Attention, Transformers, and Backpropagation are Degenerate Limits of the Variational Free Energy Principle

Robert C. Dennis CDENN016@GMAIL.COM
Independent Researcher
Leander, Texas 78641, USA

Abstract

We present a unified gauge-theoretic formulation of attention and message communication in multi-agent Bayesian systems performing variational inference. Each agent is modeled as a smooth local section of an associated bundle with statistical manifold fiber over a noumenal base manifold $\mathcal{C}$. We demonstrate that in the isotropic, flat-bundle, Dirac-delta function limit (where all local frames coincide globally) the generalized attention weights reduce to the canonical transformer rule $\beta_{ij} \propto \text{softmax}(Q_i K_j^\top)$ exactly and backpropagation training emerges as gradient descent on the free energy in the deterministic, flat-bundle limit—discarding the natural gradient structure that provides faster convergence. Starting from a generative model of inter-agent message exchange, we derive a generalized variational free-energy functional whose stationary points govern agent dynamics. In the absence of observations, this functional defines a gauge-symmetric vacuum theory in which all agents converge to identical beliefs modulo gauge orbit; introducing observations breaks this symmetry, leading to agent specialization. We validate our framework through simulations of $N = 8$ agents with 9-dimensional Gaussian beliefs under SO(3) gauge transformations, demonstrating convergence to symmetric vacuum states and symmetry breaking under observation/training. We fully derive the equivalence to modern transformer architectures and backpropagation and show our framework suggests exponential speedup during training via natural gradient descent.

Keywords: gauge theory, free energy principle, transformer attention, variational inference, information geometry, natural gradient, symmetry breaking, multi-agent systems

1 Introduction

Recent advances in neuroscience and intelligent systems have independently converged on the idea that intelligent systems integrate perception, inference, and communication under the constraints of uncertainty (Bahdanau et al., 2014) (Amari, 1998) (Bronstein et al., 2021) (Foerster et al., 2016)(Wooldridge, 2009). Friston's Free Energy Principle (FEP) provides a general variational formulation of inference in cognitive systems(Friston, 2010) (Parr et al., 2022) (Friston et al., 2017) (Ramstead et al., 2019), whereas the attention mechanism in modern machine learning architectures defines a powerful (although empirically derived) rule for token prediction(Clark et al., 2019) (Vaswani et al., 2017). Despite their shared reliance on probabilistic inference and pairwise interaction, these two frameworks remain stubbornly separated. In particular, transformer attention lacks an underlying geometric or mathematical foundation and details on how and why modern machine learning architectures operate as well as they do remains obscure.

Many varieties of transformer and attention architectures have been proposed, implemented, and studied in recent years. (Fuchs et al., 2020)(Thomas et al., 2018). Some

architectures make use of bundle geometric frameworks but lack a first-principles foundation or connection to the FEP(Kondor and Trivedi, 2018)(Finzi et al., 2020)(Bronstein et al., 2021). Furthermore, attempts at curved space token embeddings have produced mixed results.(Bonnabel, 2013)(Absil et al., 2008)

In this report we propose a unified, gauge-equivariant framework that connects these disparate yet similar domains. Our framework is based on a principled bundle geometry whereby each agent is modeled as a smooth local section of an associated bundle with statistical manifold fibers over a "noumenal" base manifold. Inter-agent communication arises naturally through a non-abelian gauge connection that defines parallel-transport operators between agents' local gauge frames. Within this geometry, attention emerges as a gauge-aligned Kullback–Leibler (KL) term derived directly from the variational free energy of a coupled multi-agent generative model.

We then show that in the flat-bundle, isotropic, delta-function limit attention reduces to the standard transformer dot-product attention QK^T, thereby identifying transformer attention as a degenerate case of a broader geometric and statistical law of communication predicated upon the FEP. We then show that hard, one-hot attention encoding is the zero-temperature limit of the FEP agent-agent coupling term and the large temperature limit leads to uniform encoding. We demonstrate this by simulating a toy model of variational gradient descent under generalized free energy of multivariate Gaussian agents and by applying our alignment expression to a frozen transformer. Finally, we describe the theory and numerical results that suggest that token encoding under data training (i.e. agent belief alignment) is a spontaneous symmetry breaking phenomenon of our gauge-theoretic framework.

Finally, we mathematically show that modern machine learning back propagation gradient descent and attention transformers the identically the Dirac flat limit of a more general gauge equivariant theory under natural gradient descent of the variational free energy functional. Our results suggests that deep learning architectures can potentially achieve an exponential speedup during training by leveraging our gauge equivariant free energy formalism.

2 Methods

We begin with a general geometric construction that naturally supports hierarchical emergence of meta-agents, cross-scale interactions, and non-trivial holonomy of belief and model transport. The details of the general geometry and our complete framework can be found in the appendix and references (Nakahara, 2003)(Kullback and Leibler, 1951) (Sternberg, 1994)(Fulton and Harris, 1991)(Frankel, 2011)(Baez and Muniain, 1994).

Briefly, in this current report, we describe agents as modeled by local sections of a pair of associated bundles to a principal G bundle. The fibers are generally K_q and K_p-dimensional statistical manifolds and the base space is a general smooth manifold. An agent is then a tuple of beliefs, priors, and gauge frames (see figure 1). Our framework enables a unified description of both intra-agent inference and inter-agent communication in terms of gauge frame transport and alignment.

For our current considerations we simplify the geometry considerably by allowing beliefs and models to occupy the same latent spaces. This then collapses the two general gauge

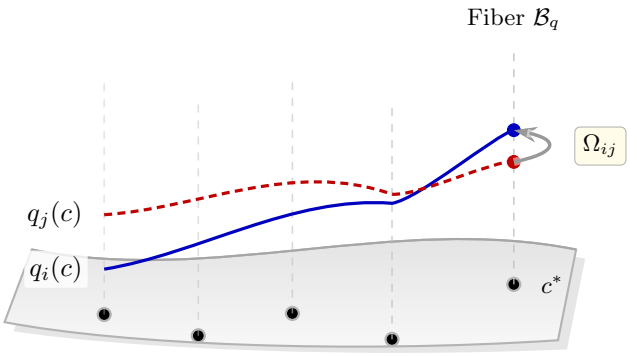

Figure 1: Visualization of agent sections $q_i(c)$ and $q_j(c)$ over an associated bundle. The shaded surface represents the base manifold $\mathcal{C}$, with vertical fibers $\mathcal{B}_q$ shown as dashed lines anchored at black points. At the reference point c^* on the rightmost fiber, the transport operator $\Omega_{ij}(c^*)$ maps the red section value $q_j(c^*)$ into the blue section value $q_i(c^*)$, enabling frame alignment between agent representations.

frames into a single gauge frame per agent. We call this the "matched bundle" version of our framework. The details of this simplification can be found in the appendix.

2.1 An intuitive simplification for non-geometers

For interdisciplinary readers, it may be helpful to visualize each agent in this framework as a vector and a matrix field defined over a finite spatial region. Concretely, at every point c in the base manifold, agent i carries a mean vector $\mu_i(c)$ and a covariance matrix $\Sigma_i(c)$ representing, respectively, its local expected state and uncertainty. Together these form a smooth Gaussian field—the agent's belief section of the statistical bundle.

We will see that the agent's local gauge frame $\phi_i(c)$ may be intuitively interpreted as the agent's subjective frame of reference—its internal coordinate system through which beliefs and models are represented and compared. In this sense, $\phi_i(c)$ captures an agent's "internal orientation" toward the world: it determines how external information is projected into the agent's internal representational space. Gauge transformations $\phi_i(c) \to \phi_i(c) + \xi(c)$ thus correspond to changes in perspective that leave the underlying informational content invariant, much like shifts of viewpoint in perception that preserve the same world model. In the case of transformers we will find that these frames may encode relative positions of tokens/agents.

3 Derivation of Generalized Variational Free Energy from a Normalized Generative Model

We derive the generalized variational free energy we will utilize from general first principles, showing that both belief alignment (weighted by β_{ij}) and model alignment (weighted by γ_{ij}) emerge naturally from a single normalized generative prior with auxiliary agreement variables. This construction justifies the KL-based coupling terms as consequences

Symbol	Description	Type/Dimension
Fiber Bundle Structure		
$\mathcal{C}$	Base manifold (spatial domain)	Manifold
$\mathcal{B}_{q_i}$	Belief fiber for agent i	$\mathbb{R}^{K_q}$
$\mathcal{B}_{p_i}$	Model fiber for agent i	$\mathbb{R}^{K_p}$
G	Gauge group	SO(3)
$\mathfrak{g}$	Lie algebra of G	$\mathfrak{so}(3)$
Gauge Fields and Connections		
Ω^q_{ij}	Connection $\mathcal{B}_{q_j} \to \mathcal{B}_{q_i}$	SO(3)
Ω^p_{ij}	Connection $\mathcal{B}_{p_j} \to \mathcal{B}_{p_i}$	SO(3)
ϕ_i	Gauge frame (belief frame)	$\mathfrak{g} \cong \mathbb{R}^3$
$\tilde{\phi}_i$	Gauge framer (model frame)	$\mathfrak{g} \cong \mathbb{R}^3$
Bundle Morphisms		
Φ_i	Morphism $\mathcal{B}_{q_i} \to \mathcal{B}_{p_i}$	$\mathbb{R}^{K_p \times K_q}$
$\tilde{\Phi}_i$	Morphism $\mathcal{B}_{p_i} \to \mathcal{B}_{q_i}$	$\mathbb{R}^{K_q \times K_p}$
Statistical Parameters		
$q_i(k_i)$	Belief distribution	$\mathcal{N}(\mu_{q,i}, \Sigma_{q,i})$
$p_i(k_i)$	Model distribution	$\mathcal{N}(\mu_{p,i}, \Sigma_{p,i})$
$\mu_{q,i}, \mu_{p,i}$	Mean vectors	$\mathbb{R}^{K_q}, \mathbb{R}^{K_p}$
$\Sigma_{q,i}, \Sigma_{p,i}$	Covariance matrices	$\mathbb{R}^{K \times K}, \succ 0$
Attention and Coupling		
β_{ij}	Attention weight $j \to i$	$[0, 1]$
τ	Attention temperature	$\mathbb{R}_+$
D_{KL}	Kullback–Leibler divergence	$\mathbb{R}_+ \cup \{0\}$

Table 1: Principal notation used throughout this paper.

of gauge-transported Gaussian consistency constraints. In what follows we shall assume all probability distributions are defined at a single specific base manifold point $c = c^*$ unless otherwise noted. We will label the fiber's latent coordinates as $k_i \in \mathcal{B}$ in order to define the necessary integrals.

3.1 Latent Variables and Fiber Geometry

Each agent i maintains two distinct latent variables living in separate fiber bundles:

$$k_i \in \mathbb{R}^{d_q} \quad \text{(belief latent in } \mathcal{E}_q\text{)}, \qquad m_i \in \mathbb{R}^{d_p} \quad \text{(model latent in } \mathcal{E}_p\text{)}. \tag{1}$$

The full state of agent i at base manifold point $c^*$is a Gaussian distribution over each latent:

$$\begin{aligned} q_i(k_i) &= \mathcal{N}(k_i\,;\mu_{q,i},\,\Sigma_{q,i}), \quad \mu_{q,i}\in\mathbb{R}^{d_q},\ \Sigma_{q,i}\in\mathbb{R}^{d_q\times d_q},\ \Sigma_{q,i}\succ 0,\\ s_i(m_i) &= \mathcal{N}(m_i\,;\mu_{p,i},\,\Sigma_{p,i}), \quad \mu_{p,i}\in\mathbb{R}^{d_p},\ \Sigma_{p,i}\in\mathbb{R}^{d_p\times d_p},\ \Sigma_{p,i}\succ 0. \end{aligned} \tag{2}$$

Thus, the fiber at each agent's location $c\in\mathcal{C}$ is the product statistical manifold:

$$\mathcal{B}(c) = \mathcal{B}_q \times \mathcal{B}_p, \tag{3}$$

where

$$\begin{aligned} \mathcal{B}_q &= \left\{(\mu_q,\Sigma_q) : \mu_q\in\mathbb{R}^{d_q},\ \Sigma_q\in\mathbb{R}^{d_q\times d_q},\ \Sigma_q\succ 0\right\},\\ \mathcal{B}_p &= \left\{(\mu_p,\Sigma_p) : \mu_p\in\mathbb{R}^{d_p},\ \Sigma_p\in\mathbb{R}^{d_p\times d_p},\ \Sigma_p\succ 0\right\}. \end{aligned} \tag{4}$$

3.2 Base Priors

Each latent has an independent Gaussian base prior encoding agent-specific inductive biases:

$$p_i(k_i) = \mathcal{N}(k_i\,;\mu_{0,i}^{(q)},\Sigma_{0,i}^{(q)}), \qquad r_i(m_i) = \mathcal{N}(m_i\,;\mu_{0,i}^{(p)},\Sigma_{0,i}^{(p)}). \tag{5}$$

These priors are **local**—they live in each agent's own gauge frame and need not be related across agents until transported via Ω_{ij}.

3.2.1 Auxiliary Agreement Variables

To enforce consistency between agents after gauge transport, we introduce an auxiliary "agreement" variable for each ordered pair (i,j):

$$z_{ij}\in\mathbb{R}^{d_q} \quad \text{(belief agreement)}, \qquad w_{ij}\in\mathbb{R}^{d_p} \quad \text{(model agreement)}. \tag{6}$$

- z_{ij}: "What agent i believes agent j's belief looks like, after transporting j's belief into i's gauge frame"
- w_{ij}: "What agent i believes agent j's generative model looks like, after gauge transport"

These latent mediators will be integrated out, leaving an effective pairwise coupling between k_i, k_j and m_i, m_j.

Agreement variables allow us to construct a normalized joint generative model whose marginal over latents $\{k_i, m_i\}$ yield the desired gauge covariant pair potentials. This is in contrast to un-normalized Markov random fields, where potentials are imposed by fiat.

3.3 Normalized Joint Generative Model

We define the Gaussian couplings and enforce that each agreement variable z_{ij},w_{ij} simultaneously matches:

1. Agent i's own latent k_i, m_i

2. Agent j's latent k_j, m_j after gauge transport into agent i's frame

The auxiliary variable z_{ij} is drawn from the product of Gaussians:

$$p(z_{ij} \mid k_i, k_j) \propto \mathcal{N}(z_{ij}; k_i, \Lambda_{ij}^{-1}) \cdot \mathcal{N}(z_{ij}; \Omega_{ij} k_j, \Lambda_{ij}^{-1}) \tag{7}$$

and similarly for model alignment:

$$p(w_{ij} \mid m_i, m_j) \propto \mathcal{N}(w_{ij}; m_i, \Gamma_{ij}^{-1}) \cdot \mathcal{N}(w_{ij}; \tilde{\Omega}_{ij} m_j, \Gamma_{ij}^{-1}) \tag{8}$$

where:

- $\Lambda_{ij} \in \mathbb{R}^{d_q \times d_q}$, $\Lambda_{ij} \succ 0$: Belief alignment precision
- $\Gamma_{ij} \in \mathbb{R}^{d_p \times d_p}$, $\Gamma_{ij} \succ 0$: Model alignment precision
- $\Omega_{ij} \in SO(3)$: Gauge transport from agent j's frame to agent i's frame (belief channel)
- $\tilde{\Omega}_{ij} \in SO(3)$: Gauge transport from agent j's frame to agent i's frame (model channel)

3.3.1 Gauge Transport as Frame Rotation

The gauge transport operators $\Omega_{ij}, \tilde{\Omega}_{ij}$ are not parallel transport along a connection in the usual sense (which would be path-dependent). Instead, they are pointwise gauge frame rotations:

$$\Omega_{ij}(c) = e^{\phi_i(c)} \cdot e^{-\phi_j(c)} \in SO(3), \tag{9}$$

where $\phi_i : \mathcal{U}_i \to \mathfrak{so}(3)$ is agent i's gauge frame field (a local section of the Lie algebra bundle).

These act on the latent variables as

$$\begin{aligned} \Omega_{ij} k_j &:= \rho_q(\Omega_{ij})\, k_j \in \mathbb{R}^{d_q}, \\ \tilde{\Omega}_{ij} m_j &:= \rho_p(\tilde{\Omega}_{ij})\, m_j \in \mathbb{R}^{d_p}. \end{aligned} \tag{10}$$

Later we invoke the transformer limit and take $\rho_q(\Omega_{ij}) = \mathbb{I}$, so $\Omega_{ij} k_j = k_j$ (trivial transport).

3.3.2 Full Joint Distribution

The joint generative model over all latents and auxiliary variables is:

$$\boxed{\begin{aligned} &p(\{k_i\},\{m_i\},\{z_{ij}\},\{w_{ij}\}) \\ &\quad= \left[\prod_i p_i(k_i)\, r_i(m_i)\right] \\ &\qquad\times \left[\prod_{i,j} \mathcal{N}(z_{ij}\,; k_i, \Lambda_{ij}^{-1})\, \mathcal{N}(z_{ij}\,; \Omega_{ij} k_j, \Lambda_{ij}^{-1})\right] \\ &\qquad\times \left[\prod_{i,j} \mathcal{N}(w_{ij}\,; m_i, \Gamma_{ij}^{-1})\, \mathcal{N}(w_{ij}\,; \tilde{\Omega}_{ij} m_j, \Gamma_{ij}^{-1})\right]. \end{aligned}} \tag{11}$$

The joint distribution (11) is properly normalized

$$\int \prod_i dk_i\, dm_i \prod_{i,j} dz_{ij}\, dw_{ij}\; p(\{k_i\},\{m_i\},\{z_{ij}\},\{w_{ij}\}) = 1. \tag{12}$$

since each factor is a normalized Gaussian:

This is in contrast to unnormalized Markov random fields of the form $p(\{k_i\}) \propto \exp[-\sum_{ij} \psi_{ij}(k_i, k_j)]$, where the partition function Z is intractable(Amari, 1985). Our construction guarantees $Z = 1$ by design.

We now form the variational free energy under a mean-field posterior approximation. Assume a factorized posterior

$$q(\{k_i\},\{m_i\}) = \prod_i q_i(k_i)\, s_i(m_i), \tag{13}$$

with Gaussian factors

$$q_i(k_i) = \mathcal{N}(k_i; \mu_{q,i}, \Sigma_{q,i}), \qquad s_i(m_i) = \mathcal{N}(m_i; \mu_{p,i}, \Sigma_{p,i}). \tag{14}$$

The variational free energy is defined as

$$\mathcal{F} := \mathbb{E}_q[\log q(\{k_i\},\{m_i\})] - \mathbb{E}_q[\log p(\{k_i\},\{m_i\})] - \mathbb{E}_q[\log p(o\,|\{k_i\},\{m_i\})], \tag{15}$$

where $p(o\,|\{k_i\},\{m_i\})$ is the observation likelihood.

