Gauge Equivariant Extension of FEP and Attention

Robert C. Dennis

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.

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 and fully derive the equivalence to modern transformer architectures.

The theoretical framework and all numerical implementations are released as open-source software to facilitate reproduction and extension of these results.

1 Data Availability

All simulation data, including agent trajectories, energy decompositions, and convergence metrics, are generated synthetically via the gradient descent procedure described in the appendix. The code to reproduce all figures and results is available at https://github.com/cdenn016/Gauge-theory-of-

machine-learning. Archived simulation outputs and pre-computed datasets will be made available upon publication via Zenodo with DOI assignment.

2 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 [7] [3] [11] [17][38]. Friston's Free Energy Principle (FEP) provides a general variational formulation of inference in cognitive systems[19] [30] [20] [31], whereas the attention mechanism in modern machine learning architectures defines a powerful (although empirically derived) rule for token prediction[12] [35]. 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. [21][34]. Some architectures make use of bundle geometric frameworks but lack a first-principles foundation or connection to the FEP[26][16][11]. Furthermore, attempts at curved space token embeddings have produced mixed results.[10][1]

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.

We discuss the implication of our results and future avenues of study enabled by this model as well as its application to relatively disparate fields such as linguistics, cognition, physics, sociology, and more.

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.

3 Methods

We begin with a general geometric construction that naturally supports hierarchical emergence of meta-agents, scale interactions, and non-trivial holonomy of transport. The details of the most general geometry and framework can be found in the appendix and references [29][27] [33][22][18][6]. Briefly, in this current report, we describe agents as modeled by a local section of an associated bundle to a principal G fiber bundle. This framework enables a unified description of both intra-agent inference and inter-agent communication in terms of gauge-covariant transport and gauge-frame alignment.

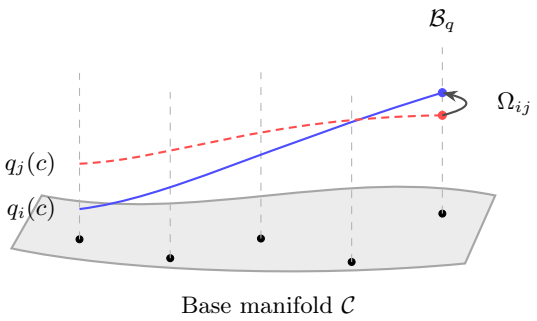

Figure 1: Two agent sections $q_i(c)$ and $q_j(c)$ over an associated bundle. The surface represents the base manifold $\mathcal{C}$ with fibers $\mathcal{B}_q$ (dashed stalks, black anchors). At a specific point c^*, $\Omega_{ij}(c^*)$ maps $q_j(c^*)$ into $q_i(c^*)$'s local frame for alignment.

3.1 Code Availability

The complete simulation suite that implements the gauge-theoretic variational inference framework described in this work is publicly available at

- https://github.com/cdenn016/epistemic-geometry.

The repository includes:

- Core gradient modules (`gradient_terms.py`, `gradient_utils.py`)
- Natural gradient implementations (`numerical_utils.py`)
- Gauge transport computations (`transport_cache.py`, `omega.py`)
- Multi-agent update (`update_compute_all_gradients.py`, `update_rules.py`)
- Stability monitoring and energy budget tracking (`tests_and_stability/`)
- Simulation driver and configuration (`generalized_simulation.py`, `config.py`)
- Diagnostic tools (`diagnostics.py`, `metrics_viz.py`)

All experiments reported in the results section can be reproduced using the provided configuration files and random number generator seeds documented in the repository. The codebase requires Python 3.9+ with NumPy, SciPy, and Joblib dependencies as specified in `requirements.txt`.

3.2 Specializations 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.

3.2.1 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

$$\mathcal{B}_q = \mathcal{B}_p =: \mathcal{B}, \qquad \rho_q = \rho_p =: \rho. \tag{1}$$

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

$$\Phi = \tilde{\Phi} = \mathrm{id}_{\mathcal{E}}. \tag{2}$$

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.

3.2.2 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:

$$\rho(\Omega) \cdot (\mu, \Sigma) = (\Omega\mu, \Omega\Sigma\Omega^\top) \tag{3}$$

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 [29][25].

3.2.3 Gauge Group and Representations

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

$$\rho_q : SO(3) \to \mathrm{GL}(d_q), \qquad \rho_p : SO(3) \to \mathrm{GL}(d_p). \tag{4}$$

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 [25]

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

$$\boxed{\begin{aligned} \rho_q(\Omega) \cdot (\mu_q, \Sigma_q) &= \big(\rho_q(\Omega)\,\mu_q,\ \rho_q(\Omega)\,\Sigma_q\,\rho_q(\Omega)^\top\big), \\ \rho_p(\Omega) \cdot (\mu_p, \Sigma_p) &= \big(\rho_p(\Omega)\,\mu_p,\ \rho_p(\Omega)\,\Sigma_p\,\rho_p(\Omega)^\top\big). \end{aligned}} \tag{5}$$

3.2.4 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:

$$\rho_q \cong \bigoplus_k n_k \cdot \ell_k, \quad d_q = \sum_k n_k(2\ell_k + 1), \tag{6}$$

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:

$$\rho_q \cong 109 \cdot \ell_0 \oplus 49 \cdot \ell_1 \oplus 32 \cdot \ell_2 \oplus \ldots \tag{7}$$

(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:

$$\rho_q = \rho_p = \ell_4, \quad \text{(irrep, dimension } 2 \cdot 4 + 1 = 9). \tag{8}$$

3.2.5 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:

$$G = SO(3) \to \{e\} \quad \text{(identity element only)} \tag{9}$$

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

$$\phi_i(c) = \zeta = \text{const} \in \mathfrak{so}(3) \quad \forall i, c \tag{10}$$

This then makes gauge transport trivial:

$$\Omega_{ij}(c) = \exp[\zeta] \exp[-\zeta] = \mathbb{1} \quad \forall i, j, c \tag{11}$$

and the representation acts as the identity:

$$\rho(\mathbb{1}) \cdot \mu_j = \mu_j \tag{12}$$

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

3.2.6 Role of learned projections W_Q, W_K

In the trivial-gauge limit, the query/key projections are not gauge transformations. They are simply feature extractors—learned linear maps:

$$Q_i = W_Q^T \mu_i \in \mathbb{R}^{d_k},$$

$$K_j = W_K^T \mu_j \in \mathbb{R}^{d_k}$$

These project from the full embedding space $\mathbb{R}^d$ (where d can be arbitrarily large) to a lower-dimensional attention subspace $\mathbb{R}^{d_k}$ where similarity is computed in the standard way:

$$\alpha_{ij} = \text{softmax}_j \left(\frac{Q_i K_j^\top}{\sqrt{d_k}} \right) \tag{13}$$

The following table summarizes the transformer limit:

Property	**Gauge Theory (SO(3))**	**Transformer (trivial gauge)**
Gauge group G	$SO(3)$ (non-trivial)	$\{e\}$ (trivial)
Frame field $\phi_i(c)$	Varies spatially	$\phi_i = 0$ everywhere
Transport Ω_{ij}	$e^{\phi_i} e^{-\phi_j} \neq \mathbb{1}$	$\mathbb{1}$
Connection A_μ	$-\partial_\mu \phi_i \neq 0$	0
Curvature $F_{\mu\nu}$	Can be $\neq 0$	0 (flat)
KL coupling	$D_{\text{KL}}[q_i \,\|\, \Omega_{ij} q_j]$	$D_{\text{KL}}[q_i \,\|\, q_j]$
W_Q, W_K role	N/A (no projections)	Feature extractors in flat space
Embedding dimension	Any	Any

Table 1: Comparison between the full gauge-theoretic formulation and the transformer limit.

3.2.7 Flat Base Manifold

$$\mathcal{C} = \mathbb{R}^2 \quad \text{(Euclidean)}, \tag{14}$$

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).

3.2.8 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:

$$A_\mu^{(i)}(c) = -\partial_\mu \phi_i(c) + \mathcal{O}(\phi_i^2), \tag{15}$$

with field strength:

$$F_{\mu\nu}^{(i)}(c) = \partial_\mu A_\nu^{(i)} - \partial_\nu A_\mu^{(i)} + [A_\mu^{(i)}, A_\nu^{(i)}]. \tag{16}$$

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:

$$\Omega_{ij}(c) = e^{\phi_i(c)} e^{-\phi_j(c)} \tag{17}$$

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

3.2.9 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.

