Preprints.org

Review

Not peer-reviewed version

Memory in Vision-Language Models: Taxonomy, Mechanisms, and Applications

Shao-Jun Xia , Yizhuo He † , Jiashen Liu † , Yuner Zhang † , Yifan Jiang * , Xiaoyang Chen * , Liangxi Liu , Chenwen Luo , Jinbao Wang *

Posted Date: 21 July 2026

doi: 10.20944/preprints202607.1539.v1

Keywords: vision-language models; memory mechanisms; memory taxonomy; multimodal intelligence; retrieval-augmented generation; long-video understanding; embodied AI; continual learning

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

Memory in Vision-Language Models: Taxonomy, Mechanisms, and Applications

Shao-Jun Xia [1], Yizhuo He [2,3,†], Jiashen Liu [4,†], Yuner Zhang [5,†], Yifan Jiang [6,*], Xiaoyang Chen [7,*], Liangxi Liu [8], Chenwen Luo [9] and Jinbao Wang [9,*]

[1] Duke University, Durham, North Carolina, USA
[2] Google, Mountain View, California, USA
[3] Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[4] University of Warwick, Coventry, United Kingdom
[5] University of Pennsylvania, Philadelphia, Pennsylvania, USA
[6] University of Southern California, Los Angeles, California, USA
[7] University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[8] Northeastern University, Boston, Massachusetts, USA
[9] Shenzhen University, Shenzhen, Guangdong, China
* Correspondence: yjiang44@usc.edu (Y.J.); xiaoyang_chen@med.unc.edu (X.C.); wangjb@szu.edu.cn (J.W.)
† Yizhuo He, Jiashen Liu, and Yuner Zhang contributed equally to this work.

Abstract

Vision-language models (VLMs) have achieved remarkable progress in multimodal understanding and generation by integrating visual and linguistic representations. However, most current VLMs lack explicit mechanisms for persistent memory, limiting their ability to maintain contextual coherence, accumulate knowledge over time, and support long-term reasoning across extended interactions. To address these limitations, a diverse set of memory mechanisms has emerged, including latent memory, key-value caches, external memory stores, retrieval-augmented memory, and hybrid memory systems. Despite rapid advances, the design space of memory in VLMs remains fragmented, and a unified understanding of its design and application remains lacking. In this survey, we provide a comprehensive review of memory mechanisms in VLMs from a system-oriented perspective. We introduce a novel four-dimensional (4D) taxonomy that organizes existing approaches along four orthogonal aspects: when memory is maintained (temporal scope), where memory is stored (storage location), what memory encodes (information stored), and how memory is accessed, updated, and utilized (memory operations). Using this taxonomy as a unifying framework, we systematically analyze representative memory-enhanced VLM architectures, review evaluation protocols for memory capabilities, and summarize key application areas, including long-video understanding, multimodal dialogue, embodied agents, and robotic reasoning. We further discuss key challenges such as scalability, memory efficiency, continual updating and forgetting, and multimodal grounding, and outline promising directions toward adaptive and unified memory systems. Overall, this survey provides a comprehensive foundation for understanding memory in VLMs and serves as a roadmap for developing next-generation multimodal systems with persistent memory and long-term reasoning capabilities. A repository associated with this survey is available at https://github.com/Xzcv-hub/Awesome-Memory-in-VLM.

Keywords: vision-language models; memory mechanisms; memory taxonomy; multimodal intelligence; retrieval-augmented generation; long-video understanding; embodied AI; continual learning

1. Introduction

Vision-language models (VLMs) have recently achieved remarkable progress across a broad spectrum of multimodal tasks [1], including image captioning [2], visual question answering (VQA) [3], video understanding [4], multimodal dialogue [5], and embodied intelligence [6]. By combining powerful visual encoders with large language models, modern VLMs demonstrate impressive capabilities

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

in multimodal perception, reasoning, and generation. However, most existing VLMs fundamentally operate in a stateless manner, where information processing is largely restricted to the current input and a finite context window. This limitation prevents models from maintaining long-term temporal coherence, accumulating knowledge from previous interactions, and performing persistent reasoning over extended periods.

Memory has emerged as a promising paradigm for overcoming these limitations. Motivated by cognitive memory systems and enabled by recent advances in neural architectures, a growing body of research has integrated diverse memory mechanisms into VLMs, including key-value caches [7], latent memory representations [8], recurrent memory states [9], memory banks [10], external memory stores [11], and retrieval-augmented modules [12]. These approaches span a spectrum from internal architectural mechanisms that maintain persistent hidden representations to external episodic or semantic memory stores that organize and retrieve textual, visual, or multimodal knowledge. By enabling models to retain historical context, access relevant prior information, and dynamically update their representations over time, memory mechanisms enhance long-range perception, reasoning, and adaptive decision-making. Consequently, memory-augmented VLMs have shown promising performance across a wide range of applications, including long-video understanding, multimodal dialogue, robotic manipulation, and embodied agents.

Table 1. Positioning of our survey among related surveys on memory. ● = explicit coverage; ◐ = partial or implicit coverage; ○ = not covered.

	Memory Taxonomy (Sec. 2)				Vision & Multimodal		Evaluation & Analysis	
Related Survey	Temporal Scope	Storage Location	Memory Entities	Memory Operations	Visual Memory	Multimodal Apps. (≥5)	Benchmarks / Metrics	Coupling Insights
Cognitive Architectures for Language Agents [13]	○	○	●	◐	○	○	○	○
Large Multimodal Agents [14]	○	○	◐	○	●	◐	◐	○
Agent Memory Mechanism [15]	◐	◐	●	●	○	○	○	○
World Models Survey [16]	○	○	◐	○	◐	◐	○	○
Rethinking Memory in AI [17]	◐	◐	●	●	◐	●	○	○
Memory in the Age of AI Agents [18]	●	○	●	◐	◐	○	○	○
Memory for Autonomous LLM Agents [19]	●	●	◐	●	○	○	◐	○
From Storage to Experience [20]	◐	○	◐	○	○	○	○	○
Video-LLM Survey [4]	○	○	○	○	●	○	●	○
Continual Learning of LLMs [21]	◐	○	◐	●	○	○	●	○
The AI Hippocampus [22]	◐	◐	●	◐	◐	●	○	○
Ours	●	●	●	●	●	●	●	●

Despite these advances, the landscape of memory mechanisms in VLMs remains fragmented. Existing studies often investigate memory from task-specific [23,24], scenario-specific [25,26], or architecture-specific [27,28] perspectives, resulting in inconsistent terminology, diverse design choices, and limited understanding of how different memory paradigms relate to each other. Moreover, memory is frequently characterized from isolated viewpoints, such as temporal scope, storage location, memory representation, or retrieval strategy, without a unified framework that captures the broader design space. Although several surveys have explored memory mechanisms in large language models, multimodal agents, and embodied systems [4,13–22], a systematic investigation of memory mechanisms specifically in vision-language models remains limited. Consequently, a comprehensive understanding of memory design, architectural integration, evaluation methodologies, and application scenarios in VLMs is still needed.

To address this gap, we present a comprehensive survey of memory mechanisms in VLMs. We introduce a unified four-dimensional taxonomy that characterizes memory from four perspectives: **when** memory is maintained, **where** memory is stored, **what** information memory represents, and **how** memory is accessed, updated, and utilized. Based on this taxonomy, we systematically review the architectural components of memory-enhanced VLMs, including memory encoding, storage, retrieval,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

and integration with multimodal backbones. We further summarize representative benchmarks and evaluation protocols, and analyze how memory facilitates a wide range of applications, including long-video reasoning, multimodal dialogue, robotics, and embodied intelligence. Finally, we discuss open challenges and future directions toward scalable, adaptive, and unified memory systems for persistent multimodal intelligence.

The main contributions of this survey are summarized as follows:

1. ***Unified Perspective.*** We provide a comprehensive review of memory mechanisms in vision-language models and establish a unified framework for understanding memory from a multi-dimensional design perspective.
2. ***4D Taxonomy.*** We propose a novel four-dimensional taxonomy that organizes memory mechanisms according to temporal scope, storage location, memory content, and memory operations, providing a principled framework for analyzing existing approaches.
3. ***Comprehensive Analysis.*** We systematically review memory architectures, evaluation protocols, and representative applications, highlighting common design principles, advantages, and trade-offs among different memory paradigms.
4. ***Challenges and Future Directions.*** We identify key challenges and emerging research opportunities toward scalable, adaptive, and cognitively inspired memory systems.
5. ***Public Repository.*** To facilitate future research, we maintain a continuously updated repository containing representative publications, benchmarks, and resources related to memory-enhanced vision-language models.

The remainder of this survey is organized as follows. Section 2 introduces the proposed four-dimensional taxonomy of memory in VLMs. Section 3 reviews architectural components and underlying mechanisms of memory-enhanced models. Section 4 summarizes representative evaluation protocols and benchmarks. Section 5 discusses major application domains enabled by memory. Section 6 highlights current challenges and limitations, while Section 7 presents future research directions. Finally, Section 8 concludes the survey and discusses promising opportunities for future work.

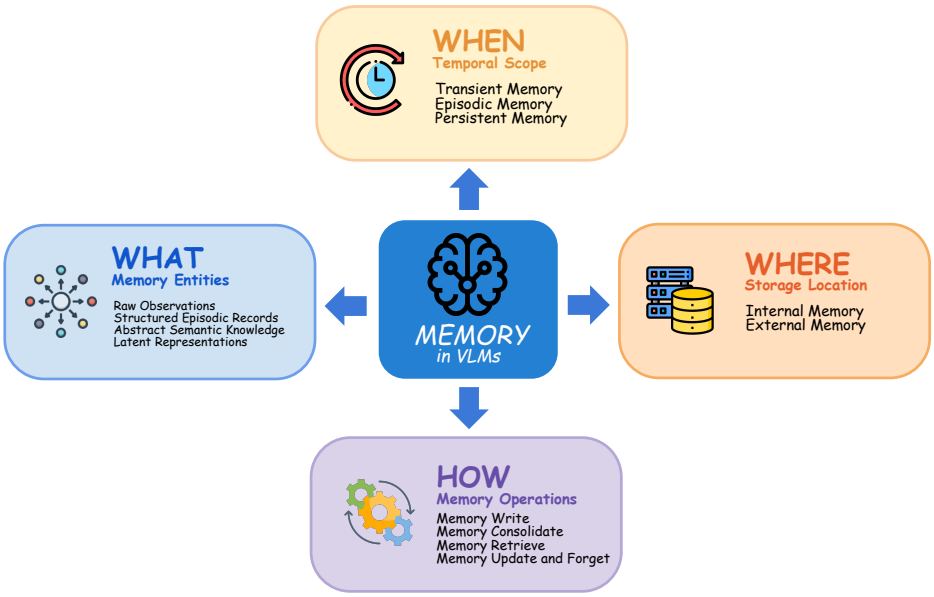

Figure 1. 4D Taxonomy of Memory Mechanisms

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

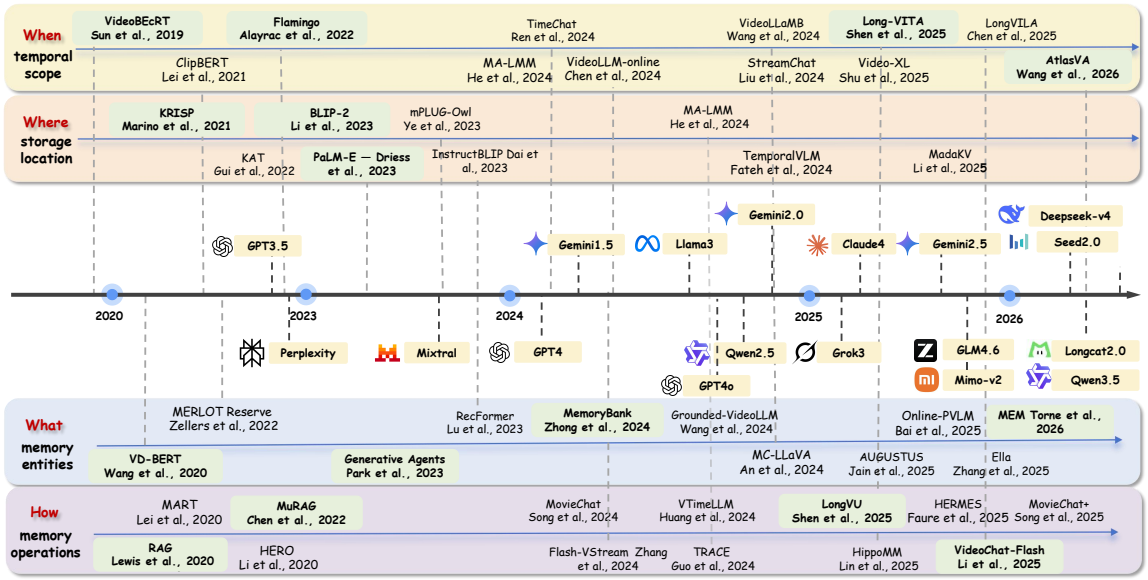

Figure 2. The proposed 4D taxonomy and the timeline of representative memory-enabled VLMs, illustrating the evolution of representative methods across the four taxonomy dimensions.

2. 4D Taxonomy of Memory Mechanisms

Memory in VLMs is not a monolithic architectural component but a multi-dimensional design space. To systematically compare diverse memory mechanisms without conflating distinct design choices, we organize the taxonomy around four orthogonal dimensions:

1. **Temporal Scope** (*When*): The temporal horizon over which stored information remains addressable.
2. **Storage Location** (*Where*): Where memory is physically or logically stored.
3. **Memory Entities** (*What*): The representational form of the stored information.
4. **Memory Operations** (*How*): How memory is written, maintained, retrieved, and updated.

2.1. Temporal Scope (When)

Temporal scope classifies memory according to the boundary across which stored information remains available for reuse. Depending on when stored information ceases to be addressable, memory may be transient, remaining available only within a single inference; episodic, persisting across multiple turns or reasoning steps within a bounded interaction; or persistent, surviving beyond the current episode and remaining reloadable for future sessions.

2.1.1. Transient Memory

Transient memory refers to information that remains available only within the current inference process. In VLMs, the active context window, including input tokens, visual representations, and intermediate caches, functions as a form of working memory: it stores task-relevant evidence for immediate reasoning but is discarded once inference concludes. Existing approaches primarily differ in how they construct, scale, and manage this temporary memory state.

Active Context Construction. Transient memory is formed through the representation and integration of multimodal information into the current model state. Early video-language systems explore this process through different strategies for representing and incorporating visual evidence. VideoBERT [29] places discretized video and language tokens within a bounded Transformer sequence, while ClipBERT [30] demonstrates that sparse clip sampling could improve the efficiency of visual evidence utilization. Flamingo [31] represents a prototypical example by integrating visual and textual

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

information through interleaved few-shot prompts rather than an external memory store. Its Perceiver Resampler compresses visual features into a fixed number of visual tokens, which are incorporated into the language model through gated cross-attention during generation, making this visual evidence available only within the active context window and cache.

Active Context Scaling. Rather than introducing persistent memory, a subsequent line of work focuses on expanding transient memory capacity within a single inference. Representative efforts include adapting long-context language modeling to multimodal inputs [32], scaling active contexts toward million-token lengths [33], and jointly optimizing model architectures and system support to enable hour-scale video understanding [34,35]. Although these approaches substantially increase the amount of information that can be processed during an interaction, enlarging the active context window alone does not establish episodic or persistent memory. Unless the resulting state is explicitly retained and reused across subsequent inference sessions, the system remains stateless beyond the current interaction. The same distinction applies to methods that maintain temporary buffers or online memory banks for streaming video analysis [36]; these mechanisms serve as transient workspaces rather than persistent stores, reflecting a broader trend in long-video modeling that substitutes expanded tracking for true memory retention [37].

Active Memory Management. Beyond scaling transient memory capacity, a critical challenge is determining which information should remain accessible within the limited active context during inference. Some recent works focus on improving the efficiency of transient memory by compressing multimodal KV caches [38], adopting modality-aware cache eviction strategies [39], and dynamically updating visual context during streaming generation [40]. These methods aim to preserve the most relevant information while operating within a finite context budget. Nevertheless, deciding what to retain remains challenging. Recent benchmarks show that simply increasing context length does not guarantee effective information utilization; models often struggle to retrieve and reason over relevant visual evidence when the active context becomes very large [41–44]. Consequently, transient memory management must balance two competing objectives: maximizing the amount of available evidence while maintaining efficient access to the information that matters most. Regardless of its capacity or management strategy, however, transient memory remains ephemeral and is discarded after inference unless its state is explicitly preserved and reused.

2.1.2. Episodic Memory

Episodic memory refers to interaction-specific information or internal state that captures context-rich experiences by preserving the sequential "what, when, and where" of an ongoing interaction. Unlike transient working context, which exists only during a single inference, episodic memory persists across multiple turns, observations, or decision steps within the same session, but is not intended to influence future independent interactions unless explicitly consolidated into persistent memory. Conceptually, it bridges transient context expansion and long-term knowledge retention.

Dialogue-bound State. In dialogue settings, episodic memory often takes the form of an evolving representation of grounded references and interaction history. VD-BERT [45] integrates image evidence with accumulated dialogue history, while multimodal dialogue state tracking explicitly models evolving visual states as new utterances and visual observations arrive [46]. RecFormer [47] maintains recurrent dialogue representations for visual dialogue, while MMCR [48] and Taking Notes [49] address multi-image, multi-turn scenarios in which previously observed visual information remains accessible across dialogue turns.

Streaming Video Sessions. Streaming video naturally exemplifies episodic memory because models must retain information from earlier video segments while processing subsequent segments within the same interaction. VideoLLM-online [50] maintains a rolling representation of previously observed video content to support later queries during an ongoing streaming session. Unlike transient context expansion, where information is discarded after a single inference, episodic memory preserves representations throughout the interaction. Mechanistically, this can be achieved either implicitly through recurrent state propagation or explicitly through event segmentation, which partitions con-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

tinuous visual streams into discrete semantic episodes [51]. Sequence-focused benchmarks further evaluate whether models can reason over temporally ordered visual evidence within an episode rather than merely recognize isolated frames [52,53].

Embodied Episodes. Embodied agents provide another natural setting for episodic memory because completing a task often requires accumulating and binding information across multiple sequential decisions. During an episode, an agent must remember previous observations, actions, instructions, and spatial relationships to determine its next action. Vision-Dialog Navigation [54] is an early example in which the agent integrates language instructions, dialogue history, and visual observations throughout a navigation task. Subsequent approaches maintain similar episode-level state through navigation histories and topological representations, enabling agents to reason about previously visited locations and past interactions [55,56]. More recent embodied systems extend this idea beyond navigation to robotics and continuous environments, where agents continually reuse task-specific experiences during ongoing interactions [57–59].

Session closure provides the operational distinction between episodic and persistent memory. Dialogue histories, streaming summaries, and embodied trajectories remain episodic as long as they are confined to the temporal lifecycle of the current interaction. Once such information is deliberately consolidated, indexed, or retained to influence future independent sessions, it becomes persistent memory.

2.1.3. Persistent Memory

Persistent memory refers to information that survives beyond the temporal boundary of a single interaction and can be retrieved, reloaded, or reused to influence future independent sessions. Unlike episodic memory, which is tied to the current interaction, persistent memory remains available after session termination and supports continual accumulation of knowledge and experience.

Cross-session Retrieval. A common form of persistent memory stores raw or structured interaction history in an external repository that survives across sessions. Retrieval-augmented language models established this paradigm by retrieving textual knowledge from an external corpus rather than relying solely on parametric knowledge [60,61]. MuRAG [62] extended this paradigm to multimodal retrieval by incorporating image-text evidence into the external corpus. MemoryBank [10] further applied this principle to conversational agents, where previous interactions are archived and later retrieved to inform future conversations. These systems are persistent because the stored information remains reusable across decoupled interactions. Whether memory resides in an external database, a vector store, or directly within model parameters belongs to the *storage substrate* dimension rather than its temporal scope.

Cross-session Experience Accumulation. Persistent memory may also take the form of an evolving cognitive profile that accumulates experience over time instead of preserving raw interaction logs. Generative Agents [63] demonstrates this by storing observations, synthesizing them into higher-level reflections, and retrieving those reflections to guide planning and social interactions across simulated days. Unlike episodic memory, these abstract records survive session termination and continue to influence future behavior. Recent personalization systems make this distinction even more explicit by maintaining user preferences and behavioral profiles across highly disjointed interactions [64,65]. Similarly, Voyager [66] accumulates a reusable library of executable skills and experiences, enabling continual improvement over successive independent tasks rather than optimization within a single episode.

Embodied Long-term Memory. Embodied agents likewise benefit from retaining knowledge acquired through historical interactions to support a continuous operational lifetime. LifelongMemory [67] stores egocentric observations that can be retrieved to answer queries long after the original experiences have ended. This contrasts with episodic memory, where interaction-specific observations are discarded once the current task concludes. KARMA [68] explicitly combines short-term episodic and long-term persistent memory modules for embodied agents, while other robotic systems maintain reusable spatial and experiential memories to support later navigation, collaboration, and reason-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

ing [69–71]. AtlasVA [72] similarly preserves visual skills as spatial heatmaps, visual exemplars, and symbolic representations that remain reusable across distinct embodied deployments.

Persistent memory does not need to reside in an explicit external repository. It can also be internalized through continual adaptation, localized fine-tuning, or dynamic parameter updates. Although such knowledge is encoded directly in model parameters rather than stored in a discrete memory structure, its defining property remains unchanged: it survives individual interactions and shapes future behavior. This further illustrates that temporal persistence (*when*) and storage substrate (*where*) are complementary but independent dimensions in our taxonomy.

2.2. Storage Location (Where)

Storage location partitions memory according to where remembered content resides and how it is accessed. Unlike temporal scope, which describes how long information persists, storage location distinguishes whether memory is maintained as *model-resident state* or as an *independently addressable store*. The same temporal behavior may therefore arise from different storage mechanisms: information that persists across requests may be encoded in model parameters or maintained in an external database, while short-lived state may appear either as intermediate activations or as temporary external records. Accordingly, we organize existing approaches into two categories. *Internal memory* stores information within the model's computation, including parameters, hidden states, latent tokens, and inference caches that are accessed only through forward execution. *External memory* stores information in standalone repositories (such as documents, embeddings, records, and user profiles) that can be queried, updated, inspected, and replaced independently of the model itself.

2.2.1. Internal Memory

Internal memory encompasses all forms of model-resident state. Although parameters, hidden activations, latent tokens, and KV caches differ in lifetime and functionality, they are unified by one property: remembered content exists only inside the model's computation and is accessed through its forward dynamics rather than through an independently queryable store.

Latent Arrays and Query Tokens. Learned latent arrays and query tokens compress visual observations into compact internal representations before language reasoning begins. Perceiver [73] provides a canonical example by introducing a learned latent array that repeatedly attends to high-dimensional inputs while remaining entirely inside network computation. The latent array serves as an internal substrate that separates memory from raw observations without introducing an external store. BLIP-2 [74] extends this idea to VLMs by using learned query tokens to bridge a frozen image encoder and a frozen language model. Its Q-Former retains visual evidence as query-conditioned latent state rather than retrievable records. Flamingo [31] similarly employs resampled visual latents as internal state, while TokenLearner [75], learnable-memory image transformers [76], Perceiver IO [77], InstructBLIP [78], and mPLUG-Owl [79] instantiate related latent bottlenecks within model computation.