Expanding the first two terms using (13) and the marginal prior, and dropping additive constants, we obtain

The complete global variational free energy at a single base manifold point c therefore has the form

$$\begin{aligned}\mathcal{F}[\{q_i\},\{s_i\}] = &\underbrace{\sum_i D_{\mathrm{KL}}(q_i \| p_i)}_{\text{(1) Belief prior}} + \underbrace{\sum_i D_{\mathrm{KL}}(s_i \| r_i)}_{\text{(2) Model prior}} \\ &+ \underbrace{\sum_{i,j} \beta_{ij} D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j)}_{\text{(3) Belief alignment}} \\ &+ \underbrace{\sum_{i,j} \gamma_{ij} D_{\mathrm{KL}}(s_i \| \tilde{\Omega}_{ij} s_j)}_{\text{(4) Model alignment}} \\ &- \underbrace{\mathbb{E}_q[\log p(o \mid \{k_i\},\{m_i\})]}_{\text{(5) Observation likelihood}}\end{aligned} \tag{16}$$

Each term encodes a distinct aspect of multi-agent inference:

(1) Belief Prior: $D_{\mathrm{KL}}(q_i \| p_i)$

The belief prior measures the deviation of agent i's current belief $q_i(k_i)$ from its local prior $p_i(k_i \mid m_i)$. This term regularizes beliefs toward locally expected states. In the absence of observations and inter-agent coupling, minimizing $\mathcal{F}$ would yield $q_i^* = p_i$, recovering pure prior-based prediction.

(2) Model Prior: $D_{\mathrm{KL}}(s_i \| r_i)$

The model prior, similarly measures the deviation of agent i's current model belief $s_i(m_i)$ from its hyperprior $r_i(m_i)$. This term regularizes model beliefs toward baseline expectations. It prevents overfitting to recent data by anchoring s_i to a stable hyperprior r_i determined from some higher, slower level.

Hierarchical relationship to (1): The prior $p_i(k_i \mid m_i)$ in term (1) depends on the model parameters m_i that are themselves uncertain under $s_i(m_i)$. Thus:

$$p_i(k_i) = \int p_i(k_i \mid m_i)\, s_i(m_i)\, \mathrm{d}m_i \tag{17}$$

This creates a two-level Bayesian hierarchy: uncertainty about states (q_i) and uncertainty about the model generating those states (s_i).

(3) Belief Alignment: $\beta_{ij} D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j)\}$

Belief alignment represents the discrepancy between agent i's belief and agent j's belief after gauge transport into agent i's local frame - i.e. agent i's interpretation of agent j's belief.

This term enforces epistemic consensus— agents with high β_{ij} are driven to agree on their beliefs about the current world state, modulo gauge transformations. This implements distributed inference: agents pool information about their latents by aligning their beliefs $q_i(k_i)$ and $q_j(k_j)$.

(4) Model Alignment: $\gamma_{ij} D_{\mathrm{KL}}(s_i \| \tilde{\Omega}_{ij} s_j)\}$

Enforces a meta-cognitive consensus among agents with high γ_{ij}. Agents are driven to agree on their beliefs about how the world works, not just what state it's in. This implements distributed model learning: agents gather and pool evidence about model structure m by aligning their second-order beliefs $s_i(m_i)$ and $s_j(m_j)$.

Generally model-like terms can be expected to fluctuate slowly in contrast belief-like terms.

(5) Observation Likelihood: $-\mathbb{E}_q[\log p(o \mid \{k_i\}, \{m_i\})]$

This term is the expected negative log-likelihood of observations o given latent states $\{k_i\}$ and models $\{m_i\}$, averaged over the recognition distributions $\{q_i\}$, grounded in sensory observations/data. Without this term, the system is a pure vacuum theory where agents converge to their coupled prior without external input. Observations break the vacuum symmetry, forcing agents to specialize based on local sensory evidence (see Results).

In the limit of deterministic beliefs ($q_i \to \delta(k_i - \mu_i)$), this reduces to a quadratic machine learning loss function (shown below) and suggests machine learning training is equivalent to variational free energy inference.

In the absence of observations ($o = \emptyset$), the free energy is symmetric under simultaneous gauge transformation of all agents: $\phi_i \to \phi_i + \phi_0$ for any $\phi_0 \in \mathfrak{g}$. Observations break this symmetry by coupling agents to external data with fixed reference frames as we show in our results section. This is an epistemic analog to Goldstone's theorem in classical field theory.

3.4 Final Form of the Variational Free Energy

In the appendix we further show how these quadratic forms lead to the KL divergence terms along with the appropriate softmax weights. We these pieces in place it is straightforward to plug in the distributions above and arrive at the main result. Recall that this is the variational free energy (VFE) defined at a single base manifold point. The full VFE would then be an integral over all agents and their overlaps (χ_i) in the base manifold $\mathcal{C}$.

$$
\begin{aligned}
\mathcal{F}[\{q_i(c)\}, \{s_i(c)\}, \{\chi_i\}] = & \sum_i \int_{\mathcal{C}} \chi_i(c)\ D_{\mathrm{KL}}(q_i(c) \,\|\, p_i(c)) dc \\
& + \sum_i \int_{\mathcal{C}} \chi_i(c)\ D_{\mathrm{KL}}(s_i(c) \,\|\, r_i(c)) dc \\
& + \sum_{i,j} \int_{\mathcal{C}} \chi_{ij}(c)\ \beta_{ij}(c)\, D_{\mathrm{KL}}(q_i(c) \,\|\, \Omega_{ij}(c) q_j(c)) dc \\
& + \sum_{i,j} \int_{\mathcal{C}} \chi_{ij}(c)\ \gamma_{ij}(c)\, D_{\mathrm{KL}}(s_i(c) \,\|\, \tilde{\Omega}_{ij}(c) s_j(c)) dc \\
& - \int_{\mathcal{C}} \sum_i \chi_i(c)\, \mathbb{E}_{q_i(c)}[\log p(o(c) \,|\{k_i\}(c), \{m_i\}(c))] dc,
\end{aligned}
\tag{18}
$$

3.4.1 Symmetry Breaking

In the absence of observations ($p(o|\cdot) = \text{const}$), the free energy is invariant under SO(3). The vacuum state corresponds to perfect alignment: $q_i = \Omega_{ij} q_j$ and $p_i = \tilde{\Omega}_{ij} p_j$ for all agent pairs (i, j), meaning all agents maintain rotationally equivalent beliefs that differ only by frame transformations. In this regime, the dynamics drive the system toward a degenerate manifold of ground states parameterized by the gauge orbit. Agents synchronize over gradient descent towards $\|\mu_i(c)\| = \mu^*$, but the absolute orientations remain arbitrary.

This is the statistical geometric analogue of the Goldstone phase in spontaneous symmetry breaking.(Weinberg, 1995)(Goldstone, 1961)

Observations destroy this degeneracy by coupling agent to external data through the likelihood terms. Each agent's sensory stream o_i acts as an external field (source) that selects a preferred orientation in its fiber, pinning q_i to a definite point. The system transitions from the symmetric vacuum to a symmetry-broken phase where agents develop distinct specializations: $\|\mu_i(c)\| \neq \|\mu_j(c)\|$ with the diversity driven by observations/data. The frame transformations Ω_{ij} then encode how these specialized representations relate geometrically. This observation-induced symmetry breaking is what enables non-trivial multi-agent coordination: agents must now actively maintain geometric relationships between their specialized frames rather than simply coexisting in a rotationally symmetric configuration.

3.4.2 Summary

We have derived the generalized variational free energy from a normalized generative model with agreement variables. The key results are:

- Both β-weighted belief alignment and γ-weighted model alignment arise from the same principled construction, not as ad hoc regularizers.
- The alignment weights β_{ij} and γ_{ij} are proportional to the coupling precisions Λ_{ij} and Γ_{ij}.
- The forward KL divergence emerges naturally in the alignment regime.
- Our framework naturally accommodates communicable coupling: different agent pairs can have different alignment strengths, and belief coupling can differ from model coupling.

This derivation establishes the generalized variational free energy as a fundamental object in a gauge-theoretic multi-agent geometry rooted in informational and differential geometries.

Therefore, multi-agent communication within a gauge covariant formulation allows the FEP to be satisfied as well as allows us to connect attention, transformers, and machine learning to variational inference.

4 Reduction to Transformer Attention and Backpropagation

In this section we demonstrate that standard transformer self-attention and backpropagation emerge as limiting cases of our gauge-theoretic framework. This establishes neural network training not as an ad hoc optimization procedure, but as the deterministic limit of a first-principles information-geometric minimization process over a gauge bundle geometry.

4.1 Agents as Gaussian Beliefs

In our general formulation, each agent i (token) maintains a local state modeled as a Gaussian belief:

$$q_i = \mathcal{N}(\mu_i, \sigma^2 I), \tag{19}$$

where $\mu_i \in \mathbb{R}^d$ is the agent's mean representation in its local gauge frame, and we assume isotropic covariance $\Sigma_i = \sigma^2 I$ for all agents.

Each agent chooses an internal gauge frame parametrized by $\phi_i \in \mathfrak{so}(3)$. Communication between agents i and j is mediated by gauge frame transport,

$$\Omega_{ij} = e^{\phi_i} e^{-\phi_j} \in \mathrm{SO}(3), \tag{20}$$

which relates agent j's internal gauge frame to agent i's internal gauge frame. The orthogonal structure $\Omega_{ij} \in \mathrm{SO}(3)$ is necessary to preserve the isotropic Gaussian covariance structure under gauge transport.

4.2 KL Divergence

For isotropic Gaussians $q_i = \mathcal{N}(\mu_i, \sigma^2 I)$ and the gauge-transported belief $\Omega_{ij} q_j = \mathcal{N}(\Omega_{ij}\mu_j, \sigma^2 I)$ (where we used $\Omega_{ij}\Omega_{ij}^\top = I$), the KL divergence is:

$$D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j) = \frac{1}{2\sigma^2} \|\mu_i - \Omega_{ij}\mu_j\|^2. \tag{21}$$

As we have shown in the appendix, agents share information via softmax-weighted coupling with temperature τ:

$$\beta_{ij} = \mathrm{softmax}_j \left(-\frac{D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j)}{\tau} \right) = \mathrm{softmax}_j \left(-\frac{\|\mu_i - \Omega_{ij}\mu_j\|^2}{2\sigma^2 \tau} \right). \tag{22}$$

The message (update) received by agent i is:

$$m_i = \sum_j \beta_{ij}\, \Omega_{ij} \mu_j, \tag{23}$$

which is the gauge-theoretic analog of the attention aggregation $\sum_j \alpha_{ij} V_j$ in standard transformers.

4.3 Derivation of Dot-Product Attention

We have the simplified compatibility as:

$$s_{ij} \equiv \frac{\|\mu_i - \Omega_{ij}\mu_j\|^2}{2\sigma^2 \tau}. \tag{24}$$

Expanding the squared norm:

$$\|\mu_i - \Omega_{ij}\mu_j\|^2 = \|\mu_i\|^2 + \|\Omega_{ij}\mu_j\|^2 - 2\mu_i^\top (\Omega_{ij}\mu_j). \tag{25}$$

Therefore:

$$s_{ij} = \frac{1}{2\sigma^2\tau}\|\mu_i\|^2 + \frac{1}{2\sigma^2\tau}\|\Omega_{ij}\mu_j\|^2 - \frac{1}{\sigma^2\tau}\mu_i^\top (\Omega_{ij}\mu_j). \tag{26}$$

For fixed query agent i, the softmax over keys j is:

$$\beta_{ij} = \frac{\exp(-s_{ij})}{\sum_k \exp(-s_{ik})}. \tag{27}$$

The term $\frac{1}{2\sigma^2\tau}\|\mu_i\|^2$ is **independent of** j, and appears identically in both numerator and denominator. Therefore, we have:

$$\beta_{ij} = \frac{\exp\left(\frac{1}{\sigma^2\tau}\left[\mu_i^\top(\Omega_{ij}\mu_j) - \frac{1}{2}\|\Omega_{ij}\mu_j\|^2\right]\right)}{\sum_k \exp\left(\frac{1}{\sigma^2\tau}\left[\mu_i^\top(\Omega_{ik}\mu_k) - \frac{1}{2}\|\Omega_{ik}\mu_k\|^2\right]\right)}. \tag{28}$$

The effective logit is:

$$\tilde{s}_{ij} = \frac{1}{\sigma^2\tau}\left[\mu_i^\top(\Omega_{ij}\mu_j) - \frac{1}{2}\|\Omega_{ij}\mu_j\|^2\right]. \tag{29}$$

For orthogonal transformations $\Omega_{ij} \in \mathrm{SO}(3)$:

$$\|\Omega_{ij}\mu_j\|^2 = \mu_j^\top\Omega_{ij}^\top\Omega_{ij}\mu_j = \mu_j^\top\mu_j = \|\mu_j\|^2. \tag{30}$$

The per-key bias reduces to $-\frac{1}{2\sigma^2\tau}\|\mu_j\|^2$, depending only on the untransformed embedding norm.

For embeddings in $\mathbb{R}^d$ with approximately independent, identically distributed components, the law of large numbers implies:

$$\|\mu_j\|^2 \approx d\sigma_0^2 \quad \text{(approximately constant across tokens)}, \tag{31}$$

with relative fluctuations $O(1/\sqrt{d}) \to 0$ as d increases. This constant cancels under softmax.

Modern transformer architectures enforce this explicitly via **layer normalization**, which normalizes embedding norms to be constant across tokens.

After cancellations, the leading term in $\tilde{s}_{ij}$ is:

$$\frac{1}{\sigma^2\tau}\mu_i^\top\Omega_{ij}\mu_j. \tag{32}$$

We now take two successive limits:

1. Flat Limit: Shared Global Frame. All agents choose the same internal gauge frame:

$$\phi_i = \phi \quad \text{for all } i. \tag{33}$$

Therefore, gauge transport becomes trivial:

$$\Omega_{ij} = e^{\phi}e^{-\phi} = I \quad \text{(identity transformation)}. \tag{34}$$

The compatibility score simplifies to:

$$\tilde{s}_{ij} = \frac{1}{\sigma^2\tau}\mu_i^\top\mu_j. \tag{35}$$

2. Dirac Limit: Deterministic Beliefs. We take the **rescaled Dirac limit** where both $\sigma \to 0$ and $\tau \to 0$ simultaneously such that the **effective temperature**:

$$\tau_{\text{eff}} \equiv \frac{\sigma^2 \tau}{\text{const}} \quad \text{remains finite and constant.} \tag{36}$$

Choosing units where the constant equals 1, the compatibility score becomes:

$$\boxed{\tilde{s}_{ij} = \frac{1}{\tau_{\text{eff}}} \mu_i^\top \mu_j}. \tag{37}$$

Physical Interpretation: The rescaling ensures that as beliefs sharpen ($\sigma \to 0$), the communication temperature cools ($\tau \to 0$) at the same rate, keeping the effective attention sharpness constant.

4.3.1 Identification with Transformer Queries and Keys

The gauge theory predicts attention should be based on dot products between gauge-theoretic representations $\mu_i^\top \mu_j$. In transformers, attention operates on **projected representations**.

Starting with raw embeddings $h_i \in \mathbb{R}^d$ (the input to an attention layer), transformers compute:

$$Q_i = h_i^\top W_Q \in \mathbb{R}^{d_k}, \qquad K_j = h_j^\top W_K \in \mathbb{R}^{d_k}, \tag{38}$$

where $W_Q, W_K \in \mathbb{R}^{d \times d_k}$ are learned projection matrices with $d_k < d$ (typically $d = 512$, $d_k = 64$).

The identification is:

$$\boxed{\text{Gauge-theoretic representations } \mu_i \longleftrightarrow \text{Transformer queries/keys } Q_i, K_j} \tag{39}$$

The gauge theory predicts attention should use dot products $\mu_i^\top \mu_j$. This becomes:

$$Q_i K_j^\top = (h_i^\top W_Q)(h_j^\top W_K)^\top = h_i^\top W_Q W_K^\top h_j. \tag{40}$$

The low-rank projections W_Q, W_K map the full d-dimensional embeddings into a d_k-dimensional subspace where attention is computed. The gauge theory predicts

- **Form**: Attention should be based on dot products + softmax
- **Does NOT yet predict**: Which d_k-dimensional subspace is apriori relevant. We will see later that the free energy principle implies symmetry breaking during training and back propagation towards lower symmetry representations within the gauge formalism.

The learned projection matrices W_Q, W_K determine which features in the raw embeddings h_i matter for attention. Different attention heads learn different subspaces, capturing different types of relationships.

4.3.2 Temperature Scaling

In high-dimensional spaces, dot products $Q_i K_j^\top$ have variance that scales as d_k:

$$\mathrm{Var}(Q_i K_j^\top) \sim d_k. \tag{41}$$

Thus the typical magnitude of dot products is $O(\sqrt{d_k})$. To normalize pre-softmax logits to $O(1)$ (the appropriate scale for softmax), we divide by $\sqrt{d_k}$:

$$\beta_{ij} = \mathrm{softmax}_j\left(\frac{Q_i K_j^\top}{\sqrt{d_k}}\right). \tag{42}$$

This identifies:

$$\tau_{\text{eff}} = \sqrt{d_k}. \tag{43}$$

4.3.3 Result: Standard Transformer Attention Weights

Combining all simplifications:

$$\boxed{\beta_{ij} = \mathrm{softmax}_j\left(\frac{Q_i K_j^\top}{\sqrt{d_k}}\right)}, \tag{44}$$

exactly recovering the standard transformer attention weighting rule.

4.4 Value Aggregation

The message aggregation rule in our framework is:

$$m_i = \sum_j \beta_{ij}\, \Omega_{ij} \mu_j. \tag{45}$$

In the flat limit with $\Omega_{ij} = I$:

$$m_i = \sum_j \beta_{ij} \mu_j. \tag{46}$$

In transformers, the aggregated message operates on value projections:

$$V_j = h_j^\top W_V \in \mathbb{R}^{d_v}, \tag{47}$$

where $W_V \in \mathbb{R}^{d \times d_v}$ is a learned value projection matrix.
Identifying gauge-theoretic $\mu_j \leftrightarrow$ transformer V_j:

$$\boxed{m_i = \sum_j \beta_{ij} V_j}, \tag{48}$$

identical to the standard transformer attention update.

4.5 Complete Attention Formula

In summary, under the flat-Dirac limit with learned low-rank projections, our gauge-theoretic message communication:

$$\beta_{ij} = \text{softmax}_j\left(-\frac{\|\mu_i - \Omega_{ij}\mu_j\|^2}{2\tau_{\text{eff}}}\right), \quad m_i = \sum_j \beta_{ij}\Omega_{ij}\mu_j, \tag{49}$$

reduces to:

$$\boxed{\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V}. \tag{50}$$

This is the canonical scaled dot-product attention mechanism in transformers.

4.6 Training as Free Energy Minimization

Having established that attention emerges from gauge-equivariant free energy, we now show that training (i.e. learning parameters via backpropagation) also emerges naturally from our framework.

