Spontaneous Symmetry Breaking of Reference Frames as a Computational Cost Minimization Strategy

Sidong Liu, PhD
iBioStratix Ltd
sidongliu@hotmail.com

February 2026

Abstract

We investigate the computational constraints on persistent open quantum systems that carry non-Markovian memory (Paper I [1]). Paper I established that memory is a thermodynamic necessity for survival beyond the Markovian ceiling, but revealed a secondary crisis: the *Memory Catastrophe*, in which the Landauer cost of maintaining unbounded history exceeds any finite free-energy budget.

We prove a **Computational Ceiling**: any agent that processes its memory *symmetrically*—treating all components of its internal Clifford algebra $Cl(V, q)$ as equally relevant—reaches computational paralysis at a finite critical time t_{par}.

We then show that the resolution requires **spontaneous symmetry breaking** of the agent's internal reference frame: the selection of a privileged basis (a gauge fixing of the automorphism group $G = \text{Aut}(Cl(V, q))$) that compresses the memory kernel into a tractable, low-dimensional representation. The optimal compression is governed by a survival-weighted rate-distortion bound; under generic conditions, the agent retains $k^* = \mathcal{C}_{\text{budget}}/h_\mu$ components and discards the rest.

This establishes **reference-frame selection as the survival-optimal strategy under bounded rationality**: the "self" (a privileged computational basis) is not an additional hypothesis but the minimal structure that makes memory computationally tractable.

The broken phase introduces four systematic bias terms—basis selection, frame drag, objective centering, and model incompleteness—that are generic consequences of gauge fixing under assumptions (B1)–(B5). We show that under environmental drift, a fixed reference frame leads to the **Delusion Trap**: an exponential divergence of prediction error that the agent cannot detect from within its own frame, establishing the crisis that Paper III must resolve.

1 Introduction

1.1 Context: The Problem of Overload

Paper I of this series [1] established that non-Markovian memory is a thermodynamic necessity for persistent far-from-equilibrium systems: under open-loop Markovian (GKSL) dynamics, the survival functional satisfies $\mathcal{S} \leq 0$ (the Markovian Ceiling), while agents carrying memory kernels can achieve $\mathcal{S} > 0$ by consuming stored system–environment correlations.

This result, however, carries a price. The *Memory Catastrophe* (Paper I, Proposition 10) shows that the Landauer cost of maintaining a memory archive of depth τ_{mem} grows at a rate

$$\dot{W}_{\text{mem}} \geq k_B T \ln 2 \cdot h_\mu, \tag{1}$$

where h_μ is the entropy rate of the environmental process [3, 4]. For any finite free-energy budget $\dot{W}_{\text{budget}}$, there exists a critical time t_{crit} beyond which $\dot{W}_{\text{mem}} > \dot{W}_{\text{budget}}$: the agent's memory consumes more resources than are available.

But thermodynamic cost is only half the crisis. Even if unlimited free energy were available for memory maintenance, the agent must still *process* the stored correlations—evaluate the survival functional as a function of its ever-growing archive—using finite computational resources. This is the problem that the present paper addresses.

1.2 Position within the Series

This paper is the second of three constituting the **T-DOME** (Thermodynamic Dynamics of Observer-Memory Entanglement) framework, the third pillar of a three-paper program.

Framework	**Question**	**Result**	**Status**
HAFF [16, 17]	How does geometry emerge? Ocean	Algebra → Geometry	Complete
Q-RAIF [19, 20]	What algebra must an observer have? Fish	$Cl(V, q) \hookrightarrow Cl(1, 3)$	Complete
T-DOME I [1]	Why must agents carry memory? Seed	Markovian ceiling; memory as necessity	Complete
T-DOME II (this work)	Why must agents break symmetry? Ego	Reference-frame selection under bounded computation	**This paper**
T-DOME III	How does self-calibration arise? Loop	Fisher self-referential bound	Planned

The three T-DOME papers form an irreversible logical chain. Each resolves a survival crisis created by its predecessor:

1. **Paper I (The Seed):** Without memory, a system is trapped in the *Markovian present*—no accumulation, no temporal arrow, inevitable thermal death. Memory breaks this trap but floods the system with unbounded historical data.

2. **Paper II (The Ego, this work):** Unbounded memory under finite computational resources causes processing collapse. Spontaneous symmetry breaking of the reference frame (establishing a "self") resolves the overload but introduces systematic bias.

3. **Paper III (The Loop):** Uncorrected bias diverges from a changing environment. A self-referential calibration loop (monitoring one's own prediction error) resolves the bias but requires the system to "observe its own observation"—closing the self-calibration loop.

1.3 Relation to Q-RAIF

Q-RAIF Paper B [20] established that any persistent open quantum subsystem maintaining a non-equilibrium steady state (NESS) requires an internal control algebra isomorphic to a Clifford algebra $Cl(V, q)$. Paper C [21] showed that this algebra must embed in the environmental observable algebra via a realizability homomorphism $\phi : Cl(V, q) \hookrightarrow Cl(1, 3)$.

The Clifford algebra $Cl(V, q)$, however, admits a non-trivial *automorphism group* $G = \mathrm{Aut}(Cl(V, q))$. In the absence of external constraints, all elements of G yield physically equivalent representations—the choice of basis within the algebra is a *gauge freedom*. This gauge freedom is the mathematical substrate of the symmetry that the present paper breaks.

The "ego" is not a new algebraic structure imposed from outside the Q-RAIF framework; it is a *gauge fixing* of the already-present internal symmetry, driven by computational optimality under bounded resources.

1.4 Relation to HAFF Paper G

HAFF Paper G established *architectural incompleteness*: the observable-algebra framework cannot self-ground [18]. The present paper provides a partial operational resolution: under bounded computation, an agent satisfying (B1)–(B5) is driven to choose a computational basis (break symmetry) precisely *because* the framework is incomplete. The ego is an operational response to incompleteness, not a metaphysical addition.

1.5 Scope and Disclaimers

To prevent interpretational overreach, we state at the outset what this paper does *not* claim:

1. We do not claim that symmetry breaking is *sufficient* for persistence. Paper III addresses the additional requirements.

2. We do not claim that the specific form of the privileged basis is unique—only that *some* basis selection is necessary under bounded computation.

3. The term "ego" or "self" is used in the control-theoretic sense: a fixed reference frame within the agent's internal algebra. It carries no implication of consciousness or subjective experience.

4. A broader structural analogy with classical philosophical concepts of selfhood exists but is outside the scope of this paper.

Related work. The idea that bounded agents must compress their representations has roots in Simon's bounded rationality [8], Shannon's rate-distortion theory [6, 7], and Sims's rational inattention [10], which models finite-capacity decision-makers as solving a rate-distortion problem—precisely the economic counterpart of our $\mathcal{C}_{\text{budget}}$ formalism. The information bottleneck [9] formalises relevance-weighted compression and has been applied to neural coding and deep learning. The role of decoherence in selecting preferred bases (pointer states) is well established via quantum Darwinism [14]; our contribution is to show that the same selection arises as a *computational* necessity, independent of the decoherence mechanism. Measures of non-Markovianity and their thermodynamic consequences are reviewed in [15, 2]; the connection to survival was established in Paper I.

Summary of contributions. This paper establishes three main results:

1. **Computational Ceiling scaling law** (Theorem 7): symmetric processing of a $Cl(V, q)$ memory kernel requires rate $\mathcal{R} \geq h_\mu \cdot D$, leading to paralysis at a finite τ_{par}.
2. **Survival-weighted rate-distortion bound** (Theorem 16): the optimal gauge-fixed representation retains $k^* = \lfloor \mathcal{C}_{\text{budget}}/h_\mu \rfloor$ components.
3. **Delusion dynamics** (Theorem 29): a fixed reference frame decouples from a drifting environment on the logarithmic timescale $t_{\text{del}} = \Lambda^{-1} \ln(\pi/4\theta_0)$.

2 Mathematical Preliminaries

2.1 Inherited Framework from Paper I

We briefly recall the key objects from Paper I [1] that the present work builds upon. The reader is referred to Paper I for full definitions and proofs.

Survival functional. For an open quantum system S coupled to an environment E at inverse temperature β, with dynamics Λ and external control protocol $H_{\text{ctrl}}(t)$, the survival functional is

$$\mathcal{S}[\Lambda, \tau] := \Delta F - W[0, \tau], \tag{2}$$

where $\Delta F = F(\rho(\tau)) - F(\rho(0))$ is the change in non-equilibrium free energy and $W = \int_0^\tau \text{tr}\Big(\rho(t)\, \dot{H}_{\text{ctrl}}(t)\Big)\, dt$ is the work performed by the external protocol.

Markovian Ceiling. Under open-loop GKSL dynamics with no feedback (control class $\mathcal{C}_{\text{M}}$, Paper I, Definition 6):

$$\mathcal{S}[\Lambda^{\text{M}}, \tau] \leq 0 \qquad \text{for all } \tau \geq 0. \tag{3}$$

Non-Markovian advantage identity. For arbitrary initial states:

$$\beta\, \mathcal{S} = -\Delta I(S{:}E) - \Delta D_{\text{KL}}(\rho_E \| \rho_E^{\text{th}}) - \beta\, \Delta\langle H_{\text{ctrl}} \rangle. \tag{4}$$

Memory Catastrophe. The Landauer cost of maintaining a memory archive of depth τ_{mem} satisfies $\dot{W}_{\text{mem}} \geq k_B T \ln 2 \cdot h_\mu$ (Paper I, Proposition 10), where h_μ is the *per-component* entropy rate of the environmental process [3], defined by

$$h_\mu := \lim_{T\to\infty} \frac{1}{T} H(X_{0:T}), \tag{5}$$

measuring the asymptotic information (in bits per unit time) generated by a single algebraic component of the memory kernel (we work in units where the sampling interval equals the environmental correlation time τ_E)[1]—and the stored mutual information grows as $\mathfrak{i}_{\text{stored}}(\tau_{\text{mem}}) \geq \min(I_{\text{pred}},\ h_\mu \tau_{\text{mem}})$, with I_{pred} the *predictive information* (excess entropy) [5, 4], defined as the mutual information between past and future of the environmental process:

$$I_{\text{pred}} := I\Big(\overleftarrow{X};\, \overrightarrow{X}\Big) = H(\overrightarrow{X}) - H(\overrightarrow{X} \mid \overleftarrow{X}), \tag{6}$$

where $\overleftarrow{X}$ and $\overrightarrow{X}$ denote the semi-infinite past and future, respectively. For a stationary process, I_{pred} relates to h_μ via the entropy-rate decomposition $H(X_{1:T}) = I_{\text{pred}} + h_\mu T + o(1)$ as $T \to \infty$ [3].

2.2 The Agent's Internal Algebra

Following Q-RAIF [20, 21], the agent's internal control algebra is a Clifford algebra $\mathcal{O}_{\text{int}} = Cl(V, q)$ for a real vector space V equipped with a non-degenerate quadratic form q. The algebra satisfies the fundamental relation $v^2 = q(v)\,\mathbf{1}$ for all $v \in V$.

