Memory Mechanisms in Advanced AI Architectures: A Unified Cross-Domain Analysis

Anonymous authors
Paper under double-blind review

Abstract

Memory mechanisms have become essential components in advanced AI architectures, significantly impacting performance, efficiency, and adaptability across diverse domains. This survey presents a unified theoretical framework for analyzing memory systems through three complementary lenses: retrieval mechanisms, memory structures, and update schemas. We systematically examine memory implementations across four key domains: Large Language Models (LLMs), Vision-Language Models (VLMs), Visual Prompt Tuning (VPT), and Video Understanding systems.

Our analysis reveals both universal memory patterns that transcend domains and domain-specific optimizations that address unique challenges in each field. We identify significant evolutionary trends, including the increasing prevalence of hybrid retrieval approaches, progression toward sophisticated hierarchical memory structures, and development of multi-factor update schemas that balance stability with adaptability. Through cross-domain comparisons, we identify transferable principles, highlight remaining challenges, and propose promising research directions for next-generation memory systems. These include theoretical frameworks for memory capacity optimization, cognitive-aligned architectures, cross-modal knowledge abstraction, and privacy-preserving memory systems. This comprehensive analysis provides valuable insights for researchers working on memory-enhanced AI systems across diverse application domains, offering both theoretical foundations and practical design considerations.

1 Introduction

1.1 Memory Systems in Modern AI Architectures

Modern AI systems face a common set of challenges that make sophisticated memory mechanisms necessary:

- **Contextual Understanding**: AI systems must maintain coherent understanding across extended interactions or processing streams, requiring effective recall of previously encountered information.
- **Computational Efficiency**: As models grow in size and complexity, memory mechanisms must balance expressiveness with computational requirements, especially for deployment in resource-constrained environments.
- **Adaptability to New Information**: Systems must effectively integrate new knowledge with existing information without catastrophic forgetting or performance degradation.
- **Cross-Modal Integration**: Many contemporary AI systems operate across multiple modalities (text, vision, audio), requiring memory structures that can effectively bridge these different information types.

1.2 Cross-Domain Memory Challenges

While memory mechanisms share common challenges across domains, each field has developed specialized approaches to address domain-specific requirements:

- **Large Language Models (LLMs)** face challenges with context length limitations, knowledge integration, and maintaining coherence across extended interactions. Recent advances have introduced innovative solutions ranging from retrieval-augmented generation to explicit memory architectures inspired by cognitive science.
- **Vision-Language Models (VLMs)** must maintain alignments between visual and textual representations while managing modality-specific processing. Memory systems in this domain have evolved from simple prompt banks to sophisticated cross-modal attention mechanisms.
- **Visual Prompt Tuning (VPT)** approaches must learn new tasks sequentially without forgetting previous ones, requiring memory structures that carefully balance stability with plasticity. Parameter efficiency is particularly critical in this domain.
- **Video Understanding Systems** face unique challenges with temporal data and extended context maintenance. Memory systems in this domain have progressed from simple temporal buffers to hierarchical architectures that mimic human cognition.

1.3 A Unified Framework for Analysis

To systematically analyze memory systems across these diverse domains, we propose a unified theoretical framework that examines memory systems through three complementary lenses:

1. **Retrieval Mechanisms**: How information is accessed from memory, including similarity-based, prompt-based, temporal-spatial, and hybrid approaches.
2. **Memory Structures**: How information is organized and stored, including static, dynamic, hierarchical, and distributed architectures.
3. **Update Schemas**: How memory content evolves over time, including frequency-based, recency-based, importance-weighted, and privacy-preserving approaches.

This framework enables us to identify both universal patterns that transcend domains and domain-specific optimizations that address unique challenges in each field.

1.4 Contributions and Paper Organization

This survey makes the following contributions:

- Provides a comprehensive analysis of memory systems across four key AI domains
- Identifies universal memory patterns and domain-specific optimizations
- Tracks evolutionary trends in memory system design
- Proposes promising directions for next-generation memory systems

The remainder of this paper is organized as follows: Section 2 reviews related work across the four domains. Section 3 presents our unified framework for analyzing memory systems. Sections 4.1-4.4 provide domain-specific analyses of memory systems in LLMs, VLMs, VPT, and video understanding, respectively. Section 5 presents a cross-domain analysis identifying common patterns and unique adaptations. Finally, Section 6 discusses future research directions and opportunities for cross-domain knowledge transfer.

2 Related Work

2.1 Large Language Models (LLMs)

Large Language Models face fundamental constraints in context management and long-term coherence. Recent surveys have systematically analyzed memory mechanisms that address these limitations. Zhang et al. (2024) provide a taxonomy of memory mechanisms for LLM-based agents, categorizing them by information sources, storage forms, and operation mechanisms. Their work emphasizes how memory enables coherence across extended interactions and knowledge integration.

Zhao et al. (2023) trace LLM evolution from statistical models to transformer architectures, examining how scaling affects memory capabilities and emergent abilities. Minaee et al. (2024) review prominent LLM families and augmentation techniques that overcome fixed-context limitations.

Guo et al. (2023) specifically explore working memory frameworks inspired by cognitive psychology, proposing centralized Memory Hubs and Episodic Buffers to overcome traditional LLM memory limitations. These surveys collectively demonstrate progression from simple context windows to sophisticated architectures inspired by human cognitive systems, highlighting memory's critical role in advancing LLM capabilities.

2.2 Vision-Language Models (VLMs)

Vision-Language Models (VLMs) have emerged as powerful frameworks for connecting visual and textual modalities, enabling various downstream tasks like image classification, object detection, and semantic segmentation Zhang et al. (2024) [ArXiv GitHub]. Recent advances in VLMs have increasingly focused on memory mechanisms as a crucial component for enhancing model capabilities and efficiency.

The fundamental architecture of VLMs typically consists of separate encoders for visual and textual inputs, with various mechanisms to facilitate cross-modal interaction Encord (2024) [Encord]. These models leverage contrastive learning methods to establish connections between the visual and language domains, allowing them to perform zero-shot predictions without task-specific fine-tuning Zhang et al. (2023) [ArXiv].

2.3 Visual Prompt Tuning (VPT)

Visual Prompt Tuning (VPT) has emerged as a parameter-efficient alternative to full fine-tuning for adapting vision models to downstream tasks Jia et al. (2022) [ArXiv SpringerLink]. VPT introduces only a small amount of trainable parameters (less than 1% of model parameters) in the input space while keeping the model backbone frozen.

The core mechanism of VPT involves prepending learnable prompt tokens to the input sequence of each Transformer layer Papers With Code [Paperswithcode]. These tokens are learned together with a linear head during fine-tuning, allowing the model to adapt to specific tasks with minimal parameter updates.

Research has demonstrated that VPT achieves significant performance gains compared to other parameter-efficient tuning protocols and even outperforms full fine-tuning in many cases across model capacities and training data scales.

2.4 Memory in Video Understanding Systems

Video understanding systems face unique memory challenges due to temporal data and the need for extended context maintenance. Recent surveys explore specialized memory architectures for video processing.

Nguyen et al. (2024) examine video-language understanding systems from architectural, training, and data perspectives, highlighting memory mechanisms for temporal coherence across frames. Tang et al. (2023) focus specifically on video understanding with Large Language Models, categorizing approaches based on temporal information handling—from frame-based encoders to sophisticated temporal encoders maintaining context across sequences.

Koprinska and Carrato established foundational principles for temporal video segmentation that continue to influence modern approaches. Research on Hierarchical Temporal Memory (HTM) examines biologically-inspired architectures for video processing requiring temporal coherence.

The evolution of video memory systems has progressed from simple temporal buffers to sophisticated hierarchical architectures mimicking human cognition, with recent advances leveraging large language models to integrate multimodal information across extended temporal sequences.

3 Unified Framework for Cross-Domain Memory Systems

Memory systems have become a critical component in advanced AI architectures across domains. To systematically analyze these systems, we propose a unified theoretical framework that examines memory systems through three complementary lenses: retrieval mechanisms, memory structures, and update schemas. This framework enables us to identify common patterns, domain-specific optimizations, and opportunities for cross-domain knowledge transfer.

3.1 Retrieval Mechanism Taxonomy

The retrieval mechanism determines how information is accessed from memory and significantly impacts both performance and efficiency. We categorize retrieval mechanisms into four primary approaches based on their underlying principles and implementation strategies.

Table 1: Taxonomy of Retrieval Mechanisms Across AI Domains

Category	**Key Approaches**	**Analysis Dimensions**
Similarity-Based		
	• Semantic search	• Retrieval precision
	• Contextual similarity	• Context sensitivity
	• Embedding-based retrieval	• Computational complexity
Prompt-Based		
	• Direct prompt selection	• Prompt transferability
	• Compositional prompting	• Task adaptation capability
	• Task-specific prompt banks	• Memory efficiency
Temporal-Spatial		
	• Temporal correlation	• Temporal consistency
	• Spatial attention	• Spatial coherence
	• Sequence modeling	• Processing efficiency
Hybrid Methods		
	• Multi-modal integration	• Integration effectiveness
	• Cross-attention mechanisms	• Modal alignment
	• Ensemble strategies	• Robustness to domain shift

3.1.1 Similarity-Based Retrieval

Similarity-based retrieval mechanisms use vector representations to find information in memory by computing similarity measures between queries and stored items. These approaches are prevalent in embedding-based systems and often serve as the foundation for more complex retrieval methods. The effectiveness of similarity-based retrieval depends on:

1. **Retrieval Precision**: How accurately the system identifies relevant information based on similarity metrics
2. **Context Sensitivity**: The ability to adapt similarity computations based on contextual factors
3. **Computational Complexity**: The efficiency of similarity calculations, especially for large-scale memory systems

3.1.2 Prompt-Based Retrieval

Prompt-based retrieval mechanisms rely on structured prompts to access information from memory. These approaches are particularly common in language and vision-language models where prompts serve as interfaces for knowledge retrieval. Key considerations include:

1. **Prompt Transferability**: How well prompts generalize across different tasks or domains
2. **Task Adaptation Capability**: The ability to customize prompts for specific tasks
3. **Memory Efficiency**: How efficiently prompts can be stored and accessed

3.1.3 Temporal-Spatial Retrieval

Temporal-spatial retrieval mechanisms leverage temporal and spatial relationships to access information. These approaches are especially important in video understanding systems and sequential data processing. Critical factors include:

1. **Temporal Consistency**: Maintaining coherent information retrieval across time steps
2. **Spatial Coherence**: Preserving spatial relationships during retrieval
3. **Processing Efficiency**: The computational overhead of tracking temporal-spatial relationships

3.1.4 Hybrid Methods

Hybrid retrieval approaches combine multiple mechanism types to leverage their complementary strengths. These methods are increasingly common in complex multi-modal systems. Key considerations include:

1. **Integration Effectiveness**: How well different retrieval mechanisms are combined
2. **Modal Alignment**: Aligning information across different modalities
3. **Robustness to Domain Shift**: Maintaining performance when distribution changes occur

3.2 Memory Structure Classification

Memory structures define how information is organized, stored, and accessed. Different structural approaches offer various trade-offs between efficiency, adaptability, and complexity.

Table 2: Classification of Memory Structures Across AI Domains

Structure Type	Architectural Patterns	Reported Advantages/Limitations
Static Memory	• Fixed capacity memories • Pre-trained prompt repositories • Immutable knowledge bases	• + Low maintenance overhead • + Predictable performance • - Limited adaptability • - Scale constraints
Dynamic Memory	• Expandable memory banks • Continually updated embeddings • Adaptive storage allocation	• + Adaptability to new data • + Scalability with task growth • - Higher computational overhead • - Potential for memory corruption
Hierarchical Memory	• Multi-level access structures • Priority-based organization • Cache-like architectures	• + Efficient information access • + Organized knowledge storage • - Complex management logic • - Increased design complexity
Distributed Memory	• Federated storage systems • Shared knowledge repositories • Decentralized architectures	• + Collaborative knowledge sharing • + Enhanced privacy potential • - Synchronization challenges • - Consistency maintenance costs

3.2.1 Static Memory

Static memory structures maintain fixed memory configurations that remain largely unchanged during operation. These structures are common in systems with well-defined, stable knowledge requirements. Key characteristics include:

1. **Fixed Capacity**: Pre-determined memory allocation that doesn't change during operation
2. **Immutable Content**: Memory contents that remain stable over time
3. **Optimized Access Patterns**: Access mechanisms tailored to the fixed structure

Static memory offers predictable performance and low maintenance overhead but suffers from limited adaptability and potential scale constraints as requirements grow.

3.2.2 Dynamic Memory

Dynamic memory structures can expand, contract, or reorganize during operation to accommodate changing information needs. These structures are essential for systems that must adapt to new data or evolving tasks. Key aspects include:

1. **Expandable Storage**: The ability to allocate additional memory as needed
2. **Continuous Updates**: Mechanisms for modifying memory contents over time
3. **Adaptive Organization**: Reorganization capabilities based on usage patterns

Dynamic memory provides excellent adaptability and scalability but typically requires more complex management and may incur higher computational costs.

3.2.3 Hierarchical Memory

Hierarchical memory structures organize information in multiple levels, often with different access characteristics at each level. This organization resembles human memory systems and cache hierarchies in computer architecture. Important features include:

1. **Multi-Level Access**: Different access patterns and speeds at each level
2. **Priority Organization**: Information organized by importance or relevance
3. **Efficient Retrieval Paths**: Optimized paths for accessing different types of information

Hierarchical structures enable efficient information access and organized knowledge storage but require more complex management logic and design.

3.2.4 Distributed Memory

Distributed memory structures spread information across multiple storage locations, often geographically or functionally separated. These approaches are increasingly important in collaborative and privacy-conscious systems. Key characteristics include:

1. **Federated Storage**: Information distributed across multiple locations
2. **Shared Access Protocols**: Mechanisms for accessing distributed information
3. **Localized Processing**: Computation performed near storage when possible

Distributed memory supports collaborative knowledge sharing and enhanced privacy but faces challenges in synchronization and consistency maintenance.

3.3 Memory Update Schema Analysis

Memory update schemas determine how information in memory evolves over time. These mechanisms significantly impact a system's ability to adapt to new information while preserving critical knowledge.

3.3.1 Frequency-Based Updates

Frequency-based update schemas prioritize information based on how often it is accessed or used. These approaches are inspired by caching algorithms like Least Frequently Used (LFU) and are common in systems where usage patterns are strong indicators of importance. Key mechanisms include:

1. **Usage Tracking**: Monitoring how frequently items are accessed

Table 3: Analysis of Memory Update Schemas Across AI Domains

Update Approach	Core Mechanisms	Theoretical Implications
Frequency-Based	• Usage tracking • Popularity-based retention • LFU/LRU-inspired approaches	• Potential bias toward common patterns • Efficient for repetitive tasks • Performance in long-tail scenarios
Recency-Based	• Temporal prioritization • Time-decay functions • Recent-first strategies	• Adaptation to concept drift • Handling temporal dynamics • Historical information loss
Importance-Weighted	• Value estimation models • Critical information retention • Salience detection	• Attention mechanism effectiveness • Information preservation quality • Computational overhead for value assessment
Privacy-Preserving	• Anonymization techniques • Differential privacy approaches • Federated updates	• Privacy-utility tradeoffs • Information leakage risks • Compliance with privacy standards

2. **Popularity-Based Retention**: Preserving frequently accessed information
3. **Access Pattern Analysis**: Identifying and optimizing for common access sequences

Frequency-based approaches work well for repetitive tasks but may struggle with long-tail scenarios and can develop biases toward common patterns.