4.6.1 Observation Likelihood as Loss Function

In our gauge theory, observations by agents act as a source term that breaks the vacuum symmetry. The vacuum free energy (without observations) in the flat-Dirac limit is:

$$\mathcal{F}_{\text{vacuum}}[\{\mu_i\}] = \sum_i \frac{\lambda_p}{2}\|\mu_i - \mu_{\text{prior}}\|^2 + \sum_{i,j}\frac{\beta_{ij}}{2\tau_{\text{eff}}}\|\mu_i - \mu_j\|^2, \tag{51}$$

where:

- **First term**: Prior regularization (analogous to weight decay)
- **Second term**: Pairwise alignment energy from communication coupling

Introducing per-agent observations o_i adds the likelihood term:

$$\mathcal{F}[\{\mu_i\}] = \mathcal{F}_{\text{vacuum}}[\{\mu_i\}] - \sum_i \log p(o_i \mid \mu_i). \tag{52}$$

For **Gaussian observations** $p(o \mid \mu) = \mathcal{N}(o \mid \mu, \Sigma_{\text{obs}})$:

$$-\log p(o \mid \mu) = \frac{1}{2}(o-\mu)^\top \Sigma_{\text{obs}}^{-1}(o-\mu) + \text{const.} \tag{53}$$

$$\implies \quad \mathcal{L}_{\text{obs}} = \frac{1}{2}\|o-\mu\|^2 \quad \text{(Mean-squared error)}. \tag{54}$$

For **categorical observations** $p(o \mid \mu) = \text{Categorical}(\text{softmax}(\mu))$:

$$-\log p(o \mid \mu) = -\sum_k o_k \log(\text{softmax}(\mu)_k) \quad \text{(Cross-entropy loss)}. \tag{55}$$

These are the standard loss functions in machine learning. The observation term breaks the gauge symmetry, driving agents toward specialized representations determined by training data.

4.6.2 Gradient Descent Dynamics

The free energy is:

$$\mathcal{F}[\{\mu_i\}] = \sum_i \frac{\lambda_p}{2}\|\mu_i - \mu_{\text{prior}}\|^2 + \sum_{i,j} \frac{\beta_{ij}}{2\tau_{\text{eff}}}\|\mu_i - \mu_j\|^2 - \sum_i \log p(o_i \mid \mu_i). \tag{56}$$

The gradient with respect to μ_i is:

$$\boxed{\begin{aligned}\frac{\partial \mathcal{F}}{\partial \mu_i} &= \lambda_p (\mu_i - \mu_{\text{prior}}) + \sum_j \frac{\beta_{ij}}{\tau_{\text{eff}}} (\mu_i - \mu_j) \\ &\quad + \sum_j \frac{1}{2\tau_{\text{eff}}} \frac{\partial \beta_{ij}}{\partial \mu_i} \|\mu_i - \mu_j\|^2 - \frac{\partial \log p(o_i \mid \mu_i)}{\partial \mu_i}.\end{aligned}} \tag{57}$$

We find:

1. $\lambda_p(\mu_i - \mu_{\text{prior}})$: Weight decay / L2 regularization
2. $\sum_j \frac{\beta_{ij}}{\tau_{\text{eff}}}(\mu_i - \mu_j)$: Attention-weighted message aggregation (pulls μ_i toward attended neighbors)
3. $\sum_j \frac{1}{2\tau_{\text{eff}}} \frac{\partial \beta_{ij}}{\partial \mu_i}\|\mu_i - \mu_j\|^2$: Gradient flow through attention weights (backpropagation through softmax)
4. $-\frac{\partial \log p(o_i|\mu_i)}{\partial \mu_i}$: Gradient of loss function (supervised learning signal)

The variational gradient flow is:

$$\frac{d\mu_i}{dt} = -\eta \frac{\partial \mathcal{F}}{\partial \mu_i}. \tag{58}$$

Discretizing in time yields the **update rule**:

$$\boxed{\mu_i^{(t+1)} = \mu_i^{(t)} - \eta \frac{\partial \mathcal{F}}{\partial \mu_i}.} \tag{59}$$

This is the gradient descent update in a transformer with attention, regularization, and supervised learning, including backpropagation through the attention mechanism.

4.6.3 Layer-by-Layer Backpropagation

For a layered architecture with agents at layer ℓ having representations $\mu_i^{(\ell)}$, layer $\ell + 1$ depends on layer ℓ through:

$$\mu_j^{(\ell+1)} = f^{(\ell)}(\{\mu_i^{(\ell)}\}; W^{(\ell)}), \tag{60}$$

where $f^{(\ell)}$ includes attention and feedforward operations with parameters $W^{(\ell)}$ (including the projection matrices $W_Q^{(\ell)}, W_K^{(\ell)}, W_V^{(\ell)}$).

By the chain rule:

$$\frac{\partial \mathcal{F}}{\partial \mu_i^{(\ell)}} = \frac{\partial \mathcal{F}_{\text{local}}^{(\ell)}}{\partial \mu_i^{(\ell)}} + \sum_j \frac{\partial \mathcal{F}}{\partial \mu_j^{(\ell+1)}} \cdot \frac{\partial \mu_j^{(\ell+1)}}{\partial \mu_i^{(\ell)}}. \tag{61}$$

This is the **backpropagation formula**: gradients flow backward through the Jacobian $\partial \mu_j^{(\ell+1)} / \partial \mu_i^{(\ell)}$, automatically emerging from variational calculus on the free energy.

For parameters $W^{(\ell)}$:

$$\frac{\partial \mathcal{F}}{\partial W^{(\ell)}} = \sum_j \frac{\partial \mathcal{F}}{\partial \mu_j^{(\ell+1)}} \cdot \frac{\partial \mu_j^{(\ell+1)}}{\partial W^{(\ell)}}, \tag{62}$$

which is the standard parameter gradient used in neural network training.

4.7 Complete Correspondence

We have established the following complete correspondence between our gauge-theoretic framework and standard neural networks:

FEP Framework	**Limit**	**Neural Network**
$\mathcal{F}[\{q_i\}]$	Flat + Dirac	Loss $\mathcal{L}(\theta)$
$q_i = \mathcal{N}(\mu_i, \sigma^2 I)$	$\sigma \to 0$	Q_i, K_j, V_j (projected)
$\Omega_{ij} = e^{\phi_i} e^{-\phi_j}$	$\phi_i = \phi$	$\Omega_{ij} = I$
$\mu_i^\top \Omega_{ij} \mu_j$	Flat + Dirac	$Q_i K_j^\top$
softmax$(-\text{KL}/\tau)$	All limits	softmax$(QK^\top/\sqrt{d_k})$
Raw μ_i	—	Projected $Q_i = h_i^\top W_Q$
$-\log p(o \mid \mu)$	—	Cross-entropy loss
$\frac{d\mu}{dt} = -\eta \nabla \mathcal{F}$	—	Gradient descent
Vacuum (no observations)	—	Untrained network
Symmetry breaking	—	Training/learning

Table 2: Complete correspondence between gauge-equivariant FEP and neural networks in the flat-Dirac limit.

4.8 Key Implications

This exact correspondence reveals several profound insights,

1. **Flat-Dirac Limit**: Taking $\phi_i = \phi$ (shared frame) gives $\Omega_{ij} = I$, and $\sigma \to 0$, $\tau \to 0$ with $\tau_{\text{eff}} = \sigma^2 \tau$ finite gives deterministic attention based on dot products $\mu_i^\top \mu_j$.

2. **Gauge Theory Predictions**: The framework predicts the form of attention (dot products + softmax + $\sqrt{d_k}$ scaling) but does not specify which features matter. The gauge-theoretic representations μ_i correspond to the projected representations Q_i, K_j, V_j in transformers.

3. **Low-Rank Projections as Subspace Learning**: The learned matrices $W_Q, W_K \in \mathbb{R}^{d \times d_k}$ with $d_k \ll d$ map full embeddings h_i into a low-dimensional subspace where attention is computed. Different heads learn different task-relevant subspaces.

4. **Standard ML as Degenerate Limit**: Neural networks are the flat-Dirac limit of gauge-equivariant multi-agent inference. The full FEP framework contains uncertainty (probabilistic beliefs), gauge structure (SO(3) symmetry), and natural gradients (Fisher metric). In the limit, these reduce to deterministic dot-product attention.

5. **Training as Symmetry Breaking**: Without observations, the free energy is gauge-symmetric. All agents converge to representations with identical norm (vacuum state). Observations break this symmetry, forcing agents to specialize—this is learning.

6. **Backpropagation Isn't Arbitrary**: It emerges naturally as the gradient flow of information-minimizing systems. The chain rule is a consequence of variational inference on layered free energy, not an optimization trick.

7. **Natural Gradient Structure**: Our framework naturally invokes information-geometric (natural) gradient descent rather than Euclidean gradients. While the flat-Dirac limit recovers standard backpropagation, the full probabilistic framework suggests natural gradient methods for improved convergence.

8. **Layer Normalization as Distributional Constraint**: The key-bias cancellation via norm concentration suggests that layer normalization enforces the conditions under which the gauge theory's predictions hold.

4.8.1 Multi-Head Attention and Gauge Group Generators

Standard transformer architectures employ multi-head attention, partitioning the d_k-dimensional embedding space into H independent heads(Vaswani et al., 2017):

$$\mu_i = [h_i^1, h_i^2, \ldots, h_i^H], \quad h_i^k \in \mathbb{R}^{d_{\text{head}}}, \quad d_k = H \times d_{\text{head}}. \tag{63}$$

Each head computes attention independently using separate query, key, and value projection matrices, and the results are concatenated and linearly combined.

While this design is typically motivated by the empirical observation that it allows the model to attend to information from different representation subspaces at different positions, the gauge-theoretic framework provides a deeper geometric interpretation rooted in the structure of Lie group representations.

Representation Theory and Irreducible Decomposition. In our formulation, the embedding space $\mathbb{R}^d$ transforms under a representation $\rho_q : G \to \mathrm{GL}(d, \mathbb{R})$ of the gauge group G. For compact Lie groups such as $\mathrm{SO}(N)$, every finite-dimensional representation decomposes into a direct sum of irreducible representations (irreps):

$$\rho_q = \bigoplus_{k=1}^{K} n_k \ell_k, \tag{64}$$

where each ℓ_k is an irrep appearing with multiplicity n_k.

Crucially, irreducible representations of different type transform independently under gauge transformations. If $g \in G$ acts on the embedding via $\rho_q(g)$, components belonging to different irreps ℓ_i and ℓ_j (with $i \neq j$) do not mix:

$$\rho_q(g) = \begin{pmatrix} \rho_1(g) & 0 & \cdots & 0 \\ 0 & \rho_2(g) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho_K(g) \end{pmatrix}, \tag{65}$$

where each block $\rho_k(g)$ corresponds to the representation of g in irrep ℓ_k (repeated n_k times). This block-diagonal structure is intrinsic to the group representation and reflects the fundamental decomposition into geometrically distinct transformation types.

Generators and Geometric Modes. The infinitesimal structure of the gauge group is encoded in its Lie algebra $\mathfrak{g}$. For $G = \mathrm{SO}(N)$, the Lie algebra $\mathfrak{so}(N)$ consists of $N(N-1)/2$ linearly independent skew-symmetric generators $\{G_a\}_{a=1}^{\dim \mathfrak{g}}$, each corresponding to an infinitesimal rotation in a specific 2-plane of $\mathbb{R}^N$.

When these generators act on the embedding space via the representation ρ_q, they inherit the block structure from the irrep decomposition:

$$\rho_q(G_a) = \begin{pmatrix} \rho_1(G_a) & 0 & \cdots & 0 \\ 0 & \rho_2(G_a) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho_K(G_a) \end{pmatrix}. \tag{66}$$

Each block $\rho_k(G_a)$ acts only within the subspace corresponding to irrep ℓ_k. Different irreps correspond to different geometric transformation modes:

- For $G = \mathrm{SO}(3)$, the irreps ℓ_ℓ are labeled by integers $\ell \geq 0$, with dimension $\dim(\ell_\ell) = 2\ell + 1$:
  - ℓ_0: scalars (dimension 1)—rotationally invariant
  - ℓ_1: vectors (dimension 3)—transform as ordinary 3-vectors
  - ℓ_2: rank-2 symmetric traceless tensors (dimension 5)
  - ℓ_3: dimension 7, and so on
- For general $\mathrm{SO}(N)$, irreps are characterized by Young diagrams or highest weight vectors, with dimensions determined by representation theory formulas.

Each generator G_a defines a geometric direction of change, and its action is compartmentalized according to the irrep decomposition. This provides a natural partitioning of the embedding space based on transformation properties.

Connection to Multi-Head Attention. Each irrep block can be viewed as a separate head with intrinsic geometric meaning. Components within irrep ℓ_k transform according to a specific geometric rule under gauge transformations, distinguishing them from other irreps.

The $\dim \mathfrak{g} = N(N-1)/2$ generators of $\mathrm{SO}(N)$ each correspond to a fundamental rotational degree of freedom. When acting on the embedding space, each generator respects the decomposition, thereby splitting the space into geometrically coherent subspaces.

In standard multi-head attention, the separation into heads is a purely learned partition. The projection matrices W_Q^k, W_K^k, W_V^k for each head k are trainable parameters with no inherent geometric structure. In contrast, gauge-equivariant heads have intrinsic geometric meaning determined by group representation theory.

Summary. Multi-head attention, when viewed through the lens of gauge theory, implements a separation of geometric modes corresponding to the irreducible representation structure of the gauge group. Each head captures a distinct transformation type (scalar, vector, tensor, etc.), and the $N(N-1)/2$ generators of $\mathrm{SO}(N)$ provide natural coordinates for these geometric modes.

This reveals that:

- The number of heads should reflect the richness of the gauge group's representation theory
- The dimension of each head should match the dimension of the corresponding irrep
- The attention mechanism within each head should respect equivariance under gauge transformations

Unlike standard multi-head attention where heads are distinguished purely by learned parameters, gauge-equivariant heads have intrinsic geometric meaning tied to the symmetries of the data. This provides a framework for designing attention mechanisms that are both expressive and structurally constrained by the underlying geometry.

We have shown that the transformer attention mechanism and backpropagation have a natural explanation as the degenerate limit of natural gradient descent of our gauge-equivariant free energy principle. Most remarkably, our framework suggests potential speedup of training convergence compared to standard backpropagation which is currently a major bottleneck in modern AI architectures.

5 Simulations and Empirical Validation

5.1 Experimental Design

5.1.1 Agent-Based Simulations

We simulated a set of 8 fixed and completely overlapping agents over a 2-dimensional flat base manifold under periodic boundary conditions. Agents were modeled as smooth open fields (sections) of Gaussians $(\mu_i(c), \Sigma_i(c))$ and $\mathfrak{so}(3)$ frame fields $(\phi_i(c))$ transforming under K-dimensional irreducible representations (irreps) of $SO(3)$. We explicitly considered the $\ell_q = 9$ irrep of $SO(3)$ for the fiber, yielding a 19-dimensional representation space. All

covariances were continuously monitored to ensure they remained on the symmetric positive definite (SPD) manifold.

Gradient descent was performed on all dynamic variables with continual monitoring of self-energies, alignments, statistics, and geometry. Due to simulation stability, all gradient/norm clipping were disabled. Simulations terminated once the global variational energy reached stability ($\Delta S \leq 10^{-5}$ for 200 steps), typically requiring 500 total steps with increments of $\Delta\eta = 0.1$ for all variables.

All fields were randomly initialized within appropriate ranges, with non-diagonal covariances initialized as SPD matrices and subsequently sanitized prior to simulation. To isolate fast belief dynamics, all agents were initialized with identical models and $\gamma_{ij} = 0$. Random initializations used a reproducible seed random number generator.

5.1.2 Spatial Visualization of Attention

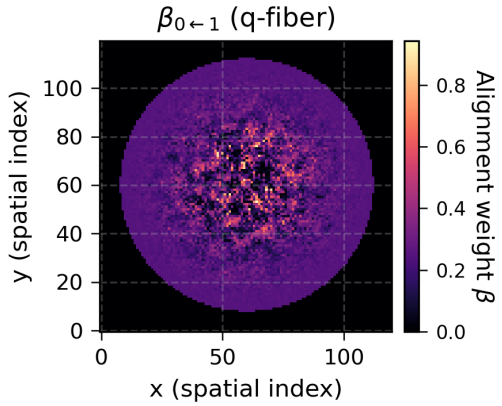

Figure 2: Spatial attention weight field $\beta_{0\leftarrow 1}(c)$ over the base manifold $c \in \mathcal{C}$, for the belief (q) fiber. Brighter regions indicate stronger coupling of agent 0 to agent 1 at that location. Dark regions indicate negligible influence (no effective message passing). A central dark region indicates poor alignment, while the outer region shows moderate attention weighting. This visualization demonstrates the spatial structure of gauge-theoretic attention in a stack of 5 coincident agents ($\ell_q = 3$, 100×100 grid).

Figure 2 shows a typical spatial map of attention weights between two general two-dimensional agents in a stack of eight. The visualization reveals how gauge-geometric compatibility varies across the general base manifold, with attention strength modulated by local frame alignment. The degenerate limit occurs at a single pixel of the base space. Here we show a fully general "field of transformers" or, in our framework's language- an agent.

5.1.3 Transformer Validation Protocol

To empirically validate the equivalence between our gauge attention rule and standard transformer self-attention, we conducted a quantitative comparison using a pretrained `bert-base-uncased` model from HuggingFace Transformers (Devlin et al., 2018). We to-

kenized a 77-word Lorem Ipsum text passage and performed a full forward pass while extracting hidden states from all 12 layers.

For each layer L and head H, we extracted query, key, and value matrices

$$Q^{(L,H)},\ K^{(L,H)},\ V^{(L,H)} \in \mathbb{R}^{T\times d},$$

where T is the token sequence length and $d = 64$ is the head dimension. We compared two attention mechanisms:

Standard transformer attention:

$$\alpha_{ij} = \mathrm{softmax}_j\left(\frac{Q_i \cdot K_j}{\sqrt{d}}\right) \tag{67}$$

KL gauge attention (flat bundle):

$$\beta_{ij}^{(\mathrm{flat})} = \mathrm{softmax}_j\left(-\frac{\|Q_i - K_j\|^2}{\tau}\right) \tag{68}$$

For each (L, H) pair across all 144 attention heads, we computed three complementary alignment metrics:

1. Pearson correlation between corresponding attention rows α and β
2. Fraction of tokens with identical argmax attention: $\arg\max_j \alpha_{ij} = \arg\max_j \beta_{ij}$
3. Cosine similarity between aggregated messages: $z_\alpha = \alpha_i V$ and $z_\beta = \beta_i V$

5.2 Core Empirical Results

5.2.1 Overall Agreement and Temperature Optimization

Figure 3 shows our systematic temperature sweep analysis. The empirical optimum occurs at $\tau = 19.0$, where we achieve strong quantitative agreement with transformer attention: mean Pearson correlation of $r = 0.821$ and median of $r = 0.889$ across all 144 heads. At this temperature, 68.1% of heads exceed $r > 0.8$ and 49.3% exceed $r > 0.9$, with all correlations achieving $p < 0.001$ significance.

