Gauge Equivariant Extension of FEP and Attention

Robert C. Dennis

Abstract

We present a unified gauge-theoretic formulation of attention and message communication in transformer architectures and postulate this as a potential bridge to Friston's free energy principle in neuroscience and cognition.

In our framework, each agent is modeled as a smooth local section of an associated bundle over a base manifold $\mathcal{C}$, carrying an exponential family belief field (specifically we consider multi-variate Gaussians $q_i = \mathcal{N}(\mu_i, \Sigma_i)$ whose mean and covariance define local epistemic states and informational metrics in the fiber).

Inter-agent communication is mediated by a non-abelian gauge connection A with parallel-transport operators Ω_{ij} (which we consider specifically under $SO(3)$, representing orientation changes between local gauge frames).

Attention weights arise naturally as normalized exponentials of information-geometric divergences, $\beta_{ij} \propto \mathrm{softmax}_j\big(-\mathrm{KL}\big(q_i \,\|\, \Omega_{ij} q_j\big)\big)$, such that belief alignment, rather than ad hoc dot products, governs agent-agent interaction. We show this form can be found by a max-entropy optimization.

We demonstrate that in the isotropic, flat-bundle, delta-function limit (where all local frames coincide globally) this expression reduces exactly to the canonical transformer rule $\beta_{ij} \propto \mathrm{softmax}(Q_i K_j^\top)$, where $Q_i = \mu_i^T e^{\phi_i}$ and $K_j = \mu_j^T e^{-\phi_j}$ (where ϕ_i are the gauge frames). Additionally, this decomposition can, in principle, occur via any decomposition we please (SVD, low-rank, etc) such that $AB^T = \Omega_{ij}$ for chosen A and B operators.

Finally, we show that statistical observations correspond to a symmetry breaking term allowing agents/tokens to flow towards unique embedding vectors. In this view an "attention" term $\beta_{ij}\mathrm{KL}\big(q_i \,\|\, \Omega_{ij} q_j\big)$ is

derived under the free energy principle. If this can be made rigorous then our model will realize the unification of the free energy principle and machine learning. Surprisingly, our framework suggests an analogy to the Grand thermodynamic potential $\Upsilon = U - TS + \mu N$ (where the chemical potential analogs to the attention softmax weights and the number of objects analogs to the alignment term between agents). Our construction therefore fuses information geometry, non-abelian gauge theory, and transformer architecture into a single mathematical formalism, opening a path toward curvature-aware, uncertainty-sensitive neural architectures grounded in standard geometric and variational principles.

1 Introduction

Recent advances in both neuroscience and artificial intelligence have independently converged on the idea that intelligent systems must integrate perception, inference, and communication under constraints of uncertainty. Friston's Free Energy Principle (FEP) provides a general variational formulation of inference in cognitive systems, whereas the attention mechanism in modern machine learning architectures defines a powerful (although empirically derived) rule for token prediction. Despite their shared reliance on probabilistic inference and message passing, 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 work as well as they do as well as why architectures can learn untrained features remain obscured in a cloud of mystery. Perhaps the most important feature of these architectures is scaling.

In this report we propose a unified, gauge-equivariant framework that connects these disparate yet similar domains. Our framework is based on geometry whereby each agent is modeled as a smooth local section of an associated bundle with statistical manifold fibers over a 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 further show that in the flat-bundle, isotropic, delta-function-agent limit this general rule reduces exactly to the standard transformer dot-

product attention QK^T, thereby identifying attention as the degenerate case of a broader geometric law of communication predicated upon the FEP. We further 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 multi-variate Gaussian agents.

This suggests that curvature, holonomy, and uncertainty in the latent manifold correspond to structured patterns of communication and emergent organization among agents. Thus, this gauge-equivariant formulation of the FEP opens a novel path toward curvature-aware, uncertainty-sensitive neural architectures and seeks a general unification of cognition, geometry, and machine learning under a single geometry.

2 Bridging the Free Energy Principle and Machine Learning

We begin by defining the mathematical model and then show that attention is a limiting case of this more general geometric framework. This suggests that a bridge exists between Friston's free energy principle and the methods of deep learning and machine learning architectures. If such a bridge does exist then machine learning may be able to leverage gauge theoretic methods and geometry to advance deep learning architectures and similarly advance new insights into cognition and neuroscience.

2.1 The Mathematical Model

We begin by introducing a general formal model which a priori bears no resemblance to machine learning architectures. Our model is geometrically and mathematically rich allowing hierarchical emergence of meta-agents and evolution and non-trivial holonomy of transport.

We model each agent as a local section of an associated bundle to a principal fiber bundle whose base space $\mathcal{C}$ which we define clearly and rigorously below.

Beliefs and models are encoded in the associated bundle constructed from a structure group G acting on the fiber $\mathcal{B}_\Pi$: a statistical manifold. These

structures allow us to formalize both intra-agent inference and inter-agent communication using gauge-theoretic transport.

2.2 Defining the Model

Let $\pi : \mathcal{N} \to \mathcal{C}$ be a smooth principal G-bundle where $\mathcal{C}$ is a smooth manifold, G is a Lie group acting freely and transitively on the right on $\mathcal{N}$.

The projection satisfies $\pi(n \cdot g) = \pi(n)$ for all $g \in G, n \in \mathcal{N}$.

