Closing the Loop: A Survey of Memory in Vision-Language Embodied Agents

Anonymous ACL submission

Abstract

Large language models and vision-language models deliver powerful reasoning but are fundamentally stateless—a limitation that becomes acute when these models are deployed as the core of agents operating in partially observable, temporally extended physical environments. Memory, the ability to acquire, organize, and retrieve information across time steps, is the bridge from instantaneous reasoning to sustained embodied competence. Existing surveys of LLM-based agents and embodied AI treat memory as one component among many; none analyzes it as a first-class design problem spanning the full perception–planning–action–learning loop. We survey over 80 recent works through a functional-role taxonomy organized around these four roles, analyzing each along shared design dimensions: what is stored, how it is maintained, and how it interfaces with foundation models. Cross-functional comparison reveals two systematic gaps: almost no system addresses when or what to forget, allowing memory to grow without bound; and memory quality degrades without detection, as stale maps, irrelevant retrieved experiences, and noisy self-reflections accumulate unchecked across episodes. Connecting these four roles into a single analytical cycle reveals that cross-role isolation compounds both gaps.

1 Introduction

Foundation models—large language models (LLMs), vision-language models (VLMs), and vision-language-action models (VLAs)—are increasingly deployed as the reasoning core of agents that perceive and act in physical environments (Duan et al., 2022; Yang et al., 2023). These models bring remarkable instantaneous capabilities: open-vocabulary recognition, commonsense reasoning about unseen situations, and zero-shot instruction following. Yet these capabilities vanish between time steps. A model that identifies every object in an unfamiliar kitchen and produces a multi-step cooking plan cannot, three actions later, recall where it set the eggs down. In the physical world, environments are partially observable, tasks unfold over extended horizons, and past actions reshape future possibilities. *Memory*, the ability to acquire, organize, and retrieve information across time, is what closes this gap, converting one-shot reasoning into sustained, grounded competence.

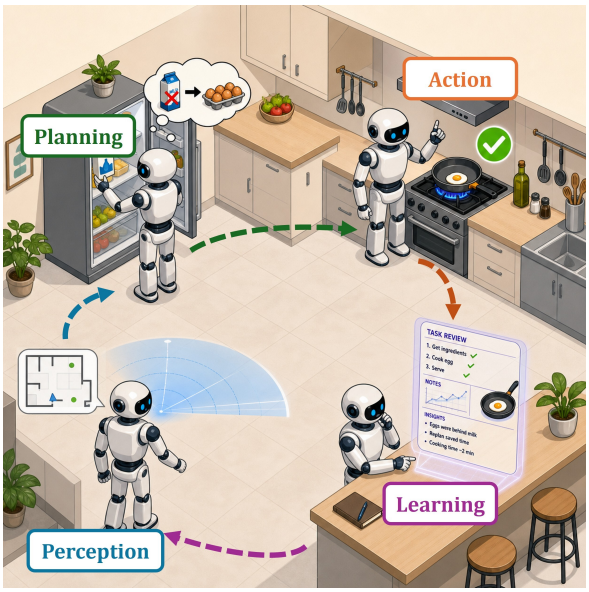

Figure 1: A household robot preparing breakfast illustrates the four memory roles: perceiving the kitchen layout, replanning when ingredients are misplaced, executing while tracking completed steps, and consolidating experience for future tasks.

The demands on such memory are multifaceted (Figure 1). A household robot preparing breakfast must accumulate spatial observations across rooms into a persistent map—memory *for perception*. When it discovers the eggs behind the milk rather than on the counter, it must draw on prior experience and commonsense kitchen knowledge to replan—memory *for planning*. It must track that the stove is already on and the pan already

placed, lest it re-execute completed steps, a failure mode shown to collapse task success to near zero (Liu et al., 2024)—memory *for action*. And when deployed in a new kitchen the next day, yesterday's skills and spatial priors should carry over rather than be relearned from scratch—memory *for learning*. These four roles form an agent loop (Figure 2) that recurs across navigation (Anderson et al., 2018), manipulation (Ahn et al., 2022), embodied question answering (Das et al., 2018), and open-world exploration (Wang et al., 2023a).

The rapid adoption of foundation models for embodied agents has produced a surge of memory-centric work across these task families, yet the resulting literature remains fragmented. Surveys on LLM-based agents (Wang et al., 2024; Xi et al., 2025) cover memory broadly but focus on text retrieval and tool-use settings, without addressing spatial reasoning or real-time control. Embodied AI surveys (Duan et al., 2022; Liu et al., 2025) treat memory as one component alongside perception and planning, without analyzing how foundation-model interfaces reshape memory design. More fundamentally, memory research within embodied AI is itself siloed: navigation communities build spatial maps, manipulation communities design skill libraries, and learning communities develop replay buffers, with shared design challenges going unrecognized.

