THERMODYNAMIC BINDING: FREEZING CHIMERIC STATES IN MULTI-MODAL ASSOCIATIVE MEMORIES

Elena Agliari[1,3,4,5], **Adriano Barra**[1,2,4,5], **Andrea Ladiana**[1,2,*], **Andrea Lepre**[1,3]
[1]Istituto Nazionale d'Alta Matematica (GNFM), Rome, Italy
[2]Dept. of Basic and Applied Sciences for Engineering, Sapienza Università di Roma, Rome, Italy
[3]Department of Mathematics, Sapienza Università di Roma, Rome, Italy
[4]Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Lecce, Italy
[5]Consiglio Nazionale delle Ricerche (CNR), Sezione di Lecce, Italy

ABSTRACT

Multi-modal inference requires heterogeneous perceptual streams to converge to a single, internally consistent interpretation. Standard cross-attention does not enforce this consistency: each modality maintains an independent posterior over a shared memory bank, which admits chimeric states, namely stable configurations in which different modalities retrieve different prototypes from the shared memory. We introduce the Multi-modal Transformer Associative Memory (mTAM)[1], an energy-based architecture that precludes chimeric states by construction. Its core mechanism, Consensus Split-Bank Attention (CSA), aggregates query–key evidence across modalities into a single global score, produces one shared distribution over memory, and broadcasts it synchronously to every modality. The resulting dynamics correspond to the Concave–Convex Procedure applied to a Difference-of-Convex energy, which guarantees monotonic descent and convergence of each trajectory to a stationary point. A graph-lifting construction maps the model to a Modern Hopfield Network and yields a topology-dependent critical load through an extreme-value capacity analysis in the spirit of the Random Energy Model. Synthetic experiments show retrieval transitions, one-step chimera resolution where standard baselines fail, and topology-dependent capacity scaling consistent with the theory.

1 INTRODUCTION

Multi-modal architectures must maintain globally coherent internal representations. 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 ($\mu_1 = \cdots = \mu_L$); the remaining configurations are *chimeric states*, namely internally contradictory retrievals in which different modalities lock onto incompatible prototypes. This is not a training failure but an architectural pathology: decoupled attention parametrizes the binding decision as a free tuple $(\mu_1, \ldots, \mu_L) \in [K]^L$, and no loss function can structurally exclude chimeric fixed points from the inference hypothesis class (Greff et al., 2020).

Associative-memory models offer a natural framework for this problem, because retrieval is governed by a global energy landscape rather than independent local decisions. The connection between Dense Associative Memories (Krotov & Hopfield, 2016) and Transformer attention (Vaswani et al., 2017; Ramsauer et al., 2020) reinterprets attention as energy-based retrieval, but remains confined to the auto-associative, single-modality setting. Hetero-associative memories (Kosko, 2002; Agliari et al., 2025a;b; Alessandrelli et al., 2025a;b) extend naturally to multi-modal binding through shared attractors, suggesting that consistency should be imposed at the level of retrieval dynamics itself.

We introduce the *Multi-modal Transformer Associative Memory* (mTAM), an energy-based architecture for hetero-associative multi-modal retrieval. Its core mechanism, *Consensus Split-Bank Attention* (CSA), replaces per-modality softmax distributions with a single shared distribution: rather than letting each modality compute its own posterior, CSA pools all query–key evidence into one global score, forms a consensus, and broadcasts it synchronously to every modality, so

*Corresponding author: andrea.ladiana@uniroma1.it
[1]Code: https://github.com/andrea-ladiana/mtam_layers

chimeric states lie outside the model's hypothesis class. The CSA update arises as the unique prescription of the Concave–Convex Procedure (CCCP) (Yuille & Rangarajan, 2003) applied to a global Difference-of-Convex (DC) energy, guaranteeing monotonic descent and convergence of each trajectory to a stationary point (Theorem 1; $L{=}1$ recovers the Modern Hopfield Network). An extreme-value capacity analysis in the spirit of the Random Energy Model yields an explicit critical load $\alpha_{\mathrm{tot},c}$ depending on graph topology through an effective variance Σ^2_{eff} (Theorem 5), with balanced graphs expanding and hub-dominated graphs contracting the retrieval region. A compact notation summary is provided in Appendix B for ease of reference.

2 THE MTAM FRAMEWORK

We consider a system with L modalities (hereafter also called layers) interacting over a directed graph with edge set $\mathcal{E}_{\mathrm{het}} \subseteq [L] \times [L]$: an edge $(a \leftarrow b)$ indicates that layer a reads from the memory bank of layer b. Each layer a has a state $z^{(a)} \in \mathbb{R}^d$ and a bank of K stored patterns $K^{(b)} \in \mathbb{R}^{d\times K}$ with $\|k_\mu^{(b)}\| = \sqrt{d}$. Let $\bar{A}_{ab} := A_{ab}/d_a^{\mathrm{in}}$ (rows sum to one, d_a^{in} = in-degree of a) and $\tilde{\beta} := \beta/\sqrt{d}$.

Consensus Split-Bank Attention (CSA). Instead of letting each modality form its own posterior and update independently, CSA computes one shared posterior. Query–key similarities from every modality and every graph edge are summed into a single global score vector; one softmax produces the shared distribution $\mathbf{p}$; all modalities update synchronously. When one modality has strong evidence for a prototype, that evidence propagates immediately to all others; when modalities disagree, $\mathbf{p}$ remains diffuse, preventing any chimeric lock-in.

The *Global Pattern Score* for prototype μ is:

$$S_\mu(\{z\}) := \sum_{a=1}^{L}\sum_{b=1}^{L} \bar{A}_{ab}\,\langle z^{(a)}, k_\mu^{(b)}\rangle, \tag{1}$$

from which a single shared distribution is derived at temperature $\tilde{\beta}$:

$$p_\mu(\{z\}) = \frac{e^{\tilde{\beta} S_\mu}}{\sum_\nu e^{\tilde{\beta} S_\nu}}. \tag{2}$$

The state update for each layer is then a consensus-weighted retrieval from its connected banks:

$$z_{\mathrm{new}}^{(a)} = \sum_{(a\leftarrow b)\in\mathcal{E}_{\mathrm{het}}} \bar{A}_{ab}\, K^{(b)}\,\mathbf{p}. \tag{3}$$

Product-of-Experts reading. The global score decomposes as $S_\mu = \sum_a \ell_\mu^{(a)}$, where $\ell_\mu^{(a)} := \sum_b \bar{A}_{ab}\langle z^{(a)}, k_\mu^{(b)}\rangle$ is the per-layer logit contribution. The consensus distribution therefore factorizes as $p_\mu \propto \prod_a \exp(\tilde{\beta}\,\ell_\mu^{(a)})$: each modality acts as an expert whose evidence multiplies in probability space, realizing a Product-of-Experts fusion (Hinton, 2002; Joshi et al., 2022). Agreement concentrates $\mathbf{p}$ sharply; disagreement diffuses it. A strong signal in one modality pulls all others toward the coherent pattern; conversely, cross-modal disagreement collapses confidence into a high-entropy consensus rather than locking into a chimeric attractor.

Global energy and CCCP structure. The CSA update 3 is the unique outcome of the CCCP (Yuille & Rangarajan, 2003) applied to the DC energy (C_{shift} ensures $\mathcal{E} \geq 0$):

$$\mathcal{E}(\{z\}) := \underbrace{\sum_{a=1}^{L}\frac{1}{2}\|z^{(a)}\|^2}_{\mathcal{U}(z)\text{ (convex)}} \underbrace{-\frac{1}{\tilde{\beta}}\log\sum_{\mu=1}^{K} e^{\tilde{\beta} S_\mu(\{z\})}}_{\mathcal{V}(z)\text{ (convex)}} + C_{\mathrm{shift}}. \tag{4}$$

The decomposition is valid: $\mathcal{U}$ is strongly convex (Hessian $= I$), and $\mathcal{V}$ is convex as the composition of the Log-Sum-Exp function with a linear map. The CCCP prescription $z_{\mathrm{new}}^{(a)} = \nabla_{z^{(a)}}\mathcal{V}(z_{\mathrm{old}})$ applied to this interaction term yields, via the chain rule, the softmax consensus $\mathbf{p}$ and the linear bank-readout in equation 3.

Theorem 1 (Monotonic Descent and Strong Convergence)**.** *Let $\{z_t\}_{t\geq 0}$ be the sequence of states generated by the mTAM dynamics 3.* ***(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. The trajectory is bounded within the convex*

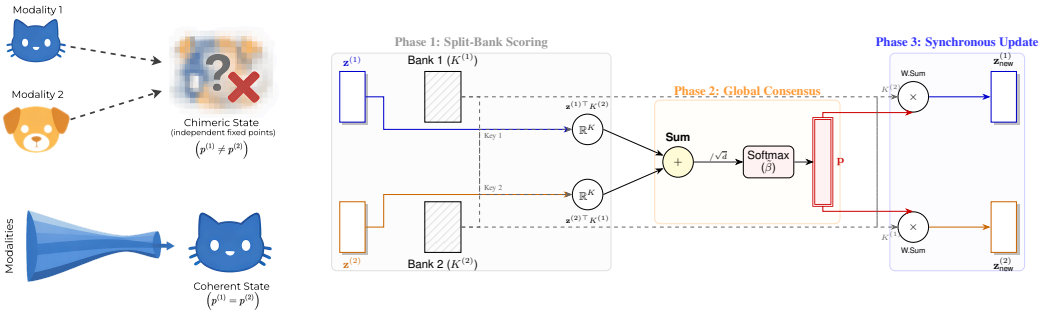

Figure 1: The mTAM framework. *(Left)* The binding problem in multi-modal retrieval. Under decoupled attention, each modality maintains an independent distribution over prototypes, allowing different modalities to settle into conflicting attractors (chimeric states). mTAM projects all modalities onto a shared energy landscape, constraining them to a single coherent basin. *(Right)* Computational graph of one CSA update step (shown for L=2 modalities). Queries $z^{(a)}$ are scored against cross-modal key banks; the resulting per-edge similarities are summed into a global score vector S_μ; a single softmax produces the shared distribution $\mathbf{p}$; all layers are updated synchronously using $\mathbf{p}$.

hull of the memory banks. ***(ii)*** *Since $\mathcal{E}$ is real-analytic (composed of exponential and polynomial operations), it satisfies the Kurdyka–Łojasiewicz inequality (Bolte et al., 2014). Consequently, the trajectory has finite arc length $\sum_t \|z_{t+1} - z_t\| < \infty$ and converges to a single stationary point z^*.*

The energy decreases monotonically ($\Delta\mathcal{E} \leq -\frac{1}{2}\|z_{t+1} - z_t\|^2$), trajectories are bounded, and the KŁ property ensures convergence of each trajectory to a single stationary point rather than oscillation among limit points. Each such fixed point defines a coherent multi-modal retrieval that can be read off directly. Standard cross-attention has no global energy and offers no comparable convergence guarantee.

3 Capacity Theory: 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), in which the graph topology controls the effective noise level seen by the retrieval mechanism.

Graph-lifted Hilbert space. Define the *edge-lifted Hilbert space* $\mathcal{H}_\mathcal{E} := \bigoplus_{(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \mathbb{R}^d$ with lifting operators $(PZ)_{ab} := \sqrt{\bar{A}_{ab}}\, z^{(a)}$ and $(QK_\mu)_{ab} := \sqrt{\bar{A}_{ab}}\, k_\mu^{(b)}$. Then $S_\mu(\{z\}) = \langle PZ, QK_\mu\rangle_{\mathcal{H}_\mathcal{E}}$ is a single inner product in $\mathcal{H}_\mathcal{E}$. The state lift P is an isometry ($P^\top P = I$) by row-stochasticity of $\bar{A}$; the pattern lift satisfies $Q^\top Q = \text{diag}(\bar{d}^{\text{src}}) \otimes I_d$ ($\bar{d}^{\text{src}}_b := \sum_a \bar{A}_{ab}$): heavily-queried banks are expanded, lightly-queried banks compressed. The graph thus reappears as a metric distortion of the pattern ensemble, reducing mTAM to a Modern Hopfield Network in a graph-warped space. This reduction is exact (Appendix): one can import the single-modal MHN capacity machinery by replacing the original patterns with the lifted patterns $\widehat{\Xi}_\mu := QK_\mu$; topology enters exclusively through the Gram factor $Q^\top Q = \text{diag}(\bar{d}^{\text{src}}) \otimes I_d$.

Effective topological variance. When distractor keys are drawn from $\mathcal{N}(\mathbf{0}, \Gamma \otimes I_d)$ with cross-modal covariance Γ ($\Gamma_{aa} = 1$), this metric distortion propagates directly into the noise statistics.

Theorem 2 (Effective Topological Variance Σ^2_{eff})**.** *The consensus score of a generic distractor ν is Gaussian: $S_\nu(Z) \sim \mathcal{N}(0, \Sigma^2_{\text{eff}}(Z))$, with*

$$\Sigma^2_{\text{eff}}(Z) = d \cdot \text{Tr}(M\,\mathsf{Q})\,, \qquad M := \bar{A}\,\Gamma\,\bar{A}^\top, \tag{5}$$

where $\mathsf{Q}_{ij} := d^{-1}\langle z^{(i)}, z^{(j)}\rangle$ is the query overlap matrix. The matrix $M = \bar{A}\,\Gamma\,\bar{A}^\top$ encodes the joint effect of graph topology (through $\bar{A}$) and cross-modal correlations (through Γ) on the distractor noise. For sub-Gaussian keys with covariance Γ, the normalized score $S_\nu/\Sigma_{\text{eff}}$ converges to $\mathcal{N}(0,1)$ as $d \to \infty$ (Berry–Esseen).

Balanced graphs yield small Σ^2_{eff} (low noise, high capacity); imbalanced graphs inflate it. Since M plays the role of an effective noise covariance, choosing regular or doubly-stochastic topologies reduces $\text{Tr}(M\,\mathsf{Q})$ at a fixed edge budget, suggesting a principled criterion for graph design.

Critical load. Let $s := d^{-1}S_{\text{signal}}$ (deterministic target score density), $\rho_{\text{eff}} := \Sigma^2_{\text{eff}}/d$, and $K = \exp(\alpha_{\text{tot}} \cdot Ld)$. By Gaussian extreme-value scaling, the noise floor grows as $\sqrt{2\alpha_{\text{tot}} L\rho_{\text{eff}}} \cdot d$. The

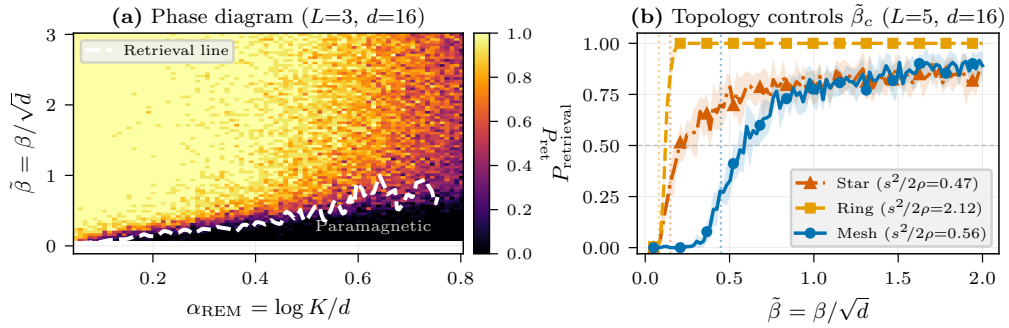

Figure 2: Retrieval landscape and topology of mTAM. (a) Empirical retrieval diagram in the $(\alpha_{\mathrm{REM}}, \tilde{\beta})$ plane for a fully connected graph (L=3, d=16). Colour encodes retrieval probability P_{ret}. The dashed white curve is the predicted retrieval boundary (Eq. 6). Two regimes are visible: diffuse (low $\tilde{\beta}$, high entropy) and retrieval (upper-left, $P_{\mathrm{ret}} \approx 1$). **(b)** Retrieval probability vs. $\tilde{\beta}$ for three topologies at fixed exponential load α_{REM}=0.4 (L=5, d=16). The Ring topology (lowest Σ^2_{eff}) retrieves at the lowest $\tilde{\beta}$, followed by Mesh and Star, confirming that graph topology controls the effective noise level (Theorem 5). Vertical dotted lines mark the predicted retrieval thresholds.

critical load at which noise overwhelms the signal is:

$$\alpha_{\mathrm{tot},c} = \frac{s^2}{2\,L\,\rho_{\mathrm{eff}}} = \frac{\left[\mathrm{Tr}(\bar{A}^\top \Gamma_{\mathrm{align}})\right]^2}{2L\,\mathrm{Tr}(M\,\mathsf{Q})}, \tag{6}$$

where $(\Gamma_{\mathrm{align}})_{ab} := d^{-1}\langle z^{(a)}, k^{(b)}_{\mu^*}\rangle$ encodes the edgewise query-target alignment.

