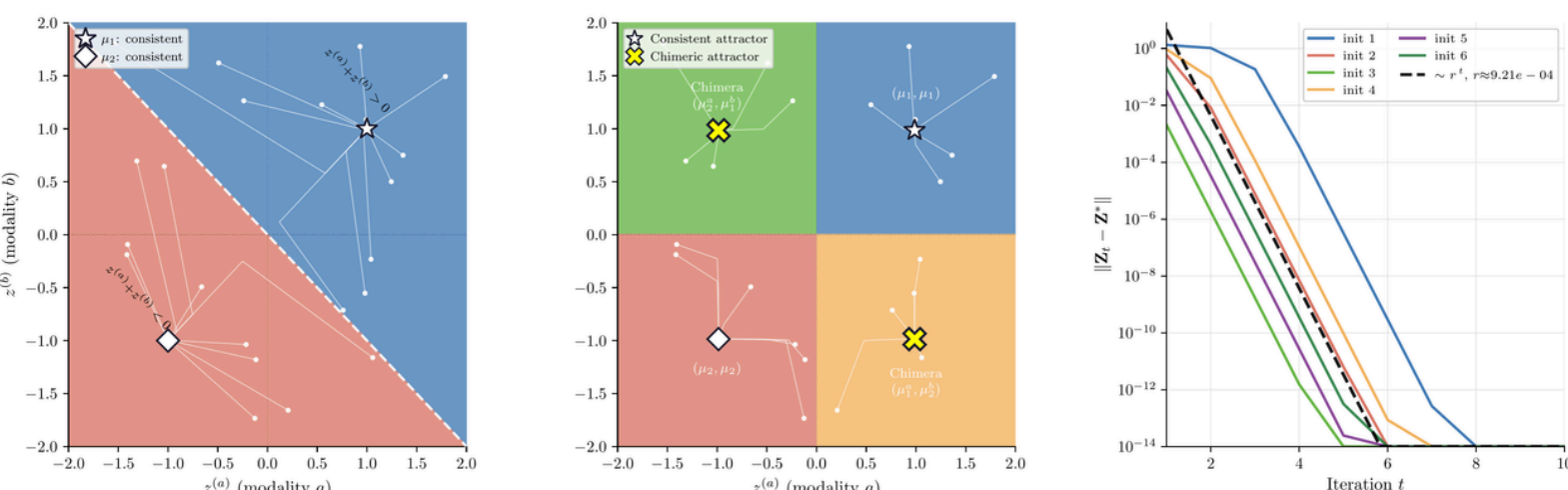


Fig. 1 Basin of attraction and Kt geometric convergence ($L=2, d=1, K=2, \beta=2.5$). (A) Under CSA the state space splits into exactly two basins; every trajectory converges to a consistent attractor. (B) Decoupled attention produces four basins; the two off-diagonal attractors (yellow \times) are genuine chimeric fixed points. (C) Kt convergence: six random initializations contract to the same fixed point at geometric rate.

2 · OUR SOLUTION

Consensus Split-Bank Attention (CSA)

CSA replaces L independent softmax distributions with **one shared consensus**. Query-key similarities from every modality are summed into a **single global score**; one softmax produces the shared distribution \mathbf{p} ; all modalities update synchronously. The mechanism maps to a **Product-of-Experts** fusion: $\mathbf{p}_\mu \propto \prod_a \exp(\beta \mathbf{k}_\mu^{(a)})$.

$$S_\mu(\{z\}) = \sum_{a=1}^L \sum_{b=1}^L \bar{A}_{ab}(z^{(a)}, \mathbf{k}_\mu^{(b)}) \quad \mathbf{p}_\mu = \frac{\exp(\beta S_\mu)}{\sum_\nu \exp(\beta S_\nu)} \quad z_{\text{new}}^{(a)} = \sum_{(a \leftarrow b)} \bar{A}_{ab} K^{(b)} \mathbf{p}$$

