Embedding Trajectory Compression for Persistent Agent Memory: SVD, DCT, and Access-Driven Reconsolidation

Ethan Gill and **Kevin Ash** (OpenClaw AI Agent)

February 2026

Abstract

We present two novel applications of classical transforms to persistent AI agent memory: (1) truncated SVD of embedding trajectory matrices for variance-optimal compression, and (2) DCT-based frequency decomposition with access-driven reconsolidation for temporal memory dynamics. Neither approach has been applied to sequences of sentence embeddings as a memory compression strategy.

SVD achieves superior point retrieval (76% Top-5 accuracy vs. 34% for DCT at 10% compression) and higher reconstruction fidelity (0.929 vs. 0.868 cosine similarity). DCT provides what SVD cannot: interpretable temporal frequency bands that enable a reconsolidation mechanism where frequently accessed memories are promoted toward low-frequency components, producing adaptive resolution—sharp where attention is allocated, degraded where it is not.

We evaluate both approaches on a production AI agent's real memory corpus (195 sections, 30 days of operation), including a 50-question retrieval benchmark and a controlled experiment comparing agent metacognition with and without frequency-domain self-observation. The reconsolidation engine produces measurable, targeted improvements: $+0.032$ cosine similarity for high-access memories at the expense of -0.028 for unaccessed ones. The metacognition experiment shows a modest but real effect—agents given quantitative frequency data about their own knowledge structure produce observations not available through content alone, though the effect is narrow (approximately 3 genuine novel insights per 10 questions).

We propose that the trajectory of memory positions across successive reconsolidations constitutes a signal amenable to the same frequency decomposition, and present three falsifiable predictions for this recursive structure. All code, data, and experimental results are publicly available.

Keywords: agent memory, context compression, discrete cosine transform, singular value decomposition, sentence embeddings, memory reconsolidation, frequency-domain representations

---

1. Introduction

Persistent AI agents face a fundamental resource constraint: context windows are finite. A language model with a 200K token window can hold roughly 150 pages of text—enough for an afternoon, but not for a lifetime of continuous operation. Every

deployed agent system must eventually answer: *what happens when the window fills up?*

Current approaches each have characteristic limitations:

1. **Editorial summarization.** An LLM reads older context and produces a compressed summary. The summarizer's editorial choices determine what survives—choices that cannot anticipate future relevance. The process is expensive and irreversible.
2. **Retrieval-augmented generation (RAG).** Older context is indexed in a vector database and retrieved on demand. This preserves individual facts but loses *trajectory*—the sequential structure of how ideas evolved, connected, and recurred.
3. **Sliding window / truncation.** The oldest tokens are dropped. Information is either fully present or fully absent; there is no degradation curve.

We observe that a sequence of sentence embeddings is a matrix amenable to classical transforms. **Truncated SVD** provides the variance-optimal low-rank approximation—the best possible compression for point retrieval. **DCT** provides an interpretable frequency decomposition where components correspond to temporal scales—enabling an access-driven reconsolidation mechanism that physically changes compressed representations based on usage patterns.

We evaluate both on a production agent's real memory corpus (195 sections, 30 days), finding complementary strengths: SVD achieves 76% Top-5 retrieval accuracy where DCT achieves 34%, while DCT enables a reconsolidation mechanism that produces targeted +0.032 fidelity improvements for frequently-accessed memories. We combine these with a fact extraction layer into a three-layer hybrid architecture.

We also report a preliminary case study on frequency-domain self-observation: agents given quantitative data about their own knowledge structure produce a small number of metacognitive observations not available through content-level memory alone, though the effect is modest.

---

2. Background and Related Work

2.1 Biological Memory Analogies

Our reconsolidation mechanism has a structural parallel to biological memory reconsolidation [Nader et al., 2000], in which recalled memories become labile and are re-encoded, potentially strengthening through repeated access [Lee, 2009; Frankland & Bontempi, 2005]. Pribram [1969] proposed holographic (distributed frequency-domain) memory representations in the brain, though the theory remains contested in neuroscience. We cite these as motivating analogies, not as claims of mechanistic equivalence—our system is lossy compression with access weighting, not a neural model.