Simplified regime and the harmonic in-degree. Under weakly correlated queries ($\mathsf{Q} \approx I$), isotropic noise ($\Gamma = I$), and homogeneous alignment ($\gamma_{ab} \approx \bar{\gamma}$), Eq. 6 simplifies to:

$$\alpha_{\mathrm{tot},c} \;\approx\; \frac{\bar{\gamma}^2}{2}\,k_{\mathrm{harm}}, \qquad k_{\mathrm{harm}} := \frac{L}{\sum_{a=1}^{L} 1/d_a^{\mathrm{in}}}, \tag{7}$$

where k_{harm} is the *harmonic mean* of in-degrees. A single in-degree-poor node bottlenecks the entire system; a k-regular graph achieves $k_{\mathrm{harm}} = k$. Note that $\bar{\gamma}$ depends on cross-modal key alignment. With *correlated* keys ($\gamma_{ab} \approx \bar{\gamma}$ on all edges), capacity grows with k_{harm} (each additional edge contributes coherent signal). With *independent* keys, the retrieval fixed point yields $\gamma_{ab} \approx \bar{A}_{ab} = 1/d_a^{\mathrm{in}}$, so equation 6 reduces to $\alpha^{(\mathrm{ind})}_{\mathrm{tot},c} = 1/(2\,k_{\mathrm{harm}})$: capacity decreases with connectivity because signal dilution through row-normalization dominates noise averaging.

At finite $\tilde{\beta}$, a noise-condensation threshold at $\tilde{\beta}_* = \sqrt{2L\alpha_{\mathrm{tot}}/\rho_{\mathrm{eff}}}$ separates a diffuse (high-entropy) regime from a retrieval regime; the resulting retrieval boundary, analogous to the REM phase structure, is validated empirically in Figure 2.

4 Experiments

All simulations use normalized keys ($\|k^{(b)}_\mu\| = \sqrt{d}$) and synchronous CCCP iterations. Synthetic data is used by design: it allows precise control over signal-to-noise conditions and exact ground-truth verification of the theoretical predictions.

Experiment 1: Retrieval landscape. We swept $\tilde{\beta}$ and exponential load $\alpha_{\mathrm{REM}} = \log K/d$ (per-dimension; note $\alpha_{\mathrm{REM}} = L \cdot \alpha_{\mathrm{tot}}$) on a fully connected graph (L=3, d=16; Figure 2a). The system exhibits a clear retrieval transition: at low $\tilde{\beta}$, $\mathbf{p}$ is diffuse; above a critical $\tilde{\beta}$, it locks onto the correct attractor. The predicted retrieval boundary (Eq. 6) closely tracks the empirical threshold. Panel (b) validates Theorem 5: at fixed load (α_{REM}=0.4, L=5, d=16), Ring retrieves at lower $\tilde{\beta}$ than Mesh or Star, matching the predicted ordering of Σ^2_{eff}.

Experiment 2: Chimera stress test. We initialized a system (L=3, K=12, d=128, $\tilde{\beta} \approx 3.98$) in a deliberate three-way chimeric state: each layer near a different stored pattern, with layer 0 receiving a slight boost (1.2×) toward target A. Four methods were compared (Figure 3): **CSA** (ours, Product-of-Experts logit sum); **DEC** (decoupled, each layer uses own $p^{(a)}$); **PF** (probability fusion, updates use $p_{\mathrm{avg}} = L^{-1}\sum_a p^{(a)}$, Mean-of-Experts); **LF** (late fusion, fuses only at last step).

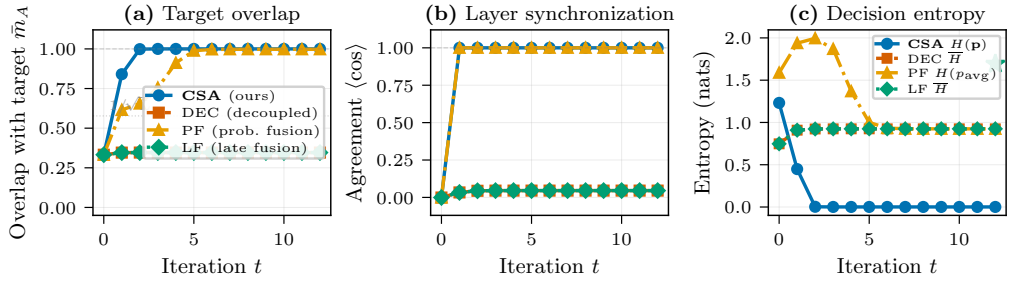

Figure 3: Chimera stress test: CSA vs. baselines (L=3, K=12, d=128, $\tilde{\beta}\approx$3.98; all methods start from the same three-way chimeric initialization). **(a)** Mean cosine overlap with target pattern A over iterations. CSA reaches >99% overlap by t=2; PF requires $t\geq$6; DEC and LF plateau at ∼0.35. **(b)** Layer synchronization (mean pairwise cosine similarity between layer states). CSA achieves full agreement in one step; DEC and LF remain desynchronized. **(c)** Entropy of the consensus/decision distribution. CSA drives entropy to near-zero by t=2; PF initially increases entropy (evidence dilution) before slowly concentrating.

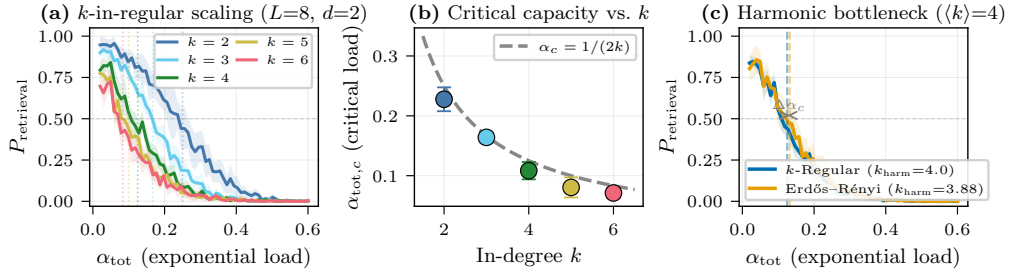

Figure 4: Topological control of capacity (L=8, d=2, $\tilde{\beta}$=1.6; independent keys, no self-loops; regime $\alpha_{\text{tot},c} = 1/(2k_{\text{harm}})$). **(a)** Retrieval probability P_{ret} vs. exponential load α_{tot} for k-in-regular graphs. Higher k shifts the retrieval collapse leftward, consistent with per-edge signal dilution $\bar{\gamma} \sim 1/k$. **(b)** Empirical critical load α_c (dots) vs. theoretical prediction $1/(2k)$ (dashed). Theory and experiment agree quantitatively. **(c)** 4-Regular (k_{harm}=4) vs. Erdős–Rényi at matching mean degree 4 ($k_{\text{harm}}\approx$2.74): lower k_{harm} yields higher α_c, confirming the harmonic bottleneck.

CSA resolves the chimera in one step and reaches >99% overlap with the target by t=2: the PoE logit-sum amplifies even a slight single-layer bias into a peaked consensus. PF synchronizes but retrieves slowly ($t\geq$6), because averaging initially erases the strongest evidence. DEC and LF remain chimeric throughout (overlap ∼0.35): without a shared distribution, each layer's attention is anchored to its own initialization and cross-modal evidence never transfers, so the three independent attractors are all locally stable. Only CSA achieves both synchronization and immediate selection of the correct attractor in this stress test.

Experiment 3: Topology-controlled capacity. Using independent keys without self-loops places the setup in the regime $\alpha_{\text{tot},c} = 1/(2k_{\text{harm}})$. On directed k-in-regular graphs (L=8, d=2, $\tilde{\beta}$=1.6, $k \in \{2,\ldots,6\}$), the critical load decreases with k as predicted (each edge dilutes the per-edge signal via row-normalization); empirical thresholds $0.25, 0.15, 0.10, 0.08, 0.07$ closely track $1/(2k)$ (Figure 4a–b). A more discriminating test compares two topologies at equal mean degree 4 but different harmonic means (Figure 4c): a 4-regular graph (k_{harm}=4, α_c=0.125) vs. an Erdős–Rényi graph ($k_{\text{harm}}\approx$2.74, $\alpha_c\approx$0.182). As predicted, the ER graph's lower harmonic mean yields a higher critical load; empirical thresholds confirm that k_{harm}, not the arithmetic mean degree, is the correct topological summary statistic. The reason is direct: in the independent-key regime $\gamma_{ab} = 1/d_a^{\text{in}}$, so low-degree nodes carry stronger per-edge signal; a heterogeneous topology retains these high-signal vertices, lifting the average capacity above the regular-graph prediction.

5 Conclusion, Limitations & Future Work

We introduced mTAM, an energy-based architecture whose CSA mechanism eliminates chimeric states by construction, collapsing the equilibrium binding space from K^L to K. The CCCP structure guarantees that each trajectory converges to a stationary point (Theorem 1), and the graph-lifting construction yields an explicit, topology-dependent critical load $\alpha_{\text{tot},c}$ governed by Σ^2_{eff} and k_{harm} (Theorem 5). Synthetic experiments confirm pronounced retrieval transitions consistent with the REM-type prediction, one-step chimera resolution where all decoupled baselines fail, and quantitative agreement on topology-dependent capacity scaling.

Integration and positioning. The forward pass (linear projections + one global softmax) is fully differentiable, so mTAM extends the Modern Hopfield framework (Ramsauer et al., 2020) to multi-modal consensus with standard backpropagation. The energy functional $\mathcal{E}$ admits a Lyapunov interpretation that suggests a natural regularizing role: by penalizing states outside the convex hull of the memory banks, it may encourage coherent multi-modal bindings without auxiliary contrastive objectives. Unlike architectures that mitigate binding failures via directional streams (Tsai et al., 2019), bottleneck latents (Nagrani et al., 2021), or contrastive losses (Radford et al., 2021), CSA enforces consistency structurally through the energy surface.

Limitations & future work. This paper characterizes theoretical retrieval limits under synthetic distributional assumptions. Validating mTAM with learned representations on real-world benchmarks, extending the capacity analysis to non-Gaussian key distributions, and scaling to large pattern sets via hierarchical or sparse retrieval are natural next steps. We also note that CCCP guarantees monotonic descent but convergence to a *local* minimum: the fixed point z^* is initialization-dependent, and starting from a chimeric state CSA finds the nearest coherent attractor, which need not be globally optimal. Characterizing the basin structure under practical learned representations is an open and important question.

REFERENCES

Elena Agliari, Andrea Alessandrelli, Adriano Barra, Martino Salomone Centonze, and Federico Ricci-Tersenghi. Generalized hetero-associative neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(1):013302, 2025a.

Elena Agliari, Andrea Alessandrelli, Adriano Barra, Martino Salomone Centonze, and Federico Ricci-Tersenghi. Networks of neural networks: more is different. *arXiv preprint arXiv:2501.16789*, 2025b.

Andrea Alessandrelli, Adriano Barra, Andrea Ladiana, Andrea Lepre, and Federico Ricci-Tersenghi. Supervised and unsupervised protocols for hetero-associative neural networks. *Physica A: Statistical Mechanics and its Applications*, 676:130871, 2025a. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2025.130871. URL https://www.sciencedirect.com/science/article/pii/S0378437125005230.

Andrea Alessandrelli, Adriano Barra, Andrea Ladiana, Andrea Lepre, and Federico Ricci-Tersenghi. Beyond disorder: Unveiling cooperativeness in multidirectional associative memories. In *New Frontiers in Associative Memories*, 2025b.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Abhinav Joshi, Naman Gupta, Jinang Shah, Binod Bhattarai, Ashutosh Modi, and Danail Stoyanov. Generalized product-of-experts for learning multimodal representations in noisy environments. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 83–93, 2022.

Bart Kosko. Bidirectional associative memories. *IEEE Transactions on Systems, man, and Cybernetics*, 18(1):49–60, 2002.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/eaae339c4d89fc102edd9dbdb6a28915-Paper.pdf.

Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Physical Review Letters*, 132(7):077301, 2024.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.

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

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

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

Alan L Yuille and Anand Rangarajan. The concave-convex procedure. *Neural computation*, 15(4): 915–936, 2003.

A mTAM/CSA Pseudocode

Algorithm 1 mTAM Dynamics via Consensus Split-Bank Attention (CSA)

Require:

Initial modality states $\mathcal{Z}^{(0)} = \{z_0^{(1)}, \dots, z_0^{(L)}\}$ with $z^{(a)} \in \mathbb{R}^d$
Memory Banks $\mathcal{K} = \{K^{(1)}, \dots, K^{(L)}\}$ with $K^{(b)} \in \mathbb{R}^{d \times K}$
Graph Adjacency $A \in \{0,1\}^{L \times L}$ where $A_{ab} = 1 \iff (a \leftarrow b)$
Inverse temperature β, Time steps T

Ensure: Converged states $\mathcal{Z}^{(T)}$, Global Consensus p

Initialization:
Compute row-normalized adjacency: $\overline{A}_{ab} \leftarrow A_{ab} / \sum_k A_{ak}$
Set effective temperature: $\tilde{\beta} \leftarrow \beta / \sqrt{d}$
Pre-compute Effective Keys (Graph-Mixing)
for each modality $a \in \{1, \dots, L\}$ **do**
    $\tilde{K}^{(a)} \leftarrow \sum_{b=1}^{L} \overline{A}_{ab} K^{(b)}$
end for

Optimization Loop (CCCP):
for $t = 0$ to $T - 1$ **do**
    // 1. Compute Global Pattern Scores
    Compute scores vector $S \in \mathbb{R}^K$:
    $S_\mu \leftarrow \sum_{a=1}^{L} \langle z_t^{(a)}, \tilde{K}_{\cdot,\mu}^{(a)} \rangle \quad \forall \mu \in \{1, \dots, K\}$
    // 2. Compute Shared Consensus Distribution
    $p \leftarrow \text{softmax}(\tilde{\beta} S)$
    $p_\mu \leftarrow \frac{\exp(\tilde{\beta} S_\mu)}{\sum_{\nu=1}^{K} \exp(\tilde{\beta} S_\nu)}$
    // 3. Synchronous Update (Gradient of Energy)
    for each modality $a \in \{1, \dots, L\}$ **do**
        $z_{t+1}^{(a)} \leftarrow \tilde{K}^{(a)} p = \sum_{\mu=1}^{K} p_\mu \tilde{K}_{\cdot,\mu}^{(a)}$
    end for
end for
return $\mathcal{Z}^{(T)} = \{z_T^{(1)}, \dots, z_T^{(L)}\}, p$

The pre-computation loop runs in $O(L^2Kd)$; each CCCP iteration costs $O(LKd)$. Total cost for T iterations: $O((L^2 + TL)Kd)$, which matches standard cross-attention in the typical regime $T \ll L$.