Let $\rho : G \to \mathrm{Aut}(\mathcal{B}_q)$ be a representation of the Lie group G on a smooth statistical manifold $\mathcal{B}_q$. Depending on context, $\mathcal{B}_q$ may be modeled as a K-dimensional probability simplex Δ^K (e.g., for categorical distributions) or as a statistical manifold equipped with a suitable information geometry. In our investigations we specifically choose a Gaussian manifold with gauge group $SO(3)$ for simplicity and clarity.

Importantly, despite the statistical manifold not being a vector space, the relevant geometric structures — such as divergence measures, metrics, and connections — remain well-defined on these fibers even when they lack such linear structure (see e.g. Amari's information geometry for examples of dually flat but non-linear statistical manifolds).

Next, given our principal bundle $\mathcal{N}$ we define the associated bundle as:

$$\mathcal{E}_q := \mathcal{N} \times_\rho \mathcal{B}_q = (\mathcal{N} \times \mathcal{B}_q)/\sim,$$

where

$$(n \cdot g, b) \sim (n, \rho(g)b).$$

This yields a fiber bundle $\pi_{\mathcal{E}_u} : \mathcal{E}_u \to \mathcal{C}$ with fiber $\mathcal{B}_q$. Given the associated bundle we can define smooth sections over the bundle.

Definition:

An agent is an open local section of the associated bundle $\mathcal{E}_q$ over $\mathcal{C}$. Specifically,

$$\mathcal{A}^i = \sigma_q^i(c) = q_i(c)$$

$$\sigma_q^i : \mathcal{U}_i \subset \mathcal{C} \to \mathcal{B}_q,$$

where $\mathcal{U}_i$ are open local subsets of $\mathcal{C}$ and i is the agent label. In the case that $\mathcal{B}_q$ is a statistical Gaussian manifold an agent will generally comprise

a field of sufficient statistics or, transparently, field $\mu(c), \Sigma(c)$ over the base manifold.

Definition:

A multi-agent $\mathcal{M}$ over $\mathcal{C}$ is a tuple of agents (where $\mathcal{I}$ is an index set)

$$\mathcal{M} = \{A^i = \sigma_q^i(c)\}_{i \in \mathcal{I}}.$$

This constitutes a collection of open smooth sections (fields). Agents will generally overlap in the open intersection $\mathcal{U}_i \bigcap \mathcal{U}_j$.

Definition:

A meta-agent is a multi-agent whose agents share identical section values. For example, $\mu_i(c) = \mu_j(c)$ and $\Sigma_i(c) = \Sigma_j(c)$ (or succintly $q_i(c) = q_j(c)$) for some overlap point c.

Furthermore, we say that a set of agents is "epistemically dead" if they identically share both beliefs and models. Note, importantly that even if the agents composing a meta-agent are epistemically dead this does NOT imply the meta-agent itself is epistemically dead. In fact, this is a feature of our framework - epistemically dead agents can generally be integrated out.

Next, in the standard way, via horizontal lifting from $\mathcal{N}$ to $\mathcal{E}_q$ we generate a variety of morphisms and induced connections across scales (i, j). In particular parallel transport and cross-scale transport operators:

1. $\Omega^s : \Gamma^s(\mathcal{B}_q) \to \Gamma^s(\mathcal{B}_q)$
2. $\Lambda^s_{s'} : \Gamma^s(\mathcal{B}_q) \to \Gamma^{s'}(\mathcal{B}_q)$

where $\Gamma(\mathcal{B}_q)$ denotes the space of smooth sections of $\mathcal{B}_q$ over the relevant open subset of $\mathcal{C}$. Additionally Λ^i_j represents the parallel transport operator between agents at scale s to s'. In our current considerations we will say nothing further about cross-scale connections and meta-agents.

2.3 Types of Parallel Transport in Epistemic Geometry

In our framework, parallel transport arises in several distinct but interconnected settings, depending on whether the transport occurs along the base manifold $\mathcal{C}$ or across agent frames. We distinguish the following cases:

2.3.1 Horizontal Transport Along the Base Manifold $\mathcal{C}$

Let $\pi : \mathcal{N} \to \mathcal{C}$ be a principal G-bundle, and let $\mathcal{E} = \mathcal{N} \times_\rho \mathcal{B}$ be an associated bundle with fiber $\mathcal{B}$. Given a path $\gamma : [0,1] \to \mathcal{C}$, a connection 1-form $A \in \Omega^1(\mathcal{C}, \mathfrak{g})$ defines a notion of parallel transport along γ via the path-ordered exponential:

$$T_\gamma = \mathcal{P} \exp\left(- \int_\gamma A_\mu(c)\, dc^\mu \right) \in G.$$

This operator maps fiber elements between different base points:

$$b(c_0) \in \mathcal{B}_{c_0} \quad \mapsto \quad b(c_1) = \rho(T_\gamma) \cdot b(c_0) \in \mathcal{B}_{c_1}.$$

This is the canonical notion of parallel transport along the base, lifted via the connection to fibers. We will need this when we study tokens as the delta-limiting case of agents.

2.3.2 2. Vertical Transport Within a Single Fiber $\mathcal{B}_c$

At a fixed point $c \in \mathcal{C}$, the fiber $\mathcal{B}_q(c)$ is a manifold in its own right. In our setting, this fiber typically represents a statistical manifold (e.g., probability simplex Δ^K, or an exponential family), equipped with an intrinsic Riemannian or dual-affine geometry.