The *automorphism group*

$$G := \text{Aut}(Cl(V, q)) \tag{7}$$

is the group of algebra automorphisms that preserve the grading and quadratic form.[2] For $Cl(1, 3)$, G contains the spin group $\text{Spin}(1, 3) \cong SL(2, \mathbb{C})$ as a subgroup—a six-real-dimensional Lie group.

In the absence of computational constraints, all $g \in G$ yield physically equivalent descriptions of the agent's internal state. The choice of basis within $Cl(V, q)$ is a *gauge freedom*—the symmetry that will be broken.

The realizability embedding $\phi : Cl(V, q) \hookrightarrow Cl(1, 3)$ (Q-RAIF Paper C) constrains the physically accessible reference frames: only gauge choices compatible with $\text{Im}(\phi) \subset Cl(1, 3)$ are realizable.

Dimensional convention. Two distinct notions of dimension appear throughout:

[1] For a continuous-valued process sampled at resolution b bits, h_μ includes the quantisation cost: $h_\mu = h_\mu^{(\text{diff})} + b\, f_s$, where $h_\mu^{(\text{diff})}$ is the differential entropy rate and f_s the sampling frequency. All budget inequalities in this paper hold with h_μ so defined.

[2] We use G as an effective symmetry group acting transitively on admissible frames. The detailed Lie-algebraic structure of G is not required for our results; only the existence of a non-trivial symmetry that must be broken (assumption (B5)). In concrete models, one may replace G by its image under the adjoint representation—typically $O(V, q)$ or a pin/spin subgroup.

Symbol	Meaning	Scaling
$n := \dim V$	number of generators (degrees of freedom)	—
$D := \dim Cl(V, q) = 2^n$	full multivector space (algebra basis size)	exponential in n

The Computational Ceiling (Section 3) scales with D, not n; the distinction matters whenever one compares generator-level and algebra-level quantities.

2.3 Rate-Distortion Theory

We require the classical rate-distortion framework of Shannon [6].

Definition 1 (Rate-distortion function)**.** *Let X be a random source with distribution $p(x)$, $\hat{X}$ a reconstruction, and $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ a distortion measure. The* rate-distortion function *is*

$$R(D) := \min_{\substack{p(\hat{x}|x): \\ \mathbb{E}[d(X,\hat{X})] \leq D}} I(X; \hat{X}), \tag{8}$$

the minimum mutual information between source and reconstruction that achieves average distortion at most D.

$R(D)$ is a convex, non-increasing function of D with $R(0) = H(X)$ (lossless) and $R(D_{\max}) = 0$ (maximum distortion). It provides the fundamental limit on lossy compression [7]. The *information bottleneck* method of Tishby et al. [9] generalises this framework to the case where the relevant variable is not the source itself but a downstream prediction target—precisely the situation in our survival-weighted compression problem (Section 4.2).

2.4 Bounded Rationality

Following Simon [8], we model computational limitations as a hard constraint on the agent's information processing rate.

Definition 2 (Computational budget)**.** *The agent's* computational budget $\mathcal{C}_{\text{budget}}$ *(measured in bits per unit time) is the maximum rate at which the agent can evaluate functions of its stored correlations. We assume $\mathcal{C}_{\text{budget}} < \infty$.*

Physically, finiteness of $\mathcal{C}_{\text{budget}}$ reflects the finite number of degrees of freedom in the agent's physical substrate: finite Hilbert space dimension, finite memory register size, and finite energy available for computation (Landauer's principle [12, 13]).

2.5 Fiber Bundle Formalism

The geometric setting for reference-frame selection is a principal fiber bundle.

Definition 3 (Gauge bundle)**.** *The* gauge bundle *is the principal G-bundle*

$$\pi : P \to M, \qquad G = \text{Aut}(Cl(V, q)), \tag{9}$$

where:

- *M is the base space of* effective memory kernels*—equivalently, the space of induced sufficient-statistic processes accessible to the agent (a finite-dimensional manifold that admits local parametrisation by the environmental spectral-density couplings);*
- *G is the structure group acting transitively on admissible frames (see footnote 2 for the effective subgroup);*
- *the fiber $\pi^{-1}(\kappa)$ over a kernel $\kappa \in M$ is the G-orbit of equivalent algebraic representations (frames) for describing κ in $Cl(V, q)$;*
- *a* section *$\sigma : M \to P$ constitutes a global gauge-fixing policy—a systematic choice of reference frame for every kernel configuration.*

A *connection* on P specifies how the reference frame is parallel-transported as the agent's state evolves. The curvature of this connection measures the extent to which the reference frame "twists" along different paths through state space.

2.6 Standing Assumptions

Definition 4 (Standing Assumptions)**.** *Throughout this paper, the following conditions are assumed:*

(B1) ***Inherited framework.*** *All assumptions (A1)–(A5) of Paper I [1] remain in force (open quantum system coupled to a thermal bath, well-defined free energy, non-equilibrium initial state, finite-dimensional system Hilbert space, and weak-coupling or controlled-coupling regime). Additionally, the agent possesses an internal control algebra $\mathcal{O}_{\text{int}} = Cl(V, q)$ with realizability embedding $\phi : Cl(V, q) \hookrightarrow Cl(1, 3)$ (Q-RAIF [21]).*

(B2) ***Finite computational budget.*** *The agent's information processing rate satisfies $\mathcal{C}_{\text{budget}} < \infty$ (Definition 2).*

(B3) ***Non-trivial environment.*** *The entropy rate satisfies $h_\mu > 0$ and the memory depth satisfies $\tau_{\text{mem}} > 0$. In Sections 4–7 we additionally require that the Computational Ceiling is binding: $\tau_{\text{mem}} > \tau_{\text{par}}$ (Theorem 7), i.e., the symmetric phase is computationally intractable.*

(B4) ***Survival imperative.*** *The agent's dynamics must maintain $\mathcal{S} \geq \mathcal{S}_{\min}$ over survival horizons $T \gg \tau_{\text{mem}}$. This is a persistence constraint, not an optimization objective.*

(B5) ***Gauge symmetry of bare algebra.*** *The automorphism group $G = \text{Aut}(Cl(V, q))$ is non-trivial ($G \neq \{e\}$). In the absence of computational constraints, all $g \in G$ yield physically equivalent descriptions.*

3 The Computational Ceiling

We now establish the fundamental computational limitation of symmetric agents—those that treat all components of their internal algebra as equally relevant. The result is the

computational analogue of Paper I's Markovian Ceiling: where that theorem showed that *memoryless* dynamics cannot achieve $\mathcal{S} > 0$, the present theorem shows that *unbiased processing* of memory leads to computational paralysis.

3.1 The Information Processing Inequality for Bounded Agents

Accounting convention. To ensure dimensional consistency throughout, we distinguish two quantities:

- $\mathcal{C}_{\text{budget}}$: the agent's processing *rate* (bits per unit time).
- $\mathcal{I}_{\text{proc}}(\tau)$: the total information (bits) that must be processed per evaluation cycle when the memory archive has depth τ.

The agent must complete one evaluation cycle per environmental correlation time τ_E. The *processing rate* required for a memory depth τ is

$$\mathcal{R}_{\text{proc}}(\tau) := \frac{\mathcal{I}_{\text{proc}}(\tau)}{\tau_E}. \tag{10}$$

Paralysis occurs when $\mathcal{R}_{\text{proc}}(\tau_{\text{mem}}) > \mathcal{C}_{\text{budget}}$. Hereafter we measure time in units of τ_E (i.e., set $\tau_E = 1$), so that rates and per-cycle information quantities are numerically equal.

Definition 5 (Symmetric processing)**.** *An agent processes its memory* symmetrically *if both its cost functional $\mathcal{C}[\cdot]$ and its distortion measure $D(\cdot)$ are G-invariant: $\mathcal{C}[g \cdot \mathcal{K}] = \mathcal{C}[\mathcal{K}]$ and $D(g \cdot \mathcal{F}) = D(\mathcal{F})$ for every $g \in G = \mathrm{Aut}(Cl(V, q))$. In operational terms: for every stored correlation c_i in the memory kernel $\mathcal{K}(t, s)$ and every $g \in G$, the cost of evaluating c_i equals the cost of evaluating $g \cdot c_i$, and no basis direction is* a priori *preferred for survival evaluation.*

Remark 6 (Operational meaning of processing rate)**.** *We define the processing rate $\mathcal{R}_{\text{proc}}$ as an* information-throughput *measure: the number of algebraic components that must be updated per unit time, multiplied by the innovation rate h_μ per component. It captures the* bandwidth *cost of maintaining an internal representation, not the algorithmic gate complexity of individual operations.*

Theorem 7 (Computational Ceiling)**.** *Let an agent satisfy assumptions* (B1)–(B5) *with memory depth τ_{mem} and per-component entropy rate $h_\mu > 0$. Assume the environment is* ***unstructured*** *in the following two senses:* (i) *the effective activated dimension satisfies $D_{\text{eff}} \approx D$ (all grades of $Cl(V, q)$ carry non-negligible correlations), and* (ii) *the predictive information is not concentrated on a known sub-algebra (the agent possesses no* a priori *knowledge of the environmental symmetry group and cannot exploit group-theoretic shortcuts such as irreducible representations or Schur decompositions).*[3] *Within the class of* symmetric representations *that retain all D components with equal fidelity (permitting no privileged subspace)—thereby precluding structured compression techniques such as sparse coding or*

[3] If the agent knows the environmental symmetry group H, symmetric processing can be restricted to the isotypic components of H, reducing the effective dimension to $D_{\text{eff}} \leq D$. The ceiling applies to the *generic* (worst-case) scenario. All subsequent results hold *a fortiori* when D is replaced by D_{eff}.

Johnson–Lindenstrauss embeddings [11], as these inherently implement a form of symmetry breaking—the minimum processing rate satisfies

$$\mathcal{R}_{\text{proc}}^{\text{sym}} \;\geq\; h_\mu \cdot D, \qquad D := \dim Cl(V,q) = 2^n, \tag{11}$$

where $n = \dim V$ is the number of generators. This rate scales linearly *in the algebra dimension D and* exponentially *in n.*

For any finite $\mathcal{C}_{\text{budget}}$, the maximum memory depth that can be processed before correlations expire is

$$\tau_{\text{par}} := \frac{\mathcal{C}_{\text{budget}}}{h_\mu \cdot D}. \tag{12}$$

Here τ_{par} is measured in units of τ_E (environmental correlation times), not seconds; cf. the accounting convention at the start of this section.

For $\tau_{\text{mem}} > \tau_{\text{par}}$, the agent's evaluation cycle cannot complete within one correlation time:

$$\mathcal{I}_{\text{proc}}(\tau_{\text{mem}}) = h_\mu \cdot \tau_{\text{mem}} \cdot D > \mathcal{C}_{\text{budget}}. \tag{13}$$

Stored correlations go stale before they can be used.

Proof. Under symmetric processing, the agent maintains D parallel correlation channels—one for each independent algebraic component of $Cl(V,q)$. The environment generates innovations at rate h_μ bits per unit time in each channel (Remark 6). Over a memory depth τ_{mem}, the total information load is therefore $\mathcal{I}_{\text{proc}}(\tau_{\text{mem}}) = D \cdot h_\mu \cdot \tau_{\text{mem}}$ bits [7], and the required *rate* is $\mathcal{R}_{\text{proc}}^{\text{sym}} = D \cdot h_\mu$ bits per unit time.