B Notation Table

Table 1: Principal notation used throughout the paper.

Symbol	Definition / meaning
L	Number of modalities
K	Number of stored prototypes
d	Per-modality embedding dimension
$\mathbf{z}^{(a)} \in \mathbb{R}^d$	State of modality a
$\mathbf{Z} \in \mathbb{R}^{Ld}$	Global state (stacked)
$K^{(b)} \in \mathbb{R}^{d\times K}$	Key bank of modality b
$\mathbf{k}_\mu^{(b)} \in \mathbb{R}^d$	μ-th column of $K^{(b)}$
$A \in \{0,1\}^{L\times L}$	Binary interaction adjacency
$\mathcal{E}_{\text{het}} \subseteq [L]\times[L]$	Edge set of the interaction graph
$\bar{A}_{ab} = A_{ab}/d_a^{\text{in}}$	Row-normalised adjacency
$d_a^{\text{in}} = \sum_b A_{ab}$	In-degree of modality a
$\bar{d}_b^{\text{src}} = \sum_a \bar{A}_{ab}$	Normalised source degree of bank b
β	Raw inverse temperature
$\tilde{\beta} = \beta/\sqrt{d}$	Scaled (effective) inverse temperature
$S_\mu(\mathbf{Z})$	Global pattern score for prototype μ
$\mathbf{p}(\mathbf{Z}) \in \Delta^{K-1}$	Consensus posterior over prototypes
$\mathcal{E}(\mathbf{Z})$	mTAM energy (DC Lyapunov function)
$\mathcal{U}(\mathbf{Z})$	Confinement term ($\sum_a \frac{1}{2}\|\mathbf{z}^{(a)}\|^2$)
$\mathcal{V}(\mathbf{Z})$	Log-sum-exp interaction term
$\mathcal{H}_\mathcal{E} = \bigoplus_{(a\leftarrow b)} \mathbb{R}^d$	Graph-lifted Hilbert space
P	State-lifting map $\mathbb{R}^{Ld} \to \mathcal{H}_\mathcal{E}$
Q	Pattern-lifting map $\mathbb{R}^{Ld} \to \mathcal{H}_\mathcal{E}$
$\alpha_{\text{tot}} = \frac{\log K}{Ld}$	Exponential load (per dimension)
$\Gamma \in \mathbb{R}^{L\times L}$	Cross-modal key covariance
Σ^2_{eff}	Effective topological variance
$\mathsf{Q}_{ij} = \frac{1}{d}\langle \mathbf{z}^{(i)}, \mathbf{z}^{(j)}\rangle$	Query overlap matrix
$\alpha_{\text{tot},c}$	Critical load threshold
$k_{\text{harm}} = L/\sum_a 1/d_a^{\text{in}}$	Harmonic mean in-degree
$\theta \in [0,1)$	KL (Łojasiewicz) exponent at $\mathbf{Z}^*$

C End-to-End Trainable mTAM

This section provides a detailed discussion of how the mTAM/CSA mechanism can be integrated into end-to-end trainable deep learning architectures, including the precise mapping to standard Transformer components, gradient flow analysis, and a proof-of-concept experiment.

C.1 Architecture Mapping: CSA ↔ Transformer Attention

The CSA mechanism maps directly to standard multi-head attention Vaswani et al. (2017), and to Hopfield layers Ramsauer et al. (2020) for the single-modality case. We detail the correspondence here.

Standard Transformer Attention (single modality). Given queries $Q = W_Q x$, keys $K = W_K x$, values $V = W_V x$:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V. \tag{8}$$

Modern Hopfield Network (Ramsauer et al., 2020). In Hopfield terminology, the "state pattern" (query) ξ attends to "stored patterns" (keys) $\{\xi_\mu\}_{\mu=1}^K$ and retrieves "pattern projections" (values):

$$\xi_{\text{new}} = \sum_\mu p_\mu \, \xi_\mu, \quad p = \text{softmax}(\beta \cdot \xi^\top \Xi), \tag{9}$$

where $\Xi = [\xi_1 \mid \ldots \mid \xi_K]$ is the stored pattern matrix and β plays the role of $1/\sqrt{d_k}$. This is equivalent to single-head attention with the "scaling" parameter β.

mTAM / CSA (this work). CSA extends this to L modalities by introducing per-modality query projections $W_Q^{(a)} \in \mathbb{R}^{d \times N_a}$, per-modality key banks $K^{(b)} \in \mathbb{R}^{d \times K}$, per-modality value banks $V^{(a)} \in \mathbb{R}^{d \times K}$, and a graph-weighted coupling structure $\bar{A}$.
The full mapping is:

Transformer	**Hopfield Layer**	**mTAM / CSA**
$Q = W_Q x$	ξ (state pattern)	$z^{(a)} = W_Q^{(a)} x^{(a)}$
$K = W_K x$	Ξ (stored patterns)	$K^{(b)}$ (learnable banks)
$V = W_V x$	V (pattern proj.)	$V^{(a)}$ (learnable banks)
softmax$(QK^\top/\sqrt{d})$	$p = \text{softmax}(\beta \xi^\top \Xi)$	$p = \text{softmax}(\tilde{\beta} \cdot S)$
per-token	*per-query*	**shared consensus** $\forall a$

The key difference is that CSA computes a *single global* consensus distribution $p \in \Delta^{K-1}$ shared across all L modalities (via the Product-of-Experts score $S_\mu = \sum_{a,b} \bar{A}_{ab} \langle z^{(a)}, k_\mu^{(b)} \rangle$), as opposed to L independent attention distributions.

Differentiability and Gradient Flow. The forward pass of each CCCP iteration consists of:

1. Linear projections: $z^{(a)} = W_Q^{(a)} x^{(a)}$ (differentiable)
2. Score computation: $S_\mu = \sum_{a,b} \bar{A}_{ab} \langle z^{(a)}, k_\mu^{(b)} \rangle$ (bilinear $\Rightarrow$ differentiable)
3. Consensus: $p = \text{softmax}(\tilde{\beta} \cdot S)$ (differentiable)
4. Update: $z_{\text{new}}^{(a)} = \tilde{K}^{(a)} p$ (linear $\Rightarrow$ differentiable)
5. Value readout: $x_{\text{out}}^{(a)} = V^{(a)} p$ (linear $\Rightarrow$ differentiable)

Every operation is composed of standard differentiable primitives. When T CCCP iterations are unrolled, gradients propagate through the entire chain via the standard chain rule (backpropagation through time), exactly as in the iterated attention updates of Ramsauer et al. Ramsauer et al. (2020).

mTAMBlock: Drop-in Multi-Modal Transformer Block. We provide a reference implementation (in the companion code, at `https://github.com/andrea-ladiana/mtam_layers`) of an `mTAMBlock` that follows the standard Transformer encoder block pattern:

$$z^{(a)} = \text{LayerNorm}(W_Q^{(a)} \cdot x^{(a)}) \quad \text{(query projection)} \tag{10}$$

$$\hat{x}^{(a)} = W_O^{(a)} \cdot V^{(a)} \cdot p_{\text{CSA}} \quad \text{(CSA readout + output proj.)} \tag{11}$$

$$x^{(a)} \leftarrow x^{(a)} + \text{Dropout}(\hat{x}^{(a)}) \quad \text{(residual connection)} \tag{12}$$

$$x^{(a)} \leftarrow x^{(a)} + \text{Dropout}(\text{FFN}(\text{LayerNorm}(x^{(a)}))) \quad \text{(feed-forward sub-layer)} \tag{13}$$

Multiple `mTAMBlocks` can be stacked into an `mTAMEncoder`, analogous to `nn.TransformerEncoder` in PyTorch. We verify that gradients flow correctly through all components, including the K and V banks and the query projections $W_Q^{(a)}$.

Energy as Architectural Regularizer. The CSA energy $E(z)$ (Eq. 4) provides a natural Lyapunov regularizer during training. Adding $\lambda E(z)$ to the task loss encourages the model to converge toward coherent multi-modal bindings, without requiring explicit contrastive or alignment losses. This is a structural advantage over auxiliary-loss-based approaches such as CLIP Radford et al. (2021).

C.2 Proof-of-Concept: Multi-Modal Bit Pattern Binding

To empirically validate end-to-end trainability, we designed a multi-modal extension of the bit pattern Multiple Instance Learning (MIL) experiment from Ramsauer et al. Ramsauer et al. (2020).

Task. Each sample consists of $L = 3$ "modality bags", each containing I binary bit-vector instances (b bits each). In positive samples, one "signal" instance is implanted into *each* modality bag; the signals across modalities are semantically correlated (drawn from the same class out of 8 signal classes). Negative samples contain no signal instances. The task is binary classification: detect whether the multi-modal input contains a coherently bound signal.

This directly tests the binding hypothesis: a single modality's signal can be confused with random noise, but the *joint* evidence from all L modalities (if correctly fused) disambiguates.

We evaluate in two regimes of increasing difficulty:

- **Easy** (b=8, I=16): With $2^8 = 256$ possible patterns for 8 signal classes, each modality's signal is highly distinctive. Per-modality detectors suffice.
- **Hard** (b=4, I=32): With only $2^4 = 16$ possible patterns for 8 signal classes, signals are nearly indistinguishable from random noise within any single modality (probability of a random instance matching a signal pattern: $\approx 50\%$). Detection requires cross-modal evidence aggregation, exactly the binding problem CSA is designed to solve.

Models. We compare four architectures, all trained end-to-end with AdamW (lr $= 10^{-3}$, gradient clipping at 1.0) for 150 epochs on 2048 samples (following the same protocol as Ramsauer et al.):

- **mTAM-CSA**: Per-modality instance-level attention pooling $\rightarrow$ CSA consensus binding (T=3, K=16, β=4) $\rightarrow$ classifier.
- **Hopfield-DEC**: Independent Hopfield association per modality with separate key/value banks, then late concatenation.
- **Hopfield-CONCAT**: Single Hopfield association on concatenated modality features (early fusion).
- **MLP**: Simple MLP baseline on flattened raw concatenated features.

All Hopfield-based methods share the same embedding dimension (d=32), number of prototypes (K=16), and classifier head architecture. The MLP has a comparable or larger parameter count.

Gradient Flow Verification. Before reporting classification results, we verify that gradients propagate correctly through the entire mTAMBlock. Using a 3-modality mTAMBlock (d=32, K=16, T=3 CCCP iterations), we measured the ℓ_2 gradient norms after a single backward pass:

Component	$\|\nabla\|_2$
K banks (CSA keys)	0.40
V banks (CSA values)	7.26
$W_Q^{(a)}$ (query projections)	0.04–0.05
$W_O^{(a)}$ (output projections)	7.5–8.0
FFN sub-layer	49.9–55.5

All components receive non-zero gradients, confirming that backpropagation flows end-to-end through the T=3 unrolled CCCP iterations, including through the shared consensus softmax. The K banks receive smaller but non-zero gradients because they influence the output only indirectly through the consensus distribution p.

Results.

	Easy (b=8, I=16)		**Hard** (b=4, I=32)	
Method	**Best (%)**	**Final (%)**	**Best (%)**	**Final (%)**
mTAM-CSA (ours)	99.2	98.7	**68.0**	**65.4**
Hopfield-DEC	**99.4**	**99.0**	62.8	60.5
Hopfield-CONCAT	62.9	58.8	60.3	56.4
MLP (baseline)	52.9	50.4	50.7	49.6

Table 2: Multi-Modal Bit Pattern MIL (L=3, 8 signal classes, 150 epochs, AdamW). *Easy regime*: high per-modality SNR; DEC's independent detectors slightly outperform CSA. *Hard regime*: low per-modality SNR; the ranking reverses: mTAM-CSA's consensus binding provides the best accuracy. In both regimes, early fusion (CONCAT) and MLP fail, confirming that associative memory with per-modality processing is essential.

Discussion. The dual-regime results reveal the complementary nature of the CSA mechanism:

1. **End-to-end trainability:** mTAM-CSA trains smoothly in both regimes via standard backpropagation. Gradients flow through the T=3 unrolled CCCP iterations without difficulty, as confirmed numerically above.
2. **Associative memory is essential:** In both regimes, the attention-over-prototypes methods (CSA and DEC) substantially outperform early fusion and MLP, validating the Hopfield-based architecture for multi-modal MIL.
3. **Regime-dependent advantage:** In the *easy* regime, per-modality signals are independently informative, DEC's per-modality detectors suffice and the classifier head can learn to fuse them, achieving 99.4% vs. CSA's 99.2%. In the *hard* regime, the ranking reverses: mTAM-CSA (68.0%) outperforms DEC (62.8%) because the consensus distribution pools evidence across modalities *during retrieval*, not just at decision time. This is precisely the scenario targeted by CSA's design.
4. **Binding guarantees:** Beyond accuracy, CSA provides a structural guarantee absent from DEC: in the retrieval experiments of Section 4, DEC remains permanently chimeric (overlap $\sim$0.35) under chimeric initialization, while CSA resolves chimeras in 1 step.

This experiment provides empirical evidence that the theoretical properties of CSA (convergence, chimera elimination) are preserved when the mechanism is embedded in a trainable architecture, and that the consensus advantage is amplified in low-SNR settings where cross-modal binding is critical.

C.3 Computational Scaling

To assess practical applicability, we measured CSA's wall-clock time per forward iteration as a function of embedding dimension d and number of stored prototypes K (Table 3). The benchmark uses the fully vectorized PyTorch implementation of `mTAM_Core` on CPU, with batch size B=32 and L=3 modalities in a fully connected graph.

d	$K = 100$	$K = 500$	$K = 1000$	$K = 5000$	$K = 10000$
128	0.25	0.79	1.44	3.78	6.84
256	0.35	0.57	0.92	6.75	12.30
512	2.49	1.59	3.05	16.50	25.42
768	2.26	2.91	5.19	19.28	35.29
1024	0.51	3.87	6.94	22.49	51.00
2048	1.26	5.69	10.53	51.19	112.28
4096	2.44	11.84	19.45	120.23	227.88

Table 3: CSA Forward Pass Scaling (ms/iter). Wall-clock time per forward step for L=3, batch size =32. The cost scales linearly in both d and K, remaining practical even at d=4096 and K=10,000.

The computational cost scales linearly with both d and K. For realistic embedding dimensions (e.g., CLIP: d=768; ViT-Large: d=1024) and large memory banks (K=10,000), a single forward iteration requires less than 51 ms. This confirms that the CSA architecture is computationally viable for integration into modern multi-modal pipelines.

C.4 When do chimeras emerge? A theoretical perspective.

Multimodal chimera states can be viewed as a retrieval-time instance of the broader binding problem Greff et al. (2020): the model must not only detect which features are present, but must also assign cross-modal evidence to the same latent concept. Existing multimodal architectures address this issue in different ways, for example through directional cross-modal attention Tsai et al. (2019), bottleneck-mediated fusion Nagrani et al. (2021), or auxiliary alignment objectives Radford et al. (2021). However, whenever concept scores remain modality-specific and are converted into separate posteriors over a shared set of prototypes, discordant modal assignments remain admissible states of the retrieval map.

Formal setup. Consider L modalities, each equipped with a memory bank $K^{(a)} \in \mathbb{R}^{d\times K}$, and a shared set of K multimodal prototypes indexed by μ. The relevant chimera-susceptible class is not "multimodal transformers" in full generality, but rather architectures that maintain separate per-modality posteriors over a common prototype set. In our notation, this corresponds exactly to the DEC baseline, where each modality computes its own score vector and its own softmax distribution.