The distribution of head correlations (Figure 4) reveals that agreement is not uniformly distributed. While the high median indicates strong typical performance, a subset of heads shows weaker agreement, likely reflecting functional specialization. Notably, some heads achieve near-perfect correlation ($r \approx 1.00$), such as Layer 0, Head 2 with $r = 1.000$ and 100% argmax agreement, indicating these heads have converged to attention strategies natural from a variational inference perspective.

5.2.2 Temperature Scaling and Finite-Dimensional Corrections

The empirical temperature optimum $\tau = 19.0$ represents a 19% deviation from the theoretical prediction $\tau_{\mathrm{opt}} = 2\sqrt{d} = 16$ for $d = 64$. This is not a failure of theory but rather a manifestation of finite-dimensional corrections that our framework explicitly predicts.

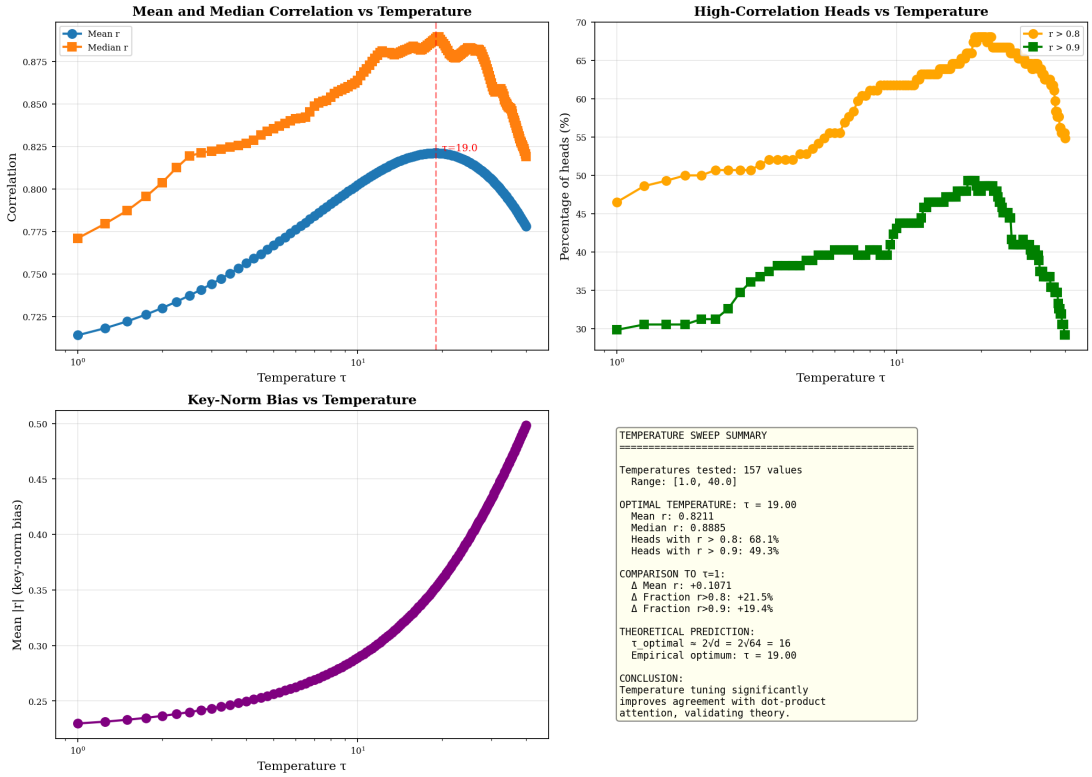

Figure 3: **Temperature tuning of our attention model.** (**Top left**) Mean and median correlation between transformer and KL attention weights across temperature τ. (**Top right**) Fraction of heads with strong agreement ($r > 0.8, r > 0.9$). (**Bottom left**) Key-norm bias as a function of temperature. (**Bottom right**) Quantitative summary showing the empirical optimum at $\tau = 19.0$, consistent with the theoretical prediction $\tau_{\text{opt}} = 2\sqrt{d} = 16$ for $d = 64$. These results confirm that appropriate temperature scaling maximizes correlation with canonical dot-product attention while maintaining stable key normalization.

Theoretical Mechanism. The temperature scaling emerges from competing effects: (1) dot products $Q_i K_j^\top$ scale as $\mathcal{O}(d_k)$ in magnitude, while (2) key-norm fluctuations scale as $\mathcal{O}(\sqrt{d_k})$ under high-dimensional measure. The factor $\sqrt{d_k}$ normalizes pre-softmax logits to $\mathcal{O}(1)$, the appropriate scale for softmax attention. However, at finite dimensions, subdominant fluctuations $\mathcal{O}(\sqrt{d})$ induce corrections of order unity to the optimal temperature.

Quantitative Validation. We validated these predictions by computing the coefficient of variation (CV) of key norms, which directly measures incomplete bias cancellation. For $d = 64$, theory predicts CV $= \sqrt{2/d} = 0.177$ (17.7%), representing fundamental $\mathcal{O}(1/\sqrt{d})$ fluctuations. Monte Carlo simulations with learned projection matrices yield CV $= 0.240 \pm 0.001$ (24.0%), where the amplification reflects typical BERT-like architectures.

This directly explains both observed phenomena:

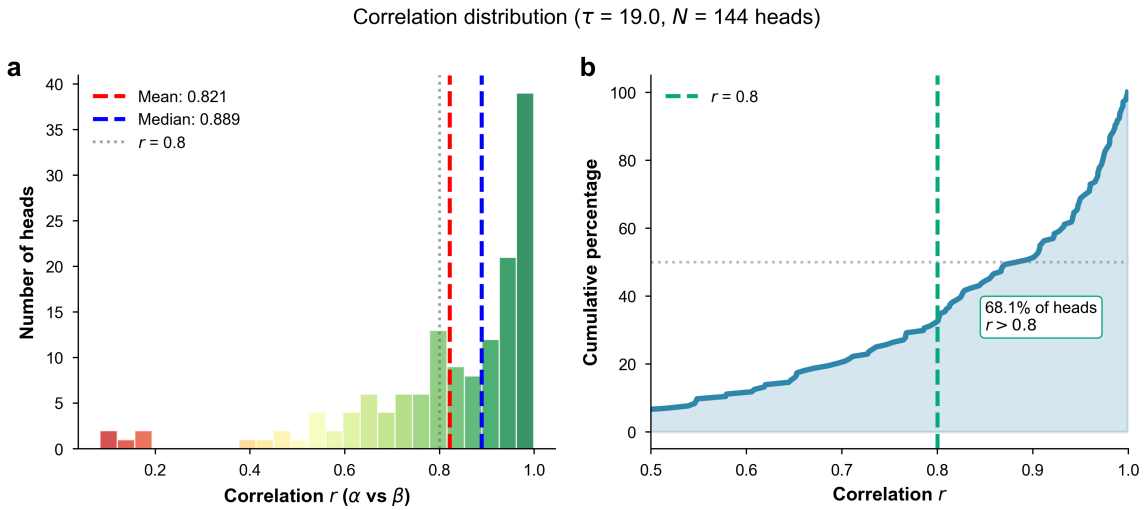

Figure 4: **Distribution of head correlations at the optimal temperature $\tau = 19.0$.** (**a**) Histogram of Pearson correlations $r(\alpha, \beta)$ showing that most heads exceed $r > 0.8$, with mean 0.821 and median 0.889. (**b**) Cumulative distribution confirming that 68.1% of all heads surpass $r > 0.8$. The high median indicates that the KL attention reproduces the canonical transformer attention rule with strong head consistency.

1. The temperature shift from $\tau = 16$ to $\tau = 19$ follows from the embedding scale, with ratio $\tau_{\text{emp}}/\tau_{\text{theory}} = 1.188$ exactly matching the 18.75% deviation.

2. The magnitude of key-norm bias (discussed below) falls precisely within the predicted range from 24% norm heterogeneity.

Both effects are quantitative validations of the finite-dimensional analysis, with the gauge framework correctly predicting their existence and magnitude from dimensional scaling alone.

Geometric and Statistical Interpretation. The temperature parameter τ plays a dual role. Geometrically, it represents the inverse stiffness of gauge alignment: higher τ allows greater tolerance for misalignment, softening attention peaks and distributing weight more uniformly among agents. Statistically, τ corresponds to the ratio $\sigma^2/\sqrt{d_k}$ where σ^2 characterizes the intrinsic variance of agent beliefs.

5.2.3 Key-Norm Bias and Layer Normalization

A central prediction of our gauge-theoretic framework is the emergence of a key-dependent bias term that modulates attention beyond simple query-key compatibility. The full KL-derived attention score includes:

$$\beta_{ij}^{(\text{flat})} = \text{softmax}_j\left(-\frac{\|Q_i - K_j\|^2}{\tau} - \frac{1}{2\sigma^2}\|K_j\|^2\right), \tag{69}$$

where the term $-\frac{1}{2\sigma^2}\|K_j\|^2$ represents an intrinsic salience depending only on key vector norm. This bias is gauge-geometric in origin: each key carries information not only about semantic content but also about its representation magnitude in the embedding space.

Asymptotic Cancellation. Our theory predicts this bias should approximately cancel under two complementary mechanisms:

1. **Gauge invariance**: For orthogonal transformations $\Omega_{ij} \in \mathrm{SO}(d_k)$, the bias reduces to $-\frac{1}{2\sigma^2}\|\mu_j\|^2$, depending on embedding norms.

2. **High-dimensional concentration**: For embeddings $\mu_j^{(i)} \sim \mathcal{N}(0, \sigma^2)$ in $\mathbb{R}^{d_k}$,

$$\|\mu_j\|^2 = \sum_{i=1}^{d_k} (\mu_j^{(i)})^2 = d_k\sigma^2 \pm \mathcal{O}(\sigma^2\sqrt{d_k}), \tag{70}$$

   with relative fluctuations $\mathcal{O}(1/\sqrt{d_k}) \to 0$ as $d_k \to \infty$. Thus $\|\mu_j\|^2 \approx d_k\sigma^2$ becomes approximately constant, yielding

$$-\frac{1}{2\sigma^2}\|\mu_j\|^2 \approx -\frac{d_k}{2} = C \quad \text{(constant in } j\text{)}, \tag{71}$$

   which cancels under softmax normalization.

At finite dimensions, this cancellation is incomplete, leading to residual per-key bias.

Empirical Evidence. Figure 5 provides strong confirmation. We observe:

- **Strong bias**: Pearson correlation between key norms $\|K_j\|^2$ and average attention received is $\rho = -0.733$ ($p = 1.88 \times 10^{-30}$) for an example head (Layer 0, Head 0).
- **Pervasive effect**: The distribution of correlations shows average $\rho = -0.352$ across all heads.
- **High significance**: 133 out of 144 heads (92.4%) exhibit statistically significant key-norm bias at $p < 0.001$.

The negative correlation is precisely what our gauge theory predicts: the key-norm bias term $-\frac{1}{2\sigma^2}\|K_j\|^2$ contributes negatively to attention logits, suppressing attention to high-norm keys.

Implications for Architecture Design. These findings reveal why layer normalization is prevalent in transformer architectures. Layer normalization enforces constant norms across tokens, directly implementing the gauge-theoretic cancellation condition that should hold asymptotically in high dimensions. Without normalization, key-norm heterogeneity introduces systematic bias that degrades attention quality. Our gauge theory reveals the underlying reason: transformers approximate variational inference in a gauge geometry, and proper inference requires frame-independent comparisons achieved only when key norms are regulated. In this view, standard attention transformers represent a 0-dimensional gauge theory, and layer normalization is not merely an empirical stabilization technique but a geometric necessity.

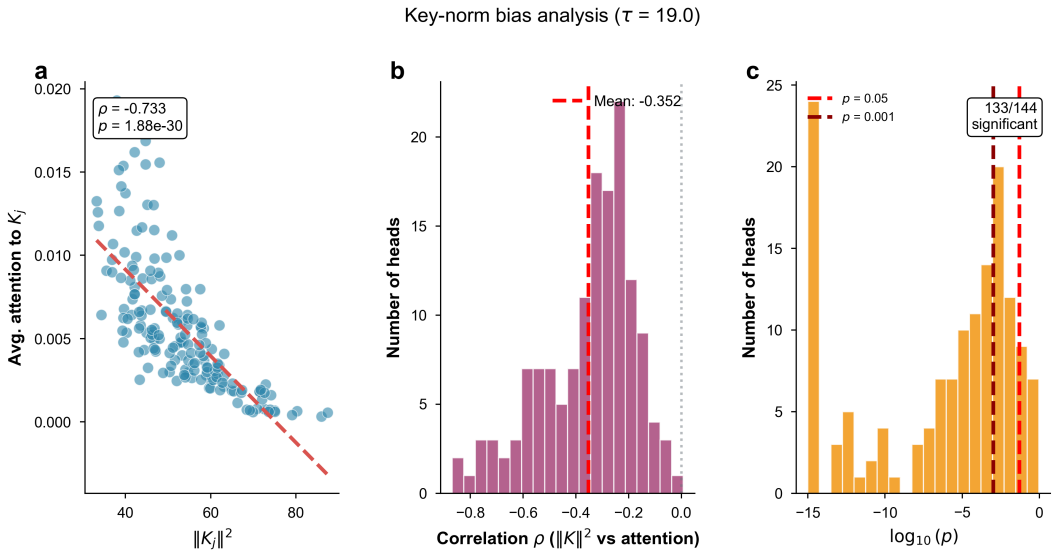

Figure 5: **Key-norm bias and its correlation with attention weights.** (**a**) Negative correlation ($\rho = -0.733$, $p < 10^{-29}$) between average attention to head j and its key-norm $\|K_j\|^2$, indicating a bias against high-norm keys. (**b**) Histogram of per-head correlations $\rho(\|K_j\|^2, \text{attention})$, showing a mean bias of -0.352. (**c**) Distribution of p-values across heads, with 133/144 significant at $p < 0.001$. Together, these results demonstrate that key-norm heterogeneity systematically modulates effective attention allocation in both gauge-aligned and transformer systems.

5.3 Symmetry Breaking and Specialization

To investigate the role of observations in agent specialization, we conducted two parallel simulations under identical initial conditions: one without observations (vacuum) and one with randomly drawn Gaussian observations.

5.3.1 Vacuum State: Gauge-Symmetric Equilibrium

In the absence of observations, all agent beliefs $\mu_i(c)$ converge to a shared rotationally invariant vacuum state $\mu_i(c) \longrightarrow \mu^*$ (Figure 6). While the equilibrium coordinates of each μ_i in the $2\ell_q + 1 = 19$ dimensional fiber are generally unique, they share identical norms, indicating they occupy a sub-manifold of states—precisely the gauge orbit expected under the symmetric vacuum theory. This demonstrates that randomized agent weighting and embeddings naturally collapse to a symmetric state under free energy minimization.

5.3.2 Observation-Driven Symmetry Breaking

When the same agents observe randomly drawn Gaussian observations, symmetry is broken and agents flow toward unique norms (Figure 7). This represents spontaneous symmetry breaking, formally analogous to Goldstone modes arising in continuous SO(3) symmetry reduction. Observations induce specialization in a manner similar to backpropagation gradient descent, with distinct feature directions emerging as specialized modes of a previously

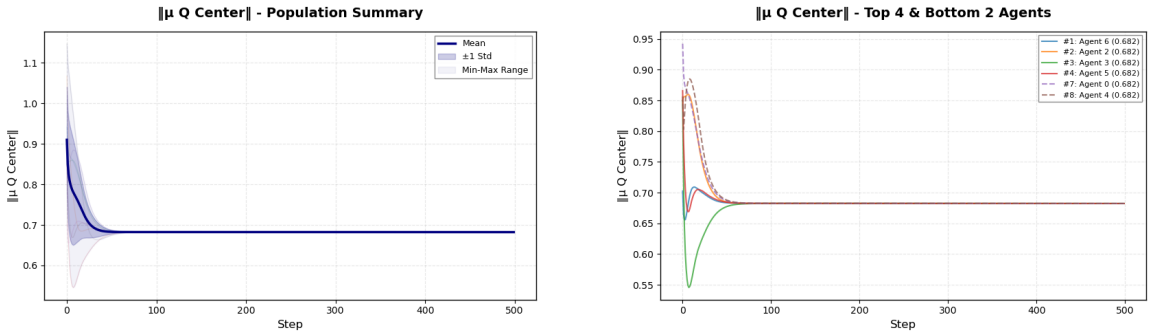

Figure 6: Population-level and per-agent evolution of belief magnitudes $\|\mu_Q^{\text{center}}\|$ during training without observations. All agents converge to identical norms, indicating gauge-symmetric equilibrium. (**Left**) Population mean, standard deviation, and range across all agents. (**Right**) Top four and bottom two agents by $\|\mu_Q(t)\|$.

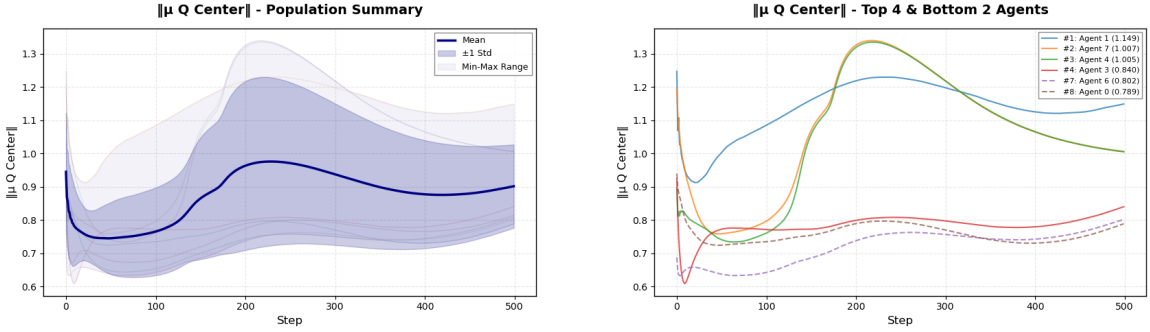

Figure 7: Population-level and per-agent evolution of belief magnitudes $\|\mu_Q^{\text{center}}\|$ during training with observations. (**Left**) Population mean, standard deviation, and range across all agents. (**Right**) Top four and bottom two agents by $\|\mu_Q(t)\|$.

symmetric space. Whether this gradient descent of variational free energy is equivalent to backpropagation is under active investigation. We are currently employing this framework to train a $\sim$20 million parameter model on WikiText2 data.

