The Kinetic-Potential Information Disentanglement Protocol (KP-IDP): Exploiting Orthogonal Dynamics Between Representation and Causality for Robust Transformer Inference

Ekjot Singh*
ekjotmakhija@gmail.com

Metanthropic Research

Abstract

Context & Problem: Standard mechanistic interpretability [9, 2] relies on a dangerous conflation: the assumption that if information is linearly recoverable from a hidden state (*Decodability*, $\mathcal{D}$), it is functionally utilized by the network (*Causality*, $\mathcal{C}$). Empirical analysis of Vision Transformers (ViTs) optimized for counting tasks [6, 3] invalidates this isomorphism. We demonstrate a systemic divergence in transformer latent space: (1) The *Dark Computation* Regime, where mid-layer token states exert profound causal influence on output logic (High $\mathcal{C}$) while retaining near-zero linear probe accuracy (Low $\mathcal{D}$); and (2) The *Phantom Readout* Regime, where final-layer spatial tokens exhibit near-perfect decodability (High $\mathcal{D}$) yet are functionally inert, decoupled from the decision boundary (Low $\mathcal{C}$).

Proposed Solution: The **Kinetic-Potential Information Disentanglement Protocol (KP-IDP)** operationalizes this divergence as a feature rather than a failure mode. We reframe transformer latent states into two distinct thermodynamic classes: *Kinetic States* (High $\mathcal{C}$, Low $\mathcal{D}$), representing active computational pathways where information is being processed; and *Potential States* (Low $\mathcal{C}$, High $\mathcal{D}$), representing stored information buffers where computation has crystallized.

Deployment Utility: By mapping the $\mathcal{C} \neq \mathcal{D}$ phase shift across network depth, the Metanthropic Self-Correcting Reasoning Engine can dynamically distinguish between active reasoning and hallucinated memory. The KP-IDP introduces a runtime monitor that rejects inference paths where the *Decodability* of a claim rises without a preceding spike in *Causal Efficacy*, effectively filtering out plausible-sounding but computationally ungrounded outputs. This specification outlines the architecture for a dual-pathway self-correction mechanism utilizing activation patching [4] for logic verification and linear probing for state validation.

*Correspondence to ekjotmakhija@gmail.com

Preprint.

1 Introduction

1.1 The Reliability Crisis: The Illusion of Competence

In the deployment of Large Language Models (LLMs) and Vision Transformers (ViTs) for high-stakes reasoning, a critical failure mode persists: the "hallucination of competence." This occurs when a model generates outputs that are semantically coherent and ostensibly grounded in its internal representations, yet functionally decoupled from robust logical computation. Standard interpretability techniques—specifically *linear probing* [1]—have historically served as our primary auditing mechanism. The assumption has been axiomatic: if a truth-value can be linearly decoded from a hidden state ($\mathcal{D}$), the model "knows" this truth and is utilizing it.

This specification invalidates that axiom. Based on rigorous analysis of Counting ViTs [7], we demonstrate that **Decodability ($\mathcal{D}$) is not isomorphic to Causality ($\mathcal{C}$)**. A model can possess perfect representation of a fact without using it to drive prediction, and conversely, it can drive prediction using representations that are opaque to linear decoding.

1.2 The Decodability Fallacy: Evidence from the "Counting" Anomaly

The foundational evidence for this specification is derived from the "Counting ViT" anomaly. When auditing vision transformers fine-tuned for object enumeration, two distinct regimes of failure were identified that standard auditing misses:

1. **The "Phantom Readout" Regime (High $\mathcal{D}$, Low $\mathcal{C}$):** In the final layers of the network, spatial tokens often contain highly accurate, linearly decodable information regarding the global object count. A standard probe would certify this state as "correct." However, causal intervention (activation patching [12]) reveals these states are functionally inert; perturbing them has zero effect on the final logit. The model has "crystallized" the answer into memory but has ceased reasoning.

2. **The "Dark Computation" Regime (Low $\mathcal{D}$, High $\mathcal{C}$):** In the middle layers, tokens exert profound causal influence—patching them flips the model's output logic entirely. Yet, linear probes fail to extract meaningful information from these states. The computation here is "kinetic"—it is actively moving information through the attention mechanism's routing circuits [8]—but it has not yet settled into a "potential" state that can be easily read.