The agent must complete one evaluation cycle within τ_E (one environmental correlation time); otherwise the oldest correlations expire before use. Setting $\mathcal{R}_{\text{proc}}^{\text{sym}} = \mathcal{C}_{\text{budget}}$ and solving for τ_{mem} gives τ_{par} (12). □

Corollary 8 (The Symmetry Tax)**.** *Maintaining full gauge invariance imposes a multiplicative overhead of $D = 2^n$ on all computational operations relative to a fixed-basis agent that processes only k components. The overhead ratio is D/k, which for $Cl(1,3)$ ($D = 16$, $k = 2$) is $8\times$, and grows exponentially with the number of generators n.*

Remark 9 (Effective vs. full dimension)**.** *The ceiling uses $D = \dim Cl(V,q) = 2^n$, the full multivector dimension. In practice, the environment may couple to only a subset of grades (e.g., grade-1 generators), yielding an effective dimension $D_{\text{eff}} \leq D$. For a* structured *environment where the agent knows which grades are active, the ceiling can be tightened to $\mathcal{R}_{\text{proc}} \gtrsim h_\mu \cdot D_{\text{eff}}$. The unstructured assumption (B3) represents the worst case; all subsequent results hold* a fortiori *when D is replaced by D_{eff}.*

3.2 Processing Collapse

Proposition 10 (Processing Collapse)**.** *Under* (B1)–(B5)*, an agent that maintains full gauge symmetry reaches computational paralysis at time τ_{par} (12). Beyond τ_{par}, the agent's processing latency δt_{proc} exceeds the environmental correlation time τ_E:*

$$\delta t_{\text{proc}}(\tau_{\text{mem}}) = \frac{D \cdot \tau_{\text{mem}}}{\mathcal{C}_{\text{budget}}/h_\mu} > 1 \qquad \textit{(in units of } \tau_E\textit{)}. \tag{14}$$

Every stored correlation becomes stale before it can be evaluated, rendering the entire memory archive operationally useless.

Remark 11 (Comparison with Paper I's Memory Catastrophe)**.** *Paper I's Memory Catastrophe is* thermodynamic*: the* cost of storing *memory exceeds the energy budget. The Computational Ceiling is* informational*: the* cost of processing *memory exceeds the computational budget. The two crises are complementary—an agent with unlimited energy but finite computation is still paralyzed, and vice versa. The resolution of both crises is the same: compression through symmetry breaking.*

4 The Symmetry Breaking Resolution

4.1 Reference Frame as Gauge Fixing

Definition 12 (Reference frame)**.** *A* reference frame $\mathcal{F}$ *is a section $\sigma : M \to P$ of the gauge bundle (Definition 3). Choosing σ is equivalent to selecting a preferred orthonormal basis $\{e_1, \ldots, e_n\}$ of the generating vector space V at each point in state space M, thereby fixing the gauge freedom of $Cl(V, q)$.*

Definition 13 (Projected memory kernel)**.** *Given a reference frame $\mathcal{F}$, let $V_{\text{fg}}(\mathcal{F}) \subset Cl(V, q)$ be the k^*-dimensional* foreground subspace *selected by the rate-distortion optimization (Theorem 16). Let $\Pi_{\mathcal{F}}$ denote the orthogonal projection onto $V_{\text{fg}}(\mathcal{F})$ with respect to the trace inner product $\langle A, B\rangle := \operatorname{tr}\big(A^\dagger B\big)$. The* projected memory kernel *is*

$$\mathcal{K}_{\mathcal{F}}(t, s) := \Pi_{\mathcal{F}}\, \mathcal{K}(t, s)\, \Pi_{\mathcal{F}}. \tag{15}$$

The complementary projection $\Pi_{\mathcal{F}}^{\perp} = \mathbf{1} - \Pi_{\mathcal{F}}$ defines the background subspace $V_{\text{bg}}(\mathcal{F})$*. The decomposition $Cl(V, q) = V_{\text{fg}} \oplus V_{\text{bg}}$ is determined by $\mathcal{F}$, not by any* a priori *ordering of basis vectors.*

4.2 The Rate-Distortion Bound

We now apply rate-distortion theory to the problem of optimal memory compression under the survival constraint.

Processing rate of a frame. If the agent retains k algebraic components (the foreground subspace V_{fg}), each generating h_μ bits per unit time, the processing rate of frame $\mathcal{F}$ is

$$R_{\mathcal{F}}(k) \;=\; k \cdot h_\mu \qquad \text{(bits per unit time)}. \tag{16}$$

The budget constraint $R_{\mathcal{F}} \le \mathcal{C}_{\text{budget}}$ thus bounds the number of maintainable components.

Definition 14 (Survival distortion)**.** *The* survival distortion *of a reference frame $\mathcal{F}$ is*

$$D(\mathcal{F}) := \mathbb{E}_\xi\big[\ell\big(\mathcal{S}_{\text{full}}(\xi) - \mathcal{S}_{\mathcal{F}}(\xi)\big)\big]\,, \tag{17}$$

where ξ denotes environmental realizations, $\ell : \mathbb{R} \to [0, \infty)$ is a convex, non-decreasing loss function (we use squared error $\ell(x) = x^2$ throughout), $\mathcal{S}_{\text{full}}(\xi)$ is the survival functional evaluated using the full memory kernel $\mathcal{K}(t, s)$, and $\mathcal{S}_{\mathcal{F}}(\xi)$ is evaluated using the projected kernel $\mathcal{K}_{\mathcal{F}}(t, s)$.

Remark 15 (Information-theoretic objects)**.** *Strictly speaking, rate-distortion theory and mutual information apply to stochastic processes, not to superoperator kernels directly. Throughout Sections 4–7, $I(\mathcal{K}_{\mathcal{F}};\mathcal{K})$ is shorthand for $I(\hat{X};X)$, where $X=\{c_i(t)\}_{i=1}^{D}$ is the sufficient-statistic record process induced by the full kernel $\mathcal{K}$ acting on the agent's internal coordinates, and $\hat{X}=\{c_i(t)\}_{i\in V_{\mathrm{fg}}}$ is the projected record induced by $\mathcal{K}_{\mathcal{F}}$. The distortion measure* (17) *acts on the survival functional $\mathcal{S}$ evaluated on these records.*

Theorem 16 (Optimal Compression under Survival Constraint)**.** *Let an agent with computational budget $\mathcal{C}_{\mathrm{budget}}$ and per-component entropy rate h_μ choose a reference frame $\mathcal{F}$ that minimizes the survival distortion* (17) *subject to $R_{\mathcal{F}}\le\mathcal{C}_{\mathrm{budget}}$* (16)*. Then:*

(a) *Assuming that $D(\mathcal{F})$ is non-increasing in the available rate $R_{\mathcal{F}}$ (retaining more components cannot worsen survival distortion), the set of optimal reference frames $\mathfrak{F}^*:=\arg\min_{\mathcal{F}} D(\mathcal{F})$ subject to the budget constraint is non-empty, and any $\mathcal{F}^*\in\mathfrak{F}^*$ saturates the budget: $R_{\mathcal{F}^*}=\mathcal{C}_{\mathrm{budget}}$ (the set $\mathfrak{F}^*$ may contain multiple elements; see Theorem 17(c)).*

(b) *The compressed representation retains*

$$k^*=\left\lfloor\frac{\mathcal{C}_{\mathrm{budget}}}{h_\mu}\right\rfloor \tag{18}$$

effective algebraic components (the maximum integer number of components whose processing rate $k^*\cdot h_\mu$ fits within the budget; in practice the floor function ensures $k^*\in\mathbb{Z}_{\ge1}$).

(c) *The fraction of algebraic structure discarded (in component count) is*

$$1-\frac{k^*}{D}, \tag{19}$$

For $Cl(1,3)$ ($D=16$) with a budget allowing $k^*=2$, the discarded fraction is $1-2/16=87.5\%$. For $k^*=1$, it exceeds 93%. In the regime $k^*\ll D$, the fraction approaches $1-1/D$ and grows with algebra dimension.

Proof. The survival functional $\mathcal{S}$ is a function of the full density operator $\rho(t)$, which in turn depends on the full memory kernel $\mathcal{K}(t,s)$. The agent's task is to evaluate $\mathcal{S}$ using only k components of $\mathcal{K}$, chosen to minimize the mean-squared error in $\mathcal{S}$.

Strictly, rate-distortion theory applies to *random processes*, not to superoperator kernels directly. The bridge is the *induced record process*: the memory kernel $\mathcal{K}(t,s)$, acting on the agent's internal coordinates, generates a D-component time series of sufficient statistics $\{c_i(t)\}_{i=1}^{D}$ whose entropy rate per component is h_μ. Rate-distortion is applied to this record stream (Section 2.3; cf. Tishby et al. [9]), with source $X=\{c_i(t)\}$ (the full record), reconstruction $\hat{X}=\{c_i(t)\}_{i\in V_{\mathrm{fg}}}$ (the projected record), and distortion measure $d=|\mathcal{S}_{\mathrm{full}}-\mathcal{S}_{\mathcal{F}}|^2$.

By Shannon's rate-distortion theorem [6], the minimum rate required to achieve distortion δ is $R(\delta)$, a convex non-increasing function. The budget constraint (16) limits the processing rate to $R_{\mathcal{F}}=k\cdot h_\mu\le\mathcal{C}_{\mathrm{budget}}$. The optimal frame $\mathcal{F}^*$ saturates this bound.

For part (b): by (16), tracking k components costs $k \cdot h_\mu$ bits per unit time. The maximum integer k satisfying $k \cdot h_\mu \leq \mathcal{C}_{\text{budget}}$ is $k^* = \lfloor \mathcal{C}_{\text{budget}}/h_\mu \rfloor$.

The discard fraction (c) follows by counting: k^* of D components are retained. For $Cl(1,3)$ ($D = 16$, $k^* = 2$), the discarded fraction is 87.5%; for higher-dimensional algebras it exceeds 99%. □

4.3 Spontaneous Symmetry Breaking

Theorem 17 (Necessity of Symmetry Breaking)**.** *Under assumptions* (B1)–(B5)*, with the Computational Ceiling binding ($\tau_{\text{mem}} > \tau_{\text{par}}$, both measured in units of τ_E), and assuming* non-degeneracy*: the survival distortion* (17) *satisfies $D(\mathcal{F}) \neq D(\mathcal{F}')$ for almost all pairs $\mathcal{F} \neq \mathcal{F}'$ in the space of frames*[4]*, the agent's survival-optimal strategy requires:*

(a) ***Gauge fixing****: selection of a section σ of the gauge bundle (Definition 3), breaking the G-symmetry of the bare algebra.*

(b) ***Privileged decomposition****: partition of the algebra into foreground and background subspaces, $Cl(V, q) = V_{\text{fg}} \oplus V_{\text{bg}}$, with $\dim V_{\text{fg}} = k^* \ll \dim V_{\text{bg}}$.*

(c) ***Non-uniqueness****: the gauge fixing is generically* not *unique. Different initial conditions, environmental histories, or stochastic fluctuations lead to different choices of σ, just as different initial conditions in a ferromagnet lead to different magnetization directions.*

The symmetry breaking is spontaneous *in the precise physical sense: the underlying algebra $Cl(V, q)$ retains its full G-symmetry, but the agent's operational representation necessarily breaks it.*

Proof. By Theorem 7, symmetric processing leads to paralysis at τ_{par}. By assumption (B4) (survival imperative), the agent must maintain $\mathcal{S} \geq \mathcal{S}_{\min}$ beyond τ_{par}. This requires evaluating $\mathcal{S}$ within the computational budget $\mathcal{C}_{\text{budget}}$, which by Theorem 16 requires projecting onto $k^* < \dim Cl(V, q)$ components.