This survey provides a unified analysis of these fragmented efforts. We analyze recent works through a functional-role taxonomy organized into four complementary roles—**Memory for Perception** (§3), **Planning** (§4), **Action** (§5), and **Learning** (§6)—whose definitions, organizing framework, task scope, and paper selection are developed in §2. Our contributions are threefold:

- We bridge fragmented memory research across navigation, manipulation, and open-world learning through a unified functional-role lens that spans the full perception–planning–action–learning loop, connecting mechanisms—maps, scene graphs, skill libraries, history encoders, and replay buffers—that serve analogous roles but are rarely compared.

- We propose a functional-role taxonomy that analyzes each role along shared design axes—what to store, how to maintain, and how to interface with foundation models—surfacing the core design tensions that govern each role.

- Through cross-functional comparison, we identify three recurring design patterns—the universality of retrieval-mediated memory, within-role architectural convergence alongside cross-role isolation, and the emergence of memory as the primary adaptation layer for frozen foundation models—and expose systematic gaps that motivate concrete open challenges.

2 The Agent Loop

Memory is a recurring concept wherever foundation models are deployed beyond static text, yet in physical environments the term spans mechanisms as different as spatial maps, KV caches, and skill libraries, often without a shared definition that enables cross-system comparison. Arriving at such a definition requires first understanding the distinctive context in which embodied memory operates. Foundation models serve embodied agents as modular planners (Ahn et al., 2022; Huang et al., 2022) or end-to-end VLA policies (Zitkovich et al., 2023; Kim et al., 2024), but both paradigms differ from text-based LLM agents in three ways that directly constrain memory design: inputs are continuous egocentric visual streams subject to sensor noise, requiring memory to encode spatial structure; outputs must map to physical actions under real-time constraints, imposing strict latency budgets; and the environment changes through the agent's own actions, demanding that memory stay consistent under closed-loop dynamics.

Embodied Memory. We define **embodied memory** as any mechanism through which an agent dynamically acquires, organizes, maintains, and retrieves information during interaction with a physical environment, to support perception, reasoning, or action across time steps. Two criteria are central. First, memory must be *dynamic*: constructed or updated as the agent interacts, not fixed before deployment. A foundation model's pretrained parameters encode vast world knowledge but do not change during interaction; they fall outside our scope, though when those parameters are updated online (e.g., via experience replay), the update mechanism itself constitutes memory. Second, memory must operate *across time steps*: past information influences future decisions, distinguishing memory from single-step perception.

Four Functional Roles. We organize this survey by **functional role**—the question that memory an-