3.3.2 Recency-Based Updates

Recency-based update schemas prioritize recent information over older data. These approaches are inspired by cache replacement policies like Least Recently Used (LRU) and are valuable in dynamic environments where relevance changes over time. Important aspects include:

1. **Temporal Prioritization**: Giving preference to recently accessed information
2. **Time-Decay Functions**: Gradually reducing the importance of older information
3. **Temporal Windowing**: Focusing on information within recent time frames

Recency-based methods adapt well to concept drift and changing environments but risk losing valuable historical information.

3.3.3 Importance-Weighted Updates

Importance-weighted update schemas evaluate the significance of information using learned or heuristic value functions. These approaches attempt to preserve critical information regardless of recency or frequency. Key elements include:

1. **Value Estimation**: Assessing the importance of each piece of information
2. **Critical Information Preservation**: Maintaining essential knowledge regardless of usage patterns
3. **Attention Mechanisms**: Using learned attention to identify significant information

Importance-weighted approaches excel at preserving crucial information but typically require more complex computational mechanisms to assess value.

3.3.4 Privacy-Preserving Updates

Privacy-preserving update schemas incorporate techniques to protect sensitive information during memory updates. These approaches are increasingly important as AI systems handle more personal and confidential data. Critical mechanisms include:

1. **Anonymization**: Removing or obscuring identifying information
2. **Differential Privacy**: Adding calibrated noise to protect individual data points
3. **Federated Updates**: Performing updates locally before aggregating changes

Privacy-preserving methods must carefully balance privacy protection with utility, managing the inherent trade-offs between these competing objectives.

Our unified framework provides a comprehensive approach for analyzing memory systems across diverse AI domains. By examining systems through these three complementary perspectives—retrieval mechanisms, memory structures, and update schemas—we can identify common patterns, unique innovations, and transferable techniques that span domains from language processing to vision-language models and video understanding systems.

4 Single domain analysis

4.1 Large Language Models (LLMs)

4.1.1 Overview of Memory Requirements in LLMs

Large Language Models present unique memory challenges stemming from several critical factors:

- **Context length limitations**: Standard transformer architectures have fixed context windows, limiting their ability to access information from distant parts of a conversation or document.
- **Knowledge integration challenges**: Models must effectively integrate parametric knowledge (learned during training) with external knowledge retrieved at inference time.
- **Computational efficiency concerns**: Memory mechanisms must balance expressiveness with computational requirements, especially for deployment in resource-constrained environments.

- **Coherence across extended interactions**: LLMs need to maintain consistent understanding over long conversations, requiring effective mechanisms to recall previous interactions.

Recent advances have introduced innovative memory mechanisms to address these challenges, ranging from retrieval-augmented generation to explicit memory architectures and neuroscience-inspired approaches.

4.1.2 Retrieval Mechanism Analysis

Our analysis of LLM memory systems reveals several key patterns in retrieval mechanisms across different approaches. Table 4.1 provides a comparative overview of retrieval mechanisms employed by various LLM systems, organized by their primary approach to illustrate the evolution of research trends.

Table 4: Comparison of Retrieval Mechanisms in LLM Memory Systems

System	Year	Primary Approach	Key Mechanisms	Notable Characteristics
Similarity-Based Approaches				
MemGPT	2023	Similarity-based	Context chunking, LLM-based retrieval selection	Autonomous memory management; virtual context expansion beyond model limits
A-Mem	2023	Similarity-based	Approximate nearest neighbor, Concept tree searching	Fast indexing; generalizable retrieval through concept mapping
memoryLLM	2023	Similarity-based	Semantic chunking, Key-value memory, Multi-vector indexing	Multi-vector approach enhances retrieval precision
ChatDB	2024	Similarity-based	Structured database as memory, SQL-based retrieval	Combines symbolic and neural approaches; strong structured memory capabilities
Temporal-Based Approaches				
MemBank	2023	Temporal-based	Recency weighting, Episodic buffer	Explicit splitting of episodic and semantic memory
SCM	2023	Temporal-based	Streaming pruning, Stream compression	Progressive summarization for extended dialogues
Hybrid Approaches				
Memory LLM-agent	2023	Hybrid (Similarity + Temporal)	Multi-granularity memory bank, Attention-based retrieval	Autonomous memory management with multi-level organization
RET-LLM	2023	Hybrid (Similarity + Prompt-based)	Reflective trigger mechanism, Chain-of-thought retrieval	Self-reflection determines when to retrieve from memory
Think-in-Mem	2024	Hybrid (Similarity + Prompt-based)	Deliberate thinking, Memory-centric reasoning paths	Mimics human memory utilization during complex reasoning
MyAgent	2024	Hybrid (Temporal + Similarity)	Event-based recollection, Self-distillation	Memory compression based on behavioral outcomes

4.1.3 Dominant Retrieval Patterns

The chronological organization of retrieval approaches reveals clear evolutionary trends in how LLM memory systems access information:

1. **Movement Toward Hybrid Approaches**: Early systems primarily employed single-strategy approaches (either similarity or temporal-based), while more recent systems increasingly combine

multiple retrieval mechanisms. This trend reflects growing recognition that no single retrieval strategy adequately addresses the diverse information needs of complex language tasks. For instance, RET-LLM (2023) integrates similarity-based retrieval with prompt-based reflection, while Think-in-Mem (2024) employs a sophisticated combination of deliberate thinking and memory-centric reasoning.

2. **Increasing Autonomy in Retrieval Decisions**: A noticeable progression toward systems that autonomously determine when and what to retrieve, rather than using fixed retrieval strategies. MemGPT (2023) introduced basic autonomous memory management, while later systems like RET-LLM (2023) implement sophisticated self-reflection mechanisms to determine appropriate retrieval timing.

3. **Growing Integration of Structured Knowledge**: Recent approaches increasingly incorporate structured representations alongside traditional vector embeddings. ChatDB (2024) implements a fully structured database approach to memory, while A-Mem (2023) uses concept trees to enable more generalizable retrieval.

4. **Evolution of Semantic Granularity**: Earlier systems operated primarily at the level of full paragraphs or documents, while newer approaches like memoryLLM (2023) and Memory LLM-agent (2023) implement multi-granular approaches that can retrieve at different levels of semantic abstraction, from key concepts to detailed passages.

4.1.4 Innovative Retrieval Mechanisms

Several particularly innovative retrieval mechanisms have emerged in recent systems:

- **RET-LLM's Reflective Triggers**: RET-LLM introduced a "reflective trigger" mechanism that enables the model to determine when memory retrieval is necessary, similar to the tip-of-the-tongue phenomenon in human cognition. This approach significantly enhances retrieval precision by avoiding unnecessary retrievals that might derail coherent reasoning.

- **Think-in-Mem's Deliberate Thinking**: Think-in-Mem implements a novel approach where the model explicitly "thinks" about what it needs to retrieve before conducting memory search, leading to more focused and relevant retrievals.

- **MyAgent's Event-Based Recollection**: MyAgent organizes memories around events rather than just semantic content, enabling more contextually appropriate retrievals based on both what happened and when it occurred.

- **A-Mem's Concept Tree**: A-Mem employs a hierarchical concept tree to enable retrieval of semantically related information even when exact matches are unavailable, addressing a key limitation of pure embedding-based approaches.

These innovations demonstrate a clear trend toward retrieval mechanisms that more closely mimic human memory processes, with increased metacognitive awareness and contextual sensitivity.

4.1.5 Memory Structure Classification

LLM memory systems employ diverse organizational structures to manage information effectively. Table 4.2 provides a comparative overview of memory structures implemented in various LLM systems, organized by their primary structure type to illustrate structural evolution over time.

4.1.6 Structural Evolution Patterns

The chronological analysis of memory structures reveals several key evolutionary trends:

Table 5: Comparison of Memory Structures in LLM Memory Systems

System	Year	Primary Structure	Key Components	Notable Characteristics
Dynamic Memory Approaches				
MemGPT	2023	Dynamic Memory	Main context, Virtual memory swap	System-inspired memory management with paging mechanisms
memoryLLM	2023	Dynamic Memory	Expandable memory slots, Dynamic indexing	Efficient scaling to large memory stores
ChatDB	2024	Dynamic + Structured	SQL database, Dynamic schema adaptation	Symbolic structure with neural interface
Hierarchical Memory Approaches				
MemBank	2023	Hierarchical Memory	Episodic buffer, Long-term store, Knowledge base	Three-tier system inspired by human memory models
Memory LLM-agent	2023	Hierarchical Memory	Core memory, Archival memory, Knowledge tools	Progressive organization from immediate to archived information
A-Mem	2023	Hierarchical Memory	Concept trees, Instance stores	Ontological organization enhancing generalization
Specialized Memory Approaches				
SCM	2023	Streaming Memory	Stream buffer, Compressed representations	Optimized for continuous information flow
RET-LLM	2023	Associative Memory	Experience store, Reflection module	Connects experiences through associative links
Think-in-Mem	2024	Reasoning-Oriented Memory	Reasoning cache, Factual store	Specialized structures for different thinking modes
MyAgent	2024	Event-Based Memory	Experience logs, Behavioral models	Organizes memories around discrete events and outcomes

1. **Increasing Structural Sophistication**: Early approaches implemented relatively straightforward memory structures, while more recent systems feature complex, multi-component architectures. MemGPT (2023) introduced basic paging mechanisms inspired by operating systems, while later systems like Memory LLM-agent (2023) implement sophisticated multi-tier architectures with specialized components for different types of information.

2. **Growing Biological Inspiration**: Recent memory structures increasingly draw inspiration from cognitive science and neuroscience models of human memory. MemBank (2023) explicitly implements components inspired by episodic, semantic, and working memory, while Think-in-Mem (2024) models different reasoning processes associated with distinct memory systems.

3. **Shift Toward Specialized Components**: Later systems tend to implement specialized memory components for different types of information or different cognitive functions. Memory LLM-agent (2023) uses distinct stores for core, archival, and knowledge information, while MyAgent (2024) maintains separate structures for experiences and behavioral patterns.

4. **Integration of Symbolic and Neural Approaches**: Recent systems increasingly combine neural representations with symbolic structures. ChatDB (2024) implements a fully symbolic database structure with a neural interface, while A-Mem (2023) uses concept trees to provide symbolic organization of neural embeddings.

4.1.7 Key Architectural Innovations

Several particularly innovative structural approaches have emerged:

- **MemGPT's Virtual Memory**: MemGPT implements a virtual memory system inspired by operating system design, with paging mechanisms that move information between active context

and external storage. This approach effectively extends the context window while maintaining computational efficiency.

- **ChatDB's SQL Integration**: ChatDB integrates a fully-featured SQL database as its memory structure, enabling powerful structured queries and complex relational operations that are difficult to implement in pure vector stores.
- **Think-in-Mem's Reasoning Cache**: Think-in-Mem introduces a specialized memory structure for caching intermediate reasoning steps, allowing the model to revisit and refine its reasoning process over time.
- **A-Mem's Concept Trees**: A-Mem organizes memories in hierarchical concept trees that enable generalization across related concepts, addressing a key limitation of flat memory structures.

These structural innovations demonstrate a trend toward increasingly sophisticated organization that balances computational efficiency with cognitive plausibility.

4.1.8 Memory Update Schema Analysis

LLM memory systems employ diverse strategies for updating stored information. Table 4.3 provides a comparative overview of update schemas implemented across various LLM approaches, organized by their primary update approach to illustrate the evolution of update strategies.

4.1.9 Update Strategy Evolution

The chronological organization of update schemas reveals important trends in how LLM memory systems determine what information to preserve or modify:

1. **From Simple to Sophisticated Criteria**: Early systems primarily employed straightforward criteria like recency (MemGPT, 2023), while later approaches introduced increasingly sophisticated mechanisms for determining information value. Think-in-Mem (2024) implements deliberate reasoning about information utility, while MyAgent (2024) employs a complex multi-factor approach considering event significance, outcome correlation, and usage patterns.
2. **Increasing Agency in Memory Management**: A clear trend toward systems that actively manage their own memories rather than using fixed update rules. Memory LLM-agent (2023) explicitly models memory management as agent actions, while Think-in-Mem (2024) implements deliberate decision-making about what to retain or forget.
3. **Growing Emphasis on Causal Understanding**: Recent systems increasingly preserve information based on its causal significance rather than just its semantic importance. MyAgent (2024) explicitly tracks correlations between memories and behavioral outcomes, preferentially preserving memories that led to successful interactions.
4. **Integration of Multiple Update Criteria**: Later systems tend to combine multiple update criteria rather than relying on a single approach. memoryLLM (2023) balances recency with semantic importance, while MyAgent (2024) integrates recency, importance, and frequency considerations in a comprehensive update strategy.

4.1.10 Notable Update Mechanisms

Several particularly innovative update mechanisms have emerged:

- **RET-LLM's Reflection-Based Consolidation**: RET-LLM implements a novel approach where the model explicitly reflects on its experiences to determine which should be preserved in long-term memory, mimicking human memory consolidation during reflection.

Table 6: Comparison of Memory Update Approaches in LLM Memory Systems

System	Year	Primary Approach	Key Mechanisms	Notable Characteristics
Recency-Based Approaches				
MemGPT	2023	Recency-Based	FIFO context management, LRU-based eviction	Automated swapping between active and archived memory
SCM	2023	Recency-Based	Streaming compression, Temporal decay	Progressive summarization of older information
Importance-Weighted Approaches				
A-Mem	2023	Importance-Weighted	Concept importance scoring, Tree restructuring	Preserves conceptually significant information
MemBank	2023	Importance-Weighted	Salience detection, Semantic consolidation	Selective information persistence based on importance
RET-LLM	2023	Importance-Weighted	Reflection-based consolidation, Experience tagging	Self-evaluation of memory importance
Hybrid Approaches				
memoryLLM	2023	Hybrid (Recency + Importance)	Semantic chunking, Dual-phase indexing	Balances temporal relevance with semantic importance
Memory LLM-agent	2023	Hybrid (Recency + Importance)	Active forgetting, Importance-based consolidation	Explicit memory management through agent actions
Think-in-Mem	2024	Hybrid (Recency + Importance)	Deliberate consolidation, Utility assessment	Strategic memory updates based on reasoning utility
MyAgent	2024	Hybrid (Recency + Importance + Frequency)	Event significance, Outcome correlation, Usage tracking	Multi-factor consolidation mimicking human memory formation
Structured (Schema-Based) Approaches				
ChatDB	2024	Structured (Schema-Based)	Schema evolution, Relational updates	Formal database update operations with neural guidance

- **Think-in-Mem's Utility Assessment**: Think-in-Mem evaluates the utility of memories based on their contribution to successful reasoning, preferentially preserving information that has proven valuable for problem-solving.
- **MyAgent's Outcome Correlation**: MyAgent tracks correlations between memories and interaction outcomes, strengthening memories that led to successful outcomes and weakening those associated with failures.
- **ChatDB's Schema Evolution**: ChatDB implements a structured approach to memory updates by evolving its database schema over time to better represent accumulated knowledge, combining formal database principles with neural flexibility.