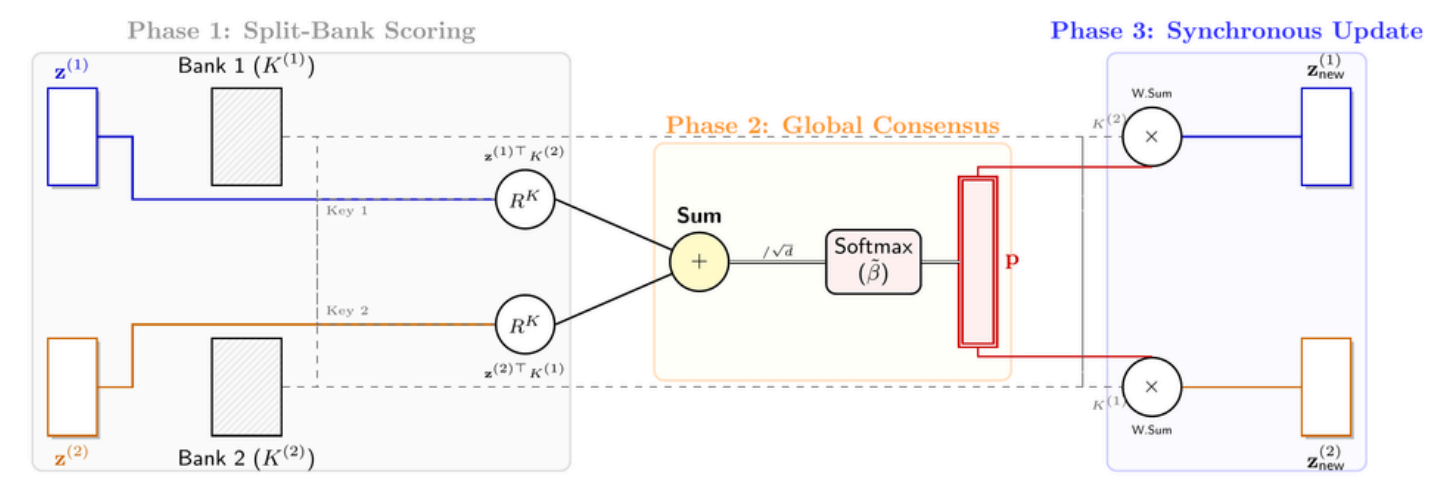


Fig. 2 Query-key inner products from all modality pairs are summed into a single global score vector S ; a unique shared distribution \mathbf{p} (the "semantic bottleneck") is then broadcast back to update every modality synchronously. This three-phase structure collapses the equilibrium binding space from K^L cross-modal assignments to exactly K coherent ones.

3 · THEORY

Energy, Convergence & Capacity

Monotonic Energy Descent

The CSA update is the unique outcome of the **Concave-Convex Procedure (CCCP)** applied to a global **Difference-of-Convex energy**:

$$\mathcal{E}(\{z\}) = \underbrace{\sum_{a=1}^L \frac{1}{2} \|z^{(a)}\|^2}_{\mathcal{U}(z) \text{ (convex; Hessian = I)}} - \frac{1}{\beta} \log \sum_{\mu=1}^K e^{\beta S_\mu(\{z\})} + C_{\text{shift}} \quad \mathcal{V}(z) \text{ (convex; LSE } \circ \text{ linear)}$$

The CCCP prescription $\nabla U(z_{\text{new}}) = \nabla V(z_{\text{old}})$ yields, via the chain rule, the softmax consensus \mathbf{p} and the linear bank-readout of the CSA update rule.

Theorem 1 — Monotonic Descent & Strong Convergence

Let $(z_t)_{t \geq 0}$ be the mTAM dynamics. (i) The energy is strictly non-increasing:

$$\mathcal{E}(z_{t+1}) \leq \mathcal{E}(z_t) - \frac{1}{2} \|z_{t+1} - z_t\|^2$$

with equality iff z_t is a stationary point. Trajectories are bounded within the convex hull of the memory banks.

(ii) Since \mathcal{E} is real-analytic, it satisfies the **Kurdyka-Łojasiewicz inequality**. Consequently, the trajectory has **finite arc length** $\sum_t \|z_{t+1} - z_t\| < \infty$ and converges to a **single stationary point** z^* .

Each fixed point defines a coherent multi-modal retrieval. Standard cross-attention has no global energy and offers no comparable guarantee.