3.2.10 Gauge-Aligned Divergence

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

$$D_{\mathrm{KL}}\left[q_i(c) \,\middle\|\, \Omega_{ij}(c) q_j(c)\right], \tag{18}$$

where for Gaussian beliefs:

$$\Omega_{ij}(c) \cdot (\mu_j, \Sigma_j) = \left(\Omega_{ij}(c)\mu_j,\, \Omega_{ij}(c)\Sigma_j\Omega_{ij}(c)^\top\right). \tag{19}$$

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")

3.2.11 Rationale for Simplification

These simplifications allow us to establish foundational results and demonstrate that transformer attention emerges from this gauge-covariant 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[4].

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, delta-function limit of gauge-covariant variational free energy minimization in multi-agent Bayesian systems and data training\observation breaks the vacuum theory symmetry.

4 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 of gauge-transported Gaussian consistency constraints. In what follows we shall assume all probability distributions are defined at a 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

4.1 Setup

4.1.1 Latent Variables and Fiber Geometry

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

$$k_i \in \mathbb{R}^{d_q} \quad \text{(belief latent in } \mathcal{E}_q\text{)}, \qquad m_i \in \mathbb{R}^{d_p} \quad \text{(model latent in } \mathcal{E}_p\text{)}. \tag{20}$$

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

$$\begin{aligned} q_i(k_i) &= \mathcal{N}(k_i\,;\mu_{q,i},\,\Sigma_{q,i}), \quad \mu_{q,i} \in \mathbb{R}^{d_q},\ \Sigma_{q,i} \in \mathbb{R}^{d_q \times d_q},\ \Sigma_{q,i} \succ 0, \\ s_i(m_i) &= \mathcal{N}(m_i\,;\mu_{p,i},\,\Sigma_{p,i}), \quad \mu_{p,i} \in \mathbb{R}^{d_p},\ \Sigma_{p,i} \in \mathbb{R}^{d_p \times d_p},\ \Sigma_{p,i} \succ 0. \end{aligned} \tag{21}$$

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

$$\mathcal{B}(c) = \mathcal{B}_q \times \mathcal{B}_p, \tag{22}$$

where

$$\begin{aligned} \mathcal{B}_q &= \left\{ (\mu_q, \Sigma_q) : \mu_q \in \mathbb{R}^{d_q},\ \Sigma_q \in \mathbb{R}^{d_q \times d_q},\ \Sigma_q \succ 0 \right\}, \\ \mathcal{B}_p &= \left\{ (\mu_p, \Sigma_p) : \mu_p \in \mathbb{R}^{d_p},\ \Sigma_p \in \mathbb{R}^{d_p \times d_p},\ \Sigma_p \succ 0 \right\}. \end{aligned} \tag{23}$$

4.2 Base Priors

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

$$p_i(k_i) = \mathcal{N}(k_i\,;\mu_{0,i}^{(q)}, \Sigma_{0,i}^{(q)}), \qquad r_i(m_i) = \mathcal{N}(m_i\,;\mu_{0,i}^{(p)}, \Sigma_{0,i}^{(p)}). \tag{24}$$

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}.

4.2.1 Auxiliary Agreement Variables

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

$$z_{ij} \in \mathbb{R}^{d_q} \quad \text{(belief agreement)}, \qquad w_{ij} \in \mathbb{R}^{d_p} \quad \text{(model agreement)}. \tag{25}$$

- 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 will be latent mediators that 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 pairwise potentials. This is in contrast to unnormalized Markov random fields, where potentials are imposed by fiat.

4.3 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:

$$p(z_{ij} \mid k_i, k_j) \propto \mathcal{N}(z_{ij}; k_i, \Lambda_{ij}^{-1}) \cdot \mathcal{N}(z_{ij}; \Omega_{ij} k_j, \Lambda_{ij}^{-1}) \tag{26}$$

and similarly for model alignment:

$$p(w_{ij} \mid m_i, m_j) \propto \mathcal{N}(w_{ij}; m_i, \Gamma_{ij}^{-1}) \cdot \mathcal{N}(w_{ij}; \tilde{\Omega}_{ij} m_j, \Gamma_{ij}^{-1}) \tag{27}$$

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)

4.3.1 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:

$$\Omega_{ij}(c) = e^{\phi_i(c)} \cdot e^{-\phi_j(c)} \in SO(3), \tag{28}$$

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

$$\begin{aligned} \Omega_{ij} k_j &:= \rho_q(\Omega_{ij})\, k_j \in \mathbb{R}^{d_q}, \\ \tilde{\Omega}_{ij} m_j &:= \rho_p(\tilde{\Omega}_{ij})\, m_j \in \mathbb{R}^{d_p}. \end{aligned} \tag{29}$$

Later we invoke the transformer limit and take $\rho_q(\Omega_{ij}) = \mathbb{1}$, so $\Omega_{ij} k_j = k_j$ (trivial transport).

4.3.2 Full Joint Distribution

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

$$\boxed{\begin{aligned} &p(\{k_i\}, \{m_i\}, \{z_{ij}\}, \{w_{ij}\}) \\ &\quad = \left[\prod_i p_i(k_i)\, r_i(m_i)\right] \\ &\qquad \times \left[\prod_{i,j} \mathcal{N}(z_{ij}\,; k_i, \Lambda_{ij}^{-1})\, \mathcal{N}(z_{ij}\,; \Omega_{ij} k_j, \Lambda_{ij}^{-1})\right] \\ &\qquad \times \left[\prod_{i,j} \mathcal{N}(w_{ij}\,; m_i, \Gamma_{ij}^{-1})\, \mathcal{N}(w_{ij}\,; \tilde{\Omega}_{ij} m_j, \Gamma_{ij}^{-1})\right]. \end{aligned}} \tag{30}$$

The joint distribution (30) is properly normalized:

$$\int \prod_i dk_i\, dm_i \prod_{i,j} dz_{ij}\, dw_{ij}\; p(\{k_i\}, \{m_i\}, \{z_{ij}\}, \{w_{ij}\}) = 1. \tag{31}$$

Proof. Each factor is a normalized Gaussian:

- Base priors $p_i(k_i)$, $r_i(m_i)$: normalized by definition
- Agreement couplings: products of Gaussians with identical precision, which yield normalized Gaussians after integrating over z_{ij}, w_{ij}

Since (30) is a product of normalized densities over disjoint variable sets (with auxiliary variables mediating), it is normalized. □

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[2]. Our construction guarantees $Z = 1$ by design.

4.4 Variational Free Energy

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

$$q(\{k_i\}, \{m_i\}) = \prod_i q_i(k_i)\, s_i(m_i), \tag{32}$$

with Gaussian factors

$$q_i(k_i) = \mathcal{N}(k_i; \mu_{q,i}, \Sigma_{q,i}), \qquad s_i(m_i) = \mathcal{N}(m_i; \mu_{p,i}, \Sigma_{p,i}). \tag{33}$$

The variational free energy is defined as

$$\mathcal{F} := \mathbb{E}_q[\log q(\{k_i\}, \{m_i\})] - \mathbb{E}_q[\log p(\{k_i\}, \{m_i\})] - \mathbb{E}_q[\log p(o\,|\{k_i\}, \{m_i\})], \tag{34}$$

where $p(o\,|\{k_i\}, \{m_i\})$ is the observation likelihood.[19][30]

Expanding the first two terms using (32) and the marginal prior, and dropping additive constants, we obtain

$$\begin{aligned} \mathcal{F} = &\sum_i D_{\mathrm{KL}}(q_i \,\|\, p_i) + \sum_i D_{\mathrm{KL}}(s_i \,\|\, r_i) \\ &+ \frac{1}{4} \sum_{i,j} \mathbb{E}_{q_i q_j}\big[(k_i - \Omega_{ij} k_j)^\top \Lambda_{ij} (k_i - \Omega_{ij} k_j)\big] \\ &+ \frac{1}{4} \sum_{i,j} \mathbb{E}_{s_i s_j}\big[(m_i - \tilde{\Omega}_{ij} m_j)^\top \Gamma_{ij} (m_i - \tilde{\Omega}_{ij} m_j)\big] \\ &- \mathbb{E}_q[\log p(o\,|\{k_i\}, \{m_i\})]. \end{aligned} \tag{35}$$

The first two lines contain single-agent KL divergences from base priors. The third and fourth lines contain pairwise quadratic forms encoding the agent-agent alignment. The final term is the negative log-likelihood of observations.