Parallel transport within $\mathcal{B}_q(c)$ may occur along a curve $\eta(\tau) \subset \mathcal{B}_q(c)$ with tangent vector $\dot{\eta}(\tau)$ governed by a connection ∇ intrinsic to $\mathcal{B}$:

$$\nabla_{\dot{\eta}} V = 0, \quad \text{for parallel vector field } V(\tau).$$

In the information geometry setting (Amari), $\mathcal{B}_q(c)$ may carry a pair of dual connections $(\nabla^{(e)}, \nabla^{(m)})$ associated with exponential and mixture families. Curvature in this fiber space is defined via the Riemann tensor $R^{\mathcal{B}}$:

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

Thus, even at a single point $c \in \mathcal{C}$, the fiber $\mathcal{B}_c$ may be a curved internal space with nontrivial geodesics and holonomies. These paths represent purely epistemic transformations of beliefs or models that do not involve movement along $\mathcal{C}$.

2.3.3 3. Intra-Agent Spatial Transport (Within Fiber Bundle)

To compare beliefs $q(c_1)$ and $q(c_2)$ held by the same agent over two nearby points $c_1, c_2 \in \mathcal{U}_i \subset \mathcal{C}$, we apply the agent's gauge frame $\phi_i(c)$ to define a connection $A_\mu^{(i)} = \partial_\mu \phi_i(c)$ and use:

$$T_{c_1 \to c_2}^{(i)} := \mathcal{P} \exp\left(- \int_{c_1}^{c_2} A_\mu^{(i)} \, dc^\mu \right). \tag{1}$$

This defines gauge-covariant parallel transport between fibers over space, for a fixed agent.

2.3.4 4. Inter-Agent Frame Transport (At a Shared Point)

When two agents $\mathcal{A}_i$ and $\mathcal{A}_j$ overlap at a point $c \in \mathcal{U}_i \cap \mathcal{U}_j$, each has its own gauge frame $\phi_i(c), \phi_j(c) \in \mathfrak{g}$. The inter-agent gauge transformation is given by:

$$\Omega_{ij}(c) := \exp(\phi_i(c)) \cdot \exp(-\phi_j(c)) \in G,$$

and transports beliefs or models from agent j's frame to agent i's:

$$q_j(c) \mapsto q_i^{(j)}(c) := \rho(\Omega_{ij}(c)) \cdot q_j(c).$$

This is essential for defining KL alignment between agents:

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

2.3.5 5. Composite Transport and Holonomy

More generally, a belief may be transported along a path $\gamma = [(i_1, c_1) \to \cdots \to (i_n, c_n)]$ consisting of both spatial transport within agents and inter-agent frame shifts. The total parallel transport operator is:

$$P_\gamma = T_{c_1 \leftarrow c_n}^{(i_n)} \cdot \Omega_{i_n i_{n-1}} \cdots \Omega_{i_2 i_1} \cdot T_{c_2 \leftarrow c_1}^{(i_1)},$$

This defines the holonomy of belief transport:

$$\mathrm{Hol}_\gamma(q) := P_\gamma \cdot q.$$

If $\mathrm{Hol}_\gamma(q) \neq \mathrm{Id}$, the loop encloses nontrivial epistemic curvature.

2.3.6 Summary

Transport Type	Domain	Operator	Purpose
Horizontal (base)	$c_1 \to c_2 \in \mathcal{C}$	$T_\gamma = \mathcal{P}\exp(-\int A)$	Compare beliefs across space
Vertical (fiber)	$\eta(\tau) \subset \mathcal{B}_c$	∇ or $\rho(\exp(\phi))$	Transform within belief space
Intra-agent	$q_i(c_1) \to q_i(c_2)$	$T^{(i)}$ from $A_\mu^{(i)}$	Spatial update in same agent
Inter-agent	$q_j(c) \mapsto q_i^{(j)}(c)$	$\Omega_{ij}(c)$	Frame alignment across agents
Composite	$\gamma : (i_1, c_1) \to \cdots$	P_γ	Meta-agent transport and holonomy

Table 1: Types of parallel transport in epistemic gauge geometry.

3 Derivation of Variational Energy

3.1 Obtaining $D_{KL}(q_i|\Omega_{ij}q_j)$ from FEP

The generalized $D_{KL}(q_i|\Omega_{ij}q_j)$ term isn't necessarily a new feature in the free energy principle. What is new is promoting probabilities into an associated bundle. We can obtain the $D_{KL}(q_i|\Omega_{ij}q_j)$ term directly from the FEP as follows:

For a single latent variable c and observation o we have the standard variational free energy as

$$\mathcal{F}[q] = D_{KL}(q(c)|p(c)) - \mathbb{E}_q[\log p(o|c)]$$

In a multi-variable case with latents $c_1, ..., c_N$ the exact free energy under an approximate posterior $q(c_1, ..., c_N)$ is given by

$$\mathcal{F}[q] = \mathbb{E}_q[\log q(c_1, ..., c_N)] - \mathbb{E}_q[\log p(c_1, ..., c_N; o_1, ..., o_N)]$$

Now, if we invoke the mean field assumption on agent's beliefs we have

$$q(c_1, ..., c_N) = \prod_i q_i(c_i)$$

Next, we assume the world does not treat agents as independent - i.e. the generative model encodes relationships between agents' latent states.

Therefore we write

$$p(c_1, ..., c_N) \propto \prod_i p_i(c_i) \prod_{ij} \psi_{ij}(c_i, c_j)$$

where $p_i(c_i)$ is the local prior for agent i's latent cause. $\psi_{ij}(c_i, c_j)$ meanwhile is a pairwise compatibility describing how consistent the pairs' latent states are expected to be.