Hidden-state Carryover. Sequential tasks require information from previous observations to remain available during subsequent computation. In these settings, hidden activations naturally function as memory by propagating contextual information across time:

$$m_t = h_t, \qquad h_t = f_\theta(h_{t-1}, x_t), \tag{1}$$

where the hidden state h_t summarizes previous observations while incorporating the current input x_t. This formulation captures the memory mechanisms employed by MART [80], Multimodal Transformer with Variable-Length Memory [81], and Perceiver IO [77], in which memory is carried entirely through evolving hidden representations rather than a separate storage module.

Parametric Retention. Some methods preserve knowledge by modifying and protecting model parameters instead of maintaining explicit retrievable memories. Learning Without Forgetting for Vision-Language Models [82] treats previously acquired visual-language knowledge as parametric

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

state that should survive continual updates. LLaMA-Adapter [83] stores task adaptation within lightweight attention and adapter parameters, while Memory-Space Visual Prompting [84] introduces a learned memory space embedded within model parameters for efficient adaptation. Although these mechanisms often support persistent or continually evolving memory, the remembered content remains internal because it is encoded in model weights and accessed only through model computation.

Cache-resident Memory. During long-context inference, key-value (KV) caches constitute another form of internal memory. Unlike external retrieval databases, cached keys and values exist only as intermediate activation state generated during inference. AKVQ-VL [85] illustrates this category by quantizing cached keys and values while preserving their role within the model's attention computation. Likewise, KV-cache compression and eviction methods improve inference efficiency by reorganizing or compressing activation state without introducing an independently addressable memory store [38,39]. Internal caches therefore improve computational efficiency while remaining tightly coupled to model execution, making their contents difficult to inspect, edit, or reuse outside the forward pass.

2.2.2. External Memory

External memory stores remembered content outside the model itself. Rather than encoding information in activations or parameters, these approaches maintain independently addressable repositories that can be queried, updated, inspected, and replaced without modifying the model's internal computation.

Retriever-generator Separation. Retrieval-augmented generation (RAG) exemplifies external memory by separating the storage of knowledge from its use during generation. REALM [60] established an early non-parametric memory in which retrieved documents complemented model parameters. RAG [61] subsequently became the canonical framework by coupling a generator with passages retrieved from an external corpus instead of relying solely on parametric knowledge. Because retrieved content resides outside the model, the underlying memory can be refreshed, re-indexed, or replaced without changing generator weights. This retrieval process can be abstracted as

$$m_t = \text{Retrieve}(q_t, \mathcal{M}), \tag{2}$$

where a query q_t accesses an external memory store $\mathcal{M}$. MuRAG [62] extends this paradigm to multimodal image-text retrieval, while Reveal [86] further incorporates multi-source multimodal knowledge stores during pretraining. More recent benchmarks and generalized retrieval frameworks continue to reinforce this separation between external storage and model computation [87,88].

External Knowledge Stores. For vision-language reasoning, external memory often consists of symbolic facts, image-text corpora, and multi-source knowledge repositories rather than hidden activations or model parameters. KRISP [89] and KAT [90] illustrate explicit knowledge storage for visual question answering and vision-language transformers. Retrieval-Augmented Multimodal Language Modeling [91] similarly treats retrieved multimodal examples as external context for language modeling. Symbolic reasoning, prompted retrieval, knowledge graphs, and information-seeking systems instantiate the same principle under different forms: remembered knowledge resides in an external store that is queried during inference rather than encoded within the model itself [92–95].

Web and Document Retrieval. External memory can also originate from continuously evolving information sources beyond curated retrieval databases. SearchLVLMs [96] accesses the web as an up-to-date knowledge source for LVLMs. V* [97] treats visual search results as externally retrieved evidence. R4 [98] and VisRAG [99] further extend retrieval to spatio-temporal environments and multimodal documents, while End-to-End Optimization for Multimodal RAG [100] jointly optimizes retrieval and generation without altering the fundamental separation between external storage and model inference.

Agent Memory Stores. Long-lived agents introduce another motivation for external storage: remembered experiences must remain editable, inspectable, and reusable long after an interaction

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

has ended. LifelongMemory [67] stores egocentric experiences as external records rather than latent activations. MemoryBank [10] similarly maintains conversational histories as an independently accessible memory store. Although these systems are categorized as persistent memory under the temporal dimension, their defining property here is that remembered content exists as external data that can be queried independently of the model's forward computation.

External memory improves inspectability, editability, and updateability while enabling knowledge to evolve independently of model parameters. These advantages come with additional challenges, including retrieval quality, index drift, stale evidence, and maintaining consistency between the external store and the model's internal reasoning.

2.3. Memory Entity (What)

Memory entities describe the representational form of remembered information in a VLM, independent of where it is stored or how long it persists. This dimension characterizes memory according to the nature of the stored content. We distinguish four complementary entity types that span increasing levels of abstraction: *raw observations*, which preserve input evidence with minimal processing; *structured episodic records*, which organize observations into identifiable events; *abstract semantic knowledge*, which generalize information into reusable facts, concepts, and preferences; and *latent representations*, which encode remembered information as distributed neural states rather than directly interpretable records. These entity types are not mutually exclusive, and many systems employ multiple forms simultaneously.

2.3.1. Raw Observations

Raw observations preserve multimodal evidence in a form that remains close to the original input, including image frames, video clips, audio tokens, text tokens, and patch-level visual streams before semantic abstraction or latent compression. Their defining characteristic is fidelity to the incoming signal rather than explicit structure or semantic interpretation.

Direct Sequence Storage. Some VLMs retain visual and textual tokens directly in memory, postponing information loss until later context-management stages rather than converting observations into higher-level records. VideoBERT [29] exemplifies this design by representing multimodal experience as a discretized video-language token sequence, allowing visual units and words to occupy a shared Transformer sequence before any durable record or semantic abstraction is formed. ClipBERT [30] similarly treats sparse video clips as the primary visual evidence for downstream reasoning, while HERO [101] preserves raw video-language streams through hierarchical pretraining. Flamingo [31] interleaves images directly into the prompt before resampling, and Long Context Transfer [32] propagates visual tokens across extended contexts without first converting them into structured memories. More recent long-context VLMs extend the same principle to much larger temporal scales. LongVITA [33] preserves multimodal token streams across long contexts, Video-XL [35] retains hour-scale frame evidence, and LongVILA [34] keeps large collections of frames directly addressable as raw video evidence.

Streaming Raw Observations. Streaming settings make the distinction between raw observations and higher-level memory particularly clear because new frames arrive before the model can reliably determine which moments deserve long-term retention. TimeChat [102] keeps recent video observations in a high-fidelity buffer for temporal grounding, VideoLLM-online [50] processes incoming chunks sequentially as raw streaming evidence, and StreamChat [40] treats streaming frames as an active, dynamically updating context during response generation. While subsequent downstream modules may compress, retrieve, or discard segments of this stream, the foundational memory entity at this stage is fundamentally tied to the original sensory input.

Raw observations preserve the highest level of information fidelity and remain valuable whenever downstream reasoning requires direct access to original evidence. Their limitation is that maintaining large collections of unprocessed observations quickly becomes expensive, requiring later memory mechanisms to determine what should be retained, summarized, or discarded.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

2.3.2. Structured Episodic Records

Structured episodic records organize past multimodal experience into bounded, addressable units such as dialogue turns, state slots, route steps, graph nodes, video chunks, and event logs. Unlike raw observations, these memories carry explicit temporal boundaries or structural organization that make individual experiences directly addressable. An observation becomes episodic once it is associated with a particular event, interaction, or temporal context rather than being stored as unstructured input evidence.

Dialogue Records. Dialogue-oriented VLMs naturally organize conversational history into bounded multimodal records instead of maintaining a flat prompt history. VD-BERT [45] exemplifies this design by jointly encoding dialogue history and visual context into a unified vision-dialogue representation that subsequent turns can systematically revisit. Multimodal dialogue state tracking [46] further formalizes this by introducing structured visual-object slots that evolve across turns, effectively transforming the conversational context into an explicit episodic state rather than raw dialogue text. This design choice, preserving structured interaction history over raw context windows, is mirrored across several recent paradigms. These include context-aware visual dialogue reasoning [103], recurrent dialogue history in RecFormer [47], cross-modal conversational representations [104], focused dialogue notes in Taking Notes [49], and multimodal turn-level context in MMCR [48].

Navigation Records. Embodied navigation systems similarly convert continuous perception into route-indexed records that associate observations with positions, actions, and traversal history. Vision-Dialog Navigation [54] stores route episodes that combine visual observations with conversational guidance, while history-aware navigation methods [55] explicitly maintain trajectory histories for subsequent planning. Dual-scale graph transformers [56] organize navigation experience into complementary global and local route representations, and ETPNav [59] maintains topological plans that remain tied to a specific traversal episode. More broadly, embodied-memory systems preserve action trajectories, object-goal routes, spatial-cognition memories, and robot spatiotemporal logs as reusable episodic records [57,58,69,105].

Streaming Records. Streaming video architectures transition raw observations into ordered session records once data frames acquire a stable temporal identity. Rather than processing isolated clips, these systems format historical context sequentially. For instance, VideoLLM-online [50] maintains rolling, causally linked video chunks across a continuous live session, while Qian et al. [51] propagate condensed chunk-level summaries recursively across extended video horizons to form a coherent chronological session log. Similarly, VideoLLaMB [106] introduces a recurrent memory bridge layer combined with a semantic segmentation algorithm to structurally link successive video segments without losing historical context. This requirement for ordered structure is highlighted by the Mementos framework [52], which evaluates a model's sequential visual reasoning strictly across temporally ordered image sequences where the explicit order defines the episodic memory. By shifting from ephemeral inputs to ordered series, these approaches construct a persistent temporal archive of the stream.

Agent Logs. Long-lived autonomous agents frequently structure historical context into event-centered records that pair raw observations with the discrete interactions or metadata that generated them. Rather than treating history as an undifferentiated stream, these architectures serialize memory as discrete episodic units. For instance, LifelongMemory [67] indexes continuous egocentric video feeds by extracting localized visual descriptions into distinct, searchable text-based event traces. Similarly, Generative Agents [63] structures an agent's memory stream as a ledger of timestamped event observations, preserving the chronological ground-truth before triggering higher-level reflection or planning layers. This pairing of raw interaction histories with structural metadata is also utilized by MemoryBank [10], which buffers sequential user-agent dialogues alongside personality and mood tracking before applying cognitive forgetting models to abstract the history into long-term user profiles. By anchoring memory to specific situational contexts, agent logs preserve the raw environmental history in a queryable, object-oriented format.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

Structured episodic records improve addressability, temporal grounding, and event-level reasoning by organizing experience into coherent units. Their primary limitation is that segmentation decisions are made before the future importance of individual details is known, so information omitted during episode construction cannot always be recovered later.

2.3.3. Abstract Semantic Knowledge

Abstract semantic knowledge stores information that has been generalized beyond individual experiences, including facts, concepts, rules, preferences, summaries, and symbolic relations. Unlike structured episodic records, semantic knowledge is intended to remain valid independent of the specific observation or event from which it originated. Its defining characteristic is reusable knowledge rather than event-specific recollection.

Retrieval Knowledge Stores. Retrieval-augmented systems represent semantic memory by storing interpretable knowledge objects, such as documents, text passages, image-text pairs, or structured knowledge bases, rather than opaque latent representations. Text-centric foundations such as RAG [61] and REALM [60] pioneered retrieving discrete textual passages as reusable semantic knowledge. Modern vision-language systems extend this paradigm to multimodal knowledge sources. Reveal [86] integrates image-text corpora and external knowledge into a unified multimodal memory for vision-language pretraining, while MuRAG [62] extends retrieval to multimodal image-text knowledge entries. Retrieval-augmented multimodal language modeling [91] further stores retrieved examples as explicit semantic priors. SearchLVLMs [96] grounds generation using real-time web knowledge, whereas VisRAG [99] retrieves complete multimodal documents as reusable knowledge objects rather than relying solely on textual evidence. More recent multimodal RAG frameworks further generalize this paradigm by integrating graph-structured memories, heterogeneous knowledge sources, and unified retrieval engines for multimodal knowledge access [88,94,95].

Symbolic and Explicit VQA Memories. Knowledge-based VQA makes semantic memory explicit by leveraging structured facts, symbolic relations, and localized knowledge evidence that cannot be reliably inferred from visual observations alone. KRISP [89] exemplifies this paradigm by linking detected visual concepts with external symbolic knowledge graphs to support open-domain reasoning. Similarly, KAT [90] augments vision-language reasoning by retrieving and integrating explicit textual knowledge aligned with visual concepts. Related approaches further enrich VQA by representing context as interpretable evidence rather than relying entirely on opaque latent features. These include architectures that incorporate plug-and-play textual knowledge blocks into the prompt space [92], systems that optimize visual prompting through localized external references [93], and frameworks that maintain explicit visual references through guided search and region-level retrieval mechanisms [97]. By representing memory through discrete knowledge tokens, symbolic relations, or explicit visual references, these approaches make stored context more interpretable and auditable than purely latent representations.

Personalized Knowledge. Long-lived agents often transform repeated interactions into durable knowledge about users, environments, and social conventions. MERLOT RESERVE [107] learns script-like multimodal knowledge that supports reasoning beyond individual video clips. Generative Agents [63] converts accumulated event traces into higher-level semantic reflections, while MemoryBank [10] abstracts interaction histories into persistent user facts and preferences. Similarly, AUGUSTUS [64] maintains contextual user profiles, and Ella [65] transforms embodied social experiences into long-term personalized knowledge.

Semantic Knowledge Maintenance through Editing and Pruning. Post-training knowledge maintenance further reveals the properties of semantic memory by examining how stored visual-language associations can be modified, removed, or preserved after initial training. Multimodal knowledge editing [108] directly targets specific semantic associations in VLMs while minimizing unintended changes to unrelated concepts. Complementing editing approaches, unified knowledge-maintenance frameworks [109] study how semantic capabilities encoded in model parameters can be selectively pruned, evaluated, and recovered after intervention. Together, these paradigms demonstrate

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

that semantic associations are not fixed after training but can be explicitly targeted, refined, and maintained through controlled post-training operations.

Semantic knowledge enables compact, reusable representations that readily transfer across tasks and interactions. Its primary limitation, however, is the gradual erosion of provenance: once raw observations and distinct episodes are abstracted into generalized knowledge, the underlying supporting evidence and its associated calibration or uncertainty are no longer directly recoverable.

2.3.4. Latent Representations

Latent representations store remembered information as distributed neural states rather than directly interpretable observations, episodic records, or symbolic knowledge. Unlike the previous categories, which distinguish memory according to the semantic form of the stored content, latent representations characterize memory according to its representational form. Their defining property is representational opacity: downstream computation can exploit these states, but the information they preserve is not directly accessible as a human-readable record or explicit fact. Latent representations include encoder outputs, learned tokens, query states, compressed activations, and inference-time cache tensors.

Latent Arrays. Latent-array architectures mediate high-dimensional inputs through compact sets of learned memory slots that serve as intermediate computational states. Perceiver IO [77] exemplifies this design by introducing a structured latent array that decouples computational complexity from input size, mediating arbitrary modalities through a fixed number of learned latent variables. Rather than retaining dense, uncompressed visual tokens, TokenLearner [75] dynamically constructs a sparse set of learned visual tokens from image and video features to provide compact representations for downstream processing. Related approaches similarly employ learned vectors as task-specific latent memory states, including the original Perceiver architecture [73], learnable-memory image transformers [76], and Memory-Space Visual Prompting [84], which parameterizes prompt information within a highly compressed latent memory space.

Query-token Visual States. Many vision-language models maintain visual information through learned query tokens that bridge vision encoders and language models. Flamingo [31] uses a Perceiver Resampler to transform variable-length visual features into a small, fixed set of interface tokens consumed by the language model. BLIP-2 [74] similarly introduces a Q-Former to extract compact visual query states from a frozen image encoder. Building on this paradigm, InstructBLIP [78] conditions these visual queries on task instructions to produce instruction-aware latent tokens, while mPLUG-Owl [79] employs a visual abstractor module to compress dense visual features into a compact sequence of modular latent tokens.

Compressed Video Latents. Long-video understanding increasingly relies on compressing dense frame-level observations into compact latent states rather than retaining all raw visual tokens. MovieChat [110] exemplifies this paradigm by compressing dense video sequences into sparse token representations for efficient long-video reasoning. Similarly, LongVU [111] stores adaptive spatiotemporal representations that reduce spatial and temporal redundancy while preserving long-range video dependencies, whereas VideoChat-Flash [112] introduces a hierarchical video compression strategy that transforms long-duration visual context into compact latent representations for extended video understanding.

Recurrent Latent Bridges. Recurrent memory mechanisms propagate latent states across successive observations, allowing models to access historical context without replaying the complete input sequence. Variable-Length Memory architectures [81] instantiate this design by encoding embodied experiences into adaptive latent states for autonomous navigation. Similarly, MA-LMM [36] maintains long-video histories through a memory-augmented latent bank that supports information accumulation over extended temporal horizons, while VideoLLaMB [106] extends this recurrent latent bridging paradigm to long-context video-language modeling over extended temporal sequences.

Cache Representations. Transformer-based key-value (KV) cache mechanisms expose latent memory during inference because the stored information consists of activation tensors rather than

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

interpretable records. Cache-centric frameworks explore how these latent states can be selectively retained, evicted, compressed, or quantized to extend multimodal context windows [38,39,85]. Although these caches do not correspond to explicit semantic entities, they function as transient memory states that preserve task-relevant intermediate information required for continued sequence generation and multimodal inference.

Latent representations provide computationally efficient memory by compressing information into forms compatible with modern neural architectures. However, this efficiency comes at the cost of interpretability, provenance, and direct editability, requiring additional mechanisms to determine what information has actually been preserved.

2.4. Memory Operations (How)

Memory operations describe the actions through which a VLM interacts with memory, independent of the memory content and storage location. This axis follows the operational lifecycle of memory: *writing* admits new information into a memory-bearing representation, *consolidation* organizes and stabilizes existing memory, *retrieval* selects relevant information at use time, and *updating* modifies or removes previously stored content. Together, these operations characterize how memory systems transform, maintain, and exploit information over time.

2.4.1. Memory Writing

Memory writing refers to the process through which new multimodal evidence enters a memory-bearing state. The resulting representation may be transient, persistent, internal, or external, but the defining operation is the admission of new information rather than its later organization or modification. Writing mechanisms can be implemented through encoders, query-token interfaces, recurrent controllers, note generators, adapters, or combinations of these components.

Encoding Controllers. Encoding controllers introduce incoming multimodal evidence into intermediate states that support later reasoning or retrieval. HERO [101] provides an early example of hierarchical video-language representation construction, where clip-level and video-level contexts are integrated into shared representations for downstream grounding tasks. Rather than forming a persistent memory store, its memory mechanism illustrates how encoded context can be maintained as a memory-bearing state for subsequent computation. MART [80] similarly introduces a recurrent memory mechanism that updates a latent state while generating grounded video descriptions, demonstrating how sequential visual information can be accumulated during generation. More recent long-video systems explicitly treat incoming visual streams as information to be admitted into memory representations: MA-LMM [36] maintains long-video context through a memory-augmented architecture, VideoLLM-online [50] converts streaming video chunks into online memory representations, and streaming long-video encoding approaches such as TimeChat [102] progressively incorporate temporal observations for later reasoning [51].

Query-token-based Admission. Query-token mechanisms provide a compact interface for admitting visual information into language-model-compatible representations. Flamingo [31] introduces a cross-modal resampler that converts visual features into a small set of latent representations that can be consumed by the language model. The Perceiver architecture [73] similarly demonstrates how high-dimensional inputs can be projected into compact latent arrays through learned attention-based bottlenecks. TokenLearner [75] further explores adaptive token selection by learning compact visual tokens from dense feature maps. In vision-language models, BLIP-2 [74] employs Q-Former queries to extract task-relevant visual representations before interaction with a frozen language model, while InstructBLIP [78] extends this mechanism by conditioning the visual query process on instructions. These methods differ in implementation, but share the operational principle of transforming incoming visual evidence into compact states that can be subsequently accessed by downstream modules.

Recording Information during Interaction. Interaction-time writing captures information generated dynamically during multi-turn conversations or agent trajectories and transforms it into reusable memory representations for future interactions. MemoryBank [10] exemplifies this process by explicitly

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

writing conversation-derived memories into a long-term memory structure to support personalized response generation across sessions. Generative Agents [63] similarly records observations, reflections, and behavioral experiences into persistent memory streams that can later support planning and decision-making. Taking Notes [49] extends this paradigm to multimodal dialogue by treating automatic note generation as an explicit writing mechanism that transforms long interaction histories into concise and reusable summaries. Multimodal dialogue state tracking frameworks [46] further formalize interaction-time writing by encoding turn-level evidence into structured slots that represent evolving dialogue states.

Writing into Model State. Some memory mechanisms incorporate information directly into model parameters or internal representations rather than maintaining an external memory store. Parameter-efficient adaptation approaches, including adapter-based and prompt-based methods, introduce lightweight trainable components that modify how multimodal information is processed and retained within the model itself [83,84,113]. Unlike explicit memory stores that maintain retrievable records, these parametric approaches encode information implicitly through changes to model states, enabling subsequent inference to reflect previously incorporated knowledge. This form of writing differs from external memory construction because information is stored as distributed model behavior rather than as individually addressable memory entries.

Writing determines the information available to subsequent memory operations. However, admission is inherently selective: early representations may discard details, introduce biases, or constrain what can later be recovered. Effective writing therefore requires balancing compression with preservation of future utility.

2.4.2. Memory Consolidation

Memory consolidation refers to downstream operations that maintain, reorganize, and stabilize information after it has entered a memory-bearing state. Unlike writing mechanisms that admit novel external evidence, consolidation operates strictly on existing internal states to optimize their usability, efficiency, and architectural longevity. Within modern vision-language and autonomous agent systems, these strategies typically materialize as context compression, hierarchical aggregation, semantic indexing, or graph-based restructuring.

Compressing Memory after Encoding. Compression-based consolidation reduces redundancy within existing memory representations while retaining information that remains useful for future access. MovieChat [110] provides a representative example of this paradigm by organizing long-video understanding through a two-stage memory mechanism that aggregates dense visual observations into compact short-term and long-term representations. LongVU [111] similarly introduces adaptive visual compression to maintain extended video context by reducing redundant visual tokens after encoding, while VideoChat-Flash [112] extends this principle through hierarchical compression across multiple levels of video representation. In these architectures, compression functions not as an initial admission mechanism, but as a maintenance process that determines how previously encoded information is retained under limited memory budgets. LongVILA [34] and MA-LMM [36] further investigate efficient maintenance of long-horizon visual context through optimized memory representations for extended sequences.

Cache Consolidation. KV-cache management represents a distinct form of memory consolidation in which memory is primarily maintained as a transient inference state. Under finite computational budgets, cache consolidation selectively maintains, compresses, and discards key and value representations to reduce memory growth while preserving useful context for subsequent generation. LOOK-M [38] formulates multimodal KV-cache management as a retention problem, selecting informative visual tokens under strict memory constraints. MadaKV [39] and AKVQ-VL [85] further introduce modality-aware cache selection and compression strategies, demonstrating how transient inference memory can be efficiently maintained through adaptive eviction and runtime quantization.

Retrieval Index Construction. Index construction shifts memory from admitted content to organized stores that can survive repeated retrieval. When maintenance is the focus, REALM [60]

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

stores evidence in an indexed retrieval corpus, and as consolidation policy, RAG [61] preserves passages for repeated access. From the storage-operation view, MuRAG [62] organized multimodal evidence into retrievable memory, while Reveal [86] maintained retrieved visual-linguistic knowledge in an external store. VisRAG [99] consolidated document images as visual retrieval units rather than converting them to text first. Multimodal RAG frameworks broaden this storage family through graph and all-format stores [88,94].