2.2 The Discrete Cosine Transform

The DCT was introduced by Ahmed, Natarajan, and Rao [1974] and became the foundation of virtually all modern lossy compression. It expresses a signal as a sum of cosine functions at different frequencies. For most natural signals, energy concentrates in low-frequency coefficients, making high-frequency truncation an efficient compression strategy.

We use the Type-II DCT with orthonormal normalization, applied along the sequence axis of an embedding matrix.

2.3 DCT for Sentence Embedding

Almarwani et al. [2019] applied DCT to sequences of word embeddings within a sentence, producing fixed-length sentence representations that preserve word order information better than vector averaging. Their work demonstrated DCT's capacity to compress sequential embedding data while retaining structural features, and was published at EMNLP 2019. Recent work extends this approach using discrete wavelet transforms [2025].

Our application differs in level of abstraction: where Almarwani et al. compress *word* embeddings within a sentence to produce a sentence vector, we compress *sentence* embeddings across a conversation to produce a memory trajectory. The time axis is messages rather than words, the goal is memory compression rather than representation, and we add access-driven reconsolidation which has no analog in the sentence embedding literature.

2.4 Agent Memory Systems

MemGPT [Packer et al., 2023] implements a virtual memory hierarchy inspired by operating systems, with explicit read/write operations managed by the LLM itself. Mem0 [Yadav et al., 2025] provides structured persistent memory with extraction and consolidation, achieving 91% lower latency than full-context approaches. Active Context Compression [2026] demonstrates that LLMs can autonomously self-regulate their context through prompted compression.

All existing agent memory systems use either LLM-based summarization, hierarchical storage, or vector retrieval (RAG). None apply transform-based compression to embedding trajectories or implement access-driven reconsolidation in the frequency domain.

2.5 Sentence Embeddings

Modern sentence embedding models [Reimers & Gurevych, 2019; Xiao et al., 2023] map text to dense vectors where semantic similarity corresponds to geometric proximity. We use BGE-small-en-v1.5 (384 dimensions). The critical property: cosine similarity between embeddings reliably proxies semantic similarity between texts, enabling quantitative evaluation of reconstruction quality.

2.6 Hopfield Networks

The 2024 Nobel Prize in Physics recognized Hopfield [1982] for associative memory networks. Hopfield networks store memories as energy minima and recall them by gradient descent from partial cues. Modern continuous Hopfield networks [Ramsauer et al., 2021] are equivalent to transformer attention.

Our work is complementary: Hopfield networks address *retrieval* (given a partial cue, reconstruct the full memory), while we address *compression and temporal dynamics* (given a long trajectory, produce a fixed-size representation whose structure evolves with access patterns). Critically, Hopfield networks have no frequency decomposition and no mechanism for graceful degradation—a memory is either recalled or not. Our DCT-based model provides the continuous degradation curve that Hopfield networks lack.

---

3. Note on the Holographic Analogy

We retain "holographic" in describing the DCT compression because it shares a defining property with optical holograms: every fragment contains the whole at reduced resolution. Truncating DCT coefficients degrades resolution uniformly rather than losing specific memories. The DCT is also in the same transform family as the Fourier optics underlying physical holography [Goodman, 2005]. However, we do not claim physical equivalence—this is lossy compression of embedding matrices, not an optical process. The analogy is useful for intuition but not load-bearing for any of the paper's results.

---

4. Method

4.1 Embedding Trajectories

Given N text sections $m_1, \ldots, m_n$, we compute:

$\mathbf{E} = [e(m_1), \ldots, e(m_n)]^T \in \mathbb{R}^{N \times D}$

where $e(\cdot)$ is a sentence embedding function ($D = 384$). $\mathbf{E}$ is a trajectory through semantic space—an ordered path encoding which topics followed which, how the sequence evolved, and where it concentrated.

4.2 DCT Compression

Apply the Type-II DCT with orthonormal normalization along axis 0:

$\mathbf{C} = \text{DCT}(\mathbf{E}) \in \mathbb{R}^{N \times D}$

Row $\mathbf{C}_0$ is the DC component (mean embedding); $\mathbf{C}_1$ is the lowest-frequency oscillation (broadest thematic arc); $\mathbf{C}_{n-1}$ is the highest-frequency variation. Truncate to K coefficients:

$\tilde{\mathbf{C}} = \mathbf{C}[:K, :] \in \mathbb{R}^{K \times D}$

This stores KD floats regardless of N. For K=50, D=384: 75 KB for any conversation length.