Next, we choose

$$\psi_{ij}(c_i, c_j) \propto e^{-\lambda_{ij} d(c_i, \Omega_{ij} c_j)}$$

where $d(c_i, \Omega_{ij} c_j)$ is a distance-like function between agent i and agent i's interpretation of c_j through the transport operator Ω_{ij}. λ_{ij} is a coupling strength (precision) to be determined later.

Next, we have

$$p(c_1, ..., c_N; o_1, ..., o_N) = p(c_1, ..., c_N) \prod_i p_i(o_i|c_i)$$

such that each agent has its own local observation o_i and additionally, that, given c_i this observation o_i is independent of the other agents.

Therefore, combining terms we have

$$p(c, o) \propto \prod p_i(c_i) \prod_{ij} \psi_{ij}(c_i, c_j) \prod_i p_i(o_i|c_i)$$

Note: all constants will ultimately be absorbed into the partition function Z.

We now write the VFE as

$$\mathcal{F}[q] = \mathbb{E}_q[\log q(c)] - \mathbb{E}_q[\log p(c, o)]$$

and expand the terms.

First, since $q(c) = \prod_i q_i(c_i)$

$$\mathbb{E}_q[\log q(c)] = \sum_i \mathbb{E}_{q_i}[\log q_i(c_i)]$$

Next, from above we expand

$$\log p(c, o) = \sum_i \log p_i(c_i) + \sum_{i,j} \log \psi_{ij}(c_i, c_j) + \sum_i \log p_i(o_i|c_i) - \log Z$$

Then we take the expectation over $q(c) = \prod_k q_k(c_k)$:

$$\mathbb{E}_{q(c)}[\log p(c,o)] = \sum_i \mathbb{E}_{q_i(c_i)} \log p_i(c_i) + \sum_{i,j} \mathbb{E}_{q_i(c_i),q_j(c_j)} \log \psi_{ij}(c_i,c_j) + \sum_i \mathbb{E}_{q_i(c_i)} \log p_i(o_i|c_i) - \mathbb{E}_{q_i(c_i)} \log$$

Plugging these into our VFE we find

$$\mathcal{F}[q] = \sum_i D_{KL}(q_i(c_i)|p_i(c_i)) - \sum_i \mathbb{E}_{q_i}[\log p_i(o_i|c_i)] - \sum_{i,j} \mathbb{E}_{q_i q_j} \log \psi_{ij}(c_i,c_j)$$

We shall focus our attention on the cross-term

$$\sum_{i,j} \mathbb{E}_{q_i q_j} \log \psi_{ij}(c_i,c_j)$$

We previously defined

$$\psi_{ij}(c_i,c_j) \propto e^{-\lambda_{ij} d(c_i,\Omega_{ij} c_j)}$$

Therefore our term becomes

$$\sum_{i,j} \lambda_{ij} \mathbb{E}_{q_i q_j} d(c_i, \Omega c_j) + \text{const}$$

Again, the constant can be absorbed by the partition function Z so we shall not consider it further.

Now, in general, any function $d(c_i,c_j)$ satisfying

$$d \geq 0$$

$$d(c,c) = 0$$

$$d \leq \infty \quad \text{and integrable}$$

will suit our compatibility. Generally we can choose any f-divergence here. f-divergences (and especially the KL-divergence)) are uniquely suited to produce exponential family compatibilities.

A natural choice for multi-variate Gaussians is the Mahalanobis distance

$$d(c_i, \Omega c_j) = \sqrt{(\Omega_{ij} c_j - c_i)^T \Sigma_i^{-1} (\Omega_{ij} c_j - c_i)}$$

The expectation over q_i, q_j then leads to $D_{KL}(q_i(c_i)|\Omega_{ij} q_j(c_j))$ as the second order expansion of the log-density ration - namely the KL term. Therefore, we find

$$\mathcal{F}[q] = \sum_i D_{KL}(q_i(c_i)|p_i(c_i)) - \sum_i \mathbb{E}_{q_i}[\log p_i(o_i|c_i)] + \sum_{i,j} \lambda_{ij} D_{KL}(q_i(c_i)|\Omega_{ij} q_j(c_j))$$

which is what we wished to show.

The attention-coupling term in our generalized energy functional is just the mean-field variational free energy of a multi-agent generative prior with gauge-aligned pairwise factors.

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.

3.2 Obtaining β_{ij}

We shall treat β_{ij} as a set of weights agent i assigns to all other agents j. For fixed i β_{ij} should form a probability distribution. Namely,

$$\sum_j \beta_{ij} = 1$$

We derive the form of β_{ij} by letting agent i choose β_{ij} to minimize its own expected coupling cost to other agents.

First, we define the mismatch cost between agents as

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

Next, we define the "expected disagreement" as

$$\mathcal{L}_i(\beta_i) = \sum_j \beta_{ij} C_{ij}$$

Upon minimization this, being linear in β_{ij}, would put all weight in the best neighbor j as $j^* = \arg\min_j C_{ij}$. This is a hard attention corresponding

to "one-hot" encoding. To obtain a soft attention we introduce an "uncertainty cost" on β_i as

$$\mathcal{L}_i^{ent}(\beta_i) = \kappa \sum_j \beta_{ij} \log \beta_{ij}$$

Incorporating this term penalizes low entropy distributions. Our total objective then becomes

$$\mathcal{J}_i(\beta_i) = \sum_j \beta_{ij} C_{ij} + \kappa \sum_j \beta_{ij} \log \beta_{ij}$$

This now prevents agents from collapsing β_{ij} to a single term: accuracy versus complexity.