Embodied Memory Consolidation. Embodied consolidation maintains interaction and spatial memories as organized stores rather than transient action context. For long-term preservation, MemoryBank [10] keeps user histories beyond a single turn, and LifelongMemory [67] records multimodal agent experience as a maintained store. Mem4Nav [105] and KARMA [68] organize navigation and embodied memories for later use without treating each access as a new write. AtlasVA [72] further separates embodied memory into visual and symbolic layers, using trajectory-derived atlases as reusable skill stores rather than one-step action context. Related robotics systems consolidate spatial-temporal and multi-memory traces for long-horizon behavior [69–71].

Consolidating Embodied and Spatial Memory. Embodied memory consolidation transforms transient, action-level perception into stabilized, persistent spatial-temporal structures capable of surviving extended agent lifetimes. For personalized agentic workflows, MemoryBank [10] and LifelongMemory [67] explicitly write and continuously maintain user histories and multimodal behavioral traces across disjointed, long-term sessions rather than treating them as isolated context window prompts. In autonomous navigation, architectures like Mem4Nav [105] and KARMA [68] organize streaming spatial trajectories and environmental maps into structured episodic memory buffers that bypass redundant write operations during re-localization. Pushing this division further, AtlasVA [72] bifurcates embodied memory into decoupled visual and symbolic layers, distilling raw trajectory histories into reusable, hierarchical skill atlases that serve as enduring behavioral primitives. This paradigm is mirrored in contemporary robotic frameworks, which systematically consolidate multi-memory traces to stabilize long-horizon planning and manipulate objects over extended, continuous operational timelines [69–71].

Consolidation policy turns capacity into a design variable: compression, indexing, and hierarchical organization can each extend the useful horizon, but each choice decides which distinctions remain recoverable. Strong systems make preservation explicit rather than hiding it in context length.

2.4.3. Memory Retrieval

Memory retrieval refers to the operation of selecting and accessing previously stored information when it is needed for reasoning. Unlike writing and consolidation, which determine how information enters and remains in memory, retrieval governs what the model reads at inference time. It may involve forming a query, routing to an external source, accessing symbolic knowledge, recalling episodic records, or attending over stored internal states.

Search-based Retrieval. Search-based retrieval treats memory access as a query-driven selection problem over external information sources. SearchLVLMs [96] provides a representative example by introducing web search as a plug-and-play retrieval component for large vision-language models, allowing models to acquire external evidence only when needed during inference. VisRAG [99] extends retrieval to document images by directly retrieving visually encoded documents rather than requiring conversion into text representations beforehand. Earlier retrieval-augmented architectures such as RAG [61] and REALM [60] similarly demonstrate how language models can access external corpora through query-dependent retrieval mechanisms. For multimodal settings, MuRAG [62] incorporates both visual and textual evidence retrieval, while REVEAL [86] enables retrieval over large-scale visual-linguistic knowledge sources. Recent retrieval-augmented multimodal systems further extend this paradigm through improved visual search, retrieval control, and multimodal evidence integration [91,97,98,100].

Symbolic and Knowledge-based Access. Symbolic retrieval mechanisms complement visual representations by accessing structured knowledge sources that are difficult to infer from visual inputs

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

alone. KRISP [89] combines visual reasoning with symbolic knowledge retrieval to improve vision-language reasoning capabilities, while KAT [90] introduces external knowledge retrieval for knowledge-intensive vision-language tasks. Plug-and-Play VQA [92] further illustrates this access paradigm by combining pretrained vision and language components at inference time without requiring additional end-to-end model modification. These approaches highlight retrieval as a mechanism for selectively incorporating information beyond the immediate visual context.

Episodic Memory Retrieval. Episodic retrieval enables agents and dialogue systems to access previously recorded experiences when making new decisions. MemoryBank [10] retrieves previously stored user memories to support personalized responses across interactions, while Generative Agents [63] accesses stored observations and reflections to guide planning and future behavior. In embodied settings, ReMEmbR [69] and Mem4Nav [105] retrieve previously encountered spatial-temporal experiences to support navigation and decision-making over long horizons. These systems demonstrate that retrieval transforms accumulated experience from passive storage into actionable evidence.

Evaluating Retrieval Capability. Retrieval-oriented benchmarks assess whether models can effectively locate relevant evidence from established information sources or memory representations. Multimodal retrieval-augmented generation benchmarks evaluate evidence selection across visual, textual, and structured knowledge sources [87,88,94,95]. Long-video benchmarks further examine whether models can retrieve relevant evidence distributed across extended temporal sequences [41,43]. These evaluations reveal retrieval failures that are distinct from memory-writing or storage limitations, including inaccurate ranking, incomplete evidence coverage, and inefficient retrieval strategies.

Retrieval quality determines whether stored memory can be effectively utilized as evidence under realistic inference budgets. While effective retrieval reduces context overhead and improves reasoning efficiency, retrieval failures in ranking, routing, or evidence coverage remain distinct from failures in memory writing and storage.

2.4.4. Memory Updating and Forgetting

Memory updating and forgetting describe operations that modify existing memory after it has been created while controlling the retention of previously acquired information. These operations include editing stored knowledge, incorporating new information, preserving prior knowledge against interference, refreshing transient states, and selectively removing obsolete content. Unlike writing, which introduces new information into memory, updating changes the content, representation, or accessibility of previously retained memory.

Editing Stored Knowledge. Editing mechanisms modify existing model knowledge without retraining the entire model from scratch. Can We Edit [108] provides a representative study of whether visual knowledge encoded in multimodal models can be edited while controlling unintended side effects. This line of work treats updating as a targeted modification problem, where the goal is to revise specific associations while preserving unrelated capabilities. Visual-oriented fine-tuning and unified knowledge maintenance approaches further explore how vision-language models can integrate corrected information while minimizing disruption to previously acquired knowledge [109,114].

Continual Learning and Forgetting Control. Continual learning frames sequential adaptation as an incremental memory update problem, where catastrophic forgetting arises when model updates interfere with previously acquired knowledge. Learning without Forgetting for VLMs [82] provides a representative example by balancing the acquisition of new multimodal tasks with the preservation of prior capabilities. Other approaches investigate interference reduction, replay-based learning, and synthetic data generation to improve knowledge retention during sequential updates [115,116]. Instruction-aware preservation and visual conditioning strategies further explore controlled adaptation under competing old and new information [117,118].

Updating Transient Inference Memory. Updating can also occur within short-lived memory states during inference, where finite computational budgets require memory to be continually refreshed, compressed, or selectively forgotten. KV-cache management methods modify retained attention states

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

by selectively removing, compressing, or refreshing stored key-value representations. LOOK-M [38] updates cache contents through importance-based retention policies, while MadaKV [39] adapts cache eviction according to modality-aware importance. AKVQ-VL [85] further updates transient memory through adaptive quantization and selective retention of cache representations. Streaming systems such as StreamChat [40] and VideoLLM-online [50] continuously refresh memory states as new visual observations arrive, enabling long-context interaction under changing environments.

Updating Persistent Memory Stores. Persistent memory systems require mechanisms for revising previously recorded experiences while maintaining consistency over time. MemoryBank [10] updates long-term user memories across multiple interactions to support personalized assistance, while KARMA [68] updates embodied memories to improve future action selection. Complementing these explicit memory stores, parameter-efficient adapter approaches update persistent knowledge implicitly by modifying lightweight parameter components rather than maintaining separate external memory structures [76,83,84].

Controlled updating enables memory systems to adapt over time, but every modification introduces the possibility of unintended information loss. Effective memory updating therefore requires balancing plasticity and stability: incorporating new evidence while preserving previously useful knowledge and controlling collateral changes across modalities.

3. Memory as a Coupled System

In practical VLMs, the four dimensions of our memory taxonomy do not represent independent design choices but jointly define a single memory system. The previous subsections introduced each dimension separately for analytical clarity, yet real systems exhibit strong dependencies across them: decisions along one dimension often constrain, or even preclude, the choices available along the others. Rather than arising from the model architecture itself, these dependencies are primarily driven by the task. Specifically, a task determines how long information must remain available, that is, its temporal scope. This temporal requirement then influences the appropriate storage location, the forms of memory that can be stored, and the operations required to manage them. In particular, storage location must support the required lifetime of the memory, while the representation of the memory entity is closely tied to where it is stored and which operations can be performed on it. Guided by this observation, we examine the taxonomy along the temporal dimension and consider the three regimes (transient, episodic, and persistent) in turn, showing how each temporal scope constrains the remaining three dimensions.

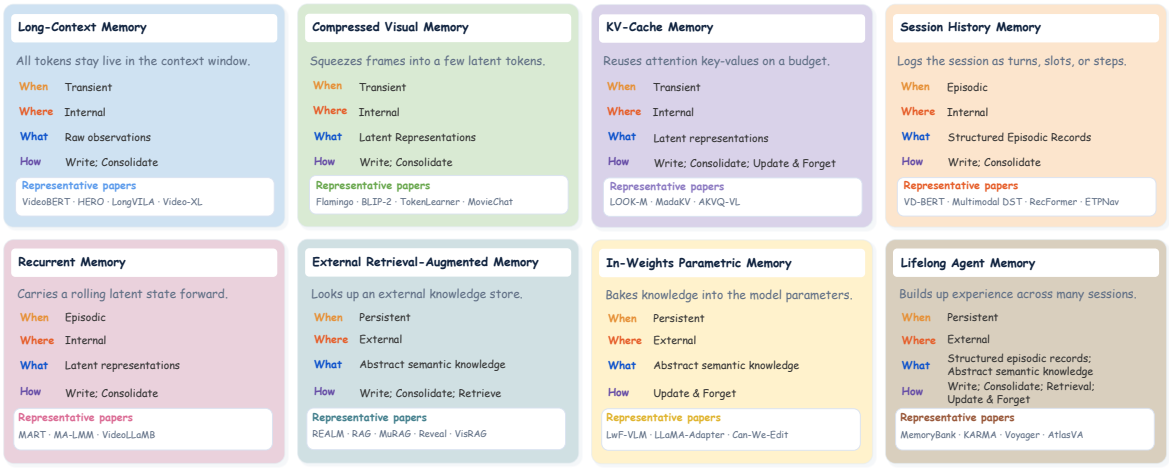

Figure 3. Eight representative memory archetypes in VLMs under the proposed 4D taxonomy, each summarized by its *when*, *where*, *what*, and *how* values with representative methods.

When information is required only within a single execution, such as answering questions over an extended video or conditioning generation on interleaved demonstrations, the temporal scope is

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

transient. In this regime, internal memory is sufficient because the data does not need to outlive the active inference session. The memory entity therefore consists of evidence that can be maintained directly within the model, either as raw observations retained close to the input [29,34,35,101] or as latent representations that compress this evidence into a compact set of learned tokens [31,73,74]. Memory management is correspondingly limited to writing and consolidation, including compression, eviction, and reorganization of the key-value cache under capacity constraints [38,39,85]. External retrieval, updating, and forgetting are unnecessary because no persistent memory store is maintained.

When information must remain available throughout a bounded interaction, such as a multi-turn grounded dialogue or a streaming video session, the temporal scope becomes episodic. Memory still remains internal because the interaction is self-contained, but the longer temporal horizon motivates more structured memory entities. These include explicit episodic records, such as dialogue turns, state slots, or navigation steps, which partition past interactions into addressable units [45,46,59], together with recurrent or compressed latent representations that propagate context without replaying the full interaction history [50,106,110]. Memory operations remain centered on writing and consolidation across turns, while external retrieval is still unnecessary because information is preserved within the ongoing interaction rather than an external memory store.

Only when information must persist across multiple sessions, for example to personalize interactions over time or support lifelong embodied agents, does the temporal scope become persistent. At this point, maintaining memory outside the current interaction becomes beneficial, either through an explicit external memory store or through parametric memory embedded in model weights. External memory stores primarily retain semantic knowledge that remains meaningful independently of the interaction in which it was acquired, and are accessed through retrieval rather than by replaying previous computations [60–62,99]. Parametric memory likewise stores semantic knowledge but cannot be queried explicitly; instead, it is modified through parameter updates or model editing, making updating and forgetting its defining operations [82,83,108]. Consequently, the complete lifecycle of writing, consolidation, retrieval, updating, and forgetting emerges only in persistent memory systems, particularly lifelong agents that continuously maintain and revise knowledge across sessions [10,63,68].

Viewed through this temporal perspective, the taxonomy occupies a constrained region of the four-dimensional design space rather than an unconstrained Cartesian product. Many combinations are absent because they violate natural couplings between dimensions. For example, latent representations are inherently internal and therefore cannot be stored directly in an external memory, while transient hidden states cannot maintain information beyond the lifetime of an inference. Other regions remain vacant not because they are fundamentally impossible, but because they are under-explored. For instance, an external memory system that stores raw multimodal observations and retrieves them on demand could relax existing design assumptions, occupying an otherwise sparse region of the taxonomy. This perspective not only explains why current VLM memory architectures converge on a small set of recurring designs, but also highlights opportunities for future research by mapping the boundaries of the remaining design space.

4. Benchmarks & Evaluation Metrics

4.1. Benchmarks and Datasets

Memory-related evaluation spans diverse tasks that probe different aspects of memory, ranging from long-context reasoning to continual adaptation and personalized interaction. To provide a structured overview, we categorize existing benchmarks and datasets into eight groups according to the memory being evaluated, the task setting, and the primary evaluation objective: long video understanding and generation, image reasoning and description, continual learning, embodied AI, multimodal RAG, egocentric vision, personalized assistants, and virtual-environment interaction. Table 2 summarizes representative benchmarks and datasets together with their data scales and task types.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

Long Video Understanding and Generation. These benchmarks evaluate memory over long or streaming videos, where relevant information may be separated by substantial temporal gaps. They assess temporal retention, cross-segment retrieval, long-range reasoning, and generation or prediction conditioned on historical observations.

Table 2. Memory-related benchmarks and datasets across vision-language applications.

Application	Benchmark / Dataset	Data Size	Task Type
Long Video Understanding and Generation	LVBench [119]	103 videos; 1.5K MCQs	Offline long-video understanding
	SceneBench [120]	2,485 videos; 8,903 QA	Offline long-video understanding
	LongVideoBench [41]	3.8K videos; 6.7K MCQs	Offline long-video understanding
	OVBench [121]	16 subtasks	Online long-video understanding
	OVO-Bench [122]	644 videos; 2.8K annotations	Online long-video understanding
	StreamingBench [123]	900 videos; 4,500 QA	Online long-video understanding
	Neptune [124]	2,405 videos; 3,268 questions	Reasoning over long videos
	CG-Bench [125]	1,219 videos; 12,129 QAC triplets	Reasoning over long videos
	MAG-Bench [126]	176 videos	Long video generation
	MBench [127]	12 memory sub-dimensions	Long video prediction
	MIND [128]	250 videos; 8 scenes	Long video prediction
Image Reasoning and Description	Mementos [52]	4,761 image sequences	Multi-image reasoning
	VisChainBench [129]	1,457 tasks; 20K+ images	Multi-image reasoning
	Instagram dataset [130]	1.1M posts; 6.3K users	Image captioning
Continual Learning	MTIL [131]	11 datasets; 1,201 classes	Incremental recognition
	VLCL [132]	8 image-caption datasets; 3 tracks	Incremental recognition
	MLLM-CL [133]	Domain and ability continual settings	Continual knowledge updating
	CLeaRS [134]	207K+ image-text pairs; 10 subsets	Continual knowledge updating
	CLiMB [135]	13 vision-language tasks	Continual task adaptation
	CoIN [136]	10 datasets; 8 task types	Continual task adaptation
	MLLM-CTBench [137]	16 datasets; 7 tasks; 6 domains	Continual task adaptation
	CL-CrossVQA [138]	5 cross-domain VQA datasets	Continual task adaptation
	ViLCo-Bench [139]	10-min videos with language queries	Continual task adaptation
	CL-VISTA [140]	8 video tasks; 6 protocols	Continual task adaptation
Embodied Navigation and Action	FindingDory [141]	60 memory-intensive tasks	Vision-language navigation
	LMEE-Bench [142]	166 tasks; 828 goals; 406 QA	Vision-language navigation
	3DMem-Bench [143]	26K+ trajectories; 2,892 tasks	Vision-language navigation
	RMBench [144]	9 manipulation tasks	Vision-language action
	RoboMME [145]	16 manipulation tasks; 770K timesteps	Vision-language action
	RoboMemArena [146]	26 tasks; 1K+ steps per trajectory	Vision-language action
	LIBERO-Mem [147]	Non-Markov manipulation suite	Vision-language action
	DRIVESPATIAL [148]	15.6K QA; 20 tasks; 5 datasets	Autonomous driving
Multimodal RAG	Dyn-VQA [149]	1,452 questions; 9 domains	Knowledge-augmented visual QA
	Visual-RAG [150]	400 queries; 103.8K images	Knowledge-augmented visual QA
	M4-RAG [151]	80K+ image-question pairs; 42 languages	Knowledge-augmented visual QA
	M^2RAG [152]	4 multimodal RAG tasks	Knowledge-augmented visual QA
	MMDocRAG [153]	4,055 QA	Multimodal document retrieval and QA
	UniDoc-Bench [154]	70K PDF pages; 1,600 QA	Multimodal document retrieval and QA
	VisDoMBench [155]	2,271 queries; 1,277 docs	Multimodal document retrieval and QA
	Chart-MRAG Bench [156]	267 docs, 627 charts, 4,738 QA	Multimodal document retrieval and QA
	MRAG-Bench [157]	16,130 images; 1,353 MCQs	Scenario Grounded multimodal RAG
	RAVENEA [158]	10K+ Wikipedia documents	Scenario Grounded multimodal RAG
	FinRAGBench-V [159]	60.8K CN pages, 51.2K EN pages	Scenario Grounded multimodal RAG
Egocentric Vision	Ego4D [160]	3,670h video; 931 wearers	Episodic retrieval
	AMB [161]	20K+ visual queries	Episodic retrieval
	EgoMemoria [162]	629 videos; 7,026 questions	Episodic retrieval
	EgoMemReason [163]	500 questions	Episodic retrieval
	EGOSTREAM [164]	2,250 questions	Episodic retrieval
	TeleEgo [165]	70h+ video; 3,291 QA	Wearable assistants
	EgoLifeQA [166]	6K QA; 266h video	Wearable assistants
	SuperMemory-VQA [167]	52.9h video; 4,853 QA	Wearable assistants
	MyEgo [168]	541 videos; 5K questions	Wearable assistants
	EgoIntrospect [169]	180h; 60 participants	Wearable assistants
Personalized Assistants	MMPB [170]	10K pairs; 111 concepts	Concept personalization
	MyVLM [171]	29 concepts	Concept personalization
	Yo' LLaVA [172]	40 concepts	Concept personalization
	MC-LLaVA [173]	2,119 images; 16,694 QA	Concept personalization
	CoViP [174]	2.8K train; 1.3K test	Concept personalization
	Persona-MME [175]	2K+ cases; 200 personas	Conversational memory
	ATM-Bench [176]	1,038 questions; 4-year history	Conversational memory
	LCMP [177]	30 concepts; 420 dialogues	Conversational memory
	MemLens [178]	789 questions; 32K–256K context	Conversational memory
	Mem-Gallery [179]	240 sessions; 1.7K QA	Conversational memory
Virtual-environment Interaction	EMemBench [180]	15 text games	Open-world agents
	MineNPC-Task [181]	216 subtasks; 8 players	Open-world agents
	MineExplorer [182]	1,497 atomic tasks; 813 instances	Open-world agents
	MemoryArena [183]	5 multi-session task suites	GUI agents
	MemGym [184]	5 tracks; 4 agentic regimes	GUI agents
	MemGUI-Bench [185]	128 tasks; 26 applications	GUI agents

Image Reasoning and Description. These benchmarks evaluate memory across multiple images, image sequences, or visual contexts. Compared with long-video tasks, they place less emphasis on continuous temporal dynamics and instead focus on cross-image reasoning, visual context integration, and memory-aware description generation.

Continual Learning. Continual learning benchmarks evaluate memory across sequential training stages, where models must acquire new knowledge, domains, tasks, or instructions while preserving previously learned capabilities. They primarily measure knowledge retention, adaptation to new tasks, and resistance to catastrophic forgetting.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

Embodied AI. Embodied AI benchmarks evaluate memory in agents operating in physical or simulated environments. They require models to maintain and utilize historical observations, object states, goals, and action histories to support navigation, manipulation, autonomous driving, and long-horizon decision-making.

Multimodal RAG. Multimodal RAG benchmarks evaluate retrieval-based external memory by measuring how effectively models retrieve, integrate, and reason over multimodal evidence, such as images, documents, charts, and structured knowledge sources, with evaluation emphasizing evidence grounding and cross-modal reasoning.

Egocentric Vision. Egocentric vision benchmarks evaluate memory from first-person visual streams, particularly those capturing daily activities and personal experiences. They emphasize episodic retrieval and first-person assistant capabilities, including recalling previously observed people, objects, events, and contextual information.

Personalized Assistants. Personalized-assistant benchmarks evaluate persistent user memory, including personalized visual concepts, preferences, experiences, and interaction histories. They assess whether models can retain user-specific information across interactions and use it to generate personalized responses.

Virtual-environment Interaction. Virtual-environment benchmarks evaluate memory in digital environments such as games, graphical user interfaces, web applications, and computer-use tasks. They assess whether agents can maintain task states, remember previous actions, and leverage long-term interaction histories during multi-step or multi-session execution.

4.2. Metric Taxonomy

Existing memory benchmarks employ diverse evaluation metrics because they target different aspects of multimodal memory, including long-context reasoning, egocentric recall, continual adaptation, embodied interaction, retrieval-augmented generation, personalization, and virtual-environment tasks. Consequently, their metrics are not directly comparable under a single unified scale. Instead, they capture complementary properties of memory, such as long-context utilization, information retention, knowledge preservation, action execution, evidence retrieval, generation consistency, and task success.

Based on their evaluation objectives, we organize existing metrics into the following eight categories:

- **Task Performance** measures overall task success, including accuracy, QA accuracy, MCQ accuracy, task score, and pass rate.
- **Retrieval and Grounding** measures the ability to retrieve, localize, and correctly ground relevant visual, textual, or multimodal evidence.
- **Memory Recall and Retention** evaluates whether models can successfully recover and utilize information from previous observations, sessions, or user interaction histories.
- **Temporal and Long-Context Evaluation** measures performance across varying video durations, context lengths, temporal dependencies, and streaming settings.
- **Continual Learning** measures the ability to preserve previously acquired knowledge while adapting to new tasks, domains, instructions, or knowledge.
- **Generation Quality and Consistency** evaluates the fidelity, visual quality, and temporal consistency of generated multimodal content.
- **Embodied and Agentic Execution** measures how effectively memory supports decision-making and action execution in embodied, robotic, or interactive environments.
- **Efficiency and Robustness** evaluates computational efficiency and robustness, including memory overhead, latency, computational cost, and performance under distracting or challenging memory conditions.

Representative metrics within each category are summarized in Table 3.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

Table 3. Unified metric taxonomy for memory evaluation in multimodal models.