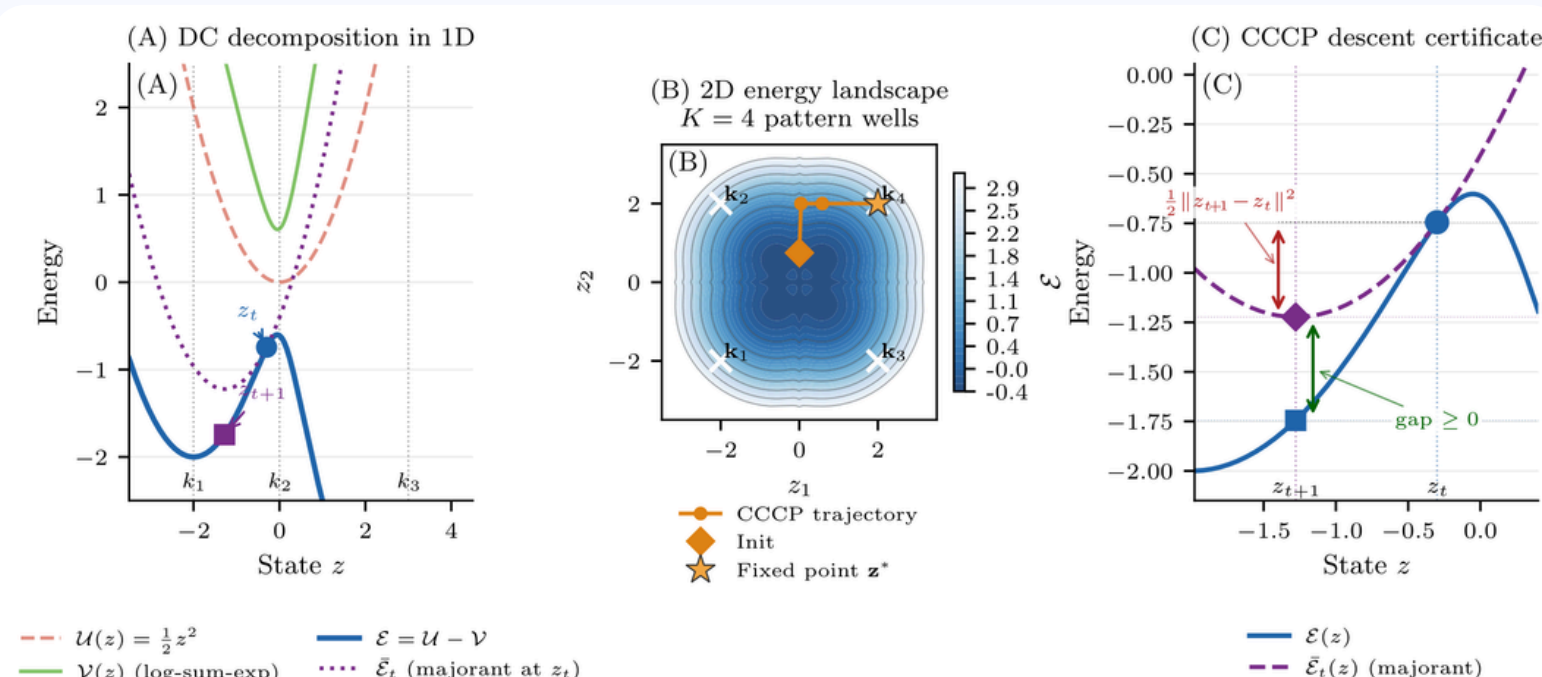


Fig. 3 (A) 1D slice of $E(z) = U(z) - V(z)$ for $K=3$ patterns (dashed vertical). The quadratic confinement $U = \frac{1}{2} z^2$ pulls toward the origin; the log-sum-exp interaction V carves memory wells at each pattern location. The CCCP surrogate \tilde{E}_t (dashed) is obtained by linearizing V at z_t , producing a convex upper bound that is tight and tangent to E at z_t , minimizing it yields z_{t+1} . (B) 2D contour of $E(z)$ for $K=4$ patterns showing four attractor wells; a random initialization converges to the nearest pattern via CCCP. (C) Descent decomposition: $E(z_t) - E(z_{t+1}) = [E(z_t) - \tilde{E}_t(z_{t+1})] + [\tilde{E}_t(z_{t+1}) - E(z_{t+1})]$. The first bracket is the surrogate decrease $\frac{1}{2} \|z_{t+1} - z_t\|^2$ (in the Hessian = I case); the second is the nonnegative majorization gap.