Next we optimize β_{ij} via Lagrange multipliers in the standard manner. Define the Lagrangian

$$\mathscr{L}_i(\beta_i, \xi_i) = \sum_j \beta_{ij} C_{ij} + \kappa \sum_j \beta_{ij} \log \beta_{ij} + \xi_i (\sum_j \beta_{ij} - 1)$$

then solve

$$\frac{\partial \mathscr{L}}{\partial \beta_{ij}} = 0$$

and

$$\frac{\partial \mathscr{L}}{\partial \xi_i} = 0$$

We find

$$\frac{\partial \mathscr{L}}{\partial \beta_{ij}} = C_{ij} + \kappa(\log \beta_{ij} + 1) + \xi_i = 0$$

Solving for β_{ij} we find

$$\beta_{ij} = e^{-1} e^{-\frac{\xi_i}{\kappa}} e^{-\frac{C_{ij}}{\kappa}}$$

The first two terms do not depend on j so we have

$$\beta_{ij} = K_i e^{-\frac{C_{ij}}{\kappa}}$$

Now we impose normalization $\sum_j \beta_{ij} = 1$ to find

$$K_i = \frac{1}{\sum_k e^{-\frac{C_{ik}}{\kappa}}}$$

and thus we arrive at our final form

$$\beta_{ij} = \frac{e^{-\frac{C_{ij}}{\kappa}}}{\sum_k e^{-\frac{C_{ik}}{\kappa}}}$$

$$\beta_{ij} = \frac{e^{-\frac{D_{KL}(q_i|\Omega_{ij}q_j)}{\kappa}}}{\sum_k e^{-\frac{D_{KL}(q_i|\Omega_{ik}q_k)}{\kappa}}}$$

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

4 Transformer Attention as a Special Case of Gauge-Covariant Probabilistic Message Communication

In this section we show that standard transformer self-attention is recovered as a δ-function limiting case of our gauge-covariant, uncertainty-aware framework where, in this view, tokens are generally "fuzzy" embeddings (that is, a field vector μ and Σ ellipse under gauge frame ϕ). 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 non-probabilistic vector.

In particular, we apply the following to our framework:

(i) a discrete base space (discrete tokens instead of extended open sections - δ-function localization),

(ii) a flat bundle / trivial connection (shared globally defined frame, no curvature),

and

(iii) isotropic uncertainty (identical spherical covariance for each agent which we keep for the derivation and then take the limit as tokens become delta-function distributions in the fiber).

Under these assumptions, our KL-based attention law, β_{ij} reduces exactly to the standard $(QK^\top)V$ form of transformer attention thereby illuminating the geometric source of the ad-hoc dot product similarity. We show this below.

4.1 Setup: Agents as Gaussian Beliefs in Local Frames

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

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

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

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

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

which maps representations expressed in agent j's local frame into agent i's local frame.

In our general framework agents share beliefs via KL coupling defined by the (negative) Kullback–Leibler divergence between i's belief and j's belief transported into i's frame:

That is, in our variational energy function (to be defined below - see appendix) we have the term

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

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.

4.2 Assumptions for the Transformer limit

We now impose the three simplifying assumptions.

4.2.1 Discrete base / tokens as agents.

We consider a base space $\mathcal{C}$ to be a finite index set of positions $\{1, \ldots, N\}$ (e.g. token positions in a sequence).

Each agent i is now just "the token at position i". There is no spatial overlap integral; all quantities are evaluated at a single site thereby considerably simplify our variational energies.

4.2.2 Flat Bundle / Trivial Connection.

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

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

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

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

4.2.3 Isotropic and identical uncertainty.

We assume all agents have the same spherical covariance:

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

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

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

To make contact with the simplest transformer form, we can absorb $\Omega\Omega^\top$ into a learned rescaling of Ω , into the variance σ^2, or under an $SO(3)$ gauge group this term becomes the identity. This is equivalent to the standard freedom in attention to choose arbitrary learned projection matrices and overall temperature scaling.

Therefore,

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

Note: we are considering $SO(3)$ for simplicity due to multi-variate Gaussians and KL-divergence being invariant under rotation (this is straightforward to show). Interestingly, in the general $SO(3)$ case global token attention cannnot exist due to $SO(3)$ presenting an obstruction, $\pi_1(SO(3)) = \mathbb{Z}_2$. To define a global attention in this case requires lifting to $SU(2)$ - otherwise we can only assign localized agent attentions!

4.2.4 Emergence of the Dot Product Attention

For two Gaussians with identical isotropic covariance $\sigma^2\mathbb{1}$, the Kullback–Leibler 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 that fixed i) will pull 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 in 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

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

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

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

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

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

The remaining term $-\frac{1}{2\sigma^2}\|\Omega\mu_j\|^2$ depends only on j. This acts as a key-dependent bias. Such additive, key-specific biases do not change the functional form of attention: they simply shift each column of the attention score matrix before the row-wise softmax. Standard Transformers typically omit this bias, but including or omitting it does not alter the bilinear dot-product structure that defines attention.