These innovations demonstrate a trend toward update mechanisms that more closely mimic human memory consolidation, with increased metacognitive awareness and goal-directed memory formation.

4.1.11 Discussion and Key Findings

Our analysis of LLM memory systems through the unified framework reveals several important insights and trends:

Universal Memory Patterns Several universal memory patterns identified in our framework are consistently present across LLM systems throughout the analyzed time period:

1. **Multi-Tier Organization**: The most effective systems consistently implement multi-tier memory structures that separate immediate context from longer-term storage, enabling efficient processing while maintaining access to extended history.

2. **Metacognitive Control**: Advanced systems increasingly implement metacognitive mechanisms that regulate memory operations, determining when to retrieve information and what to retain based on task demands.

3. **Balancing Generalization and Specificity**: Successful memory systems carefully balance specific episodic memories with generalized semantic knowledge, enabling both precise recall and broad application of learned patterns.

Domain-Specific Optimizations While adhering to universal memory patterns, LLM systems have implemented distinctive optimizations to address the specific challenges of language processing:

1. **Semantic Chunking Strategies**: LLM memory systems employ sophisticated chunking strategies that preserve semantic coherence rather than using fixed-size divisions, significantly enhancing retrieval precision.

2. **Contextual Relevance Filtering**: Advanced systems implement context-aware filtering mechanisms that retrieve information based on relevance to the current conversation state rather than just semantic similarity.

3. **Narrative Coherence Preservation**: Recent approaches prioritize maintaining narrative coherence across extended interactions, preserving causal relationships and conversational flow even when retrieving from distant parts of a dialogue.

Emerging Research Directions Based on our analysis, several promising research directions emerge for next-generation LLM memory systems:

1. **Cognitive Architecture Integration**: Future systems could more deeply integrate principles from cognitive architectures like ACT-R or SOAR, implementing unified theories of memory that span working, episodic, semantic, and procedural components.

2. **Explainable Memory Operations**: An important frontier is developing memory systems that can explain their own retrieval and storage decisions, enabling users to understand why certain information was remembered or forgotten.

3. **Adaptive Memory Allocation**: Next-generation systems could dynamically adjust memory allocation based on task demands, dedicating more resources to memory-intensive tasks while operating efficiently for simpler interactions.

4. **Cross-Episode Knowledge Transfer**: A critical challenge is developing systems that can effectively transfer knowledge between different conversations or documents, building cumulative understanding across separate interactions.

5. **Personalized Memory Adaptation**: Future systems could adapt their memory mechanisms to individual users, learning which types of information particular users tend to reference or find valuable.

Technical Challenges and Solutions Our analysis highlights several technical challenges facing LLM memory systems and emerging solutions:

1. **Embedding Drift Over Time**: As conversations evolve, semantic drift can reduce retrieval effectiveness. Recent systems address this through periodic re-embedding (memoryLLM) or hierarchical concept structures (A-Mem).

2. **Computational Efficiency at Scale**: Memory operations become computationally expensive as memory stores grow. Solutions include hierarchical indexing (Memory LLM-agent), progressive compression (SCM), and selective retention (Think-in-Mem).

3. **Balancing Precision and Recall**: Memory systems must balance retrieving exactly relevant information (precision) with finding all potentially relevant information (recall). Hybrid approaches like RET-LLM's reflective retrieval and Think-in-Mem's deliberate thinking address this trade-off effectively.

4. **Integration with External Knowledge**: Integrating parametric knowledge (learned during training) with external memory presents ongoing challenges. Systems like ChatDB demonstrate promising approaches through structured integration of symbolic and neural components.

The field of LLM memory systems has evolved rapidly from relatively simple approaches to sophisticated architectures that increasingly resemble human memory in both structure and function. As these systems continue to develop, we can expect further convergence with cognitive science models and increased capabilities for long-term reasoning, learning from experience, and maintaining coherence across extended interactions.

4.2 Vision-Language Models (VLMs)

4.2.1 Overview of Memory Requirements in VLMs

Vision-language models have distinctive memory requirements stemming from several factors:

- **Cross-modal alignment**: VLMs must maintain alignments between visual and textual representations, requiring memory systems that can establish and preserve these cross-modal relationships.

- **Modality-specific processing**: Different modalities often require specialized processing before integration, necessitating memory structures that can accommodate heterogeneous data types.

- **Efficient parameter management**: Due to the computational cost of processing visual information, VLM memory systems must be highly parameter-efficient.

- **Adaptability to diverse tasks**: VLMs are applied to a wide range of tasks (classification, retrieval, captioning), requiring flexible memory architectures.

Recent advances have introduced novel memory mechanisms to address these challenges, with approaches ranging from memory-inspired temporal interactions to dual memory architectures and memory-space visual prompting.

4.2.2 Retrieval Mechanism Analysis

Our analysis of VLM memory systems reveals several key patterns in retrieval mechanisms across different approaches. Table 4.1 provides a comparative overview of retrieval mechanisms employed by various VLM systems, organized by their primary approach to illustrate the evolution of research trends.

Table 7: Comparison of Retrieval Mechanisms in VLM Memory Systems

System	Year	Primary Approach	Key Mechanisms	Notable Characteristics
Similarity-Based Approaches				
Memory-Space Visual Prompting (MemVP)	2023	Similarity-based	Key-value memory function in FFNs, Memory-space integration	Injects visual information directly into LLM memory space rather than input space
Prompt-Based Approaches				
Generalizable Prompt Tuning	2022	Prompt-based	Mutual information maximization, Class-wise augmentation	Balances hand-crafted and learnable prompts for both performance and generalization
Conditional Prompt Tuning	2023	Prompt-based	Mixture of Prompt Experts (MoPE), Dynamic routing	Uses one modality to guide the prompting of another for cross-modal knowledge transfer
Hybrid Approaches				
Dual-Memory Model	2021	Hybrid	Feature-based retrieval (CNN), Random forest classification, Sequential processing	Follows a pathway from STM to WM to LTM, mimicking human cognitive processing
MITP	2022	Hybrid	Similarity-based temporal prompt calculations, Cross-modal memory hub	Uses activation vectors to control information flow across modalities
SynapticRAG	2023	Hybrid	Cosine similarity filtering, Dynamic time warping for temporal representations	Combines semantic and temporal similarity through a binding-score mechanism
Dual Memory Networks (DMN)	2024	Hybrid	Cosine similarity filtering, Attention-based memory interaction	Generates sample-adaptive classifiers for each test sample

4.2.3 Dominant Retrieval Patterns

The chronological organization of retrieval approaches reveals a clear evolution from simpler, single-approach methods toward increasingly sophisticated hybrid mechanisms:

1. **Trend Toward Hybrid Approaches**: The field has shifted from specialized retrieval methods toward hybrid approaches that combine multiple mechanisms. This trend reflects the growing recognition that no single retrieval strategy is sufficient for the complex demands of vision-language tasks. Early models like the Dual-Memory Model (2021) implemented basic hybrid approaches, while more recent systems like DMN (2024) feature sophisticated combinations of similarity-based filtering, cross-attention mechanisms, and adaptive classification.

2. **Increasing Focus on Cross-Modal Integration**: Over time, VLM retrieval mechanisms have placed greater emphasis on effective cross-modal integration. While earlier approaches treated visual and textual information more separately, newer methods like Conditional Prompt Tuning (2023) and MITP (2022) explicitly model the interactions between modalities, enabling more effective knowledge transfer.

3. **Growing Incorporation of Temporal-Spatial Information**: Recent models increasingly incorporate temporal or sequential information in their retrieval mechanisms. SynapticRAG (2023) implements dynamic time warping to differentiate memories based on temporal occurrence, while MITP (2022) leverages temporal prompts across intermediate layers to capture sequential information flow.

4.2.4 Evolution Toward Adaptive Routing

A notable progression in VLM retrieval mechanisms is the evolution from fixed retrieval patterns toward adaptive, context-sensitive routing strategies:

- Early systems relied on static prompts or fixed similarity calculations
- Intermediate approaches like MITP (2022) introduced more flexible cross-modal attention mechanisms
- Recent systems like Conditional Prompt Tuning (2023) implement dynamic routing through mechanisms like Mixture of Prompt Experts (MoPE)
- The latest models like DMN (2024) generate completely sample-adaptive classifiers tailored to each test point

This evolution demonstrates the field's recognition that context-sensitive, adaptive retrieval is essential for handling the diverse and multimodal nature of vision-language tasks.

4.2.5 Memory Structure Classification

Vision-language models employ diverse memory structures to handle cross-modal information. Table 4.2 provides a comparative overview of memory structures implemented in various VLM systems, organized by their primary structure type to illustrate structural evolution over time.

4.2.6 Structural Patterns and Trends

The chronological organization of memory structures reveals several key trends in VLM memory architecture:

1. **Evolution Toward Hybrid Architectures**: The field has progressively moved toward hybrid memory structures that combine static and dynamic components. The earliest analyzed approach (Dual-Memory Model, 2021) utilized a primarily hierarchical structure, while more recent approaches like Conditional Prompt Tuning (2023) and DMN (2024) implement sophisticated hybrid architectures that combine multiple structure types to balance stability with adaptability.

2. **Increasing Structural Sophistication**: Over time, VLM memory structures have grown more complex, with specialized components for different functions. Early models employed simpler memory organizations, while newer approaches like Conditional Prompt Tuning (2023) disentangle prompt vectors into static, dynamic, and mapped types and incorporate distributed memory elements through mixture of experts.

3. **Growing Biological Inspiration**: A notable trend is the increasing influence of cognitive science and neuroscience on VLM memory structures. The Dual-Memory Model (2021) drew explicit inspiration from Baddeley's psychological theory of human memory, while SynapticRAG (2023) incorporated neuroscience-inspired mechanisms like weighted spike trains and stimulus-based node activation.

4.2.7 Key Trade-offs in Memory Structure Design

Throughout the evolution of VLM memory structures, several consistent trade-offs have been addressed:

Table 8: Comparison of Memory Structures in VLM Memory Systems

System	Year	Primary Structure	Key Components	Notable Characteristics
Static Memory Approaches				
No pure static memory approaches found in analyzed VLM systems				
Dynamic Memory Approaches				
MITP	2022	Dynamic + Hierarchical	Temporal prompts, Memory hub	Layer-wise information flow with bidirectional exchange between modalities
SynapticRAG	2023	Dynamic + Hierarchical	Weighted spike trains, Parent-child node activation	Memory nodes adapt based on stimulation patterns
Hierarchical Memory Approaches				
Dual-Memory Model	2021	Hierarchical	STM (FIFO queue), WM (processing components), LTM (explicit/implicit memory)	Multi-level structure inspired by human memory psychology
Hybrid Memory Approaches				
Generalizable Prompt Tuning	2022	Hybrid (Static + Dynamic)	Hand-crafted prompts (static), Learnable soft prompts (dynamic)	Balances stability of hand-crafted prompts with adaptability of learnable prompts
Memory-Space Visual Prompting (MemVP)	2023	Hybrid	Original FFN weights (static), Position-embedded visual prompts (dynamic)	Maintains core LLM knowledge while injecting visual information
Dual Memory Networks (DMN)	2024	Hybrid (Static + Dynamic)	Dynamic memory (test samples), Static memory (training data)	Preserves historical information while maintaining stable knowledge
Conditional Prompt Tuning	2023	Hybrid + Distributed	Static, Dynamic, and Mapped prompts; Mixture of experts	Different experts specialize in different instance types

- **Adaptability vs. Stability**: Dynamic components provide adaptability to new information and tasks, while static elements ensure stability and consistent performance. This trade-off is evident in the progression from predominantly dynamic or hierarchical approaches to hybrid structures that carefully balance these aspects.

- **Efficiency vs. Expressiveness**: As structures have grown more sophisticated, models have implemented various mechanisms to maintain computational efficiency while enabling expressive representations. These include hierarchical organizations (MITP, 2022), specialized components (Conditional Prompt Tuning, 2023), and efficient integration techniques (MemVP, 2023).

- **Complexity vs. Interpretability**: While memory structures have grown more complex over time, many recent approaches have addressed interpretability concerns through clear functional separation of components, as seen in DMN's (2024) explicit division of memory for historical test samples and training data.

4.2.8 Memory Update Schema Analysis

The update mechanisms employed by VLM memory systems reveal sophisticated strategies for balancing multiple factors. Table 4.3 provides a comparative overview of update schemas implemented across various

VLM approaches, organized by their primary update approach to illustrate the evolution of update strategies.

Table 9: Comparison of Memory Update Schemas in VLM Memory Systems

System	Year	Primary Approach	Key Mechanisms	Notable Characteristics
Frequency-Based Approaches				
No pure frequency-based approaches found in analyzed VLM systems				
Recency-Based Approaches				
SynapticRAG	2023	Combined (Recency, Importance, Frequency)	Time constant updates, Binding score, Stimulus-based firing	Implements natural decay and frequency-based prioritization
Importance-Weighted Approaches				
Generalizable Prompt Tuning	2022	Importance-Weighted	MI estimator, Class-wise augmentation	Preserves information through mutual information maximization
MITP	2022	Importance-Weighted + Recency-Based	SoftMax activation vectors, Layer-wise temporal prompts	Requires only 2.0M trainable parameters (1% of foundation model)
Conditional Prompt Tuning	2023	Importance-Weighted	Learned routing scores, Importance loss regularization	Routing scores: $r = \text{Softmax}(W_r \psi_y / \tau + \epsilon)$
Memory-Space Visual Prompting (MemVP)	2023	Importance-Weighted	Position embeddings, Key-value integration	Reduces training time by 1.7× and inference by 1.4×
Dual Memory Networks (DMN)	2024	Importance-Weighted	Attention-based weighting, Multi-source integration	Combines knowledge from text input, historical data, and training data
Combined Approaches				
Dual-Memory Model	2021	Combined	ULS, IGT, Class-conditional weighting, Sequential backward selection	Balances new information against existing knowledge

4.2.9 Update Patterns and Trends

The chronological organization of update schemas reveals several important trends in VLM memory update mechanisms:

1. **Dominance of Importance-Weighted Approaches**: Across the timeline, importance-weighted update mechanisms have remained the dominant approach in VLM systems. This persistence highlights the fundamental importance of selectively prioritizing information based on its significance, regardless of other architectural changes.

2. **Increasing Sophistication of Weighting Mechanisms**: While importance-weighted approaches have remained prevalent, their implementation has grown more sophisticated over time. Early approaches used simpler weighting mechanisms, while recent models like Conditional Prompt Tuning (2023) implement complex routing scores with regularization, and DMN (2024) features multi-source integration.