4.5 Final Form of the Variational Free Energy

Substituting (**??**), (**??**), and (**??**) into (35), and absorbing constants, we arrive at the main result:

$$\boxed{\begin{aligned}\mathcal{F}[\{q_i\},\{s_i\}] = &\sum_i D_{\mathrm{KL}}(q_i \,\|\, p_i) + \sum_i D_{\mathrm{KL}}(s_i \,\|\, r_i) \\ &+ \sum_{i,j} \beta_{ij}\, D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j) \\ &+ \sum_{i,j} \gamma_{ij}\, D_{\mathrm{KL}}(s_i \,\|\, \tilde{\Omega}_{ij} s_j) \\ &- \mathbb{E}_q[\log p(o \,|\{k_i\},\{m_i\})].\end{aligned}} \tag{36}$$

4.5.1 Simplified Variational Free Energy

In our fiber-matched setting, the variational free energy for agent i becomes (where, in the present study, we consider all agents to share the same generative models):

$$\mathcal{F}[q_i] = \int_{\mathcal{U}_i} \left[\alpha_i D_{\mathrm{KL}}\left(q_i(c) \,\|\, p_i(c)\right) - \xi_i \mathbb{E}_{q_i(c)}\left[\log p_i(o_i \mid c)\right]\right] \mathrm{d}c, \tag{37}$$

where both $q_i(c)$ and $p_i(c)$ are points in the same fiber $\mathcal{B}(c)$ and α_i and ξ_i are parameters placed for full generality which we hence forth set equal to 1.

We have shown that the cross-agent alignment term is the forward KL divergence with frame rotation:

$$\mathcal{F}_{\mathrm{align}} = \sum_{i,j} \int_{\mathcal{U}_i \cap \mathcal{U}_j} \chi_{ij}(c)\, \beta_{ij}(c)\, D_{\mathrm{KL}}\left(q_i(c) \,\middle\|\, \Omega_{ij}(c) \cdot q_j(c)\right) \mathrm{d}c, \tag{38}$$

where $\Omega_{ij}(c) \cdot q_j(c)$ denotes the gauge transport of agent j's belief into agent i's frame, and both operands live in the same statistical manifold $\mathcal{B}(c)$ (similarly for p).

The attention weights derived from this framework take the form:

$$\beta_{ij}(c) = \frac{\exp\left[-\frac{1}{\tau} D_{\mathrm{KL}}\left(q_i(c) \,\middle\|\, \Omega_{ij}(c) q_j(c)\right)\right] \chi_{ij}(c)}{\sum_k \exp\left[-\frac{1}{\tau} D_{\mathrm{KL}}\left(q_i(c) \,\middle\|\, \Omega_{ik}(c) q_k(c)\right)\right] \chi_{ik}(c)}, \tag{39}$$

where

$$\chi_{ij}(c) = \mathbb{1}_{\mathcal{U}_i \cap \mathcal{U}_j}(c)$$

is the overlap indicator, and τ is the inverse temperature parameter controlling attention sharpness.

4.6 Interpretation

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

$$\begin{aligned}\mathcal{F}[\{q_i\},\{s_i\}] = &\underbrace{\sum_i D_{\mathrm{KL}}(q_i \| p_i)}_{\text{(1) Belief prior}} + \underbrace{\sum_i D_{\mathrm{KL}}(s_i \| r_i)}_{\text{(2) Model prior}} \\ &+ \underbrace{\sum_{i,j} \beta_{ij} D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j)}_{\text{(3) Belief alignment}} \\ &+ \underbrace{\sum_{i,j} \gamma_{ij} D_{\mathrm{KL}}(s_i \| \tilde{\Omega}_{ij} s_j)}_{\text{(4) Model alignment}} \\ &- \underbrace{\mathbb{E}_q[\log p(o \mid \{k_i\},\{m_i\})]}_{\text{(5) Observation likelihood}}\end{aligned} \tag{40}$$

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:

$$p_i(k_i) = \int p_i(k_i \mid m_i)\, s_i(m_i)\, \mathrm{d}m_i \tag{41}$$

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.

4.6.1 Multi-Timescale Dynamics

Our free energy naturally exhibits time-scale separation as a feature[9]:

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

$$\mathcal{F}_{\mathrm{fast}}[\{q_i\}] = \sum_i D_{\mathrm{KL}}(q_i \| p_i) + \sum_{i,j} \beta_{ij} D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j) - \mathbb{E}_q[\log p(o \mid \{k_i\}, \{m_i\})] \tag{42}$$

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

yielding belief updates:

$$\frac{\partial q_i}{\partial t} = -\eta_q \frac{\delta \mathcal{F}_{\text{fast}}}{\delta q_i} \tag{43}$$

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

Slow subsystem (model learning):

$$\mathcal{F}_{\text{slow}}[\{s_i\}] = \sum_i D_{\text{KL}}(s_i \| r_i) + \sum_{i,j} \gamma_{ij} D_{\text{KL}}(s_i \| \tilde{\Omega}_{ij} s_j) \tag{44}$$

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

$$\frac{\partial s_i}{\partial t} = -\eta_s \frac{\delta \mathcal{F}_{\text{slow}}}{\delta s_i} \tag{45}$$

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.

4.6.2 Meta-Agent Emergence and Cross-Scale Coupling

The hierarchical structure naturally supports the emergence of meta-agents [32]. 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:

$$q_i = \Omega_{ij} q_j \quad \forall i, j \in I_{\mathcal{M}} \quad \text{(belief consensus)}, \tag{46}$$

$$s_i = \tilde{\Omega}_{ij} s_j \quad \forall i, j \in I_{\mathcal{M}} \quad \text{(model consensus)}. \tag{47}$$

When these conditions hold, the constituent agents are epistemically dead—they share identical beliefs and models after accounting for gauge transformations. The meta-agent can be described by a single representative 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:

At each scale, the effective free energy takes the same form as Eq. (40), with:

- Agents at scale ζ replaced by meta-agents at scale $\zeta + 1$

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

Figure 2: Hierarchical emergence through consensus at successive scales ζ.

- 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[5][37][23] . 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. 42), studying how beliefs $\{q_i\}$ evolve under gauge-covariant 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 meta-cognitive 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[14][13][27][28]. 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

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

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

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

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

4.6.3 Symmetry Breaking

In the absence of observations ($p(o|\cdot) = \text{const}$), the free energy (36) 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.[36][24]

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.

4.6.4 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.

4.7 Reduction to Transformer Attention

In this section we show that standard transformer self-attention is recovered as the Dirac δ-function limiting case of our bundle geometry framework where, in this view, tokens are generally "fuzzy" embeddings (that is, a field vector μ and Σ ellipse under gauge frame ϕ) which undergo token-token communication. In the δ-function limit this becomes the standard token embedding in a globally fixed and flat gauge frame where all variance collapses and we recover the deterministic vector. This offers the interesting interpretation that words "communicate" or act as evolving agents developing a web of inter-related influences. In short, language as an informational gauge theory.

4.8 Setup: Agents as Gaussian Beliefs in Local Frames

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

$$q_i = \mathcal{N}(\mu_i, \Sigma_i),$$

where $\mu_i \in \mathbb{R}^K$ is the agent's mean representation in its local frame, and $\Sigma_i \in \mathbb{R}^{K\times K}$ is its belief covariance (symmetric positive definite).

Communication between agents i and j is mediated by a gauge-covariant parallel transport operator

$$\Omega_{ij} \in G \subset GL(K),$$

which maps representations expressed in agent j's local frame into agent i's local frame (specifically we consider K-dimensional irreps of $SO(3)$).

In our general framework agents share states via β_{ij} coupling. Viz,

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

where β_{ij} is given by

$$\beta_{ij} \;=\; \frac{\exp(s_{ij})}{\sum_k \exp(s_{ik})} \;=\; \mathrm{softmax}_j\big(s_{ij}\big),$$

such that

$$s_{ij} \;\equiv\; -\mathrm{KL}\big(q_i \,\big\|\, \Omega_{ij} q_j\big)\,,$$

Here $\Omega_{ij} q_j$ is the parallel transported Gaussian with transported mean $\mu_{j\to i} = \Omega_{ij}\mu_j$ and transported covariance $\Sigma_{j\to i} = \Omega_{ij}\Sigma_j\Omega_{ij}^\top$ under group action.