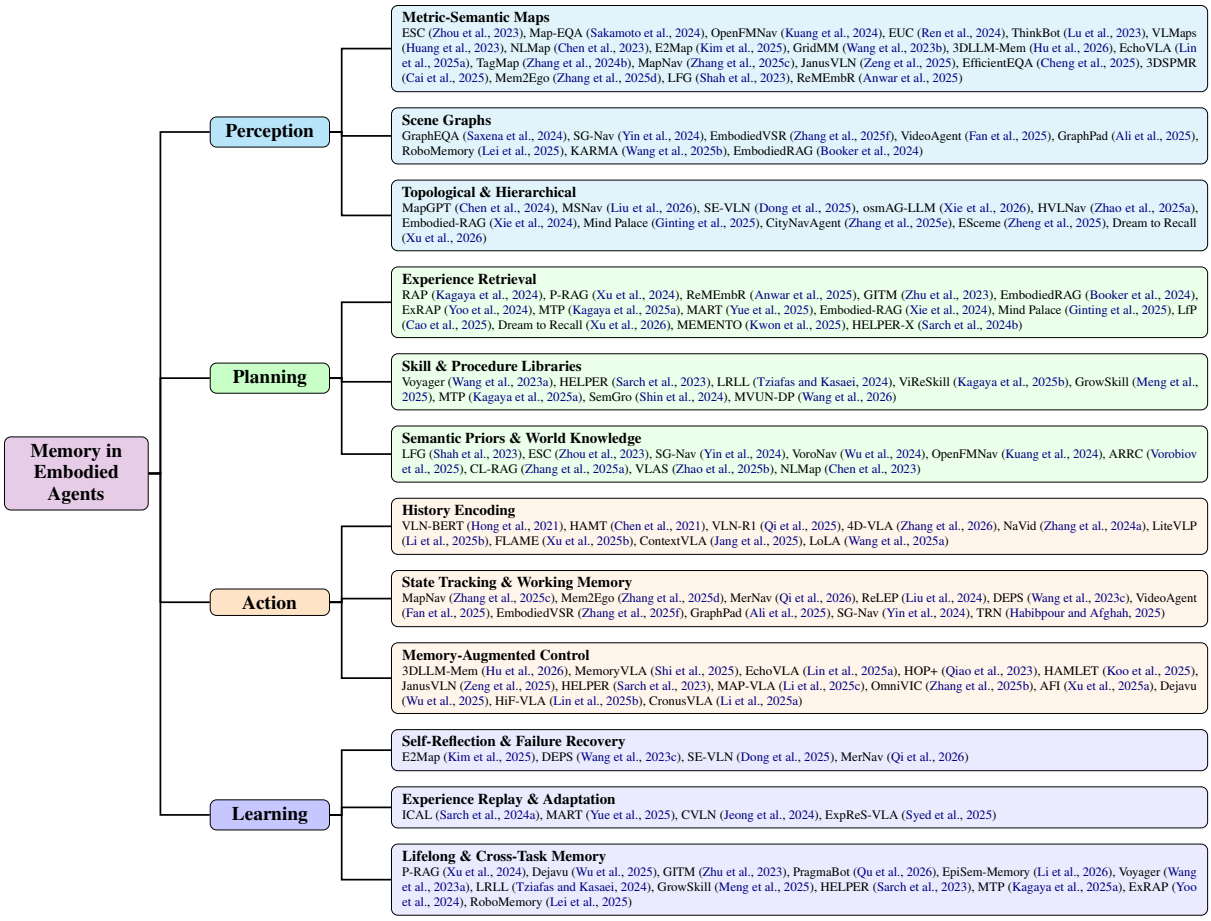

Figure 2: Complete taxonomy of surveyed works organized by functional role and sub-category. Methods appearing under multiple roles employ memory mechanisms that serve different functions.

swers within the agent's operational loop—rather than by representation type (maps, graphs, code) or task domain (navigation, manipulation). A functional taxonomy groups systems that serve the same purpose regardless of implementation, surfacing design tradeoffs obscured by surface-level similarities in representation or application.

Memory for Perception (§3) asks *what is the world like?*—accumulating observations into persistent world representations. **Memory for Planning** (§4) asks *what should I do?*—drawing on experience, procedures, and world knowledge beyond any single observation. **Memory for Action** (§5) asks *how do I execute this step?*—encoding interaction history into step-level decisions under tight latency. **Memory for Learning** (§6) asks *how do I improve over time?*—consolidating experience for cross-task and cross-environment transfer.

These four roles form a closed loop: perception feeds planning, which guides action, whose interaction experience learning distills into durable knowledge that refines all three preceding roles. Within each role, we analyze the method space along three shared dimensions—*what* information is stored, *how* it is written and maintained, and how it *interfaces* with the foundation model—surfacing design tradeoffs rather than enumerating individual systems. Across all four roles, this interface is predominantly linguistic: maps are serialized into text, graphs into structured descriptions, experiences into natural-language summaries, and skills retrieved via language queries—tying embodied memory design to core NLP challenges of grounding and serialization. Each work is categorized by the primary function its memory serves; a complete classification of all surveyed works is provided in Figure 2.

Task Scope. The memory mechanisms surveyed here span five task families—vision-language navigation, embodied question answering, object-goal navigation, open-world exploration, and language-conditioned manipulation—sharing three properties that make memory indispensable: multi-step interaction, partial observability, and language-

conditioned reasoning via foundation models (see Appendix C for benchmarks).

Corpus Construction. We collected papers from major venues in robotics, ML, CV, and NLP (2020–early 2026) using keyword search and snowball sampling. Inclusion requires a 3D or physical environment, an explicit memory mechanism persisting across time steps, and integration with a foundation model; we excluded pure text agents, classical SLAM without FM integration, and single-step perception systems. The final corpus contains 84 papers, each assigned to one or more functional roles based on the primary purpose of its memory mechanism; papers serving multiple roles appear under each relevant category in Figure 2.

3 Memory for Perception

When vision-language models are deployed in partially observable 3D environments, no single viewpoint reveals the full scene. Memory for perception—the first station in the agent loop—accumulates observations into persistent world representations that planning, action, and learning can draw on. We organize these representations by abstraction level: metric-semantic maps, scene graphs, and topological graphs.

Metric-Semantic Maps. The most prevalent form of spatial memory projects egocentric RGB-D streams into a persistent bird's-eye-view grid or 3D voxel structure (Zhou et al., 2023; Kuang et al., 2024). The *what-to-store* axis is governed by a tension between representation generality and computational cost: early systems maintain fixed-category semantic channels from trained detectors (Zhou et al., 2023), while open-vocabulary approaches store dense visual-language embeddings enabling arbitrary text queries at higher storage cost (Huang et al., 2023; Kim et al., 2025). More recent work stores dense visual features with instruction-conditioned retrieval via learned cross-attention (Hu et al., 2026; Lin et al., 2025a), or bypasses dense representations entirely by storing natural-language annotations that a VLM can "read" directly (Zhang et al., 2024b, 2025c). At the far end, JanusVLN (Zeng et al., 2025) eliminates explicit maps altogether, caching transformer key-value states in a fixed-size sliding window.