Hence, modulo a (learnable) per-key bias, we have shown:

$$\tilde{s}_{ij} \;\approx\; Q_i K_j^\top.$$

Consequently,

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

which is the standard Transformer attention weighting rule.

4.3 The Aggregation Becomes the V-Projection Sum

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
$\alpha_{ij} = \text{softmax}_j(Q_i K_j^\top)$ 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.

4.4 Conclusion

We have shown the following:

1. Under our assumptions , the score $s_{ij} = -\text{KL}(q_i \| \Omega_{ij} q_j)$ 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 logits $QK^\top$.

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.

3. Therefore, in the limit of (i) discrete sites interpreted as tokens, (ii) a flat global frame with trivial parallel transport, and (iii) isotropic identical uncertainty so that all covariances collapse to $\sigma^2 \mathbb{1}$, our gauge-covariant, uncertainty-aware message passing law (or $q_i = \delta$ if you like)

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

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

becomes exactly the canonical Transformer attention

$$\alpha_{ij} = \text{softmax}_j(Q_i K_j^\top),$$

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

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, discrete-base space ($1D$ lattice) of our gauge-covariant KL-attention rule.

4.5 Generalized Variational Energy

Following Friston[ˆ1] we define a variational free energy as

$$\mathcal{F}[q] = D_{KL}[\ q(c) \mid p(c)\ ] - \mathbb{E}_{q(c)}[\log p(o|c)].$$

where $q(c)$ represents an agents beliefs over c, $p(c)$ is the agent's prior, and $p(o|c)$ consitutes the agent's likelihood mapping.

For simplicity, in this study, we will not consider "active" inference - we shall only concern ourselves with how agents passively condense into meta-agents and quasi-meta agents and its connection to machine learning architectures.

Given our generalized gauge theory model we extend this principle to agents who may share information/beliefs/models via parallel transport. We define a generalized variational energy for an agent A as:

$$\mathscr{V}_A = D_{KL}[\ q_A(c) \mid\mid \Phi p_A(c)\ ] - \mathbb{E}_{q_A}[\log p_A(o|c)] + \sum_i \mathscr{V}_{\Lambda_i},$$

where $\mathscr{V}_{\Lambda_i}$ represents possible interactions between other agents and (quasi)-meta-agents mediated by the various gauge connections above.

In our current considerations we will only focus on the induced connection Ω such that agents can compare beliefs between different sections of $\mathcal{B}_q$ where, again,

$$\Omega : \Gamma(\mathcal{B}_q) \to \Gamma(\mathcal{B}_q).$$

The induced connection provides a parallel transport of beliefs which in turn defines the coordination structure used in our generalized variational energy term. This construction is well defined even when the fibers are not vector spaces so long as they carry a smooth G-action[ˆ34].

Therefore, we have as a generalization

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

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

$$-\sum_i \mathbb{E}_{q_i(c)}[\log p_i(o_i|c)].$$

where α, β represents general couplings (we generally take $\alpha = 1$. We note in passing that this expression bears remarkable similarities to the Grand potential $\Upsilon = U - TS + \mu N$ in standard thermodynamics (where μN is analogous to the Ω term)!

The β_{ij} coupling warrants more consideration. We tentatively assume that agents will favor alignment with other agents who share their beliefs. A natural candidate for β_{ij} is then a Boltzmann distribution given by

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

where $m_{ij}(x)$ is the agents support (or mask).

Note: for full generality we could construct analogous terms for agent-agent models.

4.6 Gauge Transport via Lie Algebra Frames

The inter-agent transport operator Ω_{ij} serve to compare beliefs (and models if desired) between agents i and j. These operators are defined pointwise over the agent's support in $\mathcal{C}$, and are constructed from local gauge frames associated with each agent.

Let $\phi_i(c) \in \mathfrak{g}$ be a smooth field over $\mathcal{U}_i \subset \mathcal{C}$ valued in the Lie algebra $\mathfrak{g} = \mathrm{Lie}(G)$, representing the gauge frame of agent i. Then, for each pair of agents (i, j), we define the induced transport operator as:

$$\Omega_{ij}(c) := \exp\left(\phi_i(c)\right) \cdot \exp\left(-\phi_j(c)\right),$$

where the exponential map $\exp : \mathfrak{g} \to G$ maps Lie algebra elements to the Lie group G, and the product is taken in the group.

This construction defines a group-valued map $\Omega_{ij}(c) \in G$, which acts on the belief fiber $\mathcal{B}_q$ via the representation ρ. That is, for any section $q_j(c) \in \Gamma(\mathcal{B}_q)$, the transported belief is:

$$\Omega_{ij}(c) \cdot q_j(c) := \rho\left(\Omega_{ij}(c)\right) \cdot q_j(c).$$

These induced transport operators implement a form of gauge-covariant parallel transport between agents. The field $\phi_i(c)$ serves as the agent's gauge frame, and differences in these frames determine the epistemic misalignment between agents.

By construction, $\Omega_{ii} = \mathrm{Id}$, and in general $\Omega_{ij} \neq \Omega_{ji}^{-1}$ unless the gauge fields are flat. This asymmetry encodes epistemic curvature, which we will explore in subsequent sections.

In our framework, each agent maintains a fixed gauge frame $\phi_i(c)$, which defines their local epistemic perspective. This frame serves as the reference against which other agents' beliefs are compared. The term $D_{\mathrm{KL}}[\, q_i(c) \,\|\, \Omega_{ij}(c) \cdot q_j(c)\,]$ then quantifies the epistemic misalignment between agent i and agent j, measured in agent i's own frame. Operationally, this can be interpreted as a communicative act: agent i attempts to transform agent j's beliefs into its own representational basis and evaluates their coherence.