Metric Type	Representative Metrics	Evaluation Focus
Task Performance	Accuracy, QA accuracy, MCQ accuracy, open-ended QA score, task score, pass rate	Task outcome
Retrieval and Grounding	Recall@K, R@K, MRR, nDCG, mAP, mIoU, grounding accuracy, citation precision/recall	Evidence use
Memory Recall and Retention	Recall, memory recall accuracy, episodic memory score, memory persistence time, performance over memory length	Memory retention
Temporal and Long-Context Evaluation	Accuracy by video length, accuracy by context length, accuracy by temporal category, real-time accuracy	Time-aware use
Continual Learning	Average accuracy, final accuracy, forgetting, backward transfer, forward transfer, zero-shot performance	Old-new balance
Generation Quality and Consistency	FVD, CLIP score, VBench score, subject consistency, background consistency, motion smoothness	Visual fidelity
Embodied and Agentic Execution	Success rate, task completion rate, subtask success rate, SPL, episode return, coverage	Action success
Efficiency and Robustness	Latency, memory footprint, inference cost, memory encoding time, robustness under distractors	Cost and stability

Figure 4. Major application domains of memory in VLMs

5. VLM Memory Applications

5.1. Long Video Understanding and Generation

Offline long-video understanding assumes the complete video is available before inference, allowing memory to be constructed over the full temporal context rather than incrementally during streaming. Within this setting, memory primarily supports retrieving, organizing, and reasoning over information distributed across extended temporal horizons.

5.1.1. Offline Long-Video Understanding

[TODO: summarize these part.]

Long-form Video QA. Long-form Video QA requires preserving and retrieving question-relevant evidence across extended video histories, rather than merely understanding individual clips. Early work already connected this task with explicit memory access, as exemplified by DeepStory [186], which represents video history as a retrievable memory. Subsequent methods externalized long videos

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

into textual histories [187] and further organized them into more concise and structured language repositories for long-range reasoning and question answering [188]. Later work shifted the focus from retaining more history to retaining the history most relevant to the query. MovieChat [110] and MovieChat+ [189] improve long-term memory through sparse compression and query-aware retention, respectively, while Glance and Focus [190] progressively retrieves evidence from global event summaries to local details. More recent methods, including MA-LMM [191] and MemVid [192], further organize memory into more structured and selective representations that facilitate retrieval and reasoning. Overall, recent methods increasingly shift from passively storing long-video history toward selectively organizing and retrieving question-relevant evidence.

Long-video Event Understanding. Long-video event understanding requires organizing long videos at the event level rather than treating them as flat temporal histories. Rather than compressing an entire video into a single representation, earlier work introduced event-centered memory structures. HERMES [193] builds long-form understanding through episodic accumulation and semantic retrieval, while hierarchical event-based memory [194] explicitly models event segmentation together with intra- and inter-event memory. Later methods emphasize persistent event memory through active memory construction and continuous-time consolidation. AMEGO [161] builds a self-contained active memory for very long egocentric videos, and ∞-Video [195] continuously consolidates observations for training-free long-video understanding. Recent work further develops structured episodic memory for modeling event relationships. HippoMM [196] combines episodic and semantic consolidation for long audiovisual event understanding, Video-EM [197] represents key moments as temporally ordered episodic events, and GCAgent [198] models schematic and narrative episodic memory to capture causal and temporal relationships across events. Overall, recent work suggests that organizing memory around semantically meaningful events provides a more effective representation than compressing entire videos into flat histories.

Temporal Grounding. Temporal grounding requires retrieving the video moments relevant to a language query, making effective temporal memory organization essential for locating evidence across long videos. Early efforts primarily improved temporal awareness through timestamp-aware encoding, boundary-aware training, and explicit temporal representations, as demonstrated by TimeChat [102], VTimeLLM [199], and Seq2Time [200]. Later work increasingly organized temporal evidence into more structured representations. Grounded-VideoLLM [201] introduces an additional temporal stream together with discrete temporal tokens to improve grounding, TRACE [202] represents videos as causal event sequences, and Number it [203] improves temporal retrieval through explicit frame indexing. More recent methods increasingly combine structured temporal evidence with reasoning. VideoExpert [204] separates temporal grounding from content generation through specialized temporal and spatial experts, Time-R1 [205] improves grounding via reasoning-oriented post-training, VideoITG [206] adaptively selects relevant frames through instruction-guided temporal grounding, and E.M.Ground [207] further improves grounding through holistic event perception and matching. Overall, recent work increasingly treats temporal grounding as retrieving and reasoning over structured temporal evidence rather than simply localizing timestamps.

5.1.2. Online / Streaming Video Understanding

Unlike offline long-video understanding, streaming video understanding requires models to process continuously arriving video without access to the complete sequence, making memory essential for retaining and retrieving information over time. Early work primarily adapted Video-LLMs to streaming settings through sequential processing and temporally aligned interaction, as demonstrated by Streaming Long Video Understanding with Large Language Models [208] and VideoLLM-online [209]. Subsequent work made explicit memory a central component of online understanding. Flash-VStream [210] maintains memory for real-time long-video understanding, while Streaming Video Understanding and Multi-round Interaction with Memory-enhanced Knowledge [211] supports multi-round interaction through hierarchical memory organization. Later methods increasingly emphasized selective and efficient memory management under limited computational budgets. Streaming

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

VQA with In-context Video KV-Cache Retrieval [212] retrieves relevant historical context from cached representations, StreamMem [213] dynamically manages streaming memory, Memory-efficient Streaming VideoLLMs for Real-time Procedural Video Understanding [214] reduces memory consumption through efficient memory design, and CacheFlow [215] maintains fixed-size caches for long-duration streaming. More recent work further integrates memory with adaptive reasoning and response planning. WAT [216] decouples observation from reasoning to better utilize accumulated memory, while StreamReady [217] jointly learns what information to retain and when to respond during continuous interaction. Overall, recent methods increasingly treat streaming memory as an actively managed resource that selectively retains, retrieves, and reasons over historical observations rather than simply buffering incoming video.

5.1.3. Reasoning over Long Videos

Reasoning over long videos has increasingly emerged as a distinct challenge beyond generic long-video understanding. Rather than only recognizing events or retrieving relevant moments, the objective is to connect temporally dispersed evidence across extended video histories to support multi-step, causal, and narrative reasoning. Recent benchmarks have made this transition more explicit. MoviePuzzle [218] studies visual narrative reasoning through multimodal order learning, while VRBench [219] and VCRBench [220] explicitly evaluate multi-step and causal reasoning over long videos. More recent benchmarks further expand long-video reasoning to larger-scale and more diverse scenarios, including NA-VQA together with its reasoning framework Video-NaRA [221], MINERVA-Cultural [222], and VBVR [223]. Motivated by these increasingly demanding reasoning tasks, recent methods have shifted from storing long-video context toward organizing memory for evidence integration and inference. TemporalVLM [224] builds temporally structured representations for long-range reasoning, while WorldMM [225] and LongVideoAgent [226] organize and retrieve relevant evidence across extended video histories. More recent agentic methods, including A4VL [227] and SAGE [228], further integrate memory with multi-stage planning and reinforcement-trained decision making to support complex long-video reasoning. Overall, recent work increasingly treats memory as an active reasoning substrate that organizes, retrieves, and integrates temporally distributed evidence rather than simply retaining long-video context.

5.1.4. Long Video Generation

Long video generation aims to synthesize extended video sequences while maintaining scene, entity, motion, and camera consistency over long horizons. This requires preserving and reusing historical context beyond locally plausible clip generation. Consequently, memory has become a key mechanism for maintaining coherent visual narratives across long sequences. In scene- and world-consistent generation, GEN3C [229], Context as Memory [230], and WorldWeaver [231] leverage 3D caches, retrieved historical frames, and depth-informed memory banks to reduce structural and temporal drift. For long-form autoregressive generation, Pack and Force Your Memory [232], LongLive [233], Context Forcing [234], and Relax Forcing [235] preserve long-range dependencies through memory packing, context reuse, and KV-memory designs. More recent methods make memory increasingly structured and semantically meaningful: VideoMemory [236] maintains entity-centric memories across shots, while AnchorWeave [237] and MemCam [238] retrieve spatial and camera-aware memories to preserve world consistency. Collectively, these approaches illustrate a transition from reusing recent visual history to maintaining structured representations of scenes, entities, geometry, and camera state that support coherent long-form video generation.

5.1.5. Long Video Prediction

Long video prediction focuses on forecasting future visual observations over extended horizons, with recent work increasingly framing the problem as learning interactive world models. Rather than merely generating plausible future clips, these models must preserve and update an internal representation of the environment under actions, viewpoint changes, and evolving scene dynamics.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

EVA [239] couples reasoning and generation for embodied future anticipation, while Vid2World [240] repurposes pre-trained video diffusion models into action-conditioned interactive world models. As prediction horizons increase, persistent memory becomes essential for maintaining world state over time. Video World Models with Long-term Spatial Memory [241], RELIC [242], and LIVE [243] introduce geometry-grounded spatial memory, compressed camera-aware KV memory, and cycle-consistent long-horizon training to mitigate forgetting and error accumulation. More recent work further expands the scope of memory beyond static scene structure. PAN [244] and Astra [245] advance more general and controllable action-conditioned world simulation, while VerseCrafter [246] and hybrid-memory methods [247] integrate geometric world representations with dynamic subject memory. Together, these developments reflect a shift from maintaining short-term temporal context to explicitly modeling persistent, evolving world states that enable long-horizon interactive prediction.

5.2. Image Reasoning and Captioning

5.2.1. Image Reasoning

Long-chain multimodal reasoning often suffers from visual forgetting, where attention to the visual input gradually diminishes as the reasoning trajectory grows longer, causing models to rely increasingly on previously generated text rather than the image itself [248]. Memory mechanisms address this problem by preserving visual evidence throughout inference. Take-along Visual Conditioning (TVC) maintains access to visual information by dynamically pruning redundant visual tokens while retaining informative ones during reasoning [248]. In multi-image reasoning, CMMCoT [249] caches decoder-layer key-value representations for each image in an external memory bank and retrieves them on demand, enabling models to revisit relevant visual evidence across multiple images. Rather than relying on external memory, VisMem [250] incorporates both short-term perceptual memory and long-term semantic memory within the model to preserve fine-grained visual details while progressively consolidating higher-level concepts during reasoning and generation. Collectively, these approaches demonstrate that memory helps maintain visual grounding over long reasoning chains by preserving perceptual evidence, semantic representations, and cross-image correspondences throughout inference.

5.2.2. Image Captioning

Image captioning benefits from memory by augmenting visual features with semantic information accumulated beyond the current image, including previous training examples and external knowledge sources. Standard attention mechanisms compute representations solely from the input image, limiting their ability to exploit knowledge acquired from previous examples or external resources [251]. The Meshed-Memory Transformer introduces learnable memory vectors into the visual encoder to encode dataset-level prior knowledge that complements region-level visual representations [252]. Moving beyond static dataset-level priors, Prototypical Memory Networks summarize representations from previously observed training samples into compact prototype vectors, allowing the model to retrieve discriminative visual patterns learned from similar examples [251]. Memory can also be maintained externally. Sarto et al. [253] retrieve captions from visually similar images in an external corpus and combine the retrieved text with the input image using a kNN-augmented language model. Overall, memory in image captioning has evolved from learnable global memory toward retrieving instance-level visual and textual knowledge that enriches image understanding and improves caption generation.

5.3. Continual Learning

5.3.1. Incremental Recognition

Incremental recognition requires VLMs to continually expand their recognition space as new categories, domains, and open-vocabulary concepts are introduced, while preserving previously acquired vision–language semantic alignment. Early studies on frozen CLIP suggest that its pretrained vision–language embedding space already exhibits considerable robustness to continual recognition,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

indicating that the embedding space itself functions as an implicit semantic memory [254]. However, continual finetuning can perturb this space and degrade zero-shot transfer. ZSCL [131] therefore shifts the focus from mitigating forgetting of old classes to preserving CLIP's open-vocabulary prior, reducing semantic drift through feature distillation and weight regularization. Subsequent methods move from protecting the original CLIP space toward more controllable incremental retention mechanisms. One line of work combines CLIP's zero-shot semantics with exemplar memory to balance unseen-class generalization and seen-class consolidation [255]. MoE-Adapters [256] organize stage-wise knowledge into callable expert modules to reduce interference across incremental stages, while CLAP4CLIP [257] and RAPF [258] further stabilize incremental adaptation through probabilistic finetuning and representation adjustment with parameter fusion, respectively. More recent methods, including CoLeCLIP [259], LADA [260], and incremental prompt tuning with intrinsic textual anchors [261], further extend memory into scalable representations at the category, open-vocabulary, and semantic-anchoring levels. Overall, the object of memory in incremental recognition evolves from preserving CLIP's pretrained semantic embedding space toward scalable mechanisms that retain and organize category-level, open-vocabulary, and semantic knowledge across incremental stages.

5.3.2. Continual Knowledge Updating

Building on category-level recognition, continual knowledge updating requires VLMs to adapt their pretrained cross-modal representations from continually arriving image–text corpora, evolving data distributions, and newly emerging domains, while retaining prior knowledge, cross-modal alignment, and previously acquired capabilities. Early work primarily viewed memory as a means of preserving CLIP's pretrained capabilities during continual pretraining. VR-LwF [262] maintains zero-shot and image–text matching abilities through replayed vocabulary, while IncCLIP [263] uses generative negative text replay and multimodal knowledge distillation to retain previous pretraining knowledge. Subsequent methods move beyond replaying samples or vocabularies and instead treat the geometry of the cross-modal representation space itself as the primary object of memory. Mod-X [264] preserves off-diagonal information to maintain multimodal alignment on previous domains, whereas CTP [265] combines compatible momentum contrast with topology preservation to retain cross-task embedding structures when absorbing new image–text data. As continual pretraining moves closer to real deployment, memory must also accommodate temporally evolving data streams and accumulated model states. TiC-CLIP [266] and FoMo-in-Flux [267] push the problem toward web-scale temporal data streams and realistic deployment constraints, while TIME [268] provides a model-level memory route through temporal model merging to integrate multimodal expert knowledge accumulated at different stages. Recent methods further preserve prior knowledge without fully storing historical data. GIFT [269] uses synthetic image–text pairs as replay surrogates, while GNSP [270] protects existing representations through gradient null-space projection and cross-modal alignment constraints. Overall, the object of memory evolves from replaying historical data or vocabularies, to preserving cross-modal representation geometry, and ultimately to maintaining temporally accumulated multimodal knowledge under limited access to historical data.

5.3.3. Continual Task Adaptation

Unlike continual updating of pretrained representations, continual task adaptation requires VLMs to acquire new multimodal capabilities as new tasks, interaction patterns, and instruction formats emerge, while retaining previously learned visual reasoning, question-answering strategies, and instruction-following behaviors. CLiMB [135] and VQACL [271] characterize this problem from continual vision–language tasks and VQA skills, respectively, making task interaction patterns a primary object of memory. Later methods transform these interaction patterns into more structured memory. Scene-graph-based symbolic replay [272] converts visual relations and question–answer patterns into scene graph prompts, while QUAD [273] uses questions-only memory to retain question patterns with reduced dependence on stored images. Subsequent methods increasingly parameterize memory as reusable task-adaptation states. TRIPLET [274] preserves modality- and task-specific

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

factors through decoupled prompts, whereas CL-MoE [275] selectively retrieves previous task abilities through expert routing. As VLMs are increasingly adapted through instruction tuning in MLLM–style architectures, the emergence of continual instruction tuning [276] further highlights forgetting in multimodal instruction following, extending the object of memory to instruction-following behavior, visual reasoning, and task-specific parameters. SMoLoRA [277] mitigates the dual forgetting of visual understanding and instruction following through separable routing, while HiDe-LLaVA [278] separates task-private knowledge from cross-task general ability through task-specific expansion and task-general fusion. Overall, the object of memory evolves from replaying previous task experiences to maintaining reusable question structures, prompt and expert states, and instruction-adaptation parameters that support continual acquisition of new multimodal skills.

5.4. Embodied Navigation and Action

5.4.1. Vision-Language Navigation

A. Instruction-following and Spatial Memory Navigation

Instruction-following Vision-Language Navigation (VLN) with spatial memory requires agents to ground language instructions in accumulated observations and spatial context under partial observability, rather than relying solely on the current visual view. Early approaches introduced lightweight navigation memory through progress estimation and trajectory tracking. Self-Monitoring [279] estimates navigation progress to align the agent's internal state with instruction execution, while Regretful Agent [280] exploits progress signals for backtracking and path correction. Subsequent methods move beyond compact state summaries toward explicit history modeling: VLN-BERT [281] and HAMT [282] incorporate past observations, actions, and instruction context to improve sequential decision making. As navigation tasks become more challenging, agents move beyond storing observation histories and begin to organize memory into spatial representations, such as topological graphs and environment maps. DUET [283] and BEVBert [284] organize explored trajectories as topological or map-based representations, while GridMM [285] maintains a dynamically expanding grid memory to preserve spatial relationships and instruction-relevant information. Recent studies further integrate memory with high-level reasoning and compact representation. MC-GPT [286] constructs a memory-based topological map over viewpoints, objects, and spatial relations, whereas JanusVLN [287] separates geometric and semantic knowledge through dual implicit memory representations. Overall, instruction-following VLN memory has evolved from trajectory-level progress tracking toward structured, semantics-aware spatial representations that support planning and reasoning.

B. Dialogue-based Navigation

Dialogue-based navigation extends VLN memory beyond instruction execution by requiring agents to maintain conversational context, user intent, environmental observations, and spatial state across multiple interaction rounds. Talk the Walk [288] and CVDN [289] formulate navigation as grounded dialogue, where conversation history assists localization, clarification, and action selection. CMN [290] explicitly separates language and visual memory to associate current dialogue with previous utterances, observations, and actions, while RMM [291] further models collaborator state by reasoning about the knowledge shared between the guide and navigator. More realistic collaborative settings emphasize the alignment among linguistic descriptions, environmental semantics, and agent location. RobotSlang [292] and DRAGON [293] model correspondences between dialogue and embodied states, while DialNav [294] considers remote collaborative navigation where the guide must infer the navigator's location through dialogue alone. Overall, memory in dialogue-based navigation has progressed from storing conversation history to maintaining cross-modal interaction states that connect dialogue, perception, and spatial reasoning.

C. Long-horizon and Persistent Navigation

Long-horizon and persistent VLN focuses on accumulating and reusing knowledge across extended routes, multi-stage instructions, and repeated interactions with the same or related environments. Early work explored future-aware and map-based memory mechanisms: Look Before You Leap [295] predicts future navigation states to support planning, Chasing Ghosts [296] formu-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

lates instruction following as Bayesian state estimation with an online semantic spatial map, and Talk2Nav [297] uses spatial memory to associate landmarks with directional transitions in large-scale urban navigation. Memory subsequently became important for managing long and compositional instructions. M-TRACK [298] tracks intermediate milestones to support subgoal completion and reduce navigation failures caused by missed targets or premature termination. Beyond individual episodes, persistent navigation methods focus on reusable environmental knowledge. Structured Scene Memory [299] organizes historical observations into persistent scene representations, IVLN [300] and ESceme [301] retrieve prior experiences in revisited environments, and OVER-NAV [302] leverages open-vocabulary perception with structured representations to integrate visual, semantic, and spatial information over long tours. More recently, MSNav [303] combines dynamic memory management with LLM-based spatial reasoning, using selective updates to reduce memory redundancy in long-horizon tasks. Overall, persistent VLN memory extends beyond episode-specific state tracking toward reusable environmental knowledge that supports navigation across stages, routes, and experiences.

5.4.2. Vision-Language Action

Memory in Vision-Language Action (VLA) models enables robots to perform long-horizon tasks in partially observable and dynamic environments, where the appropriate action depends not only on the current observation but also on previous interactions, task progress, and execution feedback. Unlike short-horizon control, real-world manipulation is inherently non-Markovian: identical visual inputs may require different actions depending on prior actions, hidden states, object changes, or completed subtasks. Early episodic memory approaches for robotic manipulation demonstrate the importance of retaining state transitions, action trajectories, and subtask structures from demonstrations [304]. Building on this idea, subsequent VLA methods incorporate interaction histories to monitor task progress, recover from failures, and improve action generation.

Recent approaches increasingly transform raw interaction histories into structured perceptual and spatial memories. SAM2Act+ [28] introduces a memory bank with attention mechanisms to enhance spatial reasoning, while MemoryVLA [305] combines working memory with perceptual-cognitive memory to preserve both visual details and high-level task semantics. Beyond storing past observations, memory also enables experience retrieval and reuse across tasks. MAP-VLA [306] retrieves stage-specific memory prompts from previous demonstrations to guide action prediction, whereas EchoVLA [307] integrates scene memory with episodic memory for mobile manipulation by maintaining spatial-semantic representations and task-level experiences.

More recent VLA models focus on scalable and temporally consistent memory mechanisms for extended real-world deployment. OptimusVLA [308] separates memory into Global Prior Memory and Local Consistency Memory to balance reusable knowledge with current-task adaptation. MEM [309] introduces complementary short-term visual memory and long-term textual memory for sustained interaction, while ReMem-VLA [310] maintains multi-scale temporal context through frame-level and chunk-level recurrent memory queries. Overall, memory in VLA systems has evolved from storing demonstrations and interaction histories toward structured perceptual-spatial memory, experience-based retrieval, and multi-scale recurrent mechanisms that support robust long-horizon robot behavior.

5.4.3. Autonomous Driving

Autonomous driving requires memory mechanisms that support continuous perception, reasoning, and decision making in dynamic and safety-critical environments. Unlike embodied navigation and manipulation, where memory primarily helps maintain spatial and task context, autonomous driving must track rapidly changing traffic conditions, preserve interactions with surrounding agents, and incorporate prior experience to handle rare or long-tail scenarios. Since driving decisions often depend on events that are not fully observable from the current sensory input, memory provides a mechanism for integrating historical observations, contextual knowledge, and previously encountered situations.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

Recent vision-language driving approaches explore memory as a bridge between perception and high-level reasoning. VLP [311] incorporates language-model-based reasoning into an end-to-end driving planner, where a query-based transformer decoder selectively accesses memory representations containing task-relevant contextual information. By combining retrieved memory with the current scene understanding, VLP improves generalization to uncommon scenarios beyond those directly observed during training. Beyond implicit contextual memory, recent methods introduce explicit experience-based memory for scenario retrieval and decision support. LeapVAD [312] adopts a dual-process architecture consisting of a heuristic process and an analytic process. Its reflection mechanism stores previous driving decisions and outcomes in a memory bank, allowing the analytic process to accumulate scenario knowledge while the heuristic process retrieves relevant experiences for efficient decision making. Similarly, MTRDrive [313] proposes a memory-tool synergy reasoning framework, where a VLM interacts with an external memory bank and an active toolkit engine to gather complementary information. Retrieved experiences and tool-generated evidence are then integrated through reasoning to produce driving decisions.

Overall, memory in autonomous driving has evolved from maintaining temporal scene context toward explicit experience-based and reasoning-oriented memory systems that enable adaptive planning, long-tail scenario handling, and safer decision making in complex environments.

5.5. Multimodal RAG

5.5.1. Knowledge-Augmented Visual QA

Knowledge-augmented visual QA requires models to retrieve external knowledge beyond the image content, since many questions depend on named entities, commonsense relations, world knowledge, or fine-grained factual information. Benchmarks such as KVQA [314], OK-VQA [315], A-OKVQA [316], and InfoSeek [317] collectively demonstrate that visual evidence alone is often insufficient, motivating external knowledge as a retrievable memory that complements visual perception. Early retrieval-augmented approaches organize structured knowledge bases or textual corpora as searchable external memory for answer generation [318]. Subsequent work improves the alignment between image-question representations and retrieved knowledge by enabling fine-grained retrieval over visual entities and semantic relations [319]. More recently, retrieval has become tightly integrated with multimodal generation. ReAuSE [320] enables MLLMs to autoregressively generate document identifiers before reasoning over the retrieved evidence, coupling retrieval directly with generation. MI-RAG [321] further introduces iterative multimodal retrieval, maintaining an evolving reasoning memory that progressively refines retrieval queries and integrates evidence from heterogeneous knowledge sources. Overall, memory in knowledge-augmented VQA has evolved from static external knowledge repositories to retrieval memory that is increasingly entity-aware, tightly coupled with multimodal generation, and iteratively updated throughout the reasoning process.

