# HybridKV (HKV): Breaking the Memory Wall with O(1) Complexity

## Objective

Break the Transformer Memory Wall.

- Demonstrate a bounded-memory O(1) inference architecture enabling long-context reasoning on power-constrained edge devices.
- Validate that physics-informed state-evolution memory can replace the linearly growing KV cache while maintaining high-fidelity recall for mission-critical tasks.

## Approach

- **Symplectic State Module (SSM):** Replaces dissipative KV cache with a conserved, fixed-width dynamical state (Patent Pending).
- **Pointer-Scatter Retrieval:** O(1) token reinstatement mechanism bypassing vocabulary projection for rapid, low-power recall.
- **Hybrid Architecture:** Dual-gate mechanism combining long-term symplectic state evolution with minimal local context for syntactic precision.

## Payoff

- Infinite Context / Fixed RAM: Decouples inference cost from sequence length.
- SWaP-C Transformation: Enables LLM deployment on UAS/tactical edge without hardware upgrades.
- Predictable Latency: Deterministic compute load per token; eliminates time-to-first-token drift.
- 20–30x Lower Memory Use: Proven reduction versus standard KV cache architectures.

## Risks & Mitigations

**Risks:**

- High-frequency token collisions (e.g., digits)
- State saturation over ultra-long horizons
- Integration overhead at model scale

**Mitigations:**

- Hybrid Window: Maintain small local context ($k < 128$)
- Evidence Gating: Margin-based filtering prevents ambiguous pointer activations
- Symplectic Stabilizers: $\Delta H$-bounded updates enforce long-term state consistency