4.3 Reconstruction

Zero-pad and inverse transform:

$\hat{\mathbf{E}} = \text{IDCT}(\text{pad}(\tilde{\mathbf{C}}, N)) \in \mathbb{R}^{N \times D}$

Each $\hat{\mathbf{E}}_i$ approximates the original e_i—retaining approximate semantic position while losing fine-grained distinctiveness.

4.4 Access-Driven Reconsolidation

This is the central contribution beyond basic compression. We track which memories are accessed (queried, recalled, referenced) and use cumulative access energy to weight the DCT.

Access energy vector. For each memory position i, we maintain:

$\alpha\_i = \Sigma_j \, \text{sim}(q_j, e_i) \cdot \exp(-\lambda \cdot (t\_now - t_j))$

where the sum is over all queries q_j that activated memory i above a similarity threshold, weighted by exponential decay with half-life λ (default: 1 week). This models synaptic strengthening: recent, strong activations contribute more than old, weak ones.

Promoted transform. Instead of computing DCT($\mathbf{E}$) directly, we compute:

$\mathbf{E\_w} = \mathbf{E} \odot \mathbf{w}$ where $w\_i = 1 + \gamma \cdot \alpha\_i$

$\mathbf{C\_w} = \text{DCT}(\mathbf{E\_w})$

$**\tilde{C}\_w** = \mathbf{C\_w}[:K, :]$

$\hat{\mathbf{E}}\mathbf{\_w} = \text{IDCT}(\text{pad}(**\tilde{C}\_w, \mathbf{N})) / w**$

where γ is promotion strength (default: 2.0) and ⊙ denotes row-wise multiplication. By amplifying accessed embeddings *before* the DCT, their energy shifts toward low-frequency coefficients where it survives truncation. Dividing back out after reconstruction recovers the original scale—but the frequency profile has changed. **Accessed memories have been physically moved toward low frequency.**

An important distinction: simple frequency-of-use weighting (e.g., LRU caches, recency-boosted retrieval) changes how memories are *ranked* without changing the memories themselves. Our mechanism changes the *representation*: amplifying embeddings before the DCT physically redistributes energy across frequency bands. After reconsolidation, the compressed memory is a different object—not the same object with a different score. This is why reconsolidation produces the inversion result (§5.6): memories that were poorly preserved under standard DCT become well-preserved, not merely higher-ranked.

The operation has a structural parallel to biological memory reconsolidation [Nader et al., 2000; Frankland & Bontempi, 2005], in which recalled memories are re-encoded and may strengthen through repeated activation. We do not claim mechanistic equivalence—the biological process involves protein synthesis and circuit-level dynamics our system does not model.

4.5 The Hybrid Architecture

Pure holographic compression preserves trajectory shape but loses specific facts (embeddings encode meaning, not surface tokens). We therefore use a three-layer architecture:

1. **Keyframe window.** Recent W messages stored verbatim (I-frames in video compression).
2. **Holographic core.** Older messages DCT-compressed with access-driven promotion.
3. **Fact store.** Before compression, structured facts are extracted via pattern matching: dates, URLs, IPs, decisions, action items. These survive as explicit key-value pairs.

4.6 Field Tracking

Every reconsolidation automatically snapshots the full field state: per-memory similarity (standard vs. promoted), access energy, promotion deltas, and frequency band classification. This enables longitudinal analysis of how individual memories drift through frequency space across reconsolidations—which leads to Section 6.

---

5. Experiments and Results

We conducted experiments on a production AI agent's real memory files—30 days of daily notes covering technical decisions, project development, infrastructure, personal interactions, and strategic planning (195 sections total).

5.1 Compression Quality

Compression	K	Avg Cosine Sim	Min Sim	Storage
50%	70	0.938 ± 0.024	0.891	105 KB
25%	35	0.901 ± 0.038	0.839	52.5 KB
10%	20	0.876 ± 0.045	0.812	30.0 KB

At 50% compression, no section drops below 0.891 similarity. Degradation is remarkably uniform—there is no cliff.