Graph Topology Controls the Noise

Convergence guarantees say nothing about how many patterns can be reliably stored. We address this via an **extreme-value analysis** inspired by the **Random Energy Model (REM)**.

Key idea: Define the edge-lifted Hilbert space $H_E = \bigoplus_{(a,b)} \mathbb{R}^d$. The lifting operators P (states) and Q (patterns) reduce the global score S_μ to a **single inner product** in H_E . The graph reappears as a **metric distortion**: $Q^*Q = \text{diag}(\bar{d}^{(a)}) \otimes I_d$ — heavily queried banks are expanded, lightly queried banks are compressed. This reduction is **exact**: mTAM becomes a Modern Hopfield Network in graph-warped space.

Theorem 2 — Effective Topological Variance

Model the $K-1$ non-target keys as i.i.d. draws from $N(0, \Gamma \otimes I_d)$, where $\Gamma \in \mathbb{R}^{(K-1) \times (K-1)}$ is the symmetric PSD cross-modal covariance ($\Gamma_{aa}=1$). Define the query overlap matrix $Q_d := d^{-1}(z^{\otimes 2}, z^{\otimes 2})$. For any fixed Z , the consensus score of a distractor is Gaussian: $S_\nu(Z) \sim N(0, \Sigma_{\text{eff}}^2(Z))$, where

$$\Sigma_{\text{eff}}^2(Z) = d \cdot \text{Tr}(MQ), \quad M = \bar{A} \Gamma \bar{A}^T$$

$M = \bar{A} \bar{A}^T$ is the topology-covariance sandwich: \bar{A} mixes source banks via the graph, Γ couples cross-modal correlations, and Q projects onto current query directions.

Theorem 3 — REM Critical Capacity

By extreme-value theory, the maximum of $K-1$ i.i.d. $N(0, \Sigma_{\text{eff}}^2)$ scores grows as $\Sigma_{\text{eff}} \sqrt{2 \log K}$. Retrieval succeeds iff S_{target} exceeds this maximum. With $\alpha_{\text{tot}} := (Ld)^{-1} \log K$ and $\rho_{\text{eff}} := \Sigma_{\text{eff}}^2/d$, in the thermodynamic limit ($K, d \rightarrow \infty$ with α_{tot} fixed), retrieval of μ^* succeeds w.h.p. iff $\alpha_{\text{tot}} < \alpha_{\text{crit}}(\alpha_{\text{tot}})$, where:

$$\alpha_{\text{tot},c} = \frac{s^2}{2L\rho_{\text{eff}}} = \frac{\text{Tr}^2[\bar{A}^T \Gamma \bar{A}]}{2L \text{Tr}[\bar{A} \Gamma \bar{A}^T]}$$

Numerator s^2 : signal depth squared (maximized at the CCCP fixed point). Denominator $2L\rho_{\text{eff}}$: noise floor scaled by architecture depth.