Chimera susceptibility of independent posteriors. A chimera is not generic disagreement, but a fixed-point inconsistency in which different modalities assign maximal score to different prototypes. Let

$$S_\mu^{(a)} = \langle q^{(a)}, k_\mu^{(a)} \rangle \tag{14}$$

denote the score assigned by modality a to prototype μ. A chimera between two concepts μ and ν occurs whenever

$$S_\mu^{(a)} > S_\nu^{(a)} \qquad \text{but} \qquad S_\nu^{(b)} > S_\mu^{(b)} \tag{15}$$

for some pair of modalities (a, b). Under the planted-REM assumptions of Section 3, distractor scores are Gaussian with variance Σ_a^2. In the high-load regime, when noise fluctuations become comparable to or larger than the signal gap in at least one modality, independent retrieval errors need not align across modalities. Under a simplifying independence approximation, this suggests the heuristic scaling

$$P(\text{chimera}) \;\approx\; 1 - \prod_{a=1}^{L} P(\text{modality } a \text{ selects the same prototype}), \tag{16}$$

which becomes non-negligible near capacity. Equivalently, if each modality is correct with probability $1-\varepsilon$, then the probability of unanimous agreement is only $(1-\varepsilon)^L$. Thus even modest per-modality error rates are amplified by the number of modalities into an appreciable probability of cross-modal disagreement.

Resolution via shared consensus. In practice, the chimera condition arises naturally whenever the memory contains semantically nearby but distinct multimodal concepts. In such cases, different modalities may favor different prototypes, yielding a stable but semantically inconsistent assignment under decoupled retrieval. CSA removes this failure mode because all modalities are updated from the same shared posterior

$$p_\mu \propto \exp\Big(\tilde{\beta} \sum_{a,b} \bar{A}_{ab} \, \langle z^{(a)}, k_\mu^{(b)} \rangle\Big), \tag{17}$$

so the selected prototype is common to all modalities by construction. Bottlenecks and alignment losses may encourage coherence, but they do not in general exclude discordant modal argmax assignments at the level of the retrieval map. By contrast, in CSA such assignments lie outside the image of the update itself.

This is the sense in which Figure 3 should be interpreted: not as an isolated toy failure mode, but as the controlled manifestation of a structural ambiguity inherent to decoupled multimodal retrieval, and of its removal under a shared-consensus dynamics.

D Formal Derivations: Introduction & Convergence Analysis

D.1 Structural Elimination of Chimeric States via Consensus

Proposition 1 (Structural Elimination of Chimeric States via Consensus)**.** *Consider the mTAM architecture with Consensus Split-Bank Attention (CSA). Let $\pi^{(a)} \in \Delta^{K-1}$ denote the effective*

attention distribution driving the update of the a-th modality, and let $\mathcal{M}^{(a)} := \arg\max_{\mu \in [K]} \pi_\mu^{(a)}$ be the set of dominant prototype indices for that modality. In the CSA framework, for any state configuration $\{\mathbf{z}\}$, the modal decision sets are identical across all layers:

$$\mathcal{M}^{(a)} = \mathcal{M}^{(b)} \quad \forall a, b \in 1, \dots, L. \tag{18}$$

Consequently, the system admits no chimeric states (configurations where $\mathcal{M}^{(a)} \cap \mathcal{M}^{(b)} = \emptyset$).

Proof. By definition of the CSA mechanism, the attention distribution $\mathbf{p}$ is computed from the unique global score vector $\mathbf{S}(\{\mathbf{z}\})$:

$$\mathbf{p} = \mathrm{softmax}(\tilde{\beta}\mathbf{S}), \quad \text{with } p_\mu = \frac{e^{\tilde{\beta} S_\mu}}{\sum_\nu e^{\tilde{\beta} S_\nu}}. \tag{19}$$

The update rule for the a-th layer is explicitly defined as:

$$\mathbf{z}_{\text{new}}^{(a)} = \sum_{b:(a \leftarrow b) \in \mathcal{E}_{\text{het}}} \bar{A}_{ab} K^{(b)} \mathbf{p}. \tag{20}$$

This implies that the effective attention weight assigned to the μ-th prototype by the a-th modality is exactly $\pi_\mu^{(a)} \equiv p_\mu$. Crucially, the functional dependence $\mu \mapsto p_\mu$ is independent of the target layer index a. Since $\pi_\mu^{(a)} = p_\mu$ for all a, it follows that:

$$\mathcal{M}^{(a)} = \arg\max_\mu p_\mu = \mathcal{M}_{\text{global}} \quad \forall a. \tag{21}$$

The set of maximizing indices is thus invariant under permutation of the modality index. A chimeric state, defined as a configuration where distinct modalities possess disjoint sets of maximizers ($\mathcal{M}^{(a)} \cap \mathcal{M}^{(b)} = \emptyset$), is therefore algebraically impossible by construction. □

D.2 DC Decomposition of the mTAM Energy

Lemma 1 (DC Decomposition of the mTAM Energy)**.** *Let the configuration state be denoted by the collection $\mathbf{z} \coloneqq \{\mathbf{z}^{(a)}\}_{a=1}^L \in \mathbb{R}^{D_{\text{tot}}}$, where $D_{\text{tot}} = \sum_a d_a$. The global energy functional $\mathcal{E}(\mathbf{z})$ admits a Difference-of-Convex (DC) decomposition:*

$$\mathcal{E}(\mathbf{z}) = \mathcal{U}(\mathbf{z}) - \mathcal{V}(\mathbf{z}), \tag{22}$$

where $\mathcal{U} : \mathbb{R}^{D_{\text{tot}}} \to \mathbb{R}$ is strictly convex and $\mathcal{V} : \mathbb{R}^{D_{\text{tot}}} \to \mathbb{R}$ is convex.

Proof. We identify the two components of the energy function as:

$$\mathcal{U}(\mathbf{z}) \coloneqq \sum_{a=1}^{L} \frac{1}{2} \|\mathbf{z}^{(a)}\|^2, \tag{23}$$

$$\mathcal{V}(\mathbf{z}) \coloneqq \tilde{\beta}^{-1} \log \sum_{\mu=1}^{K} \exp\left(\tilde{\beta}\, S_\mu(\mathbf{z})\right) - C_{\text{shift}}. \tag{24}$$

Convexity of $\mathcal{U}$. The function $\mathcal{U}(\mathbf{z})$ is a sum of squared Euclidean norms. The Hessian of $\mathcal{U}$ with respect to the global state vector $\mathbf{z}$ is the identity matrix $I_{D_{\text{tot}}}$. Since $\nabla^2 \mathcal{U} = I \succ 0$, $\mathcal{U}$ is strictly convex (and strongly convex with parameter 1).

Convexity of $\mathcal{V}$. The convexity of $\mathcal{V}$ follows from the properties of the Log-Sum-Exp function composed with linear mappings. Recall that the global pattern score $S_\mu(\mathbf{z})$ is defined as:

$$S_\mu(\mathbf{z}) = \sum_{a,b} \bar{A}_{ab} \langle \mathbf{z}^{(a)}, \mathbf{k}_\mu^{(b)} \rangle. \tag{25}$$

Observe that $S_\mu(\mathbf{z})$ is a linear function of the state variables $\mathbf{z}^{(a)}$. We can express the vector of scores $\mathbf{S}(\mathbf{z}) = [S_1(\mathbf{z}), \dots, S_K(\mathbf{z})]^\top$ as a linear transformation:

$$\mathbf{S}(\mathbf{z}) = \mathcal{K}\mathbf{z}, \tag{26}$$

where $\mathcal{K}$ is a linear operator representing the weighted connectivity and dot-products with key matrices.

Now, consider the function $f(\mathbf{y}) = \tilde{\beta}^{-1} \log \sum_{\mu=1}^{K} \exp(\tilde{\beta}\, y_\mu)$, which is the standard Log-Sum-Exp (LSE) function with effective inverse temperature $\tilde{\beta} = \beta/\sqrt{d}$. It is a well-known result in convex analysis that the LSE function is convex on $\mathbb{R}^K$ (its Hessian is positive semi-definite, corresponding to the covariance matrix of the softmax distribution). Since $\mathcal{V}(\mathbf{z}) = f(\mathcal{K}\mathbf{z}) - C_{\text{shift}}$, and the composition of a convex function with a linear map preserves convexity, it follows that $\mathcal{V}(\mathbf{z})$ is convex on $\mathbb{R}^{D_{\text{tot}}}$. Thus, $\mathcal{E}$ is the difference of a strictly convex function $\mathcal{U}$ and a convex function $\mathcal{V}$. □

Definition 1 (mTAM Dynamical System)**.** Consider a system with L layers, state variables $\mathbf{z}^{(a)} \in \mathbb{R}^{d_a}$, and finite key matrices $K^{(b)}$ with bounded norms. Let the edge set $\mathcal{E}_{\text{het}}$ and inverse temperature $\beta > 0$ be fixed. Define $\tilde{\beta} := \beta/\sqrt{d}$ as the effective inverse temperature. The energy landscape is defined as:

$$\mathcal{E}(\mathbf{z}) \;=\; \underbrace{\sum_{a=1}^{L} \frac{1}{2} \|\mathbf{z}^{(a)}\|^2}_{\mathcal{U}(\mathbf{z})\text{ (Convex)}} \;-\; \underbrace{\tilde{\beta}^{-1} \log \sum_{\mu=1}^{K} \exp\left(\tilde{\beta}\, S_\mu(\mathbf{z})\right)}_{\mathcal{V}(\mathbf{z})\text{ (Convex)}} + C_{\text{shift}}. \tag{27}$$

Since $\mathcal{U}$ and $\mathcal{V}$ are both convex functions, $\mathcal{E}$ is a Difference-of-Convex (DC) function.

D.3 Gradient Consistency of the Consensus Update

Lemma 2 (Gradient Consistency of the Consensus Update)**.** *Consider the interaction term $\mathcal{V}(\mathbf{z})$ defined in Lemma 1. The partial gradient of $\mathcal{V}$ with respect to the state of layer a, denoted as $\mathbf{z}^{(a)}$, is given by:*

$$\nabla_{\mathbf{z}^{(a)}} \mathcal{V}(\mathbf{z}) = \sum_{b:(a \leftarrow b) \in \mathcal{E}_{\text{het}}} \bar{A}_{ab}\, K^{(b)}\, \mathbf{p}(\mathbf{z}), \tag{28}$$

where $\mathbf{p}(\mathbf{z}) = \text{softmax}(\tilde{\beta}\, \mathbf{S}(\mathbf{z}))$ is the consensus probability vector with $\tilde{\beta} = \beta/\sqrt{d}$. Consequently, the synchronous mTAM update rule is exactly equivalent to the CCCP prescription:

$$\mathbf{z}^{(a)}_{\text{new}} = \nabla_{\mathbf{z}^{(a)}} \mathcal{V}(\mathbf{z}_{\text{old}}). \tag{29}$$

Proof. Recall the definition $\mathcal{V}(\mathbf{z}) = \tilde{\beta}^{-1} \log \sum_{\nu=1}^{K} \exp(\tilde{\beta}\, S_\nu(\mathbf{z}))$. We apply the chain rule to compute the gradient with respect to $\mathbf{z}^{(a)}$. Let $Z = \sum_{\nu=1}^{K} \exp(\tilde{\beta}\, S_\nu(\mathbf{z}))$ be the partition function. The derivative is:

$$\frac{\partial \mathcal{V}}{\partial \mathbf{z}^{(a)}} = \tilde{\beta}^{-1} \frac{1}{Z} \sum_{\mu=1}^{K} \frac{\partial}{\partial \mathbf{z}^{(a)}} \left(e^{\tilde{\beta}\, S_\mu(\mathbf{z})}\right). \tag{30}$$

Computing the inner derivative:

$$\frac{\partial}{\partial \mathbf{z}^{(a)}} e^{\tilde{\beta}\, S_\mu(\mathbf{z})} = \tilde{\beta}\, e^{\tilde{\beta}\, S_\mu(\mathbf{z})} \cdot \nabla_{\mathbf{z}^{(a)}} S_\mu(\mathbf{z}). \tag{31}$$

The factor $\tilde{\beta}$ cancels with the prefactor $\tilde{\beta}^{-1}$. We identify the softmax probability $p_\mu = \frac{e^{\tilde{\beta}\, S_\mu}}{Z}$. Thus:

$$\nabla_{\mathbf{z}^{(a)}} \mathcal{V}(\mathbf{z}) = \sum_{\mu=1}^{K} p_\mu(\mathbf{z})\, \nabla_{\mathbf{z}^{(a)}} S_\mu(\mathbf{z}). \tag{32}$$

Now we evaluate the gradient of the global score S_μ defined in Eq. 1. Since S_μ is a weighted sum over all edges in the graph, only the terms involving $\mathbf{z}^{(a)}$ (where layer a is the target of an edge) contribute to the gradient:

$$S_\mu(\mathbf{z}) = \sum_{u,v} \bar{A}_{uv}\, (\mathbf{z}^{(u)})^\top \mathbf{k}^{(v)}_\mu. \tag{33}$$

Taking the gradient with respect to $\mathbf{z}^{(a)}$ isolates the terms where $u = a$:

$$\nabla_{\mathbf{z}^{(a)}} S_\mu(\mathbf{z}) = \sum_{b:(a \leftarrow b) \in \mathcal{E}_{\text{het}}} \bar{A}_{ab}\, \mathbf{k}^{(b)}_\mu. \tag{34}$$

Substituting this back into Eq. 32 and swapping the summation order (linearity):

$$\nabla_{\mathbf{z}^{(a)}}\mathcal{V}(\mathbf{z}) = \sum_{\mu=1}^{K} p_\mu(\mathbf{z}) \left(\sum_{b:(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \bar{A}_{ab}\, \mathbf{k}_\mu^{(b)} \right) = \sum_{b:(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \bar{A}_{ab} \sum_{\mu=1}^{K} \mathbf{k}_\mu^{(b)} p_\mu(\mathbf{z}). \tag{35}$$

Recognizing that $\sum_{\mu=1}^{K} \mathbf{k}_\mu^{(b)} p_\mu$ is the matrix-vector product $K^{(b)}\mathbf{p}$, we obtain:

$$\nabla_{\mathbf{z}^{(a)}}\mathcal{V}(\mathbf{z}) = \sum_{b:(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \bar{A}_{ab}\, K^{(b)}\mathbf{p}(\mathbf{z}). \tag{36}$$

This matches the right-hand side of the mTAM update rule. Since $\nabla_{\mathbf{z}^{(a)}}\mathcal{U}(\mathbf{z}_{\text{new}}) = \mathbf{z}_{\text{new}}^{(a)}$ (from Lemma 1), the update satisfies the CCCP condition $\nabla\mathcal{U}(\mathbf{z}_{\text{new}}) = \nabla\mathcal{V}(\mathbf{z}_{\text{old}})$. □

D.4 Monotonic Descent and Stationarity

Theorem 3 (Monotonic Descent and Stationarity)**.** *Let $\{\mathbf{z}_t\}_{t\geq 0}$ be the sequence of states generated by the mTAM update rule starting from any bounded initialization $\mathbf{z}_0$. The following properties hold:*

1. ***Strict Monotonicity:*** *The energy is strictly non-increasing along the trajectory, i.e., $\mathcal{E}(\mathbf{z}_{t+1}) \leq \mathcal{E}(\mathbf{z}_t)$. The equality holds if and only if $\mathbf{z}_t$ is a stationary point of $\mathcal{E}$.*
2. ***Boundedness:*** *The trajectory is contained within a compact set $\mathcal{B} \subset \mathbb{R}^{D_{\text{tot}}}$. Specifically, for $t \geq 1$, each layer state lies in the weighted Minkowski sum of the memory banks' convex hulls:*

   $$\mathbf{z}_t^{(a)} \in \sum_{b:(a\leftarrow b)} \bar{A}_{ab}\, \text{conv}\{\mathbf{k}_\mu^{(b)}\}_{\mu=1}^{K},$$

   hence $\|\mathbf{z}_t^{(a)}\| \leq M_a^{\text{in}} := \sum_b \bar{A}_{ab} \max_\mu \|\mathbf{k}_\mu^{(b)}\| \leq \max_b \max_\mu \|\mathbf{k}_\mu^{(b)}\|$.
3. ***Stationary Limit Points:*** *The set of accumulation points $\omega(\mathbf{z}_0)$ is non-empty. Every $\mathbf{z}^* \in \omega(\mathbf{z}_0)$ satisfies the fixed-point equation:*

   $$\mathbf{z}^{(a),*} = \sum_{b:(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \bar{A}_{ab}\, K^{(b)}\, \text{softmax}\left(\tilde{\beta}\, \mathbf{S}(\mathbf{z}^*)\right), \tag{37}$$

   where $\mathbf{S}(\mathbf{z}) \in \mathbb{R}^K$ is the vector of global scores and $\tilde{\beta} = \beta/\sqrt{d}$.