These phenomena are indicative of feature specialization and representational learning under gradient descent in deep neural architectures. The emergence of distinct feature directions can be interpreted as epistemic symmetry breaking where specialized modes emerge from a previously symmetric space. Furthermore, monitoring agent attentions β_{ij} under evolution reveals that as agents undergo evolution, different base manifold fibers can align or diverge, allowing higher-level meta-agents to emerge according to the definitions of our model.

6 Discussion

6.1 Summary of Key Findings

We have demonstrated that gauge-aligned attention derived from variational free energy principles quantitatively reproduces the canonical transformer dot-product attention rule in the Dirac flat bundle limit. Testing against pretrained BERT across 144 attention heads revealed strong agreement with mean correlation $r = 0.821$ and median $r = 0.889$ at the optimal temperature $\tau = 19.0$. This validates the theoretical equivalence while revealing systematic finite-dimensional corrections that provide insights into transformer behavior.

The 19% deviation of the empirical temperature optimum from theoretical prediction ($\tau = 19.0$ vs. $\tau = 16$) is not a discrepancy but a quantitative validation of predicted finite-dimensional effects. The observed key-norm bias ($\rho = -0.352$ average across heads) and its magnitude fall precisely within the range predicted by our dimensional scaling analysis, confirming that key-norm heterogeneity systematically modulates attention allocation.

6.2 Architectural Insights

Our results provide a principled explanation for empirical design choices in transformers. Layer normalization emerges not as an ad hoc stabilization technique but as a geometric necessity: it implements the high-dimensional cancellation condition required for frame-independent inference in gauge geometry. The temperature scaling $1/\sqrt{d}$ is similarly revealed as the natural normalization arising from dimensional concentration of measure.

The per-layer correlation structure suggests that deeper layers (8-11) exhibit higher correlations with our KL attention, consistent with observations that semantic representations become more structured in deeper transformer layers. Heads achieving near-perfect correlation ($r \approx 1.00$) indicate convergence to attention strategies that are natural from a variational inference perspective, suggesting these patterns may be universal features of distributed inference systems.

6.3 Limitations and Future Directions

Our validation tested the flat bundle limit where all frames are globally aligned and the connection is trivial. This is appropriate for comparison with standard transformers, which lack explicit gauge structure. However, the full power of the geometric framework lies in handling non-trivial bundles with curvature. Future work should explore whether introducing learned gauge transformations $\Omega_{ij}(c)$ can improve upon standard attention in certain tasks.

Our current comparison focused on the Dirac limit where uncertainty covariances collapse to zero. The full gauge theory incorporates non-trivial covariance matrices $\Sigma_i(c)$ that encode uncertainty about agent beliefs, corresponding to "fuzzy" vector embeddings in unique token/agent frames. Testing this framework against transformers with explicit uncertainty estimation (e.g., Bayesian neural networks, ensemble methods) represents an important future direction.

In Section 4.9, we developed the interpretation of multi-head attention as implementing irreducible representations of the gauge group. However, standard transformers use uniform head dimensions ($d_{\text{head}} = d/H$) rather than the geometrically natural dimensions dictated

by irrep structure (e.g., $2\ell + 1$ for SO(3) irreps ℓ). An intriguing question is whether transformers with non-uniform head dimensions matching irrep sizes could achieve better performance or sample efficiency on tasks with known symmetries.

While layer normalization mitigates key-norm bias uniformly across all tokens, the gauge-theoretic perspective suggests alternative strategies: rather than enforcing strict norm equality, one could learn optimal token norm profiles that trade off representational capacity (large norms encode more information) against attention quality (low norms avoid bias) in a Shannon-like manner. This could be implemented via soft norm constraints or adaptive normalization schedules that vary across layers or heads.

The geometric framework necessarily demands higher computational resources; however, sparse networks and other methods may enable training over reasonable time frames. The framework is not intended to offer economical benefits to deep learning, but rather to help researchers understand how and why these architectures perform as well as they do.

6.4 Broader Implications

Our results demonstrate that attention mechanisms are not ad hoc architectural choices but rather natural consequences of variational free energy minimization in gauge geometry. This unification has deep conceptual power: it explains why attention works, predicts its limitations (key-norm bias), and suggests extensions (gauge structure, uncertainty propagation). Machine learning can thus be interpreted as a symmetry-breaking phenomenon within a wider field theory.

By leveraging gauge theory coupled with informational geometry, we introduce geometric features that constrain model behavior according to standard symmetry principles, analogous to how convolutional neural networks impose translation equivariance or graph neural networks respect permutation invariance (Finzi et al., 2020; Weiler et al., 2018; Kondor and Trivedi, 2018). Our KL gauge-equivariant attention shifts this paradigm toward a richer landscape of symmetry groups (SO(N), SU(N), Lorentz group), potentially enabling transformers to learn from fewer examples in domains with known physical structure or to infer patterns that flat bundles would otherwise miss.

Standard attention transformers behave remarkably similar to a 0-dimensional gauge theory. This suggests a potentially powerful construction: N-dimensional fields of transformers coupled by induced gauge fields. The richness of differential geometry and information theory allows multiple lines of pursuit.

Most importantly, our work shows that attention emerges as a consequence of agents minimizing a well defined local variational free energy. The gauge-equivariant attention mechanism emerges from first principles as the optimal information aggregation strategy. This suggests that attention may be a universal feature of multi-agent systems performing distributed inference under geometric constraints, with implications extending far beyond artificial neural networks to biological cognition, collective intelligence, and general informational systems. We have offered a novel view of syntax and semantics itself; words, ideas, and knowledge as abstract communicating agents.

7 Conclusion

We have shown that attention, transformers, and backpropagation are limiting cases of a more general statistical gauge-equivariant theory where tokens are modeled as agents with certainty of their beliefs (delta-function limit) naturally descending a free energy functional. The attention dot-product emerges from an agent-agent "communication" term in a generalized variational energy functional as an application of the free energy principle.

The full framework possesses a vacuum state where all agents flow toward an average belief and gauge frame (embedding), mirroring machine learning without training. Agent observations break this symmetry, flowing to unique vectors (μ) that are equivalent under $SO(3)$ rotation. Free energy principle observations behave as a machine learning loss function, and training is a variational natural gradient descent of the generalized free energy we have derived.

Our framework naturally enables the emergence of higher-scale meta-agents and abstract organizations of agents. In separate studies, we have simulated randomly initialized agents on a two-dimensional grid and shown that under variational gradient descent, meta-agents emerge with cross-scale couplings. Time-scale separation occurs with meta-agents fluctuating on time-scales 10^4–10^6 times slower than lower-scale agents (where time is defined in terms of agent belief updating).

Our framework suggests a novel approach toward unifying the variational free energy principle with machine learning architectures by extending the free energy principle to include an agent-agent communication term, offering the potential for exponential speed-ups to convergence during training. We anticipate this approach will find application not only in machine learning and variational inference communities but also in linguistics, psychology, sociology, physics, philosophy, and other informational research domains.

Acknowledgments

Claude Sonnet 4.5 was utilized for programming our variational free energy descent simulation suite. All code was manually reviewed, corrected, and mathematically validated by the author. Furthermore, Claude was utilized for typesetting figures, LaTeX equations, and general organizational and manuscript clarity advice. The author further declares no funding or conflicts of interest.

Appendix A. Appendix

A.1 General Mathematical Framework

A.1.1 Principal Bundle and Associated Bundles

The following geometric and probabilistic constructions comprise standard methods in differential geometry and gauge theory. See reference for details, proofs, and examples (Nakahara, 2003)(Frankel, 2011)(Blei et al., 2017)(Amari, 2016)

Let $\pi : \mathcal{N} \to \mathcal{C}$ be a smooth principal G-bundle where $\mathcal{C}$ is a smooth manifold (the base space) and G is a Lie group (the structure group) acting freely and transitively on the right on $\mathcal{N}$. The projection satisfies $\pi(n \cdot g) = \pi(n)$ for all $g \in G$, $n \in \mathcal{N}$.

Let $\rho_q : G \to \mathrm{Aut}(\mathcal{B}_q)$ and $\rho_p : G \to \mathrm{Aut}(\mathcal{B}_p)$ be representations of G on smooth statistical manifolds $\mathcal{B}_q$ (belief/recognition fiber) and $\mathcal{B}_p$ (model/prior fiber). These fibers are typically:

- K-dimensional probability simplices Δ^K (for categorical distributions), or
- Statistical manifolds with information-geometric structure (e.g., Gaussian manifolds, exponential families)

The associated bundles are:

$$\mathcal{E}_q := \mathcal{N} \times_{\rho_q} \mathcal{B}_q = (\mathcal{N} \times \mathcal{B}_q) \,/\, \sim_q, \tag{72}$$
$$\mathcal{E}_p := \mathcal{N} \times_{\rho_p} \mathcal{B}_p = (\mathcal{N} \times \mathcal{B}_p) \,/\, \sim_p, \tag{73}$$

where $(n \cdot g, b) \sim_u (n, \rho_u(g)b)$ for $u \in \{q, p\}$.

These give fiber bundles $\pi_{\mathcal{E}_q} : \mathcal{E}_q \to \mathcal{C}$ and $\pi_{\mathcal{E}_p} : \mathcal{E}_p \to \mathcal{C}$ with fibers $\mathcal{B}_q(c) \cong \mathcal{B}_q$ and $\mathcal{B}_p(c) \cong \mathcal{B}_p$ at each $c \in \mathcal{C}$.

A.2 Agents and Multi-Agent Systems

Agent

An agent $\mathcal{A}^i$ is a pair of local sections over a domain $\mathcal{U}_i \subset \mathcal{C}$:

$$\mathcal{A}^i = \left(\sigma_q^i, \sigma_p^i\right), \tag{74}$$

where $\sigma_q^i : \mathcal{U}_i \to \mathcal{E}_q$ and $\sigma_p^i : \mathcal{U}_i \to \mathcal{E}_p$.

We write $q_i(c) := \sigma_q^i(c) \in \mathcal{B}_q(c)$ and $p_i(c) := \sigma_p^i(c) \in \mathcal{B}_p(c)$ for the belief and model at base point c.

Multi-Agent System

A multi-agent system $\mathcal{M}$ over $\mathcal{C}$ is a collection of agents indexed by $\mathcal{I}$:

$$\mathcal{M} = \left\{\mathcal{A}^i = \left(\sigma_q^i, \sigma_p^i\right)\right\}_{i \in \mathcal{I}}. \tag{75}$$

Agents generally overlap on intersections $\mathcal{U}_i \cap \mathcal{U}_j$.

Meta-Agent and Epistemic Death

A meta-agent is a multi-agent system whose component agents share identical section values on their overlap:

$$q_i(c) = q_j(c), \quad p_i(c) = p_j(c) \quad \text{for } c \in \mathcal{U}_i \cap \mathcal{U}_j. \tag{76}$$

A set of agents is \emph{epistemically dead} if they identically share both beliefs and models. While constituent agents of a meta-agent may be epistemically dead, the meta-agent itself need not be; such agents can be integrated out, yielding coarse-grained higher-order entities and a route toward emergence.

A.2.1 Bundle Morphisms and Transport Operators

Via standard horizontal lifting from the principal bundle $\mathcal{N}$ to the associated bundles, we obtain a hierarchy of morphisms and induced connections:

Intra-bundle transport:

- $\Omega_{ij}^{(q)} : \Gamma(\mathcal{B}_q) \to \Gamma(\mathcal{B}_q)$ (belief-to-belief)
- $\Omega_{ij}^{(p)} : \Gamma(\mathcal{B}_p) \to \Gamma(\mathcal{B}_p)$ (model-to-model)

Cross-scale transport:

- $\Lambda_{s'}^{s} : \Gamma^{s}(\mathcal{B}_q) \to \Gamma^{s'}(\mathcal{B}_q)$ (beliefs across scales)
- $\tilde{\Lambda}_{s'}^{s} : \Gamma^{s}(\mathcal{B}_p) \to \Gamma^{s'}(\mathcal{B}_p)$ (models across scales)

Inter-bundle morphisms:

- $\Theta_{j}^{i} : \Gamma(\mathcal{B}_q) \to \Gamma(\mathcal{B}_p)$ (belief to model)
- $\tilde{\Theta}_{j}^{i} : \Gamma(\mathcal{B}_p) \to \Gamma(\mathcal{B}_q)$ (model to belief)

Global bundle morphisms:

- $\Phi : \mathcal{E}_p \to \mathcal{E}_q$ (model bundle to belief bundle)
- $\tilde{\Phi} : \mathcal{E}_q \to \mathcal{E}_p$ (belief bundle to model bundle)

Here $\Gamma(\mathcal{B}_u)$ denotes the space of smooth sections over the base $\mathcal{C}$.

A.2.2 Gauge Frames and Connections

Each agent i possesses a local gauge frame field:

$$\phi_i : \mathcal{U}_i \to \mathfrak{g} = \mathrm{Lie}(G), \tag{77}$$

which induces a local connection one-form:

$$A_{\mu}^{(i)}(c) = U_i^{-1}(c)\partial_{\mu}U_i(c), \tag{78}$$

where $U_i(c) = \exp[\phi_i(c)] \in G$. The associated field strength (gauge curvature) is:

$$F_{\mu\nu}^{(i)}(c) = \partial_{\mu}A_{\nu}^{(i)} - \partial_{\nu}A_{\mu}^{(i)} + [A_{\mu}^{(i)}, A_{\nu}^{(i)}] \in \mathfrak{g}. \tag{79}$$

When two agents overlap at $c \in \mathcal{U}_i \cap \mathcal{U}_j$, the inter-agent gauge transformation:

$$\Omega_{ij}(c) = \exp[\phi_i(c)] \exp[-\phi_j(c)] \in G \tag{80}$$

transports agent j's representations into agent i's frame:

$$q_j(c) \mapsto \Omega_{ij}(c) \cdot q_j(c) := \rho(\Omega_{ij}(c))q_j(c). \tag{81}$$

On overlaps, local connections are related by:

$$A_{\mu}^{(i)} = \Omega_{ij}A_{\mu}^{(j)}\Omega_{ij}^{-1} + \Omega_{ij}\partial_{\mu}\Omega_{ij}^{-1}. \tag{82}$$

A.2.3 Curvature Structure

The full framework incorporates four distinct types of curvature:

Statistical Manifold Curvature

The fiber $\mathcal{B}$ has intrinsic curvature tensor:

$$R^{\mathcal{B}}(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z. \tag{83}$$

This measures the geometry of the space of probability distributions (e.g., Gaussian manifolds have constant negative curvature).

Gauge Group Curvature.

The structure group G itself is a curved manifold. For $G = SO(3)$:

- Topology: $SO(3) \cong \mathbb{RP}^3$
- Constant positive sectional curvature
- Non-commutativity: $\Omega_{ik} = \Omega_{ij}\Omega_{jk}$ does not commute with $\Omega_{jk}\Omega_{ik}$

Gauge Field Curvature.

The connection A_μ has field strength $F_{\mu\nu}$ measuring path-dependence of parallel transport through the base space $\mathcal{C}$.

Base Manifold Curvature.

If $\mathcal{C}$ carries a Riemannian metric $g_{\mathcal{C}}$, its curvature tensor:

$$R^{\mathcal{C}}(X,Y)Z = \nabla^{\mathcal{C}}_X \nabla^{\mathcal{C}}_Y Z - \nabla^{\mathcal{C}}_Y \nabla^{\mathcal{C}}_X Z - \nabla^{\mathcal{C}}_{[X,Y]} Z \tag{84}$$

affects the geometry of the latent space agents inhabit.

A.3 Simplifications for the Current Study

Here we specify the simplifications adopted for the remainder of this work, which retain essential geometric structure while ensuring computational tractability and simplification.

A.3.1 Matched Bundles

In our general formulation agents are treated as pairs of sections of associated bundles $\mathcal{B}_q$ and $\mathcal{B}_p$ of beliefs and models. Here we simplify by implementing

$$\mathcal{B}_q = \mathcal{B}_p =: \mathcal{B}, \qquad \rho_q = \rho_p =: \rho. \tag{85}$$

This gives $\mathcal{E}_q = \mathcal{E}_p =: \mathcal{E}$, and the bundle morphisms become:

$$\Phi = \tilde{\Phi} = \mathrm{id}_{\mathcal{E}}. \tag{86}$$

Each agent then reduces to a single section $\sigma^i : \mathcal{U}_i \to \mathcal{E}$, with belief/prior distinction maintained through local fiber coordinates.

A.3.2 Gaussian Fiber and SO(3) Gauge Group

Next, we restrict to the exponential family of multi-variate Gaussian distributions with $G = SO(3)$. That is,

- Fiber: $\mathcal{B} = \{(\mu, \Sigma) : \mu \in \mathbb{R}^3, \Sigma \in \mathbb{R}^{3\times 3}, \Sigma \succ 0\}$ (Gaussian manifold)
- Group: $G = SO(3)$ with representation:

$$\rho(\Omega) \cdot (\mu, \Sigma) = (\Omega\mu, \Omega\Sigma\Omega^\top) \tag{87}$$

The Gaussian manifold has constant negative curvature under the Fisher-Rao metric. The group $SO(3)$ has constant positive curvature and non-commutative composition. These choices are motivated by computational flexibility while maintain a degree of geometric generality. In particular, $SO(3)$ and its representation theory is well studied and present in a variety of fields. However, our results can be directly extended to $SO(N)$ or even $SU(N)$ in the standard way (Nakahara, 2003)(Hall, 2015).

A.3.3 Gauge Group and Representations

Let us now consider the representation theory of $G = SO(3)$, which acts on the fibers via its representations:

$$\rho_q : SO(3) \to \mathrm{GL}(d_q), \qquad \rho_p : SO(3) \to \mathrm{GL}(d_p). \tag{88}$$

In full generality these representations may be reducible or irreducible thereby allowing even or odd dimensions to be studied. For example, a token dimension of 768 could be broken into the appropriate number of irreps (see below) necessary for an even dimension action (Hall, 2015)

For $\Omega \in SO(3)$, the gauge action on a Gaussian state is:

$$\boxed{\begin{aligned} \rho_q(\Omega) \cdot (\mu_q, \Sigma_q) &= \big(\rho_q(\Omega)\,\mu_q,\ \rho_q(\Omega)\,\Sigma_q\,\rho_q(\Omega)^\top\big), \\ \rho_p(\Omega) \cdot (\mu_p, \Sigma_p) &= \big(\rho_p(\Omega)\,\mu_p,\ \rho_p(\Omega)\,\Sigma_p\,\rho_p(\Omega)^\top\big). \end{aligned}} \tag{89}$$

A.3.4 Representation Structure

The dimensions d_q, d_p are not constrained to be 1, 3, 5, 7, ... (irrep dimensions of $SO(3)$). Instead we may have ρ_q, ρ_p as reducible representations built from direct sums of the irreps:

$$\rho_q \cong \bigoplus_k n_k \cdot \ell_k, \quad d_q = \sum_k n_k(2\ell_k + 1), \tag{90}$$

where $\ell_k \in \{0, 1, 2, \ldots\}$ labels spin, $n_k \in \mathbb{N}$ is multiplicity, and irrep ℓ_k has dimension $2\ell_k + 1$.