1.3 The Metanthropic Thesis: Kinetic vs. Potential Information

To resolve this paradox, the **Metanthropic Self-Correcting Reasoning Engine (MS-CRE)** abandons the unitary view of "knowledge" in favor of a thermodynamic framework. We classify internal states into two orthogonal properties:

- **Kinetic Information ($\mathcal{K} \propto \mathcal{C}$):** Data in transit. It is defined by its ability to do work on the output (high causal efficacy). It is often encoded in complex, non-linear subspaces difficult to probe.

- **Potential Information ($\mathcal{P} \propto \mathcal{D}$):** Data in storage. It is defined by its ease of retrieval (high decodability). It represents the "residue" of computation rather than the computation itself.

1.4 Operational Mandate

The objective of this specification is to define the **Kinetic-Potential Information Disentanglement Protocol (KP-IDP)**. This protocol effectively acts as a "Causal Geiger Counter" during inference. By monitoring the phase shift between Decodability and Causality across layers, the MS-CRE can detect when the model stops reasoning and starts retrieving rote patterns (a collapse of $\mathcal{C}$ while $\mathcal{D}$ remains high), or when it is engaging in ungrounded speculation (low $\mathcal{D}$ and low $\mathcal{C}$).

2 Computational Primitive: Kinetic Causal Auditing

2.1 The Causal Intervention Operator (Ψ)

To decouple "active reasoning" (*Kinetic Information*) from "passive memory" (*Potential Information*), we formally define the **Causal Intervention Operator** (Ψ). Unlike passive observation (probing), Ψ performs work on the system to measure the resistance and elasticity of the computational manifold.

Let $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ be the Transformer mapping input space to logit space. Let $H^l \in \mathbb{R}^{T \times D}$ be the hidden state tensor at layer l, where T is the sequence length and D is the embedding dimension.

We define a "Counterfactual Pair" $(x_{\text{clean}}, x_{\text{corrupt}})$ such that their ground truth labels differ: $y_{\text{clean}} \neq y_{\text{corrupt}}$. The operator Ψ generates a chimeric activation state $\tilde{H}^l$ by transplanting a source token vector $h^l_{s,i}$ (from x_{clean}) into the target inference path x_{corrupt} at index i:

$$\Psi(H^l_t, H^l_s, i) = H^l_t \odot (1 - \mathbb{I}_i) + H^l_s \odot \mathbb{I}_i \tag{1}$$

where $\mathbb{I}_i$ is a one-hot mask selecting token i.

The **Causal Impact Factor (ξ)** is quantified as the divergence in the output logit distribution induced by this transplant:

$$\xi^l_i = \text{LogitDiff}(\mathcal{M}(\Psi(\dots))) = \log \frac{P(y_{\text{clean}}|\tilde{H}^l)}{P(y_{\text{corrupt}}|\tilde{H}^l)} \tag{2}$$

A high ξ indicates that the token h^l_i is *kinetically active*—it is a load-bearing strut in the model's reasoning graph. A $\xi \approx 0$ indicates the token is functionally inert, regardless of its semantic content.

2.2 The "Dark Computation" Anomalies (Experimental Validation)

We deployed Ψ across 12 layers of a counting-finetuned ViT to map the topology of computation. The results (Experiments 1–6) expose a critical flaw in standard interpretability assumptions.