Proof. We prove the properties sequentially using the DC structure and update consistency established previously.

1. Strict Monotonicity. Let $\Delta\mathcal{E}_t := \mathcal{E}(\mathbf{z}_{t+1}) - \mathcal{E}(\mathbf{z}_t)$. Recall that $\mathcal{E}(\mathbf{z}) = \mathcal{U}(\mathbf{z}) - \mathcal{V}(\mathbf{z})$ with $\mathcal{U}(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|^2$ and $\mathcal{V}$ convex. By the convexity of $\mathcal{V}$, we have the subgradient inequality:

$$\mathcal{V}(\mathbf{z}_{t+1}) \geq \mathcal{V}(\mathbf{z}_t) + \langle \nabla\mathcal{V}(\mathbf{z}_t), \mathbf{z}_{t+1} - \mathbf{z}_t \rangle. \tag{38}$$

Substituting this into the energy difference:

$$\Delta\mathcal{E}_t = \mathcal{U}(\mathbf{z}_{t+1}) - \mathcal{U}(\mathbf{z}_t) - [\mathcal{V}(\mathbf{z}_{t+1}) - \mathcal{V}(\mathbf{z}_t)] \tag{39}$$

$$\leq \frac{1}{2}\|\mathbf{z}_{t+1}\|^2 - \frac{1}{2}\|\mathbf{z}_t\|^2 - \langle \nabla\mathcal{V}(\mathbf{z}_t), \mathbf{z}_{t+1} - \mathbf{z}_t \rangle. \tag{40}$$

Using the update rule $\mathbf{z}_{t+1} = \nabla\mathcal{V}(\mathbf{z}_t)$, the inner product becomes $\langle \mathbf{z}_{t+1}, \mathbf{z}_{t+1} - \mathbf{z}_t \rangle = \|\mathbf{z}_{t+1}\|^2 - \langle \mathbf{z}_{t+1}, \mathbf{z}_t \rangle$. Substituting back:

$$\Delta\mathcal{E}_t \leq \frac{1}{2}\|\mathbf{z}_{t+1}\|^2 - \frac{1}{2}\|\mathbf{z}_t\|^2 - \|\mathbf{z}_{t+1}\|^2 + \langle \mathbf{z}_{t+1}, \mathbf{z}_t \rangle \tag{41}$$

$$= -\frac{1}{2}\|\mathbf{z}_{t+1}\|^2 - \frac{1}{2}\|\mathbf{z}_t\|^2 + \langle \mathbf{z}_{t+1}, \mathbf{z}_t \rangle. \tag{42}$$

Recognizing the expansion of the squared Euclidean distance $-\frac{1}{2}\|\mathbf{a} - \mathbf{b}\|^2 = -\frac{1}{2}\|\mathbf{a}\|^2 - \frac{1}{2}\|\mathbf{b}\|^2 + \langle \mathbf{a}, \mathbf{b} \rangle$, we obtain the *sufficient decrease* condition:

$$\mathcal{E}(\mathbf{z}_{t+1}) - \mathcal{E}(\mathbf{z}_t) \leq -\frac{1}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2. \tag{43}$$

Moreover, equation 43 shows that $\mathcal{E}(\mathbf{z}_{t+1}) < \mathcal{E}(\mathbf{z}_t)$ holds strictly if and only if $\mathbf{z}_{t+1} \neq \mathbf{z}_t$ (i.e., the dynamics is not at a fixed point). This rules out non-trivial limit cycles.

2. Boundedness. We establish boundedness directly from the update rule, independent of energy coercivity. For any layer a, the update is:

$$\mathbf{z}_{t+1}^{(a)} = \sum_{b:(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \bar{A}_{ab}\, K^{(b)}\mathbf{p}_t, \tag{44}$$

where $\mathbf{p}_t \in \Delta^{K-1}$ is a probability vector ($\sum_\mu p_{\mu,t} = 1, p_{\mu,t} \geq 0$). Let $R_b \coloneqq \max_\mu \|\mathbf{k}_\mu^{(b)}\|$ be the maximum norm of a key in bank b. By the convexity of the norm:

$$\|K^{(b)}\mathbf{p}_t\| = \left\|\sum_{\mu=1}^{K} p_{\mu,t}\mathbf{k}_\mu^{(b)}\right\| \leq \sum_{\mu=1}^{K} p_{\mu,t}\|\mathbf{k}_\mu^{(b)}\| \leq R_b \sum_{\mu=1}^{K} p_{\mu,t} = R_b. \tag{45}$$

Therefore, for any bounded initialization $\mathbf{z}_0$, the trajectory enters after one update and then remains in a compact set for all $t \geq 1$; in particular, the state norm is uniformly bounded by the structural constants of the memory banks:

$$\|\mathbf{z}_{t+1}^{(a)}\| \leq \sum_{b:(a\leftarrow b)} \bar{A}_{ab}\, \|K^{(b)}\mathbf{p}_t\| \leq \sum_{b:(a\leftarrow b)} \bar{A}_{ab}\, R_b \leq \max_b R_b. \tag{46}$$

The entire trajectory $\{\mathbf{z}_t\}_{t\geq 1}$ is thus contained in a compact set $\Omega \subset \mathbb{R}^{D_{\text{tot}}}$. Since $\mathcal{E}$ is continuous on this compact set, it is also bounded below.

3. Convergence to Stationary Points. Since the sequence $\{\mathcal{E}(\mathbf{z}_t)\}$ is decreasing and bounded below, it converges. Summing the sufficient decrease inequality (Eq. 43) over t:

$$\sum_{t=0}^{\infty} \frac{1}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 \leq \mathcal{E}(\mathbf{z}_0) - \lim_{t\to\infty} \mathcal{E}(\mathbf{z}_t) < \infty. \tag{47}$$

This implies that the step size vanishes: $\|\mathbf{z}_{t+1} - \mathbf{z}_t\| \to 0$ as $t \to \infty$. Since $\{\mathbf{z}_t\}$ lies in a compact set Ω, there exists a convergent subsequence $\mathbf{z}_{t_j} \to \mathbf{z}^*$ as $j \to \infty$. Define the update map $T(\mathbf{z}) \coloneqq \nabla\mathcal{V}(\mathbf{z})$. The map is continuous. Applying the update to the subsequence:

$$\mathbf{z}_{t_j+1} = T(\mathbf{z}_{t_j}). \tag{48}$$

Taking the limit $j \to \infty$:

- LHS: $\mathbf{z}_{t_j+1} = (\mathbf{z}_{t_j+1} - \mathbf{z}_{t_j}) + \mathbf{z}_{t_j} \to \mathbf{0} + \mathbf{z}^* = \mathbf{z}^*$.
- RHS: $T(\mathbf{z}_{t_j}) \to T(\mathbf{z}^*)$ by continuity.

Thus, $\mathbf{z}^* = T(\mathbf{z}^*) = \nabla\mathcal{V}(\mathbf{z}^*)$. Substituting this into the gradient definition:

$$\nabla\mathcal{E}(\mathbf{z}^*) = \nabla\mathcal{U}(\mathbf{z}^*) - \nabla\mathcal{V}(\mathbf{z}^*) = \mathbf{z}^* - T(\mathbf{z}^*) = \mathbf{0}. \tag{49}$$

This proves that every accumulation point of the trajectory is a stationary point of the energy. □

D.5 Energy is constant on the limit set

Lemma 3 (Energy is constant on the limit set)**.** *Assume the hypotheses of Theorem 3. Let $\omega(\mathbf{z}_0)$ denote the set of limit points of $\{\mathbf{z}_t\}_{t\geq 0}$ and let $\mathcal{E}_t \coloneqq \mathcal{E}(\mathbf{z}_t)$. Then $\{\mathcal{E}_t\}_{t\geq 0}$ converges to a finite limit $\mathcal{E}_\infty$, and*

$$\mathcal{E}(\mathbf{z}) = \mathcal{E}_\infty \qquad \forall\, \mathbf{z} \in \omega(\mathbf{z}_0).$$

In particular, $\mathcal{E}$ is constant on $\omega(\mathbf{z}_0)$.

Proof. $\{\mathcal{E}_t\}$ is non-increasing. As proved before the trajectory remains in a compact sublevel set, hence $\mathcal{E}$ is bounded below along $\{\mathbf{z}_t\}$. Therefore $\mathcal{E}_t \downarrow \mathcal{E}_\infty$ for some finite $\mathcal{E}_\infty$. Now take any $\mathbf{z} \in \omega(\mathbf{z}_0)$. There exists a subsequence $\mathbf{z}_{t_k} \to \mathbf{z}$. By continuity of $\mathcal{E}$,

$$\mathcal{E}(\mathbf{z}) = \lim_{k\to\infty} \mathcal{E}(\mathbf{z}_{t_k}) = \lim_{k\to\infty} \mathcal{E}_{t_k} = \mathcal{E}_\infty,$$

which proves the claim. □

D.6 Global Strong Convergence via Łojasiewicz Property

Theorem 4 (Global Strong Convergence via Łojasiewicz Property)**.** *The energy function $\mathcal{E}(z)$ is real-analytic on its domain and the trajectory lies within a compact set $\mathcal{B}$ (by Theorem 3). Therefore, $\mathcal{E}$ satisfies the Kurdyka-Łojasiewicz (KL) property on any compact set. Consequently:*

1. *The trajectory $\{\mathbf{z}_t\}_{t\geq 0}$ has finite length: $\sum_{t=0}^{\infty} \|\mathbf{z}_{t+1} - \mathbf{z}_t\| < \infty$.*
2. *The entire sequence converges to a* single *stationary point $\mathbf{z}^*$ (which depends on initialization), without oscillation.*
3. *Asymptotic convergence rate: $\mathcal{E}(\mathbf{z}_t) \to \mathcal{E}(\mathbf{z}^*)$ with a rate that depends on the Łojasiewicz exponent $\theta \in [0, 1)$ of the critical point.*

Proof. We rely on the convergence framework for descent algorithms satisfying the Kurdyka-Łojasiewicz (KL) property, following Attouch et al. (2013).

The energy function $\mathcal{E}(\mathbf{z})$ is formed by sums and compositions of polynomials (squared norms, linear scores), exponentials, and logarithms. Note that the argument of the logarithm, $\sum_\mu \exp(\tilde{\beta}\, S_\mu(\mathbf{z}))$, is strictly positive for all $\mathbf{z} \in \mathbb{R}^{D_{\text{tot}}}$. Since the logarithm is real-analytic on $(0, +\infty)$ and compositions of real-analytic functions remain real-analytic, $\mathcal{E}$ is a real-analytic function on its entire domain.

Let $\omega(\mathbf{z}_0)$ be the set of accumulation points of the bounded sequence $\{\mathbf{z}_t\}_{t\geq 0}$. Since $\mathcal{E}(\mathbf{z}_t)$ is monotonically decreasing and bounded below, it converges to a value $\mathcal{E}_\infty$. By Lemma 3, $\mathcal{E}$ takes the constant value $\mathcal{E}_\infty$ on the limit set $\omega(\mathbf{z}_0)$. According to the **Uniform KL Property** (Bolte et al., 2014), since $\omega(\mathbf{z}_0)$ is a compact set on which $\mathcal{E}$ is constant, there exist $\epsilon > 0$, $\eta > 0$, and an exponent $\theta \in [0, 1)$ such that for all $\mathbf{z}$ in the neighborhood $\{\mathbf{z} : \text{dist}(\mathbf{z}, \omega(\mathbf{z}_0)) < \epsilon\} \cap \{\mathbf{z} : \mathcal{E}_\infty < \mathcal{E}(\mathbf{z}) < \mathcal{E}_\infty + \eta\}$, the following inequality holds:

$$|\mathcal{E}(\mathbf{z}) - \mathcal{E}_\infty|^\theta \leq C\|\nabla\mathcal{E}(\mathbf{z})\|. \tag{50}$$

Recall the DC structure $\mathcal{E} = \mathcal{U} - \mathcal{V}$ with $\mathcal{U}(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|^2$. The gradient is $\nabla\mathcal{E}(\mathbf{z}_t) = \mathbf{z}_t - \nabla\mathcal{V}(\mathbf{z}_t)$. Since the update rule is defined as $\mathbf{z}_{t+1} = \nabla\mathcal{V}(\mathbf{z}_t)$, we obtain the fundamental identity relating the gradient norm to the step size:

$$\nabla\mathcal{E}(\mathbf{z}_t) = \mathbf{z}_t - \mathbf{z}_{t+1}. \tag{51}$$

We prove that the trajectory has finite length, i.e., $\sum_{t=0}^{\infty} \|\mathbf{z}_{t+1} - \mathbf{z}_t\| < \infty$. If $\mathcal{E}(\mathbf{z}_t) = \mathcal{E}_\infty$ for some t, the algorithm terminates. Assume $\mathcal{E}(\mathbf{z}_t) > \mathcal{E}_\infty$ for all t. Since $\mathbf{z}_t$ approaches the compact set $\omega(\mathbf{z}_0)$, for sufficiently large $t > T$, the uniform KL inequality 50 applies. Consider the concave function $\phi(s) = \frac{C}{1-\theta}(s - \mathcal{E}_\infty)^{1-\theta}$. By concavity of ϕ (using $\phi(x) - \phi(y) \geq \phi'(x)(x - y)$ for $x > y$):

$$\phi(\mathcal{E}(\mathbf{z}_t)) - \phi(\mathcal{E}(\mathbf{z}_{t+1})) \geq C(\mathcal{E}(\mathbf{z}_t) - \mathcal{E}_\infty)^{-\theta}(\mathcal{E}(\mathbf{z}_t) - \mathcal{E}(\mathbf{z}_{t+1})). \tag{52}$$

From the KL inequality 50, we rearrange terms to bound the inverse gradient:

$$(\mathcal{E}(\mathbf{z}_t) - \mathcal{E}_\infty)^{-\theta} \geq \frac{1}{C\|\nabla\mathcal{E}(\mathbf{z}_t)\|}. \tag{53}$$

From the monotone theorem, we use the energy decrease bound:

$$\mathcal{E}(\mathbf{z}_t) - \mathcal{E}(\mathbf{z}_{t+1}) \geq \frac{1}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2. \tag{54}$$

Combining these three relations:

$$\begin{aligned}
\phi(\mathcal{E}(\mathbf{z}_t)) - \phi(\mathcal{E}(\mathbf{z}_{t+1})) &\geq \frac{C}{C\|\nabla\mathcal{E}(\mathbf{z}_t)\|} \cdot \frac{1}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 && (55)\\
&= \frac{1}{2\|\mathbf{z}_t - \mathbf{z}_{t+1}\|}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 \quad \text{(using Eq. 51)} && (56)\\
&= \frac{1}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|. && (57)
\end{aligned}$$

Rearranging gives $\|\mathbf{z}_{t+1} - \mathbf{z}_t\| \leq 2[\phi(\mathcal{E}(\mathbf{z}_t)) - \phi(\mathcal{E}(\mathbf{z}_{t+1}))]$. Summing over $t \geq T$:

$$\sum_{t=T}^{M} \|\mathbf{z}_{t+1} - \mathbf{z}_t\| \leq 2\sum_{t=T}^{M}[\phi(\mathcal{E}(\mathbf{z}_t)) - \phi(\mathcal{E}(\mathbf{z}_{t+1}))] \leq 2\phi(\mathcal{E}(\mathbf{z}_T)). \tag{58}$$

The series converges, implying the trajectory has finite length.