The *update* axis exposes a second tension: cumulative projection is simple but indiscriminate (Huang et al., 2023), while selective writing preserves quality but requires a relevance signal. Novelty-based keyframe filtering (Cai et al., 2025) reduces stored images by 16–31% without performance loss, and discrepancy-driven updates (Lin et al., 2025a) enable adaptation to object rearrangements. Foundation models increasingly guide what gets emphasized—VLM-derived relevance scores bias frontier selection (Ren et al., 2024).

The *foundation-model interface* axis reflects a tension between interpretability and integration efficiency, spanning explicit modalities—visual rendering (Zhang et al., 2025c), text serialization (Kuang et al., 2024), and function-call APIs (Zhang et al., 2024b)—and implicit integration via embedding retrieval (Huang et al., 2023) or direct attention (Zeng et al., 2025).

Scene Graphs. Per-cell maps capture *where* things are but not *how* they relate; scene graphs trade spatial precision for relational structure, organizing perception around discrete entities as nodes and relationships as edges (Saxena et al., 2024; Yin et al., 2024). Construction follows *perception-driven* bottom-up assembly from detectors (Saxena et al., 2024) or *reasoning-driven* FM-guided gap-filling (Ali et al., 2025). A key advantage is entity-level state tracking: action-conditioned updates maintain object states through interactions (Fan et al., 2025). Querying ranges from full-graph serialization (Wang et al., 2025b) to subgraph retrieval reducing tokens by an order of magnitude (Booker et al., 2024) and function-call APIs (Ali et al., 2025).

Topological and Hierarchical Representations. Neither metric detail nor entity relations are the bottleneck for large-scale navigation; *connectivity* is. Topological maps sacrifice geometric precision for scalability and LLM compatibility, abstracting environments into graphs of viewpoints and traversable edges (Chen et al., 2024; Liu et al., 2026). Semantic augmentation enriches nodes with VLM features or scene descriptions (Dong et al., 2025; Xie et al., 2026), and hierarchical clustering enables scalable traversal over large graphs (Xie et al., 2024). A unique strength is *cross-episode persistence*: nodes from successive trajectories merge by proximity into ever-richer models (Zhang et al., 2025e; Xu et al., 2026).

Design Patterns. No single representation dominates; the community is converging toward *hybrid, multi-granularity* designs that combine families

(Wu et al., 2024; Xie et al., 2026), though most systems still assume static environments (Jiang et al., 2025) and few address error propagation from upstream perception. This trajectory also poses an unresolved tension: representations are becoming increasingly FM-native (from dense features to language annotations to KV caches), but greater queryability comes at the cost of the geometric precision that makes spatial memory valuable in the first place.

4 Memory for Planning

Perceptual memory tells the agent what the world looks like, but not what to do about it. Effective planning requires three complementary knowledge sources that perception alone cannot provide: *what has worked before* (experience retrieval), *how to do things* (skill and procedure libraries), and *how the world generally works* (semantic priors from foundation-model knowledge or external knowledge bases).

Experience Retrieval. Retrieval-mediated planning maintains a store of past experiences and retrieves relevant entries at decision time (Kagaya et al., 2024; Anwar et al., 2025). The central tension in *what to store* is between fidelity and compactness: raw trajectories preserve rich context but are noisy (Kagaya et al., 2024); natural-language summaries are compact but risk discarding situation-specific detail (Anwar et al., 2025); and structured representations such as scene-graph entities (Booker et al., 2024) or planner code (Kagaya et al., 2025a) maximize reusability at the cost of construction effort.

A parallel tension governs retrieval indexing: embedding similarity can mislead (Kagaya et al., 2025a), and MART (Yue et al., 2025) addresses this with preference-trained retrieval that shifts from "what looks similar" to "what actually helps," while multimodal matching (Kagaya et al., 2024) and hierarchical traversal (Xie et al., 2024) offer intermediate solutions.

Retrieved experience enters planning via prompt injection (Kagaya et al., 2024; Cao et al., 2025), structured self-query loops (Booker et al., 2024), or latent-space fusion (Xu et al., 2026). A critical finding: MEMENTO (Kwon et al., 2025) shows that *increasing retrieved memories degrades performance*—noise from irrelevant entries overwhelms attention even when the gold-relevant memory is included, establishing retrieval *quality* over quantity as the binding constraint.

Skill and Procedure Libraries. Skill libraries externalize procedural knowledge as named, callable abstractions (Wang et al., 2023a; Tziafas and Kasaei, 2024). The central tradeoff is *skill granularity*: code-level skills are precise but brittle across environments; language-level specifications transfer readily but lack executability; policy-level skills handle continuous control but resist composition. *Code-level* skills store verified executable functions in embedding-indexed libraries, enabling continual discovery via LLM-driven curricula and zero-shot transfer to unseen environments (Wang et al., 2023a); extensions add failure-driven repair (Kagaya et al., 2025b), human-in-the-loop refinement (Meng et al., 2025), and cross-environment code adaptation (Kagaya et al., 2025a). *Language-level* skills store natural-language specifications in hierarchical ontologies, trading executability for cross-domain flexibility: SemGro (Shin et al., 2024) adaptively decomposes high-level skills into executable sub-skills via a multimodal executability critic. *Policy-level* skills store trained neural policies producing continuous actions, tightly coupled to the planner through structured documentation (Wang et al., 2026).