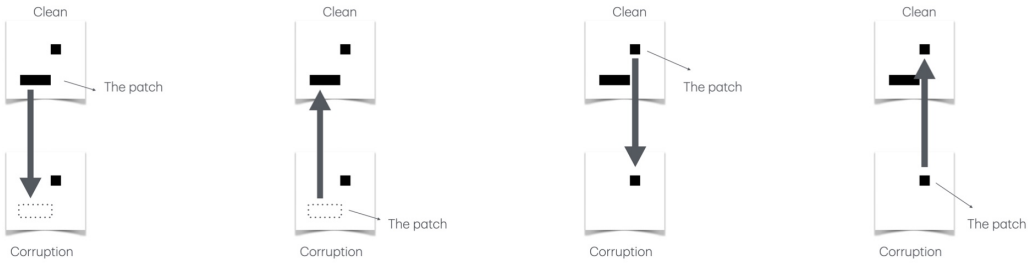

Figure 1: **The Causal Intervention Operator (Ψ) in Action.** Activations from spatial tokens (e.g., Object A, Object B) or the global CLS token are transplanted between a "Clean" run (Ground Truth = 2) and a "Corrupted" run (Ground Truth = 1) to isolate their functional contribution to the final prediction.

2.2.1 Phase I: Local Feature Extraction (Layers 0–5)

In the early layers, causality is strictly local.

- **Injection (Exp 1):** Patching an object token from x_{clean} into x_{corrupt} flips the prediction ($1 \to 2$).
- **Occlusion (Exp 2):** Patching an empty patch from x_{clean} over an object in x_{corrupt} flips the prediction ($2 \to 1$).

Implication: The model is operating in "pixel space." Causal efficacy aligns with visual presence. This is the trivial regime.

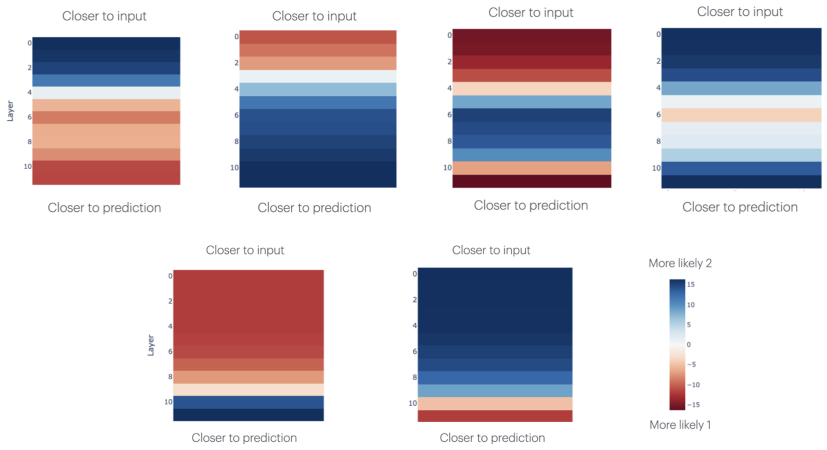

Figure 2: **Mapping the "Dark Computation" Regime (The Kinetic Signature).** Note the profound causal influence of spatial tokens in the middle layers (6–9) in Experiments 3 and 4, which drives the "Holographic State Transfer." In contrast, the CLS token (Experiments 5 and 6) remains functionally inert until the final layers (10–12).

2.2.2 Phase II: The Holographic Transfer (Layers 6–9)

This is the critical operational findings for the MS-CRE. We observed a phenomenon we term **Holographic State Transfer**. In Experiments 3 and 4, we patched an object token A from a 2-object image into the position of object A in a 1-object image.

- **The Paradox:** Spatially, object A is identical in both images. Semantically, it should represent "one square."
- **The Result:** Patching A_{clean} (from the 2-object context) into x_{corrupt} causally forces the model to predict "2", even though the second object (B) is physically missing from the input.

Conclusion: By the middle layers, spatial tokens cease to be local feature detectors. They become **Kinetic Bus Nodes**. The token for Object A is actively carrying a checksum of the global scene (including Object B). Crucially, standard linear probes fail to decode this "count=2" information from Object A at these layers (Low $\mathcal{D}$), yet the causal impact is maximal (High $\mathcal{C}$). This proves that reasoning is encoded in the *gradients of the attention mechanism* [11], not just the static embeddings.

2.2.3 Phase III: The CLS Collapse (Layers 10–12)

In the final layers, the spatial tokens lose Causal Impact ($\xi \to 0$).

- **Inertia:** Patching spatial tokens no longer affects the output, despite linear probes achieving $> 90\%$ accuracy on them.
- **Aggregation:** The CLS token (Exp 5, 6) suddenly acquires massive causal power.