3. **Trend Toward Multi-Factor Integration**: The field has gradually moved toward update schemas that balance multiple factors simultaneously. The Dual-Memory Model (2021) implemented a basic combined approach with multiple update strategies, while SynapticRAG (2023) features a sophisticated integration of recency, importance, and frequency considerations.

4.2.10 Efficiency Innovations

A consistent focus across the evolution of VLM update schemas has been on computational efficiency:

- **Parameter Efficiency**: There has been a steady improvement in parameter efficiency, with MITP (2022) requiring only 2.0M trainable parameters (approximately 1% of the foundation model) and Conditional Prompt Tuning (2023) achieving state-of-the-art performance with only 0.7% of trainable parameters.
- **Computational Efficiency**: Recent approaches have made significant strides in reducing computational requirements, with MemVP (2023) achieving 1.7× faster training and 1.4× faster inference compared to full fine-tuning.
- **Balanced Efficiency-Effectiveness Trade-offs**: The chronological progression shows increasingly sophisticated approaches to balancing efficiency with effectiveness, as seen in DMN's (2024) ability to achieve state-of-the-art performance across multiple adaptation settings with minimal parameter requirements.

4.2.11 Discussion and Key Findings

Our analysis of VLM memory systems through the unified framework reveals several important insights and trends:

Universal Memory Patterns Several universal memory patterns identified in our framework are consistently present across VLM systems throughout the analyzed time period:

1. **Temporal Priority Mechanisms**: From the Dual-Memory Model's (2021) sequential pipeline to MITP's (2022) layered approach to SynapticRAG's (2023) treatment of temporal information as a first-class feature, temporal prioritization has remained a constant theme in VLM memory systems.
2. **Context-Sensitive Retrieval**: All systems implement context-aware information retrieval, though the sophistication has increased over time, from basic context sensitivity in early models to DMN's (2024) highly adaptive classifiers tailored to each test sample.
3. **Compression-Accuracy Balancing**: Throughout the timeline, VLM systems have demonstrated effective trade-offs between parameter efficiency and performance, though the efficiency has steadily improved with technological advances.

Domain-Specific Optimizations While adhering to universal memory patterns, VLM systems have implemented increasingly sophisticated domain-specific optimizations to address the unique challenges of vision-language integration:

1. **Modality Integration Strategies**: The approaches to integrating visual and language modalities have diversified over time, from simpler integration in early models to sophisticated mechanisms like MITP's (2022) cross-attention memory hub, Conditional Prompt Tuning's (2023) modality-guided prompting, and MemVP's (2023) direct injection of visual information into language model memory.
2. **Parameter Efficiency Focus**: A consistent thread across the timeline has been the emphasis on parameter efficiency, with a clear trend toward achieving comparable or better performance with progressively fewer parameters.

3. **Biological Inspiration**: The influence of cognitive science and neuroscience on VLM memory systems has grown over time, from the Dual-Memory Model's (2021) explicit modeling of psychological theory to SynapticRAG's (2023) incorporation of neuroscience-inspired spike timing mechanisms.

Performance Insights The performance of these VLM memory systems has steadily improved over time:

- **Improved Accuracy**: Recent models like DMN (2024) outperform competitors by over 3% in zero-shot settings without using external training data
- **Enhanced Efficiency**: Newer approaches like MemVP (2023) match full fine-tuning performance while significantly reducing computational requirements
- **Better Generalization**: Recent systems demonstrate improved cross-dataset generalization and better scaling with training data

Future Directions Based on my chronological analysis, I've identified several promising research directions for VLM memory systems that go beyond the superficial:

1. **Dynamic Memory Allocation Strategies** Current systems show a trade-off between adaptability and stability in memory structures. Future research could develop more sophisticated dynamic allocation strategies that:
   - Automatically adjust memory capacity based on task complexity
   - Implement resource-aware memory management that optimizes for hardware constraints
   - Develop pruning techniques specifically designed for multimodal memory to eliminate redundancies across modalities
2. **Cross-modal Calibration and Alignment** While current approaches have sophisticated modality integration strategies, they still struggle with alignment problems. Future work could develop:
   - Self-supervised calibration mechanisms that continuously refine cross-modal alignments
   - Uncertainty-aware memory retrieval that accounts for confidence differences across modalities
   - Contrastive memory mechanisms that explicitly model differences between modalities to better understand their relationships
3. **Temporal-Aware Retrieval for Dynamic Environments** Recent models have begun incorporating temporal-spatial information in their retrieval mechanisms. Further development could include:
   - Memory systems specialized for video understanding that explicitly model object permanence
   - Event-based memory architectures that organize information around temporal events rather than static features
   - Causal memory structures that capture not just correlations between visual and language elements but causal relationships over time
4. **Hierarchical Compression with Guaranteed Fidelity** Compression-accuracy balancing has been a constant theme in these systems. Future work could develop:
   - Theoretically-grounded approaches to memory compression with formal guarantees on information retention
   - Task-aware compression techniques that preserve information relevant to downstream tasks
   - Progressive refinement memory mechanisms that store coarse representations with the ability to refine details when needed

The field of VLM memory systems has evolved rapidly from relatively simple architectures to sophisticated, hybrid systems that carefully balance multiple retrieval mechanisms, memory structures, and update schemas. This progression reflects the growing understanding of the unique challenges posed by vision-language tasks and the innovative approaches developed to address them. As the field continues to mature, we can expect to see further integration of interdisciplinary insights and increasingly efficient and effective memory architectures.

4.3 Visual Prompt Tuning (VPT)

Visual Prompt Tuning represents a significant advancement in parameter-efficient fine-tuning for computer vision models. Unlike Vision-Language Models that focus on cross-modal interactions, VPT methods primarily address how to efficiently adapt vision models to new tasks without catastrophic forgetting. This section analyzes recent advances in VPT memory systems using our unified framework, examining how these approaches implement retrieval mechanisms, memory structures, and update schemas.

4.3.1 Overview of Memory Requirements in VPT

Visual Prompt Tuning methods have unique memory requirements stemming from several factors:

- **Continual learning challenges**: VPT methods must learn new tasks sequentially without forgetting previous ones, requiring careful memory organization.
- **Parameter efficiency**: Models must adapt to new tasks while modifying only a tiny fraction of the pre-trained model parameters.
- **Task-specific knowledge representation**: Memory systems must store and retrieve task-specific information effectively.
- **Knowledge transfer**: Effective VPT methods need to leverage knowledge from previous tasks to improve performance on new ones.

Recent advances have introduced novel memory mechanisms to address these challenges, with approaches ranging from prompt pools to dual memory structures and attribute-based learning.

4.3.2 Retrieval Mechanism Analysis

Our analysis of VPT memory systems reveals several key patterns in retrieval mechanisms across different approaches. Table 4.4 provides a comparative overview of retrieval mechanisms employed by various VPT systems, organized by their primary approach to illustrate the evolution of research trends.

4.3.3 Evolutionary Trends in Retrieval Mechanisms

The chronological organization of retrieval approaches reveals a clear evolution in VPT retrieval mechanisms:

1. **Early Focus on Discrete Selection**: Early approaches like L2P (2022) introduced key-value paired prompt pools with instance-wise selection, using a query function to identify the most relevant prompts for each input.
2. **Separation of Knowledge Types**: DualPrompt (2022) advanced the paradigm by creating a two-tier system that explicitly separates task-invariant knowledge (G-Prompt) from task-specific knowledge (E-Prompt), reflecting a deeper understanding of how different types of knowledge support continual learning.
3. **Shift to Compositional Approaches**: CODA-Prompt (2023) moved beyond discrete prompt selection to weighted combinations of prompt components through attention mechanisms, making the selection process fully differentiable and more expressive.

Table 10: Comparison of Retrieval Mechanisms in VPT Memory Systems

System	Year	Primary Approach	Key Mechanisms	Notable Characteristics
Prompt-Based Approaches				
Learning to Prompt (L2P)	2022	Prompt-based	Key-value paired prompt pool, Instance-wise query function	Selects top-N matching prompts dynamically for each input; task identity not required at test time
DualPrompt	2022	Prompt-based	G-Prompt (task-invariant) and E-Prompt (task-specific), Feature similarity matching	E-Prompts matched to test instances based on feature similarity; G-Prompt shared across all tasks
CODA-Prompt	2023	Prompt-based	Weighted component combinations, Attention-based querying	Uses weighted sum of prompt components based on similarity scores; fully differentiable selection process
Similarity-Based Approaches				
STAR-Prompt	2023	Similarity-based + Prompt-based	CLIP's embedding space for matching, Two-level prompting strategy	Uses CLIP for stable similarity computation; class-specific prototypes serve as keys
Hybrid Approaches				
PromptFusion	2023	Hybrid	Split modules (Stabilizer and Booster), Text-visual similarity computation	Stabilizer (CoOp) for stability; Booster (VPT) for plasticity; balance parameter λ between modules
AttriCLIP	2024	Hybrid	Attribute word bank, Key-prompt pairs, Multi-modal matching	Selects attributes based on image-key similarity; focuses on generalizable attributes across categories

4. **Integration with Pre-trained Models**: STAR-Prompt (2023) leveraged CLIP's multi-modal embedding space to create a more stable foundation for similarity computation and prompt selection, addressing the vulnerability of learned selection mechanisms to catastrophic forgetting.

5. **Module Specialization**: PromptFusion (2023) explicitly separated stability and plasticity into distinct modules (CoOp for stability, VPT for plasticity), directly addressing the fundamental trade-off through architectural specialization.

6. **Focus on Cross-Cutting Features**: AttriCLIP (2024) shifted focus entirely to learning generalizable attributes that transcend task boundaries, using an attribute word bank that associates visual attributes with textual descriptions.

This evolution demonstrates a progressive refinement of retrieval mechanisms, from simple selection to sophisticated compositional and specialized approaches, with a growing emphasis on leveraging pre-trained models and focusing on transferable knowledge.

4.3.4 Dominant Patterns in Retrieval Design

Several key patterns emerge across the evolution of VPT retrieval mechanisms:

1. **Increasing Context Sensitivity**: Newer approaches demonstrate enhanced ability to adapt retrieval based on input characteristics, with CODA-Prompt's attention-based mechanism and AttriCLIP's attribute selection providing highly context-sensitive retrieval.

2. **Growing Emphasis on Knowledge Transfer**: More recent methods like STAR-Prompt and AttriCLIP explicitly design retrieval mechanisms to leverage knowledge transfer between tasks, rather than treating tasks in isolation.
3. **Shift Toward Multi-Level Retrieval**: The trend toward multi-level retrieval structures (STAR-Prompt's two-level approach, PromptFusion's dual modules) indicates recognition that effective continual learning requires handling different aspects of knowledge through specialized mechanisms.
4. **Integration with Pre-trained Knowledge**: The increasing use of pre-trained models like CLIP as foundations for retrieval (STAR-Prompt, AttriCLIP) leverages the rich, stable representations these models provide.

4.3.5 Memory Structure Classification

VPT methods employ diverse memory structures to handle task-specific and generalizable knowledge. Table 4.5 provides a comparative overview of memory structures implemented in various VPT systems, organized by their primary structure type to illustrate structural evolution over time.

Table 11: Comparison of Memory Structures in VPT Memory Systems

System	Year	Primary Structure	Key Components	Notable Characteristics
Dynamic Memory Approaches				
Learning to Prompt (L2P)	2022	Dynamic Memory	Prompt pool, Shared memory space	Expandable memory bank; Grows with number of tasks; Knowledge transfer through shared prompts
CODA-Prompt	2023	Dynamic Memory	Decomposable prompt components; Layer-specific prompts	Expandable component bank; Hierarchical organization through layer-specific prompts
Hierarchical Memory Approaches				
DualPrompt	2022	Hierarchical Memory	G-Prompt (static, shared); E-Prompts (dynamic, task-specific)	Two-level hierarchy separating general and specific knowledge; Different prompts attached to different layers
STAR-Prompt	2023	Hierarchical Memory	First-level class-specific prompts; Second-level task-adaptive prompts	CLIP conditions first level for stability; ViT adaptation through second level for plasticity
Hybrid Memory Approaches				
PromptFusion	2023	Hybrid (Static + Dynamic)	Stabilizer module (CoOp); Booster module (VPT)	Complete separation of stability and plasticity into dedicated modules
Distributed Memory Approaches				
AttriCLIP	2024	Distributed + Fixed-Capacity	Attribute word bank (keys + prompts)	Fixed number of attribute pairs; Constant parameter count regardless of tasks; Knowledge shared across categories

4.3.6 Structural Evolution Patterns

The chronological analysis of memory structures reveals several key evolutionary trends:

1. **From Simple to Sophisticated**: Early approaches like L2P (2022) used relatively simple dynamic memory structures (prompt pools), while later methods introduced increasingly sophisticated architectures with explicit separation of different knowledge types.

2. **Increasing Specialization**: The progression shows growing specialization of memory components, from DualPrompt's (2022) separation of general and specific knowledge to PromptFusion's (2023) complete module specialization for stability and plasticity.

3. **From Expanding to Fixed Capacity**: While early methods (L2P, DualPrompt) expanded memory linearly with the number of tasks, later approaches like AttriCLIP (2024) achieved constant parameter counts regardless of task number by focusing on transferable attributes.

4. **Integration with External Models**: More recent methods leverage pre-trained models as part of their memory architecture, with STAR-Prompt (2023) using CLIP for stable class representations and AttriCLIP (2024) leveraging CLIP for attribute learning.

4.3.7 Key Design Principles in Memory Structures

Several important design principles emerge across VPT memory structures:

1. **Knowledge Separation**: The explicit separation of different types of knowledge (general vs. specific, stable vs. plastic) appears consistently beneficial across multiple approaches.

2. **Hierarchical Organization**: Organizing memory in hierarchical layers with different components serving different roles (DualPrompt, STAR-Prompt) enhances both stability and adaptability.

3. **Cross-Task Knowledge Sharing**: More advanced structures focus on facilitating knowledge sharing across tasks through either shared components (CODA-Prompt) or transferable features (AttriCLIP).

4. **Balancing Expansion and Efficiency**: The evolution shows a constant tension between expanding memory for better task-specific representation and maintaining efficiency, culminating in fixed-capacity approaches like AttriCLIP.

4.3.8 Memory Update Schema Analysis

The update mechanisms employed by VPT memory systems reveal sophisticated strategies for balancing stability and plasticity. Table 4.6 provides a comparative overview of update schemas implemented across various VPT approaches, organized by their primary update approach to illustrate the evolution of update strategies.

4.3.9 Update Strategy Evolution

The chronological organization of update schemas reveals important trends in how VPT approaches handle the critical challenge of updating knowledge while avoiding catastrophic forgetting:

1. **From Simple to Multi-Objective**: Early approaches like L2P (2022) used relatively straightforward end-to-end update strategies, while later methods introduced increasingly sophisticated multi-objective learning with carefully designed regularization terms.

2. **Increasing Explicitness in Knowledge Separation**: DualPrompt (2022) introduced explicit separation of task-invariant and task-specific updates, a trend that continued with more sophisticated separations in later approaches.