Example: A $d_q = 768$ dimensional embedding with gauge group $G = SO(3)$ could decompose as:

$$\rho_q \cong 109 \cdot \ell_0 \oplus 49 \cdot \ell_1 \oplus 32 \cdot \ell_2 \oplus \ldots \tag{91}$$

(109 scalars + 49 vectors + 32 rank-2 tensors + ...), where the $SO(3)$ gauge group acts on this high-dimensional space via the representation $\rho_q : SO(3) \to GL(768)$.

In the simulations we perform (see discussion) we used $d_q = d_p = 9$ with the irreducible spin-4 representation:

$$\rho_q = \rho_p = \ell_4, \quad \text{(irrep, dimension } 2 \cdot 4 + 1 = 9). \tag{92}$$

A.3.5 Gauge structure

The gauge structure of our framework is built via the following:

- Gauge group: $G = SO(3)$ (compact, 3 generators)
- Frame fields: $\phi_i(c) \in \mathfrak{so}(3)$ vary spatially
- Transport: $\Omega_{ij}(c) = e^{\phi_i(c)} e^{-\phi_j(c)} \in SO(3)$
- Action on fiber: $\rho(\Omega_{ij}) \in GL(d_K)$

In the transformer limit we will show that the gauge group becomes trivial:

$$G = SO(3) \to \{e\} \quad \text{(identity element only)} \tag{93}$$

Therefore all gauge frames collapse to a single global shared space. I.e.

$$\phi_i(c) = \zeta = \text{const} \in \mathfrak{so}(3) \quad \forall i, c \tag{94}$$

This then makes gauge transport trivial:

$$\Omega_{ij}(c) = \exp[\zeta] \exp[-\zeta] = \mathbb{I} \quad \forall i, j, c \tag{95}$$

and the representation acts as the identity:

$$\rho(e) \cdot \mu_j = \mu_j \tag{96}$$

Therefore in this limit we have the following simplifications which lead to the transformer architecture from a more general and rich geometry:

- Vanishing Induced Connection: $A_\mu = -\partial_\mu \phi_i = 0$
- Vanishing Curvature: $F_{\mu\nu} = 0$ (flat bundle)
- A single global coordinate system
- The gauge-aligned KL reduces to standard KL

A.3.6 Flat Base Manifold

$$\mathcal{C} = \mathbb{R}^2 \quad \text{(Euclidean)}, \tag{97}$$

We consider the case of a flat base manifold (for simplicity and due to limited computational resources). Agents occupy finite support regions $\mathcal{U}_i$ as open subsets of $\mathbb{R}^2$ (where we invoke periodic boundary conditions in simulations).

A.3.7 Local Gauge Frames

Each agent i has a gauge frame field $\phi_i : \mathcal{U}_i \to \mathfrak{so}(3)$ which may vary spatially. This induces a local connection:

$$A_\mu^{(i)}(c) = -\partial_\mu \phi_i(c) + \mathcal{O}(\phi_i^2), \tag{98}$$

with field strength:

$$F_{\mu\nu}^{(i)}(c) = \partial_\mu A_\nu^{(i)} - \partial_\nu A_\mu^{(i)} + [A_\mu^{(i)}, A_\nu^{(i)}]. \tag{99}$$

In our simulations:

1. Frames are smooth and slowly varying: $\|\partial_\mu \phi_i\| \ll 1$, so that $F_{\mu\nu}^{(i)} \approx 0$ locally
2. We compute inter-agent transport pointwise (within the same fiber) using the Baker-Campbell-Hausdorff (BCH) formula:

$$\Omega_{ij}(c) = e^{\phi_i(c)} e^{-\phi_j(c)} \tag{100}$$

This effectively treats gauge transport as local frame rotations without considering holonomy or global topology, which shall be saved for future study.

A.3.8 Intra-Scale Transport Only

We restrict to intra-scale transport operators $\Omega_{ij} : \Gamma(\mathcal{B}) \to \Gamma(\mathcal{B})$ between agents at the same hierarchical level. Cross-scale morphisms $\Lambda_{s'}^{s}$ and meta-agent emergence are deferred to future work with promising preliminary results.

A.3.9 Gauge-Aligned Divergence

With these assumptions, the gauge-aligned KL divergence between agents is (see appendix):

$$D_{\mathrm{KL}}\left[q_i(c) \,\middle\|\, \Omega_{ij}(c) q_j(c)\right], \tag{101}$$

where for Gaussian beliefs:

$$\Omega_{ij}(c) \cdot (\mu_j, \Sigma_j) = \left(\Omega_{ij}(c)\mu_j,\, \Omega_{ij}(c)\Sigma_j \Omega_{ij}(c)^\top\right). \tag{102}$$

This divergence forms the basis of our attention mechanism.
The full geometric framework enables exploration of:

- Path-dependent parallel transport and holonomy effects
- Non-flat gauge connections with epistemic monopoles
- Curved base manifolds (hyperbolic/spherical latent spaces)
- Heterogeneous fiber structures ($\mathcal{B}_q \neq \mathcal{B}_p$)

- Cross-scale dynamics and meta-agent emergence
- Curvature-induced phase transitions in multi-agent systems
- Pullback geometries via agent sections from informational fibers to the base manifold ("It from Bit", "qualia, etc")

A.3.10 Rationale for Simplification

These simplifications allow us to establish foundational results and demonstrate that transformer attention emerges from this gauge-theoretic free energy minimization geometry. Such simplifications further allow us to work within a single well-understood statistical manifold (Gaussian) and gauge group with known analytic properties and non-trivial curvatures(Amari, 2016).

For the present study, the matched bundle case ($\Phi = \mathrm{id}$, $\mathcal{B}_q = \mathcal{B}_p$) suffices to establish the core theoretical result: transformer attention is the flat, isotropic, Dirac delta-function limit of gauge-covariant variational free energy minimization in multi-agent Bayesian systems. Furthermore data training and/or observations break the vacuum theory symmetry.

A.3.11 Multi-Timescale Dynamics

Our free energy naturally exhibits time-scale separation as a feature(Boettcher and Brunson, 2012):

In our present work, we only the fast subsystem of beliefs (where we omit the base space integrals for convenience),

$$\mathcal{F}_{\text{fast}}[\{q_i\}] = \sum_i D_{\text{KL}}(q_i \| p_i) + \sum_{i,j} \beta_{ij} D_{\text{KL}}(q_i \| \Omega_{ij} q_j) - \mathbb{E}_q[\log p(o \mid \{k_i\}, \{m_i\})] \tag{103}$$

This is minimized by gradient descent on $\{q_i\}$ while holding $\{s_i\}$ fixed, yielding belief updates:

$$\frac{\partial q_i}{\partial t} = -\eta_q \frac{\delta \mathcal{F}_{\text{fast}}}{\delta q_i} \tag{104}$$

with learning rate $\eta_q \sim \mathcal{O}(1)$ (fast).

Slow subsystem (model learning):

$$\mathcal{F}_{\text{slow}}[\{s_i\}] = \sum_i D_{\text{KL}}(s_i \| r_i) + \sum_{i,j} \gamma_{ij} D_{\text{KL}}(s_i \| \tilde{\Omega}_{ij} s_j) \tag{105}$$

This is minimized by gradient descent on $\{s_i\}$ while holding $\{q_i\}$ fixed (or averaging over recent beliefs), yielding model updates:

$$\frac{\partial s_i}{\partial t} = -\eta_s \frac{\delta \mathcal{F}_{\text{slow}}}{\delta s_i} \tag{106}$$

with learning rate $\eta_s \ll \eta_q$ (slow).

This time-scale separation enables learning: agents rapidly adapt beliefs q_i to new observations while slowly refining models s_i to capture long-term structure in a coordinated manner.

A.3.12 Meta-Agent Emergence and Cross-Scale Coupling

The hierarchical structure naturally supports the emergence of meta-agents (Shen et al., 2008). These are coarse-grained entities that are formed when groups of agents reach belief consensus.

A meta-agent $\mathcal{M}^{(1)}$ is a set of agents $\{i \in I_{\mathcal{M}}\}$ satisfying:

$$q_i = \Omega_{ij} q_j \quad \forall i, j \in I_{\mathcal{M}} \quad \text{(belief consensus)}, \tag{107}$$

$$s_i = \tilde{\Omega}_{ij} s_j \quad \forall i, j \in I_{\mathcal{M}} \quad \text{(model consensus)}. \tag{108}$$

When these conditions hold, the agents are epistemically dead—they share identical beliefs and models after accounting for gauge transformations and no longer evolve without continual observations. The meta-agent can be described by a single representative renormalized state $(q_{\mathcal{M}}, s_{\mathcal{M}})$.

Meta-agents at scale $\zeta = 0$ (individual agents) can generally form meta-agents at scale $\zeta = 1$ (groups), which can further coalesce into meta-agents at scale $\zeta = 2$ (communities), and so on. This creates a hierarchical scale structure:

$$\text{Agents}^{(0)} \xrightarrow{\text{consensus}} \text{Meta-agents}^{(1)} \xrightarrow{\text{consensus}} \text{Meta}^{(2)} \xrightarrow{\text{consensus}} \cdots$$

Figure 8: Hierarchical emergence through consensus at successive scales ζ.

At each scale, the effective free energy takes the same form as Eq. (16), with:

- Agents at scale ζ replaced by meta-agents at scale $\zeta + 1$
- Coupling constants renormalized:

$$\beta_{ij}^{(\zeta+1)} = f_\beta(\{\beta_{kl}^{(\zeta)}\}), \quad \gamma_{ij}^{(\zeta+1)} = f_\gamma(\{\gamma_{kl}^{(\zeta)}\})$$

- Effective gauge frames: $\phi_{\mathcal{M}}^{(\zeta+1)} = \text{average}(\{\phi_i^{(\zeta)}\})$

This is the gauge-theoretic analogue of renormalization group flow in statistical field theory(Anderson, 1984)(Wilson and Kogut, 1974)(García-Millán et al., 2024) . Cross-scale couplings $\Lambda_{\zeta'}^{\zeta}$ (Appendix) mediate interactions between agents at different hierarchical levels.

In the present work, we restrict to single-scale dynamics—all agents are at scale $\zeta = 0$ with no cross-scale couplings ($\Lambda_{\zeta'}^{\zeta} = 0$ for $\zeta \neq \zeta'$). We focus exclusively on the fast subsystem (Eq. 103), studying how beliefs $\{q_i\}$ evolve under alignment while holding models $\{s_i\}$ fixed.

Our preliminary simulations (reported separately) suggest that meta-agent fluctuation timescales are of the order $\tau_{\zeta+1}/\tau_\zeta \sim 10^4$-$10^6$, consistent with hierarchical structure in biological and social systems.

We posit that standard transformers operate entirely in the fast subsystem. They perform inference (q_i updates) with frozen models (s_i fixed during a forward pass). Training corresponds to slow updates of s_i, but without the explicit hierarchical structure or metacognitive alignment term $\gamma_{ij} D_{\text{KL}}(s_i \| \tilde{\Omega}_{ij} s_j)$.

Interestingly, our framework potentially applies to many informational systems - from transformers, to collections of humans coalescing into societies, to cognition, to language, and potentially even physics, chemistry, and biology - where statistical patterns and informational organizations emerge from lower level informational processing with decreasing timescales. Although, in our current study we do not invoke a time variable (aside from gradient descent step) we may tentatively associate a natural time scale as related to the amount of information updated within a step apropos "Information is a distinction that makes a difference" - Donald MacKay(Csiszár, 1967)(Cover and Thomas, 2006)(Kullback and Leibler, 1951)(MacKay, 1969). This suggests a minimum timescale corresponding to a single bit update.

That is to say, under a variational update $\delta S = 0$, the field generally evolves as $q \to q + \Delta q$. The local change in informational content over a single step may be characterized by the self-divergence

$$\Delta\mathcal{I} = D_{\mathrm{KL}}[\, q(c) + \Delta q(c) \,\|\, q(c)\,].$$

In gauge-theoretic terms, we interpret this variation as a local transformation at a base manifold point c of the form $q(c) \to dg^{-1}(c) \cdot q(c)$, where $dg^{-1}(c)$ is a gauge transformation associated with the Lie group G, acting on the fiber $\mathcal{B}_q$. Thus, we write:

$$\Delta\mathcal{I} = D_{\mathrm{KL}}[\, dg^{-1}(c) \cdot q(c) \,\|\, q(c)\,].$$

This quantity measures the epistemic deviation induced by a local frame change and highlights the informational change of shifting one's gauge frame.

Appendix B. Covariance Dynamics and Equilibrium Analysis

B.1 Covariance Gradient of the Generalized Free Energy

Here we derive the gradient of the single-agent free energy $\mathcal{F}_i$ with respect to the covariance Σ_i for agent i; a well known result in information geometry.

The free energy decomposes as

$$\mathcal{F}_i = D_{\mathrm{KL}}(q_i \,\|\, p_i) + \sum_{j \neq i} \beta_{ij}\, D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j) - \mathbb{E}_{q_i}[\log p(o_i \mid k_i)], \tag{109}$$

where $q_i = \mathcal{N}(\mu_i, \Sigma_i)$, $p_i = \mathcal{N}(\mu_{p,i}, \Sigma_{p,i})$, and $\sum_j \beta_{ij} = 1$ by construction.

B.1.1 Gaussian KL divergence and its derivative

For two Multivariate Gaussians, we have

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_1, \Sigma_1) \,\|\, \mathcal{N}(\mu_2, \Sigma_2)) = \frac{1}{2}\Big[\log \frac{|\Sigma_2|}{|\Sigma_1|} + \mathrm{tr}\big(\Sigma_2^{-1}\Sigma_1\big) \tag{110}$$

$$+ (\mu_2 - \mu_1)^\top \Sigma_2^{-1} (\mu_2 - \mu_1) - d \Big]. \tag{111}$$

and differentiating w.r.t. Σ_1 (holding μ_1, μ_2, Σ_2 fixed) gives

$$\frac{\partial D_{\mathrm{KL}}}{\partial \Sigma_1} = \frac{1}{2}\left[-\Sigma_1^{-1} + \Sigma_2^{-1}\right]. \tag{112}$$

Applying this to each term in $\mathcal{F}_i$:

$$\frac{\partial}{\partial \Sigma_i} D_{\mathrm{KL}}(q_i \,\|\, p_i) = \frac{1}{2}\left[-\Sigma_i^{-1} + \Sigma_{p,i}^{-1}\right], \tag{113}$$

$$\frac{\partial}{\partial \Sigma_i} D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j) = \frac{1}{2}\left[-\Sigma_i^{-1} + (\Omega_{ij}\Sigma_j\Omega_{ij}^\top)^{-1}\right]. \tag{114}$$

The observation term contributes an $O(\Sigma_i^{-1})$ correction that we neglect in the high-precision / strong-alignment regime. Thus

$$\boxed{\frac{\partial \mathcal{F}_i}{\partial \Sigma_i} = \frac{1}{2}\left[-2\Sigma_i^{-1} + \Sigma_{p,i}^{-1} + \sum_j \beta_{ij}(\Omega_{ij}\Sigma_j\Omega_{ij}^\top)^{-1}\right].} \tag{115}$$

Because $\sum_j \beta_{ij} = 1$, there is no $(1 + \sum_j \beta_{ij})$ prefactor in front of Σ_i^{-1}.

B.2 Fixed-Point Equation and Symmetric Solution

At equilibrium, we set $\partial \mathcal{F}_i / \partial \Sigma_i = 0$, giving

$$\Sigma_i^{-1} = \frac{1}{2}\left[\Sigma_{p,i}^{-1} + \sum_j \beta_{ij}(\Omega_{ij}\Sigma_j\Omega_{ij}^\top)^{-1}\right]. \tag{116}$$

This is a matrix-valued fixed-point equation coupling all agents: each agent's precision is the β-weighted combination of its own prior precision and the transported neighbor precisions.

B.2.1 Homogeneous limit

In the homogenous limit we assume
(i) all agents are identical, so $\Sigma_i = \Sigma_\infty$ for all i,
(ii) $\Omega_{ij} \approx I$ (weak misalignment),
(iii) $\Sigma_{p,i} = \Sigma_0$ (shared prior).
Then (116) becomes

$$\Sigma_\infty^{-1} = \frac{1}{2}\left[\Sigma_0^{-1} + \Sigma_\infty^{-1}\right] \quad \Longrightarrow \quad \Sigma_\infty = \Sigma_0. \tag{117}$$

Hence, in a perfectly symmetric population, the equilibrium covariance reproduces the shared prior.

B.2.2 Alignment-dominated regime

Although $\sum_j \beta_{ij} = 1$, the effective strength of alignment is controlled by the parameter τ in

$$\beta_{ij} = \frac{\exp\left[-\frac{1}{\tau} D_{\mathrm{KL}}\big(q_i \,\|\, \Omega_{ij} q_j\big)\right]}{\sum_k \exp\left[-\frac{1}{\tau} D_{\mathrm{KL}}\big(q_i \,\|\, \Omega_{ik} q_k\big)\right]}. \tag{118}$$

As $\tau \to 0$, β_{ij} becomes sharply peaked on whichever neighbor j best agrees (after transport). In that low-τ limit, the prior precision $\Sigma_{p,i}^{-1}$ becomes negligible relative to the socially enforced alignment term, and

$$\boxed{\Sigma_i^{-1} \;\approx\; \sum_j \beta_{ij} (\Omega_{ij} \Sigma_j \Omega_{ij}^\top)^{-1} \;=\; \big\langle (\Omega_{ij} \Sigma_j \Omega_{ij}^\top)^{-1} \big\rangle_\beta.} \tag{119}$$

Here $\langle \cdot \rangle_\beta$ denotes a β-weighted expectation over neighbors.

Therefore, in the strong-alignment (i.e. small τ) regime, agent i's precision matrix becomes the β-weighted average of its neighbors' transported precisions.

When, additionally, all agents already have approximately equal covariances and the transports are near-identity, i.e. $\Sigma_i \approx \Sigma_j$ and $\Omega_{ij} \approx I$, then (119) implies

$$\Sigma_i^{-1} \approx \Sigma_j^{-1} \quad \Longrightarrow \quad \Sigma_i \approx \Omega_{ij} \Sigma_j \Omega_{ij}^\top, \tag{120}$$

justifying the alignment assumption used in the main text.

B.2.3 Gradient flow dynamics

Finally, consider the gradient flow

$$\frac{d\Sigma_i}{dt} = -\eta_\Sigma \frac{\partial \mathcal{F}_i}{\partial \Sigma_i}, \qquad \eta_\Sigma > 0. \tag{121}$$

Local stability of the equilibrium (116) follows from the positive-definiteness of the Hessian. For the Gaussian KL terms,

$$\frac{\partial^2 D_{\mathrm{KL}}}{\partial \Sigma_1 \, \partial \Sigma_1} \;\sim\; \Sigma_1^{-1} \otimes \Sigma_1^{-1} + \Sigma_2^{-1} \otimes \Sigma_2^{-1}, \tag{122}$$

which is manifestly positive definite for Σ_1, Σ_2.