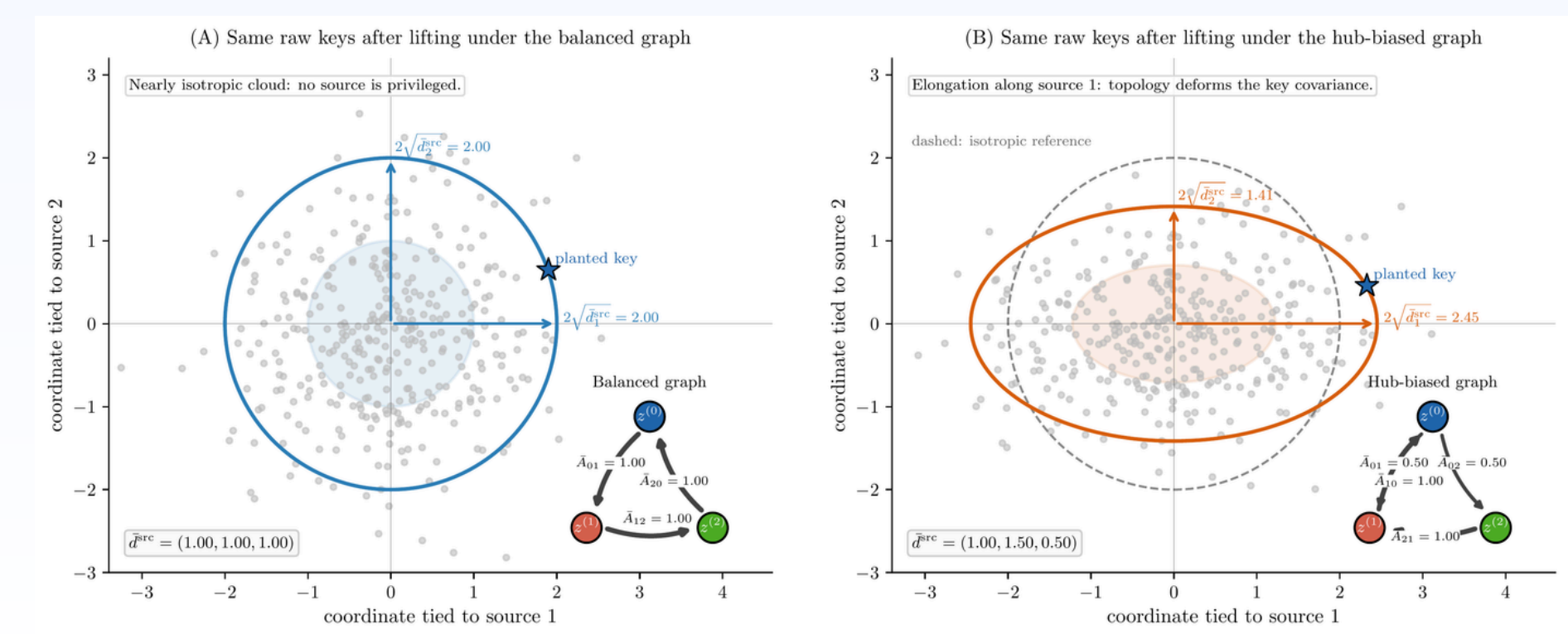


Fig. 4 Anisotropic deformation in H_E . (A) Balanced graph: isotropic key cloud. (B) Hub-biased graph: Q stretches high-degree sources \rightarrow controls Σ_{eff}^2 .