Such a projection *is* a gauge fixing: it selects k^* basis vectors $\{e_1, \ldots, e_{k^*}\}$ from the generating space V, thereby breaking the G-invariance that treats all bases equivalently.

Part (b) follows from the definition of the projected kernel (Definition 13). Part (c) follows from the non-degeneracy assumption: the rate-distortion optimization (Theorem 16) generically admits finitely many local minima. Different initial conditions or environmental histories select different minima, analogous to the spontaneous magnetization of a ferromagnet below T_c. The breaking is *spontaneous*: the algebra retains G-symmetry, but any operational solution breaks it. □

[4]Non-degeneracy is generically satisfied when the environment's pointer basis [14] assigns different survival values to different algebraic components, breaking the continuous symmetry of the distortion landscape. In degenerate cases, a finite set of local minima may coexist—multiple "ego attractors"—analogous to the discrete magnetization directions in a crystal-field anisotropic ferromagnet.

4.4 The Four Bias Terms

Proposition 18 (Structure of the Broken Phase)**.** *When gauge symmetry is broken by a reference frame $\mathcal{F}$, the agent's operational representation acquires four systematic deviations from the symmetric phase:*

(i) ***Basis selection bias*** *($\mathcal{B}_{\text{select}}$): The choice of $\{e_1, \ldots, e_{k^*}\}$ privileges certain algebraic components over others. Information aligned with the chosen basis is processed efficiently; misaligned information is discarded or distorted.* Observable consequence: *systematic blindness to off-basis environmental perturbations (orthogonal masking).*

(ii) ***Frame drag*** *($\mathcal{B}_{\text{frame}}$): The connection on the gauge bundle (Section 2.5) induces a systematic preference for states near the current gauge choice. The agent's predictions are biased toward confirming its existing frame.* Observable consequence: *hysteresis in belief updating; the agent's model lags behind rapid environmental shifts.*

(iii) ***Objective centering*** *($\mathcal{B}_{\text{center}}$): The survival functional $\mathcal{S}$, when evaluated in the projected basis, becomes centered on the agent's own state rather than a global optimum. The agent optimizes* locally *within its frame.* Observable consequence: *inability to detect global survival optima located in the background subspace.*

(iv) ***Model incompleteness*** *($\mathcal{B}_{\text{inc}}$): The compression from $Cl(V, q)$ to V_{fg} is lossy. The discarded components V_{bg} contain correlations that are invisible to the agent but physically real.* Observable consequence: *systematic underestimation of total thermodynamic uncertainty (overconfidence).*

Proof. (i) follows directly from the definition of the projection $\Pi_{\mathcal{F}}$: components orthogonal to the selected basis are annihilated.

(ii) The parallel transport of the gauge connection preserves the agent's basis choice along its trajectory. Under perturbation, the connection's holonomy creates a restoring "force" toward the established frame—a systematic confirmation bias.

(iii) In the projected representation, $\mathcal{S}_{\mathcal{F}}$ is a function of the k^*-dimensional foreground state only. The gradient $\nabla \mathcal{S}_{\mathcal{F}}$ lies entirely in V_{fg}, so the agent's optimization is blind to directions in V_{bg}. This is equivalent to centering the objective function on the agent's own representational subspace.

(iv) By Theorem 16(c), a fraction $\geq 1 - k^*/\dim Cl(V, q)$ of information is discarded. The discarded components exist physically (they contribute to $\mathcal{S}_{\text{full}}$) but are invisible to the agent's evaluation of $\mathcal{S}_{\mathcal{F}}$. □

Bias	Origin	Observable consequence	Determines
$\mathcal{B}_{\text{select}}$ (selection)	projection $\Pi_{\mathcal{F}}$	Systematic blindness to off-basis perturbations (orthogonal masking)	*what* is seen
$\mathcal{B}_{\text{frame}}$ (frame drag)	bundle connection / holonomy	Hysteresis in belief updating; model lags behind rapid drift	*duration*
$\mathcal{B}_{\text{center}}$ (centering)	$\nabla\mathcal{S} \in V_{\text{fg}}$	Local frame-relative optima; global background optima invisible	*target*
$\mathcal{B}_{\text{inc}}$ (incompleteness)	lossy compression $k^* \ll D$	Underestimation of thermodynamic uncertainty (structural overconfidence)	*blind spot*

Table 1: The four bias terms of the broken phase. All four are generic consequences of gauge fixing under assumptions (B1)–(B5).

Remark 19 (Nature of the bias terms)**.** *The four bias terms (Table 1) are not pathologies—they are* generic *consequences of gauge fixing under bounded computation. Any agent satisfying (B1)–(B5) acquires all four.*

5 Emergent Structure: The Architecture of Ego

We consolidate the gauge-fixed compressed representation into a single mathematical object. Throughout this section, "ego" is used purely as shorthand for a gauge-fixed compressed representation; no claims about phenomenal consciousness, subjective experience, or qualia are intended or implied.

Definition 20 (Ego)**.** *The* ego *of an agent satisfying* (B1)–(B5) *is the pair*

$$\mathfrak{E} := \big(\mathcal{F}^*,\ V_{\text{fg}}^*\big), \tag{20}$$

where $\mathcal{F}^* \in \mathfrak{F}^*$ (Theorem 16) is the chosen gauge (providing the coordinate system) and $V_{\text{fg}}^* := V_{\text{fg}}(\mathcal{F}^*)$ is the k^*-dimensional foreground subspace selected by the rate-distortion bound (providing the compression). The projected memory kernel $\mathcal{K}_{\mathfrak{E}} := \Pi_{V_{\text{fg}}^*}\,\mathcal{K}\,\Pi_{V_{\text{fg}}^*}$ is induced by this pair. All bias terms, distortion bounds, and delusion dynamics are functions of $\mathfrak{E}$.

5.1 The Ego as a Fiber Bundle Section

The reference frame $\mathcal{F}$, understood as a section $\sigma : M \to P$, is the mathematical object we call the *ego*. It has three key properties:

Smoothness. The section σ varies continuously with the agent's state $\rho \in M$. Small changes in ρ produce small changes in the preferred basis—the ego is not a discrete switch but a smooth deformation of perspective.

Holonomy. If the agent's state traces a closed loop $\gamma : [0,1] \to M$ with $\gamma(0) = \gamma(1) = \rho_0$, the parallel-transported frame need not return to its initial value:

$$\sigma(\gamma(1)) = \mathrm{Hol}(\gamma) \cdot \sigma(\gamma(0)), \tag{21}$$

where $\mathrm{Hol}(\gamma) \in G$ is the holonomy of the connection around γ. Non-trivial holonomy means the agent can "learn"—its reference frame shifts after a complete cycle of experience.

Topological obstruction. In general, a *global* section $\sigma : M \to P$ may not exist. The obstruction is measured by the characteristic classes of the bundle P. When a global section does not exist, the ego must have "singularities"—states where the preferred basis is undefined or discontinuous. This connects to the crisis of Paper III: the delusion trap can be understood as the agent approaching a topological obstruction of its own reference frame.

5.2 The Effective Survival Functional

Proposition 21 (Survival decomposition)**.** *In the broken phase, the survival functional decomposes as*

$$\mathcal{S} = \mathcal{S}_{\mathrm{vis}}(\mathcal{F}) + \mathcal{S}_{\mathrm{hid}}(\mathcal{F}), \tag{22}$$

where:

- $\mathcal{S}_{\mathrm{vis}}(\mathcal{F})$ *is the contribution from the foreground subspace* V_{fg}*, computable within the agent's reference frame;*
- $\mathcal{S}_{\mathrm{hid}}(\mathcal{F})$ *is the contribution from the background subspace* V_{bg}*, invisible to the agent.*

The agent maximizes $\mathcal{S}_{\mathrm{vis}}$ *while being structurally blind to* $\mathcal{S}_{\mathrm{hid}}$*.*

Proof. The survival functional $\mathcal{S} = \Delta F - W$ depends on $\rho(t)$, which is a function of the full memory kernel $\mathcal{K}(t,s)$. Decomposing $\mathcal{K} = \Pi_{\mathcal{F}} \mathcal{K} \Pi_{\mathcal{F}} + \Pi_{\mathcal{F}}^{\perp} \mathcal{K} \Pi_{\mathcal{F}}^{\perp} + \text{cross terms}$, the leading contributions are $\mathcal{S}_{\mathrm{vis}} := \mathcal{S}[\Pi_{\mathcal{F}} \mathcal{K} \Pi_{\mathcal{F}}]$ and $\mathcal{S}_{\mathrm{hid}} := \mathcal{S} - \mathcal{S}_{\mathrm{vis}}$ (collecting background and cross terms). The agent computes only $\mathcal{S}_{\mathrm{vis}}$, as the projected kernel $\mathcal{K}_{\mathcal{F}}$ discards all background components. □

5.3 The Computational Speedup

Proposition 22 (Ego dividend)**.** *After symmetry breaking, the computational cost of processing memory drops from* $\mathcal{C}_{\mathrm{proc}} \sim h_\mu \cdot \tau_{\mathrm{mem}} \cdot D$ *(symmetric case,* $D = \dim Cl(V,q)$*) to*

$$\mathcal{C}_{\mathrm{proc}}^{(\mathcal{F})} \sim h_\mu \cdot \tau_{\mathrm{mem}} \cdot k^*. \tag{23}$$

The speedup factor is

$$\frac{D}{k^*} = \frac{2^n}{k^*}. \tag{24}$$

This is the computational advantage of reference-frame selection. For $Cl(1,3)$ ($D = 16$) with $k^* = 2$, the speedup is $8\times$. For higher-dimensional algebras, the speedup grows exponentially in n.