5.5.2. Multimodal Document Retrieval and QA

Multimodal document retrieval and QA require models to retrieve and integrate evidence from visually rich documents, where critical information may reside not only in text but also in layouts, tables, charts, figures, and page-level visual structures. Early work on multimodal document understanding, including LayoutLM [322], DocVQA [323], InfographicVQA [324], and ChartQA [325], demonstrates that document representations must preserve both textual semantics and visual layout. Hierarchical document modeling [326] further highlights the need to aggregate evidence across multiple pages as documents become longer and more complex. Building upon these representations, recent multimodal RAG methods organize document collections as retrievable visual memory. ColPali [327] learns multi-vector embeddings directly from page images, enabling retrieval without relying on OCR pipelines. M3DocRAG [328] extends this idea to multi-page and multi-document settings by jointly retrieving relevant visual pages from document memory and generating grounded responses. VDocRAG [329] further unifies PDFs, presentation slides, and other visually rich documents as image-based retrieval units, reducing information loss introduced by text parsing. MMDocRAG [153] systematically eval-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

uates cross-modal evidence retrieval and grounded generation across long, multi-page documents, highlighting visual grounding as a central challenge. Overall, memory in multimodal document RAG has evolved from layout-aware document representations to retrievable visual page memory that enables evidence aggregation across pages, documents, and modalities.

5.5.3. Domain-Specific and Scenario-Grounded Multimodal RAG

Domain-specific and scenario-grounded multimodal RAG extends retrieval memory to specialized environments where reasoning depends on domain knowledge, scenario experience, or structured contextual evidence. In medical vision-language tasks, datasets such as VQA-RAD [330], SLAKE [331], and PMC-VQA [332] demonstrate that accurate reasoning requires integrating medical terminology, structured clinical knowledge, and visual evidence. MED-VRAG [333] further advances medical RAG by retrieving document page images rather than isolated text passages, iteratively accumulating retrieved clinical evidence within a memory bank to support diagnosis-oriented reasoning. In autonomous driving, DriveLM [334] formulates driving reasoning across perception, prediction, and planning, highlighting the need to retrieve and reuse structured scenario memory. RAG-Driver [335] retrieves similar driving cases as external memory for in-context learning, improving both interpretability and generalization across traffic scenarios, while RealGen [336] retrieves real driving scenes to guide controllable traffic scene generation, extending retrieval memory beyond question answering to multimodal synthesis. Beyond medicine and autonomous driving, similar retrieval paradigms have also emerged in other specialized domains. RS-RAG [337] organizes multimodal geographic knowledge as retrievable memory for remote sensing reasoning, whereas Video-RAG [338] constructs event-level retrieval memory from multimodal auxiliary cues, including audio transcripts, OCR, and object detections, to support long-video understanding. Overall, multimodal RAG in domain-specific settings is evolving from generic external knowledge retrieval toward scenario-grounded retrieval memory that organizes and reuses clinical evidence, driving experience, geographic knowledge, and temporally structured video events for grounded multimodal reasoning.

5.6. Egocentric Vision

5.6.1. Episodic Retrieval

First-person videos provide a natural setting for studying episodic memory, requiring models to recall where an object was previously observed, when an interaction occurred, or what happened earlier in a long stream of experience. The Ego4D dataset introduces an episodic memory benchmark that requires localizing visual evidence answering a natural-language query within a user's past video [160]. Moving beyond evaluation, AMEGO [161] constructs an online memory of object-centric interaction events using structured representations without explicit semantic labels, enabling efficient retrieval from long egocentric videos. Embodied VideoAgent [339] further extends episodic memory beyond retrieval by maintaining a persistent scene memory that is continuously updated through observed object interactions and subsequently used for reasoning and long-horizon planning in embodied environments.

5.6.2. Wearable Assistants

Wearable assistants require long-term memory that retains previous observations and retrieves relevant experiences in real time. EgoLife builds such an assistant, EgoButler, around an explicit two-stage memory architecture: EgoGPT continuously converts multimodal observations into structured memory entries, while EgoRAG organizes them into hierarchical hourly and daily summaries and retrieves time-stamped evidence for answering questions about events that occurred days earlier [166]. The benchmark highlights remaining challenges, including robust identity recognition and retrieval over extended temporal horizons. Complementing this direction, TeleEgo evaluates egocentric assistants on continuous video streams under real-time constraints, explicitly assessing long-term memory retention as a core capability alongside response quality [165].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

5.7. Personalized Assistants

5.7.1. Concept Personalization

Personalized assistants require VLMs to acquire user-specific concepts and reason about their visual appearance, textual descriptions, and semantic relationships. Early approaches encode personalized knowledge directly into model parameters. MyVLM [171] augments a VLM with concept-specific classification heads and learned embeddings for personalized concepts, while Yo'LLaVA [172] learns latent representations of personalized subjects from only a few example images. Although effective, these methods require additional optimization whenever new concepts are introduced, making continual personalization increasingly expensive as the number of users and personalized concepts grows.

Recent methods instead externalize personalized knowledge into explicit memory. Retrieval-Augmented Personalization (RAP) [340] stores personalized concepts in an external key–value memory, allowing user-specific knowledge to be updated by modifying the memory rather than retraining the underlying VLM. Online-PVLM [341] further treats personalization as dynamic memory, generating concept embeddings on the fly at inference time without per-instance optimization and continuously storing and retrieving user-specific visual features across interactions. By decoupling personalized concepts from model parameters, these memory-augmented approaches improve scalability to large numbers of users and continually expanding concept vocabularies.

5.7.2. Conversational Memory

Recognizing personalized concepts alone is insufficient for long-term assistance; an assistant must also remember what has been said and observed throughout its interactions with the user. The challenge is that conversational history is open-ended and inherently multimodal. Relevant information may have appeared weeks earlier, in a different conversation session, or within an image or video, making it impractical to retain the entire history in the context window. Consequently, recent systems increasingly maintain explicit long-term memory that stores historical interactions separately from the language model and retrieves only the most relevant information at inference time. M2A [342] combines an immutable message log with a higher-level semantic memory that is updated online, while M3-Agent [343] extends this idea to streaming audiovisual observations by organizing them into an entity-centric memory graph for retrieval and reasoning. PersonaVLM [175] similarly focuses on maintaining long-term multimodal personal context within a personalized vision-language assistant. Evaluating such capabilities remains challenging because conventional short-context benchmarks rarely assess retrieval over extended interaction histories. Mem-Gallery [179] addresses this gap by directly measuring multimodal long-term conversational memory, evaluating whether assistants can accurately recall information introduced many dialogue turns earlier.

5.8. Virtual-Environment Interaction

5.8.1. Open-World Agents

In games and open-world environments, agents perceive the environment through visual observations and human instructions while performing tasks over extended horizons. A key challenge is enabling agents to leverage information beyond the current observation by reflecting on previous interactions. Memory can operate at multiple levels: within a single trial, recent observations and actions provide a working context for immediate decision-making; across trials, stored experiences can be retrieved to guide future planning and improve task execution. In addition to experiential memory, agents can incorporate external knowledge sources, such as crafting rules and world descriptions from the Minecraft wiki through API-based access [344].

JARVIS-1 incorporates a multimodal memory module that stores successful interaction experiences by associating task scenarios with corresponding action plans. During inference, the agent retrieves relevant experiences as in-context examples to support reasoning and planning, improving decision consistency and performance over long-horizon tasks [345]. Optimus-1 [346] further intro-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

duces a Hybrid Multimodal Memory module that summarizes multimodal observations collected during execution into a compact experience pool for efficient storage and retrieval. Beyond episodic experiences, it represents important environmental knowledge, such as crafting requirements, using a hierarchical knowledge graph, enabling agents to reason based on both prior interactions and structured knowledge of the virtual world.

5.8.2. GUI Agents

A. Web Agents

Web agents interact with complex websites by interpreting visual interfaces and executing sequences of actions. WebVoyager [347] uses trajectory-based context as a form of working memory, maintaining previous actions and annotated screenshots to support ongoing task completion. However, such within-session context does not provide persistent memory across tasks, limiting the ability of agents to reuse previous experiences or avoid recurring failures.

ICAL [348] addresses this limitation by transforming episodic interaction histories into reusable semantic knowledge. Rather than directly replaying previous trajectories, it uses a VLM to extract higher-level information, including causal relationships, object state transitions, and temporal subgoals, and stores these abstractions for retrieval in future tasks. This enables cross-task knowledge transfer and improves performance on visually grounded web benchmarks such as VisualWebArena [349] and VideoWebArena [350].

B. Mobile Agents

Mobile agents operate in dynamic environments involving multiple applications and long sequences of UI interactions. Mobile-Agent-v2 introduces a memory unit that maintains task-relevant historical information from previous screenshots. During execution, the decision-making agent updates this memory with newly identified focus content, allowing it to preserve important context across interactions. This historical information is particularly useful for multi-step tasks involving multiple applications, where remembering the locations and states of relevant UI elements improves action localization and execution reliability [351].

To support longer-horizon mobile tasks, Mobile-Agent-E [352] introduces a persistent long-term memory composed of Tips and Shortcuts. Tips capture generalizable lessons from previous tasks, while Shortcuts represent reusable sequences of atomic operations for recurring subtasks. These memory components allow the agent to improve interaction efficiency by reusing previously acquired procedural knowledge. The design is motivated by the distinction between episodic experiences and procedural knowledge in cognitive science.

C. Computer Agents

Computer-use agents extend Graphical User Interface (GUI) interaction beyond specific applications by enabling autonomous control of general-purpose computer environments through keyboard and mouse actions. Agent S [353] introduces a self-supervised memory mechanism that continually updates from interaction experiences. Its narrative memory stores abstracted task knowledge that provides contextual information for future planning, while episodic memory preserves detailed task trajectories that can support later refinement and improvement of agent behavior.

Similarly, UI-TARS [354] is a native computer-use agent that directly interprets screenshots and generates actions through keyboard and mouse control. Through iterative training and reflection, it leverages accumulated experience to improve decision-making and adapt to new tasks. Together, these approaches demonstrate how integrating long-term knowledge and persistent memory can transform GUI agents from reactive systems into agents capable of accumulating and reusing experiences.

6. Challenges & Limitations

Memory extends VLMs beyond isolated image–text inference by enabling models to accumulate, organize, and reuse multimodal experiences. However, current memory-augmented VLMs remain limited by challenges that span four fundamental dimensions: temporal scope, storage location, memory entity, and memory operation. These challenges become particularly critical in long videos,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

retrieval-augmented generation, embodied environments, and personalized assistants, where memory must be scalable, grounded, editable, secure, and reliably evaluated.

Scalable Memory for Long-horizon Multimodal Experiences. Long-context VLMs must process increasingly long streams of visual observations, language interactions, audio signals, and intermediate reasoning states. Recent systems, including Long Context Transfer, LongVILA, Long-VITA, and Video-XL, demonstrate substantial progress toward extending active context windows and processing hour-scale videos [32–35]. However, increasing context length alone does not guarantee effective memory. A model may accept more tokens while still failing to retrieve a brief but critical event, identify its temporal location, or use it efficiently under realistic latency and storage constraints. This challenge becomes more pronounced in streaming settings, where the memory horizon is potentially unbounded and the model must continuously decide what information should remain accessible [355]. The central problem is therefore not simply enlarging context windows, but developing mechanisms for selective retention, indexing, and access over long multimodal histories.

Capacity–efficiency Tradeoffs. Memory capacity inevitably introduces computational and storage costs. KV-cache based approaches make this tension explicit: methods such as LOOK-M, MadaKV, and AKVQ-VL reduce inference memory through token pruning, adaptive eviction, or quantization, but these strategies rely on assumptions about which visual or linguistic states can be safely discarded [38, 39,85]. Full-stack systems such as LongVILA further suggest that effective long-context memory requires joint optimization across model architectures, training strategies, and inference systems [34]. A fundamental challenge remains that memory importance is often query-dependent: information that appears redundant during storage may later become essential for answering questions about a specific object, event, or temporal relationship.

Faithful Compression and Temporal Preservation. Because storing every observation is impractical, memory systems must compress multimodal experiences while preserving future utility. Approaches such as MovieChat, LongVU, and VideoChat-Flash improve scalability through sparse, adaptive, or hierarchical video memory mechanisms [110–112]. However, compression decisions are inherently difficult because future queries are unknown at storage time. Existing approaches often preserve high-level semantics while losing fine-grained temporal structure, including event order, duration, and visual details. Although TimeChat and VideoLLaMB introduce mechanisms for temporal awareness and recurrent context modeling [102,106], maintaining both semantic abstraction and precise temporal grounding under limited memory remains unresolved.

Reliable Retrieval and Trustworthy External Memory. External memory expands the knowledge available to VLMs but introduces new failure modes in retrieval, ranking, and evidence integration. RAG established the retrieval–generation paradigm, while MuRAG and REVEAL extended this framework to multimodal knowledge sources [61,62,86]. More recent systems, including SearchLVLMs and VisRAG, incorporate web-scale retrieval and visually grounded document memory [96,99]. Nevertheless, retrieved evidence may be irrelevant, outdated, incomplete, or inconsistent with the current visual context. Unlike traditional databases, VLMs may generate plausible responses even when retrieved memory is incorrect, making retrieval reliability and evidence verification critical challenges. MRAG-Bench highlights this issue by evaluating whether retrieved visual evidence genuinely improves multimodal reasoning [87].

Grounding, Provenance, and Semantic Alignment. As memory representations become increasingly abstract, maintaining a clear connection between stored information and original observations becomes challenging. Latent memory slots, compressed summaries, and user profiles improve efficiency but reduce interpretability and traceability. Knowledge-grounded approaches such as KRISP and REVEAL demonstrate the benefit of explicit evidence connections [86,89], while VisRAG preserves visual document representations to avoid excessive loss during textual conversion [99]. Meanwhile, benchmarks such as LongVideoBench, Video-MME, and MMLongBench show that models may fail to associate answers with the correct visual evidence even when relevant information exists in mem-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

ory [41,42,356]. Future memory systems require not only compact storage and accurate retrieval, but also provenance tracking, evidence attribution, and uncertainty estimation.

Memory Updating, Forgetting, and Privacy. Persistent memory requires models to continuously incorporate new information while avoiding harmful accumulation of outdated or incorrect knowledge. Multimodal editing and continual learning studies identify updating and forgetting as fundamental challenges rather than secondary concerns [82,108,109]. These issues become more sensitive in personalized assistants, where long-term memory may contain user preferences, social interactions, and private experiences. Systems such as MemoryBank, personalized multimodal LLM frameworks, and AUGUSTUS demonstrate the value of persistent user memory [10,64,357]. However, persistent memory also creates risks including unauthorized retention, adversarial memory injection, profile drift, and unintended personalization. Designing memory systems therefore requires mechanisms for controllable updating, selective forgetting, user consent, and memory integrity verification.

Comprehensive Evaluation of Memory Behavior. Evaluation of multimodal memory remains fragmented across individual capabilities. Needle-style benchmarks examine hidden information retrieval, long-video benchmarks measure temporal understanding, and multimodal RAG benchmarks evaluate evidence retrieval and integration [41,42,44,87]. MMLongBench and OVO-Bench further investigate long-context reasoning and online video understanding [53,356]. However, real-world memory failures often involve multiple stages: incorrect storage, excessive compression, unreliable retrieval, and improper updating or forgetting. Future benchmarks should therefore evaluate the complete memory lifecycle across multiple sessions, modalities, environments, and evolving user objectives.

7. Future Directions

Future progress in memory-augmented VLMs will require more than longer context windows or larger retrieval stores. Instead, memory should be viewed as a controllable lifecycle that determines what information is acquired, how it is represented, when it is reused, how it is revised, and when it should be forgotten. Future memory-native VLMs will likely require coordinated designs across memory formation, representation, retrieval, consolidation, update, and governance, rather than treating memory as an external module attached to a pretrained model.

Cognitive-inspired Memory Architectures. Future VLMs may benefit from cognitively inspired but engineering-oriented memory hierarchies that separate transient perception, episodic experience, semantic abstraction, and reflective reasoning. Generative Agents demonstrated how observations and reflections can be organized into a persistent memory stream for later behavior [63], while MemoryBank converted interaction histories into long-term conversational memory [10]. KARMA introduced explicit short- and long-term memory modules for embodied agents [68], and brain-inspired multi-memory systems suggest that lifelong embodied intelligence may require several interacting memory stores [358]. Future VLMs could move beyond flat context accumulation by maintaining fast perceptual buffers, episodic records, semantic summaries, and reflective memory that are continuously consolidated over time. An important open question is how raw multimodal experiences should be transformed into reusable knowledge without losing critical evidence.

Hybrid and Adaptive Memory Substrates. Future VLMs will likely require hybrid memory architectures that dynamically combine multiple substrates rather than relying on a single storage mechanism. Raw observations preserve visual evidence, episodic memory captures event structure, semantic memory enables abstraction, latent memory provides computational efficiency, and external retrieval expands knowledge beyond model parameters. RAG and MuRAG established retrieval-based augmentation for textual and multimodal knowledge [61,62], while Reveal and retrieval-augmented multimodal language modeling demonstrated how multimodal knowledge memory can support reasoning [86,91]. VisRAG further showed that some information is better preserved as visual documents rather than converted into text-only representations [99]. Future systems should learn how to

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

route information among different memory substrates while maintaining provenance, compression efficiency, and retrieval reliability.

Learned Memory Formation and Consolidation. Most existing VLM memory systems rely on manually designed rules for selecting, summarizing, or caching information. However, deciding what deserves to become memory is itself a fundamental intelligence problem. Future models should learn memory formation policies that determine when to write new information, what level of abstraction to preserve, and how repeated experiences should be consolidated into reusable knowledge. Such mechanisms could enable lifelong adaptation while preventing uncontrolled memory growth. Similar to human consolidation, future VLMs may need mechanisms that periodically reorganize accumulated experiences, resolve redundancy, and transform episodic memories into more general semantic knowledge.

Task-adaptive Memory Lifecycle Management. Memory policies should adapt to downstream objectives rather than rely on fixed compression ratios, static retrieval thresholds, or predefined eviction rules. VQA may require fine-grained object details, long-video understanding may require event-level summaries, navigation may require persistent spatial representations, and personalization may require stable yet editable user profiles. Model editing, continual VLM learning, and knowledge maintenance have begun to treat memory update as a central operation [82,108,109]. LOOK-M and MadaKV further demonstrated that online inference memory can be revised through retention and replacement strategies [38,39]. Future VLMs require task-adaptive controllers that jointly determine what to write, compress, retrieve, revise, and forget. Since long-lived memory inevitably accumulates outdated or conflicting information, future systems must also learn when to trust, reconcile, or discard existing memories.

Grounded and Provenance-aware Memory. Future memory representations should preserve links to their original evidence rather than storing only compressed vectors or summaries. A memory entry should ideally include its source image, video segment, dialogue history, timestamp, confidence, and retrieval trajectory. KRISP and Reveal demonstrated the value of explicit evidence for knowledge-grounded reasoning [86,89], while VisRAG preserved document pages as visual evidence during retrieval [99]. LongVideoBench, Video-MME, and MMLongBench further indicate that answer accuracy alone is insufficient when models cannot identify the supporting visual evidence [41,42,356]. Provenance-aware memory could therefore improve not only grounding but also hallucination diagnosis, error analysis, and user trust.

Efficient Long-context Memory Systems. Efficiency should be incorporated into memory mechanisms from the beginning rather than treated as a post-processing optimization. LongVILA demonstrated the importance of model–system co-design for long-context VLMs [34], while LongVU and VideoChat-Flash explored complementary strategies for long-video compression and efficient reasoning [111,112]. LOOK-M, MadaKV, and AKVQ-VL showed that KV-cache pruning, eviction, and quantization can substantially influence multimodal inference efficiency [38,39,85]. Future architectures should move beyond simply extending context length by combining video-token compression before reasoning, hierarchical memory retrieval during inference, adaptive cache management during decoding, and streaming memory updates as new observations arrive.

Unified Cross-modal and Cross-task Memory Frameworks. Future memory systems should become reusable across modalities and tasks rather than being redesigned separately for images, videos, language, navigation, and robotics. Perceiver and Perceiver IO introduced general latent representations for high-dimensional inputs [73,77], while Flamingo and BLIP-2 demonstrated how compact visual representations can interface with large language models [31,74]. PaLM-E extended this interface toward embodied multimodal reasoning [359], and RoboOS-NeXT suggested shared memory infrastructures for long-lived collaborative agents [70]. A common memory interface across vision, language, audio, action, and personalization could enable knowledge transfer across tasks while preserving modality-specific grounding and update mechanisms.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

Memory-centric Evaluation and Governance. Future benchmarks should evaluate complete memory lifecycles rather than isolated retrieval or generation accuracy. Needle-in-a-haystack evaluations expose long-context retrieval failures, MRAG-Bench targets multimodal retrieval-augmented generation, OVO-Bench evaluates online video understanding, and MMLongBench expands long-context VLM evaluation [44,53,87,356]. Future benchmarks should jointly evaluate memory writing, compression, retrieval, updating, forgetting, provenance tracking, and cross-session reuse. Beyond accuracy, evaluation should measure whether users can inspect, correct, delete, and control persistent memories. This is particularly important for personalized multimodal assistants and contextualized user-memory systems [64,357]. Security mechanisms against malicious memory injection, stale information propagation, and unintended retention should become fundamental components of memory governance rather than deployment-time patches.

8. Conclusion

In this survey, we presented a systematic study of memory mechanisms in VLMs and introduced a unified 4D taxonomy that characterizes memory along four fundamental dimensions: **when** memory is formed and used, **where** it is stored, **what** information it represents, and **how** it is accessed and updated. Building upon this taxonomy, we reviewed representative memory architectures, learning strategies, evaluation benchmarks, and applications across multimodal understanding, long-video reasoning, retrieval-augmented generation, embodied intelligence, and personalized assistants.

Despite rapid advances, current memory-augmented VLMs remain far from achieving reliable lifelong multimodal intelligence. Key challenges include scalable memory management, adaptive memory formation and consolidation, efficient long-context reasoning, multimodal grounding and provenance, continual updating, and trustworthy evaluation. We envision that future VLMs will evolve from systems with extended context or external retrieval modules into memory-centric architectures capable of selectively acquiring, organizing, revising, and reusing multimodal experiences over time. Such adaptive and unified memory frameworks will be critical for building intelligent agents that can reason persistently, interact naturally, and continuously improve from experience.

To facilitate future research, we accompany this survey with a continuously maintained public repository of representative papers, benchmarks, and resources. We hope this survey provides a structured perspective on the emerging field of VLM memory and serves as a foundation for developing next-generation multimodal systems with persistent, grounded, and trustworthy intelligence.