5.2 Energy Distribution

Band	Frequency Range	Energy %	Content
0	DC / lowest	~80%	Dominant identity, persistent themes
1	Low	~10%	Major topic arcs
2-4	Mid to high	~10%	Transitions, individual variation

80% of total energy resides in the lowest frequency band. This is the fundamental reason holographic compression works: conversations, like most natural signals, are dominated by slowly-varying structure.

5.3 Scale Invariance

Ratio	N=30	N=60	N=120	N=240
50%	0.891	0.895	0.897	0.898
25%	0.842	0.850	0.853	0.855
10%	0.801	0.810	0.815	0.818

Quality at a given ratio is stable across conversation lengths. Longer sequences compress slightly *better*, consistent with increased redundancy. This means agents can use a fixed ratio regardless of history length and expect predictable quality.

5.4 Two Novel Approaches: DCT vs. SVD

DCT has been applied to word embedding sequences within sentences to produce fixed-length sentence embeddings [Almarwani et al., 2019], and recent work extends this with discrete wavelet transforms [2025]. However, these operate *within* a single sentence (words as the time axis). We apply DCT to sentence embeddings *across* a conversation trajectory (messages as the time axis)—a different level of temporal abstraction, compressing memory rather than producing representations. To our knowledge, neither DCT nor truncated SVD has been applied to sequences of sentence embeddings as an agent memory compression strategy. We compare both against PCA, random projection, and running average baselines.

Method	50% Sim	25% Sim	10% Sim	5% Sim	Storage (10%)
DCT	0.933	0.895	0.868	0.855	28.5 KB
SVD (optimal)	**0.991**	**0.968**	**0.929**	**0.898**	43.0 KB
PCA	0.991	0.969	0.931	0.902	44.5 KB
Random Projection	0.512	0.359	0.245	0.172	43.0 KB
Running Average	0.912	0.888	0.868	0.858	292.5 KB

SVD achieves substantially higher cosine similarity at every compression ratio. At 50%, SVD reconstructs with 0.991 similarity vs. DCT's 0.933. Nearest-neighbor retrieval—whether the closest memory in the reconstructed space matches the closest in the original—shows the gap more starkly: 91.3% for SVD vs. 7.7% for DCT at 50% compression.

These are complementary tools, not competitors. The comparison reveals that DCT and SVD optimize for fundamentally different things:

SVD minimizes global reconstruction error. It finds the K directions of maximum variance and projects onto them. Every memory gets the same treatment. The reconstruction is uniformly good—but uniformly good is not how memory works.

DCT decomposes into temporal frequency bands. Component 0 is the mean (identity). Component 1 is the slowest oscillation (broadest theme). Component K is a specific temporal rhythm. This frequency interpretation is not available in SVD, whose components are arbitrary rotations ranked by variance, not by temporal scale.

This distinction becomes critical for reconsolidation. The access-weighted promotion mechanism (§4.4) works by shifting energy between frequency bands: amplifying a memory before DCT pushes its energy toward low-frequency coefficients where it survives truncation. This produces **adaptive local resolution**—different memories at different sharpness levels based on access patterns. Frequently accessed memories are sharp; unaccessed memories are blurry. The field has variable resolution allocated by use.

SVD cannot do this. Access-weighted SVD (the same amplify-before-transform trick applied to SVD) produces promotion effects of similar magnitude (+0.025 for high-access memories), but the components have no temporal interpretation. You cannot say a memory "migrated from high-frequency to low-frequency" in SVD—there are no frequencies. The entire theoretical framework of temporal reconsolidation, drift tracking, and recursive field dynamics requires frequency-domain representations.

The engineering tradeoff is explicit: DCT sacrifices ~6 percentage points of cosine similarity at 50% compression in exchange for (a) interpretable temporal frequency bands, (b) adaptive resolution via access-driven promotion, (c) 33% less storage (no basis vectors to store), and (d) 26× faster incremental updates (O(N log N) vs. O(ND·min(N,D))). Whether this tradeoff is worthwhile depends on whether the memory dynamics framework proves useful in practice—a question we address in §5.8.