Since the trajectory has finite length, $\{\mathbf{z}_t\}$ is a Cauchy sequence in a complete space, and thus converges to a **single limit point** $\mathbf{z}^*$ (as opposed to oscillating within the limit set). From Theorem 3, we established that all accumulation points are stationary. Therefore, $\mathbf{z}^*$ is a stationary point of $\mathcal{E}$, and $\nabla\mathcal{E}(\mathbf{z}^*) = 0$. □

E Formal Derivations: Capacity Scaling

E.1 Proof of $P^\top P = I_{Ld}$

Proof. Since $\bar{A}$ is row-stochastic (i.e., $\sum_b \bar{A}_{ab} = 1$ for all a), we compute:

$$\begin{aligned}\|P\mathbf{Z}\|_{\mathcal{H}_\mathcal{E}}^2 &= \sum_{(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \|(P\mathbf{Z})_{ab}\|^2 = \sum_{(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \bar{A}_{ab}\,\|\mathbf{z}^{(a)}\|^2 \\ &= \sum_{a=1}^{L} \|\mathbf{z}^{(a)}\|^2 \sum_{b:(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \bar{A}_{ab} = \sum_{a=1}^{L} \|\mathbf{z}^{(a)}\|^2 \cdot 1 = \|\mathbf{Z}\|^2. \end{aligned} \tag{59}$$

Thus, P is an isometry and $P^\top P = I_{Ld}$. This holds *exactly* for any row-stochastic $\bar{A}$, without requiring in-degree regularity. □

Note that $P^\top$ denotes the adjoint operator with respect to the specific inner products defined on $\mathbb{R}^{Ld}$ (standard Euclidean) and $\mathcal{H}_\mathcal{E}$ (weighted graph-lifted inner product). This is the first rigorous point where the graph influences the geometry: with the row-stochastic normalization $\bar{A}$, the lifted norm $\|P\mathbf{Z}\|^2$ equals $\|\mathbf{Z}\|^2$ exactly. Topology enters through $Q^\top Q$ (which depends on $\bar{d}^{\text{src}}$) and through the cross-modal covariance structure of the patterns, not through variable reweighting of the state.

E.2 Proof of $Q^\mathsf{T} Q = \operatorname{diag}(\bar{d}^{\text{src}}) \otimes I_d$

Proof. Let $V = [v^{(1)}, \dots, v^{(L)}]^\mathsf{T} \in \mathbb{R}^{Ld}$. By definition $(QV)_{ab} = \sqrt{\bar{A}_{ab}}\, v^{(b)}$. Therefore

$$\|QV\|_{\mathcal{H}_\mathcal{E}}^2 = \sum_{(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \bar{A}_{ab}\,\|v^{(b)}\|^2 = \sum_{b=1}^{L} \Big(\sum_{a=1}^{L} \bar{A}_{ab}\Big)\|v^{(b)}\|^2 = \sum_{b=1}^{L} \bar{d}_b^{\text{src}}\,\|v^{(b)}\|^2 = V^\mathsf{T}\big(\operatorname{diag}(\bar{d}^{\text{src}})\otimes I_d\big)V,$$

which proves $Q^\mathsf{T} Q = \operatorname{diag}(\bar{d}^{\text{src}}) \otimes I_d$. □

E.3 Algebraic equivalence to an MHN on $\mathcal{H}_\mathcal{E}$

Lemma 4 (Algebraic equivalence to an MHN on $\mathcal{H}_\mathcal{E}$)**.** *Define the lifted state $\widehat{\mathbf{Z}} := P\mathbf{Z} \in \mathcal{H}_\mathcal{E}$ and lifted patterns $\widehat{\boldsymbol{\Xi}}_\mu := Q\mathbf{K}_\mu \in \mathcal{H}_\mathcal{E}$. Then the mTAM energy is exactly the MHN softmax energy on $\mathcal{H}_\mathcal{E}$ with patterns $\{\widehat{\boldsymbol{\Xi}}_\mu\}_{\mu=1}^K$ and effective inverse temperature $\tilde{\beta} = \beta/\sqrt{d}$. Moreover, the dynamics act on the constrained manifold $\widehat{\mathbf{Z}} \in \operatorname{Im}(P)$ (i.e., edge–lifted states whose edge blocks agree whenever they share the same* target *layer).*

Proof. Let $\mathcal{H}_\mathcal{E} := \bigoplus_{(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \mathbb{R}^d$ with inner product $\langle U, V\rangle_{\mathcal{H}_\mathcal{E}} := \sum_{(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \langle U_{ab}, V_{ab}\rangle_{\mathbb{R}^d}$. By definition of the weighted lifting operators P, Q, the lifted state $\widehat{Z} = PZ$ has blocks $\widehat{Z}_{ab} = \sqrt{\bar{A}_{ab}}\, z^{(a)}$, while the lifted pattern $\widehat{\Xi}_\mu = QK_\mu$ has blocks $\widehat{\Xi}_{\mu,ab} = \sqrt{\bar{A}_{ab}}\, k_\mu^{(b)}$. Hence

$$\langle PZ, QK_\mu\rangle_{\mathcal{H}_\mathcal{E}} = \sum_{(a\leftarrow b)\in\mathcal{E}_{\text{het}}} \bar{A}_{ab}\,\langle z^{(a)}, k_\mu^{(b)}\rangle = S_\mu(\{z\}),$$

where the last equality is exactly the definition of the weighted global score in equation 1. Setting $\tilde{\beta} := \beta/\sqrt{d}$, we obtain $\tilde{\beta}\langle PZ, QK_\mu\rangle_{\mathcal{H}_\mathcal{E}} = \tilde{\beta}\, S_\mu(\{z\})$, so the softmax/LSE term in the MHN energy on $\mathcal{H}_\mathcal{E}$ coincides with the mTAM interaction energy (up to the same additive constant, if present). If also contains a quadratic confinement term, the equivalence holds verbatim by adding $\frac{1}{2}\|PZ\|^2$ to both sides. Since $\|PZ\|^2 = \|Z\|^2$ by the row-stochastic property of $\bar{A}$, the confinement terms match exactly.

Assume the graph covers all targets, i.e. $\forall a\, \exists b : (a \leftarrow b) \in \mathcal{E}_{\text{het}}$, so that P is injective. Then $\text{Im}(P) \subset \mathcal{H}_{\mathcal{E}}$ is the linear subspace of *target-consistent* edge-lifted states:

$$U \in \text{Im}(P) \quad \Longleftrightarrow \quad \exists \{z^{(a)}\}_{a=1}^{L} \text{ s.t. } U_{ab} = \sqrt{\bar{A}_{ab}}\, z^{(a)}\ \forall (a \leftarrow b) \in \mathcal{E}_{\text{het}}.$$

Equivalently, one can verify target-consistency intrinsically:

$$\frac{U_{ab}}{\sqrt{\bar{A}_{ab}}} = \frac{U_{ac}}{\sqrt{\bar{A}_{ac}}} \quad \forall (a \leftarrow b), (a \leftarrow c) \in \mathcal{E}_{\text{het}} \text{ with } \bar{A}_{ab}, \bar{A}_{ac} > 0.$$

Since the mTAM evolution is defined on the base variable $Z(t) \in \mathbb{R}^{Ld}$, the lifted trajectory $\widehat{Z}(t) := PZ(t)$ satisfies $\widehat{Z}(t) \in \text{Im}(P)$ for all t. Therefore, mTAM is equivalent to an MHN on $\mathcal{H}_{\mathcal{E}}$ *restricted to* the invariant subspace $\text{Im}(P)$ (equivalently: to the projected MHN dynamics onto $\text{Im}(P)$). □

F Formal Derivations: Scaling Capacity via REM Analysis

F.1 Effective Topological Variance

Theorem 5 (Effective Topological Variance)**.** *Consider the mTAM model defined on the edge-lifted Hilbert space $\mathcal{H}_{\mathcal{E}}$. Let the distractor keys $\{K_\nu\}_{\nu \neq \mu}$ follow a multivariate Gaussian distribution centered at the origin:*

$$K_\nu \sim \mathcal{N}(\mathbf{0}, \Gamma \otimes I_d), \tag{60}$$

where $\Gamma \in \mathbb{R}^{L \times L}$ is a symmetric positive semidefinite cross-modal correlation matrix (in particular $\Gamma_{aa} = 1$) and I_d is the identity matrix of dimension d. We stress that $\Gamma_{aa} = 1$ implies $\mathbb{E}\|k_\nu^{(a)}\|^2 = d$, so $\|k_\nu^{(a)}\| = \sqrt{d}(1 + o_p(1))$ by concentration: this is not a spherical constraint, but the standard Gaussian normalization in $\mathbb{R}^d$.

Let $Z = [z^{(1)\top}, \dots, z^{(L)\top}]^\top \in \mathbb{R}^{Ld}$ be the fixed global query state, let $A \in \{0,1\}^{L \times L}$ be the binary adjacency matrix of the interaction graph (with $A_{ab} = 1$ denoting an edge $a \leftarrow b$), and let $\bar{A} \in \mathbb{R}^{L \times L}$ be the row-normalized adjacency matrix.

The consensus score of a generic distractor pattern ν, defined as:

$$S_\nu(Z) = \sum_{a,b=1}^{L} \bar{A}_{ab} \langle z^{(a)}, k_\nu^{(b)} \rangle, \tag{61}$$

is a scalar Gaussian random variable with zero mean, $S_\nu \sim \mathcal{N}(0, \Sigma_{\text{eff}}^2(Z))$, with the effective topological variance given by the quadratic form:

$$\Sigma_{\text{eff}}^2(Z) = Z^\top \left[(\bar{A} \Gamma \bar{A}^\top) \otimes I_d \right] Z. \tag{62}$$

Furthermore, defining the query overlap matrix $\mathsf{Q} \in \mathbb{R}^{L \times L}$ with entries $\mathsf{Q}_{ij} = \frac{1}{d}\langle z^{(i)}, z^{(j)} \rangle$, this variance can be expressed as the trace of the topology-state interaction:

$$\Sigma_{\text{eff}}^2(\mathsf{Q}) = d \cdot \text{Tr}\left((\bar{A} \Gamma \bar{A}^\top) \mathsf{Q} \right). \tag{63}$$

Proof. We proceed by direct construction, utilizing block matrix algebra and the properties of the Kronecker product.

First, we define the *aggregated query vector* $\mathbf{u} \in \mathbb{R}^{Ld}$, where the b-th block $\mathbf{u}^{(b)} \in \mathbb{R}^d$ represents the weighted sum of all queries attending to key bank b:

$$\mathbf{u}^{(b)} := \sum_{a=1}^{L} \bar{A}_{ab} z^{(a)}. \tag{64}$$

In global matrix notation, this aggregation corresponds to a linear transformation of the state Z. Since the coefficient $\bar{A}_{ab}$ connects the query row index a to the key block index b, the operator is given by the transpose of the row-normalized adjacency matrix lifted to the embedding dimension:

$$\mathbf{u} = (\bar{A}^\top \otimes I_d) Z. \tag{65}$$

Consequently, the consensus score $S_\nu(Z)$ can be rewritten as the inner product between the aggregated queries and the global key vector K_ν:

$$S_\nu(Z) = \sum_{b=1}^{L} \langle \mathbf{u}^{(b)}, k_\nu^{(b)} \rangle = \mathbf{u}^\top K_\nu. \tag{66}$$

Since S_ν is a linear transformation of the Gaussian vector K_ν, it is itself a Gaussian random variable. Its mean is trivially zero: $\mathbb{E}[S_\nu] = \mathbf{u}^\top \mathbb{E}[K_\nu] = 0$. The variance is determined by the expectation of the squared score:

$$\begin{aligned} \Sigma^2_{\text{eff}} &= \mathbb{E}\left[\left(\mathbf{u}^\top K_\nu\right) \left(K_\nu^\top \mathbf{u}\right) \right] \\ &= \mathbf{u}^\top \mathbb{E}[K_\nu K_\nu^\top] \mathbf{u}. \end{aligned} \tag{67}$$

Substituting the assumed covariance structure $\mathbb{E}[K_\nu K_\nu^\top] = \Gamma \otimes I_d$, we obtain:

$$\Sigma^2_{\text{eff}} = \mathbf{u}^\top (\Gamma \otimes I_d) \mathbf{u}. \tag{68}$$

Substituting the definition of $\mathbf{u}$ in terms of the state Z yields:

$$\Sigma^2_{\text{eff}} = \left[(\bar{A}^\top \otimes I_d) Z \right]^\top (\Gamma \otimes I_d) \left[(\bar{A}^\top \otimes I_d) Z \right]. \tag{69}$$

We apply the transpose property $(M \otimes I)^\top = M^\top \otimes I$ and the mixed-product property $(\bar{A} \otimes B)(C \otimes D) = (\bar{A}C \otimes BD)$:

$$\begin{aligned} \Sigma^2_{\text{eff}} &= Z^\top (\bar{A} \otimes I_d)(\Gamma \otimes I_d)(\bar{A}^\top \otimes I_d) Z \\ &= Z^\top \left((\bar{A} \Gamma \bar{A}^\top) \otimes I_d \right) Z. \end{aligned} \tag{70}$$

This proves Eq. 62. Finally, using the trace cyclic property $\mathbf{x}^\top M \mathbf{x} = \text{Tr}(M \mathbf{x}\mathbf{x}^\top)$, we expand the quadratic form in terms of blocks:

$$\begin{aligned} \Sigma^2_{\text{eff}} &= \sum_{i,j=1}^{L} (\bar{A} \Gamma \bar{A}^\top)_{ij} \langle z^{(i)}, z^{(j)} \rangle \\ &= d \sum_{i,j=1}^{L} (\bar{A} \Gamma \bar{A}^\top)_{ij} \mathsf{Q}_{ij} = d \cdot \text{Tr}\left((\bar{A} \Gamma \bar{A}^\top) \mathsf{Q} \right), \end{aligned} \tag{71}$$

where we used the definition $\mathsf{Q}_{ij} = d^{-1} \langle z^{(i)}, z^{(j)} \rangle$ and the symmetry $\mathsf{Q} = \mathsf{Q}^\top$. □

F.2 Universality of the distractor score

Lemma 5 (Universality of the distractor score)**.** *Fix a deterministic query state $Z \in \mathbb{R}^{Ld}$ and consider the distractor score*

$$S_\nu(Z) \;:=\; \sum_{a,b=1}^{L} \bar{A}_{ab} \, \langle z^{(a)}, k_\nu^{(b)} \rangle \;=\; \Big\langle (\bar{A}^\top \otimes I_d) Z, \; K_\nu \Big\rangle.$$

Assume that, for each ν, the d coordinate-vectors $\boldsymbol{k}_{\nu,j} := (k_{\nu,j}^{(1)}, \ldots, k_{\nu,j}^{(L)}) \in \mathbb{R}^L$ are i.i.d. over $j \in [d]$ with $\mathbb{E}[\boldsymbol{k}_{\nu,j}] = 0$, $\text{Cov}(\boldsymbol{k}_{\nu,j}) = \Gamma$, and there exists a constant $\kappa_3 < \infty$ such that for every deterministic $a \in \mathbb{R}^L$,

$$\mathbb{E}\big| a^\top \boldsymbol{k}_{\nu,1} \big|^3 \;\le\; \kappa_3 \left(a^\top \Gamma a \right)^{3/2}. \tag{72}$$

(Condition 72 holds, for instance, for Gaussian keys and more generally for sub-Gaussian keys with covariance Γ.)