3. **From Parameter Isolation to Knowledge Transfer**: While early methods focused primarily on isolating task-specific parameters to prevent interference, later approaches like AttriCLIP (2024) explicitly designed update mechanisms to facilitate knowledge transfer between tasks.

Table 12: Comparison of Memory Update Schemas in VPT Memory Systems

System	Year	Primary Approach	Key Mechanisms	Notable Characteristics
End-to-End Approaches				
Learning to Prompt (L2P)	2022	End-to-End	Prompt parameter updates, Matching loss	Natural separation of task knowledge into different prompts; end-to-end training with task data
Knowledge Decoupling Approaches				
DualPrompt	2022	Knowledge Decoupling	Explicit separation of task-invariant and task-specific updates	Multi-layer attachment based on knowledge type; complementary learning between general and specific components
Component-Based Approaches				
CODA-Prompt	2023	Component-Based	Expansion with new components, Orthogonality constraints	Freezes previous components when adding new ones; orthogonality constraint minimizes inter-task correlations
Multi-Objective Approaches				
STAR-Prompt	2023	Multi-Objective	Explicit class separation, Confidence-weighted updates	Direct supervision for class prototype separation; multi-modal generative replay for class distribution modeling
PromptFusion	2023	Module-Specific	Different strategies for stability and plasticity modules	Stabilizer preserves past knowledge; Booster adapts to new tasks; balancing parameter λ between modules
AttriCLIP	2024	Attribute-Focused	Classification, Matching, and Orthogonality losses	Learns generalizable attributes rather than task-specific features; consistent parameter count across tasks

4. **Growing Incorporation of Pre-trained Knowledge**: Recent methods increasingly leverage pre-trained models as stable foundations, with update mechanisms designed to preserve this stable knowledge while adapting to new tasks.

5. **Shift Toward Attribute-Level Learning**: The most recent approach, AttriCLIP (2024), represents a significant shift from updating task or class-specific knowledge to learning transferable attributes that generalize across different tasks and categories.

4.3.10 Catastrophic Forgetting Mitigation Strategies

VPT update schemas employ various strategies to address the fundamental challenge of catastrophic forgetting in continual learning:

1. **Knowledge Isolation**: Early approaches like L2P (2022) use separate prompts for different tasks to prevent direct interference.

2. **Knowledge Decoupling**: Methods like DualPrompt (2022) separate general from specific knowledge to minimize interference while maximizing transfer.

3. **Parameter Freezing**: CODA-Prompt (2023) locks previous components when learning new tasks to preserve existing knowledge.

4. **Stable Foundations**: STAR-Prompt (2023) uses pre-trained models like CLIP to provide stable representations that are less susceptible to forgetting.

5. **Specialized Preservation**: PromptFusion (2023) employs a dedicated stability module specifically designed to preserve past knowledge.

6. **Cross-Cutting Features**: AttriCLIP (2024) focuses on attributes that transcend task boundaries, making knowledge naturally transferable between tasks.

The progression shows an increasing sophistication in how these systems balance stability (preserving existing knowledge) with plasticity (learning new information), moving from simple parameter isolation to complex architectures with specialized components.

4.3.11 Discussion and Key Findings

Our analysis of VPT memory systems through the unified framework reveals several important insights and trends that go beyond surface-level observations, illuminating deeper principles about memory system design and continual learning.

Theoretical Foundations and Epistemological Implications The evolution of VPT memory systems reflects a fundamental shift in how we conceptualize knowledge acquisition and representation:

- **From Localist to Distributed Representations**: Early approaches like L2P employed relatively localist representations (discrete prompts), while later methods like CODA-Prompt and AttriCLIP moved toward distributed representations where knowledge is spread across components or attributes. This mirrors neuroscience findings about how the brain represents concepts through distributed neural patterns rather than individual neurons.

- **Complementary Learning Systems Theory**: DualPrompt's separation of general and specific knowledge directly implements principles from the Complementary Learning Systems theory in cognitive science, which proposes humans learn through dual systems: one for extracting general patterns (neocortex) and another for specific experiences (hippocampus). This architectural principle has proven increasingly valuable for AI systems facing similar challenges in balancing generalization with specificity.

- **Emergence of Meta-Knowledge**: Recent approaches like STAR-Prompt and AttriCLIP demonstrate the emergence of "knowledge about knowledge" – systems that learn not just task information but develop meta-cognitive awareness about what types of knowledge transfer across tasks. This represents a critical step toward more human-like learning capabilities, where strategic decisions about knowledge acquisition and retention become increasingly important.

Fundamental Trade-offs and Their Theoretical Implications Our analysis reveals how different VPT approaches navigate several deep trade-offs with significant theoretical implications:

- **The Stability-Plasticity Continuum**: Rather than a simple dichotomy, we observe that stability and plasticity exist along a continuum, with approaches like PromptFusion explicitly recognizing that optimal positions on this continuum differ by task. This suggests the need for a theoretically grounded understanding of how to determine the optimal balance for a given learning scenario.

- **Parameter Efficiency vs. Representational Capacity**: The trend toward compressed but expressive representations (CODA-Prompt's decomposable components, AttriCLIP's attributes) approaches what information theory would consider optimal coding – maximizing information content while minimizing representation size. This relates to fundamental questions about the minimum description length required to represent knowledge effectively.

- **Knowledge Isolation vs. Transfer**: The progression from strict task isolation (L2P) to explicit transfer mechanisms (AttriCLIP) reflects an increasingly sophisticated understanding of the geometry of knowledge spaces – from treating them as orthogonal to recognizing natural manifolds of related concepts that span traditional task boundaries.

Cognitive and Neuroscience Connections The most successful VPT approaches incorporate principles that align with findings from cognitive science and neuroscience:

- **Schema Theory Implementation**: The attribute-based approach of AttriCLIP parallels schema theory in cognitive psychology, where humans organize knowledge into frameworks (schemas) that help integrate new information into existing knowledge structures. By focusing on attributes rather than categories, AttriCLIP creates flexible schemas that facilitate knowledge integration.

- **Attention as a Cognitive Filter**: CODA-Prompt's attention-based component selection implements cognitive filtration mechanisms similar to those in human attention, where selective focus determines which information enters working memory. This suggests the potential for further integration of cognitive attention models into AI memory systems.

- **Meta-Cognition and Memory Allocation**: PromptFusion's explicit module selection represents a primitive form of meta-cognition, where the system makes decisions about which memory system to employ based on input characteristics. This parallels human meta-cognitive processes that allocate cognitive resources based on task demands.

Future Research Directions Based on our deep analysis, we identify several transformative research directions that go beyond incremental improvements:

1. Theoretical Framework for Memory Capacity Optimization While current approaches empirically determine memory structures, developing a theoretical framework that can predict optimal memory capacity based on task complexity, dataset properties, and transfer objectives would provide invaluable guidance for design decisions. This framework would:

- Establish mathematical relationships between prompt capacity and performance bounds
- Predict memory interference patterns before they occur
- Determine optimal component allocation for compositional memory structures
- Quantify the information-theoretic limits of parameter-efficient tuning

2. Cognitive-Aligned Memory Architectures The convergent evolution toward architectures resembling human memory systems suggests an opportunity for deeper integration of cognitive principles:

- Implementing complementary learning with faster episodic-like and slower semantic-like memory consolidation processes
- Developing memory systems with explicit metacognitive monitoring and control capabilities
- Creating memory architectures with built-in concept formation mechanisms that abstract higher-level patterns from experiences
- Integrating principles of memory reconsolidation, where retrieval makes memories temporarily malleable for updating

3. Cross-Modal Knowledge Abstraction Moving beyond attributes within a single modality, future systems could:

- Develop unified attribute spaces that span multiple modalities (vision, language, audio)
- Learn abstract conceptual structures that exist independently of specific modality representations
- Create modality-agnostic memory systems where knowledge gained in one domain automatically transfers to others
- Implement symbolic reasoning capabilities within neural memory frameworks to achieve true cross-domain abstraction

4. Privacy-Preserving and Ethical Memory Systems As these systems become more deployed, developing memory architectures with built-in privacy guarantees becomes crucial:

- Differential privacy mechanisms for prompt updating that provide formal guarantees about information leakage
- Memory systems that can selectively "forget" sensitive information without compromising overall performance
- Ethical frameworks for determining what knowledge should be preserved versus discarded
- Memory architectures designed for transparency, allowing inspection of what knowledge is stored and how it influences decisions

5. Self-Evolving Memory Architectures The ultimate progression would be memory systems that evolve their own structure:

- Meta-learning approaches that discover optimal memory architectures for specific domains
- Systems that dynamically adjust their position on the stability-plasticity spectrum based on task characteristics
- Memory structures that expand or contract based on information complexity rather than task count
- Neural-symbolic architectures that combine the flexibility of neural approaches with the symbolic reasoning capabilities needed for truly compositional knowledge

The evolution of VPT memory systems demonstrates remarkable progress in addressing the challenges of continual learning in computer vision. From simple prompt pools to sophisticated attribute-based learning systems with constant parameter counts, these approaches have steadily improved in their ability to balance stability and plasticity while maintaining high parameter efficiency. The field continues to move toward more transferable knowledge representations and more specialized architectural components, promising further advances in efficient continuous adaptation of vision models to new tasks and domains.

4.4 Memory Systems for Video Understanding

Recent advancements in memory systems for video understanding demonstrate substantial progress in AI systems' capability to efficiently capture and utilize temporal and spatial information across extensive video sequences. Video-based methods face unique challenges including maintaining temporal coherence, computational efficiency, scalability in memory requirements, and managing long-range dependencies. These developments have significant implications across multiple applications, such as autonomous driving, video surveillance, and real-time robotics. This section systematically analyzes recent approaches using our unified analytical framework, examining retrieval mechanisms, memory structures, and update schemas across leading-edge video memory systems.

4.4.1 Overview of Memory Requirements in Video Understanding

Video understanding imposes distinctive and stringent memory requirements driven by several critical factors:

Temporal coherence requires maintaining accurate temporal relationships and continuity across lengthy video sequences, crucial for tasks like action recognition and event detection. Computational efficiency demands systems operate in real-time or near-real-time, vital for practical applications such as surveillance and robotics. Scalability necessitates efficiently scaling memory with increased video lengths, essential for managing extended recordings. Dynamic and hierarchical representation involves flexible memory structures to manage detailed, immediate information alongside broad, long-term context, essential for coherent narrative understanding and predictive capabilities.

Recent methodologies employ hierarchical memory organization, streaming encoding techniques, and dynamic compression mechanisms to fulfill these stringent requirements.

4.4.2 Retrieval Mechanism Analysis

Retrieval mechanisms in contemporary video memory systems reveal prominent patterns emphasizing their approach to accessing stored information.

Table 13: Comparative Analysis of Retrieval Mechanisms

System	Year	Retrieval Approach	Key Mechanisms	Notable Characteristics
Continuous Video Process (CVP)	2024	Temporal-Spatial	Diffusion-based interpolation	Continuous interpolation between frames, significantly reducing sampling steps
Grounded-VideoLLM	2024	Temporal-Spatial + Prompt-based	Temporal tokens, dual-stream encoding	Precise temporal grounding using discrete timestamps
MeMViT	2022	Temporal-Spatial	Memory cache utilizing keys and values	Efficient retrieval of extended temporal contexts with minimal overhead
VideoStreaming	2024	Hybrid (Temporal + Similarity)	Memory-propagated streaming, adaptive memory selection	Real-time, relevance-based adaptive retrieval
VidCompress	2024	Temporal-Spatial	Dual-compressor (memory-enhanced, text-aware)	Combines short-term detail and long-term context via token compression
MemFlow	2024	Temporal-Spatial	Real-time buffer, attention-driven retrieval	Immediate optical flow estimation and historical motion retrieval
VideoLLaMB	2024	Hybrid (Temporal + Similarity)	Semantic segmentation, recurrent memory tokens	Periodic memory refresh ensuring semantic coherence
XMem	2022	Temporal-Spatial	Multi-tier memory stores, space-time attention	Distinct sensory, working, long-term memories for precise retrieval

Temporal-spatial retrieval remains predominant, critical for ensuring sequential coherence and spatial accuracy, though it can lead to increased computational demands and storage needs for extended sequences. Hybrid retrieval methods have emerged, integrating similarity-based or prompt-based mechanisms to enhance adaptability and query responsiveness, yet they may introduce additional complexity in managing multiple

retrieval criteria simultaneously. Approaches like Grounded-VideoLLM and VideoLLaMB increase semantic granularity, significantly improving retrieval precision through discrete tokens and semantic segmentation, but they can incur higher computational overhead and potential difficulties in accurately segmenting semantic units consistently. Temporal-spatial retrieval remains predominant, critical for ensuring sequential coherence and spatial accuracy. Hybrid retrieval methods have emerged, integrating similarity-based or prompt-based mechanisms to enhance adaptability and query responsiveness. Approaches like Grounded-VideoLLM and VideoLLaMB increase semantic granularity, significantly improving retrieval precision through discrete tokens and semantic segmentation.

However, reliance on semantic granularity may introduce computational overhead and complexity in system architecture, necessitating optimization strategies to maintain efficiency.

4.4.3 Memory Structure Classification

Memory structures adopted by recent video systems demonstrate diverse strategic implementations to manage temporal-spatial complexities effectively.

Table 14: Comparative Analysis of Memory Structures

System	Year	Memory Structure	Components	Notable Characteristics
Continuous Video Process	2024	Dynamic	Latent diffusion states	Continuous internal state updates, ephemeral context representation
Grounded-VideoLLM	2024	Dynamic	Temporal tokens, dual-stream encoders	Temporary tokens, focusing on brief, significant intervals
MeMViT	2022	Dynamic	Memory keys/values, compressed representation	Efficient, scalable coverage for extended durations
VideoStreaming	2024	Hierarchical	Streaming condensed memories, summary tokens	Adaptively hierarchical memory summaries
VidCompress	2024	Dynamic	Memory-enhanced dual-compressor structure	Dynamic token compression across varying temporal scales
MemFlow	2024	Dynamic	Real-time online buffer	Continuous updates for real-time processing
VideoLLaMB	2024	Hierarchical	Semantic segments, recurrent bridging tokens	Hierarchical segmentation enabling efficient context bridging
XMem	2022	Hierarchical	Sensory, working, and long-term memory stores	Cognitively inspired hierarchical memory, effectively preventing memory overload

The transition from dynamic to advanced hierarchical structures significantly enhances memory efficiency by effectively compartmentalizing short-term detailed information and long-term abstract representations. Hierarchical structures reduce computational overhead by selectively consolidating critical information, improving retrieval precision, and facilitating scalability in managing extended temporal contexts. Transition from dynamic to advanced hierarchical structures markedly improves long-term memory efficiency and contextual retention. Increased adoption of semantic segmentation contributes significantly to memory organization, enhancing the semantic consistency of retrieved information, though potentially increasing computational complexity.

4.4.4 Memory Update Schema Analysis

Video memory systems exhibit complex strategies for memory updates, balancing stability with dynamic adaptation.