References

1. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *National Science Review* **2024**, *11*, nwae403.
2. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)* **2019**, *51*, 1–36.
3. Wu, Q.; Teney, D.; Wang, P.; Shen, C.; Dick, A.; van den Hengel, A. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding* **2017**, *163*, 21–40.
4. Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology* **2025**.
5. Liu, G.; Wang, S.; Yu, J.; Yin, J. A survey on multimodal dialogue systems: recent advances and new frontiers. In Proceedings of the 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE). IEEE, 2022, pp. 845–853.
6. Liu, H.; Guo, D.; Cangelosi, A. Embodied intelligence: A synergy of morphology, action, perception and learning. *ACM Computing Surveys* **2025**, *57*, 1–36.
7. Wu, Y.; Rabe, M.N.; Hutchins, D.; Szegedy, C. Memorizing transformers. *arXiv preprint arXiv:2203.08913* **2022**.
8. Rae, J.W.; Potapenko, A.; Jayakumar, S.M.; Lillicrap, T.P. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507* **2019**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

9. Bulatov, A.; Kuratov, Y.; Burtsev, M. Recurrent memory transformer. *Advances in Neural Information Processing Systems* **2022**, *35*, 11079–11091.
10. Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; Wang, Y. Memorybank: Enhancing large language models with long-term memory. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2024, Vol. 38, pp. 19724–19731.
11. Graves, A.; Wayne, G.; Reynolds, M.; Harley, T.; Danihelka, I.; Grabska-Barwińska, A.; Colmenarejo, S.G.; Grefenstette, E.; Ramalho, T.; Agapiou, J.; et al. Hybrid computing using a neural network with dynamic external memory. *Nature* **2016**, *538*, 471–476.
12. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **2020**, *33*, 9459–9474.
13. Sumers, T.R.; Yao, S.; Narasimhan, K.; Griffiths, T.L. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427* **2023**.
14. Xie, J.; Chen, Z.; Zhang, R.; Wan, X.; Li, G. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116* **2024**.
15. Zhang, Z.; Dai, Q.; Bo, X.; Ma, C.; Li, R.; Chen, X.; Zhu, J.; Dong, Z.; Wen, J.R. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems* **2025**, *43*, 1–47.
16. Ding, J.; Zhang, Y.; Shang, Y.; Zhang, Y.; Zong, Z.; Feng, J.; Yuan, Y.; Su, H.; Li, N.; Sukiennik, N.; et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys* **2025**, *58*, 1–38.
17. Du, Y.; Huang, W.; Zheng, D.; Wang, Z.; Montella, S.; Lapata, M.; Wong, K.F.; Pan, J.Z. Rethinking memory in ai: Taxonomy, operations, topics, and future directions. *arXiv e-prints* **2025**, pp. arXiv–2505.
18. Hu, Y.; Liu, S.; Yue, Y.; Zhang, G.; Liu, B.; Zhu, F.; Lin, J.; Guo, H.; Dou, S.; Xi, Z.; et al. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564* **2025**.
19. Du, P. Memory for autonomous llm agents: Mechanisms, evaluation, and emerging frontiers. *arXiv preprint arXiv:2603.07670* **2026**.
20. Luo, J.; Tian, Y.; Cao, C.; Luo, Z.; Lin, H.; Li, K.; Kong, C.; Yang, R.; Ma, J. From Storage to Experience: A Survey on the Evolution of LLM Agent Memory Mechanisms, 2026, [arXiv:cs.AI/2605.06716].
21. Shi, H.; Xu, Z.; Wang, H.; Qin, W.; Wang, W.; Wang, Y.; Wang, Z.; Ebrahimi, S.; Wang, H. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys* **2025**, *58*, 1–42.
22. Jia, Z.; Li, J.; Kang, Y.; Wang, Y.; Wu, T.; Wang, Q.; Wang, X.; Zhang, S.; Shen, J.; Li, Q.; et al. The AI Hippocampus: How Far are We From Human Memory? *arXiv preprint arXiv:2601.09113* **2026**.
23. Maharana, A.; Lee, D.H.; Tulyakov, S.; Bansal, M.; Barbieri, F.; Fang, Y. Evaluating very long-term conversational memory of llm agents. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 13851–13870.
24. Wu, D.; Wang, H.; Yu, W.; Zhang, Y.; Chang, K.W.; Yu, D. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813* **2024**.
25. Zhang, L.; Hao, X.; Xu, Q.; Zhang, Q.; Zhang, X.; Wang, P.; Zhang, J.; Wang, Z.; Zhang, S.; Xu, R. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 13032–13056.
26. Gao, X.; Hu, C.; Chen, B.; Li, T. Chain-of-memory: Enhancing gui agents for cross-application navigation. *arXiv preprint arXiv:2506.18158* **2025**.
27. Wu, W.; Song, Z.; Zhou, K.; Shao, Y.; Hu, Z.; Huang, B. Towards general continuous memory for vision-language models. *Advances in Neural Information Processing Systems* **2026**, *38*, 128685–128710.
28. Fang, H.; Grotz, M.; Pumacay, W.; Wang, Y.R.; Fox, D.; Krishna, R.; Duan, J. Sam2act: Integrating visual foundation model with a memory architecture for robotic manipulation. *arXiv preprint arXiv:2501.18564* **2025**.
29. Sun et al.. VideoBERT: A Joint Model for Video and Language Representation Learning, 2019.
30. Lei et al.. Less Is More: ClipBERT for Video-and-Language Learning via Sparse Sampling, 2021.
31. Alayrac et al.. Flamingo: a Visual Language Model for Few-Shot Learning, 2022.
32. Zhang et al.. Long Context Transfer from Language to Vision, 2025.
33. Shen et al.. Long-VITA: Scaling Large Multi-modal Models to 1 Million Tokens with Leading Short-Context Accuracy, 2025.
34. Chen et al.. LongVILA: Scaling Long-Context Visual Language Models for Long Videos, 2025.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

35. Shu et al.. Video-XL: Extra-Long Vision Language Model for Hour-Scale Video Understanding, 2025.
36. He et al.. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding, 2024.
37. Ranasinghe et al.. Understanding Long Videos with Multimodal Language Models, 2025.
38. Wan et al.. LOOK-M: Look-Once Optimization in KV Cache for Efficient Multimodal Long-Context Inference, 2024.
39. Li et al.. MadaKV: Adaptive Modality Perception KV Cache Eviction for Efficient Multimodal Long-Context Understanding, 2025.
40. Liu et al.. StreamChat: Chatting with Streaming Video, 2024.
41. Wu, H.; Li, D.; Chen, B.; Li, J. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems* **2024**, *37*, 28828–28857.
42. Fu et al.. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis, 2025.
43. Wang et al.. LVBench: An Extreme Long Video Understanding Benchmark, 2024.
44. Wang et al.. Multimodal Needle in a Haystack: Benchmarking Long-Context Capability of Multimodal Large Language Models, 2025.
45. Wang et al.. VD-BERT: A Unified Vision and Dialog Transformer with BERT, 2020.
46. Le et al.. Multimodal Dialogue State Tracking, 2022.
47. Lu et al.. RecFormer: Recurrent Multi-modal Transformer with History-Aware Contrastive Learning for Visual Dialog, 2023.
48. Yan et al.. MMCR: Advancing Visual Language Model in Multimodal Multi-Turn Contextual Reasoning, 2025.
49. Liu et al.. Taking Notes Brings Focus: Towards Multi-Turn Multimodal Dialogue Learning, 2025.
50. Chen et al.. VideoLLM-online: Online Video Large Language Model for Streaming Video, 2024.
51. Qian et al.. Streaming Long Video Understanding with Large Language Models, 2024.
52. Wang, X.; Zhou, Y.; Liu, X.; Lu, H.; Xu, Y.; He, F.; Yoon, J.; Lu, T.; Liu, F.; Bertasius, G.; et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 416–442.
53. Niu et al.. OVO-Bench: How Far is Your Video-LLMs from Real-World Online Video Understanding?, 2025.
54. Zhu et al.. Vision-Dialog Navigation by Exploring Cross-modal Memory, 2020.
55. Chen et al.. History Aware Multimodal Transformer for Vision-and-Language Navigation, 2021.
56. Chen et al.. Think Global, Act Local: Dual-scale Graph Transformer for Vision-and-Language Navigation, 2022.
57. Shah et al.. LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action, 2022.
58. Chang et al.. GOAT: GO to Any Thing, 2024.
59. An et al.. ETPNav: Evolving Topological Planning for Vision-Language Navigation in Continuous Environments, 2025.
60. Guu et al.. REALM: Retrieval-Augmented Language Model Pre-Training, 2020.
61. Lewis et al.. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2020.
62. Chen et al.. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text, 2022.
63. Park et al.. Generative Agents: Interactive Simulacra of Human Behavior, 2023.
64. Jain et al.. AUGUSTUS: LLM-Driven Contextualized User Memory in Personalized Multimodal Agents, 2025.
65. Zhang et al.. ELLA: Embodied Social Agents with Lifelong Memory, 2025.
66. Wang et al.. Voyager: An Open-Ended Embodied Agent with Large Language Models, 2024.
67. Wang et al.. LifelongMemory: Leveraging LLMs for Answering Queries in Egocentric Videos, 2023.
68. Wang et al.. KARMA: Augmenting Embodied AI Agents with Long-and-Short Term Memory Systems, 2025.
69. Anwar et al.. ReMEmbR: Building and Reasoning Over Long-Horizon Spatio-Temporal Memory for Robot Navigation, 2025.
70. Tan et al.. RoboOS-NeXT: A Unified Memory-based Framework for Lifelong, Scalable, and Robust Multi-Robot Collaboration, 2025.
71. Long et al.. RoboMemory: A Brain-Inspired Multi-Memory Agentic Framework for Lifelong Learning, 2025.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

72. Wang, P.; Hu, Y.; Liu, X.; Yang, J.; Wang, H.; Wen, Z. AtlasVA: Self-Evolving Visual Skill Memory for Teacher-Free VLM Agents. *CoRR* **2026**, *abs/2605.17933*, [2605.17933]. https://doi.org/10.48550/ARXIV.2605.17933.
73. Jaegle et al.. Perceiver: General Perception with Iterative Attention, 2021.
74. Li et al.. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, 2023.
75. Ryoo et al.. TokenLearner: What Can 8 Learned Tokens Do for Images and Videos?, 2021.
76. Sandler et al.. Fine-tuning Image Transformers using Learnable Memory, 2022.
77. Jaegle et al.. Perceiver IO: A General Architecture for Structured Inputs and Outputs, 2022.
78. Dai et al.. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, 2023.
79. Ye et al.. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality, 2023.
80. Lei et al.. MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning, 2020.
81. Lin et al.. Multimodal Transformer with Variable-Length Memory for Vision-and-Language Navigation, 2022.
82. Zhou et al.. Learning Without Forgetting for Vision-Language Models, 2025.
83. Zhang et al.. LLaMA-Adapter: Efficient Fine-tuning of Large Language Models with Zero-initialized Attention, 2024.
84. Jie et al.. Memory-Space Visual Prompting for Efficient Vision-Language Fine-Tuning, 2024.
85. Su et al.. AKVQ-VL: Attention-Aware KV Cache Adaptive 2-Bit Quantization for Vision-Language Models, 2025.
86. Hu et al.. Reveal: Retrieval-Augmented Visual-Language Pre-Training with Multi-Source Multimodal Knowledge Memory, 2023.
87. Hsiao et al.. MRAG-Bench: Vision-Centric Evaluation for Retrieval-Augmented Multimodal Models, 2025.
88. Guo et al.. RAG-Anything: All-in-One RAG Framework, 2025.
89. Marino et al.. KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA, 2021.
90. Gui et al.. KAT: A Knowledge Augmented Transformer for Vision-and-Language, 2022.
91. Yasunaga et al.. Retrieval-Augmented Multimodal Language Modeling, 2023.
92. Tiong et al.. Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training, 2022.
93. Lin et al.. Rethinking Visual Prompting for Multimodal Large Language Models with External Knowledge, 2024.
94. Hu et al.. MegaRAG: Multimodal Knowledge Graph Based Retrieval-Augmented Generation, 2025.
95. Deng et al.. MuKA: Multimodal Knowledge Augmented Visual Information-Seeking, 2025.
96. Li et al.. SearchLVLMs: A Plug-and-Play Framework for Augmenting Large Vision-Language Models by Searching Up-to-Date Internet Knowledge, 2024.
97. Wu et al.. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs, 2024.
98. Sohn et al.. R4: Retrieval-Augmented Reasoning for Multimodal Models, 2025.
99. Yu et al.. VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents, 2025.
100. Fan et al.. End-to-End Optimization for Multimodal Retrieval-Augmented Generation via Reward Back-propagation, 2025.
101. Li et al.. HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training, 2020.
102. Ren, S.; Yao, L.; Li, S.; Sun, X.; Hou, L. Timechat: A time-sensitive multimodal large language model for long video understanding. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14313–14323.
103. Shah et al.. Reasoning over History: Context Aware Visual Dialog, 2020.
104. Chen et al.. Improving Cross-Modal Understanding in Visual Dialog Via Contrastive Learning, 2022.
105. He et al.. Mem4Nav: Boosting Vision-and-Language Navigation with Memory, 2025.
106. Wang et al.. VideoLLaMB: Long-Context Video Understanding with Recurrent Memory Bridges, 2024.
107. Zellers et al.. MERLOT Reserve: Neural Script Knowledge through Vision, Language, and Sound, 2022.
108. Cheng et al.. Can We Edit Multimodal Large Language Models?, 2023.
109. Wu et al.. Unified Knowledge Maintenance for Vision-Language Models, 2025.
110. Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. Moviechat: From dense token to sparse memory for long video understanding. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18221–18232.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

111. Shen et al.. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding, 2025.
112. Li et al.. VideoChat-Flash: Hierarchical Compression for Long-Context Video Modeling, 2025.
113. Sung, Y.L.; Cho, J.; Bansal, M. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5227–5237.
114. Zeng et al.. Visual-Oriented Fine-Grained Knowledge Editing for Multi-modal Large Language Models, 2024.
115. Tang et al.. Mind the Interference: Retaining Pre-trained Knowledge in Parameter Efficient Continual Learning of Vision-Language Models, 2024.
116. Wu et al.. Synthetic Data is an Elegant GIFT for Continual Vision-Language Models, 2025.
117. Fu et al.. IAP: Improving Continual Learning of Vision-Language Models via Instance-Aware Prompting, 2026.
118. Sun et al.. Mitigating Visual Forgetting via Take-along Visual Conditioning for Multi-modal Long CoT Reasoning, 2025.
119. Wang, W.; He, Z.; Hong, W.; Cheng, Y.; Zhang, X.; Qi, J.; Ding, M.; Gu, X.; Huang, S.; Xu, B.; et al. Lvbench: An extreme long video understanding benchmark. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 22958–22967.
120. Chen, S.N.; Chen, H.; Ho, C.; Mao, X.; Wang, J.; Zhang, Y.; Li, C. Seeing the Scene Matters: Revealing Forgetting in Video Understanding Models with a Scene-Aware Long-Video Benchmark. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 4515–4525.
121. Huang, Z.; Li, X.; Li, J.; Wang, J.; Zeng, X.; Liang, C.; Wu, T.; Chen, X.; Li, L.; Wang, L. Online video understanding: A comprehensive benchmark and memory-augmented method. *arXiv e-prints* **2024**, pp. arXiv–2501.
122. Niu, J.; Li, Y.; Miao, Z.; Ge, C.; Zhou, Y.; He, Q.; Dong, X.; Duan, H.; Ding, S.; Qian, R.; et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 18902–18913.
123. Lin, J.; Fang, Z.; Chen, C.; Cheng, H.; Wan, Z.; Luo, F.; Wang, Z.; Li, P.; Liu, Y.; Sun, M. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In Proceedings of the ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2026, pp. 12147–12151.
124. Nagrani, A.; Zhang, M.; Mehran, R.; Hornung, R.; Gundavarapu, N.B.; Jha, N.; Myers, A.; Zhou, X.; Gong, B.; Schmid, C.; et al. Neptune: The long orbit to benchmarking long video understanding. *arXiv preprint arXiv:2412.09582* **2024**.
125. Chen, G.; Liu, Y.; Huang, Y.; Pei, B.; Xu, J.; He, Y.; Lu, T.; Wang, Y.; Wang, L. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In Proceedings of the International Conference on Learning Representations, 2025, Vol. 2025, pp. 45647–45682.
126. Zhu, T.; Zhang, S.; Sun, Z.; Tian, J.; Tang, Y. Memorize-and-Generate: Towards Long-Term Consistency in Real-Time Video Generation. *arXiv preprint arXiv:2512.18741* **2025**.
127. Zhang, S.; Zhang, Z.; Huang, S.; Tang, Z.; Wang, H.; Dai, C.; Chen, M.; Li, Y.; Li, Y.; Chen, Y.; et al. MBench: A Comprehensive Benchmark on Memory Capability for Video World Models. *arXiv preprint arXiv:2606.00793* **2026**.
128. Ye, Y.; Lu, X.; Jiang, Y.; Gu, Y.; Zhao, R.; Liang, Q.; Pan, J.; Zhang, F.; Wu, W.; Wang, A.J. Mind: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025* **2026**.
129. Lyu, W.; Du, Y.; Zhao, J.; Zhen, X.; Shao, L. VisChainBench: A Benchmark for Multi-Turn, Multi-Image Visual Reasoning Beyond Language Priors. *arXiv preprint arXiv:2512.06759* **2025**.
130. Chunseong Park, C.; Kim, B.; Kim, G. Attend to you: Personalized image captioning with context sequence memory networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 895–903.
131. Zheng, Z.; Ma, M.; Wang, K.; Qin, Z.; Yue, X.; You, Y. Preventing zero-shot transfer degradation in continual learning of vision-language models. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 19125–19136.
132. Liu, W.; Zhu, F.; Wei, L.; Tian, Q. C-CLIP: Multimodal continual learning for vision-language model. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

133. Zhao, H.; Zhu, F.; Guo, H.; Wang, M.; Wang, R.; Meng, G.; Zhang, Z. Mllm-cl: Continual learning for multimodal large language models. *arXiv preprint arXiv:2506.05453* **2025**.
134. Weng, X.; Ni, R.; Pang, C.; Hao, X.; Wang, Y.; Zhang, X.; Xu, W.; Xia, G.S. Continual Vision-Language Learning for Remote Sensing: Benchmarking and Analysis. *arXiv preprint arXiv:2604.00820* **2026**.
135. Srinivasan, T.; Chang, T.Y.; Pinto Alva, L.; Chochlakis, G.; Rostami, M.; Thomason, J. Climb: A continual learning benchmark for vision-and-language tasks. *Advances in Neural Information Processing Systems* **2022**, *35*, 29440–29453.
136. Chen, C.; Zhu, J.; Luo, X.; Shen, H.T.; Song, J.; Gao, L. Coin: A benchmark of continual instruction tuning for multimodel large language models. *Advances in neural information processing systems* **2024**, *37*, 57817–57840.
137. Guo, H.; Hou, Z.; Sun, Y.; He, J.; Chen, Y.; Zhou, Y.; Jia, Y.; Wang, J.; Chua, T.S. MLLM-CTBench: A Benchmark for Continual Instruction Tuning with Reasoning Process Diagnosis. *arXiv preprint arXiv:2508.08275* **2025**.
138. Zhang, Y.; Chen, H.; Frikha, A.; Krompass, D.; Zhang, G.; Gu, J.; Tresp, V. Cl-cross vqa: A continual learning benchmark for cross-domain visual question answering. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025, pp. 6269–6278.
139. Tang, T.; Deldari, S.; Xue, H.; De Melo, C.; Salim, F. Vilco-bench: Video language continual learning benchmark. *Advances in Neural Information Processing Systems* **2024**, *37*, 70213–70229.
140. Guo, H.; Shi, Y.; Zhu, F.; Liu, W.; Zhao, H.; Zeng, F.; Ma, S.; Wang, D.H.; Zhang, X.Y. CL-VISTA: Benchmarking Continual Learning in Video Large Language Models. *arXiv preprint arXiv:2604.00677* **2026**.
141. Yadav, K.; Ali, Y.; Gupta, G.; Gal, Y.; Kira, Z. Findingdory: A benchmark to evaluate memory in embodied agents. *arXiv preprint arXiv:2506.15635* **2025**.
142. Wang, S.; Liu, B.; Gao, Z.; Ma, L.; Wang, X.; Xie, Y.; Tan, X. Explore with Long-term Memory: A Benchmark and Multimodal LLM-based Reinforcement Learning Framework for Embodied Exploration. *arXiv preprint arXiv:2601.10744* **2026**.
143. Hu, W.; Hong, Y.; Wang, Y.; Gao, L.; Wei, Z.; Yao, X.; Peng, N.; Bitton, Y.; Szpektor, I.; Chang, K.W. 3dllm-mem: Long-term spatial-temporal memory for embodied 3d large language model. *Advances in Neural Information Processing Systems* **2026**, *38*, 67856–67884.
144. Chen, T.; Wang, Y.; Li, M.; Qin, Y.; Shi, H.; Li, Z.; Hu, Y.; Zhang, Y.; Wang, K.; Chen, Y.; et al. Rmbench: Memory-dependent robotic manipulation benchmark with insights into policy design. *arXiv preprint arXiv:2603.01229* **2026**.
145. Dai, Y.; Fu, H.; Lee, J.; Liu, Y.; Zhang, H.; Yang, J.; Finn, C.; Fazeli, N.; Chai, J. RoboMME: Benchmarking and understanding memory for robotic generalist policies. *arXiv preprint arXiv:2603.04639* **2026**.
146. Lei, H.; Song, W.; Zhang, H.; Pei, J.; Chen, J.; Yan, H.; Zhao, H.; Ding, P.; Zhang, Z.; Huang, L.; et al. RoboMemArena: A Comprehensive and Challenging Robotic Memory Benchmark. *arXiv preprint arXiv:2605.10921* **2026**.
147. Chung, N.; Hanyu, T.; Nguyen, T.; Le, H.; Bumgarner, F.; Nguyen, D.M.H.; Vo, K.; Yamazaki, K.; Rainwater, C.; Kieu, T.; et al. Rethinking progression of memory state in robotic manipulation: An object-centric perspective. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 3407–3415.
148. Vo, H.; Vo, K.; Nguyen, P.L.; Tran, S.; Nguyen, D.M.; Cuong, N.X.; Gawugah, G.; Godavarthi, S.A.T.; Rainwater, C.; Bui, N.D.; et al. DRIVESPATIAL: A Benchmark for Spatiotemporal Intelligence in VLMs for Autonomous Driving. *arXiv preprint arXiv:2605.23176* **2026**.
149. Li, Y.; Li, Y.; Wang, X.; Jiang, Y.; Zhang, Z.; Zheng, X.; Wang, H.; Zheng, H.T.; Huang, F.; Zhou, J.; et al. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. In Proceedings of the International Conference on Learning Representations, 2025, Vol. 2025, pp. 95582–95604.
150. Wu, Y.; Long, Q.; Li, J.; Yu, J.; Wang, W. Visual-rag: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries. *arXiv preprint arXiv:2502.16636* **2025**.
151. Anugraha, D.; Irawan, P.A.; Singh, A.; Lee, E.S.A.; Winata, G.I. M4-RAG: A Massive-Scale Multilingual Multi-Cultural Multimodal RAG. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 23083–23094.
152. Liu, Z.; Zhu, X.; Zhou, T.; Zhang, X.; Yi, X.; Yan, Y.; Yu, G.; Sun, M. Benchmarking retrieval-augmented generation in multi-modal contexts. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 4817–4826.
153. Dong, K.; YUJING, C.; Huang, S.; Wang, Y.; Tang, R.; Liu, Y. Benchmarking retrieval-augmented multimomal generation for document question answering. *Advances in Neural Information Processing Systems* **2026**, *38*.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