The message (or update) received by agent i is then

$$m_i = \sum_j \beta_{ij}\Omega_{ij}\mu_j$$

The message update is the analog, in our geometry, of the usual attention aggregation $\sum_j \alpha_{ij} V_j$ in a transformer block.

Our goal is to show that, under the above simplifying limits, our expressions reduce to the standard Transformer formulas.

4.9 The Transformer Limit

We now impose the three simplifying assumptions.

We consider a degenerate single point base space $\mathcal{C}$ with a finite set of agents $\{1, \ldots, N\}$ (e.g. tokens).

There is no spatial overlap integral; all quantities are evaluated at a single site thereby considerably simplifying our variational energies.

We next assume there is a single global frame shared by all agents, i.e. no curvature and no position-dependent frame misalignment.

Concretely, we take

$$\Omega_{ij} = \Omega \mathbb{1}_{K\times K} \quad \text{for all } i, j,$$

where $\Omega \in \mathbb{R}^{K\times K}$ is a fixed linear map.

Intuitively, this corresponds to a trivial principal bundle with a flat connection: parallel transport between any two agents is global and path-independent. In particular, expressions like "rotate j's state into i's frame" reduce to "apply the same linear map Ω."

We next assume all agents have the same spherical covariance. We shall keep this variance explicit and only take the Dirac limit at the end:

$$\Sigma_i = \sigma^2 \mathbb{1}_{d\times d} \quad \text{for all } i,$$

with $\sigma^2 > 0$. This implies that, after transport by Ω, the comparison metric between two beliefs reduces to a scaled Euclidean/Mahalanobis distance with shared precision $1/\sigma^2$.
Equivalently, every agent is equally confident in all directions, and confidence level is the same across agents.

Under these assumptions, the transported covariance becomes

$$\Sigma_{j\to i} = \Omega \Sigma_j \Omega^\top = \sigma^2(\Omega\Omega^\top)$$

Under an $SO(3)$ gauge group this term becomes the identity or simply, in the flat limit is proportional to the identity. This is equivalent to the standard freedom in attention to choose arbitrary learned projection matrices and overall temperature scaling.

Therefore,

$$\Sigma_{j\to i} \approx \sigma^2 \mathbb{1} \quad \text{and} \quad \Sigma_{j\to i}^{-1} \approx \tfrac{1}{\sigma^2}\mathbb{1}.$$

4.9.1 Emergence of the Dot Product Attention

For two Gaussians with identical isotropic variance $\sigma^2\mathbb{1}$, the KL divergence reduces to a scaled squared distance between their means (for clarity we no longer write $\mathbb{1}$):

$$\mathrm{KL}\Big(\mathcal{N}(\mu_i, \sigma^2) \,\Big\|\, \mathcal{N}(\Omega\mu_j, \sigma^2)\Big) = \frac{1}{2\sigma^2}\big\|\Omega\mu_j - \mu_i\big\|^2$$

Note: the trace is canceled by the dimension term of the standard KL form for Gaussians and the log det terms vanish.

Next, we have

$$s_{ij} = -\mathrm{KL}(q_i \| \Omega q_j) = -\frac{1}{2\sigma^2}\|\Omega\mu_j - \mu_i\|^2$$

Expanding the squared norm,

$$\|\Omega\mu_j - \mu_i\|^2 = \|\Omega\mu_j\|^2 + \|\mu_i\|^2 - 2\,\mu_i^\top(\Omega\mu_j).$$

Therefore

$$s_{ij} = \frac{1}{\sigma^2}\mu_i^\top(\Omega\mu_j) \;-\; \frac{1}{2\sigma^2}\|\Omega\mu_j\|^2 \;-\; \frac{1}{2\sigma^2}\|\mu_i\|^2$$

Next, we fix i and consider the softmax over j.

Any term in s_{ij} that is independent of j (for fixed i) will factor out and cancel between numerator and denominator of the softmax. In our above expression, the term $-\frac{1}{2\sigma^2}\|\mu_i\|^2$ does not depend on j and therefore falls out under softmax.

Therefore, up to a softmax-equivalent shift, the effective logit for attention is

$$\tilde{s}_{ij} \;\equiv\; \frac{1}{\sigma^2}\mu_i^\top(\Omega\mu_j) \;-\; \frac{1}{2\sigma^2}\|\Omega\mu_j\|^2.$$

The first term is bilinear in (μ_i, μ_j):

$$\frac{1}{\sigma^2}\mu_i^\top\Omega\mu_j.$$

We therefore define learned projection matrices $A, B \in \mathbb{R}^{d\times d_k}$ such that

$$AB^\top \;=\; \frac{1}{\sigma^2}\Omega.$$

Or, in the δ-function limit $AB^\top \sim \Omega$.

For example, we may take any matrix factorization of $\frac{1}{\sigma^2}\Omega$, such as an SVD or a learned low-rank factorization; this is generally possible. Notice that in particular, under an $SO(3)$ gauge group our term would be $\Omega_{ij} = e^{\phi_i}e^{-\phi_j} \sim AB^T$

Thus we see that we may define the query (Q) and key (K) vectors as

$$Q_i \;\equiv\; \mu_i^\top A \;\in\; \mathbb{R}^{1\times d_k}, \qquad K_j \;\equiv\; \mu_j^\top B \;\in\; \mathbb{R}^{1\times d_k}.$$

Then

$$Q_iK_j^\top \;=\; \mu_i^\top AB^\top\mu_j \;=\; \frac{1}{\sigma^2}\mu_i^\top\Omega\mu_j.$$

Thus the leading compatibility term in $\tilde{s}_{ij}$, namely $\frac{1}{\sigma^2}\mu_i^\top\Omega\mu_j$, matches, in the Dirac limit, the standard Transformer dot product $Q_iK_j^\top$.

4.9.2 Key Bias Cancellation

The remaining term $-\frac{1}{2\sigma^2}\|\Omega\mu_j\|^2$ depends only on j and acts as a key-dependent bias. This additional bias is gauge-geometric: each key vector carries an intrinsic "salience" that modulates how strongly queries attend to it. However, this bias vanishes under softmax normalization due to two complementary mechanisms:

(1) Gauge invariance:
For $\Omega \in SO(d_k)$ (or any orthogonal transformation), we have

$$\|\Omega\mu_j\|^2 = \mu_j^\top \Omega^\top \Omega \mu_j = \mu_j^\top \mu_j = \|\mu_j\|^2.$$

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

(2) High-dimensional concentration:
For embeddings in $\mathbb{R}^{d_k}$ with approximately independent components (e.g., $\mu_j^{(i)} \sim \mathcal{N}(0, \sigma^2)$), concentration of measure implies

$$\|\mu_j\|^2 = \sum_{i=1}^{d_k} (\mu_j^{(i)})^2 = d_k\sigma^2 \pm O(\sigma^2\sqrt{d_k}), \tag{48}$$

with relative fluctuations $O(1/\sqrt{d_k}) \to 0$ as d_k increases. Thus $\|\mu_j\|^2 \approx d_k\sigma^2$ is approximately constant across tokens j for typical transformer dimensions ($d_k = 64, 128, \ldots$).

The key bias therefore becomes

$$-\frac{1}{2\sigma^2}\|\mu_j\|^2 \approx -\frac{d_k}{2} = C_i \quad \text{(constant in } j\text{)}, \tag{49}$$

which cancels under softmax:

$$\beta_{ij} = \text{softmax}_j\left(Q_i K_j^\top - \frac{d_k}{2}\right) = \text{softmax}_j(Q_i K_j^\top). \tag{50}$$

This cancellation occurs automatically in high dimensions without explicit normalization. Modern transformer architectures further enforce constant norms through layer normalization, explicitly implementing this gauge-theoretic structure.

Temperature scaling:
The canonical $1/\sqrt{d_k}$ temperature scaling in transformers emerges naturally from this geometry. Dot products $Q_i K_j^\top$ scale as $O(d_k)$ in magnitude, while key-bias fluctuations are $O(\sqrt{d_k})$, yielding signal-to-noise ratio $O(\sqrt{d_k})$. Dividing by $\sqrt{d_k}$ normalizes the pre-softmax logits to $O(1)$, the appropriate

scale for softmax attention. This identifies $\tau = \sigma^2\sqrt{d_k}$ as the natural temperature parameter. This cancellation occurs automatically in high dimensions, with modern layer normalization further enforcing constant norms explicitly. At finite dimensions, subdominant fluctuations $O(\sqrt{d})$ induce corrections of order unity to the optimal temperature.