Let $M := \bar{A} \, \Gamma \, \bar{A}^\top$ and define, for each coordinate $j \in [d]$, the vector $\boldsymbol{z}_j := (z_j^{(1)}, \ldots, z_j^{(L)}) \in \mathbb{R}^L$ and the associated variance contribution $v_j := \boldsymbol{z}_j^\top M \boldsymbol{z}_j$. Then $\Sigma^2_{\text{eff}}(Z) = \sum_{j=1}^d v_j$ as in Theorem 5. If the (Lindeberg) no-dominant-coordinate condition holds,

$$\frac{\max_{1 \le j \le d} v_j}{\sum_{j=1}^d v_j} \xrightarrow[d \to \infty]{} 0, \tag{73}$$

then the normalized score is asymptotically Gaussian:

$$\frac{S_\nu(Z)}{\Sigma_{\text{eff}}(Z)} \xrightarrow[d\to\infty]{} \mathcal{N}(0,1).$$

Moreover, a Berry–Esseen bound yields

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{S_\nu(Z)}{\Sigma_{\text{eff}}(Z)}\le x\right)-\Phi(x)\right| \;\le\; C\,\frac{\sum_{j=1}^d v_j^{3/2}}{\left(\sum_{j=1}^d v_j\right)^{3/2}} \;\le\; C\,\sqrt{\frac{\max_j v_j}{\sum_{j=1}^d v_j}}, \tag{74}$$

where C depends only on κ_3. Consequently, at leading order in d the REM noise floor (and hence the $T=0$ capacity threshold) depends only on $\Sigma_{\text{eff}}^2(Z)$ and not on the exact key distribution.

Assumptions (coordinate-wise universality). Fix the query state $Z=\{z^{(a)}\}_{a=1}^L$ and condition on it. For each distractor ν, write the L-vector of j-th coordinates $\boldsymbol{k}_{\nu,j} := (k_{\nu,j}^{(1)},\dots,k_{\nu,j}^{(L)}) \in \mathbb{R}^L$. Assume:

(U1) (*Independence across coordinates*) the vectors $\{\boldsymbol{k}_{\nu,j}\}_{j=1}^d$ are independent (for fixed ν);

(U2) (*Second moments*) $\mathbb{E}[\boldsymbol{k}_{\nu,j}]=0$ and $\mathbb{E}[\boldsymbol{k}_{\nu,j}\boldsymbol{k}_{\nu,j}^\top]=\Gamma$ for all j;

(U3) (*Uniform third moment / Lyapunov*) there exists $\kappa_3<\infty$ such that for all $a\in\mathbb{R}^L$ and all j,

$$\mathbb{E}\big|\langle a,\boldsymbol{k}_{\nu,j}\rangle\big|^3 \le \kappa_3\big(a^\top\Gamma a\big)^{3/2}.$$

Proof. Fix a query state $Z=\{z^{(a)}\}_{a=1}^L$ and consider a distractor index $\nu\neq\mu^\star$. We work conditionally on Z (so Z is deterministic throughout the argument). For each coordinate $j\in[d]$, define the L-vectors

$$\boldsymbol{z}_j := (z_j^{(1)},\dots,z_j^{(L)})\in\mathbb{R}^L, \qquad \boldsymbol{k}_{\nu,j} := (k_{\nu,j}^{(1)},\dots,k_{\nu,j}^{(L)})\in\mathbb{R}^L, \qquad a_j := \bar{A}^\top \boldsymbol{z}_j \in \mathbb{R}^L.$$

With this notation, the distractor score decomposes as

$$S_\nu(Z)=\sum_{a,b=1}^L \bar{A}_{ab}\,\langle z^{(a)},k_\nu^{(b)}\rangle = \sum_{j=1}^d\sum_{a,b=1}^L \bar{A}_{ab}\,z_j^{(a)}\,k_{\nu,j}^{(b)} = \sum_{j=1}^d X_j, \qquad X_j := a_j^\top \boldsymbol{k}_{\nu,j}.$$

Assume that the coordinate-blocks $\{\boldsymbol{k}_{\nu,j}\}_{j=1}^d$ are independent (for fixed ν), so the variables $\{X_j\}_{j=1}^d$ are independent as well. Assume also $\mathbb{E}[\boldsymbol{k}_{\nu,j}]=0$, hence $\mathbb{E}[X_j]=a_j^\top\mathbb{E}[\boldsymbol{k}_{\nu,j}]=0$.

Assume $\mathbb{E}[\boldsymbol{k}_{\nu,j}\boldsymbol{k}_{\nu,j}^\top]=\Gamma$ for all j. Then

$$\mathrm{Var}(X_j)=\mathbb{E}[(a_j^\top\boldsymbol{k}_{\nu,j})^2]=a_j^\top\Gamma a_j=(\bar{A}^\top\boldsymbol{z}_j)^\top\Gamma(\bar{A}^\top\boldsymbol{z}_j)=:\;v_j,$$

and therefore

$$\sum_{j=1}^d \mathrm{Var}(X_j)=\sum_{j=1}^d v_j \;=:\; \Sigma_{\text{eff}}^2(Z).$$

Let Φ denote the standard normal cdf. By the one-dimensional Berry–Esseen inequality for independent (not necessarily identically distributed) centered variables,

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sum_{j=1}^d X_j}{\Sigma_{\text{eff}}(Z)}\le x\right)-\Phi(x)\right| \;\le\; C_{\text{BE}}\,\frac{\sum_{j=1}^d \mathbb{E}|X_j|^3}{\Sigma_{\text{eff}}(Z)^3}.$$

Using equation 72 with $a=a_j=\bar{A}^\top\boldsymbol{z}_j$, we have

$$\mathbb{E}|X_j|^3=\mathbb{E}|\langle a_j,\boldsymbol{k}_{\nu,j}\rangle|^3 \;\le\; \kappa_3\,(a_j^\top\Gamma a_j)^{3/2}=\kappa_3\,v_j^{3/2}.$$

Substituting into Berry–Esseen yields

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{S_\nu(Z)}{\Sigma_{\text{eff}}(Z)}\le x\right)-\Phi(x)\right| \;\le\; C_{\text{BE}}\kappa_3\,\frac{\sum_{j=1}^d v_j^{3/2}}{\left(\sum_{j=1}^d v_j\right)^{3/2}}. \tag{75}$$

Consequently, if the Lyapunov (equivalently Lindeberg-type) condition

$$\frac{\sum_{j=1}^{d} v_j^{3/2}}{\left(\sum_{j=1}^{d} v_j\right)^{3/2}} \xrightarrow[d\to\infty]{} 0, \tag{76}$$

holds (for the sequence of query states under consideration), then the RHS of equation 75 vanishes and hence

$$\frac{S_\nu(Z)}{\Sigma_{\text{eff}}(Z)} \;\Rightarrow\; \mathcal{N}(0,1), \qquad d\to\infty.$$

This establishes the universality (Gaussian limit) of the distractor score at fixed Z under the stated moment assumptions. □

Remark 1 (Simple sufficient condition). Since $\sum_{j=1}^{d} v_j^{3/2} \le \left(\max_j v_j^{1/2}\right)\sum_{j=1}^{d} v_j$, we have

$$\Delta_d(Z) \le \sqrt{\frac{\max_{1\le j\le d} v_j}{\Sigma_{\text{eff}}^2(Z)}}.$$

Hence it suffices that $\max_j v_j/\Sigma_{\text{eff}}^2(Z) \to 0$. In particular, if $\max_j v_j = O(1)$ and $\Sigma_{\text{eff}}^2(Z) = \Theta(d)$, then $\Delta_d(Z) = O(d^{-1/2})$.

Corollary 1 (Bounded coordinates imply universality). *Assume $\|z^{(a)}\|_\infty \le B$ for all a and $\|\bar{A}\|_\infty \le 1$. Then $\|a_j\| \le LB$ and thus $v_j = a_j^\top \Gamma a_j \le \lambda_{\max}(\Gamma)\, L^2B^2 = O(1)$. If moreover $\Sigma_{\text{eff}}^2(Z) = \sum_{j=1}^{d} v_j \ge c\,d$ for some $c > 0$, then $\Delta_d(Z) = O(d^{-1/2})$ and $S_\nu(Z)/\Sigma_{\text{eff}}(Z) \Rightarrow \mathcal{N}(0,1)$.*

F.3 Signal-Noise Decomposition and Topological Scaling

Proposition 2 (Signal-Noise Decomposition and Topological Scaling). *Let $Z \in \mathbb{R}^{Ld}$ be a fixed query state. Consider the retrieval competition between the planted target memory μ^* and a generic distractor $\nu \neq \mu^*$, under the assumption that distractor keys are i.i.d. centered Gaussian vectors with covariance $\Gamma \otimes I_d$. The score gap $\Delta_{\mu^*,\nu}(Z) := S_{\mu^*}(Z) - S_\nu(Z)$ admits the following decomposition:*

$$\Delta_{\mu^*,\nu}(Z) = S_{\text{signal}}(Z) - \Sigma_{\text{eff}}(Z)\cdot \xi_\nu, \tag{77}$$

where:

- $S_{\text{signal}}(Z) := \langle PZ, QK_{\mu^*}\rangle_{\mathcal{H}_\mathcal{E}}$ *is the deterministic signal component (conditional on Z and the planted key);*
- $\xi_\nu \sim \mathcal{N}(0,1)$ *is a standard normal random variable;*
- $\Sigma_{\text{eff}}(Z) := \sqrt{d\cdot \operatorname{Tr}\left((\bar{A}\Gamma\bar{A}^\top)\mathsf{Q}\right)}$ *is the* topological noise scale*, which depends solely on the interaction graph $\bar{A}$, the cross-modal covariance Γ, and the query overlap structure Q.*

Consequently, the probability that a specific distractor overshadows the signal is strictly controlled by the signal-to-topology ratio $\text{SNR}(Z) := S_{\text{signal}}(Z)/\Sigma_{\text{eff}}(Z)$.

Proof. We analyze the two terms of the gap $\Delta_{\mu^*,\nu}(Z)$ separately. The first term, $S_{\mu^*}(Z)$, is a linear functional of the planted key K_{μ^*}. Conditional on the realization of the planted memory and the resulting query Z, this term is constant with respect to the randomness of the distractor ν.

The second term is the distractor score $S_\nu(Z) = \langle PZ, QK_\nu\rangle_{\mathcal{H}_\mathcal{E}}$. Since $K_\nu \sim \mathcal{N}(\mathbf{0}, \Gamma\otimes I_d)$ and the map $K_\nu \mapsto S_\nu(Z)$ is linear (being an inner product with a fixed vector PZ), $S_\nu(Z)$ is a scalar Gaussian random variable.

The mean is $\mathbb{E}[S_\nu] = \langle PZ, Q\mathbb{E}[K_\nu]\rangle = 0$. The variance is given by Theorem 5:

$$\text{Var}(S_\nu) = \Sigma_{\text{eff}}^2(Z) = Z^\top[(\bar{A}\Gamma\bar{A}^\top)\otimes I_d]Z = d\cdot\text{Tr}((\bar{A}\Gamma\bar{A}^\top)\mathsf{Q}). \tag{78}$$

A centered Gaussian variable X with variance σ^2 can be written as $X \overset{d}{=} \sigma\xi$ with $\xi \sim \mathcal{N}(0,1)$. Applying this to the noise term:

$$S_\nu(Z) \overset{d}{=} \Sigma_{\text{eff}}(Z)\cdot\xi_\nu, \quad \xi_\nu \sim \mathcal{N}(0,1). \tag{79}$$

Substituting back into the gap definition yields the claim. The failure condition $S_\nu > S_{\mu^*}$ corresponds to $\Sigma_{\text{eff}}\xi_\nu > S_{\text{signal}}$, or equivalently $\xi_\nu > \text{SNR}(Z)$, linking the error probability directly to the tail of the standard normal distribution at $\text{SNR}(Z)$. □

F.4 REM Critical Capacity (total exponential load)

Theorem 6 (REM Critical Capacity (total exponential load))**.** *Define the* total exponential load $\alpha_{\text{tot}} := \frac{1}{Ld} \log K$*, so that the number of patterns scales as* $K = \exp(\alpha_{\text{tot}} Ld)$*. The critical capacity* $\alpha_{\text{tot},c}$ *marks the thermodynamic phase transition where the target signal is overwhelmed by the condensation of noise into spurious minima.*

Combining Definition 2 and Theorem 5, this threshold is given by:

$$\alpha_{\text{tot},c}(Z) = \frac{S_{\text{signal}}(Z)^2}{2Ld\,\Sigma_{\text{eff}}^2(Z)} = \frac{1}{L}\,\alpha_c(Z), \tag{80}$$

where $\alpha_c(Z) = \frac{S_{\text{signal}}(Z)^2}{2d\,\Sigma_{\text{eff}}^2(Z)}$ *is the per-d convention. Substituting* $S_{\text{signal}} = d\,\text{Tr}(\bar{A}^\top \Gamma_{\text{align}})$ *and* $\Sigma_{\text{eff}}^2 = d\,\text{Tr}\big((\bar{A}\Gamma_{\text{noise}}\bar{A}^\top)\,\mathsf{Q}\big)$ *(with* $\Gamma_{\text{noise}} := \Gamma$ *denoting the distractor covariance) yields the explicit bound:*

$$\alpha_{\text{tot},c}(\bar{A}, \Gamma_{\text{noise}}, \Gamma_{\text{align}}) = \frac{\big(\text{Tr}(\bar{A}^\top \Gamma_{\text{align}})\big)^2}{2L\,\text{Tr}\big((\bar{A}\Gamma_{\text{noise}}\bar{A}^\top)\,\mathsf{Q}\big)}. \tag{81}$$

Proof. The proof relies on the equivalence between the asymptotic retrieval failure and the condensation transition in the Random Energy Model (REM). Let $\mathcal{D} = \{S_\nu\}_{\nu \neq \mu^*}$ be the set of $K-1$ distractor scores. From Theorem 5, each S_ν is an identically distributed Gaussian variable with mean 0 and variance Σ_{eff}^2. Under the modeling assumption that distractor patterns are independent (conditionally on a fixed query state Z), the variables in $\mathcal{D}$ are i.i.d. Gaussians.