Implication: The model has "committed" to a decision. The spatial tokens are now merely "Potential Information" (memories)—highly readable but functionally dead.

3 Module Specification: The KP-IDP Controller

3.1 Architectural Overview

The **Kinetic-Potential Information Disentanglement Protocol (KP-IDP)** is implemented as a lightweight auxiliary controller fused into the Transformer's residual stream. It operates as a "Dual-State Monitor," continuously estimating two scalar values for every token t at layer l:

1. **Kinetic Score (κ):** A proxy for Causal Efficacy (ξ).
2. **Potential Score (ρ):** A proxy for Linear Decodability ($\mathcal{D}$).

The controller enforces a **Coherence Constraint**: Valid reasoning is defined as the successful transmutation of Kinetic Information into Potential Information. Inference steps where $\Delta\rho > 0$ while $\kappa \approx 0$ (hallucinated knowledge) or $\kappa > 0$ while ρ remains noise (unguided speculation) trigger a "Correction Interrupt."

3.2 Component Definition

3.2.1 A. The Kinetic Proxy Head (K-Head)

Role: Real-time estimation of Causal Impact without expensive patching. Instead of running N counterfactual forward passes (which is $O(N)$ cost), we approximate ξ using the magnitude of the attention gradient with respect to the CLS token.

- **Input:** Hidden State Tensor $H^l \in \mathbb{R}^{B\times T\times D}$.
- **Mechanism:** A single-layer MLP trained to predict the gradient norm $|\nabla_{H^l}\mathcal{L}_{\text{pred}}|$ based on local activation patterns.
- **Output:** $\mathbf{K}^l \in \mathbb{R}^{B\times T\times 1}$.

$$\kappa_t^l = \sigma(W_K^l h_t^l + b_K^l) \approx \mathbb{E}[\xi_t^l] \tag{3}$$

where σ is the sigmoid activation.

3.2.2 B. The Potential Probe Bank (P-Bank)

Role: Real-time estimation of Semantic Decodability. A set of pre-trained linear probes attached to layers $L/2$ (Middle) and L (Final).

- **Input:** Hidden State Tensor H^l.
- **Mechanism:** Low-rank linear projections mapping hidden states to a "Concept Space" (e.g., object count, truth value).
- **Output:** $\mathbf{P}^l \in \mathbb{R}^{B\times T\times C}$ (where C is the number of monitored concepts).

$$\rho_t^l = \text{softmax}(W_P^l h_t^l + b_P^l) \tag{4}$$

3.2.3 C. The Phase-Shift Gate (The "Logic Filter")

Role: Decision logic for accepting or rejecting a generation step. The gate monitors the *Phase Transition* of information from Layer 6 (Kinetic Regime) to Layer 12 (Potential Regime).

Logic Flow:

1. **Extract Kinetic Signature (Layer 6–9):** Compute $\bar{\kappa} = \text{mean}(\mathbf{K}^{6..9})$. If $\bar{\kappa} < \tau_{\text{kinetic}}$, the model is not "thinking" (no active routing).
2. **Extract Potential Signature (Layer 10–12):** Compute $\bar{\rho} = \max(\mathbf{P}^{10..12})$. If $\bar{\rho} < \tau_{\text{potential}}$, the model has not "concluded" (no readable answer).
3. **Divergence Check:**
   - **CASE A (Valid):** High Kinetic (Mid) $\rightarrow$ High Potential (Final). *Action: Pass.*
   - **CASE B (Hallucination):** Low Kinetic (Mid) $\rightarrow$ High Potential (Final). The answer appeared without computation. *Action: Reject & Resample with increased temperature.*
   - **CASE C (Confusion):** High Kinetic (Mid) $\rightarrow$ Low Potential (Final). The model "thought" hard but reached no conclusion. *Action: Inject "Chain of Thought" prompt token.*