154. Peng, X.; Qin, C.; Chen, Z.; Xu, R.; Xiong, C.; Wu, C.S. Unidoc-bench: A unified benchmark for document-centric multimodal rag. *arXiv preprint arXiv:2510.03663* **2025**.
155. Suri, M.; Mathur, P.; Dernoncourt, F.; Goswami, K.; Rossi, R.A.; Manocha, D. Visdom: Multi-document qa with visually rich elements using multimodal retrieval-augmented generation. In Proceedings of the Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2025, pp. 6088–6109.
156. Yang, Y.; Zhong, J.; Jin, L.; Huang, J.; Gao, J.; Liu, Q.; Bai, Y.; Zhang, J.; Jiang, R.; Wei, K. Benchmarking multimodal RAG through a chart-based document question-answering generation framework. *arXiv preprint arXiv:2502.14864* **2025**.
157. Hu, W.; Gu, J.C.; Dou, Z.Y.; Fayyaz, M.; Lu, P.; Chang, K.W.; Peng, N.V. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. In Proceedings of the International Conference on Learning Representations, 2025, Vol. 2025, pp. 95558–95581.
158. Li, J.; Yuan, Y.; Li, W.; Aliannejadi, M.; Hershcovich, D.; Søgaard, A.; Vulić, I.; Zhang, W.; Liang, P.P.; Deng, Y.; et al. Ravenea: A benchmark for multimodal retrieval-augmented visual culture understanding. *arXiv preprint arXiv:2505.14462* **2025**.
159. Zhao, S.; Jin, Z.; Li, S.; Gao, J. Finragbench-v: A benchmark for multimodal rag with visual citation in the financial domain. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 4215–4249.
160. Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 18995–19012.
161. Goletto, G.; Nagarajan, T.; Averta, G.; Damen, D. Amego: Active memory from long egocentric videos. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 92–110.
162. Ye, H.; Zhang, H.; Daxberger, E.; Chen, L.; Lin, Z.; Li, Y.; Zhang, B.; You, H.; Xu, D.; Gan, Z.; et al. MMEgo: Towards building egocentric multimodal LLMs for video QA. In Proceedings of the International Conference on Learning Representations, 2025, Vol. 2025, pp. 71705–71723.
163. Wang, Z.; Zhang, Y.; Yu, S.; Zhang, C.; Zhao, Z.; Yoon, J.; Lee, H.; Bertasius, G.; Bansal, M. EgoMemReason: A Memory-Driven Reasoning Benchmark for Long-Horizon Egocentric Video Understanding. *arXiv preprint arXiv:2605.09874* **2026**.
164. Forte, R.; Lando, G.; Furnari, A. EGOSTREAM: A Diagnostic Benchmark for Streaming Episodic Memory in Egocentric Vision. *arXiv preprint arXiv:2605.31557* **2026**.
165. Yan, J.; Ren, R.; Liu, J.; Xu, S.; Wang, L.; Wang, Y.; Zhong, X.; Wang, Y.; Zhang, L.; Chen, X.; et al. TeleEgo: Benchmarking Egocentric AI Assistants in the Wild. *arXiv preprint arXiv:2510.23981* **2025**.
166. Yang, J.; Liu, S.; Guo, H.; Dong, Y.; Zhang, X.; Zhang, S.; Wang, P.; Zhou, Z.; Xie, B.; Wang, Z.; et al. Egolife: Towards egocentric life assistant. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 28885–28900.
167. Alam, S.; Siam, S.I.; Proulx, M.J.; Fort, J.; Newcombe, R.; Kim, H.J.; Zhang, M. SuperMemory-VQA: An Egocentric Visual Question-Answering Benchmark for Long-Horizon Memory. *arXiv preprint arXiv:2606.00825* **2026**.
168. Xiao, J.; Zhang, S.; Zhu, P.; Yao, A. Ego-Grounding for Personalized Question-Answering in Egocentric Videos. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 40537–40547.
169. Wang, Z.; Liu, C.; Tjitrahardja, E.; Wang, Y.; Pavlov, B.; Gou, F.; Davila, J.M.; Shi, D.; Xu, R.; Pan, Y.; et al. EgoIntrospect: An Egocentric Dataset and Benchmark for User-Centric Internal State Reasoning. *arXiv preprint arXiv:2605.17262* **2026**.
170. Kim, J.; Kim, W.; Park, W.; Do, J. Mmpb: It's time for multi-modal personalization. *Advances in Neural Information Processing Systems* **2026**, *38*.
171. Alaluf, Y.; Richardson, E.; Tulyakov, S.; Aberman, K.; Cohen-Or, D. Myvlm: Personalizing vlms for user-specific queries. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 73–91.
172. Nguyen, T.; Liu, H.; Li, Y.; Cai, M.; Ojha, U.; Lee, Y.J. Yo'llava: Your personalized language and vision assistant. *Advances in Neural Information Processing Systems* **2024**, *37*, 40913–40951.
173. An, R.; Yang, S.; Zhang, R.; Lu, M.; Jiang, T.; Zeng, K.; Luo, Y.; Cao, J.; Liang, H.; Chen, Y.; et al. Mc-llava: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706* **2024**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

174. Oh, Y.; Yu, S.; Park, J.; Moon, H.C.; Mok, J.; Yoon, S. Contextualized Visual Personalization in Vision-Language Models. *arXiv preprint arXiv:2602.03454* **2026**.
175. Nie, C.; Fu, C.; Zhang, Y.; Yang, H.; Shan, C. Personavlm: Long-term personalized multimodal llms. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 15000–15009.
176. Mei, J.; Chen, J.; Yang, G.; Hou, X.; Li, M.; Byrne, B. According to me: Long-term personalized referential memory qa. *arXiv preprint arXiv:2603.01990* **2026**.
177. Hong, R.; Lang, J.; Zhong, T.; Wang, Y.; Zhou, F. Tameing long contexts in personalization: Towards training-free and state-aware mllm personalized assistant. In Proceedings of the Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1, 2026, pp. 452–463.
178. Ren, X.; Wang, Z.; Du, Y.; Xie, Z.; Liu, C.; Yang, X.; Feng, H.; Pan, W.; Zheng, T.; Xu, B.; et al. MEM-LENS: Benchmarking Multimodal Long-Term Memory in Large Vision-Language Models. *arXiv preprint arXiv:2605.14906* **2026**.
179. Bei, Y.; Wei, T.; Ning, X.; Zhao, Y.; Liu, Z.; Lin, X.; Zhu, Y.; Hamann, H.; He, J.; Tong, H. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. *arXiv preprint arXiv:2601.03515* **2026**.
180. Li, X.; Zhu, Z.; Liu, S.; Ma, Y.; Zang, Y.; Cao, Y.; Sun, A. EMemBench: Interactive Benchmarking of Episodic Memory for VLM Agents. *arXiv preprint arXiv:2601.16690* **2026**.
181. Doss, T.S.M.; Xu, M.; Rao, S.; Wilson, A.D.; Kumaravel, B.T. MineNPC-Task: Task Suite for Memory-Aware Minecraft Agents. *arXiv preprint arXiv:2601.05215* **2026**.
182. Ju, T.; Sun, Y.; Wu, Z.; Zhang, W.; Huo, Y.; Su, X.; Gu, Q.; Cai, X.; Liu, G.; Zhang, Z. MineExplorer: Evaluating Open-World Exploration of MLLM Agents in Minecraft. *arXiv preprint arXiv:2605.30931* **2026**.
183. He, Z.; Wang, Y.; Zhi, C.; Hu, Y.; Chen, T.P.; Yin, L.; Chen, Z.; Wu, T.A.; Ouyang, S.; Wang, Z.; et al. Memoryarena: Benchmarking agent memory in interdependent multi-session agentic tasks. *arXiv preprint arXiv:2602.16313* **2026**.
184. Xu, W.; Wang, Y.; Mei, K.; Liang, K.; Wang, Z.; Jin, M.; Zhang, H.; Zhang, S.X.; Hua, W.; Sahu, S.; et al. MemGym: a Long-Horizon Memory Environment for LLM Agents. *arXiv preprint arXiv:2605.20833* **2026**.
185. Liu, G.; Zhao, P.; Liang, Y.; Luo, Q.; Tang, S.; Chai, Y.; Lin, W.; Xiao, H.; Wang, W.; Chen, S.; et al. MemGUI-Bench: Benchmarking Memory of Mobile GUI Agents in Dynamic Environments. *arXiv preprint arXiv:2602.06075* **2026**.
186. Kim, K.M.; Heo, M.O.; Choi, S.H.; Zhang, B.T. Deepstory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836* **2017**.
187. Zhang, C.; Lu, T.; Islam, M.M.; Wang, Z.; Yu, S.; Bansal, M.; Bertasius, G. A simple llm framework for long-range video question-answering. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 21715–21737.
188. Kahatapitiya, K.; Ranasinghe, K.; Park, J.; Ryoo, M.S. Language repository for long video understanding. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 5627–5646.
189. Song, E.; Chai, W.; Ye, T.; Hwang, J.N.; Li, X.; Wang, G. Moviechat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**.
190. Bai, Z.; Wang, R.; Chen, X. Glance and focus: Memory prompting for multi-event video question answering. *Advances in Neural Information Processing Systems* **2023**, *36*, 34247–34259.
191. He, B.; Li, H.; Jang, Y.K.; Jia, M.; Cao, X.; Shah, A.; Shrivastava, A.; Lim, S.N. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 13504–13514.
192. Yuan, H.; Liu, Z.; Qin, M.; Qian, H.; Shu, Y.; Dou, Z.; Wen, J.R.; Sebe, N. Memory-enhanced retrieval augmentation for long video understanding. *arXiv preprint arXiv:2503.09149* **2025**.
193. Faure, G.J.; Yeh, J.F.; Chen, M.H.; Su, H.T.; Lai, S.H.; Hsu, W.H. Hermes: temporal-coherent long-form understanding with episodes and semantics. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 22911–22921.
194. Cheng, D.; Li, M.; Liu, J.; Guo, Y.; Jiang, B.; Liu, Q.; Chen, X.; Zhao, B. Enhancing long video understanding via hierarchical event-based memory. In Proceedings of the 2025 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2025, pp. 1–6.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

195. Santos, S.; Farinhas, A.; McNamee, D.C.; Martins, A. ∞-Video: A Training-Free Approach to Long Video Understanding via Continuous-Time Memory Consolidation. In Proceedings of the Proceedings of the 42nd International Conference on Machine Learning. PMLR, 2025, pp. 52877–52893.
196. Lin, Y.; Zhang, J.; Wang, Q.; Ye, H.; Fu, Y.; Liu, Y.; Li, H.; Chen, Y.; et al. Hippomm: Hippocampal-inspired multimodal memory for long audiovisual event understanding. *arXiv preprint arXiv:2504.10739* **2025**.
197. Wang, Y.; Zhang, L.; Liu, J.; Yan, J.; Zhang, Z.; Zheng, J.; Yang, X.; Wu, D.; Chen, X.; Li, X. Episodic memory representation for long-form video understanding. *arXiv preprint arXiv:2508.09486* **2025**.
198. Yeo, J.H.; Chung, S.; Park, S.; Kim, D.H.; Moon, J.; Ro, Y.M. Gcagent: Long-video understanding via schematic and narrative episodic memory. *arXiv preprint arXiv:2511.12027* **2025**.
199. Huang, B.; Wang, X.; Chen, H.; Song, Z.; Zhu, W. Vtimellm: Empower llm to grasp video moments. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14271–14280.
200. Deng, A.; Gao, Z.; Choudhuri, A.; Planche, B.; Zheng, M.; Wang, B.; Chen, T.; Chen, C.; Wu, Z. Seq2time: Sequential knowledge transfer for video llm temporal grounding. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 13766–13775.
201. Wang, H.; Xu, Z.; Cheng, Y.; Diao, S.; Zhou, Y.; Cao, Y.; Wang, Q.; Ge, W.; Huang, L. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290* **2024**.
202. Guo, Y.; Liu, J.; Li, M.; Liu, Q.; Chen, X.; Tang, X. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643* **2024**.
203. Wu, Y.; Hu, X.; Sun, Y.; Zhou, Y.; Zhu, W.; Rao, F.; Schiele, B.; Yang, X. Number it: Temporal grounding videos like flipping manga. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 13754–13765.
204. Zhao, H.; Ji, G.P.; Yan, R.; Xiong, H.; Li, Z. Videoexpert: Augmented llm for temporal-sensitive video understanding. *IEEE Transactions on Circuits and Systems for Video Technology* **2026**.
205. Wang, Y.; Wang, Z.; Xu, B.; Du, Y.; Lin, K.; Xiao, Z.; Yue, Z.; Ju, J.; Zhang, L.; Yang, D.; et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377* **2025**.
206. Wang, S.; Chen, G.; Huang, D.a.; Li, Z.; Li, M.; Li, G.; Alvarez, J.M.; Zhang, L.; Yu, Z. VideoITG: Multimodal Video Understanding with Instructed Temporal Grounding. *arXiv preprint arXiv:2507.13353* **2025**.
207. Nie, J.; An, W.; Zhang, G.; Xu, Y.; Tan, Y.P.; Kot, A.C.; Lu, S. EM Ground: A Temporal Grounding Vid-LLM with Holistic Event Perception and Matching. *arXiv preprint arXiv:2602.05215* **2026**.
208. Qian, R.; Dong, X.; Zhang, P.; Zang, Y.; Ding, S.; Lin, D.; Wang, J. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems* **2024**, *37*, 119336–119360.
209. Chen, J.; Lv, Z.; Wu, S.; Lin, K.Q.; Song, C.; Gao, D.; Liu, J.W.; Gao, Z.; Mao, D.; Shou, M.Z. Videollm-online: Online video large language model for streaming video. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18407–18418.
210. Zhang, H.; Wang, Y.; Tang, Y.; Liu, Y.; Feng, J.; Dai, J.; Jin, X. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085* **2024**.
211. Xiong, H.; Yang, Z.; Yu, J.; Zhuge, Y.; Zhang, L.; Zhu, J.; Lu, H. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. *arXiv preprint arXiv:2501.13468* **2025**.
212. Di, S.; Yu, Z.; Zhang, G.; Li, H.; Zhong, T.; Cheng, H.; Li, B.; He, W.; Shu, F.; Jiang, H. Streaming video question-answering with in-context video kv-cache retrieval. *arXiv preprint arXiv:2503.00540* **2025**.
213. Yang, Y.; Zhao, Z.; Shukla, S.N.; Singh, A.; Mishra, S.K.; Zhang, L.; Ren, M. Streammem: Query-agnostic kv cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717* **2025**.
214. Chatterjee, D.; Remelli, E.; Song, Y.; Tekin, B.; Mittal, A.; Bhatnagar, B.; CamgÃķz, N.C.; Hampali, S.; Sauser, E.; Ma, S.; et al. Memory-efficient streaming videollms for real-time procedural video understanding. *arXiv preprint arXiv:2504.13915* **2025**.
215. Patel, S.; Patel, D. CacheFlow: Compressive Streaming Memory for Efficient Long-Form Video Understanding. *arXiv preprint arXiv:2511.13644* **2025**.
216. Han, Z.; Sun, H.; Xu, J.; Tang, C.; Lei, Y.; Zhang, X.; Sun, H.; He, Z.; Sun, H. WAT: Online Video Understanding Needs Watching Before Thinking. *arXiv preprint arXiv:2603.13412* **2026**.
217. Azad, S.; Vineet, V.; Rawat, Y.S. StreamReady: Learning What to Answer and When in Long Streaming Videos. *arXiv preprint arXiv:2603.08620* **2026**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

218. Wang, J.; Wang, Y.; Zhao, D.; Zheng, Z. MoviePuzzle: Visual narrative reasoning through multimodal order learning. *arXiv preprint arXiv:2306.02252* **2023**.
219. Yu, J.; Wu, Y.; Chu, M.; Ren, Z.; Huang, Z.; Chu, P.; Zhang, R.; He, Y.; Li, Q.; Li, S.; et al. Vrbench: A benchmark for multi-step reasoning in long narrative videos. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 21655–21666.
220. Sarkar, P.; Etemad, A. Vcrbench: Exploring long-form causal reasoning capabilities of large video language models. *arXiv preprint arXiv:2505.08455* **2025**.
221. Jain, R.; Doshi, K.; Uzkent, B.; Kessler, G. Narrative Aligned Long Form Video Question Answering. *arXiv preprint arXiv:2603.19481* **2026**.
222. Singh, D.; Nagrani, A.; Manikantan, K.; Singh, H.; Tewari, D.; Weyand, T.; Schmid, C.; Angelova, A.; Dave, S. CURVE: A Benchmark for Cultural and Multilingual Long Video Reasoning. *arXiv preprint arXiv:2601.10649* **2026**.
223. Wang, M.; Wang, R.; Lin, J.; Ji, R.; Wiedemer, T.; Gao, Q.; Luo, D.; Qian, Y.; Huang, L.; Hong, Z.; et al. A Very Big Video Reasoning Suite. *arXiv preprint arXiv:2602.20159* **2026**.
224. Fateh, F.J.; Ahmed, U.; Khan, H.; Zia, M.Z.; Tran, Q.H. Video llms for temporal reasoning in long videos. *arXiv preprint arXiv:2412.02930* **2024**.
225. Yeo, W.; Kim, K.; Yoon, J.; Hwang, S.J. Worldmm: Dynamic multimodal memory agent for long video reasoning. *arXiv preprint arXiv:2512.02425* **2025**.
226. Liu, R.; Liu, Z.; Tang, J.; Ma, Y.; Pi, R.; Zhang, J.; Chen, Q. LongVideoAgent: Multi-Agent Reasoning with Long Videos. *arXiv preprint arXiv:2512.20618* **2025**.
227. Xu, Y.; Liu, G.; Kompella, R.R.; Huang, T.; Hu, S.; Ilhan, F.; Tekin, S.F.; Yahn, Z.; Liu, L. A Multi-Agent Perception-Action Alliance for Efficient Long Video Reasoning. *arXiv preprint arXiv:2603.14052* **2026**.
228. Jain, J.; Li, J.; Ma, Z.; Zhang, J.; Kim, C.D.; Lee, S.; Tripathi, R.; Gupta, T.; Clark, C.; Shi, H. SAGE: Training Smart Any-Horizon Agents for Long Video Reasoning with Reinforcement Learning. *arXiv preprint arXiv:2512.13874* **2025**.
229. Ren, X.; Shen, T.; Huang, J.; Ling, H.; Lu, Y.; Nimier-David, M.; Müller, T.; Keller, A.; Fidler, S.; Gao, J. Gen3c: 3d-informed world-consistent video generation with precise camera control. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 6121–6132.
230. Yu, J.; Bai, J.; Qin, Y.; Liu, Q.; Wang, X.; Wan, P.; Zhang, D.; Liu, X. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In Proceedings of the Proceedings of the SIGGRAPH Asia 2025 Conference Papers, 2025, pp. 1–11.
231. Liu, Z.; Deng, X.; Chen, S.; Wang, A.; Guo, Q.; Han, M.; Xue, Z.; Chen, M.; Luo, P.; Yang, L. Worldweaver: Generating long-horizon video worlds via rich perception. *arXiv preprint arXiv:2508.15720* **2025**.
232. Wu, X.; Zhang, G.; Xu, Z.; Zhou, Y.; Lu, Q.; He, X. Pack and force your memory: Long-form and consistent video generation. *arXiv preprint arXiv:2510.01784* **2025**.
233. Yang, S.; Huang, W.; Chu, R.; Xiao, Y.; Zhao, Y.; Wang, X.; Li, M.; Xie, E.; Chen, Y.; Lu, Y.; et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622* **2025**.
234. Chen, S.; Wei, C.; Sun, S.; Nie, P.; Zhou, K.; Zhang, G.; Yang, M.H.; Chen, W. Context Forcing: Consistent Autoregressive Video Generation with Long Context. *arXiv preprint arXiv:2602.06028* **2026**.
235. Zhao, Z.; Lu, Y.; Liu, Z.; Song, J.; Deng, J.; Patras, I. Relax Forcing: Relaxed KV-Memory for Consistent Long Video Generation. *arXiv preprint arXiv:2603.21366* **2026**.
236. Zhou, J.; Du, Y.; Xu, X.; Wang, L.; Zhuang, Z.; Zhang, Y.; Li, S.; Hu, X.; Su, B.; Chen, Y.c. VideoMemory: Toward Consistent Video Generation via Memory Integration. *arXiv preprint arXiv:2601.03655* **2026**.
237. Wang, Z.; Lin, H.; Yoon, J.; Cho, J.; Zhang, Y.; Bansal, M. AnchorWeave: World-Consistent Video Generation with Retrieved Local Spatial Memories. *arXiv preprint arXiv:2602.14941* **2026**.
238. Gao, X.; Guan, J.; Luo, S.; Li, W.; Tan, G.; Wang, J. MemCam: Memory-Augmented Camera Control for Consistent Video Generation. *arXiv preprint arXiv:2603.26193* **2026**.
239. Chi, X.; Fan, C.K.; Zhang, H.; Qi, X.; Zhang, R.; Chen, A.; Chan, C.m.; Xue, W.; Liu, Q.; Zhang, S.; et al. Eva: An embodied world model for future video anticipation. *arXiv preprint arXiv:2410.15461* **2024**.
240. Huang, S.; Wu, J.; Zhou, Q.; Miao, S.; Long, M. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357* **2025**.
241. Wu, T.; Yang, S.; Po, R.; Xu, Y.; Liu, Z.; Lin, D.; Wetzstein, G. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284* **2025**.
242. Hong, Y.; Mei, Y.; Ge, C.; Xu, Y.; Zhou, Y.; Bi, S.; Hold-Geoffroy, Y.; Roberts, M.; Fisher, M.; Shechtman, E.; et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040* **2025**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