Crosscutting challenges include *discovery* (from manual specification to automated curricula (Wang et al., 2023a)), *composition* (predominantly LLM-mediated), and *transfer* (strongly mediated by abstraction level (Kagaya et al., 2025a)). Only LRLL (Tziafas and Kasaei, 2024) implements principled capacity management, clustering code by AST structure with $\sim 8\times$ compression.

Semantic Priors and World Knowledge. A third planning memory source encodes general knowledge about how the world works, but the choice of *what to store* here carries a reliability tension: FM priors provide zero-shot bootstrapping but encode training-data biases, while external knowledge bases are controllable but static and domain-specific. The most widely adopted approach treats the FM itself as a queryable knowledge store: LLM-derived co-occurrence scores serve as heuristic terms in frontier-based search (Shah et al., 2023; Zhou et al., 2023). When deployment-specific or personalized knowledge is needed, explicit external knowledge bases supply it: domain-specific manuals (Vorobiov et al., 2025), continuously updated sensor-derived documents (Zhang et al., 2025a), or voiceprint-keyed user pro-

files (Zhao et al., 2025b).

The *FM interface* varies accordingly: heuristic scores bias frontier selection (Shah et al., 2023), retrieved constraints restrict the planner's output space (Vorobiov et al., 2025), and grounding resolves referential ambiguity (Zhao et al., 2025b).

Design Patterns. The three memory types are complementary—experience retrieval contributes situational precedent, skill libraries supply reusable procedures, and semantic priors provide zero-shot bootstrapping—but no surveyed system jointly optimizes across all three, and their integration with perceptual memory (§3) remains almost entirely unidirectional (Booker et al., 2024). A likely reason is that the three sources operate on fundamentally different update time scales—experience grows per-episode, skills evolve through slow curation, and semantic priors are largely static—making a shared write policy difficult to design.

5 Memory for Action

Planning produces strategies; action must ground them in continuous control. Action-level memory encodes interaction history into step-level control decisions under the tightest latency constraints of any role. We organize by integration depth: implicit history encoding, explicit state tracking, and memory-augmented control.

History Encoding in End-to-End Models. The core tension is information retention versus computational cost, and three encoding strategies resolve it differently. *Recurrent compression* carries forward a fixed-size latent state (Hong et al., 2021), offering constant cost but risking information loss. *Full-sequence attention* retains the complete observation history as explicit tokens (Chen et al., 2021), with substantial gains over recurrent baselines on long-horizon settings at the cost of linear memory growth. *Selective sampling and compression*, the dominant paradigm since 2024, balances both through three mechanisms: temporal sampling that prioritizes recent frames while sparsely retaining earlier ones (Qi et al., 2025); token compression that reduces per-frame footprint by up to $\sim$60% (Zhang et al., 2024a); and attention windowing that restricts cross-modal scope to recent steps—which can *improve* performance by filtering distant noise (Xu et al., 2025b).

For VLA manipulation, a complementary axis is *fusion with proprioceptive state*: learned context tokens compress visual history for efficient inference (Jang et al., 2025), while state-aware gating modulates VLM representations to ground actions in robot state (Wang et al., 2025a).

State Tracking and Working Memory. Where implicit encoding decides *what to store*, explicit state tracking addresses *how to maintain* it—trading opacity for interpretability, with update frequency ranging from write-once spatial recording to active policies for task progress, object dynamics, and error detection. *Spatial state* is the most prevalent: annotated semantic maps with constant footprint (Zhang et al., 2025c), panoramic frontier projections (Zhang et al., 2025d), and hierarchical structures combining BEV maps with bounded observation caches (Qi et al., 2026). *Task progress* tracking proves surprisingly critical: ReLEP (Liu et al., 2024) demonstrates that removing a simple finished-steps memory causes the agent to repeatedly execute completed steps, collapsing success to near zero despite unchanged first-round performance. DEPS (Wang et al., 2023c) accumulates richer progress memory through describe–explain cycles that condition replanning. *Object and scene dynamics* require active maintenance: action-conditioned updates track objects through interactions (Fan et al., 2025), and demand-driven perception fills gaps with minimal API calls (Ali et al., 2025). *Error detection* closes the supervisory loop: anomaly detection overrides plans when the agent is stuck (Qi et al., 2026), credibility scoring abandons false-positive goal detections (Yin et al., 2024), and even a minimal action-history buffer reduces pathological oscillation by 10 SR points (Habibpour and Afghah, 2025).