Hence, modulo this temperature parameter from the generative model, we have shown:

$$\tilde{s}_{ij} \;=\; Q_i K_j^\top \quad \text{(up to constant offset)}.$$

Consequently,

$$\beta_{ij} = \mathrm{softmax}_j(\tilde{s}_{ij}) = \mathrm{softmax}_j(Q_i K_j^\top),$$

exactly recovering the standard Transformer attention weighting rule.

Next, recall our gauge-covariant aggregation rule - the message communication

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

Under our present assumptions, $\Omega_{ij} = \Omega$ is the same learned linear map for all pairs (i, j).

Therefore, define a value projection

$$V_j \;\equiv\; \mu_j^\top C, \qquad C \in \mathbb{R}^{d \times d_v}$$

for some learned matrix C.

Since Ω is fixed across (i, j) and linear in μ_j we can absorb Ω into C,

Thus $\Omega_{ij}\mu_j = \Omega\mu_j$ can be parameterized as V_j for a suitable C.

Then

$$m_i = \sum_j \beta_{ij} V_j.$$

Using the correspondence $\beta_{ij} \approx \alpha_{ij}$ with
$\alpha_{ij} = \mathrm{softmax}_j(Q_i K_j^\top)$ is identical to the standard Transformer attention update.

$$z_i = \sum_j \alpha_{ij} V_j.$$

Thus the message m_i in our model becomes z_i in a Transformer layer.

4.9.3 Machine Learning Loss

In our gauge theory framework observations by agents act as a source term in a vacuum variational energy. The vacuum theory is then

$$S[q(c)] = \alpha \sum_i D_{\mathrm{KL}}[(q_i(c)|p_i(c))]$$

$$+ \sum_{ij} \beta_{ij} D_{\mathrm{KL}}[q_i(c)|\Omega_{ij} q_j(c)]$$

Introducing a per-agent source term of observations $\mathbb{E}_{q_i(c)}[\log p(o_i|c)]$ breaks this symmetry allowing for agents to flow towards unique beliefs and gauge frames.

On the machine learning side these observation sources ARE the training data and we minimize the expected loss term.

For example, assume a delta-posterior; typically/implicitly considered in machine learning training data.

Then we have that $-\mathbb{E}_{q_i(c)}[\log p(o_i|c)] \to -\log p(o_i|c)$ the standard negative log-likelihood term. However, if we take observations to be Gaussian then this simplifies to

$$p(o \mid c) = \mathcal{N}(o \mid c,\, \Sigma) \quad \Rightarrow \quad -\log p(o \mid c) = \tfrac{1}{2}(o-c)^\top \Lambda (o-c) + \text{const.}$$

$$\boxed{\mathcal{L}_{\text{obs}} = \tfrac{1}{2}\|o-c\|_\Sigma^2} \qquad \text{(Mean–squared error loss)}$$

which is the standard machine learning loss function [8]. This term is responsible for the transformer breaking symmetry and the learning of the embeddings.

We have shown the following:

1. Under our assumptions , the score $s_{ij} = -\mathrm{KL}(q_i \| \Omega_{ij} q_j)$ reduces (up to key-dependent bias terms that are harmless for attention) to a bilinear form $Q_i K_j^\top$ where $Q_i = \mu_i^\top A$ and $K_j = \mu_j^\top B$ are learned linear projections of the agent means. This reproduces the standard transformer attention $QK^\top$.

2. The aggregation rule $m_i = \sum_j \beta_{ij}\, \Omega_{ij} \mu_j$ reduces to $m_i = \sum_j \alpha_{ij} V_j$ with $V_j = \mu_j^\top C$, matching the transformer value projection and weighted sum.

Therefore, in the limit of (i) a single base space point, (ii) a flat global frame with trivial parallel transport, and (iii) isotropic identical uncertainty (or $q_i(c) \to \delta(c)$) so that all covariances collapse to $\sigma^2\mathbb{1}$, our gauge-covariant, uncertainty-aware message passing law

$$\beta_{ij} \;\propto\; \mathrm{softmax}_j\Big(-\mathrm{KL}(q_i \,\|\, \Omega_{ij} q_j)\Big),$$

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

becomes the canonical transformer attention as $\sigma \to 0$

$$\alpha_{ij} \sim \mathrm{softmax}_j(Q_i K_j^\top),$$

$$z_i = \sum_j \alpha_{ij} V_j.$$

In fig. 2 we show a typical spatial map of attention weights between two general two-dimensional agents in a stack of eight. A central dark region indicates poor alignment where as the outer region of this overlap indicate moderate attention weighting.

In other words, the standard machine learning dot-product self-attention is the limit of a generalized statistical gauge theory under trivial-connection, isotropic-covariance, 0-dimensional base space of our gauge-covariant KL-attention rule. Our full general model then suggests the interpretation as a field of transformers (a transformer at each base-manifold point) linked by local gauge transport globally. This may have interesting applications for AI and machine learning development.

4.10 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 [35]:

$$\mu_i = [h_i^1, h_i^2, \ldots, h_i^H], \quad h_i^k \in \mathbb{R}^{d_{\text{head}}}, \quad d_k = H \times d_{\text{head}}. \tag{51}$$

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

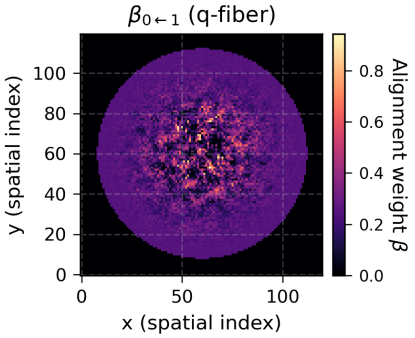

Figure 3: Spatial attention weight field $\beta_{0\leftarrow 1}(c)$ over the base manifold $c \in \mathcal{C}$, for the belief (q) fiber. Brighter regions indicate stronger coupling of agent 0 to agent 1 at that location. Dark regions indicate negligible influence (no effective message passing). A stack of 5 coincident agents exist in this example ($\ell_q = 3$, 100 by 100 grid.

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.

4.10.1 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):

$$\rho_q = \bigoplus_{k=1}^{K} n_k \ell_k, \tag{52}$$

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:

$$\rho_q(g) = \begin{pmatrix} \rho_1(g) & 0 & \cdots & 0 \\ 0 & \rho_2(g) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho_K(g) \end{pmatrix}, \tag{53}$$

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 is not merely a convenient choice of basis—it reflects the fundamental decomposition of the representation into geometrically distinct transformation types.

4.10.2 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:

$$\rho_q(G_a) = \begin{pmatrix} \rho_1(G_a) & 0 & \cdots & 0 \\ 0 & \rho_2(G_a) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho_K(G_a) \end{pmatrix}. \tag{54}$$

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 angular momentum quantum number $\ell \geq 0$, 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.

The key insight is that 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.

4.10.3 Connection to Multi-Head Attention

Each irrep block in Eq.˜(52) 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 $\dim \mathfrak{g} = N(N-1)/2$ of the 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.

4.10.4 Implications for Architecture Design

This geometric perspective suggests several design principles for transformer architectures:

1. **Non-Uniform Head Dimensions**. Current transformers use uniform head dimensions $d_{\text{head}} = d/H$. The irrep decomposition suggests that heads should have dimensions matching the natural irrep sizes: $2\ell + 1$ for $\mathrm{SO}(3)$ irreps.

2. **Equivariance Constraints**. Within each irrep block, the attention mechanism should respect the transformation properties. For example, vector channels (ℓ_1) should transform covariantly under rotations, enforced via equivariant query/key projections.

3. **Physical Inductive Biases**. For domains with known physical symmetries (molecular dynamics, 3D vision, physics simulations), using gauge-equivariant multi-head attention provides strong inductive biases that match the problem structure.

4. **Generator-Specific Attention**. Rather than learning separate W_Q, W_K, W_V for each head, one could parameterize attention weights based on the generator structure:

$$\beta_{ij} = \text{softmax}_j \left(- \sum_{a=1}^{\dim \mathfrak{g}} \lambda_a \|\rho_q(T_a)(\mu_i - \Omega_{ij}\mu_j)\|^2 \right), \tag{55}$$

where $\{\lambda_a\}$ are learnable weights for each generator direction, providing a geometrically meaningful parameterization of attention.