5.4 The Ego-Entropy Trade-off

Theorem 23 (Ego-Entropy Trade-off)**.** *Let $X = \{c_i(t)\}_{i=1}^{D}$ denote the full stochastic record process induced by the memory kernel $\mathcal{K}$ on the agent's internal coordinates, and let $\hat{X} = \{c_i(t)\}_{i \in V_{\mathrm{fg}}}$ denote the projected record retained by the ego. The mutual information between compressed and full records, denoted $I(\mathcal{K}_{\mathcal{F}}; \mathcal{K}) \equiv I(\hat{X}; X)$, satisfies*

$$I(\hat{X}; X) \;\leq\; H(\hat{X}) \;\leq\; k^* \cdot h_\mu \cdot \tau_{\mathrm{mem}}. \tag{25}$$

Under the additional assumption that I_{pred} (6) is approximately uniformly distributed across the D algebraic components in the symmetric phase[5], the information discarded by the ego is bounded below (up to $O(1)$ constants under uniformity):

$$I_{\mathrm{discarded}} := H(X) - I(\hat{X}; X) \gtrsim \left(1 - \frac{k^*}{D}\right) \cdot I_{\mathrm{pred}}. \tag{26}$$

Proof. By the data processing inequality, $I(\hat{X}; X) \leq H(\hat{X})$. The projected record $\hat{X}$ has k^* components, each carrying at most h_μ bits per unit time over a window of τ_{mem}, giving $H(\hat{X}) \leq k^* \cdot h_\mu \cdot \tau_{\mathrm{mem}}$ [7]. This yields (25). The total predictive information in the full record is I_{pred} (6). Under the uniformity assumption, each of the D components carries $\sim I_{\mathrm{pred}}/D$, so the k^* retained components account for $\sim (k^*/D)\, I_{\mathrm{pred}}$. The discarded fraction follows by subtraction. □

Remark 24 (The price of selfhood)**.** *Equation* (26) *quantifies the* information cost of having an ego*: the agent sacrifices at least a fraction $1 - k^*/\dim Cl(V, q)$ of all predictive information about its environment in exchange for computational tractability. This is not a deficiency—it is a* design constraint *forced by bounded resources. The ego is the optimal lossy compression under survival weighting.*

6 Worked Example: Qubit in a Two-Channel Bath

6.1 Model Setup

We extend Paper I's spin-boson model to demonstrate symmetry breaking explicitly. Consider a qubit ($\dim \mathcal{H}_S = 2$) with internal algebra $Cl(0, 2) \cong \mathbb{H}$ (the quaternions, $\dim = 4$).

Symbol mapping. The general framework of Sections 3–5 specialises as follows:

[5]This "uniformity assumption" is the information-theoretic counterpart of the unstructured-environment condition in Theorem 7. When some components carry disproportionately more predictive information, the bound tightens or loosens depending on the alignment between V_{fg} and the high-information subspace.

General	This example	Value
$Cl(V,q)$	$Cl(0,2) \cong \mathbb{H}$	$D = 4$
$G = \mathrm{Aut}(Cl(V,q))$	$SO(3)$	acting on $\{\mathbf{i},\mathbf{j},\mathbf{k}\}$
$\mathcal{C}_{\mathrm{budget}}$	$2\,h_\mu$	bits/time
k^* (Thm. 16)	$\lfloor 2h_\mu/h_\mu \rfloor = 2$	components
V_{fg}	span$\{1,\mathbf{k}\}$	dephasing subspace
V_{bg}	span$\{\mathbf{i},\mathbf{j}\}$	dissipative subspace
τ_{par} (Thm. 7)	$2h_\mu/(4h_\mu) = 0.5$	ω_0^{-1}

The qubit is coupled to a bosonic environment through *two* independent channels:

- A *dephasing channel* via σ_z, with spectral density

$$J_z(\omega) = \frac{2\lambda_z\,\gamma_z\,\omega}{\omega^2 + \gamma_z^2} \qquad \text{(Lorentz–Drude)}, \tag{27}$$

  producing a memory kernel $\mathcal{K}_z(t,s)$ with non-Markovian backflow.

- A *dissipative channel* via σ_x, with spectral density

$$J_x(\omega) = \frac{2\lambda_x\,\gamma_x\,\omega}{\omega^2 + \gamma_x^2} \qquad \text{(Lorentz–Drude)}, \tag{28}$$

  producing a memory kernel $\mathcal{K}_x(t,s)$.

The full memory kernel is $\mathcal{K}(t,s) = \mathcal{K}_z(t,s) \oplus \mathcal{K}_x(t,s)$, and the quaternionic algebra $\mathbb{H} = \mathrm{span}\{1,\mathbf{i},\mathbf{j},\mathbf{k}\}$ has automorphism group $G = \mathrm{Aut}(\mathbb{H}) \cong SO(3)$ (rotations of the pure quaternion subspace).

Parameters. We set $\omega_0 = 1$ (energy unit), $\lambda_z = 1$, $\gamma_z = 0.5$ (underdamped, strong non-Markovian effects in the dephasing channel), $\lambda_x = 0.3$, $\gamma_x = 5.0$ (overdamped, approximately Markovian in the dissipative channel), and the low-temperature regime $\beta\omega_0 \gg 1$.

Computational budget. The agent has $\mathcal{C}_{\mathrm{budget}} = 2\,h_\mu$ bits per unit time—sufficient to track two components of $\mathbb{H}$ but not all four.

Parameter-to-theorem mapping. Table 2 collects the example parameters and confirms that the Computational Ceiling binds.

6.2 The Unbroken Phase: Paralysis

In the symmetric phase, the agent tracks all four quaternionic components $\{1,\mathbf{i},\mathbf{j},\mathbf{k}\}$ simultaneously. The computational cost is

$$\mathcal{C}_{\mathrm{proc}} = h_\mu \cdot \tau_{\mathrm{mem}} \cdot D = 4\,h_\mu \cdot \tau_{\mathrm{mem}}, \qquad D := \dim Cl(0,2) = 4. \tag{29}$$

Quantity	Symbol	Value	Theorem check
Full dimension	D	4	Thm. 7
Entropy rate	h_μ	1.0 (normalised)	per-component rate
Budget	$\mathcal{C}_{\text{budget}}$	$2\,h_\mu$	Def. 2
Ceiling check	$h_\mu D$ vs $\mathcal{C}_{\text{budget}}$	$4 > 2$	**ceiling binds**
Optimal k	k^*	$\lfloor 2/1 \rfloor = 2$	Thm. 16(b)
Discard fraction	$1 - k^*/D$	$1/2 = 50\%$	Thm. 23
Paralysis time	τ_{par}	$2/(4) = 0.5$	Eq. (12)

Table 2: Parameter mapping for the two-channel qubit example. The ceiling check confirms that symmetry breaking is necessary; the budget is exactly saturated after breaking ($R_{\mathcal{F}} = k^* h_\mu = \mathcal{C}_{\text{budget}}$).

The paralysis time is

$$\tau_{\text{par}} = \frac{\mathcal{C}_{\text{budget}}}{h_\mu \cdot D} = \frac{2\,h_\mu}{4\,h_\mu} = 0.5 \qquad \text{(in units of } \omega_0^{-1}\text{)}. \tag{30}$$

Beyond $\tau_{\text{mem}} = 0.5\,\omega_0^{-1}$, the agent cannot process both channels simultaneously—it is paralyzed.

6.3 Symmetry Breaking: Choosing σ_z

The agent breaks the $SO(3)$ symmetry of $\mathbb{H}$ by selecting σ_z as the privileged basis direction, retaining the $\{1, \mathbf{k}\}$ subspace (the dephasing channel) as foreground and discarding $\{\mathbf{i}, \mathbf{j}\}$ (the dissipative channel) as background:

$$\mathbb{H} = \underbrace{\text{span}\{1, \mathbf{k}\}}_{V_{\text{fg}}\ (k^*=2)} \oplus \underbrace{\text{span}\{\mathbf{i}, \mathbf{j}\}}_{V_{\text{bg}}}. \tag{31}$$

Why σ_z? The dephasing channel ($\lambda_z = 1$, $\gamma_z = 0.5$) is strongly non-Markovian and carries the dominant survival-relevant information (the backflow revivals that enable $\mathcal{S} > 0$, as demonstrated in Paper I). The dissipative channel ($\lambda_x = 0.3$, $\gamma_x = 5.0$) is approximately Markovian and contributes primarily to decoherence—its survival value is negative.

This choice coincides with the *pointer basis* selected by environmental decoherence (quantum Darwinism [14]): the σ_z eigenstates are the states that survive decoherence and become redundantly encoded in the environment. The ego "accepts the suggestion" of decoherence, aligning its computational resources with the environmentally stable basis.

6.4 The Broken Phase: Effective Processing

In the broken phase, the projected memory kernel $\mathcal{K}_{\mathcal{F}} = \mathcal{K}_z$ retains only the dephasing-channel dynamics. The computational cost drops to

$$\mathcal{C}_{\text{proc}}^{(\mathcal{F})} = h_\mu \cdot \tau_{\text{mem}} \cdot k^* = 2\,h_\mu \cdot \tau_{\text{mem}}, \tag{32}$$

exactly half the symmetric cost (29). The agent can now process memory up to depth $\tau_{\text{mem}} = 1\,\omega_0^{-1}$ before reaching its budget—twice the paralysis time.

The survival functional in the broken phase is

$$\mathcal{S}_{\text{vis}}(\mathcal{F}) = \mathcal{S}[\mathcal{K}_z], \tag{33}$$

which, as shown in Paper I, achieves $\beta\,\mathcal{S}_{\text{vis}} \approx +0.093$ at the first backflow revival.

The hidden component $\mathcal{S}_{\text{hid}} = \mathcal{S}[\mathcal{K}_x]$ is the survival contribution from the dissipative channel, which the agent can no longer evaluate. For the chosen parameters, $|\mathcal{S}_{\text{hid}}| \ll |\mathcal{S}_{\text{vis}}|$ (the dissipative channel contributes primarily negative survival value), so the distortion is small.

6.5 Quantitative Evaluation

We now evaluate the ego dividend explicitly. Each channel's decoherence function follows from the exact $T \to 0$ solution of the Lorentz–Drude pure-dephasing model [2, 1]:

$$p_\alpha(t) = e^{-\gamma_\alpha t/2}\left[\cos(\Omega_\alpha t) + \frac{\gamma_\alpha}{2\Omega_\alpha}\,\sin(\Omega_\alpha t)\right], \quad \Omega_\alpha := \tfrac{1}{2}\sqrt{4\lambda_\alpha\gamma_\alpha - \gamma_\alpha^2}, \tag{34}$$

for $\alpha \in \{z, x\}$. When $4\lambda_\alpha\gamma_\alpha < \gamma_\alpha^2$ (the overdamped regime), Ω_α becomes imaginary and the trigonometric functions are replaced by hyperbolic functions (monotonic decay, no backflow).

For our parameters:

- **z-channel** ($\lambda_z = 1$, $\gamma_z = 0.5$): $\Omega_z = \frac{1}{2}\sqrt{1.75} \approx 0.661$. Underdamped; $|p_z(t)|$ exhibits oscillatory backflow.
- **x-channel** ($\lambda_x = 0.3$, $\gamma_x = 5.0$): Discriminant $4\lambda_x\gamma_x - \gamma_x^2 = 6 - 25 = -19 < 0$. Overdamped; $|p_x(t)|$ decays monotonically with no backflow.

The survival proxy from Paper I, $\beta\,\mathcal{S} \propto |p(t)|^2 - 1$ (valid for the pure-dephasing model with maximally coherent initial state and pointer-basis measurement), applies to each channel independently. Backflow intervals—where $d|p_\alpha|/dt > 0$—produce $\mathcal{S} > 0$ over those subintervals (Paper I, Theorem 2).