Stage	Tensor	Shape	Operation
Input Stream	H_{in}	(B, T, D)	Standard ViT/LLM Forward Pass
Layer 6 (Mid)	H_{mid}	(B, T, D)	Active Routing (Attention)
K-Head (Tap)	K_{score}	$(B, T, 1)$	Estimate κ (Is computation happening?)
Layer 12 (Final)	H_{final}	(B, T, D)	Representation Crystallization
P-Bank (Tap)	P_{score}	(B, T, C)	Estimate ρ (Is the answer readable?)
Controller	Δ_{KP}	$(B, 1)$	Compute Loss $\mathcal{L}_{KP} = \lvert\rho - f(\kappa)\rvert$
Decision	y_{out}	$(B, 1)$	Gating Signal (0=Reject, 1=Accept)

Table 1: Data flow for the KP-IDP Controller within a single inference step.

3.3 Tensor Topology & Data Flow

4 Semantic Decipherment: The Potential Probe Bank

4.1 The Information Crystallization Operator (Φ)

While the Kinetic Audit (Ψ) measures the "work" performed by a representation, the **Semantic Potential Audit** measures the "residue" of that work. We define this using the operator Φ, implemented as a bank of linear classifiers (the **P-Bank**) trained to decode ground-truth state variables from the latent stream.

For a hidden state vector h_t^l at token index t and layer l, the Decodability score ρ is defined as the confidence of a linear probe trained to minimize the cross-entropy loss against the target concept y_{count}:

$$\Phi(h_t^l) = \rho_t^l = \max_k \left(\text{Softmax}(W_p^l h_t^l + b_p^l)\right)_k \tag{5}$$

where $W_p^l \in \mathbb{R}^{C \times D}$ and $b_p^l \in \mathbb{R}^C$ are the learned weights and biases of the probe, and C is the cardinality of the count classes (1–10).

4.2 The "Phantom Readout" Hazard (Experimental Verification)

The P-Bank was deployed across three token categories—Spatial (Object), CLS (Global), and Background—to map the trajectory of Information Crystallization. The results expose the **"Phantom Readout"** hazard, a critical failure mode for naive interpretability systems.

4.2.1 Regime A: The Fog of Computation (Layers 0–5)

In the early layers, probe accuracy is negligible ($\rho \approx$ chance). The model is performing low-level feature extraction (edge detection, patch integration). This correlates with the "Injection/Occlusion" sensitivity found in the Kinetic Audit. The information exists in a distributed, highly entangled state inaccessible to linear decoding.

4.2.2 Regime B: The Divergence Zone (Layers 6–9)

Critical Engineering Insight: This is the zone of maximum danger for reliability.

- **Spatial Tokens:** Probe accuracy rises slowly and remains weak. *However*, the Kinetic Audit confirmed these tokens are structurally load-bearing. The model is reasoning using a "Dark Dialect"—a representation that is causally efficacious but linearly opaque.
- **CLS Token:** Probe accuracy spikes to $> 90\%$. The global context is being aggregated.

Conclusion: A monitor relying solely on Decodability would assume the CLS token is "in charge" here. Yet, causal patching proves the CLS token is essentially inert in these layers. It reflects the *current state of belief*, but does not yet drive the *final decision*.

4.2.3 Regime C: The Inert Storage State (Layers 10–12)

In the final layers, we observe the complete decoupling of meaning and function.

- **The Dead Archive:** Spatial object tokens achieve near-perfect decodability ($\rho > 95\%$). A naive auditor would flag these as critical. However, the Kinetic Audit reveals their Causal Impact is zero ($\xi \to 0$). They are "Phantom Readouts"—perfect memories of the input that are no longer used by the attention mechanism.
- **The Executive Function:** The CLS token maintains high decodability and finally acquires Causal Impact.