Memory-Augmented Control Policies. How memory *interfaces with the foundation model* at action time admits a spectrum from deep to shallow. *Feature-level integration* treats memory as additional tokens the policy attends over, from expandable episodic banks (Hu et al., 2026) to dual perceptual-cognitive banks with similarity-based merging at capacity (Shi et al., 2025). *Lightweight modulation* compresses memory into fixed-cost signals—from single-token bottlenecks (Qiao et al., 2023) through moment-token windows adding $\sim$1.02$\times$ latency (Koo et al., 2025) to KV-cache retention reducing inference time by 69–90% (Zeng et al., 2025). *Prompt-level integration* injects memories into frozen FMs as in-context exemplars (Sarch et al., 2023; Li et al., 2025c). A growing

"plug-in memory" trend augments frozen VLAs with external correction: affordance fields escape "memory traps" (Xu et al., 2025a) and experience feedback networks generate action residuals (Wu et al., 2025), both without weight updates.

Design Patterns. The most striking pattern is a *navigation–manipulation divergence*: navigation drives spatial memory (maps, graphs) for discrete actions at ∼1–5 Hz, while manipulation favors temporal memory (compressed features, motion tokens) for continuous actions at 10–50 Hz—making the action space itself a first-class design variable. Tasks requiring both, such as mobile manipulation, must therefore manage two parallel memory streams at different time scales, a design problem that remains largely unaddressed.

6 Memory for Learning

Action generates a stream of interaction experience; learning memory determines what survives beyond the current episode. It accumulates and consolidates experience for cross-task and cross-environment transfer, tracing a pipeline from reflection through replay to lifelong accumulation.

Self-Reflection and Failure Recovery. Reflection mechanisms close a *diagnose–distill–memorize* loop, and their design varies along the same three dimensions that govern earlier roles. *What is stored* spans a depth spectrum: factual event records (Kim et al., 2025), causal failure attributions via describe–explain cycles (Wang et al., 2023c), and full strategy revisions that pinpoint the first erroneous choice (Dong et al., 2025). *How it is maintained* depends on triggering policy—event-driven (Kim et al., 2025), task-driven (Wang et al., 2023c), or continuous monitoring (Qi et al., 2026)—and on persistence scope, from ephemeral signals (Qi et al., 2026) to spatial cost fields (Kim et al., 2025) to vector-database entries enabling cross-episode retrieval (Dong et al., 2025). *How reflections interface with the FM*: corrected decision traces become few-shot examples for future episodes (Dong et al., 2025), while spatial cost fields bypass the language channel by modifying planner traversal costs (Kim et al., 2025). A core tension cuts across all three dimensions: grounded corrections are precise but non-transferable, while abstract reflections transfer but risk unreliable self-evaluation—SE-VLN shows experience repository quality can degrade as noisy reflections accumulate (Dong et al., 2025).

Experience Replay and Continual Adaptation. Two adaptation pathways trade off safety against depth of change. *Retrieval-based* approaches keep parameters frozen and adapt through the FM interface—distilling demonstrations into structured programs-of-thought (Sarch et al., 2024a) or training the retriever with preference pairs to shift from "what looks similar" to "what actually helps" (Yue et al., 2025). This pathway is forgetting-free but bounded by what in-context conditioning can express. *Fine-tuning-based* approaches update model weights directly, enabling deeper change—reproducing full behavioral distributions via logit-matching (Jeong et al., 2024) or leveraging *failed* trajectories as contrastive negatives (Syed et al., 2025)—but risk catastrophic forgetting. Both pathways share an under-addressed *selection* problem: strategies range from random sampling (Jeong et al., 2024) to success gating (Sarch et al., 2024a) to preference-based curation (Yue et al., 2025), but how buffer composition affects adaptation remains largely unstudied. The dominant trend is retrieval: most systems freeze parameters, reflecting the strong in-context learning of modern FMs and the practical difficulty of safe online updates.

Lifelong and Cross-Task Memory. What to store for the long term trades detail for transferability: raw interaction records (Xu et al., 2024) are faithful but bulky; natural-language summaries (Zhu et al., 2023) and decision rules (Li et al., 2026) compress but lose context; executable code skills (Wang et al., 2023a; Tziafas and Kasaei, 2024)—the dominant resolution—are compact and composable but assume well-defined APIs. Transfer operates through compositional reuse (Wang et al., 2023a), experiential analogy (Xu et al., 2024), and cross-environment code adaptation where an LLM retargets stored programs to new conventions (Kagaya et al., 2025a). Maintaining consistency as the store grows poses an orthogonal challenge: success gating (Wang et al., 2023a) provides a coarse write filter, timestamped knowledge graphs (Yoo et al., 2024) detect outdated entries, and incremental updates with conflict resolution (Lei et al., 2025) reconcile new against existing knowledge.