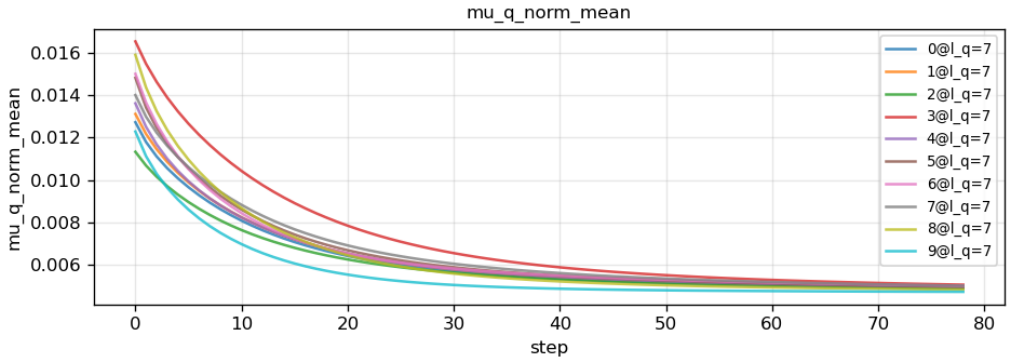
Intuitively, the gauge frame $\phi_i(c)$ can be interpreted as a geometric encoding of the agent's cognitive identity — its internally consistent perspective on the latent manifold $\mathcal{C}$. In this sense, the gauge frame acts as a formal proxy for an agent's "selfhood" or how they interpret their world, organizing all incoming beliefs and observations relative to a quasi-static internal model which gets compared with other agents.

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

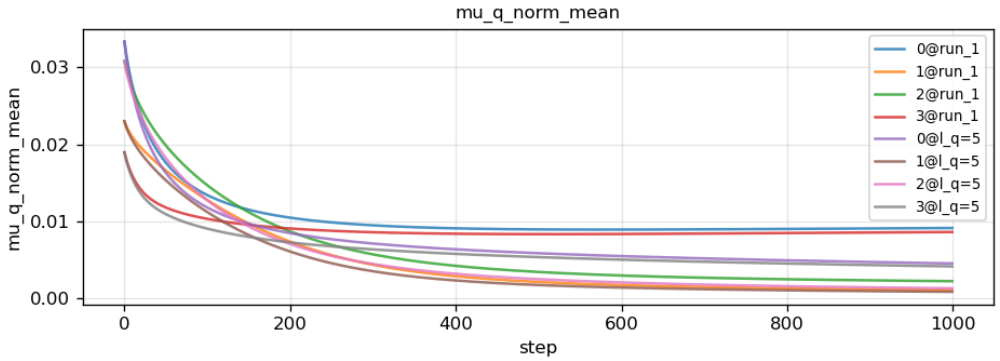
$$\mathscr{V}[q(c)] = \alpha \sum_i D_{\mathrm{KL}}[(q_i(c)|p_i(c))]$$

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

Detailed simulations and analytic derivations produce agent belief and frame flow towards a shared constant vacuum state $\mu_i(c), \phi_i(c) \longrightarrow \mu^*, \phi^*$ under variational gradient descent (all agents share the same models) in the quasi-static regime. All agents share beliefs and gauge frames becoming epistemically dead (see below - $l_q = 7$ is the spin-7 representation of $SO(3)$).



In the case below observations induce specialization.



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. Notice the following:

1. Assume a delta-posterior. This is what is typically considered in machine learning training data
2. then $-\mathbb{E}_{q_i(c)}[\log p(o_i|c)] \to -\log p(o_i|c)$ the standard negative log-likelihood term
3. but if we take observations to be Gaussian then this simplifies to

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

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

which is the standard machine learning loss function!

With only attention (which we have shown is due to the $\sum_j \beta_{ij} D_{\text{KL}}[q_i(c)|\Omega_{ij} q_j(c)]$ term and no learning from data all token embeddings will flow to the same embedding - the average of the tokens) and with structured observations will flow to unique embeddings. The mapping (in the case of Gaussians and $SO(3)$ is then complete. This can also be shown easily for categorical distributions).

In this view we hypothesize that our general framework IS a bridge between machine learning and cognition/neuroscience. Attention is possibly all you need to map from machine learning to Friston's free energy principle and a statistical gauge theory (with statistical fibers) is (possibly) all you need to get there.

One major question remains: What generative model, in the Friston FEP view, will produce this attention term - $\sum_j \beta_{ij} D_{\text{KL}}[q_i(c)|\Omega_{ij} q_j(c)]$? If we can find such a generative model then we will have fully mapped the FEP to machine learning.

5 Conclusion

We have shown that attention and transformers are a limiting case of a more general statistical gauge 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. 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 postulate then that agent observations break this symmetry by flowing to unique vector (μ) embeddings (ϕ) and acts as a machine learning loss function.

Furthermore, 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 an enticing path towards unifying the variational free energy principle with machine learning architectures by extending the free energy principle to include an agent-agent communication term - attention is all you need.

6 Appendix

6.0.1 A Tantalizing Aside

This interpretation, pregnant with possibilities, opens the door to a rich epistemic ontology: communication becomes a gauge-theoretic alignment of perspectives, and disagreement (or confusion) corresponds to curvature in some "semantic" field. Agents with similar gauge frames can align beliefs easily, while those with strongly divergent frames experience epistemic dissonance.