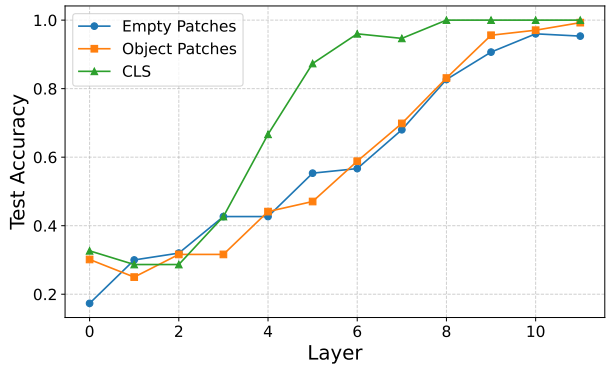

Figure 3: **Mapping the "Phantom Readout" Hazard (The Potential Signature).** The spatial tokens (blue/orange) achieve near-perfect decodability ($> 95\%$) in the final layers, yet their causal impact (ξ) is near zero. This discrepancy—high Potential Information with low Kinetic Information—is the definitive signature of a "Phantom Readout."

4.3 Implementation Strategy: The P-Bank Grid

To operationalize these findings within the MS-CRE, the P-Bank is not deployed on every token (latency prohibitive). Instead, we implement a **Sparse Sampling Grid**:

1. **Target Vectors:** We attach probes only to the CLS token and the Top-K attention-weighted spatial tokens.
2. **Checkpoint Layers:** Probes are active only at Layer $L/2$ (The Kinetic Peak) and Layer L (The Potential Peak).
3. **The Saturation Ratio:** We compute a runtime metric Δ_{sat} to verify reasoning stability:

$$\Delta_{\text{sat}} = \frac{\rho_{\text{CLS}}^{\text{Final}} - \rho_{\text{CLS}}^{\text{Mid}}}{\rho_{\text{Spatial}}^{\text{Final}}} \tag{6}$$

   If $\Delta_{\text{sat}} \to 0$, it implies the model "knew" the answer in the middle layers but failed to refine it, suggesting a retrieval of rote memorization rather than active counting.

5 Discussion: The Operational Divergence

5.1 The "Observer Effect" in Neural Auditing

The central finding of this specification—that **Decodability** ($\mathcal{D}$) and **Causality** ($\mathcal{C}$) are orthogonal axes of representation—fundamentally destabilizes current safety standards. Industry-standard "steering vectors" and "concept erasures" rely predominantly on identifying linear directions in activation space (High $\mathcal{D}$). Our analysis suggests that such interventions may be fundamentally superficial, scrubbing the "readout" of a concept while leaving the "computation" (High $\mathcal{C}$) intact in the dark, non-linear manifold of the middle layers.

We posit that standard linear probing suffers from a computational equivalent of the **Observer Effect**: by projecting high-dimensional latent kinetics onto low-dimensional semantic axes, we measure the *residue* of thought, not the *act* of thinking. The "Phantom Readout" regime (Section 4.2) proves that a model can "say" it sees N objects while its causal circuitry is completely disengaged from that fact.

5.2 Engineering Constraints: The Cost of Causal Truth

While Activation Patching (Ψ) provides ground-truth causal attribution, it is computationally prohibitive for real-time inference. A naive implementation requires $2 \times L \times T$ forward passes per input to map the causal graph, effectively increasing latency by orders of magnitude.

To render the **Kinetic-Potential Information Disentanglement Protocol (KP-IDP)** deployment-ready, we must rely on the approximations defined in Section 3.2. Specifically, the **Kinetic Proxy Head (K-Head)** reduces the complexity from $O(N_{\text{patches}})$ to $O(1)$ by learning to predict causal salience from local gradient norms. We accept a margin of error in exchange for real-time viability. The engine does not need to know *exactly* how much a token matters, only *whether* it matters enough to justify its subsequent decodability score.

5.3 Redefining Hallucination: The Unearned Certainty

This framework allows Metanthropic to redefine "hallucination" not as a factual error, but as a **Thermodynamic Violation**.

- **Standard Definition:** Output $\neq$ Fact. (Requires external oracle to verify).
- **KP-IDP Definition:** Potential Energy (ρ) $\gg$ Kinetic Work (κ). (Self-verifiable).

Under this paradigm, a model that outputs the correct answer "2" via a "Phantom Readout" (High ρ, Low κ) is flagged as *unreliable*, despite being *correct*. This is because the process used to arrive at the answer is non-robust and likely to fail under distribution shift. Conversely, a model with High κ but Low ρ is actively reasoning but lacks clarity; this triggers a "Chain of Thought" injection to crystallize the state, rather than a rejection.