Design Patterns. The reflection–replay–lifelong pipeline is conceptually coherent but *practically fragmented*: no surveyed system connects all three

Table 1: Cross-functional comparison of memory design patterns across the four roles. Each role exhibits a distinctive core tension and signature finding.

Role	Core Tension	Signature Finding
Perception	Generality vs. cost	Converging toward hybrid representations
Planning	Fidelity vs. compactness	More retrieval can *hurt*
Action	Retention vs. latency	Navigation $\neq$ manipulation memory
Learning	Precision vs. transferability	Pipeline stages remain disconnected

stages so that a single failure can trigger reflection, update a replay buffer, and eventually be distilled into a reusable skill. A key bottleneck is format incompatibility: reflections produce natural-language traces, replay buffers store structured trajectory pairs, and lifelong stores accumulate executable code—each transition requires a lossy format conversion that no current system automates.

7 Discussion and Outlook

Cross-Functional Design Patterns. Table 1 summarizes each role's core tension and signature finding; three higher-level patterns emerge only from cross-section comparison. First, *retrieval-mediated memory has become the dominant paradigm*, appearing in all four roles—yet stored items, retrieval criteria, and integration mechanisms differ fundamentally across roles. Second, architectures are converging *within* each role but remain *isolated across* roles: information flows feedforward, with reverse updates strikingly rare (Kim et al., 2025). Third, a "*memory as the adaptation layer*" paradigm is emerging: most learning systems freeze model parameters, and plug-in memory modules augment frozen VLAs without weight updates—suggesting that memory is becoming the primary adaptation mechanism for embodied foundation models.

Pervasive Gaps. Cross-role comparison reveals two systematic deficiencies. First, *capacity management* is nearly absent: almost no surveyed system addresses when or what to forget, allowing memory to grow without bound. The few exceptions—novelty-based keyframe filtering in perception (Cai et al., 2025), AST-based code clustering in planning (Tziafas and Kasaei, 2024), KV-cache windowing in action (Zeng et al., 2025)—remain isolated, role-specific solutions with no shared curation principles, contrasting sharply with classical continual learning where data curation is well-studied (Aljundi et al., 2019). Second, *memory quality degrades without detection*: systems accumulate memories but rarely verify whether stored entries remain correct, current, or relevant. Maps become stale in dynamic environments (Jiang et al., 2025; Fan et al., 2025), irrelevant retrieved experiences can overwhelm useful ones and hurt planning performance (Kwon et al., 2025), and noisy self-reflections cause experience repositories to deteriorate over time (Dong et al., 2025). Cross-role isolation compounds both gaps: without shared curation policies, each role rediscovers capacity limits independently; without feedback across role boundaries, quality degradation in one role propagates undetected into downstream decisions.

Open Challenges. *Memory-specific evaluation.* No current benchmark isolates whether improvement stems from better construction, retrieval, or utilization of memory; memory-specific metrics and diagnostic probes would enable targeted optimization rather than end-to-end hill-climbing. *Cross-functional integration.* Whether a unified memory store or heterogeneous stores with learned routing offer the better path remains open; either direction requires interfaces that let learning outcomes reshape perception, planning, and action memories. *Dynamic and non-stationary environments.* Both gaps above are amplified in dynamic settings, where the environment changes independently of the agent; real-world deployment demands memory mechanisms that detect environmental change, invalidate stale entries, and reacquire information on the fly.

8 Conclusion

This survey has organized memory in vision-language embodied agents through a functional-role taxonomy—perception, planning, action, and learning—revealing two pervasive gaps: memory grows without bound, and its quality degrades without detection. The analytical cycle connecting all four roles reveals cross-role isolation as a compounding factor behind both gaps: without shared curation policies, each role rediscovers capacity limits independently; without feedback across role boundaries, quality degradation in one role propagates undetected into downstream decisions. We hope this survey guides the community toward embodied agents that remember what matters, forget what doesn't, and learn the difference.

Limitations

This survey focuses on vision-language embodied agents and may underrepresent memory mechanisms designed for non-visual modalities (e.g., tactile or auditory). Our functional-role taxonomy is one of several possible organizations; some works span multiple roles, and boundary assignment inevitably involves judgment calls. The literature search covers works available through early 2026; the rapid pace of the field means relevant systems may have appeared since. Finally, our design dimensions and gap analysis are derived from qualitative synthesis rather than validated through controlled experiments—developing such benchmarks is itself one of the open challenges we identify.

Potential Risks. Long-term embodied memory raises concerns beyond this survey's technical scope. Persistent spatial and interaction memories may record household layouts, personal routines, and object arrangements, creating privacy risks if stored without user consent or adequate access control. Stale or erroneous memories can cause unsafe physical actions—navigating into rearranged obstacles or re-executing dangerous manipulation steps. We encourage future work to incorporate privacy-preserving storage, user-controllable memory deletion, and safety-aware retrieval alongside the performance-oriented advances surveyed here.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, and 1 others. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Muhammad Qasim Ali, Saeejith Nair, Alexander Wong, Yuchen Cui, and Yuhao Chen. 2025. Graphpad: Inference-time 3d scene graph updates for embodied question answering. *arXiv preprint arXiv:2506.01174*.

