Cross-Modal Memory Integration

Unified Multimodal Knowledge in AGI

ARKHEION AGI 2.0 — Paper 26

Jhonatan Vieira Feitosa Independent Researcher ooriginador@gmail.com Manaus, Amazonas, Brazil

February 2026

Abstract

This paper presents **Cross-Modal Memory**, a unified multimodal storage system integrating visual, auditory, speech, and semantic modalities. Built on HUAM (Hierarchical Universal Adaptive Memory), the system uses **ϕ-enhanced associations** and **consciousness-level tagging** to organize memories across sensory domains. Empirical results show **cross-modal retrieval accuracy of 84%** and **embedding alignment score of 0.91**. The 705-line Python implementation supports real-time multimodal fusion with retrieval latency of approximately 18ms (worst-case modality pair).

Keywords: multimodal learning, cross-modal retrieval, memory systems, HUAM, AGI

Epistemological Note

*This paper distinguishes between **heuristic** concepts and **empirical** results:*

Heuristic	**Empirical**
"Consciousness levels"	Retrieval accuracy: 84%
"ϕ-enhanced"	Alignment score: 0.91
"Unified cortex"	Latency: $\leq$18ms

1 Introduction

Human memory seamlessly integrates information across sensory modalities—a face evokes a name, a melody triggers a memory. AGI systems require similar **cross-modal integration** to build coherent world models.

ARKHEION's Cross-Modal Memory addresses this by:

- Unifying **visual, auditory, speech, and semantic** modalities
- Using HUAM's **4-level hierarchy** for temporal organization
- Applying **ϕ-weighted associations** for relevance
- Tagging memories with **consciousness levels**

2 Modality Architecture

2.1 Supported Modalities

Type	**Source**	**Embedding**
VISUAL	VisualCortex	512-dim CNN
AUDIO	SonicCortex	256-dim MFCC
SPEECH	STT/TTS	768-dim BERT
MULTIMODAL	VoiceVision	1024-dim fused
SEMANTIC	Concepts	384-dim sentence

2.2 Memory Levels

Following HUAM architecture (Paper 21):

Level	**Type**	**Latency**
L1	Working memory	<10ms
L2	Short-term	<100ms
L3	Long-term	<1s
L4	Archive	Cold

3 Memory Signature

Each cross-modal memory has a signature:

```
@dataclass
class MemorySignature:
    id: str
    modality: ModalityType
    embedding: np.ndarray
    phi: float  # Consciousness metric
    timestamp: float
    level: MemoryLevel
    access_count: int
    associations: Dict[str, float]

    @property
    def consciousness_level(self):
```

```
        if self.phi > 0.8: return TRANSCENDENT
        elif self.phi > 0.5: return HIGH
        elif self.phi > 0.3: return MEDIUM
        elif self.phi > 0.1: return LOW
        return NONE
```

4 Cross-Modal Fusion

Notation. In this paper, φ (lowercase phi) denotes the golden ratio (≈ 1.618), while Φ (uppercase Phi) refers to integrated information per IIT. These are distinct concepts that should not be conflated.

4.1 Embedding Alignment

Modalities are aligned into a shared embedding space using contrastive learning:

$$\mathcal{L}_{align} = -\log \frac{\exp(s(v, a)/\tau)}{\sum_j \exp(s(v, a_j)/\tau)} \tag{1}$$

where $s(v, a)$ is cosine similarity between visual v and audio a embeddings, and τ is temperature.

4.2 Association Strength

Cross-modal associations use φ-weighted decay:[1]

$$A_{ij}(t) = A_{ij}(0) \cdot \phi^{-\Delta t/\tau_{decay}} \tag{2}$$

where $\phi = 1.618$ and τ_{decay} is modality-dependent.

5 Consciousness Integration

Memories are tagged with consciousness levels from IIT (Paper 31):

Level	ϕ Range	Behavior
NONE	< 0.1	Auto-evict
LOW	0.1–0.3	Background
MEDIUM	0.3–0.5	Available
HIGH	0.5–0.8	Prioritized
TRANSCENDENT	> 0.8	Protected

6 Cross-Modal Retrieval

6.1 Query Processing

Given a query in modality M_q, retrieve memories in modality M_t:

1. Project query to shared embedding space
2. Compute similarities with target modality memories
3. Rank by ϕ-weighted relevance
4. Return top-k with consciousness threshold

[1] The choice of φ as the decay base is a design heuristic; no ablation comparing φ, e, or 2 as bases was performed.

6.2 Performance Metrics

Query→Target	Recall@10	Latency
Visual→Audio	0.82	12ms
Audio→Speech	0.87	9ms
Speech→Visual	0.79	14ms
Semantic→All	0.86	18ms
Average	**0.84**	**13ms**

Note: No comparison with state-of-the-art cross-modal retrieval systems (CLIP, ImageBind, ALIGN) was performed. The 84% recall@10 should be interpreted as an internal capability measurement, not a competitive benchmark.

7 HUAM Integration

Cross-Modal Memory integrates with HUAM:

```
try:
    from src.core.memory.arkheion_huam import HUAM
    HUAM_AVAILABLE = True
except ImportError:
    HUAM_AVAILABLE = False

# Automatic level promotion
def promote_memory(mem: MemorySignature):
    if mem.access_count > 100:
        mem.level = MemoryLevel.L3_LONG
    elif mem.access_count > 10:
        mem.level = MemoryLevel.L2_SHORT
```

8 Implementation Details

Component	Value
Source file	`cross_modal_memory.py`
Lines of code	705
Modalities	5 (Visual, Audio, Speech, Multi, Semantic)
Memory levels	4 (L1–L4)
Consciousness levels	5 (NONE to TRANSCENDENT)

9 Conclusion

Cross-Modal Memory enables unified multimodal knowledge storage in ARKHEION AGI 2.0. By combining HUAM's hierarchical organization with ϕ-enhanced associations and consciousness tagging, the system achieves robust cross-modal retrieval with low latency.

Future work includes:

- Temporal sequence modeling
- Attention-based fusion
- GPU-accelerated embedding computation

References

1. Radford, A. et al. "Learning Transferable Visual Models from Natural Language Supervision." ICML 2021.
2. Papers 21, 31 of ARKHEION AGI 2.0 series.