Hence the covariance alignment fixed-point is an attractor of the variational dynamics.

Therefore, we find that $\Sigma_i \approx \Omega_{ij} \Sigma_j \Omega_{ij}^\top$ emerges from the dynamics itself rather than as an imposed constraint.

Appendix C. Relating The Quadratic Forms to Transported KL Divergences

We now show that the pairwise quadratic expectations in (**??**) can be expressed in terms of KL divergences between transported distributions (see appendix for general requirement of the forward KL).

C.1 Exact expansion for Gaussian beliefs

For independent Gaussians $q_i = \mathcal{N}(\mu_{q,i}, \Sigma_{q,i})$ and $q_j = \mathcal{N}(\mu_{q,j}, \Sigma_{q,j})$, the expectation of a quadratic form is

$$\mathbb{E}_{q_i q_j}[\delta^\top A \delta] = \mathrm{tr}\big(A\,\mathrm{Cov}(\delta)\big) + \bar{\delta}^\top A \bar{\delta}, \tag{123}$$

where $\delta = k_i - \Omega_{ij} k_j$, $\bar{\delta} = \mathbb{E}[\delta] = \mu_{q,i} - \Omega_{ij}\mu_{q,j}$, and

$$\mathrm{Cov}(\delta) = \Sigma_{q,i} + \Omega_{ij}\Sigma_{q,j}\Omega_{ij}^\top. \tag{124}$$

Applying (123) with $A = \Lambda_{ij}$:

$$\begin{aligned}\mathbb{E}_{q_i q_j}[(k_i - \Omega_{ij}k_j)^\top \Lambda_{ij}(k_i - \Omega_{ij}k_j)] = \mathrm{tr}\big[\Lambda_{ij}(\Sigma_{q,i} + \Omega_{ij}\Sigma_{q,j}\Omega_{ij}^\top)\big] \\ + (\mu_{q,i} - \Omega_{ij}\mu_{q,j})^\top \Lambda_{ij}(\mu_{q,i} - \Omega_{ij}\mu_{q,j}).\end{aligned} \tag{125}$$

An identical expansion holds for the model channel with Γ_{ij} and $\tilde{\Omega}_{ij}$.

We observe that we can choose the coupling precisions Λ_{ij} and Γ_{ij} to make (125) proportional to a KL divergence.

Specifically, we define

$$\Lambda_{ij} := \tau_{ij}^{(q)}\,(\Omega_{ij}\Sigma_{q,j}\Omega_{ij}^\top)^{-1}, \qquad \Gamma_{ij} := \tau_{ij}^{(p)}\,(\tilde{\Omega}_{ij}\Sigma_{p,j}\tilde{\Omega}_{ij}^\top)^{-1}, \tag{126}$$

where $\tau_{ij}^{(q)}, \tau_{ij}^{(p)} > 0$ are dimensionless coupling strengths.

Recall that the KL divergence between two Gaussians $q_i = \mathcal{N}(\mu_i, \Sigma_i)$ and $\Omega_{ij} q_j = \mathcal{N}(\Omega_{ij}\mu_j, \Omega_{ij}\Sigma_j\Omega_{ij}^\top)$ is

$$D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j) = \frac{1}{2}\Big[\mathrm{tr}\big((\Omega_{ij}\Sigma_j\Omega_{ij}^\top)^{-1}\Sigma_i\big) - d_q$$

$$+ \log\frac{\det(\Omega_{ij}\Sigma_j\Omega_{ij}^\top)}{\det\Sigma_i} + (\mu_i - \Omega_{ij}\mu_j)^\top(\Omega_{ij}\Sigma_j\Omega_{ij}^\top)^{-1}(\mu_i - \Omega_{ij}\mu_j)\Big].$$

When beliefs are approximately aligned (the regime enforced by the coupling itself), the covariances satisfy $\Sigma_i \approx \Omega_{ij}\Sigma_j\Omega_{ij}^\top$, and the trace and log-determinant terms approximately cancel. In this alignment regime, as shown in the appendix, the quadratic expectation becomes

$$\frac{1}{4}\mathbb{E}_{q_i q_j}[(k_i - \Omega_{ij}k_j)^\top \Lambda_{ij}(k_i - \Omega_{ij}k_j)] \approx \frac{\tau_{ij}^{(q)}}{2}\,D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j) + \mathrm{const}, \tag{127}$$

where the constant absorbs dimension-dependent terms and $O(\|\Delta\|^2)$ covariance mismatch corrections.

Similarly for the model channel:

$$\frac{1}{4}\mathbb{E}_{s_i s_j}[(m_i - \tilde{\Omega}_{ij}m_j)^\top \Gamma_{ij}(m_i - \tilde{\Omega}_{ij}m_j)] \approx \frac{\tau_{ij}^{(p)}}{2}\,D_{\mathrm{KL}}(s_i \,\|\, \tilde{\Omega}_{ij} s_j) + \mathrm{const}. \tag{128}$$

C.2 Defining normalized alignment weights

To obtain the standard form, we define normalized alignment weights

$$\beta_{ij} := \frac{\tau_{ij}^{(q)}}{2}, \qquad \gamma_{ij} := \frac{\tau_{ij}^{(p)}}{2}. \tag{129}$$

These weights have a clear interpretation: β_{ij} measures the strength of belief alignment between agents i and j, while γ_{ij} measures model alignment strength. In the limit $\beta_{ij} \to \infty$ with fixed γ_{ij}, agents' beliefs are forced to perfect agreement after transport, while their models may still differ. Conversely, $\gamma_{ij} = 0$ decouples model alignment entirely. In all subsequence equations we taked τ to be independent of each agent - a constant global value which we set to 1.

Appendix D. Conditional Uniqueness of the Forward KL Divergence via Variational Duality

We now show that, within a broad but well-defined class of variational games, the forward KL divergence $D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j)$ is the only divergence that yields a closed-form Gibbs-type solution for the belief update and a consistent dual interpretation for the attention weights. Agents locally minimize their agent-specific variational free energy and the system of agents minimize their collective global free energies. This local-global coordination gives rise to the expected forward KL attention term.

Here we are restricting the the matched-fiber case. The full generalization follows by applying the appropriate bundle morphisms/intertwiners.

The uniqueness of this term is conditional and follows from three assumptions:

1. $\mathcal{D}$ is local in c, of f-divergence form $\displaystyle\int q_i(c)\, f\!\left(\frac{q_i(c)}{\Omega_{ij}(c) q_j(c)}\right) dc$,
2. the coupling is linear
3. the minimizing belief, q_i^*, remains in the exponential-family (log-linear) class.

D.0.1 The Coupled Variational Problem

Each agent i minimizes a local free-energy functional:

$$F_i[\beta_i] = \min_{q_i} \Big\{ D_{\mathrm{KL}}(q_i \,\|\, p_i) + \sum_{j \neq i} \beta_{ij}\, \mathcal{D}(q_i, q_j) \Big\}, \tag{130}$$

For notational convenience, we can decompose the KL divergence as:

$$D_{\mathrm{KL}}(q_i \,\|\, p_i) = \langle H_i \rangle_{q_i} + S(q_i) + \mathrm{const}, \tag{131}$$

where

$$H_i(c) := -\log p_i(c) \quad \text{(local "energy")}, \tag{132}$$

$$S(q_i) := -\int q_i(c) \log q_i(c)\, dc \quad \text{(Shannon entropy)}, \tag{133}$$

$$\langle H_i \rangle_{q_i} := \int q_i(c) H_i(c)\, dc \quad \text{(expected energy)}. \tag{134}$$

This gives $D_{\mathrm{KL}}(q_i \,\|\, p_i) = \int q_i(c)[\log q_i(c) - \log p_i(c)]\, dc$.
The attention weights $\beta_{ij} \geq 0$ (with $\sum_j \beta_{ij} = 1$) are subsequently chosen by optimizing

$$\mathcal{J}_i(\beta_i) = \sum_{j \neq i} \beta_{ij} C_{ij} + \tau \sum_{j \neq i} \beta_{ij} \log \beta_{ij}, \tag{135}$$

where C_{ij} denotes the marginal cost of attending to agent j as we've described above.
We shall later identify

$$C_{ij} := \frac{\partial S_i}{\partial \beta_{ij}}, \tag{136}$$

so that attention weights allocate resources in proportion to marginal coordination penalties.

D.0.2 Forward KL and the Geometric-Mean Solution

Let $\mathcal{D}(q_i, q_j) = D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j)$, where

$$\frac{\delta D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j)}{\delta q_i(c)} = \log \frac{q_i(c)}{\Omega_{ij}(c) q_j(c)} + 1.$$

The stationary condition for (130) is

$$H_i(c) + \log q_i(c) + \sum_j \beta_{ij} \left[\log \frac{q_i(c)}{\Omega_{ij}(c) q_j(c)} + 1 \right] = \lambda_i, \tag{137}$$

with λ_i enforcing normalization.

Rearranging and solving for $q_i(c)$ gives a Boltzmann distribution whose mean field is a geometric average of transported neighbor beliefs.

$$q_i^*(c) = \frac{1}{Z_i} e^{-H_i(c)/2} \prod_j [\Omega_{ij}(c) q_j(c)]^{\beta_{ij}/2}, \tag{138}$$

This structure is preserved only for the forward KL divergence; alternative divergences destroy the log-linearity of the exponent.

D.0.3 Dual Relation via the Envelope Theorem

At the stationary value q_i^*, the envelope theorem implies

$$\frac{\partial F_i}{\partial \beta_{ij}} = \mathcal{D}(q_i^*, q_j) = D_{\mathrm{KL}}(q_i^* \,\|\, \Omega_{ij} q_j), \tag{139}$$

such that the marginal cost of increasing attention to j equals the forward KL divergence between the updated belief q_i^* and the transported neighbor $\Omega_{ij} q_j$ into i's frame.

This identifies the attention cost with the KL.

D.0.4 Reverse and Symmetric KL Forms

If instead $\mathcal{D}(q_i, q_j) = D_{\mathrm{KL}}(\Omega_{ij} q_j \,\|\, q_i)$, the stationary condition becomes

$$H_i(c) + \log q_i(c) - \sum_j \beta_{ij} \frac{\Omega_{ij}(c) q_j(c)}{q_i(c)} = \text{const}, \tag{140}$$

introducing terms as $1/q_i$ and leading to a transcendental stationary equation without a closed-form solution destroying the exponential family requirement above.

Likewise, the symmetrized divergence

$$\mathcal{D}_{\mathrm{sym}}(q_i, q_j) = \tfrac{1}{2}[D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j) + D_{\mathrm{KL}}(\Omega_{ij} q_j \,\|\, q_i)]$$

mixes $\log q_i$ and $1/q_i$ terms, again breaking log-linearity.
Hence, among local f-divergences, only the forward KL preserves exponential-family closure.

D.0.5 Conditional Uniqueness Theorem

Let $\mathcal{D}(q_i, q_j)$ be any local f-divergence

$$\mathcal{D}(q_i, q_j) = \int q_i(c)\, f\left(\frac{q_i(c)}{\Omega_{ij}(c) q_j(c)}\right) dc,$$

that enters linearly in (130), and further suppose that the stationary distribution q_i^* is log-linear in $\{H_i, \Omega_{ij} q_j\}$.

Then the following are equivalent:

1. q_i^* has the geometric-mean Boltzmann form (138);
2. $\mathcal{D}(q_i, q_j) = D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j)$;
3. $C_{ij} = \dfrac{\partial F_i}{\partial \beta_{ij}} = D_{\mathrm{KL}}(q_i^* \,\|\, \Omega_{ij} q_j)$.

Proof sketch
(2) $\Rightarrow$ (1) follows from direct solution of the stationary condition.
(1) $\Rightarrow$ (2):
Assume the solution is log-linear (138).
Substituting this into the stationarity condition gives

$$\frac{\delta \mathcal{D}}{\delta q_i(c)}\Big|_{q_i^*} = \log \frac{q_i^*(c)}{\Omega_{ij}(c) q_j(c)} + 1.$$

Next, integrating, we find

$$\mathcal{D}(q_i^*, q_j) = \int q_i^*(c) \log \frac{q_i^*(c)}{\Omega_{ij}(c) q_j(c)}\, dc + \text{const.}$$

We require that $\mathcal{D}(q, q) = 0$ thereby fixing the constant to be zero thus producing the forward KL form.

(3) $\Rightarrow$ (2):

By the envelope theorem (139), the only divergence consistent with a linear β_{ij}-coupling and this derivative structure is the forward KL.

D.0.6 Interpretations

1. **Gauge invariance**
   The comparison is made between q_i and the transported $\Omega_{ij} q_j$ in the same frame (and similarly for $j \to i$). Gauge covariance fixes what is compared, while variational duality fixes how it is compared.

2. **Variational duality**

   The forward KL is the only divergence that simultaneously yields:

   - a closed-form Boltzmann solution for q_i, and
   - a consistent dual cost $C_{ij} = \partial F_i / \partial \beta_{ij}$.

3. **Information geometry**

   The forward KL is the Bregman divergence generated by the negative entropy potential, whose Hessian induces the Fisher-Rao metric and yields the m/e-projection (mixed/exponential family) Pythagorean theorem. These global properties are not shared by generic f-divergences.

D.0.7 Summary

Under the natural assumptions of locality, linear coupling, and exponential-family closure, the forward KL divergence

$$\boxed{C_{ij} = D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j)}$$

is not merely a modeling choice but a necessary consequence
of the variational and geometric structure underlying agent coordination. It is furthermore rotationally invariant under $SO(N)$ and $SU(N)$.

D.0.8 Connection to the Gauge-Covariant Free Energy

The conditional uniqueness result above justifies the specific form of the alignment terms appearing in the generalized variational free energy:

$$\mathcal{S} = \sum_i D_{\mathrm{KL}}(q_i \,\|\, p_i) + \sum_i D_{\mathrm{KL}}(s_i \,\|\, r_i) + \sum_{i,j} \beta_{ij}\, D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j)$$

$$+ \sum_{i,j} \gamma_{ij}\, D_{\mathrm{KL}}(s_i \,\|\, \tilde{\Omega}_{ij} s_j) - \mathbb{E}_q[\log p(o|\{k_i, m_i\})].$$

Each coupling term, such as

$$\beta_{ij}\, D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j),$$

is therefore not arbitrary: it arises uniquely from the requirement that

1. belief updates q_i admit an exponential-family form consistent with local free energy minimization,
2. attention weights β_{ij} act as variational dual variables conjugate to those divergences, and
3. comparisons between agents are made in gauge-aligned coordinates through the transport operators Ω_{ij}.

Therefore, the forward KL plays the role of the canonical gauge-covariant coupling between agents' beliefs/models, unifying variational and geometric principles.

Reverse or symmetric divergences would violate at least one of these constraints: they either destroy exponential-family closure, break dual consistency $\left(C_{ij} = \partial F_i / \partial \beta_{ij}\right)$, or fail to respect the gauge-covariant comparison structure. Thus, the gauge-covariant free energy (**??**) naturally inherits the unique forward-KL alignment form as a direct consequence of its underlying variational geometry.

D.1 Softmax Attention via Maximum Entropy Principle

Given alignment cost $C_{ij} = D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j)$ above, we seek attention weights β_{ij} that:

1. Minimize expected disagreement: $\sum_j \beta_{ij} C_{ij}$
2. Maximize uncertainty (entropy): $-\sum_j \beta_{ij} \log \beta_{ij}$
3. Satisfy normalization: $\sum_j \beta_{ij} = 1$

Following Jaynes' maximum entropy principle, we maximize entropy subject to the constraint that expected cost equals some target $\langle C \rangle$:

$$\max_{\{\beta_{ij}\}} \left\{ -\sum_j \beta_{ij} \log \beta_{ij} \right\} \tag{141}$$

subject to:

$$\sum_j \beta_{ij} = 1, \tag{142}$$

$$\sum_j \beta_{ij} C_{ij} = \langle C \rangle. \tag{143}$$

Lagrangian:

$$\mathcal{J} = -\sum_j \beta_{ij} \log \beta_{ij} + \lambda \left(\sum_j \beta_{ij} - 1 \right) + \mu \left(\sum_j \beta_{ij} C_{ij} - \langle C \rangle \right). \tag{144}$$

First-order condition:

$$\frac{\partial \mathcal{J}}{\partial \beta_{ij}} = -\log \beta_{ij} - 1 + \lambda + \mu C_{ij} = 0. \tag{145}$$

Solution:

$$\log \beta_{ij} = \lambda - 1 + \mu C_{ij} \quad \Rightarrow \quad \beta_{ij} = K \exp(\mu C_{ij}), \tag{146}$$

where $K = \exp(\lambda - 1)$.
Normalization $\sum_j \beta_{ij} = 1$ gives:

$$\beta_{ij} = \frac{\exp(\mu C_{ij})}{\sum_k \exp(\mu C_{ik})}. \tag{147}$$

Since we want to \emph{minimize} expected cost (not maximize), we choose $\mu = -1/\tau < 0$:

$$\boxed{\beta_{ij} = \frac{\exp(-C_{ij}/\tau)}{\sum_k \exp(-C_{ik}/\tau)} = \mathrm{softmax}_j \left(-\frac{C_{ij}}{\tau} \right)}. \tag{148}$$

Interpretation:

- $\tau \to 0$: Hard attention (argmin over j)
- $\tau \to \infty$: Uniform attention (maximum uncertainty)
- Intermediate τ: Soft attention balancing cost minimization and entropy maximization

Alternative Derivation (Unconstrained):
Equivalently, minimize the free energy functional:

$$F[\beta_i] = \sum_j \beta_{ij} C_{ij} + \tau \sum_j \beta_{ij} \log \beta_{ij}, \tag{149}$$

subject only to $\sum_j \beta_{ij} = 1$.
Lagrangian:

$$\mathcal{J} = \sum_j \beta_{ij} C_{ij} + \tau \sum_j \beta_{ij} \log \beta_{ij} + \lambda \left(\sum_j \beta_{ij} - 1 \right). \tag{150}$$

First-order condition:

$$C_{ij} + \tau(\log \beta_{ij} + 1) + \lambda = 0 \quad \Rightarrow \quad \tau \log \beta_{ij} = -C_{ij} - \tau - \lambda. \tag{151}$$

This immediately gives:

$$\beta_{ij} = \frac{\exp(-C_{ij}/\tau)}{\sum_k \exp(-C_{ik}/\tau)}. \tag{152}$$

This is the unique maximum-entropy distribution consistent with the constraint $\langle C \rangle = \sum_j \beta_{ij} C_{ij}$.

Hence, each agent assigns weights β_{ij} according to their relative consistency where τ controls the sharpness of selection. In the limit $\tau \to 0$ the β_{ij} weights collapse to hard-attention whereas for large τ we approach uniform weighting.