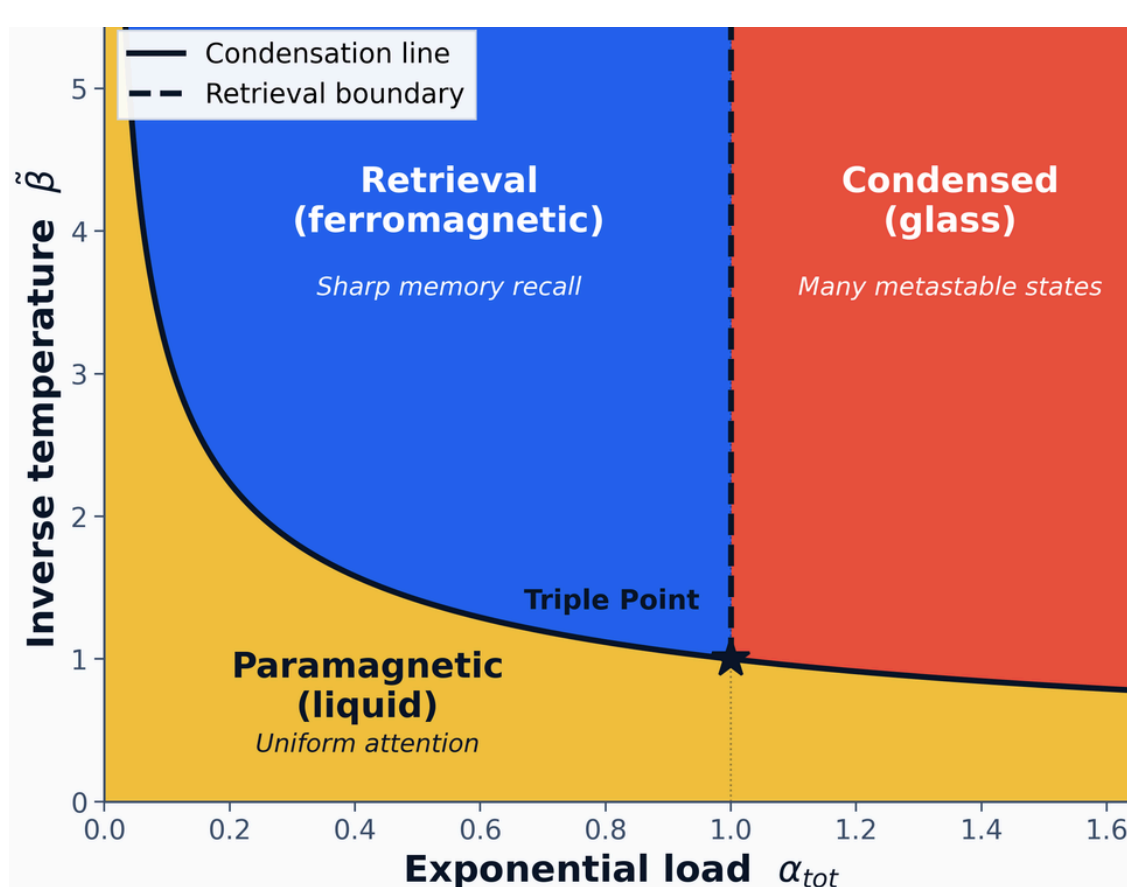
CORE INSIGHT

CSA **structurally eliminates** chimeric states by collapsing the equilibrium space from K^L to K . The CCCP structure guarantees that every trajectory converges to a coherent fixed point. The graph topology enters as a **metric distortion** of the pattern ensemble in a lifted Hilbert space H_E , controlling both convergence speed and memory capacity through a single quantity: the **effective topological variance** Σ_{eff}^2 .

4A · PHASE DIAGRAM

Retrieval Phase Diagram

The system exhibits a clear **retrieval phase transition**: at low β , the consensus \mathbf{p} is diffuse (**paramagnetic phase**); above a critical β , it locks onto the correct attractor (**retrieval phase**). A **condensation transition (glass phase)** lies between — analogous to the **Random Energy Model**. The three-phase structure is fully predicted by the capacity theory.



4B · ARCHITECTURE & GUARANTEES

Trainable Architecture & One-Step Retrieval

End-to-end trainability. CSA maps directly to standard multi-head attention. Backpropagation through T unrolled CCCP iterations is well-defined: all parameters ($K^{(b)}, V^{(a)}, W_0^{(a)}, W_0^{(b)}$) are shared across iterations; gradients accumulate across the full chain. For $L=1$, CSA recovers the Modern Hopfield Network.

Practical scaling. Per-step cost is $O(LKd)$, identical to standard dot-product attention. Wall-clock benchmarks ($L=3$, batch 32): at $d=1024$, $K=10000$ — a single forward iteration takes <30 ms on CPU. Cost grows linearly in both d and K .

One-step retrieval guarantee. Define the consensus margin $\Delta_{\mu^*}(Z) := S_{\mu^*}(Z) - \max_{\mu \neq \mu^*} S_\mu(Z)$ and the graph-mixed prototype $\tilde{K}_\mu^{(a)} := \sum_b \bar{A}_{ab} K_\mu^{(b)}$. If $\Delta_{\mu^*} > 0$, then $\mathbf{p}_\mu \approx 1/(1+(K-1)e^{-\beta \Delta_{\mu^*}})$ and $\|F(Z) - \tilde{K}_{\mu^*}\| \leq D(1-\mathbf{p}_{\mu^*})$, where $D := \max_{\mu, \nu} \|\tilde{K}_\mu - \tilde{K}_\nu\|$. A margin $\beta \Delta_{\mu^*} \gtrsim 5$ drives $>99\%$ of the attention weight onto the correct pattern, independently of K .