243. Huang, J.; Ye, Z.; Hu, X.; He, T.; Zhang, G.; Shi, S.; Bian, J.; Jiang, L. LIVE: Long-horizon Interactive Video World Modeling. *arXiv preprint arXiv:2602.03747* **2026**.
244. Xiang, J.; Gu, Y.; Liu, Z.; Feng, Z.; Gao, Q.; Hu, Y.; Huang, B.; Liu, G.; Yang, Y.; Zhou, K.; et al. Pan: A world model for general, interactable, and long-horizon world simulation. *arXiv preprint arXiv:2511.09057* **2025**.
245. Zhu, Y.; Feng, J.; Zheng, W.; Gao, Y.; Tao, X.; Wan, P.; Zhou, J.; Lu, J. Astra: General Interactive World Model with Autoregressive Denoising. *arXiv preprint arXiv:2512.08931* **2025**.
246. Zheng, S.; Yin, M.; Hu, W.; Li, X.; Shan, Y.; Fu, Y. VerseCrafter: Dynamic Realistic Video World Model with 4D Geometric Control. *arXiv preprint arXiv:2601.05138* **2026**.
247. Chen, K.; Liang, D.; Zhou, X.; Ding, Y.; Liu, X.; Wan, P.; Bai, X. Out of Sight but Not Out of Mind: Hybrid Memory for Dynamic Video World Models. *arXiv preprint arXiv:2603.25716* **2026**.
248. Sun, H.L.; Sun, Z.; Peng, H.; Ye, H.J. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 5158–5171.
249. Zhang, G.; Zhong, T.; Xia, Y.; Liu, M.; Yu, Z.; Li, H.; He, W.; She, D.; Wang, Y.; Jiang, H. Cmmcot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 12430–12438.
250. Yu, X.; Xu, C.; Zhang, G.; Chen, Z.; Zhang, Y.; He, Y.; Jiang, P.T.; Zhang, J.; Hu, X.; Yan, S. Vismem: Latent vision memory unlocks potential of vision-language models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 31544–31555.
251. Barraco, M.; Sarto, S.; Cornia, M.; Baraldi, L.; Cucchiara, R. With a little help from your own past: Prototypical memory networks for image captioning. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3021–3031.
252. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10578–10587.
253. Sarto, S.; Cornia, M.; Baraldi, L.; Nicolosi, A.; Cucchiara, R. Towards retrieval-augmented architectures for image captioning. *ACM Transactions on Multimedia Computing, Communications and Applications* **2024**, *20*, 1–22.
254. Thengane, V.; Khan, S.; Hayat, M.; Khan, F. Clip model is an efficient continual learner. *arXiv preprint arXiv:2210.03114* **2022**.
255. Zhu, Z.; Lyu, W.; Xiao, Y.; Hoiem, D. Continual learning in open-vocabulary classification with complementary memory systems. *arXiv preprint arXiv:2307.01430* **2023**.
256. Yu, J.; Zhuge, Y.; Zhang, L.; Hu, P.; Wang, D.; Lu, H.; He, Y. Boosting continual learning of vision-language models via mixture-of-experts adapters. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23219–23230.
257. Jha, S.; Gong, D.; Yao, L. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *Advances in neural information processing systems* **2024**, *37*, 129146–129186.
258. Huang, L.; Cao, X.; Lu, H.; Liu, X. Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 214–231.
259. Li, Y.; Pang, G.; Suo, W.; Jing, C.; Xi, Y.; Liu, L.; Chen, H.; Liang, G.; Wang, P. Coleclip: Open-domain continual learning via joint task prompt and vocabulary learning. *IEEE Transactions on Neural Networks and Learning Systems* **2025**.
260. Luo, M.L.; Zhou, Z.H.; Wei, T.; Zhang, M.L. LADA: Scalable label-specific CLIP adapter for continual learning. *arXiv preprint arXiv:2505.23271* **2025**.
261. Lu, H.; Zhang, X.; Moore, K.; Xue, J.; Yao, L.; Hengel, A.v.d.; Gong, D. Continual Learning on CLIP via Incremental Prompt Tuning with Intrinsic Textual Anchors. *arXiv preprint arXiv:2505.20680* **2025**.
262. Ding, Y.; Liu, L.; Tian, C.; Yang, J.; Ding, H. Don't stop learning: Towards continual learning for the clip model. *arXiv preprint arXiv:2207.09248* **2022**.
263. Yan, S.; Hong, L.; Xu, H.; Han, J.; Tuytelaars, T.; Li, Z.; He, X. Generative negative text replay for continual vision-language pretraining. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 22–38.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

264. Ni, Z.; Wei, L.; Tang, S.; Zhuang, Y.; Tian, Q. Continual vision-language representation learning with off-diagonal information. In Proceedings of the International Conference on Machine Learning. PMLR, 2023, pp. 26129–26149.
265. Zhu, H.; Wei, Y.; Liang, X.; Zhang, C.; Zhao, Y. Ctp: Towards vision-language continual pretraining via compatible momentum contrast and topology preservation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 22257–22267.
266. Garg, S.; Farajtabar, M.; Pouransari, H.; Vemulapalli, R.; Mehta, S.; Tuzel, O.; Shankar, V.; Faghri, F. Tic-clip: Continual training of clip models. *arXiv preprint arXiv:2310.16226* **2023**.
267. Roth, K.; Udandarao, V.; Dziadzio, S.; Prabhu, A.; Cherti, M.; Vinyals, O.; Hénaff, O.; Albanie, S.; Bethge, M.; Akata, Z. A Practitioner's Guide to Continual Multimodal Pretraining. *arXiv preprint arXiv:2408.14471* **2024**.
268. Dziadzio, S.; Udandarao, V.; Roth, K.; Prabhu, A.; Akata, Z.; Albanie, S.; Bethge, M. How to Merge Your Multimodal Models Over Time? In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 20479–20491.
269. Wu, B.; Shi, W.; Wang, J.; Ye, M. Synthetic data is an elegant gift for continual vision-language models. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 2813–2823.
270. Peng, T.; Liu, Y.; Yang, S.; Hong, Q.; Tian, Y. GNSP: Gradient Null Space Projection for Preserving Cross-Modal Alignment in VLMs Continual Learning. *arXiv preprint arXiv:2507.19839* **2025**.
271. Zhang, X.; Zhang, F.; Xu, C. Vqacl: A novel visual question answering continual learning setting. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19102–19112.
272. Lei, S.W.; Gao, D.; Wu, J.Z.; Wang, Y.; Liu, W.; Zhang, M.; Shou, M.Z. Symbolic replay: Scene graph as prompt for continual learning on vqa task. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2023, Vol. 37, pp. 1250–1259.
273. Marouf, I.E.; Tartaglione, E.; Lathuilière, S.; Van De Weijer, J. Ask and remember: A questions-only replay strategy for continual visual question answering. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 18078–18089.
274. Qian, Z.; Wang, X.; Duan, X.; Qin, P.; Li, Y.; Zhu, W. Decouple before interact: Multi-modal prompt learning for continual visual question answering. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2953–2962.
275. Huai, T.; Zhou, J.; Wu, X.; Chen, Q.; Bai, Q.; Zhou, Z.; He, L. Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In Proceedings of the Proceedings of the computer vision and pattern recognition conference, 2025, pp. 19608–19617.
276. He, J.; Guo, H.; Zhu, K.; Tang, M.; Wang, J. Continual instruction tuning for large multimodal models. *IEEE Transactions on Image Processing* **2026**.
277. Wang, Z.; Che, C.; Wang, Q.; Li, Y.; Shi, Z.; Wang, M. Smolora: Exploring and defying dual catastrophic forgetting in continual visual instruction tuning. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 177–186.
278. Guo, H.; Zeng, F.; Xiang, Z.; Zhu, F.; Wang, D.H.; Zhang, X.Y.; Liu, C.L. Hide-llava: Hierarchical decoupling for continual instruction tuning of multimodal large language model. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 13572–13586.
279. Ma, C.Y.; Lu, J.; Wu, Z.; AlRegib, G.; Kira, Z.; Socher, R.; Xiong, C. Self-monitoring navigation agent via auxiliary progress estimation. *arXiv preprint arXiv:1901.03035* **2019**.
280. Ma, C.Y.; Wu, Z.; AlRegib, G.; Xiong, C.; Kira, Z. The regretful agent: Heuristic-aided navigation through progress estimation. In Proceedings of the Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2019, pp. 6732–6740.
281. Hong, Y.; Wu, Q.; Qi, Y.; Rodriguez-Opazo, C.; Gould, S. Vln bert: A recurrent vision-and-language bert for navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2021, pp. 1643–1653.
282. Chen, S.; Guhur, P.L.; Schmid, C.; Laptev, I. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems* **2021**, *34*, 5834–5847.
283. Chen, S.; Guhur, P.L.; Tapaswi, M.; Schmid, C.; Laptev, I. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16537–16547.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

284. An, D.; Qi, Y.; Li, Y.; Huang, Y.; Wang, L.; Tan, T.; Shao, J. Bevbert: Multimodal map pre-training for language-guided navigation. *arXiv preprint arXiv:2212.04385* **2022**.
285. Wang, Z.; Li, X.; Yang, J.; Liu, Y.; Jiang, S. Gridmm: Grid memory map for vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF International conference on computer vision, 2023, pp. 15625–15636.
286. Zhan, Z.; Yu, L.; Yu, S.; Tan, G. Mc-gpt: Empowering vision-and-language navigation with memory map and reasoning chains. *arXiv preprint arXiv:2405.10620* **2024**.
287. Zeng, S.; Qi, D.; Chang, X.; Xiong, F.; Xie, S.; Wu, X.; Liang, S.; Xu, M.; Wei, X.; Guo, N. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. *arXiv preprint arXiv:2509.22548* **2025**.
288. De Vries, H.; Shuster, K.; Batra, D.; Parikh, D.; Weston, J.; Kiela, D. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367* **2018**.
289. Thomason, J.; Murray, M.; Cakmak, M.; Zettlemoyer, L. Vision-and-dialog navigation. In Proceedings of the Conference on Robot Learning. PMLR, 2020, pp. 394–406.
290. Zhu, Y.; Zhu, F.; Zhan, Z.; Lin, B.; Jiao, J.; Chang, X.; Liang, X. Vision-dialog navigation by exploring cross-modal memory. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10730–10739.
291. Roman, H.R.; Bisk, Y.; Thomason, J.; Celikyilmaz, A.; Gao, J. RMM: A recursive mental model for dialogue navigation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1732–1745.
292. Banerjee, S.; Thomason, J.; Corso, J. The robotslang benchmark: Dialog-guided robot localization and navigation. In Proceedings of the Conference on Robot Learning. PMLR, 2021, pp. 1384–1393.
293. Liu, S.; Hasan, A.; Hong, K.; Wang, R.; Chang, P.; Mizrachi, Z.; Lin, J.; McPherson, D.L.; Rogers, W.A.; Driggs-Campbell, K. Dragon: A dialogue-based robot for assistive navigation with visual language grounding. *IEEE Robotics and Automation Letters* **2024**, *9*, 3712–3719.
294. Han, L.; Min, H.; Hwangbo, G.; Choi, J.; Seo, P.H. DialNav: Multi-turn Dialog Navigation with a Remote Guide. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 8514–8523.
295. Wang, X.; Xiong, W.; Wang, H.; Wang, W.Y. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 37–53.
296. Anderson, P.; Shrivastava, A.; Parikh, D.; Batra, D.; Lee, S. Chasing ghosts: Instruction following as bayesian state tracking. *Advances in neural information processing systems* **2019**, *32*.
297. Vasudevan, A.B.; Dai, D.; Van Gool, L. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision* **2021**, *129*, 246–266.
298. Song, C.H.; Kil, J.; Pan, T.Y.; Sadler, B.M.; Chao, W.L.; Su, Y. One step at a time: Long-horizon vision-and-language navigation with milestones. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 15482–15491.
299. Wang, H.; Wang, W.; Liang, W.; Xiong, C.; Shen, J. Structured scene memory for vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2021, pp. 8455–8464.
300. Krantz, J.; Banerjee, S.; Zhu, W.; Corso, J.; Anderson, P.; Lee, S.; Thomason, J. Iterative vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 14921–14930.
301. Zheng, Q.; Liu, D.; Wang, C.; Zhang, J.; Wang, D.; Tao, D. Esceme: Vision-and-language navigation with episodic scene memory. *International Journal of Computer Vision* **2025**, *133*, 254–274.
302. Zhao, G.; Li, G.; Chen, W.; Yu, Y. Over-nav: Elevating iterative vision-and-language navigation with open-vocabulary detection and structured representation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16296–16306.
303. Liu, C.; Zhou, Z.; Zhang, J.; Zhang, M.; Huang, S.; Duan, H. Msnav: Zero-shot vision-and-language navigation with dynamic memory and llm spatial reasoning. In Proceedings of the ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2026, pp. 20112–20116.
304. Behbahani, S.; Chhatpar, S.; Zahrai, S.; Duggal, V.; Sukhwani, M. Episodic memory model for learning robotic manipulation tasks. *arXiv preprint arXiv:2104.10218* **2021**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

305. Shi, H.; Xie, B.; Liu, Y.; Sun, L.; Liu, F.; Wang, T.; Zhou, E.; Fan, H.; Zhang, X.; Huang, G. Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2508.19236* **2025**.
306. Li, R.; Guo, W.; Wu, Z.; Wang, C.; Deng, H.; Weng, Z.; Tan, Y.P.; Wang, Z. Map-vla: Memory-augmented prompting for vision-language-action model in robotic manipulation. *arXiv preprint arXiv:2511.09516* **2025**.
307. Lin, M.; Liang, X.; Lin, B.; Jingzhi, L.; Jiao, Z.; Li, K.; Ma, Y.; Liu, Y.; Zhao, S.; Zhuang, Y.; et al. EchoVLA: Robotic Vision-Language-Action Model with Synergistic Declarative Memory for Mobile Manipulation. *arXiv preprint arXiv:2511.18112* **2025**.
308. Li, Z.; Hu, B.; Shao, R.; Chen, G.; Jiang, D.; Xie, P.; Hao, J.; Nie, L. Global prior meets local consistency: Dual-memory augmented vision-language-action model for efficient robotic manipulation. *arXiv preprint arXiv:2602.20200* **2026**.
309. Torne, M.; Pertsch, K.; Walke, H.; Vedder, K.; Nair, S.; Ichter, B.; Ren, A.Z.; Wang, H.; Tang, J.; Stachowicz, K.; et al. Mem: Multi-scale embodied memory for vision language action models. *arXiv preprint arXiv:2603.03596* **2026**.
310. Li, H.; Shen, F.; Chen, D.; Yang, L.; Wang, X.; Shi, J.; Bing, Z.; Liu, Z.; Knoll, A. ReMem-VLA: Empowering Vision-Language-Action Model with Memory via Dual-Level Recurrent Queries. *arXiv preprint arXiv:2603.12942* **2026**.
311. Pan, C.; Yaman, B.; Nesti, T.; Mallik, A.; Allievi, A.G.; Velipasalar, S.; Ren, L. Vlp: Vision language planning for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14760–14769.
312. Ma, Y.; Wei, T.; Zhong, N.; Mei, J.; Hu, T.; Wen, L.; Yang, X.; Shi, B.; Liu, Y. LeapVAD: A leap in autonomous driving via cognitive perception and dual-process thinking. *IEEE Transactions on Neural Networks and Learning Systems* **2025**.
313. Luo, Z.; Qian, K.; Wang, J.; Luo, Y.; Miao, J.; Fu, Z.; Wang, Y.; Jiang, S.; Huang, Z.; Hu, Y.; et al. MTRDrive: Memory-Tool Synergistic Reasoning for Robust Autonomous Driving in Corner Cases. *arXiv preprint arXiv:2509.20843* **2025**.
314. Shah, S.; Mishra, A.; Yadati, N.; Talukdar, P.P. Kvqa: Knowledge-aware visual question answering. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33, pp. 8876–8884.
315. Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, 2019, pp. 3195–3204.
316. Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; Mottaghi, R. A-okvqa: A benchmark for visual question answering using world knowledge. In Proceedings of the European conference on computer vision. Springer, 2022, pp. 146–162.
317. Chen, Y.; Hu, H.; Luan, Y.; Sun, H.; Changpinyo, S.; Ritter, A.; Chang, M.W. Can pre-trained vision and language models answer visual information-seeking questions? In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 14948–14968.
318. Lin, W.; Byrne, B. Retrieval augmented visual question answering with outside knowledge. In Proceedings of the Proceedings of the 2022 conference on empirical methods in natural language processing, 2022, pp. 11238–11254.
319. Lin, W.; Chen, J.; Mei, J.; Coca, A.; Byrne, B. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems* **2023**, *36*, 22820–22840.
320. Long, X.; Ma, Z.; Hua, E.; Zhang, K.; Qi, B.; Zhou, B. Retrieval-augmented visual question answering via built-in autoregressive search engines. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 24723–24731.
321. Choi, C.; Lee, W.; Ko, J.; Rhee, W. Multimodal Iterative RAG for Knowledge Visual Question Answering. *arXiv e-prints* **2025**, pp. arXiv–2509.
322. Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; Zhou, M. Layoutlm: Pre-training of text and layout for document image understanding. In Proceedings of the Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, 2020, pp. 1192–1200.
323. Mathew, M.; Karatzas, D.; Jawahar, C. Docvqa: A dataset for vqa on document images. In Proceedings of the Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 2200–2209.
324. Mathew, M.; Bagal, V.; Tito, R.; Karatzas, D.; Valveny, E.; Jawahar, C. Infographicvqa. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1697–1706.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

325. Masry, A.; Do, X.L.; Tan, J.Q.; Joty, S.; Hoque, E. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In Proceedings of the Findings of the association for computational linguistics: ACL 2022, 2022, pp. 2263–2279.
326. Tito, R.; Karatzas, D.; Valveny, E. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition* **2023**, *144*, 109834.
327. Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; Hudelot, C.; Colombo, P. Colpali: Efficient document retrieval with vision language models. In Proceedings of the International Conference on Learning Representations, 2025, Vol. 2025, pp. 61424–61449.
328. Cho, J.; Mahata, D.; Irsoy, O.; He, Y.; Bansal, M. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952* **2024**.
329. Tanaka, R.; Iki, T.; Hasegawa, T.; Nishida, K.; Saito, K.; Suzuki, J. Vdocrag: Retrieval-augmented generation over visually-rich documents. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 24827–24837.
330. Lau, J.J.; Gayen, S.; Ben Abacha, A.; Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data* **2018**, *5*, 180251.
331. Liu, B.; Zhan, L.M.; Xu, L.; Ma, L.; Yang, Y.; Wu, X.M. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In Proceedings of the 2021 IEEE 18th international symposium on biomedical imaging (ISBI). IEEE, 2021, pp. 1650–1654.
332. Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; Xie, W. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415* **2023**.
333. Chen, X.; Shi, B.; Le, C.; Zhang, J.; Wang, K.; Gong, R.; Zhang, J.; Wang, C. Iterative Multimodal Retrieval-Augmented Generation for Medical Question Answering. *arXiv preprint arXiv:2604.27724* **2026**.
334. Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; Li, H. Drivelm: Driving with graph visual question answering. In Proceedings of the European conference on computer vision. Springer, 2024, pp. 256–274.
335. Yuan, J.; Sun, S.; Omeiza, D.; Zhao, B.; Newman, P.; Kunze, L.; Gadd, M. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828* **2024**.
336. Ding, W.; Cao, Y.; Zhao, D.; Xiao, C.; Pavone, M. Realgen: Retrieval augmented generation for controllable traffic scenarios. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 93–110.
337. Wen, C.; Lin, Y.; Qu, X.; Li, N.; Liao, Y.; Li, X.; Lin, H. Remote Sensing Retrieval-Augmented Generation: Bridging remote sensing imagery and comprehensive knowledge with a multimodal dataset and retrieval-augmented generation model. *IEEE Geoscience and Remote Sensing Magazine* **2026**.
338. Luo, Y.; Zheng, X.; Li, G.; Yin, S.; Lin, H.; Fu, C.; Huang, J.; Ji, J.; Chao, F.; Luo, J.; et al. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems* **2026**, *38*, 168008–168033.
339. Fan, Y.; Ma, X.; Su, R.; Guo, J.; Wu, R.; Chen, X.; Li, Q. Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2025, pp. 6342–6352.
340. Hao, H.; Han, J.; Li, C.; Li, Y.F.; Yue, X. Rap: Retrieval-augmented personalization for multimodal large language models. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 14538–14548.
341. Bai, H.; Wang, R.; Du, Z.; Zhao, Y.; Zhang, F.; Chen, H.; Zhu, X.; Zheng, B.; Zhao, X. Online-PVLM: Advancing Personalized VLMs with Online Concept Learning. *arXiv preprint arXiv:2511.20056* **2025**.
342. Feng, J.; Xu, B.; Chen, J.; Dai, M.; Wu, C.; Li, H.; Zeng, B.; Xie, Y.; Liang, H.; Lu, M.; et al. M2a: Multimodal memory agent with dual-layer hybrid memory for long-term personalized interactions. *arXiv preprint arXiv:2602.07624* **2026**.
343. Long, L.; He, Y.; Ye, W.; Pan, Y.; Lin, Y.; Li, H.; Zhao, J.; Li, W. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736* **2025**.
344. Zhu, X.; Chen, Y.; Tian, H.; Tao, C.; Su, W.; Yang, C.; Huang, G.; Li, B.; Lu, L.; Wang, X.; et al. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144* **2023**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2026 doi:10.20944/preprints202607.1539.v1

345. Wang, Z.; Cai, S.; Liu, A.; Jin, Y.; Hou, J.; Zhang, B.; Lin, H.; He, Z.; Zheng, Z.; Yang, Y.; et al. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *47*, 1894–1907.
346. Li, Z.; Xie, Y.; Shao, R.; Chen, G.; Jiang, D.; Nie, L. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. *Advances in neural information processing systems* **2024**, *37*, 49881–49913.
347. He, H.; Yao, W.; Ma, K.; Yu, W.; Dai, Y.; Zhang, H.; Lan, Z.; Yu, D. Webvoyager: Building an end-to-end web agent with large multimodal models. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 6864–6890.
348. Sarch, G.; Jang, L.; Tarr, M.J.; Cohen, W.W.; Marino, K.; Fragkiadaki, K. Vlm agents generate their own memories: Distilling experience into embodied programs of thought. *Advances in Neural Information Processing Systems* **2024**, *37*, 75942–75985.
349. Koh, J.Y.; Lo, R.; Jang, L.; Duvvur, V.; Lim, M.; Huang, P.Y.; Neubig, G.; Zhou, S.; Salakhutdinov, R.; Fried, D. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 881–905.
350. Jang, L.; Li, Y.; Zhao, D.; Ding, C.; Lin, J.; Liang, P.P.; Bonatti, R.; Koishida, K. Videowebarena: Evaluating long context multimodal agents with video understanding web tasks. In Proceedings of the International Conference on Learning Representations, 2025, Vol. 2025, pp. 36934–36958.
351. Wang, J.; Xu, H.; Jia, H.; Zhang, X.; Yan, M.; Shen, W.; Zhang, J.; Huang, F.; Sang, J. Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration. *Advances in Neural Information Processing Systems* **2024**, *37*, 2686–2710.
352. Wang, Z.; Xu, H.; Wang, J.; Zhang, X.; Yan, M.; Zhang, J.; Huang, F.; Ji, H. Mobile-agent-e: Self-evolving mobile assistant for complex tasks. *arXiv preprint arXiv:2501.11733* **2025**.
353. Agashe, S.; Han, J.; Gan, S.; Yang, J.; Li, A.; Wang, X. Agent s: An open agentic framework that uses computers like a human. In Proceedings of the International Conference on Learning Representations, 2025, Vol. 2025, pp. 22924–22946.
354. Qin, Y.; Ye, Y.; Fang, J.; Wang, H.; Liang, S.; Tian, S.; Zhang, J.; Li, J.; Li, Y.; Huang, S.; et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326* **2025**.
355. Xu, R.; Xiao, G.; Chen, Y.; He, L.; Peng, K.; Lu, Y.; Han, S. StreamingVLM: Real-Time Understanding for Infinite Video Streams. *CoRR* **2025**, *abs/2510.09608*, [2510.09608]. https://doi.org/10.48550/ARXIV.2510.09608.
356. Wang et al.. MMLongBench: Benchmarking Long-Context Vision-Language Models, 2025.
357. Wu, J.; Lyu, H.; Xia, Y.; Zhang, Z.; Barrow, J.; Kumar, I.; Mirtaheri, M.; Chen, H.; Rossi, R.A.; Dernoncourt, F.; et al. Personalized Multimodal Large Language Models: A Survey. *CoRR* **2024**, *abs/2412.02142*, [2412.02142]. https://doi.org/10.48550/ARXIV.2412.02142.
358. Lei, M.; Cai, H.; Que, B.; Cui, Z.; Tan, L.; Hong, J.; Hu, G.; Zhu, S.; Wu, Y.; Jiang, S.; et al. RoboMemory: A Brain-inspired Multi-memory Agentic Framework for Lifelong Learning in Physical Embodied Systems. *CoRR* **2025**, *abs/2508.01415*, [2508.01415]. https://doi.org/10.48550/ARXIV.2508.01415.
359. Driess et al.. PaLM-E: An Embodied Multimodal Language Model, 2023.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.