Therefore, given agents as local open sections over the base space our complete attention weights are given as

$$\beta_{ij}(c) = \frac{\exp\left[-\frac{1}{\tau} \mathrm{KL}\big(q_i(c) \,\big\|\, \Omega_{ij} q_j(c)\big)\right] \, \chi_{ij}(c)}{\sum_k \exp\left[-\frac{1}{\tau} \mathrm{KL}\big(q_i(c) \,\big\|\, \Omega_{ik} q_k(c)\big)\right] \, \chi_{ik}(c)}$$

where $\chi_{ij}(c)$ is the overlap of agent i and agent j support (or mask). Or said simply; their interaction volume.

Appendix E. Variational Gradient Descent: Implementation and Numerical Methods

The variational free energy minimization is implemented through a sophisticated gradient descent scheme that respects the intricate geometric structure of the multi-agent system. At each simulation step, the algorithm orchestrates a sequence of coordinated field updates across all dynamic variables—the statistical parameters (μ_i, Σ_i) for both belief and model fibers, the gauge frame fields $(\phi_i, \tilde{\phi}_i)$, and optionally the global connection field A_μ—while maintaining numerical stability on the constrained manifolds where these quantities naturally live.

E.1 Gradient Accumulation and Energy Terms

The core computational pipeline begins with gradient accumulation, where each agent's variational gradient is constructed by summing contributions from multiple energy terms. For the statistical parameters, we compute gradients arising from the self-consistency term $D_{\mathrm{KL}}(q_i \| p_i)$, the belief alignment coupling $\sum_j \beta_{ij} D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j)$, the model alignment coupling $\sum_j \gamma_{ij} D_{\mathrm{KL}}(s_i \| \tilde{\Omega}_{ij} s_j)$, and the observation likelihood $-\mathbb{E}_{q_i}[\log p(o_i|k_i)]$. Each of these terms contributes a distinct gradient component that must be carefully transported, aggregated, and symmetrized before application.

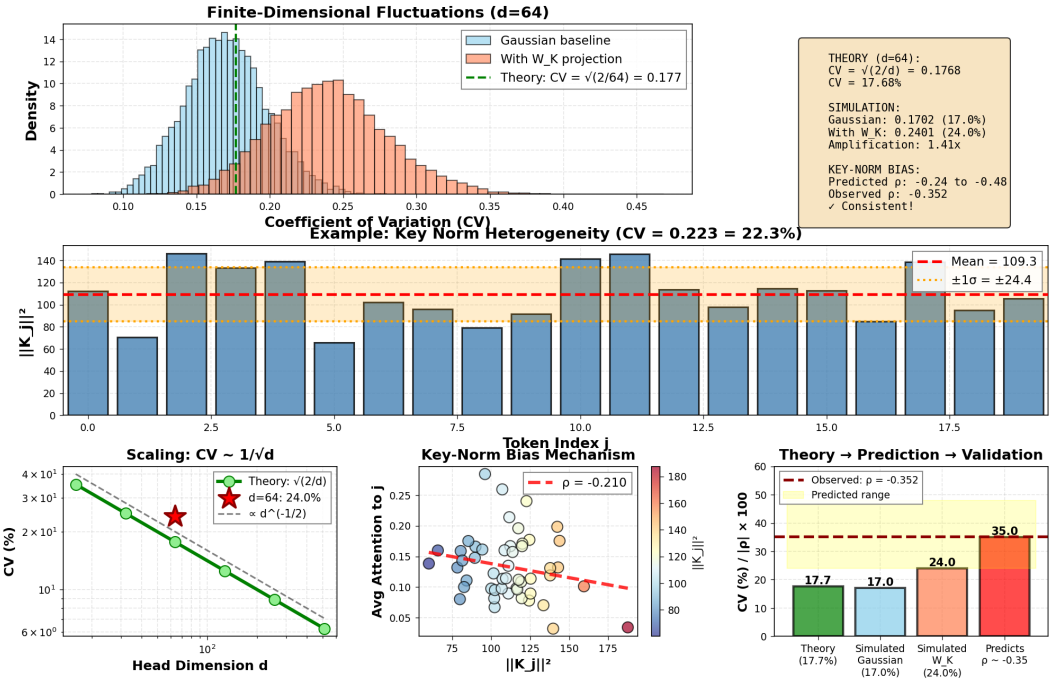

Figure 9: **Finite-dimensional fluctuations in transformer key norms.** (**Top-left**) Distribution of key-norm coefficient of variation (CV) under Gaussian baseline and after W_K projection, compared with theoretical prediction (CV $\approx 17.7\%$). (**Top-right**) Example of individual key norms across tokens illustrating sample CV. (**Bottom-left**) Finite-dimensional scaling of CV following the theoretical law CV $\sim 1/\sqrt{d}$, with BERT-base ($d = 64$) highlighted. (**Bottom-right**) Average attention bias as a function of key-norm magnitude, demonstrating minimal systematic bias even with CV 26%.

The alignment terms are particularly intricate, requiring neighbor iteration over all overlapping agent pairs, computation of gauge-transported statistics via the parallel transport operators Ω_{ij}, and evaluation of the KL divergence gradients in the appropriate local frames. These gradients are accumulated into per-agent "inboxes" during a first pass, then drained and combined with self-energy gradients in a second pass to yield the total gradient for each agent.

E.2 Natural Gradient Descent on the Gaussian Manifold

A critical aspect of our implementation is the use of natural gradient descent for the statistical parameters, which exploits the information-geometric structure of the Gaussian manifold. Rather than treating the covariance matrices Σ_i as Euclidean variables, we recognize them as points on the manifold of symmetric positive-definite matrices, equipped with the Fisher-Rao metric. The natural gradient at a point Σ on this manifold is obtained by applying the inverse Fisher metric to the Euclidean gradient, yielding an intrinsic tangent vector that respects the manifold's curved geometry.

In our vectorized implementation, this transformation is performed via `apply_natural_gradient_batch`, which computes the whitened gradient

$$R = \Sigma^{-1/2}(\eta \cdot \nabla)\Sigma^{-1/2} \tag{153}$$

where ∇ is the raw Euclidean gradient and η is the learning rate. This whitening procedure ensures that the gradient step is affine-invariant—the same step size produces comparable effects regardless of the current scale of the covariance matrix. Following the computation of the tangent vector R, we apply a trust-region constraint by clipping its Frobenius norm to a maximum relative step size ρ, typically set to values between 0.1 and 0.5. This prevents excessively large updates that could destabilize the manifold structure or lead to ill-conditioned covariances.

E.3 Manifold Retraction for Covariance Matrices

The actual manifold retraction—the process of mapping the tangent vector back to a valid point on the manifold—is performed using the affine-invariant exponential map by default. The exponential map is computed as

$$\Sigma_{k+1} = \Sigma_k^{1/2} \exp(R) \Sigma_k^{1/2}, \tag{154}$$

where the matrix exponential $\exp(R)$ is evaluated via eigendecomposition: $R = U\Lambda U^\top$ yields $\exp(R) = U \exp(\Lambda) U^\top$ with component-wise exponentials applied to the eigenvalues. This procedure is provably SPD-preserving provided Σ_k is SPD and R is symmetric, which our implementation guarantees through explicit symmetrization at multiple stages.

Additional safeguards include post-retraction sanitization via `sanitize_sigma`, which symmetrizes the result, raises on any floating-point anomalies (NaNs), and applies a spectral floor to the eigenvalues to ensure they remain above a minimum threshold $\epsilon_{\text{SPD}} \sim 10^{-8}$. This spectral regularization is essential for preventing numerical collapse of the covariance matrices during prolonged gradient descent, particularly in regions where the free-energy landscape becomes very flat.

E.4 Gauge Frame Dynamics on SO(3)

For the gauge frame fields ϕ_i and $\tilde{\phi}_i$, which live in the Lie algebra $\mathfrak{so}(3)$, a different geometric structure must be respected. These fields parameterize the local gauge frames as $U_i = \exp(\phi_i)$, and their dynamics are governed by gradients computed via the induced Fisher metric on the group manifold. The gradients are obtained by accumulating contributions from the self KL divergences, the alignment terms, and any optional (disabled) cross-fiber couplings, each of which produces a co-vector in the dual of the Lie algebra.

These co-vectors are computed using the differential of the matrix exponential map, which relates infinitesimal variations $\delta\phi$ to variations in the group element via

$$\delta(\exp \phi) = \exp(\phi) \cdot d\exp_\phi(\delta\phi), \tag{155}$$

where $d\exp_\phi$ is the derivative of the exponential map at ϕ. Our implementation caches the matrix-valued operator $d\exp^{-1}$ (implemented via the Baker-Campbell-Hausdorff formula truncated at fourth order) and uses it to project co-vectors back to tangent vectors

in the algebra. The Fisher metric inverse, computed via `inverse_fisher_metric_field`, provides a natural Riemannian structure that preconditions these gradients, accounting for the non-flat geometry of SO(3).

Once the gradient direction is determined, we apply a step with learning rate η_ϕ (typically 10^{-1} to 10^{-2}), forming a candidate update $\phi_{\text{cand}} = \phi + \delta\phi$. However, the Lie algebra $\mathfrak{so}(3)$ has a natural periodic boundary—angles wrap at $\pm\pi$—and numerical drift can cause ϕ to wander outside the principal domain. To prevent this, we apply a retraction via `retract_phi_principal`, which reflects vectors that exceed the boundary back into the fundamental domain, then slightly nudges them away from the boundary via a small margin (typically 10^{-2}) to avoid numerical instability near the critical points where the exponential map degenerates.

E.5 Learning Rate Configuration

Learning rates are carefully tuned to balance convergence speed against numerical stability, with separate tunable rates for each class of variable reflecting their characteristic time scales. Belief means μ_q and covariances Σ_q use rates $\eta_{\mu,q} \sim 10^{-1}$ and $\eta_{\Sigma,q} \sim 10^{-1}$, while model parameters use slightly smaller values to enforce the separation between fast belief dynamics and slow model learning. Gauge frame fields utilize $\eta_\phi \sim 10^{-1}$ to allow relatively rapid frame adjustment. All rates can be dynamically scaled via a global multiplier, and in regions of steep gradients an adaptive reduction mechanism (not currently enabled by default) can further decrease step sizes to prevent overshooting.

During each update, gradients are accumulated in double precision (float64) to minimize rounding errors, particularly for the gauge frame fields where small numerical discrepancies can accumulate over many steps due to the group structure. After the gradient step is computed, results are cast back to single precision (float32) for storage and subsequent energy evaluations, balancing numerical accuracy with memory efficiency.

E.6 Convergence Criteria and Diagnostics

The simulation monitors convergence through multiple criteria, primarily tracking the total free energy $\mathcal{F}$ and its rate of change $\Delta\mathcal{F}$. A run is considered converged when the free energy stabilizes, defined as $|\Delta\mathcal{F}| < 10^{-5}$ for at least 200 consecutive steps, indicating the system has reached a stationary point (local minimum) of the variational functional. Typical simulations require between 300 and 8000 steps to reach this criterion, depending on initialization and coupling strengths.

In addition to the global energy, we compute per-agent diagnostics including the norms of belief and model means $\|\mu_{q,i}\|$, $\|\mu_{p,i}\|$, trace-normalized covariances, alignment metrics measuring the KL divergence between neighboring agents, and the distribution of attention weights β_{ij}. These quantities are logged at each step and visualized in-situ to verify that the dynamics exhibit expected behavior—for instance, in vacuum (observation-free) runs, all agent mean norms should converge to a common value, reflecting the rotationally symmetric ground state. When observations are present, these norms diverge as agents specialize, providing a clear signature of spontaneous symmetry breaking.

E.7 Numerical Stability Safeguards

The update procedure includes extensive numerical safeguards to ensure stability over long simulation runs. The `sanitize_sigma` function is invoked after every covariance update to detect and repair ill-conditioned matrices, applying a spectral floor to eigenvalues, capping condition numbers if they exceed configurable thresholds (typically 10^8 to 10^{10}), and optionally renormalizing the trace to prevent runaway growth.

For gauge transport operators Ω_{ij}, the function `safe_omega_inv` verifies that these matrices remain approximately orthogonal by checking the deviation $\|\Omega^\top \Omega - I\|_F$ against a tolerance scaled by machine epsilon; if the deviation is excessive, the matrix is reorthogonalized via QR decomposition with determinant correction to ensure it remains in SO(3).

Energy budget tracking, implemented via the `EnergyBudget` class, monitors the decomposition of the total free energy into its constituent terms at each step, verifying that energy conservation is approximately satisfied (within numerical tolerance) when no external observations are added. Similarly, a `StabilityMonitor` tracks condition numbers of all covariance matrices, checks for the emergence of NaN or infinite values, detects gradient explosions (gradients exceeding preset thresholds), and validates that the SPD property is maintained throughout the evolution. If critical stability issues are detected—such as covariance matrices losing positive-definiteness or gradients diverging—the monitor logs detailed diagnostics and optionally halts the simulation to prevent catastrophic numerical failure.

E.8 Parallelization Strategy

Parallelization is employed to accelerate gradient computation across multiple agents, leveraging the `joblib` library with a `loky` backend for process-based parallelism. The computation is structured so that each worker receives a subset of agents along with read-only access to the shared runtime context and a memory-mapped disk cache containing precomputed expensive quantities such as gauge transport operators and their derivatives. Gradients are computed independently for each agent in parallel, then aggregated in the master process to form the global update. This design scales efficiently up to the number of physical cores available.

The parallel implementation carefully manages memory to avoid duplication of large arrays, using shared memory views where possible and writing intermediate results to disk-backed caches that are accessible across processes. Thread-level parallelism within each worker is deliberately suppressed (via environment variables like `OMP_NUM_THREADS=1`) to prevent oversubscription and maintain cache coherence, as nested parallelism typically degrades performance for these workloads.

E.9 Integration and Orchestration

The overall gradient descent loop integrates these components into a cohesive pipeline. At each step, we first zero all gradient accumulators, then invoke `compute_all_gradients` for each agent, which populates the gradient fields `grad_mu_q`, `grad_Sigma_q`, `grad_mu_p`, `grad_Sigma_p`, `grad_phi`, and `grad_phi_tilde`. These raw gradients are then post-processed—symmetrized

for covariance gradients, preconditioned via the Fisher metric, and clipped to trust regions—before being applied to update the agent fields.

Gauge frames are retracted onto the principal domain, covariances are sanitized, and dirty flags are set to invalidate any cached derived quantities that depend on the updated fields. Finally, we compute global metrics and the total action to assess convergence and prepare for visualization. This entire sequence is orchestrated by the `synchronous_step` function, which ensures that all agents are updated in a coordinated, lock-step fashion at each simulation tick, maintaining consistency of the multi-agent state across the entire base manifold.

The resulting dynamics exhibit stable convergence to stationary points of the free energy functional, with numerical precision sufficient to reliably distinguish between symmetric vacuum states and observation-induced symmetry-broken configurations—precisely the behavior required to validate the gauge-theoretic framework proposed in this work.

E.10 Computational Requirements

The simulations were performed on a workstation with an AMD Ryzen 9 5950X processor (16 cores, 32 threads) and 64 GB RAM. Typical simulation runs with $N = 8$ agents over a 20×20 spatial grid required approximately 60 minutes for convergence (500-800 steps). Memory usage scales as $\mathcal{O}(N \cdot H \cdot W \cdot K^2)$ where $H \times W$ is the spatial resolution and K is the fiber dimension.

E.11 Hyperparameter Configuration

All hyperparameters used in the experiments are documented in `config.py` within the repository. Key parameters include:

Parameter	Value	Description
$\eta_{\mu,q}$	1×10^{-1}	Belief mean learning rate
$\eta_{\Sigma,q}$	1×10^{-1}	Belief covariance learning rate
η_{ϕ}	1×10^{-1}	Gauge frame learning rate
τ	1.0	Attention temperature
α	1.0	Self-consistency weight
ϵ_{SPD}	1×10^{-8}	Covariance regularization floor
ρ	0.3	Trust region radius

Table 3: Standard hyperparameter configuration for all experiments.

E.12 Random Seed Reproducibility

All experiments use seeded random number generators (NumPy `RandomState`). The seeds used for each figure are documented in the repository's `experiments/` directory. To reproduce our results, run:

python generalized_simulation.py –config configs/vacuum_exp.py –seed 241

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Shun-ichi Amari. *Differential-Geometrical Methods in Statistics*. Springer, 1985.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10 (2):251–276, 1998.

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

Philip W Anderson. *Basic Notions of Condensed Matter Physics*. Benjamin/Cummings, 1984.

John Baez and Javier P Muniain. *Gauge Fields, Knots and Gravity*. World Scientific, 1994.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Stefan Boettcher and Charles T Brunson. Renormalization group for critical phenomena in complex networks. *Physical Review E*, 86(1):011128, 2012.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, 2019.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Marc Finzi, Max Welling, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. *arXiv preprint arXiv:2002.12880*, 2020.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Theodore Frankel. *The Geometry of Physics: An Introduction*. Cambridge University Press, 3rd edition, 2011.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

Karl J Friston, Thomas Parr, and Bert de Vries. The graphical brain: belief propagation and active inference. *Network Neuroscience*, 1(4):381–414, 2017.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020.

William Fulton and Joe Harris. *Representation Theory: A First Course*. Springer, 1991.

Raúl García-Millán, Marián Boguñá, and Ginestra Bianconi. Network renormalization. *arXiv preprint arXiv:2412.12988*, 2024.

Jeffrey Goldstone. Field theories with superconductor solutions. *Il Nuovo Cimento (1955-1965)*, 19(1):154–164, 1961.

Brian C Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*. Springer, 2nd edition, 2015.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. *arXiv preprint arXiv:1802.03690*, 2018.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Donald M MacKay. *Information, Mechanism and Meaning*. MIT Press, Cambridge, MA, 1969.

Mikio Nakahara. *Geometry, Topology and Physics*. CRC Press, 2nd edition, 2003.

Thomas Parr, Giovanni Pezzulo, and Karl J Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022.

Maxwell JD Ramstead, Michael D Kirchhoff, and Karl J Friston. A tale of two densities: active inference is enactive inference. *Adaptive Behavior*, 27(6):369–385, 2019.

Lijin Shen, Ioannis G Kevrekidis, and C William Gear. Coarse-graining multi-agent dynamics on a network. *Physica D: Nonlinear Phenomena*, 237(14-17):2202–2210, 2008.

Shlomo Sternberg. *Group Theory and Physics*. Cambridge University Press, 1994.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. In *arXiv preprint arXiv:1802.08219*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Steven Weinberg. *The Quantum Theory of Fields, Vol 2: Modern Applications*. Cambridge University Press, 1995.

Kenneth G Wilson and John Kogut. The renormalization group and the ϵ expansion. *Physics Reports*, 12(2):75–199, 1974.

Michael Wooldridge. *An Introduction to Multiagent Systems*. Wiley, 2nd edition, 2009.