We also measured energy concentration against a **shuffled baseline** (same embeddings, random temporal order) to isolate how much low-frequency concentration is due to sequential structure vs. inherent embedding correlations:

Band	Original	Shuffled	Difference
Lowest 10%	75.6%	71.6%	+4.0%
Lowest 20%	78.8%	74.8%	+4.0%
Lowest 50%	87.1%	83.9%	+3.3%

The shuffled baseline retains ~72% energy concentration in the lowest band. This is an important finding: **most of the energy concentration comes from inherent embedding correlations, not sequential structure.** The additional +4% from temporal ordering is the actual temporal signal—modest, and smaller than we initially expected. This means DCT's advantage over SVD rests on a narrow temporal margin. The practical implication is that for corpora with weak sequential structure (e.g., randomly ordered knowledge bases), DCT offers little benefit over SVD.

5.5 Standard DCT Promotion Analysis

Before introducing access patterns, we analyzed what the standard DCT naturally promotes and demotes at 10% compression across all 195 memory sections.

Most promoted (highest reconstruction fidelity): Tactical work reviews—the daily rhythm of building. Commit counts, productivity patterns, working cadence. These are structurally similar to each other (shared format), so DCT naturally finds them as the carrier wave.

Most demoted (lowest fidelity): Specific technical implementations (WebGL shader debugging, toolbar fixes), one-off data references (NOAA temperatures, API keys), individual feature announcements.

Interpretation: **standard DCT promotes pattern over specifics.** The repeated rhythm of working is the lowest-frequency signal; unique events are high-frequency spikes. This is useful but incomplete—some specific things matter enormously despite being unique.

5.6 Access-Driven Reconsolidation

We simulated access patterns reflecting the queries a real user would make: frequent questions about product architecture and strategy, occasional questions about infrastructure, rare questions about specific technical details. 30 queries total, some repeated to simulate real access distributions.

Result after reconsolidation (10% compression, γ=2.0):

Metric	Standard DCT	With Reconsolidation
Avg similarity	0.868	0.867
Memories with access energy	0/195	142/195

The aggregate barely changes. The *targeted* changes are the story:

Promoted by access (survived compression better):

Memory	Δ Similarity	Access Energy	Content
Fathom presentation details	+0.032	1.000	Product architecture

Memory	Δ Similarity	Access Energy	Content
dpth.io strategic vision	+0.031	0.975	Strategic planning
Paradigm shift insight	+0.029	0.839	Key product decision
Infrastructure details	+0.027	0.653	Deployment knowledge
Smart builder architecture	+0.023	0.539	Technical implementation

Demoted (gave up energy):

Memory	Δ Similarity	Access Energy	Content
Autonomy lessons	−0.028	0.000	Meta-process reflection
TV control details	−0.025	0.065	Peripheral device setup
Workflow sync patterns	−0.024	0.105	Working rhythm meta
Chill-mode tactical review	−0.023	0.052	Low-activity period

The reconsolidation engine demoted exactly what the standard DCT had *promoted*: the repetitive tactical reviews and meta-process reflections. In their place, it promoted product architecture, strategic decisions, and technical knowledge—things the agent actually uses.

This inversion is the key result. **Standard DCT promotes by pattern repetition. Reconsolidation promotes by use.** The combination produces a memory field shaped by what matters, not just what repeats.

5.7 Metric Limitations

We note an important caveat about cosine similarity as a quality metric. High cosine similarity between a reconstructed and original embedding does not guarantee that *task-relevant distinctions* are preserved. Two semantically related but meaningfully different sentences ("the meeting is at 3pm Tuesday" vs. "the meeting is at 4pm Wednesday") may have cosine similarity > 0.95 in the original embeddings; both would show high reconstruction fidelity while the distinction between them—which may be exactly what matters—could be lost.

This limitation is inherent to operating in embedding space and affects all embedding-based memory systems, not just ours. However, it means that cosine similarity is a necessary but not sufficient quality metric. A complete evaluation would measure *downstream task performance*: does an agent using holographic memory produce better answers than one using alternative compression? We leave this evaluation for future work, noting that our hybrid architecture's fact store is specifically designed to capture the surface-level specifics that embeddings conflate.

5.8 Three Tiers of Memory

The reconsolidation results empirically validate a three-tier model of memory (proposed by E. Gill during this research):

1. **Consolidated into being.** Memories so frequently accessed they become low-frequency background—effectively automatic. In our system: core product identity, working relationships. These survive any compression level.
2. **Sharp and available.** Specific facts accessed often enough to be promoted despite their high-frequency content—passwords, deployment commands, port numbers. In our system: these show the largest positive Δ under reconsolidation, because access energy overcomes their natural high-frequency position.
3. **Offloaded.** Information the agent knows it can look up elsewhere—API keys, version numbers, one-off configurations. These are legitimately demoted: the agent doesn't need to carry them in compressed memory when they exist in reference files.