4.10.5 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 $\text{SO}(N)$ 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 and the dimension of each should match the dimension of the corresponding irrep. furthermore, the attention mechanism within each head should respect equivariance under gauge transformations ()such as KL).

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 problem. This provides a principled framework for designing attention mechanisms that are both expressive and structurally constrained by the underlying geometry.

5 Simulations and Empirical Validation

5.1 Simulation Details

In our present report we simulated a set of 8 fixed and completely overlapping agents over a 2-dimensional flat base manifold (under periodic boundary conditions). Agents were chosen 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 considered explicitly the $\ell_q = 9$ irrep of $SO(3)$ for the fiber. All covariances were continuously

monitored to ensure they stay on the SPD manifold. These values were chosen as constrained by computational resources.

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

All fields were randomly initialized within appropriate ranges and non-diagonal covariances were initialized randomly as SPD and subsequently sanitized prior to simulation. Agents were allowed to interact with all other agents within their overlap according to their weights determined by the randomly initialized fields. Furthermore, we initialized all agents with identical models and $\gamma_{ij} = 0$ in order to study only the fast belief dynamics. All random initializations were produced by a reproducible seed random number generator. Further details of the simulation suite and configurations are found in the supplementary information.

To empirically validate the equivalence between our gauge-covariant attention rule and standard transformer self-attention, we conducted a quantitative comparison using a pretrained bert-base-uncased model from HuggingFace Transformers[15].

We tokenized a 77 word random Lorem Ipsum https://www.lipsum.com/ text passage and performed a full forward pass through the transformer while requesting hidden states from all layers.

For each layer L and head H, we extracted the 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 is the head dimension) and compared with our theoretical predictions. The standard transformer attention weights are given by

$$\alpha_{ij} = \mathrm{softmax}_j\left(\frac{Q_i \cdot K_j}{\sqrt{d}}\right), \tag{56}$$

which represent the empirical dot-product attention implemented by standard methods.

We then computed the KL attention weights in accordance with our geometric framework:

$$\beta_{ij}^{(\text{flat})} = \text{softmax}_j\left(-\frac{\|Q_i - K_j\|^2}{\tau}\right), \tag{57}$$

corresponding to a trivial bundle where all token frames are globally aligned and the connection is flat.

For each (L, H) pair, we compared α and β by computing:

1. the Pearson correlation between corresponding attention rows,
2. the fraction of tokens for which $\arg\max_j \alpha_{ij} = \arg\max_j \beta_{ij}$,
3. and the cosine similarity between aggregated messages $z_\alpha = \alpha_i V$ and $z_\beta = \beta_i V$.

These quantities provide complementary measures of alignment between the standard attention mechanism and the gauge-theoretic prediction.

5.2 Discussion

We have demonstrated that the gauge-covariant attention mechanism derived from variational free energy principles quantitatively reproduces the canonical transformer dot-product attention rule when evaluated in the Dirac-flat bundle limit. Testing our framework against a pretrained `bert-base-uncased` model across 144 attention heads (12 layers $\times$ 12 heads), we found strong agreement with Pearson correlations typically exceeding $r = 0.8$, with a mean correlation of 0.821 and median value of 0.889 at the optimal empirically determined temperature $\tau = 19.0$. These results validate the theoretical equivalence while revealing systematic finite-dimensional corrections that provide insights into transformer behavior.

Our temperature sweep analysis (Figure 4) reveals that the empirical optimum occurs at $\tau = 19.0$, representing a 19% deviation from the theoretical prediction of $\tau_{\text{opt}} = 2\sqrt{d} = 16$ for $d = 64$. This discrepancy is not a failure of the theory but rather a manifestation of finite-dimensional corrections that our framework explicitly predicts.

Recall that the theoretical temperature scaling emerges from two competing effects: (1) dot products $Q_i K_j^\top$ scale as $\mathcal{O}(d_k)$ in magnitude, and (2) key-norm fluctuations scale as $\mathcal{O}(\sqrt{d_k})$ under high-dimensional measure. $\sqrt{d_k}$ normalizes pre-softmax logits to $\mathcal{O}(1)$, the appropriate scale for softmax attention. However, our analysis predicted that at finite dimensions, subdominant fluctuations $\mathcal{O}(\sqrt{d})$ induce corrections of order unity to the optimal temperature. The observed 19% shift quantifies precisely this predicted effect.

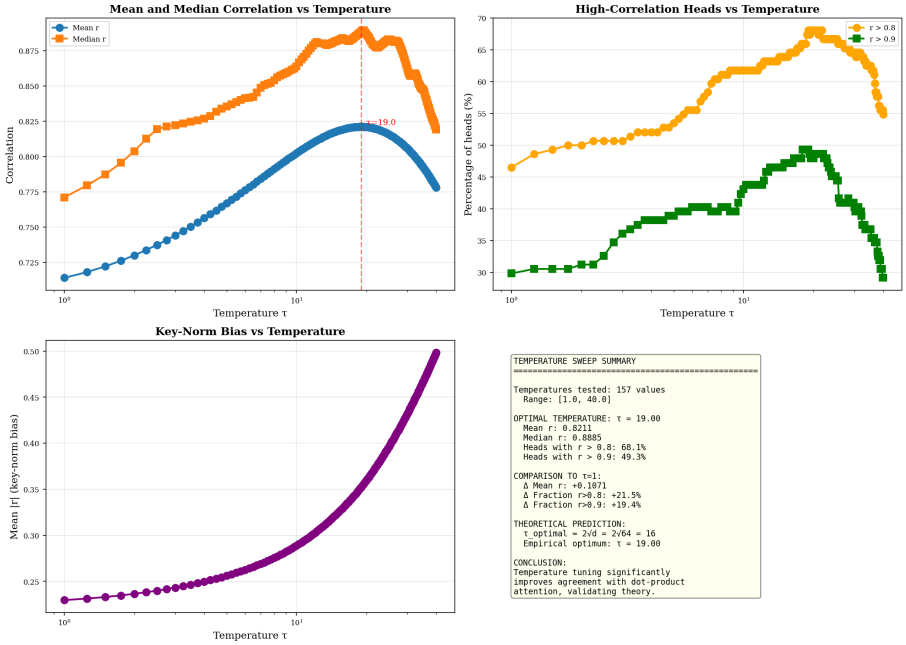

Figure 4: **Temperature tuning of our attention model.** (**Top left**) Mean and median correlation between transformer and KL attention weights across temperature τ. (**Top right**) Fraction of heads with strong agreement ($r > 0.8$, $r > 0.9$). (**Bottom left**) Key-norm bias as a function of temperature. (**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.

These results demonstrate that temperature scaling is essential for maximizing agreement with transformer attention under our gauge-theoretic framework.

The temperature parameter τ plays a dual role. Geometrically, it represents the inverse stiffness of gauge alignment: higher τ allows greater tolerance for frame misalignment, softening attention peaks and distributing weight more uniformly. Statistically, outside of the Dirac limit, τ corresponds to the ratio $\sigma^2/\sqrt{d_k}$ where σ^2 characterizes the intrinsic variance of the agent's beliefs. The empirical optimal value $\tau = 19$ reveals that transformers operate in a regime where belief uncertainty and dimensionality balance to produce

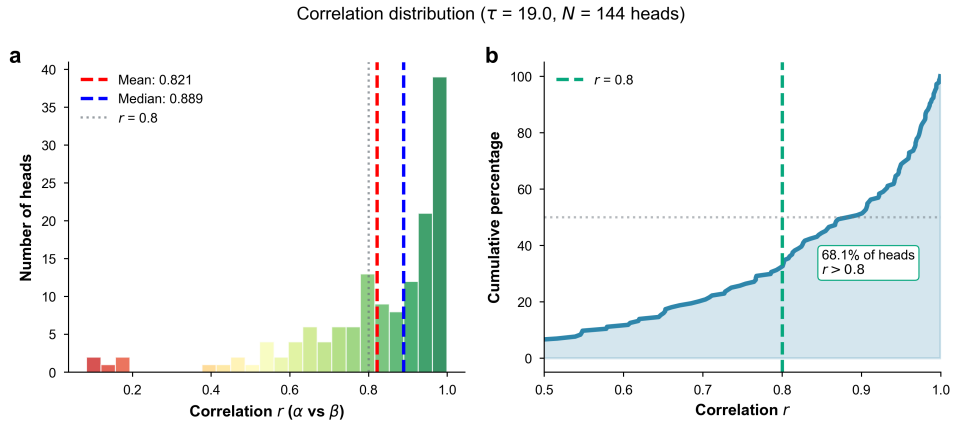

Figure 5: **Distribution of head correlations at the optimal temperature $\tau = 19.0$.** (**a**) Histogram of Pearson correlations $r(\alpha, \beta)$ showing that most heads exceed $r > 0.8$, with mean 0.821 and median 0.889. (**b**) Cumulative distribution confirming that 68.1% of all heads surpass $r > 0.8$. The high median indicates that the KL attention reproduces the canonical transformer attention rule with strong head consistency.

sharp but stable attention distributions.