In the zero-temperature limit ($\tilde{\beta} \to \infty$), the free energy of the noise term $\Phi(Z) = \tilde{\beta}^{-1} \log \sum_\nu e^{\tilde{\beta} S_\nu}$ is dominated by the maximum value of the set $\mathcal{D}$ (the ground state energy of the disorder). We define the noise floor as the random variable $M_K := \max_{\nu \neq \mu^*} S_\nu$. From standard Extreme Value Theory, for a set of K i.i.d. Gaussian variables $\mathcal{N}(0, \sigma^2)$, the maximum satisfies:

$$M_K / \sqrt{2 \ln K} \xrightarrow{P} \sigma \quad \text{as } K \to \infty. \tag{82}$$

In our setting, $\sigma = \Sigma_{\text{eff}}$ and the number of patterns scales exponentially with total dimension Ld as $K = \exp(\alpha_{\text{tot}} Ld)$, implying $\ln K = \alpha_{\text{tot}} Ld$. Thus, the noise floor concentrates in the thermodynamic limit ($d \to \infty$):

$$M_K = \Sigma_{\text{eff}}(Z)\sqrt{2\alpha_{\text{tot}} Ld}\,(1 + o(1)) \quad \text{w.h.p.} \tag{83}$$

A pattern is successfully retrieved as the global energy minimum if and only if the deterministic signal energy exceeds this noise floor:

$$S_{\text{signal}}(Z) > \max_{\nu \neq \mu^*} S_\nu \implies S_{\text{signal}}(Z) > \Sigma_{\text{eff}}(Z)\sqrt{2\alpha_{\text{tot}} Ld}. \tag{84}$$

Squaring both sides and isolating the load α_{tot} yields the critical condition:

$$\alpha_{\text{tot}} < \frac{S_{\text{signal}}(Z)^2}{2Ld\,\Sigma_{\text{eff}}^2(Z)}. \tag{85}$$

The critical capacity $\alpha_{\text{tot},c}$ is defined as the saturation point of this inequality. This derivation matches the "condensed phase" boundary condition derived in Lucibello & Mézard (2024), where the signal term intersects the asymptotic REM free energy. □

F.5 Finite-K Sufficient Condition for High-Confidence Retrieval

Corollary 2 (Finite-K Sufficient Condition for High-Confidence Retrieval)**.** *Adopt the definitions of Proposition 2 and let* Z *be a fixed query state. Let* K *be the total number of stored patterns (1 target* μ^**,* $K-1$ *distractors), and fix a confidence level* $\delta \in (0,1)$*. A sufficient condition to guarantee correct retrieval with probability at least* $1-\delta$ *(over the realization of the distractor keys) is that the deterministic signal exceeds the topological noise floor by a margin logarithmic in the failure rate:*

$$S_{\text{signal}}(Z) \;\geq\; \Sigma_{\text{eff}}(Z)\sqrt{2\log\left(\frac{K-1}{\delta}\right)}. \tag{86}$$

Specifically, under this condition, the probability of the failure event $\mathcal{F} := \{\exists \nu \neq \mu^* : S_\nu(Z) \geq S_{\mu^*}(Z)\}$ *satisfies* $\mathbb{P}(\mathcal{F} \mid Z) \leq \delta$*.*

Proof. We aim to bound the probability that any distractor score exceeds the signal. By the Union Bound (Boole's inequality):

$$\mathbb{P}(\mathcal{F}) \;=\; \mathbb{P}\left(\bigcup_{\nu\neq\mu^*}\{S_\nu(Z)\geq S_{\text{signal}}(Z)\}\right) \;\leq\; \sum_{\nu\neq\mu^*}\mathbb{P}\left(S_\nu(Z)\geq S_{\text{signal}}(Z)\right). \tag{87}$$

From Proposition 2, each distractor score scales as $S_\nu(Z)\overset{d}{=}\Sigma_{\text{eff}}(Z)\cdot\xi$, where $\xi\sim\mathcal{N}(0,1)$. Using the standard Chernoff bound for the tail of the normal distribution ($\mathbb{P}(\xi\geq t)\leq\exp(-t^2/2)$ for $t>0$), we have:

$$\mathbb{P}\left(S_\nu(Z)\geq S_{\text{signal}}(Z)\right) \;=\; \mathbb{P}\left(\xi\geq\frac{S_{\text{signal}}(Z)}{\Sigma_{\text{eff}}(Z)}\right) \;\leq\; \exp\left(-\frac{S_{\text{signal}}(Z)^2}{2\Sigma_{\text{eff}}^2(Z)}\right). \tag{88}$$

Substituting this into the sum over the $K-1$ distractors:

$$\mathbb{P}(\mathcal{F}) \;\leq\; (K-1)\exp\left(-\frac{S_{\text{signal}}(Z)^2}{2\Sigma_{\text{eff}}^2(Z)}\right). \tag{89}$$

We require this failure probability to be at most δ. Setting the upper bound to $\leq\delta$ and taking natural logarithms:

$$\log(K-1)-\frac{S_{\text{signal}}(Z)^2}{2\Sigma_{\text{eff}}^2(Z)} \;\leq\; \log\delta \quad\Longleftrightarrow\quad \frac{S_{\text{signal}}(Z)^2}{2\Sigma_{\text{eff}}^2(Z)} \;\geq\; \log(K-1)-\log\delta. \tag{90}$$

Combining logarithms gives $\log((K-1)/\delta)$. Multiplying by 2 and taking the square root (noting $S_{\text{signal}}>0$) yields the sufficient condition:

$$S_{\text{signal}}(Z) \;\geq\; \Sigma_{\text{eff}}(Z)\sqrt{2\log\left(\frac{K-1}{\delta}\right)}. \tag{91}$$

This concludes the proof. □

F.6 REM FREE ENERGY FOR GAUSSIAN DISTRACTORS (FINITE TEMPERATURE)

Definition 2 (Signal Energy)**.** Let μ^* be the index of the target pattern. For a query state Z, we define the *edgewise alignment overlaps* as $\gamma_{ab}:=d^{-1}\langle z^{(a)},k_{\mu^*}^{(b)}\rangle$ for edges $(a\leftarrow b)\in\mathcal{E}_{\text{het}}$, and set $\gamma_{ab}:=0$ whenever $\bar{A}_{ab}=0$. Collecting these in the *alignment matrix* $\Gamma_{\text{align}}\in\mathbb{R}^{L\times L}$ with entries $(\Gamma_{\text{align}})_{ab}=\gamma_{ab}$, the target consensus score is then given exactly by:

$$S_{\text{signal}}(Z)=d\sum_{a,b=1}^{L}\bar{A}_{ab}\gamma_{ab}=d\,\text{Tr}(\bar{A}^\top\Gamma_{\text{align}}). \tag{92}$$

Proposition 3 (REM free energy for Gaussian distractors (finite temperature))**.** *Fix Z and assume the distractor scores $\{S_\nu(Z)\}_{\nu\neq\mu^*}$ are i.i.d. with $S_\nu(Z)\sim\mathcal{N}(0,d\,\rho_{\text{eff}}(Z))$, and let $K=\exp(\alpha d)$. Then, as $d\to\infty$,*

$$\phi_{\text{noise}}(\tilde{\beta}) \;\xrightarrow{\;\mathbb{P}\;}\; \varphi_{\alpha,\rho_{\text{eff}}(Z)}(\tilde{\beta}):=\begin{cases}\dfrac{\rho_{\text{eff}}(Z)\,\tilde{\beta}}{2}+\dfrac{\alpha}{\tilde{\beta}}, & \tilde{\beta}<\tilde{\beta}_*(\alpha;Z),\\[2ex] \sqrt{2\,\alpha\,\rho_{\text{eff}}(Z)}, & \tilde{\beta}\geq\tilde{\beta}_*(\alpha;Z),\end{cases} \tag{93}$$

where the condensation threshold *is*

$$\tilde{\beta}_*(\alpha;Z):=\sqrt{\frac{2\alpha}{\rho_{\text{eff}}(Z)}}. \tag{94}$$

Equivalently, in total-load variables $\alpha=L\alpha_{\text{tot}}$,

$$\tilde{\beta}_*(\alpha_{\text{tot}};Z)=\sqrt{\frac{2L\alpha_{\text{tot}}}{\rho_{\text{eff}}(Z)}}. \tag{95}$$

Proof. Set $E_\nu := S_\nu/d$. Since $S_\nu \sim \mathcal{N}(0, d\rho_{\text{eff}})$, the density of E_ν is

$$f_d(u) = \sqrt{\frac{d}{2\pi\rho_{\text{eff}}}}\exp\Big(-d\,\frac{u^2}{2\rho_{\text{eff}}}\Big),$$

which exhibits a large-deviation form with rate function $I(u) = u^2/(2\rho_{\text{eff}})$ at speed d. Let $N_d([u, u+\Delta])$ be the number of indices $\nu \le K$ such that $E_\nu \in [u, u+\Delta]$. Then $N_d([u, u+\Delta])$ is binomial with parameters $K = \exp(\alpha d)$ and $p_d([u, u+\Delta]) = \int_u^{u+\Delta} f_d(v)\,dv$. For fixed u and small fixed $\Delta > 0$,

$$p_d([u, u+\Delta]) = \exp\Big(-d\,I(u) + o(d)\Big),$$

hence $\mathbb{E}[N_d([u, u+\Delta])] = \exp(d(\alpha - I(u)) + o(d))$. A standard Chernoff bound for binomials implies concentration on the exponential scale: for any $\varepsilon > 0$,

$$\mathbb{P}(\exp(d(\alpha - I(u) - \varepsilon)) \le N_d([u, u+\Delta]) \le \exp(d(\alpha - I(u) + \varepsilon))) \to 1$$

whenever $\alpha > I(u)$, while $N_d([u, u+\Delta]) = 0$ w.h.p. when $\alpha < I(u)$. Therefore, with high probability, the dominant contribution to the partition function

$$\mathcal{Z}_{\text{noise}}(\tilde\beta) = \sum_{\nu \le K} \exp(d\tilde\beta E_\nu)$$

comes from energy levels u such that $\alpha \ge I(u)$, i.e. $|u| \le u_c$ with $u_c := \sqrt{2\alpha\rho_{\text{eff}}}$. Approximating the sum by a Riemann sum over bins and using the above exponential-scale control gives

$$\frac{1}{d}\log \mathcal{Z}_{\text{noise}}(\tilde\beta) \xrightarrow{\mathbb{P}} \sup_{|u| \le u_c}\Big\{\alpha - I(u) + \tilde\beta u\Big\}.$$

Since $I(u) = u^2/(2\rho_{\text{eff}})$, the unconstrained maximizer is $u_* = \rho_{\text{eff}}\tilde\beta$. If $u_* < u_c$ (equivalently $\tilde\beta < \sqrt{2\alpha/\rho_{\text{eff}}}$), the supremum equals $\alpha + \frac{\rho_{\text{eff}}\tilde\beta^2}{2}$. If $u_* \ge u_c$ (equivalently $\tilde\beta \ge \sqrt{2\alpha/\rho_{\text{eff}}}$), the supremum is attained at the boundary $u = u_c$ and equals $\tilde\beta u_c = \tilde\beta\sqrt{2\alpha\rho_{\text{eff}}}$. Dividing by $\tilde\beta$ yields equation 93 and equation 94. □

F.7 Spectral Bound on Topological Noise Variance

Proposition 4 (Spectral Bound on Topological Noise Variance)**.** *Assume isotropic distractor noise ($\Gamma = I_L$) and let $M := \bar{A}\bar{A}^\top$ be the topology-induced Gram matrix. Let $\lambda_1(M) \ge \lambda_2(M) \ge \cdots \ge \lambda_L(M) \ge 0$ denote the eigenvalues of M. The effective topological variance $\Sigma^2_{\text{eff}}(\mathsf{Q})$ admits the following spectral decomposition bound:*

$$\Sigma^2_{\text{eff}}(\mathsf{Q}) \;\le\; d\cdot[\lambda_1(M)\cdot\mathcal{S}_{\text{sync}}(\mathsf{Q}) \;+\; \lambda_2(M)\cdot\mathcal{E}_{\text{async}}(\mathsf{Q})]\,, \tag{96}$$

where $\mathcal{S}_{\text{sync}}(\mathsf{Q}) := \frac{1}{L}\mathbf{1}^\top\mathsf{Q}\mathbf{1}$ measures the energy of the query component aligned with the uniform mode, and $\mathcal{E}_{\text{async}}(\mathsf{Q}) := \mathrm{Tr}(\mathsf{Q}) - \mathcal{S}_{\text{sync}}(\mathsf{Q})$ measures the asynchronous fluctuations.

In the specific case of doubly-stochastic topologies *(e.g., regular graphs) where $\bar{A}\mathbf{1} = \mathbf{1}$, the dominant eigenvalue is exactly $\lambda_1(M) = 1$ with eigenvector $\mathbf{1}/\sqrt{L}$. In this regime, the variance is explicitly controlled by the* spectral gap $\gamma := 1 - \lambda_2(M)$*:*

$$\Sigma^2_{\text{eff}}(\mathsf{Q}) \;\le\; d\,[\mathrm{Tr}(\mathsf{Q}) \;-\; \gamma\cdot\mathcal{E}_{\text{async}}(\mathsf{Q})]\,. \tag{97}$$

Proof. The effective variance is given by $\frac{1}{d}\Sigma^2_{\text{eff}} = \mathrm{Tr}(M\mathsf{Q})$. Since Q is positive semi-definite, we decompose it via orthogonal projection onto the subspace spanned by the uniform vector $\mathbf{u} := \frac{1}{\sqrt{L}}\mathbf{1}$. We write $\mathsf{Q} = q_{\text{sync}}\mathbf{u}\mathbf{u}^\top + \mathsf{Q}_\perp$, where $\mathsf{Q}_\perp\mathbf{u} = \mathbf{0}$. Identifying terms, we have $q_{\text{sync}} = \mathbf{u}^\top\mathsf{Q}\mathbf{u} = \frac{1}{L}\mathbf{1}^\top\mathsf{Q}\mathbf{1} = \mathcal{S}_{\text{sync}}(\mathsf{Q})$, and by linearity of the trace, $\mathrm{Tr}(\mathsf{Q}_\perp) = \mathrm{Tr}(\mathsf{Q}) - q_{\text{sync}} = \mathcal{E}_{\text{async}}(\mathsf{Q})$.
Substituting this decomposition into the variance functional:

$$\mathrm{Tr}(M\mathsf{Q}) = q_{\text{sync}}\mathbf{u}^\top M\mathbf{u} + \mathrm{Tr}(M\mathsf{Q}_\perp). \tag{98}$$

Using the variational characterization of eigenvalues (Courant-Fischer theorem): 1. The first term is bounded by the spectral radius: $\mathbf{u}^\top M\mathbf{u} \le \lambda_1(M)$. 2. For the second term, since $\mathsf{Q}_\perp$ is supported

on the subspace orthogonal to $\mathbf{u}$, the trace is bounded by the maximum eigenvalue of M restricted to that subspace. Generally, this is bounded by $\lambda_1(M)$, but if $\mathbf{u}$ is the dominant eigenvector of M (which holds if $\bar{A}$ is doubly stochastic), then the maximum eigenvalue on the orthogonal complement $\mathbf{u}^\perp$ is exactly $\lambda_2(M)$.

Thus, for doubly stochastic graphs where $\lambda_1(M) = 1$ and $\mathbf{u}$ is an eigenvector:

$$\mathrm{Tr}(M\mathsf{Q}) = 1 \cdot q_{\mathrm{sync}} + \mathrm{Tr}(M\mathsf{Q}_\perp) \leq q_{\mathrm{sync}} + \lambda_2(M)\mathrm{Tr}(\mathsf{Q}_\perp). \tag{99}$$

Substituting $\lambda_2(M) = 1 - \gamma$ yields:

$$\mathrm{Tr}(M\mathsf{Q}) \leq q_{\mathrm{sync}} + (1-\gamma)\mathcal{E}_{\mathrm{async}} = (q_{\mathrm{sync}} + \mathcal{E}_{\mathrm{async}}) - \gamma\mathcal{E}_{\mathrm{async}} = \mathrm{Tr}(\mathsf{Q}) - \gamma\mathcal{E}_{\mathrm{async}}. \tag{100}$$

Multiplying by d recovers the claim. □