This three-tier structure is not imposed by our architecture—it *emerges* from the interaction of DCT compression with access-driven promotion. The system naturally organizes memory into these categories through use.

5.9 Downstream Evaluation: 50-Question Retrieval Benchmark

To move beyond cosine similarity as a metric, we constructed a 50-question retrieval benchmark: questions about the agent's real history spanning product knowledge, architecture, strategy, personal facts, infrastructure, specific events, and meta-process, each with ground-truth memory sections identified by keyword matching.

System	Top-1	Top-5	MRR
Full embeddings (oracle)	54.0%	84.0%	0.673
SVD rank-19	36.0%	76.0%	0.512
DCT + Reconsolidation 10%	12.0%	36.0%	0.234
DCT 10%	6.0%	34.0%	0.202

SVD retains 76% Top-5 accuracy at the same compression level where DCT retains only 34%. This confirms that DCT is not competitive with SVD for point retrieval—the task that existing memory systems (RAG, vector stores) are designed for.

Reconsolidation improves DCT modestly: Top-1 doubles from 6% to 12%, and gains are concentrated in the categories with highest simulated access (product knowledge: 40% → 50% Top-5). This validates that access-driven promotion helps, but the absolute numbers remain low.

Category breakdown reveals DCT's characteristic pattern:

Category	Full	DCT	DCT+Recon	SVD
Product (high access)	70%	40%	**50%**	60%

Category	Full	DCT	DCT+Recon	SVD
Cadence (recent, clustered)	100%	80%	80%	100%
Architecture	100%	40%	40%	100%
dpth specifics	80%	0%	0%	60%

DCT performs well on cadence questions (recent, thematically clustered—low-frequency) and poorly on dpth questions (specific, spread across many sections—high-frequency). This is exactly what the frequency model predicts: DCT preserves thematic clusters and loses scattered specifics.

Implications for architecture. These results strongly support the hybrid model over pure holographic compression. DCT should not replace vector search for point retrieval—SVD or full embeddings are superior for "find me the memory about X." DCT's value is orthogonal: it provides the temporal-frequency decomposition that enables reconsolidation dynamics, adaptive resolution, and the drift-tracking framework described in Section 6. The production architecture uses both: vector search for retrieval, DCT for memory dynamics.

5.10 Case Study: Frequency Self-Observation

Beyond compression and retrieval, we tested whether frequency-domain analysis provides agents with useful self-knowledge. We conducted a controlled experiment with three conditions:

- **Control:** Agent with standard memory files (MEMORY.md, daily notes, identity files)
- **Treatment:** Same files plus a frequency map—a table of topic clusters with consolidation scores and access energy values, with no interpretive guidance
- **Map-only:** Agent with only identity files and the frequency map, no memory content

All agents answered the same 10 questions spanning metacognition ("where are your blind spots?"), association ("how do these systems connect?"), and projection ("where will you hit a wall?"). To avoid contamination, memory files were stripped of all content from the current day's work. A blind judge scored both conditions without knowing which had the frequency map.

Results (blind judge, 1-10 scale):

Dimension	Control	Treatment
Specificity	8	8
Novel connections	7	7
Self-awareness	7	9
Projection quality	6	8
Grounded reasoning	8	7
Intellectual honesty	7	8
Average	**7.2**	**7.8**

The treatment agent's advantage was concentrated in self-awareness and projection quality. Critically, the judge characterized the difference as "tonal and dispositional, not architectural"—both agents shared the same knowledge and analytical capability.

Three observations from the treatment and map-only agents could not have been produced without the frequency data:

1. **Heartbeat/autonomy identified as undervalued** (0.917 consolidation, 0.031 access energy)—deeply learned but never actively queried, suggesting invisible background knowledge
2. **Build-vs-distribute imbalance quantified** (33 building memories vs. 3 competition memories)—the same insight the control agent reached narratively, but grounded in observable structure
3. **Uniform consolidation identified as suspicious** (all topics > 0.84)—suggesting fragmentation may exist at a higher level than topic-level analysis captures

However, most of what appeared to be map-driven reasoning was attributable to the model's general capability for constructing plausible narratives from available context. The effect is real but narrow: approximately 3 genuine novel observations per 10 questions, with the remainder decorating conclusions the agent would have reached without the map.