Table 15: Comparative Analysis of Memory Update Schemas

System	Year	Update Schema	Key Mechanisms	Notable Characteristics
Continuous Video Process	2024	Recency-based	Stepwise diffusion updates	Immediate prior state significantly influences current updates
Grounded-VideoLLM	2024	Importance-weighted	Weighted updates via temporal tokens	Prioritizes updates based on relevance to current query context
MeMViT	2022	Recency-based	Periodic compression based on recent frames	Balances efficiency with detail retention
VideoStreaming	2024	Importance-weighted	Adaptive query-driven selection	Dynamically updates memory relevance for real-time queries
VidCompress	2024	Recency-based	Continuous incremental compression	Incremental updating efficiently preserves short and long-term detail
MemFlow	2024	Recency-based	Ongoing incremental motion updates	Real-time updates ensuring up-to-date context information
VideoLLaMB	2024	Importance-weighted	Periodic semantic relevance updates	Ensures critical semantic continuity and relevance in memory updates
XMem	2022	Importance-weighted	Selective long-term consolidation and potentiation	Strategic memory consolidation to preserve critical prototypes

Recency-based schemas predominate due to their computational simplicity and effectiveness in capturing recent, relevant context. Importance-weighted schemas, however, increasingly provide nuanced and selective updates, emphasizing semantically critical or query-specific information, demonstrating superior contextual and query-response performance.

4.4.5 Discussion and Key Findings

Comprehensive analysis highlights essential insights into the design and optimization of video memory systems. Hierarchical memory architectures and semantic segmentation substantially enhance long-term contextual coherence and mitigate memory explosion risks. Temporal-spatial retrieval strategies remain foundational, though increasingly sophisticated hybrid methods leveraging semantic granularity and similarity retrieval offer significant advances in precision and adaptability.

A critical balance between recency and importance weighting is crucial, with systems demonstrating increasingly sophisticated selective update strategies. Future research should address theoretical frameworks for hierarchical memory optimization, refine advanced semantic retrieval mechanisms, and develop dynamically adaptive, query-sensitive memory structures. Anticipated challenges include managing computational complexity, optimizing memory structures for real-time responsiveness, and integrating scalable hierarchical models with existing infrastructure.

5 Cross-Domain Analysis

Having analyzed memory systems across LLMs, VLMs, VPT, and Video Understanding domains individually, this section provides a cross-domain analysis that identifies universal patterns, unique adaptations, and evolutionary trends through our unified framework. This comparative analysis reveals important insights about how different AI domains have addressed similar memory challenges with domain-specific solutions.

5.1 Retrieval Mechanism Cross-Domain Patterns

Our cross-domain analysis reveals several significant patterns in retrieval mechanisms across different AI domains.

5.1.1 Evolution Toward Hybrid Approaches

A clear trend across all domains is the evolution from simple, single-strategy retrieval mechanisms toward sophisticated hybrid approaches that combine multiple retrieval methods:

- **LLMs**: Early systems like MemGPT (2023) primarily employed similarity-based retrieval, while later systems like RET-LLM (2023) and Think-in-Mem (2024) implemented sophisticated hybrids combining similarity-based, prompt-based, and temporal approaches.

- **VLMs**: The progression from single-modality approaches to cross-modal interaction mechanisms is evident, with systems like MITP (2022) introducing memory hubs for cross-modal attention and later approaches like MemVP (2023) directly injecting visual information into language model memory spaces.

- **VPT**: The evolution from discrete selection in L2P (2022) to compositional approaches in CODA-Prompt (2023) and finally to attribute-based retrieval in AttriCLIP (2024) demonstrates increasing sophistication in retrieval strategies.

- **Video Understanding**: Systems evolved from basic temporal-spatial retrieval to incorporate semantic segmentation and adaptive selection, as seen in the progression from MeMViT (2022) to VideoLLaMB (2024).

This convergent evolution toward hybrid approaches suggests that combining multiple retrieval strategies is fundamentally more effective than any single approach, regardless of domain.

5.1.2 Increasing Autonomy in Retrieval Decisions

Another cross-cutting pattern is the progression toward systems that autonomously determine when and what to retrieve:

- **LLMs**: RET-LLM's (2023) reflective trigger mechanism enables the model to determine when memory retrieval is necessary, while Think-in-Mem (2024) implements deliberate thinking about what to retrieve.

- **VLMs**: Conditional Prompt Tuning (2023) uses a learned router to dynamically determine routing scores for weighting prompt experts.

- **VPT**: The MoPE mechanism in Conditional Prompt Tuning creates context-aware information retrieval by dynamically selecting prompt experts based on input.

- **Video Understanding**: VideoStreaming (2024) implements adaptive memory selection based on relevance to the current context.

This pattern reflects a broader shift toward more metacognitive systems that can strategically manage their own memory resources.

5.1.3 Unified Retrieval Abstraction

To better understand cross-domain retrieval mechanisms, we propose a formal abstraction:

Retrieval Mechanism (R) = f(Query Representation, Memory Index, Similarity Function, Contextual Factors)

Where:

- **Query Representation** varies by domain (text embeddings in LLMs, visual features in VLMs, etc.)

- **Memory Index** ranges from simple vector stores to hierarchical structures

- **Similarity Function** includes cosine similarity, learned attention, and dynamic routing

- **Contextual Factors** represent domain-specific considerations that influence retrieval

This abstraction helps explain why similar principles manifest differently across domains based on their specific constraints and requirements.

5.1.4 Comparative Case Study: Context-Sensitive Retrieval

To illustrate cross-domain retrieval patterns, we examine how context-sensitive retrieval is implemented across domains:

- **LLMs (Think-in-Mem, 2024)**: Implements "deliberate thinking" where the model explicitly reasons about what information it needs before retrieval, creating a two-stage process that enhances precision.

- **VLMs (DMN, 2024)**: Generates sample-adaptive classifiers for each test point by dynamically weighting cached features from both static and dynamic memories, tailoring retrieval to each specific input.

- **VPT (PromptFusion, 2023)**: Dynamically balances stability and plasticity modules using a learned parameter that adjusts based on input characteristics, creating context-sensitive retrieval.

- **Video Understanding (XMem, 2022)**: Employs space-time attention mechanisms across multi-tier memory stores, allowing the system to focus on relevant temporal-spatial information based on query context.

Despite differences in implementation, all these approaches demonstrate a common principle: effective retrieval requires adapting to the specific context of the current query rather than using fixed patterns.

5.1.5 Domain-Specific Retrieval Optimizations

Despite these universal patterns, each domain has developed distinctive retrieval optimizations to address domain-specific challenges:

- **LLMs**: Focus on narrative coherence and conversational context maintenance through mechanisms like episodic buffers (MemBank, 2023) and experience tagging (RET-LLM, 2023).

- **VLMs**: Emphasis on cross-modal alignment through mechanisms like memory hubs (MITP, 2022) and direct memory space integration (MemVP, 2023).

- **VPT**: Prioritization of transferable knowledge through techniques like attribute word banks (AttriCLIP, 2024) and class prototypes (STAR-Prompt, 2023).

- **Video Understanding**: Development of specialized temporal grounding through timestamp mechanisms (Grounded-VideoLLM, 2024) and semantic segmentation (VideoLLaMB, 2024).

5.1.6 Retrieval Mechanism Trade-offs Visualization

Figure 5.1 illustrates the trade-offs between retrieval precision, computational efficiency, and context sensitivity across domains. The plot reveals that while all domains have moved toward the upper-right quadrant (high precision and sensitivity), video understanding systems typically incur higher computational costs due to temporal processing requirements. LLMs show the highest average precision, while VLMs demonstrate the strongest balance across all three dimensions.

5.2 Memory Structure Cross-Domain Patterns

Our analysis reveals important patterns in how memory structures have evolved across domains to address common challenges.

5.2.1 Convergence Toward Hierarchical and Hybrid Structures

Across all domains, there has been a consistent movement from simple, monolithic memory structures toward hierarchical and hybrid architectures:

- **LLMs**: Evolution from context windows to sophisticated multi-tier systems like MemBank's (2023) three-tier architecture and Memory LLM-agent's (2023) core/archival/knowledge organization.

- **VLMs**: Progression from straightforward prompt banks to hierarchical structures like MITP's (2022) layer-wise organization and DMN's (2024) dual memory architecture.

- **VPT**: Development from L2P's (2022) simple prompt pool to DualPrompt's (2022) general/specific separation and eventually to sophisticated structures like STAR-Prompt's (2023) two-level prompting hierarchy.

- **Video Understanding**: Advancement from simple memory buffers to multi-tier organizations like XMem's (2022) sensory/working/long-term memory stores.

This convergent evolution toward hierarchical and hybrid structures suggests fundamental advantages to organizing information at multiple levels of abstraction and combining different structure types.

5.2.2 Memory Structure Taxonomy Across Domains

We propose a cross-domain taxonomy of memory structures based on their functional properties rather than domain-specific implementations:

1. **Temporary Storage Structures**:

- LLMs: Context window (MemGPT)

- VLMs: Temporal prompts (MITP)

- VPT: Dynamic prompts (DualPrompt)

- Video: Real-time buffer (MemFlow)

2. **Working Memory Structures**:

- LLMs: Reasoning cache (Think-in-Mem)

- VLMs: Memory hub (MITP)

- VPT: Mixture of Prompt Experts (Conditional Prompt Tuning)

- Video: Working memory (XMem)

3. **Long-term Memory Structures**:

- LLMs: Knowledge store (MemBank)

- VLMs: Static memory (DMN)

- VPT: G-Prompt (DualPrompt)

- Video: Long-term memory (XMem)

4. **Integrative Memory Structures**:

- LLMs: Episodic buffer (MemBank)

- VLMs: Cross-modal memory hub (MITP)

- VPT: Stabilizer module (PromptFusion)

- Video: Memory consolidation mechanism (VideoLLaMB)

This taxonomy highlights functional similarities across domains despite different implementations, suggesting universal principles in memory organization.

5.2.3 Specialization of Memory Components

Another universal pattern is the increasing specialization of memory components for different types of information or cognitive functions:

- **LLMs**: Memory LLM-agent (2023) uses distinct stores for core, archival, and knowledge information, while ChatDB (2024) implements a structured database approach for explicit relational knowledge.

- **VLMs**: Conditional Prompt Tuning (2023) disentangles prompt vectors into static, dynamic, and mapped types for different functions.

- **VPT**: PromptFusion (2023) completely separates stability and plasticity functions into dedicated modules.

- **Video Understanding**: XMem (2022) implements distinct sensory, working, and long-term memory stores with specialized functions.

This specialization pattern mirrors the functional separation observed in human memory systems, suggesting convergent evolution toward cognitively-aligned architectures.

5.2.4 Comparative Case Study: Stability-Plasticity Balance

The challenge of balancing stability (preserving existing knowledge) with plasticity (adapting to new information) appears universally across domains:

- **LLMs (Memory LLM-agent, 2023)**: Implements a three-tier memory system where core memory maintains immediate context (high plasticity), archival memory preserves important past interactions (balanced), and knowledge tools provide stable information (high stability).

- **VLMs (DMN, 2024)**: Uses dual memory with dynamic memory for test samples (high plasticity) and static memory for training data (high stability), with attention mechanisms balancing their influence.

- **VPT (PromptFusion, 2023)**: Explicitly separates stability and plasticity into dedicated modules (CoOp for stability, VPT for plasticity) with a learnable parameter controlling their balance.

- **Video Understanding (VidCompress, 2024)**: Employs a dual-compressor architecture that balances short-term detail preservation with long-term context maintenance through separate mechanisms.

This case study demonstrates how different domains have converged on similar architectural solutions to the fundamental stability-plasticity dilemma, despite their distinct applications.

5.2.5 Biological Inspiration

A notable trend across domains is the increasing influence of cognitive science and neuroscience on memory structure design:

- **LLMs**: MemBank (2023) explicitly models components inspired by episodic, semantic, and working memory.

- **VLMs**: The Dual-Memory Model (2021) explicitly implements Baddeley's psychological theory of human memory.

- **VPT**: DualPrompt's (2022) separation of general and specific knowledge directly implements principles from Complementary Learning Systems theory.

- **Video Understanding**: XMem's (2022) multi-tier architecture directly mirrors human memory organization.

This trend toward biologically-inspired designs reflects growing recognition that human memory systems offer valuable architectural principles for AI memory.

5.2.6 Memory Structure Trade-offs Visualization

Figure 5.2 maps memory structures across domains according to their adaptability, stability, and parameter efficiency. The visualization shows that while early systems clustered toward either high stability or high adaptability, newer systems across all domains have converged toward the center-right region that balances these properties while maintaining strong parameter efficiency. VPT systems consistently demonstrate the highest parameter efficiency, while LLMs show the greatest range of adaptability-stability trade-offs.

5.3 Update Schema Cross-Domain Patterns

Our analysis identifies significant commonalities and differences in how memory update approaches have evolved across domains.

5.3.1 Evolution Toward Multi-Factor Integration

All domains show a progression from simple update criteria toward sophisticated approaches that balance multiple factors:

- **LLMs**: Early systems primarily employed straightforward criteria like recency (MemGPT, 2023), while later approaches introduced mechanisms that balance recency, importance, and frequency (MyAgent, 2024).

- **VLMs**: While importance-weighted approaches have remained prevalent, their implementation has grown more sophisticated, with recent models like SynapticRAG (2023) implementing complex integration of multiple update factors.

- **VPT**: Movement from simple end-to-end updates (L2P, 2022) to multi-objective approaches (STAR-Prompt, 2023) that balance stability and plasticity.

- **Video Understanding**: Progression from simple recency-based updates to importance-weighted schemas that prioritize semantic relevance (VideoLLaMB, 2024).

This convergent evolution suggests that effective memory management requires balancing multiple update criteria rather than relying on any single approach.

5.3.2 Update Schema Taxonomy Across Domains

We propose a functional taxonomy of update mechanisms that transcends domain boundaries:

1. **Temporal Management Mechanisms**:

- LLMs: Temporal decay functions (SCM)

- VLMs: Layer-wise temporal prompts (MITP)

- VPT: Experience replay (STAR-Prompt)

- Video: Time constant updates (SynapticRAG)

2. **Information Consolidation Mechanisms**:

- LLMs: Utility assessment (Think-in-Mem)

- VLMs: Attention-based weighting (DMN)

- VPT: Orthogonality constraints (CODA-Prompt)

- Video: Selective long-term consolidation (XMem)

3. **Novelty Detection Mechanisms**:

- LLMs: Salience detection (MemBank)

- VLMs: Mutual information estimation (GPT)

- VPT: Class-wise augmentation (GPT)

- Video: Adaptive relevance updates (VideoLLaMB)

This taxonomy highlights functional similarities in update mechanisms despite different implementations across domains.

5.3.3 Comparative Case Study: Catastrophic Forgetting Mitigation

The challenge of catastrophic forgetting (losing existing knowledge when learning new information) appears universally across domains:

- **LLMs (RET-LLM, 2023)**: Addresses forgetting through reflection-based consolidation, where the model explicitly considers which experiences should be preserved and strengthened based on their long-term utility.