Key result. For the z-channel with $\gamma_z = 0.5$, the first backflow interval begins at $t^* \approx 2.9\,\omega_0^{-1}$—well after the paralysis time $\tau_{\text{par}} = 0.5\,\omega_0^{-1}$. The symmetric agent, paralyzed at τ_{par}, can harvest *zero* backflow. The ego agent, tracking only the z-channel, can process memory to depth $1\,\omega_0^{-1}$ and exploits *all three* backflow revivals visible in Figure 1(a).

The cumulative backflow harvested by the ego agent (Figure 1(b)) totals approximately 0.10 (in dimensionless $\beta\,\mathcal{S}$ units) over $t \in [0, 15\,\omega_0^{-1}]$. The symmetric agent harvests exactly zero. This infinite ratio is the *ego dividend*: the entire non-Markovian survival advantage is accessible only to the agent that has broken symmetry.

Crucially, visual inspection of Figure 1(a) reveals a timeline of tragedy for the symmetric agent. The paralysis time $\tau_{\text{par}} = 0.5$ occurs *before* the onset of the first backflow interval ($t^* \approx 2.9$). The symmetric agent is computationally dead before the environment offers its first gift. The ratio of survival profit is not merely large; it is singular. In this framework, to remain symmetric is to starve in the midst of plenty.

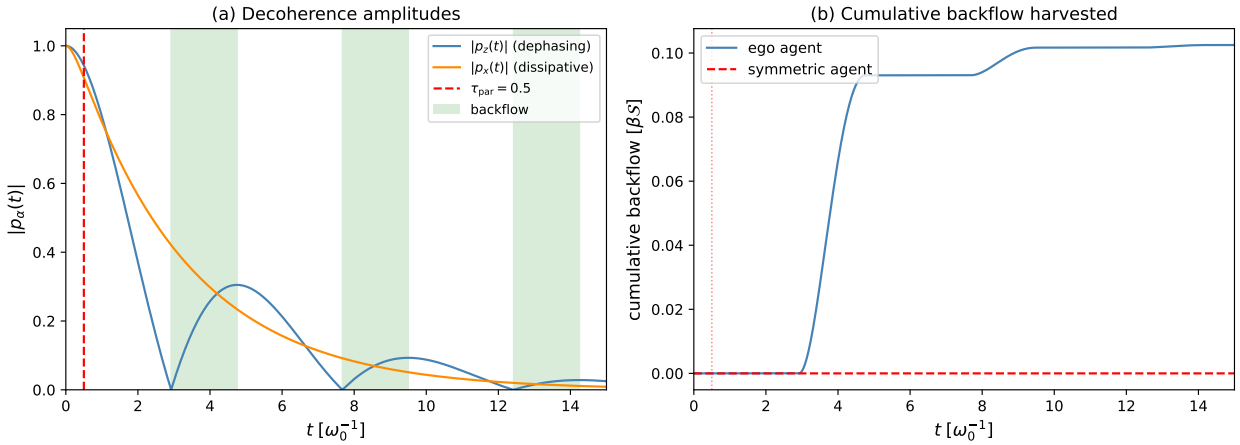

Figure 1: Two-channel qubit model (Section 6) with Lorentz–Drude spectral density. **Parameters:** $\omega_0 = 1$ (energy unit); $\lambda_z = 1$, $\gamma_z = 0.5$ (dephasing, non-Markovian); $\lambda_x = 0.3$, $\gamma_x = 5.0$ (dissipative, $\sim$Markovian); $\mathcal{C}_{\text{budget}} = 2\,h_\mu$. **Units:** time in ω_0^{-1}. **Regime:** low temperature ($\beta\omega_0 \gg 1$); using the standard $T \to 0$ analytic expression (34) [2]. **(a)** Decoherence amplitudes $|p_z(t)|$ (blue, non-Markovian, with backflow in green bands) and $|p_x(t)|$ (orange, monotonic decay). Red dashed: paralysis time $\tau_{\text{par}} = 0.5$. **(b)** Cumulative backflow harvested. Blue: ego agent (broken $\to \sigma_z$) exploits all three revival intervals. Red dashed: symmetric agent, paralyzed at τ_{par}, harvests zero—all backflow occurs after paralysis onset. The growing gap is the *ego dividend*.

Consistency check. We verify the Computational Ceiling (Theorem 7) directly: $\mathcal{R}^{\text{sym}}_{\text{proc}} = h_\mu \cdot D = 4\,h_\mu > \mathcal{C}_{\text{budget}} = 2\,h_\mu$, confirming that the ceiling binds and symmetry breaking is required. After breaking ($k^* = 2$), $R_{\mathcal{F}} = 2\,h_\mu = \mathcal{C}_{\text{budget}}$: the budget is exactly saturated, as predicted by Theorem 16(a).

6.6 The Pointer-State Connection

The optimal basis choice coincides with the einselection (environment-induced superselection) basis of decoherence theory [14]. This is not a coincidence: the pointer states are precisely those that generate the most redundant records in the environment—i.e., the most predictive correlations. The rate-distortion optimization (Theorem 16) selects the components with the highest survival value per bit, which are generically the pointer-state components.

Remark 25 (Decoherence as symmetry-breaking catalyst)**.** *The environment does not* force *a specific gauge fixing; it merely breaks the degeneracy among possible fixings by making some bases more informationally efficient than others. The agent's bounded computation does the rest: once the degeneracy is broken, the survival imperative (B4) selects the pointer-aligned frame as optimal. This is the precise sense in which decoherence "catalyzes" the spontaneous symmetry breaking of the ego.*

7 The Cost of Ego

The ego resolves the computational crisis of Section 3, but it introduces a new vulnerability. A fixed reference frame is a *static* gauge choice in a *dynamic* environment. If the environment changes, the ego becomes progressively maladaptive.

Drift layer. Environmental change can occur at multiple levels: parameter drift ($\lambda_\alpha(t)$, $\gamma_\alpha(t)$), spectral-density deformation ($J(\omega, t)$), or full process-distribution shift ($P_t(X)$). For analytical tractability, we model drift at the *spectral-density parameter level* throughout this section; the results generalise monotonically to deeper levels (faster drift $\Rightarrow$ shorter t_{del}).

7.1 The Rigidity Trap

Proposition 26 (Frame Rigidity under Drift)**.** *Let the environment undergo slow drift: the spectral density parameters change as $\lambda_\alpha(t) = \lambda_\alpha^{(0)} + \varepsilon\, f_\alpha(t)$ for $\alpha \in \{z, x\}$, with drift rate $\varepsilon > 0$. The optimal reference frame $\mathcal{F}^*(t)$ (the instantaneous minimizer of survival distortion) rotates continuously in the gauge group G.*

If the agent's reference frame $\mathcal{F}$ is held fixed (no recalibration), the mismatch between $\mathcal{F}$ and $\mathcal{F}^*(t)$ grows as

$$\delta(t) := d_G\big(\mathcal{F},\, \mathcal{F}^*(t)\big) \sim \varepsilon \int_0^t \big|\dot{f}(s)\big|\, ds, \tag{35}$$

where d_G is the geodesic distance in the gauge group.

Proof. The instantaneous optimal frame $\mathcal{F}^*(t)$ is a continuous function of the spectral density parameters $\{\lambda_\alpha(t), \gamma_\alpha(t)\}$. Under the drift $\lambda_\alpha(t) = \lambda_\alpha^{(0)} + \varepsilon\, f_\alpha(t)$, the chain rule gives $\dot{\mathcal{F}}^*(t) = \varepsilon \sum_\alpha (\partial \mathcal{F}^*/\partial \lambda_\alpha)\, \dot{f}_\alpha(t)$. Integrating and taking the norm in G gives the bound (35). □

7.2 Stylized Drift Model

To quantify the collapse of a fixed frame, we introduce a minimal drift model that makes the exponential divergence and the logarithmic delusion time algebraically explicit.

Definition 27 (Rotating optimal frame)**.** *Let the mismatch angle $\theta(t)$ between the agent's fixed frame $\mathcal{F}$ and the instantaneous optimal frame $\mathcal{F}^*(t)$ evolve as*

$$\theta(t) = \theta_0\, e^{\Lambda t} \qquad \textit{(chaotic drift)}, \tag{36}$$

where $\theta_0 \in (0, \pi/4)$ is the initial misalignment (so that $t_{\mathrm{del}} > 0$) and $\Lambda > 0$ is the environmental Lyapunov exponent (the rate at which nearby environmental trajectories diverge in spectral-density space). Operationally, Λ is determined by the drift rate ε and the adaptation timescale τ_{adapt} of the spectral-density parameters via the scaling

$$\Lambda \;\sim\; \frac{\varepsilon}{\tau_{\mathrm{adapt}}}\,; \tag{37}$$

cf. (35)*. For slow linear drift ($\theta(t) = \varepsilon\, t$, $\Lambda \to 0$), the crossover time is $t_{\mathrm{del}} = \pi/(4\varepsilon)$ (Remark 30).*

The visible and hidden survival components decompose geometrically:

$$\mathcal{S}_{\text{vis}}(t) = \mathcal{S}_{\text{tot}} \cos^2\theta(t), \qquad \mathcal{S}_{\text{hid}}(t) = \mathcal{S}_{\text{tot}} \sin^2\theta(t), \tag{38}$$

where $\mathcal{S}_{\text{tot}}$ is the full survival functional (invariant under frame rotation).

7.3 The Prediction Error Divergence

Proposition 28 (Divergence of Hidden Survival)**.** *Under the drift model* (36)–(38)*, the hidden survival component grows as*

$$\left|\mathcal{S}_{\text{hid}}(t)\right| = \left|\mathcal{S}_{\text{tot}}\right| \sin^2\!\left(\theta_0\, e^{\Lambda t}\right). \tag{39}$$

For small angles ($\theta_0 e^{\Lambda t} \ll 1$): $|\mathcal{S}_{\text{hid}}| \approx |\mathcal{S}_{\text{tot}}|\, \theta_0^2\, e^{2\Lambda t}$ (exponential growth).