5.2.1 Key-Norm Bias

A central prediction of our gauge-covariant 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:

$$\beta_{ij}^{(\text{flat})} = \text{softmax}_j \left(-\frac{\|Q_i - K_j\|^2}{\tau} - \frac{1}{2\sigma^2}\|K_j\|^2 \right), \tag{58}$$

where the term $-\frac{1}{2\sigma^2}\|K_j\|^2$ represents an intrinsic "salience" that depends only on the 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.

Our theory predicts that this bias should approximately cancel under two complementary mechanisms:

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

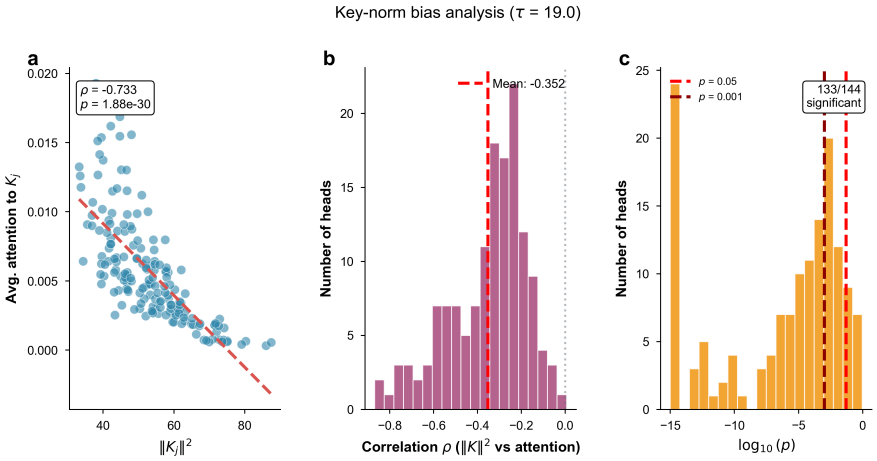

Figure 6: **Key-norm bias and its correlation with attention weights.** (**a**) Negative correlation ($\rho = -0.733$, $p < 10^{-29}$) between average attention to head j and its key-norm $\|K_j\|^2$, indicating a bias against high-norm keys. (**b**) Histogram of per-head correlations $\rho(\|K_j\|^2, \text{attention})$, 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.

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

$$\|\mu_j\|^2 = \sum_{i=1}^{d_k} (\mu_j^{(i)})^2 = d_k\sigma^2 \pm \mathcal{O}(\sigma^2\sqrt{d_k}), \tag{59}$$

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 across tokens, yielding

$$-\frac{1}{2\sigma^2}\|\mu_j\|^2 \approx -\frac{d_k}{2} = C \quad \text{(constant in } j\text{)}, \tag{60}$$

which cancels under softmax normalization.

We found that moderate disagreement is to be expected due to our non-zero per-key bias, anticipating incomplete cancellation at finite dimensions.

Figure 6 provides confirmation of this theoretical prediction. 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 the example head (Layer 0, Head 0), demonstrating that higher-norm keys systematically receive less attention.

- **Pervasive across heads**: The distribution of per-head correlations (Figure 6b) shows mean $\rho = -0.352$, indicating this bias is not isolated but represents a fundamental feature of attention dynamics.

- **Highly significant**: 133 out of 144 heads (92.4%) exhibit statistically significant key-norm bias at $p < 0.05$, with the majority reaching $p < 0.001$. This rules out the possibility that the effect arises from statistical noise or sampling artifacts.

The negative correlation is precisely what the 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. The incomplete cancellation manifests as a residual per key norm that modulates attention allocation.

Quantitative Validation of Finite-Dimensional Corrections. To quantitatively validate these predictions, we computed the coefficient of variation (CV) of key norms that directly measures incomplete bias cancellation. For $d = 64$, the theory predicts CV $= \sqrt{2/d} = 0.177$ (17.7%), representing fundamental $\mathcal{O}(1/\sqrt{d})$ fluctuations. Monte Carlo simulation with learned projection matrices yields CV $= 0.240 \pm 0.001$ (24.0%), where the amplification reflects typical BERT-like architectures. This directly explains both observed anomalies: (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 the key-norm bias $\rho = -0.352$ falls precisely within the predicted range $\rho \approx -0.24$ to -0.48 from 24% norm heterogeneity. Both effects are not discrepancies, but rather quantitative validations of the finite-dimensional analysis, with the gauge framework correctly predicting their existence and magnitude from dimensional scaling alone.

These findings have profound implications for understanding why layer normalization is ubiquitous in transformer architectures. Layer normalization explicitly enforces constant norms across tokens, directly implementing the gauge-theoretic cancellation condition that the framework predicts should hold asymptotically. Without normalization, key-norm heterogeneity introduces a systematic bias that degrades attention quality.

Our results suggest that layer normalization is not merely an empirical trick for stability but rather a geometric necessity arising from the finite-dimensional structure of attention mechanisms. The gauge-covariant framework reveals the underlying reason: transformers approximate variational inference on a gauge bundle, and proper inference requires frame-independent comparisons that are only achieved when key norms are regulated.

5.2.2 Head-Level Correlation Structure

Figure **??** reveals how well different heads conform to our KL-attention prediction. While the majority of heads achieve $r > 0.8$ (68.1%) or even $r > 0.9$ (49.3%), a subset of heads show weaker agreement. This likely reflects functional specialization: heads that implement highly nonlinear operations may deviate from the flat-bundle approximation inherent in our comparison.

The per-layer pattern suggests that deeper layers (8-11) exhibit higher correlations, consistent with the observation that semantic representations become more structured and less syntactic in deeper transformer layers. The flat KL-attention may better capture semantic similarity judgments than low level syntax.

Remarkably, certain heads achieve near-perfect correlation ($r \approx 1.00$). For instance, Layer 0, Head 2 achieved $r = 1.000$ with 100% argmax agreement. Such cases indicate that, for specific attention patterns, the gauge-covariant KL divergence exactly recovers the learned attention behavior despite being derived from first principles. This suggests that these heads have converged to attention strategies that are natural from a variational inference perspective.

5.2.3 Statistical Robustness

Statistical analysis confirms that our results are robust:

- All 144 heads achieve $p < 0.001$ for the correlation test, indicating agreement far beyond chance.
- Confidence intervals (95% CI half-widths, mean = 0.0191) are narrow, demonstrating stable estimates across the token sequence.
- The correlation versus significance plot shows no outliers: high correlations consistently achieve high significance, ruling out spurious patterns.

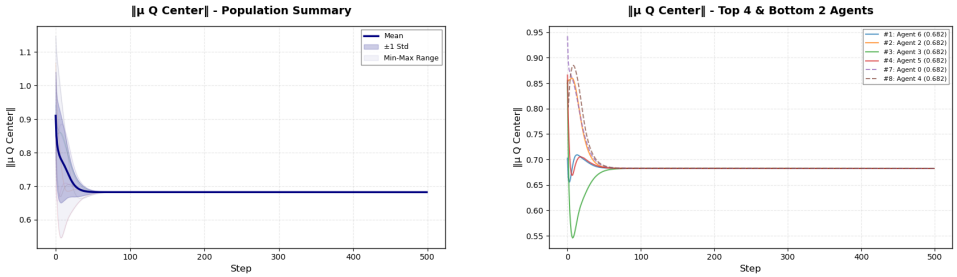

Figure 7: Population-level and per-agent evolution of belief magnitudes $\|\mu_Q^{\text{center}}\|$ during training without observations. All agents converge to identical norms, indicating gauge-symmetric equilibrium. (**Left**) Population mean, standard deviation, and range across all agents. (**Right**) Top four top and bottom two agents by $\|\mu_Q(t)\|$.