Additionally, for future exploration, human language itself may possibly be considered as a gauge theory. Each speaker maintains a local linguistic frame — a grammar, lexicon, and semantic structure — and successful communication requires parallel transport of meaning across these frames. Misunderstandings arise when the linguistic gauge transformations between individuals are ill-defined or insufficient to fully align perspectives. In this view, syntax encodes local structural constraints, semantics provides the fiber content, and discourse becomes a path-dependent operation on shared meaning spaces. The geometric framework developed here may thus offer a natural foundation for modeling language as a structured system of epistemic transport.

Interestingly and tantalizingly this framework might allow one to study the pullbacks of beliefs/models from the fiber to the base manifold. The fiber is manifestly informational. This suggests a wild interpretation that agent qualia might be modeled as the pull back of informational geometries to the base manifold. E.g. agents at a point c in $\mathcal{C}$ pullback their beliefs/models via their personal sections - they "experience" different phenomena over the same underlying base space. This, however, is fantastically speculative but

it may allow new pursuits in Wheeler's "It from Bit" ontology!

As another example, constructivist theories of physics might have a firm mathematical model with which to explore. For example, if we consider this base manifold as a Kantian noumenon then we might speculate that physics itself is a generative model of cognition that humans have evolutionarily flowed towards sharing. In this view gauge invariance in physics is necessitated by the gauge frame alignment of human agents! There is much more to say.

6.1 Generalized Variational Energy

$$
\begin{aligned}
S = & \underbrace{\sum_i \int_{\mathcal{C}} \alpha_i \, \mathrm{KL}\big(q_i(x) \,\|\, p_i(x)\big) \, \mathrm{d}x}_{\text{(1) Self term: belief} \to \text{prior alignment}} \\
& - \underbrace{\sum_i \int_{\mathcal{C}} \gamma_i \, \mathbb{E}_{q_i(x)}\big[\log p_i(o_i \mid x)\big] \, \mathrm{d}x}_{\text{(2) Observation (likelihood) term: prediction vs. sensory data}} \\
& + \underbrace{\sum_i \sum_j \int_{\mathcal{C}} \beta_{ij}(x) \, \mathrm{KL}\big(q_i(x) \,\big\|\, \Omega_{ij}^{(q)} q_j(x)\big) \, \mathrm{d}x}_{\text{(3) Alignment across belief fibers (attention coupling)}} \\
& + \underbrace{\lambda_A \int_{\mathcal{C}} \tfrac{1}{2} \sum_{a<b} \big\|F_{ab}(x)\big\|^2 \, \mathrm{d}x}_{\text{(4) Gauge curvature energy: global field consistency)}} \\
& + \underbrace{\lambda_\phi \int_{\mathcal{C}} \tfrac{1}{2} \big|D\phi(x)\big|^2 \, \mathrm{d}x}_{\text{(5) Optional gauge-fixing or } \varphi\text{-smoothness term)}} \\
& + (\text{other regularizers: Fisher metric, mass, or higher-order couplings}).
\end{aligned}
$$

References

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

[2] Friston, K., Parr, T., & de Vries, B. (2017). *The graphical brain: belief propagation and active inference.* **Network Neuroscience**, 1(4), 381–414.

[3] Parr, T., Pezzulo, G., & Friston, K. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior.* MIT Press.

[4] Ramstead, M. J. D., Kirchhoff, M. D., & Friston, K. (2019). *A tale of two densities: active inference is enactive inference.* **Adaptive Behavior**, 27(6), 369–385.

[5] Friston, K., Sajid, N., et al. (2021). *Deep temporal models and active inference.* **Neuroscience & Biobehavioral Reviews**, 128, 279–295.

[6] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). *Attention is All You Need.* In *Advances in Neural Information Processing Systems (NeurIPS).*

[7] Bahdanau, D., Cho, K., & Bengio, Y. (2015). *Neural Machine Translation by Jointly Learning to Align and Translate.* In *International Conference on Learning Representations (ICLR).*

[8] Dosovitskiy, A., et al. (2021). *An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale.* In *International Conference on Learning Representations (ICLR).*

[9] Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). *What Does BERT Look at? An Analysis of BERT's Attention.* In *BlackBoxNLP Workshop.*

[10] Amari, S. (2016). *Information Geometry and Its Applications.* Springer.

[11] Amari, S. (1985). *Differential-Geometrical Methods in Statistics.* Springer.

[12] Nakahara, M. (2003). *Geometry, Topology and Physics* (2nd ed.). CRC Press.

[13] Frankel, T. (2012). *The Geometry of Physics: An Introduction* (3rd ed.). Cambridge University Press.

[14] Yang, C. N., & Mills, R. L. (1954). *Conservation of Isotopic Spin and Isotopic Gauge Invariance.* **Physical Review**, 96(1), 191–195.

[15] Clark, A. (2013). *Whatever next? Predictive brains, situated agents, and the future of cognitive science.* **Behavioral and Brain Sciences**, 36(3), 181–204.

[16] Fuchs, C. A., & Schack, R. (2013). *Quantum-Bayesian coherence.* **Reviews of Modern Physics**, 85(4), 1693–1715.

[17] Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). *Geometric deep learning: Grids, groups, graphs, geodesics, and gauges.* **Nature**, 590, 197–205.