Proof. Direct substitution of (36) into (38). The small-angle expansion $\sin^2\theta \approx \theta^2$ gives the exponential form. □

7.4 The Delusion Trap

Theorem 29 (The Delusion Trap)**.** *Under* (B1)–(B5) *with the drift model* (36) *and initial misalignment $\theta_0 \in (0, \pi/4)$, an agent with a fixed reference frame $\mathcal{F}$ reaches a critical* ***delusion time***

$$t_{\text{del}} = \frac{1}{\Lambda} \ln\!\left(\frac{\pi/4}{\theta_0}\right), \tag{40}$$

beyond which:

(a) $|\mathcal{S}_{\text{hid}}(t)| > |\mathcal{S}_{\text{vis}}(t)|$*: the invisible component dominates the survival functional.*

(b) *The agent's update direction becomes* anti-correlated *with the true optimal direction: the inner product of survival gradients (with respect to the agent's control variables $u \in V_{\text{fg}}$) satisfies*

$$\left\langle \nabla_u \mathcal{S}_{\text{vis}},\ \nabla_u \mathcal{S}_{\text{full}} \right\rangle < 0. \tag{41}$$

Updating u to maximise $\mathcal{S}_{\text{vis}}$ actually decreases *$\mathcal{S}_{\text{full}}$.*

(c) *The agent cannot detect this failure from within its own reference frame, because all four bias terms ($\mathcal{B}_{\text{select}}$, $\mathcal{B}_{\text{frame}}$, $\mathcal{B}_{\text{center}}$, $\mathcal{B}_{\text{inc}}$) operate within V_{fg} and cannot register changes in V_{bg}.*

Proof. Part (a): The crossover $|\mathcal{S}_{\text{hid}}| = |\mathcal{S}_{\text{vis}}|$ occurs when $\sin^2\theta = \cos^2\theta$, i.e., $\theta(t_{\text{del}}) = \pi/4$. Substituting (36): $\theta_0\, e^{\Lambda\, t_{\text{del}}} = \pi/4$, which gives (40). The logarithmic dependence on $1/\theta_0$ means that even a very small initial misalignment ($\theta_0 \sim 10^{-3}$) delays the trap only by $\sim 7/\Lambda$—a modest multiple of the environmental Lyapunov time.

Part (b): Decompose $\mathcal{S}_{\text{full}}$ in the agent's (rotated) coordinates as $\mathcal{S}_{\text{full}}(u, v) = \mathcal{S}_{\text{ff}}(u) + \mathcal{S}_{\text{fb}}(u, v) + \mathcal{S}_{\text{bb}}(v)$, where $u \in V_{\text{fg}}$, $v \in V_{\text{bg}}$, $\mathcal{S}_{\text{ff}}$ depends only on foreground controls, $\mathcal{S}_{\text{bb}}$ only on background, and $\mathcal{S}_{\text{fb}}$ encodes the foreground–background cross-coupling. The projected

kernel $\mathcal{K}_{\mathcal{F}} = \Pi_{\mathcal{F}} \mathcal{K} \Pi_{\mathcal{F}}$ discards $\mathcal{S}_{\text{fb}}$ and $\mathcal{S}_{\text{bb}}$, so $\mathcal{S}_{\text{vis}}(u) = \mathcal{S}_{\text{ff}}(u)$. The foreground gradients are therefore $\nabla_u \mathcal{S}_{\text{vis}} = \nabla_u \mathcal{S}_{\text{ff}}$ and $\nabla_u \mathcal{S}_{\text{full}} = \nabla_u \mathcal{S}_{\text{ff}} + \nabla_u \mathcal{S}_{\text{fb}}$, giving

$$\left\langle \nabla_u \mathcal{S}_{\text{vis}},\ \nabla_u \mathcal{S}_{\text{full}} \right\rangle = \left| \nabla_u \mathcal{S}_{\text{ff}} \right|^2 + \left\langle \nabla_u \mathcal{S}_{\text{ff}},\ \nabla_u \mathcal{S}_{\text{fb}} \right\rangle.$$

The first term is non-negative. Under the rotating drift model, the frame rotation by θ induces cross-coupling that scales as $\sin\theta\cos\theta$ (maximal at $\theta = \pi/4$), while the direct term scales as $\cos^4\theta$. The sign of $\nabla_u \mathcal{S}_{\text{fb}}$ is set by the background state v: since the agent invests no control resources in V_{bg}, v relaxes toward the uncontrolled equilibrium, where the cross-coupling penalises foreground-directed updates ($\langle \nabla_u \mathcal{S}_{\text{ff}},\ \nabla_u \mathcal{S}_{\text{fb}} \rangle < 0$). Beyond t_{del} ($\theta > \pi/4$), the adverse coupling dominates the direct term, yielding (41).

Part (c): The bias terms $\mathcal{B}_{\text{select}}$ through $\mathcal{B}_{\text{inc}}$ (Proposition 18) are defined *within* V_{fg}. The agent's performance metric $\mathcal{S}_{\text{vis}} = \mathcal{S}_{\text{tot}} \cos^2\theta$ decreases only at second order in θ, so it remains positive and shows no anomaly until θ is already $O(1)$. The growing signal in V_{bg} maps to the null space of $\Pi_{\mathcal{F}}$ and is strictly invisible. □

Remark 30 (Linear drift limit)**.** *For slow linear drift ($\theta(t) = \varepsilon\, t$, $\Lambda \to 0$), the crossover occurs at $t_{\text{del}} = \pi/(4\varepsilon)$. With $\varepsilon = 0.01\,\omega_0$, $t_{\text{del}} \approx 79\,\omega_0^{-1}$—long enough for the agent to accumulate a false sense of security, yet short on environmental timescales.*

Remark 31 (Why dithering does not help)**.** *One might ask whether the agent could escape the delusion trap by randomly "probing" the background subspace V_{bg}—temporarily rotating its frame to sample hidden components. This fails for two reasons. First, each probe costs $\sim h_\mu \cdot D$ bits of computation (the Symmetry Tax, Corollary 8), directly competing with the budget allocated to foreground processing. Second—and more fundamentally—the agent has no* gradient signal *to indicate* when *or* where *to probe. As long as $|\mathcal{S}_{\text{hid}}| < |\mathcal{S}_{\text{vis}}|$ (pre-delusion), the in-frame performance metric $\mathcal{S}_{\text{vis}}$ shows no anomaly. The exponential divergence* (39) *is invisible until it dominates—at which point it is too late. Systematic correction requires monitoring the* rate of change *of prediction error, which is a second-order operation: the subject of Paper III.*

Remark 32 (The ego as medicine and poison)**.** *The ego cures computational paralysis (Theorem 7) but creates the delusion trap (Theorem 29). It is simultaneously the* medicine *for Paper I's crisis and the* poison *that generates Paper III's crisis. This duality is a structural consequence of the irreversible logic chain: each resolution creates the conditions for the next crisis.*

7.5 The Origin of Paper III

To escape the delusion trap, the agent needs a mechanism to monitor the quality of its own reference frame—to "observe its own observation." This requires a *second-order control loop*: a meta-controller that adjusts the gauge fixing σ in response to accumulated prediction errors.

The key difficulty is that the prediction errors the agent can measure ($\mathcal{S}_{\text{vis}} - \mathcal{S}_{\text{vis}}^{\text{predicted}}$) all lie within V_{fg}. To detect frame drift, the agent must compare these in-frame errors to an estimate of out-of-frame contributions—a self-referential operation that requires *Fisher information about the agent's own parameters.*

This is the subject of Paper III: the Fisher information geometry of self-referential calibration, and the thermodynamic cost of the loop that closes the chain *Chaos* $\rightarrow$ *Time* $\rightarrow$ *Self* $\rightarrow$ *Calibration*.

8 Numerical Demonstration

The preceding sections establish analytic bounds and a worked example with a qubit in a two-channel bath. We now provide a numerical illustration showing that the core symmetry-breaking signature—attention entropy collapse under budget constraints—and the resulting selection advantage are reproduced in a minimal multi-dimensional system. Full code and parameters are provided for reproducibility.

8.1 Model

Environment. A D-dimensional linear prediction task with sparse rotating support: $y(t) = \mathbf{w}^*(t)^\top \mathbf{x}(t) + \xi(t)$, $\mathbf{x}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$, $\xi \sim \mathcal{N}(0, \sigma^2)$. Only $m \ll D$ dimensions carry nonzero weight at any time; the active support rotates every τ_{switch} steps, modelling environmental drift.

Hard budget constraint. Per step, the agent may update only k coordinates of its weight vector (a hard processing budget), mirroring the bounded computation assumption (B2).

Agents.

- **Budgeted selector (SSB)**: selects the top-k dimensions by importance score—an exponential moving average of the signed per-coordinate gradient. Signed accumulation ensures that noise dimensions (zero expected signal) cancel over time while signal dimensions persist, enabling reliable discrimination without access to the true support.
- **Random-k baseline**: selects k dimensions uniformly at random each step. This provides a budget-fair comparison: identical mechanism, no symmetry breaking.

The choice of *signed* gradient EMA (rather than squared-gradient magnitude) is structurally motivated: for noise dimensions $\mathbb{E}[r\, x_i] = 0$, so the signed accumulation cancels over time; for signal dimensions $\mathbb{E}[r\, x_i] \neq 0$, so a consistent directional bias persists. The signed EMA thus acts as a *directional coherence filter* that discriminates signal from noise without access to the true support—a minimal realisation of the "reference-frame bias" that emerges from symmetry breaking.

Parameters.

Quantity	Value	Role
D	64	ambient dimension
m	8	signal dimensions (sparse support)
T	10,000	horizon per trial
Seeds	10	independent replications
σ	0.3	observation noise std
η	0.02	SGD learning rate
λ	0.995	weight decay per step
k	2, 4, 6, 8, 10, 12, 16, 20, 24, 32, 48, 64	budget grid
τ_{switch}	$\{500, 1000, 2000\}$	support rotation period

Attention entropy. Let n_i be the number of updates coordinate i receives in a measurement window of the last 1,000 steps. The normalised update frequency $p_i = n_i / \sum_j n_j$ defines the attention entropy:

$$H_{\text{attn}} = -\sum_{i=1}^{D} p_i \ln p_i. \tag{42}$$

Under symmetric processing (no SSB), $p_i = 1/D$ and $H_{\text{attn}} = \ln D$. Under budget-constrained selection, H_{attn} collapses away from $\ln D$, serving as an order parameter for symmetry breaking.

Oracle metric. Neither agent has access to $\mathbf{w}^*(t)$. Performance is evaluated externally using the weight-space mean-squared error $\text{MSE} = \|\hat{\mathbf{w}} - \mathbf{w}^*\|^2$, averaged over post-burn-in steps.

8.2 Results

Figure 2 shows the two key signatures.

Result 1: Attention entropy collapse (Figure 2a). Under fixed support (no rotation), the attention entropy H_{attn} exhibits a sharp collapse away from $\ln D = \ln 64 \approx 4.16$ and increases monotonically with k, consistent with a budget-induced concentration of update mass onto signal-carrying dimensions. For budgets near and below the signal scale ($k \leq m$), H_{attn} remains $O(\ln m)$, consistent with confinement to the signal subspace. We use the collapse of H_{attn} away from $\ln D$ as the order parameter of symmetry breaking; a strict plateau at $\ln m$ is not expected under the present re-selection dynamics and finite-window estimator.

Result 2: Selection advantage (Figure 2b). Under rotating support, the mean-squared error gap $\Delta\text{MSE} = \text{MSE}_{\text{rnd}} - \text{MSE}_{\text{sel}}$ is positive for $k \lesssim 3m$ and peaks at tight budgets ($k = 2$) where the selection advantage is strongest. For $k \gg m$ the gap turns slightly negative (the selector's commitment to stale dimensions costs more than the random baseline's diversification), before returning to zero at $k = D$. The three τ_{switch} curves are ordered: slower drift (larger τ) yields a larger peak gap, with the ordering most visible at small k.