- **VLMs (SynapticRAG, 2023)**: Mitigates forgetting through binding scores that combine temporal and semantic similarity, preserving important connections while allowing gradual adaptation.

- **VPT (DualPrompt, 2022)**: Prevents forgetting by separating task-invariant knowledge (G-Prompt) from task-specific knowledge (E-Prompt), allowing new tasks to be learned without interfering with general knowledge.

- **Video Understanding (XMem, 2022)**: Combats forgetting by maintaining prototype representations in long-term memory with selective consolidation and potentiation mechanisms that preserve critical visual patterns.

Despite domain differences, these approaches share common principles: separating stable knowledge from adaptive components, selective consolidation of important information, and metacognitive assessment of knowledge importance.

5.3.4 Increasing Metacognitive Control

A significant pattern across domains is the shift toward update mechanisms with explicit metacognitive components:

- **LLMs**: RET-LLM (2023) implements reflection-based consolidation where the model explicitly considers which experiences should be preserved.

- **VLMs**: DMN (2024) uses attention-based mechanisms to actively determine the importance of information from different memory sources.

- **VPT**: Think-in-Mem (2024) evaluates the utility of memories based on their contribution to successful reasoning.

- **Video Understanding**: XMem (2022) implements selective long-term consolidation based on strategic importance assessment.

This pattern reveals a broader trend toward systems with greater agency over their own memory management processes.

5.3.5 Balancing Stability and Plasticity

A fundamental challenge addressed across all domains is balancing stability (preserving existing knowledge) with plasticity (learning new information):

- **LLMs**: Memory LLM-agent (2023) explicitly models memory management as agent actions that balance preservation and updating.

- **VLMs**: Conditional Prompt Tuning (2023) implements regularization to prevent dominant experts while enabling adaptation.

- **VPT**: PromptFusion (2023) employs dedicated modules for stability and plasticity with a learnable balance parameter.

- **Video Understanding**: VidCompress (2024) balances short-term detail preservation with long-term context maintenance through its dual-compressor architecture.

The universal nature of this challenge and the comparable approaches to addressing it suggest a fundamental principle of memory system design that transcends specific domains.

5.3.6 Update Schema Trade-offs Visualization

Figure 5.3 visualizes update schemas across domains according to their stability, plasticity, and computational efficiency. The plot reveals a clear evolutionary path in all domains from the bottom corners (high stability/low plasticity or low stability/high plasticity) toward the upper center (balanced stability and plasticity with improved efficiency). VPT systems show the most dramatic improvements in computational efficiency over time, while LLMs demonstrate the most balanced approaches to the stability-plasticity trade-off.

5.4 Performance and Efficiency Trends

Beyond the architectural patterns, our cross-domain analysis reveals important trends in performance and efficiency metrics across systems.

5.4.1 Parameter Efficiency Focus

A universal trend across all domains is the progressive improvement in parameter efficiency:

- **LLMs**: Memory LLM-agent (2023) and Think-in-Mem (2024) achieve strong performance with minimal parameter overhead through efficient memory indexing and reasoning caches.

- **VLMs**: MITP (2022) requires only 2.0M trainable parameters (1% of foundation model), while Conditional Prompt Tuning (2023) achieves state-of-the-art performance with only 0.7% of trainable parameters.

- **VPT**: AttriCLIP (2024) maintains constant parameter counts regardless of task number by focusing on transferable attributes.

- **Video Understanding**: MeMViT (2022) and VidCompress (2024) achieve strong performance with compressed representations and minimal parameter overhead.

This cross-cutting focus on parameter efficiency reflects broader industry trends toward more economical AI systems.

5.4.2 Efficiency-Performance Relationship

Our analysis reveals a consistent relationship between memory efficiency and model performance across domains that can be approximated as:

$$\textbf{Performance (P)} = \textbf{log(Memory Capacity)} + \textbf{(Retrieval Efficiency)} + \textbf{(Update Sophistication)}$$

Where , , and are domain-specific constants. This relationship suggests that while memory capacity is important, the efficiency of retrieval mechanisms and the sophistication of update schemas can compensate for limited capacity, explaining why smaller, more efficient models can sometimes outperform larger ones.

5.4.3 Privacy Considerations in Memory Systems

While still emerging, considerations of privacy in memory systems are gaining prominence across domains:

- **LLMs**: ChatDB (2024) implements structured approaches that could enable better privacy controls, though explicit privacy mechanisms remain limited.

- **VLMs**: Most current approaches lack explicit privacy preservation mechanisms.

- **VPT**: No explicit privacy-preserving update mechanisms observed in current systems.

- **Video Understanding**: Some initial considerations in VideoStreaming (2024) regarding what information to retain, but limited formal privacy guarantees.

This analysis reveals a significant gap in current memory systems across domains, suggesting an important direction for future research.

5.4.4 Performance Efficiency Visualization

Figure 5.4 charts the evolution of performance (y-axis) and parameter efficiency (x-axis) across domains from 2021 to 2024. The visualization demonstrates a clear trend toward the upper-right quadrant (high performance with high parameter efficiency) across all domains. VPT and VLM systems show the steepest improvement curves in parameter efficiency, while LLMs demonstrate the most consistent performance gains.

This cross-domain analysis reveals both universal patterns in memory system design that transcend specific AI domains and unique adaptations that address domain-specific challenges. The convergent evolution toward hybrid retrieval mechanisms, hierarchical memory structures, and multi-factor update schemas suggests fundamental principles of effective memory management. At the same time, the domain-specific optimizations highlight how similar architectural patterns can be adapted to address unique challenges in different AI fields.

6 Future Directions and Cross-Domain Transfer Opportunities

Based on our comprehensive cross-domain analysis of memory systems, this section explores promising research directions and knowledge transfer opportunities that could advance memory architectures across AI domains. We identify fundamental principles that transcend domain boundaries and propose innovative research paths that leverage cross-domain insights.

6.1 Cross-Domain Knowledge Transfer Opportunities

Our analysis reveals several fertile areas where knowledge transfer across domains could yield significant advances:

6.1.1 From VLMs to LLMs: Multimodal Memory Integration

VLMs have developed sophisticated mechanisms for cross-modal alignment and integration that could benefit LLM memory systems:

- **Memory-Space Integration Implementation Proposal**: We propose extending MemVP's (2023) approach by developing a "memory projection layer" that maps arbitrary modalities (not just visual) into LLM memory space through learned projection functions. This architecture would consist of:

- Modality-specific encoders that generate normalized representations

- A shared projection layer mapping these representations to the FFN memory space of LLMs

- A modality-aware attention mechanism to weight the importance of different modalities

Implementation could begin with structured data types like tables and graphs, which have well-defined representations, before progressing to more complex modalities.

- **Cross-Modal Attention Mechanism Transfer**: MITP's (2022) memory hub for bidirectional information flow between modalities could be adapted to create a "memory router" for LLMs that enables more efficient integration of different knowledge types (factual, procedural, episodic). This would allow LLMs to maintain separate memory stores for different types of knowledge while providing a unified interface for retrieval.

- **Multi-Granularity Retrieval Adaptation**: DMN's (2024) ability to adaptively adjust retrieval granularity based on different knowledge sources could be implemented in LLMs through a "granularity controller" that dynamically determines whether to retrieve at the document, paragraph, or sentence level based on query characteristics and task requirements.

6.1.2 From VPT to Video Understanding: Parameter-Efficient Adaptation

VPT methods have achieved remarkable parameter efficiency while maintaining performance, offering valuable lessons for video understanding systems that often struggle with computational demands:

- **Temporal Attribute Bank Implementation**: We propose developing a "temporal attribute bank" inspired by AttriCLIP (2024) that maintains a fixed set of transferable temporal attributes for video understanding. This system would:

- Identify and maintain a gallery of fundamental temporal patterns (e.g., acceleration, periodicity, transitions)

- Learn to compose these patterns to represent complex video sequences

- Maintain constant parameter counts regardless of video length by operating at the pattern level rather than the frame level

This approach could be validated on video datasets of varying lengths to demonstrate parameter scaling independence.

- **Module Specialization Architecture**: Adapting PromptFusion's (2023) separation of stability and plasticity modules could lead to a "dual-stream video processor" with:

- A stability stream that captures persistent scene elements and background context

- A plasticity stream that focuses on dynamic elements and temporal changes

- A learnable balancing mechanism that adjusts the importance of each stream based on video content

This architecture would be particularly valuable for long-form video understanding where maintaining scene consistency while tracking changes is crucial.

- **Hierarchical Prompting for Temporal Data**: STAR-Prompt's (2023) two-level prompting strategy could be adapted into a "temporal prompt hierarchy" where:

- First-level prompts capture stable scene elements and general motion patterns

- Second-level prompts adapt to specific temporal dynamics and event transitions

- Different prompts attach to different temporal scales in the video processing pipeline

This approach would enable efficient processing of videos with varying temporal dynamics.

6.1.3 From LLMs to VPT: Metacognitive Memory Management

Advanced LLM memory systems have developed sophisticated metacognitive mechanisms that could enhance VPT approaches:

- **Reflection-Based Consolidation Implementation**: Adapting RET-LLM's (2023) reflective triggers could create a "prompt reflection module" for VPT that:

- Explicitly evaluates prompt effectiveness after each usage

- Determines when specific prompts need refinement based on performance metrics

- Makes strategic decisions about when to access different prompt components

This system could significantly improve prompt utilization efficiency by avoiding unnecessary retrievals and focusing computational resources on the most relevant prompts.

- **Deliberate Thinking Integration**: Think-in-Mem's (2024) explicit reasoning about retrieval could be implemented as a "prompt reasoning controller" that:

- Analyzes input characteristics to determine optimal prompt composition before retrieval

- Generates explicit reasoning paths that explain prompt selection decisions

- Continuously refines its selection strategies based on performance feedback

This approach would make prompt selection more interpretable and context-aware.

- **Experience Tagging for Prompts**: MyAgent's (2024) correlation tracking between memories and outcomes could be implemented as a "prompt effectiveness tracker" that:

- Associates prompts with task performance metrics

- Strengthens effective prompts and weakens those associated with poor outcomes

- Builds a causal model linking prompt characteristics to performance across different tasks

This mechanism would enable more strategic prompt evolution in continual learning settings.

6.1.4 From Video Understanding to VLMs: Temporal Coherence

Video understanding systems have developed specialized approaches to maintain temporal coherence that could benefit VLMs:

- **Semantic Segmentation for Visual Sequences**: VideoLLaMB's (2024) semantic segmentation approach could be adapted to create a "visual narrative segmenter" for VLMs that:

- Identifies meaningful segments in image sequences based on semantic coherence

- Maintains hierarchical representations that preserve both segment-level and sequence-level information

- Enables more efficient processing of visual stories by operating at the segment level rather than the individual image level

This would address current limitations in processing visual narratives or sequences in VLMs.

- **Multi-Tier Memory Organization Implementation**: XMem's (2022) sensory/working/long-term memory organization could be adapted as a "visual memory hierarchy" for VLMs with:

- A sensory memory that briefly stores detailed visual information

- A working memory that processes currently relevant visual context

- A long-term memory that preserves important visual concepts and relationships

This architecture would improve visual reasoning capabilities in VLMs by mirroring human visual memory processes.

- **Adaptive Condensation for Visual Information**: VideoStreaming's (2024) adaptive memory selection could inspire a "visual information condenser" for VLMs that:

- Dynamically adjusts the detail level of stored visual information based on its importance

- Preserves high-fidelity representations of important visual elements while compressing less relevant details

- Adaptively refines visual memory based on query requirements

This approach would help VLMs manage memory more efficiently when processing large volumes of visual information.

6.2 Emerging Research Directions

Our cross-domain analysis points to several promising research directions that could significantly advance memory systems across domains:

6.2.1 Theoretically-Grounded Memory Capacity Optimization

While current approaches determine memory structures empirically, developing a theoretical framework for optimal memory capacity could provide invaluable guidance:

- **Testable Hypothesis**: "The optimal memory capacity for a given domain follows a power law relationship with model size and task complexity, expressible as C = MT, where M is model size, T is task complexity, and , , are domain-specific constants."

- **Validation Framework**: This hypothesis could be tested through systematic experiments that:

- Vary model size while keeping task complexity constant

- Vary task complexity while keeping model size constant

- Measure performance across multiple domains to identify domain-specific constants

- **Information-Theoretic Bounds Research Agenda**: We propose a research program to establish mathematical relationships between memory capacity and performance bounds based on:

- Mutual information between memory contents and task requirements

- Entropy of input distributions across different domains

- Minimum description length principles for memory representations

- **Interference Prediction Model**: Developing predictive models of memory interference could prevent catastrophic forgetting through:

- Theoretical analysis of representation overlap in different memory structures

- Simulation-based assessment of interference risks before deployment

- Automated memory reorganization based on predicted interference patterns

This theoretical framework would transform memory design from an empirical process to a principled approach with predictable outcomes.

6.2.2 Cognitive-Aligned Memory Architectures

The convergent evolution toward biologically-inspired designs suggests deeper integration of cognitive principles:

- **Tri-Level Memory System Architecture**: We propose a comprehensive architecture with explicit separation of:

- Episodic memory (instance-specific experiences with high detail)

- Semantic memory (concept-level knowledge with abstracted representations)

- Procedural memory (task-specific patterns with action-oriented representations)

Each component would have distinct update rates, retrieval mechanisms, and integration functions, mirroring human memory organization.

- **Implementation Strategy**: This architecture could be implemented through:

- A transformer-based episodic store with high-dimensional contextual embeddings

- A graph-structured semantic network with concept nodes and relation edges

- A sequence-based procedural store optimized for action prediction

- **Evaluation Protocol**: We propose a specialized benchmark suite designed to test human-like memory properties, including:

- The spacing effect (better retention with spaced vs. massed repetition)

- Context-dependent recall (retrieval performance in matched vs. mismatched contexts)

- Schema-consistent learning (faster acquisition of schema-consistent information)

- **Memory Reconsolidation Mechanisms**: Implementing principles of memory reconsolidation, where retrieval makes memories temporarily malleable for updating, could enable more efficient continual learning through:

- Selective destabilization of relevant memory components during retrieval

- Controlled integration of new information into existing knowledge structures

- Stabilization processes that preserve updated memories

These cognitive-aligned architectures would bridge the gap between AI systems and human memory capabilities.

6.2.3 Cross-Modal Knowledge Abstraction

Moving beyond current approaches to multimodal integration:

- **Unified Attribute Space Implementation**: We propose developing a shared representational space for attributes across modalities through:

- Alignment of conceptual representations across text, vision, and other modalities

- Modality-invariant encoders that extract consistent attributes regardless of input type

- Cross-modal distillation techniques that transfer attribute knowledge between modalities

- **Modality-Agnostic Concept Formation**: Creating abstraction mechanisms that form conceptual structures independent of specific modality representations through:

- Unsupervised discovery of cross-modal patterns in large-scale multimodal datasets

- Representation learning techniques that isolate conceptual content from modality-specific features

- Evaluation metrics that assess concept transfer across modalities

- **Neuro-Symbolic Integration for Memory**: Implementing symbolic reasoning capabilities within neural memory frameworks through:

- Hybrid architectures that combine vector representations with symbolic structures

- Differentiable reasoning modules that operate over structured memory contents

- Mechanisms for bi-directional translation between neural and symbolic representations

These advances would enable true cross-domain abstraction and reasoning, dramatically improving knowledge transfer across modalities.