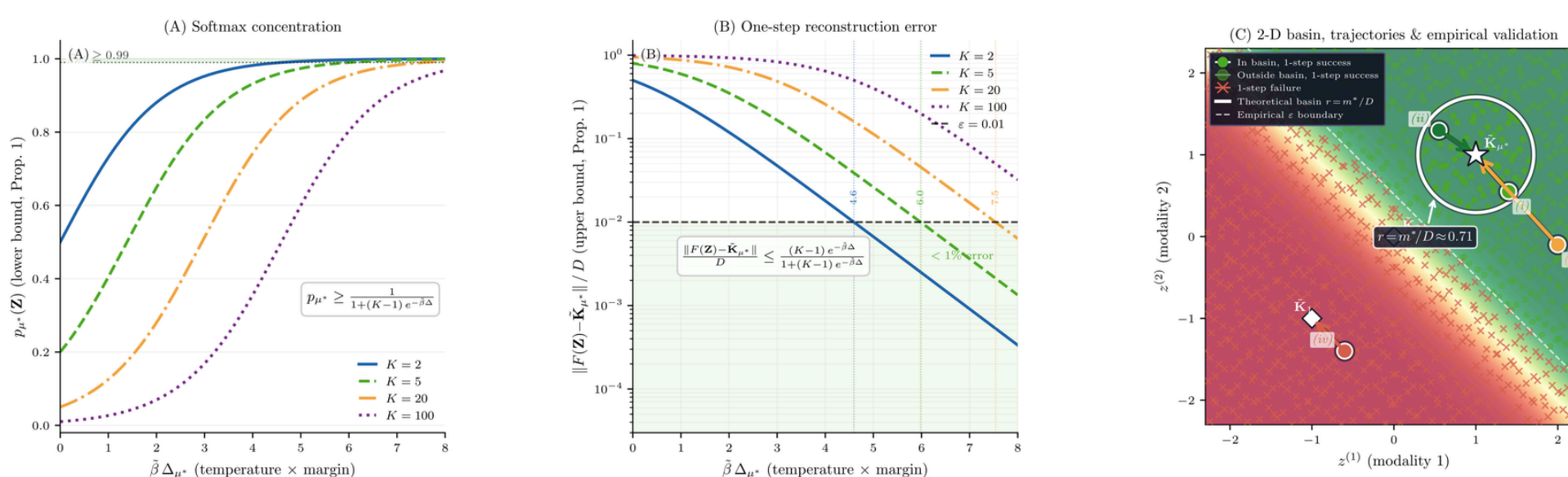


Fig. 6 One-step retrieval guarantee. (A) Softmax concentration: lower bound on \mathbf{p}_{μ^*} vs $\beta \Delta_{\mu^*}$ for $K \in \{2, 5, 20, 100\}$. (B) Reconstruction error upper bound on log scale; green zone = safe operating region. (C) Empirical validation on the minimal model ($L=2, K=3$).

5 · TAKE-HOME

Conclusions & Future Directions

- **mTAM eliminates chimeric states by construction.** CSA collapses the equilibrium binding space from K^L to K . The shared softmax enforces that all modalities converge to the same attractor — structurally, not by training.
- **CCCP guarantees monotonic convergence.** The global DC energy decreases by at least $\frac{1}{2} \|z_{t+1} - z_t\|^2$ at each step. The Kurdyka-Łojasiewicz property ensures every trajectory converges to a single coherent stationary point (Theorem 1).
- **Graph topology controls memory capacity.** The effective topological variance $\Sigma_{\text{eff}}^2 = d \cdot \text{Tr}(\bar{A} \Gamma \bar{A}^T Q)$ governs the noise floor.
- **One-step chimera resolution.** From 3-way chimeric initializations, CSA resolves to the correct coherent attractor in a single step.
- **$L=1$ recovers the Modern Hopfield Network.** mTAM generalizes the MHN framework to multi-modal consensus retrieval with standard backpropagation.

Future directions: Real-world benchmarks with learned representations · Capacity under non-Gaussian key distributions · Hierarchical and sparse retrieval for large pattern sets · Basin structure characterization.

References

- [1] Ramsauer, H. et al. "Hopfield Networks is All You Need." ICLR 2021.
- [2] Krotov, D. & Hopfield, J. "Dense Associative Memory for Pattern Recognition." NeurIPS 2016.
- [3] Yuille, A. L. & Rangarajan, A. "The Concave-Convex Procedure." Neural Computation, 2003.
- [4] Greff, K. et al. "On the Binding Problem in Artificial Neural Networks." arXiv:2012.05208, 2020.
- [5] Hinton, G. E. "Training Products of Experts by Minimizing Contrastive Divergence." Neural Computation, 2002.
- [6] Bolte, J. et al. "Proximal Alternating Linearized Minimization." Math. Programming, 2014.
- [7] Agliari, E. et al. "Generalized Hebbian Hetero-Associative Memories." 2025.
- [8] Tsai, Y.-H. H. et al. "Multimodal Transformer for Unaligned Multimodal Language Sequences." ACL 2019.

github.com/andrea-ladiana/mtam_layers