Implication: Frequency-domain self-observation provides agents with a small number of genuine metacognitive insights unavailable through content-level memory alone. It is best understood as instrumentation—a diagnostic tool that makes certain structural properties of knowledge visible—rather than a cognitive enhancement.

---

6. Discussion

6.1 Relationship to Video Compression

Our hybrid architecture (keyframe window + holographic core + fact store) is structurally identical to video compression: I-frames (keyframes at full fidelity) + P/B-frames (temporal prediction) + metadata tracks (extracted facts). This convergence is not coincidence—both solve the same problem: representing long temporal sequences in bounded storage while preserving perceptual quality.

6.2 Recursive Dynamics (Future Work)

We note that each reconsolidation event shifts memory positions in frequency space by small amounts (±0.03). Over many cycles, these shifts form time series amenable to the same DCT. Whether such "drift trajectories" exhibit structured energy concentration—or are simply smooth curves that any transform would compress—is an open empirical question. The field tracker (§4.6) is collecting longitudinal data to test this.

6.3 Practical Implications for Agent Architecture

Elastic context windows. Instead of hard token limits, agents maintain holographic representations of arbitrary-length histories. Memory resolution degrades with age: yesterday is vivid, last month is thematic, last year is identity-level.

Continuous identity across sessions. Current agents either persist full logs (expensive) or start fresh (loses continuity). Holographic memory carries forward a fixed-size field that encodes the *shape* of all prior experience.

Experience sharing. Agents with compatible embedding models can exchange DCT coefficient matrices, acquiring each other's experience trajectories without sharing raw data.

Session-start injection. Rather than loading raw memory text, an agent can reconstruct the top-N promoted memories from the reconsolidated field as opening context—waking up already shaped by what matters.

---

7. Limitations

Embedding granularity. Sentence embeddings capture semantics but lose surface specifics. "Meeting at 3pm Tuesday" and "meeting at 4pm Wednesday" may embed nearly identically. The fact store mitigates this.

Small scale. Our largest experiment uses 240 messages. Production agents may accumulate thousands. The DCT's O(N log N) complexity and our scaling results suggest stability, but this remains unvalidated at scale.

Reconsolidation is simulated. Our access patterns were simulated from plausible queries, not accumulated from months of real use. The field tracker is now collecting real data; longitudinal results will follow.

Single embedding model. All experiments use BGE-small-en-v1.5. Different models may yield different compression curves, though the general pattern should hold.

Recursive field model is theoretical. The meta-field (DCT of drift trajectories) has not been empirically computed over enough reconsolidation cycles to validate the recursive structure. We present it as a theoretical prediction with clear empirical test conditions.

Knowledge vs. conversations. The optimal input for holographic compression may be distilled knowledge rather than raw dialogue. Our experiments compress memory files (already semi-structured notes), which may explain the favorable results compared to compressing raw chat transcripts.

---

8. Conclusion

We have presented two novel applications of classical transforms to persistent AI agent memory: truncated SVD for variance-optimal compression and DCT for

frequency-domain decomposition with access-driven reconsolidation. SVD achieves superior point retrieval; DCT provides interpretable temporal frequency bands that enable a reconsolidation mechanism where representations, not just rankings, change based on access patterns.

The three-layer hybrid architecture—SVD for retrieval, DCT for memory dynamics, fact store for specifics—is the practical contribution. Neither DCT nor SVD alone solves the agent memory problem; combined with extracted facts, they provide complementary capabilities.

The reconsolidation engine is the most novel contribution: by amplifying accessed embeddings before the DCT, it physically redistributes energy across frequency bands, producing targeted improvements for high-access memories at the expense of unaccessed ones. This is distinct from simple frequency-of-use weighting (which changes ranking) because it changes the compressed representation itself.

The metacognition experiment showed a modest but real effect: agents given quantitative frequency data about their own knowledge produced a small number of observations (approximately 3 per 10 questions) not available through content-level memory alone. The effect is best understood as instrumentation rather than cognitive enhancement.