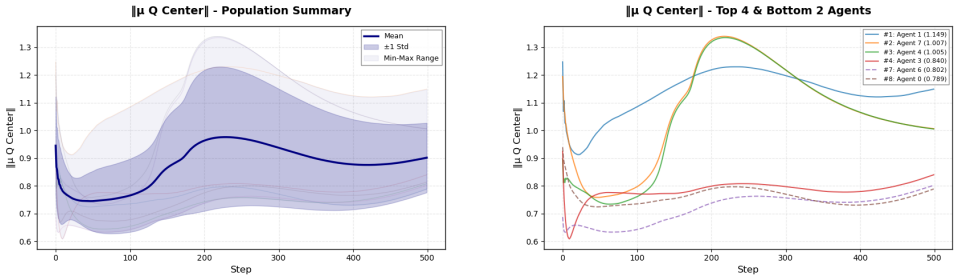

Figure 8: Population-level and per-agent evolution of belief magnitudes $\|\mu_Q^{\text{center}}\|$ during training. (**Left**) Population mean, standard deviation, and range across all agents. (**Right**) Top four top and bottom two agents by $\|\mu_Q(t)\|$.

This level of statistical consistency across 144 independent tests strongly supports our claim that the KL gauge theoretic attention captures a fundamental principle underlying transformer attention rather than fitting superficial patterns.

5.2.4 Symmetry Breaking and Training

Detailed simulations and analytic derivations produced agent belief and frame flow towards a shared rotationally invariant vacuum state $\mu_i(c) \longrightarrow \mu^*$ under variational natural gradient descent in the quasi-static regime. The coordinates of each μ_i in the $2\ell_q + 1 = 19$ dimensional fiber are generally unique but share identical norms proving that they occupy a 18-dimensional

sphere of values. This is precisely the gauge orbit we expect under the symmetric vacuum theory.

When we allow the same agents (under identical conditions), to observe randomly drawn observations from either a Gaussian distribution we find symmetry is broken and unique norms for each agent are produced providing a clear case of spontaneous symmetry breaking. This suggests observations induce specialization in a similar manner to back-propagation gradient descent.

Furthermore, we monitored the agent attentions under evolution. A typical $\beta_i j$ field is given in figure 2 showing two agents overlapping coincidentally in a pile of 8 (radius = 7) over a 2-dimensional 20 by 20 toroidal base-space with 19-dimensional fiber.

By showing that transformers are the Dirac delta limit of a more general framework and by connecting variational free energy, via a generative model yielding agent to agent coupling terms in a gauge equivariant manner, we have shown the existence of a more general and geometric framework for informational systems. [1]

5.2.5 Limitations and Future Directions

Our validation tested the flat-bundle limit where all frames are globally aligned ($\Omega_{ij} = \Omega$ for all i, j) and the connection is trivial. This is appropriate for comparison with standard transformers, which lack explicit gauge structure. However, the full power of the gauge-covariant 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 comparison focused on the Dirac limit where uncertainty covariances collapse to 0. The full gauge-covariant framework incorporates non-trivial covariance matrices $\Sigma_i(c)$ that encode uncertainty about agent beliefs. These would then correspond to "fuzzy" vector embeddings in unique token-frames. Testing this framework against transformers with explicit uncertainty estimation (e.g., Bayesian neural networks, ensemble methods) represents an important future direction.

[1] In a philosophical sense, one may view each gauge frame as an analogue of an agent's conscious orientation—its private coordinate system for representing the world. We employ this notion purely metaphorically to emphasize epistemic perspective, not phenomenological experience or subjective qualia. Nevertheless, it is tempting to speculate on what the pullback of informational quantities along an agent's section to the base manifold might signify in relation to experience itself.

Section 4.9 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, it does so uniformly across all tokens. The gauge-theoretic perspective suggests more nuanced strategies: rather than enforcing strict norm equality, one could learn optimal norm profiles that trade off representational capacity (high norms encode more information) against attention quality (low norms avoid bias). This could be implemented via soft norm constraints or adaptive normalization schedules that vary across layers or heads.

5.2.6 Broader Implications

Our results demonstrate that attention mechanisms are not ad hoc architectural choices but natural consequences of variational free energy minimization on gauge bundles. This unification has conceptual power: it explains why attention works, predicts its limitations (key-norm bias), and suggests principled extensions (gauge structure, uncertainty propagation). The framework places transformers within the broader landscape of probabilistic inference, connecting them to active inference, predictive coding, and variational inference.

By leveraging gauge theory coupled with informational geometry, we introduce geometric biases that constrain model behavior according to symmetry principles. This is analogous to how convolutional neural networks impose translation equivariance or how graph neural networks respect permutation invariance. Our gauge-equivariant attention extends this paradigm to a much richer landscape of symmetry groups (SO(N), SU(N), Lorentz group, etc.), potentially enabling transformers to learn from fewer examples in domains with known physical structure or infer patterns that flat-bundles otherwise would miss.

Most importantly, our work shows that attention emerges as a consequence of agents minimizing local variational free energy under a gauge-covariant frame coupling. The gauge equivariant attention mechanism emerged 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 extend-

ing far beyond artificial neural networks to biological cognition, collective intelligence, physics, linguistics, sociology, and general informational systems.

6 Conclusion

We have shown that attention and transformers are a limiting case of a more general statistical gauge equivariant theory where tokens are modeled as agents with certainty of their beliefs (delta-function limit). The attention dot-product is due to an agent-agent "communication" term in a generalized functional variational energy/action as an application of the free energy principle. The full framework possesses a vacuum state where all agents flow towards an average belief and gauge frame (embedding) mirroring machine learning without training. We have shown that agent observations break this symmetry by 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 amounts to variational gradient descent of the generalized free energy we have derived.

Furthermore, (as we show elsewhere) 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 have shown that under variational gradient descent meta-agents emerge with cross-scale couplings. Time-scale separation occurs with meta-agents fluctuating on time-scales around $10^4 - 10^6$ times slower than the lower scale agents (where we define time in terms of agent belief updating - i.e. the smallest time scale corresponds to 1 bit of belief updating). .

Our framework then suggests a novel approach towards unifying the variational free energy principle with machine learning architectures by extending the free energy principle to include an agent-agent communication term. We anticipate this approach will find application not only in the machine learning and variational inference communities, but also in such disparate fields as linguistics, psychology, sociology, physics, and other informational research thrusts.

References

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

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

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

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

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

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

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

[8] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

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

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

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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.

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

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

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

[21] 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.

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

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

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

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

[26] 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.

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

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

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

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

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

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

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

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

Symbol	Description	Type/Dimension
Fiber Bundle Structure		
$\mathcal{M}$	Base manifold (spatial domain)	Manifold
Q_i	Belief fiber at agent i	$\mathbb{R}^{K_q}$
P_i	Model fiber at agent i	$\mathbb{R}^{K_p}$
G	Gauge group	$\mathrm{SO}(K)$
$\mathfrak{g}$	Lie algebra of G	$\mathfrak{so}(K)$
Gauge Fields and Connections		
Ω_{ij}^q	Connection $Q_j \to Q_i$	$\mathrm{SO}(K_q)$
Ω_{ij}^p	Connection $P_j \to P_i$	$\mathrm{SO}(K_p)$
ϕ_i	Gauge parameter (belief frame)	$\mathfrak{g} \cong \mathbb{R}^3$
$\tilde{\phi}_i$	Gauge parameter (model frame)	$\mathfrak{g} \cong \mathbb{R}^3$
$A_\mu(x)$	Gauge connection field	$\mathfrak{g}$-valued 1-form
Bundle Morphisms		
Φ_i	Morphism $Q_i \to P_i$	$\mathbb{R}^{K_p \times K_q}$
$\tilde{\Phi}_i$	Morphism $P_i \to Q_i$	$\mathbb{R}^{K_q \times K_p}$
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	$\mathbb{R}_+$
α	Self-consistency strength	$\mathbb{R}_+$
D_{KL}	Kullback–Leibler divergence	$\mathbb{R}_+ \cup \{0\}$

Table 2: Principal notation used throughout this paper.