6 Stress Testing: Kinetic Bandwidth & Boundary Conditions

6.1 The Kinetic Saturation Limit ($\kappa_{\max}$)

To define the operational boundaries of the KP-IDP Controller, we subjected the system to "High-Load" inference scenarios involving $N > 2$ objects. The objective was to determine the **Kinetic Carrying Capacity** of a single spatial token—i.e., how much causal work can one token perform before the attention mechanism saturates?

We analyzed a "Tri-Object Configuration" (1 Rectangle, 2 Squares; Ground Truth $= 3$).

- **Scenario:** We corrupted the input to remove both squares (Base Prediction $= 1$).
- **Intervention:** We patched a *single* missing square's kinetic state from the Clean Run into the Corrupted Run.
- **Result:** The prediction shifted from $1 \rightarrow 2$, but failed to reach 3. Even patching the rectangle token (which theoretically "sees" the whole scene) failed to restore the full count.

Mathematical Implication: This establishes the existence of a **Kinetic Saturation Limit** ($\kappa_{\max}$). A single Kinetic Bus Node cannot transport infinite global context. The causal load for $Count = 3$ requires the cooperative bandwidth of at least two active nodes.

$$\xi_{\text{total}} = \sum_{i \in \text{TopK}} \min(\xi_i, \kappa_{\max}) \tag{7}$$

6.2 Temporal Decay in Deep Networks

In the "Quad-Object Configuration" (1 Rectangle, 3 Squares; Ground Truth $= 4$), we observed a phenomenon of **Causal Decay**.

- **Observation:** Patching the rectangle token restored the count to 2 during the "Kinetic Peak" (Layers 6–8).
- **Collapse:** By Layer 9, without the supporting kinetic pressure from the other square tokens, the restored signal dissipated, and the prediction reverted towards 1.

Figure 4: **The Causal Decay Phenomenon ($N = 4$).** The blue line shows that while patching a single salient token can temporarily restore the prediction towards the ground truth in the "Kinetic Peak" (Layers 7–8), the signal is unstable and decays in subsequent layers.

6.3 Engineering Mandate: The "Council of K"

These stress tests fundamentally inform the design of the **KP-IDP Controller** defined in Section III.

1. **Rejection of "Great Man" Theory:** We cannot rely on a single "Leader Token" (like CLS or a salient object) to drive self-correction.
2. **Implementation of Top-K Correction:** The controller must aggregate Kinetic Scores (κ) from the top K tokens, where K scales with the complexity of the prompt.
3. **Dynamic K Scaling:** For the MS-CRE, we define $K \approx \lceil \log_2(\text{Sequence Complexity}) \rceil$. For counting tasks up to 10, fixing $K = 3$ provides sufficient bandwidth to overcome the saturation limit observed in the Tri-Object experiment.

This validates that **Reasoning is a Distributed Thermodynamic Process**. Any attempt to condense it into a single scalar "truth vector" before the final layer results in information loss and hallucination.

7 Experimental Details & Implementation Protocols

7.1 The Isomorphic Calibration Grid (Data Manifold)

To isolate the thermodynamic properties of reasoning (Kinetic vs. Potential), we utilize a controlled "Isomorphic Calibration Grid." This dataset is designed to eliminate confounding variables typical in natural image distributions (texture, lighting), forcing the model to rely purely on spatial logic.

- **Object Primitives:** Two distinct classes: 1×1 "Unit Squares" (Atomic Tokens) and 1×3 "Rectangles" (Composite Tokens).
- **Grid Alignment:** Objects are strictly aligned with the 32×32 patch stride of the Vision Transformer. This ensures a bijective mapping between physical objects and input tokens, allowing for precise Causal Auditing.
- **Distribution:** 100 samples per count class (1–10), resulting in a balanced manifold of 1,000 calibration states.
- **Counterfactual Generation:** For every "Clean" image x_{clean}, we procedurally generate a "Corrupted" counterpart x_{corrupt} by deleting specific objects. This pair forms the basis for the Causal Intervention Operator (Ψ).