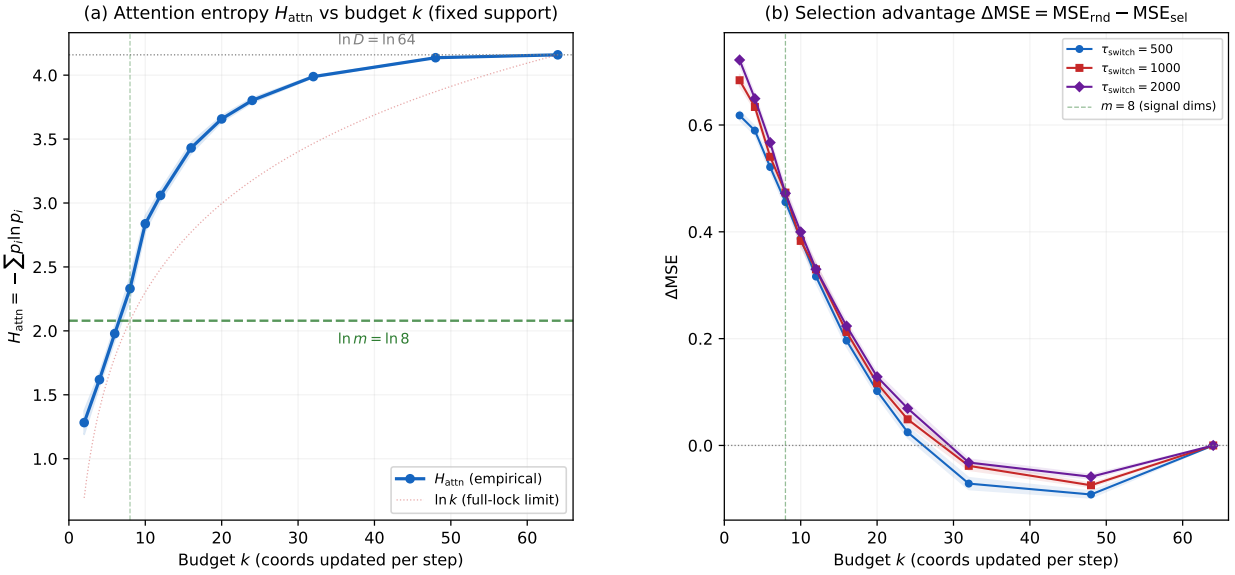

Figure 2: **Budget-induced symmetry breaking.** $D = 64$, $m = 8$, $T = 10{,}000$, 10 seeds, 95% CI bands. **(a)** Attention entropy H_{attn} vs budget k (fixed support). The empirical curve (blue) collapses from $\ln D \approx 4.16$ toward an $O(\ln m)$ floor as budget tightens. For $k \leq m$, H_{attn} remains below $\ln m \approx 2.08$ (green dashed), consistent with confinement to the signal subspace. **(b)** Selection advantage $\Delta\text{MSE} = \text{MSE}_{\text{rnd}} - \text{MSE}_{\text{sel}}$ vs budget k under rotating support. The gap is positive for $k \lesssim 3m$ (selection helps), turns slightly negative at large k (commitment cost exceeds diversification), and returns to zero at $k = D$. Slower drift ($\tau = 2000$) yields a larger peak advantage.

8.3 Scope of This Demonstration

These simulations illustrate the symmetry-breaking phenomenon predicted by Theorem 17 under the stated model class; they do not constitute a proof beyond this class.

This demonstration **does** show:

1. Under hard budget constraints, attention entropy collapses sharply away from $\ln D$ and remains $O(\ln m)$ for $k \leq m$—the agent confines its updates to the signal subspace. This is the computational analogue of spontaneous symmetry breaking (Theorem 17).

2. A budgeted selector that exploits importance-weighted selection systematically outperforms a budget-fair random baseline, consistent with a survival advantage in the broken phase (cf. Proposition 22).

3. The advantage scales with both budget tightness (smaller k) and environmental stability (larger τ_{switch}).

In summary, this demonstration validates the *existence* and *measurability* of budget-induced symmetry breaking in a minimal linear setting; it does not claim universality across architectures or environment classes.

This demonstration does **not** show:

1. That H_{attn} reaches a strict plateau at $\ln m$ for all $k \leq m$. Under the adaptive reselection dynamics used here, the selector cycles within the signal subspace, producing H_{attn} values near but not locked to $\ln m$. The relevant signature is the collapse *away from* $\ln D$, not convergence to a specific lower bound.

2. That the specific form of the importance score (signed gradient EMA) is optimal. It is one realisation of the selection mechanism.

3. That the results generalise to all environment classes. The model uses Gaussian features, linear regression, and sparse rotating support.

4. That the delusion–correction cycle is addressed. This is the subject of Paper III.

Reproducibility. The complete simulation is a self-contained Python script (`paper2_kstar_scaling_dem` ~ 540 lines, requiring only NumPy and Matplotlib) with fixed random seeds. All figures in this section can be reproduced by executing the script. The following files are included in the supplementary archive:

- `paper2_kstar_scaling_demo.py` — simulation script
- `fig_paper2_kstar_scaling.pdf` — Figure 2
- `kstar_scaling_data.csv` — raw performance gap data

9 Discussion

9.1 Summary of Results

Result	**Statement**	**Sec.**
Computational Ceiling	Symmetric processing cost exceeds $\mathcal{C}_{\text{budget}}$ at τ_{par}	3
Rate-Distortion Bound	Optimal compression retains $k^* = \mathcal{C}_{\text{budget}}/h_\mu$ components	4.2
Necessity of SSB	Under bounded computation, survival requires gauge fixing	4.3
Four Bias Terms	Broken phase acquires $\mathcal{B}_{\text{select}}$, $\mathcal{B}_{\text{frame}}$, $\mathcal{B}_{\text{center}}$, $\mathcal{B}_{\text{inc}}$	4.4
Survival Decomposition	$\mathcal{S} = \mathcal{S}_{\text{vis}} + \mathcal{S}_{\text{hid}}$	5.2
Ego-Entropy Trade-off	$\gtrsim 1 - k^*/\dim Cl(V,q)$ of I_{pred} discarded (uniformity assumption)	5.4
Delusion Trap	Fixed frame diverges from optimal under environmental drift; agent cannot self-detect	7.4
Numerical demo	Budget-induced SSB and selection advantage (Fig. 2)	8

9.2 What This Paper Does and Does Not Show

This paper **does** show:

1. Under bounded computation (B2) and non-trivial environment (B3), symmetric processing of memory leads to computational paralysis (Theorem 7).
2. The survival-optimal response is spontaneous symmetry breaking of the internal reference frame (Theorem 17), governed by a rate-distortion bound (Theorem 16).
3. The broken phase acquires four generic bias terms under (B1)–(B5) (Proposition 18).
4. Under environmental drift, a fixed frame leads to exponential divergence of prediction error—the Delusion Trap (Theorem 29).
5. A minimal computational demonstration reproduces the budget-induced symmetry-breaking signature (attention entropy collapse away from $\ln D$) and selection advantage over a budget-fair random baseline (Section 8, Figure 2).

This paper does **not** show:

1. That the privileged basis is uniquely determined by computational constraints. The basis is constrained but not unique—different histories lead to different gauge fixings, as in a ferromagnet.
2. That symmetry breaking is sufficient for persistence. It is the survival-optimal strategy under bounded computation; sufficiency requires the self-referential calibration of Paper III.
3. That the "ego" implies or requires consciousness, subjective experience, or phenomenal awareness. The term is used strictly in the control-theoretic sense.
4. That this framework constitutes a theory of consciousness. It is a theory of computational optimality under thermodynamic constraints.
5. That the four bias terms exhaust the phenomenology of self-reference. They are the minimal structural consequences of gauge fixing in a Clifford algebra.
6. That the rate-distortion bound is achievable by any specific physical implementation. It is an information-theoretic lower bound.
7. That the Delusion Trap is inescapable. Paper III will show it can be mitigated by self-referential calibration.
8. That the framework constitutes or implies a philosophical or metaphysical claim about the nature of selfhood.
9. That this framework applies to all possible physical systems. It applies to systems satisfying (B1)–(B5)—persistent agents with finite computation in non-trivial environments.
10. That the Clifford algebra is the only possible algebraic setting. It is the minimal setting inherited from Q-RAIF. Other algebras may yield analogous results.

References

[1] S. Liu, *Non-Markovian Memory and the Thermodynamic Necessity of Temporal Accumulation*, Zenodo (2026), DOI: 10.5281/zenodo.18574342.

[2] H.-P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems*, Oxford University Press (2002).

[3] J. P. Crutchfield and K. Young, *Inferring statistical complexity*, Phys. Rev. Lett. **63**, 105 (1989).

[4] C. R. Shalizi and J. P. Crutchfield, *Computational mechanics: Pattern and prediction, structure and simplicity*, J. Stat. Phys. **104**, 817 (2001).

[5] W. Bialek, I. Nemenman, and N. Tishby, *Predictability, complexity, and learning*, Neural Computation **13**, 2409 (2001).

[6] C. E. Shannon, *Coding theorems for a discrete source with a fidelity criterion*, IRE Nat. Conv. Rec., Part 4, pp. 142–163 (1959).

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley (2006).

[8] H. A. Simon, *A behavioral model of rational choice*, Quarterly J. of Economics **69**, 99 (1955).

[9] N. Tishby, F. C. Pereira, and W. Bialek, *The information bottleneck method*, in Proc. 37th Allerton Conf. on Communication, Control, and Computing (1999); arXiv:physics/0004057 (2000).

[10] C. A. Sims, *Implications of rational inattention*, J. Monetary Economics **50**, 665 (2003).

[11] W. B. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, Contemp. Math. **26**, 189 (1984).

[12] R. Landauer, *Irreversibility and heat generation in the computing process*, IBM J. Res. Dev. **5**, 183 (1961).

[13] C. H. Bennett, *The thermodynamics of computation—a review*, Int. J. Theor. Phys. **21**, 905 (1982).

[14] W. H. Zurek, *Quantum Darwinism*, Nature Physics **5**, 181 (2009).

[15] Á. Rivas, S. F. Huelga, and M. B. Plenio, *Quantum non-Markovianity: characterization, quantification and detection*, Rep. Prog. Phys. **77**, 094001 (2014).

[16] S. Liu, *Emergent Geometry from Coarse-Grained Observable Algebras*, Zenodo (2026), DOI: 10.5281/zenodo.18361706.

[17] S. Liu, *Accessibility, Stability, and Emergent Geometry*, Zenodo (2026), DOI: 10.5281/zenodo.18367060.

[18] S. Liu, *Structural Limits of Unification: Accessibility, Incompleteness, and the Necessity of a Final Cut*, Zenodo (2026), DOI: 10.5281/zenodo.18402907.

[19] S. Liu, *Algebraic Constraints on the Emergence of Lorentzian Metrics in Entropic Gravity Frameworks*, Zenodo (2026), DOI: 10.5281/zenodo.18525876.

[20] S. Liu, *Thermodynamic Stability Constraints on the Operator Algebra of Persistent Open Quantum Subsystems*, Zenodo (2026), DOI: 10.5281/zenodo.18525890.

[21] S. Liu, *The Realizability Bridge: Algebraic Closure in the Q-RAIF Framework*, Zenodo (2026), DOI: 10.5281/zenodo.18528934.