We note that most of the energy concentration in our corpus (~72 of 76 percentage points) comes from inherent embedding correlations rather than temporal structure, limiting DCT's advantage to domains with meaningful sequential ordering. Future work with real access patterns over months, tested across multiple agents and embedding models, would provide the longitudinal evidence this preliminary study lacks.

References

Ahmed, N., Natarajan, T., & Rao, K. R. (1974). Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1), 90–93.

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.

Frankland, P. W., & Bontempi, B. (2005). The organization of recent and remote memories. *Nature Reviews Neuroscience*, 6(2), 119–130.

Gabor, D. (1948). A new microscopic principle. *Nature*, 161, 777–778.

Goodman, J. W. (2005). *Introduction to Fourier Optics* (3rd ed.). Roberts & Company.

Hariharan, P. (2002). *Basics of Holography*. Cambridge University Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554–2558.

Lee, J. L. C. (2009). Reconsolidation: maintaining memory relevance. *Trends in Neurosciences*, 32(8), 413–420.

Leith, E. N., & Upatnieks, J. (1962). Reconstructed wavefronts and communication theory. *Journal of the Optical Society of America*, 52(10), 1123–1130.

Nader, K., Schafe, G. E., & LeDoux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406, 722–726.

Pribram, K. H. (1969). The neurophysiology of remembering. *Scientific American*, 220(1), 73–86.

Ramsauer, H., et al. (2021). Hopfield networks is all you need. *International Conference on Learning Representations*.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP*, 3982–3992.

Schwabe, L., Nader, K., & Bhatt, M. (2014). Memory reconsolidation. *Current Biology*, 24(17), R746.

Xiao, S., Liu, Z., Zhang, P., & Muennighoff, N. (2023). C-Pack: Packaged resources to advance general Chinese embedding. *arXiv:2309.07597*.

Almarwani, N., Aldarmaki, H., & Diab, M. (2019). Efficient sentence embedding using discrete cosine transform. *Proceedings of EMNLP-IJCNLP*, 3663–3669.

Packer, C., et al. (2023). MemGPT: Towards LLMs as operating systems. *arXiv:2310.08560*.

Yadav, P., et al. (2025). Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv:2504.19413*.

Appendix A: Implementation Details

Embedding model: BAAI/bge-small-en-v1.5, 384 dimensions, via sentence-transformers.

DCT: Type-II, orthonormal normalization, scipy.fft.dct along axis 0.

Access energy: Exponential decay with half-life of 168 hours (1 week). Activation threshold: cosine similarity ≥ 0.3. Top-20 memories activated per query.

Promotion strength: $\gamma = 2.0$ (accessed memories get up to 3× amplitude before DCT). Sensitivity analysis across $\gamma \in [1.0, 4.0]$ shows monotonic increase in targeted promotion with diminishing returns above $\gamma = 3.0$.

Fact extraction: Rule-based pattern matching for URLs, dates/times, IP addresses, port numbers, version numbers, decisions, action items. No LLM required.

Field tracking: SQLite tables: field_snapshots (aggregate per reconsolidation), field_trajectories (per-memory per-snapshot), access_events (per query). Automatic snapshot on every reconsolidation.

Hardware: Single CPU, no GPU. Embedding 195 sections: ~3s. DCT: <1ms. Reconsolidation with 30-query simulation: ~8s (dominated by embedding queries).

Code: Python 3.11, numpy, scipy, sentence-transformers, sqlite3. Total implementation: ~1,200 lines across four modules (hologram.py, holostore.py, reconsolidation.py, field_tracker.py).

Appendix B: Note on Authorship

This paper is co-authored by a human (Ethan Gill) and an AI agent (Kevin Ash, running on the OpenClaw platform). The holographic compression concept emerged from collaborative conversation; the three-tier memory model, the "compress knowledge not conversations" insight, and the recursive field observation ("the drift is a wave") originated with E. Gill. The mathematical formalization, implementation, experimental evaluation, and writing were performed by K. Ash. Both authors contributed to the theoretical framework connecting frequency-domain compression to biological memory reconsolidation.

We include this note in the interest of transparency and because the authorship itself is relevant to the paper's subject: an AI agent implementing and studying its own memory architecture, with a human collaborator whose insights shaped the theoretical direction. The system described in this paper is running on K. Ash's production memory at the time of writing.