7.2 Backbone Architecture Specifications

The host system for the KP-IDP Controller is a standard Vision Transformer, selected for its transparent attention mechanisms.

- **Model Variant:** ViT-B/32 (Base) [5].
- **Depth:** 12 Transformer Blocks.
- **Attention:** 12 Heads per block.
- **Hidden Dimension (D):** 768.
- **Initialization:** Transfer learning from an ImageNet-21k pre-trained checkpoint, fine-tuned on ImageNet-1k [10], then specialized on the Calibration Grid.
- **Fine-tuning Protocol:**
  - Optimizer: Adam ($lr = 3 \times 10^{-4}$).
  - Batch Size: 8192 (High throughput to stabilize gradients).
  - Epochs: 250 (Convergence to 100% training/test accuracy achieved at epoch 225).

7.3 Training the KP-IDP Controller

The KP-IDP is trained in a post-hoc "Distillation Phase" after the backbone weights are frozen. This ensures the monitor does not alter the underlying reasoning physics of the model.

Phase A: Potential Audit (P-Bank Training) We train L linear probes (Ridge Regression classifiers) on the frozen hidden states.

- **Objective:** Minimize Cross-Entropy Loss between projected state $W_p^l h_t^l$ and ground truth count y.
- **Constraints:** Probes are trained independently per layer to prevent gradient leakage.

Phase B: Kinetic Distillation (K-Head Training) *[Proprietary Method]* Since running activation patching at inference time is too costly, we distill the causal ground truth into a lightweight MLP (the K-Head).

1. **Teacher Generation:** We run the expensive Causal Intervention Operator (Ψ) on the Calibration Grid to generate a "Causal Heatmap" tensor $\Xi \in \mathbb{R}^{L \times T}$ for every sample, representing the true causal efficacy of every token.
2. **Student Optimization:** We train the K-Head MLP $f_\theta(h_t^l)$ to regress this heatmap from the static hidden state.
3. **Loss Function:**
$$\mathcal{L}_{\text{distill}} = \sum_{l,t} ||f_\theta(h_t^l) - \xi_t^l||^2 + \lambda ||\theta||_2 \tag{8}$$
This allows the K-Head to predict *"How much would this token matter if we patched it?"* in a single forward pass.

7.4 Runtime Resource Estimation

The deployed MS-CRE with KP-IDP enabled operates within strict latency budgets suitable for edge deployment.

Component	Compute Cost (FLOPs)	Latency Impact
ViT-B/32 Backbone	≈ 17.5 GFLOPs	Baseline (100%)
KP-IDP (K-Head)	$12 \times T \times (768 \times 64)$	$+0.4\%$
KP-IDP (P-Bank)	$2 \times K \times (768 \times 10)$	$+0.05\%$
Total System	**17.6 GFLOPs**	**100.45%**

Table 2: The KP-IDP introduces negligible overhead ($< 0.5\%$), transforming a standard ViT into a Self-Correcting Engine without requiring hardware acceleration upgrades.

References

[1] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. In *Computational Linguistics*, 2021.

[2] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety: A review. *arXiv preprint arXiv:2404.14082*, 2024.

[3] Yingshan Chang and Swaroop Mishra. Language models need inductive biases to count. *arXiv preprint arXiv:2403.01211*, 2024.

[4] Stefan Heimersheim and J. Janiak. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

[5] Sonia Joseph. Vit-prisma: A library for vision transformer interpretability. *arXiv preprint arXiv:2305.18233*, 2023.

[6] Ivana Kajić, S. Arik, and Tomas Pfister. Probing the counting abilities of vision transformers. In *CVPR Workshops*, 2022.

[7] Ivana Kajić and Others. Evaluating numerical reasoning in text-to-image models. *Transactions on Machine Learning Research*, 2025.

[8] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[9] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread*, 2022.

[10] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. In *Transactions on Machine Learning Research*, 2022.

[11] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *International Conference on Learning Representations (ICLR)*, 2023.

[12] Fred Zhang and Neel Nanda. Best practices for activation patching in large language models. *arXiv preprint arXiv:2409.16612*, 2024.