6.2.4 Privacy-Preserving Memory Systems

As memory systems become more sophisticated and deployed in sensitive contexts, privacy considerations become crucial:

- **Technical Approach: Differential Privacy for Memory Updates**: We propose developing memory update mechanisms with formal privacy guarantees through:

- Noise calibration techniques that protect individual data points while preserving aggregate patterns

- Privacy budget management across multiple memory update operations

- Theoretical bounds on information leakage from memory retrieval operations

- **Selective Forgetting Implementation**: Creating memory systems that can selectively "forget" sensitive information without compromising overall performance through:

- Fine-grained removal of specific information from neural representations

- Model editing techniques that preserve overall structure while removing targeted memories

- Verification methods that confirm successful forgetting

- **Evaluation Framework**: We propose specific metrics to measure the privacy-utility tradeoff:

- Reconstruction resistance (difficulty of recovering original data from memory)

- Membership inference resistance (difficulty of determining if specific data was used)

- Performance retention after forgetting (maintenance of capabilities despite information removal)

- **Regulatory Alignment**: These approaches would align with emerging privacy regulations by:

- Providing technical mechanisms to implement "right to be forgotten" requirements

- Enabling data minimization principles through selective memory storage

- Supporting transparency through explainable memory operations

These privacy-preserving mechanisms will be essential for deploying memory systems in domains with sensitive information.

6.2.5 Self-Evolving Memory Architectures

The ultimate progression would be memory systems that adapt their own structure:

- **Meta-Learning for Architecture Discovery**: Developing approaches that discover optimal memory architectures through:

- Neural architecture search techniques specialized for memory components

- Meta-learning algorithms that optimize memory structures across diverse tasks

- Evolutionary algorithms that explore the space of possible memory configurations

- **Dynamic Memory Allocation Implementation**: Creating systems that automatically adjust memory capacity based on information complexity through:

- Information-theoretic metrics that assess required memory capacity for current data

- Allocation algorithms that expand or contract memory resources based on task demands

- Efficiency-oriented pruning techniques that maintain performance with minimal resources

- **Adaptive Stability-Plasticity Mechanisms**: Implementing systems that dynamically adjust their position on the stability-plasticity spectrum through:

- Detection algorithms for concept drift and distribution shifts

- Dynamic regulation of learning rates based on novelty assessment

- Meta-cognitive monitoring of forgetting patterns to trigger adaptive adjustments

These self-evolving architectures would represent a significant advance toward truly adaptive AI systems.

6.3 Technical Challenges and Potential Solutions

Several technical challenges must be addressed to realize these future directions:

6.3.1 Computational Efficiency at Scale

Memory operations become increasingly expensive as memory stores grow:

- **Research Agenda**: We propose a three-stage research program focusing on:

1. **Hierarchical Indexing**: Developing multi-level indexing structures that maintain retrieval efficiency with massive memory stores

2. **Adaptive Compression**: Creating content-aware compression techniques that preserve important information while reducing storage requirements

3. **Hardware-Specific Optimizations**: Designing memory architectures tailored to specific hardware accelerators

- **Benchmark Suite**: This agenda requires a comprehensive benchmark that evaluates:

- Retrieval latency at different memory scales

- Compression ratio vs. information preservation

- Energy efficiency of memory operations

- **Success Metrics**: Specific targets include:

- 10x reduction in retrieval computation without accuracy degradation

- Linear scaling of memory efficiency with exponential increases in information volume

- 100x improvement in energy efficiency for memory operations

6.3.2 Catastrophic Forgetting Prevention

Continual learning remains a fundamental challenge across domains:

- **Memory Isolation and Transfer Implementation**: Building on DualPrompt's (2022) approach, we propose a "memory compartmentalization" architecture that:

- Creates dedicated memory regions for domain-specific knowledge

- Implements controlled pathways for knowledge transfer between regions

- Utilizes gating mechanisms to prevent interference during learning

- **Experience Replay with Generative Models**: Using foundation models to implement privacy-preserving experience replay through:

- Generative models that produce synthetic examples preserving statistical properties of original data

- Representation-level replay that operates on embeddings rather than raw data

- Prioritized replay scheduling based on estimated forgetting risk

- **Neuromodulation-Inspired Approaches**: Implementing mechanisms inspired by biological neuromodulators through:

- Adaptive learning rate regulation based on novelty assessment

- Attention-gated memory updates controlled by importance signals

- Context-sensitive plasticity that varies based on task requirements

7 Conclusion

This comprehensive survey of memory systems across LLMs, VLMs, VPT, and Video Understanding domains reveals both universal patterns in memory system design and domain-specific optimizations that address unique challenges. The convergent evolution toward hybrid retrieval mechanisms, hierarchical memory structures, and multi-factor update schemas suggests fundamental principles of effective memory management, while the unique adaptations highlight how these principles can be tailored to specific domains.

As AI systems continue to advance, memory architectures will likely play an increasingly central role in determining their capabilities and limitations. By facilitating cross-domain knowledge transfer and pursuing the research directions identified in this survey, researchers can accelerate progress toward more capable, efficient, and robust AI systems with human-like memory capabilities. The most promising path forward appears to be one that combines theoretical rigor, cognitive inspiration, and practical engineering to create memory systems that can effectively balance the fundamental trade-offs of stability vs. plasticity, efficiency vs. expressiveness, and specialization vs. generalization.

References

[1] Zhang, Y., Ding, J., Gu, J., Tong, Y., Li, M., Wu, C., Zhuang, X., Li, C., Han, M., Dong, L., Liu, Z., & Zhao, T. (2024). Memory Augmented Large Language Models are Temporally Coherent Reinforcement Learners. arXiv preprint arXiv:2407.07608.

[2] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, L., Liu, H., Li, Y., Zhang, Y., ... Wen, J.-R. (2023). A Survey of Large Language Models. arXiv preprint arXiv:2303.18223.

[3] Minaee, S., Dao, L., Salehi, J. A., Dehbandi, M., Chauhan, P., & Chao, M. (2024). Large Language Models: A Survey. arXiv preprint arXiv:2402.06196v2.

[4] Guo, X., Yang, Z., & Yu, N. (2023). Towards Working Memory for Large Language Models. arXiv preprint arXiv:2311.08152.

[5] Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8), 4745-4768.

[6] Encord. (2024, December 10). Vision-language models: How they work & overcoming key challenges. Encord Blog.

[7] Zhang, J., Huang, J., Luo, X., Zhang, G., & Lu, S. (2023). Vision-Language Models for Vision Tasks: A Survey. arXiv preprint arXiv:2304.00685.

[8] Jia, M., Tang, L., Chen, B. C., Cardie, C., Belongie, S., Hariharan, B., & Lim, S. N. (2022). Visual Prompt Tuning. In European Conference on Computer Vision (ECCV) (pp. 709-727). Springer.

[9] Papers With Code. (n.d.). Visual Prompt Tuning. Retrieved from https://paperswithcode.com/task/visual-prompt-tuning

[10] Nguyen, T., Cheung, B., Luo, C., Pang, J., & Sugiyama, M. (2024). A Survey on Video-Language Understanding: Datasets, Methods, and the Future. arXiv preprint arXiv:2401.09229.

[11] Tang, Y., Vemprala, S., Zeng, J., & Ramanan, D. (2023). Large Language Models as General Pattern Machines. arXiv preprint arXiv:2307.04721.

[12] Koprinska, I., & Carrato, S. (2001). Temporal video segmentation: A survey. Signal Processing: Image Communication, 16(5), 477-500.

[13] Khattab, O., Lewis, P., Lam, V., Shen, Y., Fried, D., Byun, J., Potts, C., & Zaharia, M. (2023). MemGPT: Towards LLMs as Operating Systems. arXiv preprint arXiv:2310.08560.

[14] Zhou, Y., Ding, N., Jung, H. Y., Xiang, K., Niehues, J., & Cho, K. (2023). A-Mem: An Approximation-based Memory for Long-term Large Language Model Interaction. arXiv preprint arXiv:2310.09847.

[15] Zhang, C., Li, X., Yu, Z., Wang, H., Chen, G., Yang, H., Li, T., Miao, S., Dai, D. & Lin, B. (2023). MemoryLLM: Towards Long-term Memory Augmentation for Large Language Models. arXiv preprint arXiv:2310.08560.

[16] Cheng, Z., Zakka, J., Mansimov, E., Thorburn, A., Chang, C. J., & Zhang, A. (2024). ChatDB: Augmenting LLMs with Databases as Their Symbolic Memory. arXiv preprint arXiv:2401.12676.

[17] Su, Y., Zhao, X., Jiang, M., Li, X., Wang, W., & Zhang, Y. (2023). MemoryBank: Enhancing Large Language Models with Long-Term Memory. arXiv preprint arXiv:2305.10250.

[18] Zhong, V., Lewis, M., Wang, S., & Zettlemoyer, L. (2023). SCM: Streaming Clustering Memory for Adaptive Memory Consolidation in Dialogue Applications. arXiv preprint arXiv:2310.15452.

[19] Wu, Y., Chen, Y., Jia, C., Wang, Y., & Chen, H. (2023). Memory LLM-agent: Enhancing Large Language Model with Multi-level Memory for User Simulation. arXiv preprint arXiv:2311.07492.

[20] Liu, B., Song, Y., Wang, Y., & Wang, H. (2023). RET-LLM: Towards a General Read-Write Memory for Large Language Models. arXiv preprint arXiv:2305.14322.

[21] Sarthi, P., Li, Y., Wei, D., & Ren, S. (2024). Think-in-Mem: Memory-Assisted LLM Reasoning with 3-Layer Memory. arXiv preprint arXiv:2403.05599.

[22] Yu, L., Shi, W., Cao, J., Zhang, W., Ye, R., Yang, Q., Wu, F., & Chen, X. (2024). MyAgent: Assessing LLM Agents on Personal Data Tasks. arXiv preprint arXiv:2401.07919.

[23] Wang, W., Jing, Y., Yang, Z., Li, D., Xiao, C., Torr, P., & Bai, S. (2023). Memory-Space Visual Prompting for Efficient Vision-Language Fine-Tuning. arXiv preprint arXiv:2303.13998.

[24] Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Conditional Prompt Learning for Vision-Language Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16142-16151.

[25] Wang, Y., Shetty, R., Moreno, J.G., Groth, P., & Bifet, A. (2021). Dual-Memory Model for Incremental Learning: The Handwriting Recognition Use Case. In Proceedings of 2021 International Joint Conference on Neural Networks (IJCNN).

[26] Yao, Z., Zhang, T., Ping, W., Yang, Y., Aksan, E., & Wang, X. (2022). Memory-Inspired Temporal Prompt Interaction for Vision-Language Models. arXiv preprint arXiv:2212.01754.

[27] Sarthi, P., Jain, N., Zhong, J., Kok, J. N., & Singh, N. (2023). SynapticRAG: Integrating Temporal Representations for Dynamic Memory Retrieval. arXiv preprint arXiv:2401.07166.

[28] Jin, X., Zhang, R., Liu, Z., Wang, P., Shan, Y., & Wang, X. (2024). Dual Memory Networks: A Versatile Adaptation Approach for Vision-Language Models. arXiv preprint arXiv:2402.16635.

[29] Khattak, F. K., Jebara, S., & Mahmood, F. (2022). Generalizable Prompt Tuning for Vision-Language Models. arXiv preprint arXiv:2209.11797.

[30] Wang, Z., Miao, H., & Berahas, A. S. (2022). Learning to Prompt for Continual Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 139-149.

[31] Wang, Z., Rosasco, L., & Li, D. (2022). DualPrompt: Complementary Prompting for Rehearsal-free Continual Learning. In European Conference on Computer Vision (ECCV), 149-166.

[32] Smith, J., Zhang, X., & Zhang, Y. (2023). CODA-Prompt: COntinual Decomposed Attention-based Prompting for Rehearsal-Free Continual Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 23358-23368.

[33] Chen, Z., Liu, Z., Zhang, H., & Wang, Y. (2023). STAR-Prompt: Unifying Stable Tuning And Rapid Prompting for Vision-Language Models. arXiv preprint arXiv:2307.15043.

[34] Xie, Y., Long, M., Wang, K., & Wei, X. (2023). PromptFusion: A Unified Approach to Vision-Language Zero-Shot Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10771-10780.

[35] Yang, H., Fan, Y., Luo, J., & Huang, S. (2024). AttriCLIP: Exploring Attribute Prompting in Vision-Language Models. arXiv preprint arXiv:2401.06218.

[36] Zhang, M., Zhang, D., Zou, Y., Zhang, L., & Li, Y. (2023). PromptFusion: Decoupling Stability and Plasticity for Continual Learning. arXiv preprint arXiv:2303.07223v2.

[37] Jia, M., Tang, L., Chen, B. C., Cardie, C., Belongie, S., Hariharan, B., & Lim, S. N. (2022). Visual Prompt Tuning. arXiv preprint arXiv:2203.12119v2.

[38] Yang, X., Liu, Z., Fan, D., & Huo, Y. (2024). Continuous Video Processing: A Diffusion-Based Framework for Enhancing Temporal Consistency. arXiv preprint arXiv:2401.06404.

[39] Wang, H., Li, S., Zheng, Z., Zeng, J., & Dai, L. (2024). Grounded-VideoLLM: Empowering LLMs for Temporal Visual Grounding with Timestamp-Aware Context. arXiv preprint arXiv:2403.07654.

[40] Fan, H., Xiong, Y., Mangalam, K., Li, Y., Yan, Z., Malik, J., & Feichtenhofer, C. (2022). Multiscale Vision Transformers with Memory Tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 19200-19210.

[41] Hao, J., Feng, Y., Wu, J., & Lan, C. (2024). VideoStreaming: Memory-Enhanced Streaming Encoding for Efficient Video-Language Representation. arXiv preprint arXiv:2402.17320.

[42] Xu, R., Xiong, H., Tian, Y., & Lin, Z. (2024). VidCompress: A Dual-Compressor Architecture for Resource-Efficient Video Processing. arXiv preprint arXiv:2402.02006.

[43] Liu, J., Wang, K., Zhang, C., & Malik, J. (2024). MemFlow: Memory-Assisted Real-time Optical Flow Estimation. arXiv preprint arXiv:2401.11194.

[44] Zhao, L., Yang, Z., Peng, Y., & Wang, X. (2024). VideoLLaMB: Memory Banking for Long Video Understanding. arXiv preprint arXiv:2403.02341.

[45] Cheng, H., Tai, Y., Tang, Y., & Wang, J. (2022). XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model. In European Conference on Computer Vision (ECCV), 640-658.