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. 2019. Online continual learning with maximal interfered retrieval. *Advances in neural information processing systems*, 32.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. 2025. Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2838–2845. IEEE.

Meghan Booker, Grayson Byrd, Bethany Kemp, Aurora Schmidt, and Corban Rivera. 2024. Embodiedrag: Dynamic 3d scene graph retrieval for efficient and scalable robot task planning. *arXiv preprint arXiv:2410.23968*.

Zhongyi Cai, Yi Du, Chen Wang, and Yu Kong. 2025. Vision to geometry: 3d spatial memory for sequential embodied mllm reasoning and exploration. *arXiv preprint arXiv:2512.02458*.

Guanqun Cao, Ryan Mckenna, Erich Graf, and John Oyekan. 2025. Learn from the past: Language-conditioned object rearrangement with large language models. In *Pacific Rim International Conference on Artificial Intelligence*, pages 434–448. Springer.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*.

Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S Ryoo, Austin Stone, and Daniel Kappler. 2023. Open-vocabulary queryable scene representations for real world planning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11509–11522. IEEE.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. 2024. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9796–9810.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems*, 34:5834–5847.

Kai Cheng, Zhengyuan Li, Xingpeng Sun, Byung-Cheol Min, Amrit Singh Bedi, and Aniket Bera. 2025. Efficienteqa: An efficient approach to open-vocabulary embodied question answering. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 20357–20364. IEEE.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions

of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, and 1 others. 2020. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3164–3174.

Xiangyu Dong, Haoran Zhao, Jiang Gao, Haozhou Li, Xiaoguang Ma, Yaoming Zhou, Fuhai Chen, and Juan Liu. 2025. Se-vln: A self-evolving vision-language navigation framework based on multimodal large language models. *arXiv preprint arXiv:2507.13152*.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR.

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. 2022. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362.

Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. 2025. Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6342–6352. IEEE.

Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. 2022. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056.

Muhammad Fadhil Ginting, Dong-Ki Kim, Xiangyun Meng, Andrzej Reinke, Bandi Jai Krishna, Navid Kayhani, Oriana Peltzer, David D Fan, Amirreza Shaban, Sung-Kyun Kim, and 1 others. 2025. Enter the mind palace: Reasoning and planning for long-term active embodied question answering. *arXiv preprint arXiv:2507.12846*.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. 2019. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*.

Mobin Habibpour and Fatemeh Afghah. 2025. Think, remember, navigate: Zero-shot object-goal navigation with vlm-powered reasoning. *arXiv preprint arXiv:2511.08942*.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1643–1653.

Wenbo Hu, Yining Hong, Yanjun Wang, Leison Gao, Zibu Wei, Xingcheng Yao, Nanyun Peng, Yonatan Bitton, Idan Szpektor, and Kai-Wei Chang. 2026. 3dllm-mem: Long-term spatial-temporal memory for embodied 3d large language model. *Advances in Neural Information Processing Systems*, 38:67856–67884.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. 2023. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, and 1 others. 2022. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*.

Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1862–1872.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. 2020. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026.

Huiwon Jang, Sihyun Yu, Heeseung Kwon, Hojin Jeon, Younggyo Seo, and Jinwoo Shin. 2025. Contextvla: Vision-language-action model with amortized multi-frame context. *arXiv preprint arXiv:2510.04246*.

Seongjun Jeong, Gi-Cheon Kang, Seongho Choi, Joochan Kim, and Byoung-Tak Zhang. 2024. Continual vision-and-language navigation. *arXiv preprint arXiv:2403.15049*.

Jiajun Jiang, Yiming Zhu, Zirui Wu, and Jie Song. 2025. Dualmap: Online open-vocabulary semantic mapping for natural language navigation in dynamic changing scenes. *IEEE Robotics and Automation Letters*.

Tomoyuki Kagaya, Subramanian Lakshmi, Yuxuan Lou, Thong Jing Yuan, Jayashree Karlekar, Sugiri Pranata, Natsuki Murakami, Akira Kinose, and Yang You. 2025a. Memory transfer planning: Llm-driven context-aware code adaptation for robot manipulation. *arXiv preprint arXiv:2509.24160*.

Tomoyuki Kagaya, Subramanian Lakshmi, Anbang Ye, Thong Jing Yuan, Jayashree Karlekar, Sugiri Pranata, Natsuki Murakami, Akira Kinose, and Yang You. 2025b. Vireskill: Vision-grounded replanning with skill memory for llm-based planning in lifelong robot learning. *arXiv preprint arXiv:2509.24219*.

Tomoyuki Kagaya, Thong Jing Yuan, Yuxuan Lou, Jayashree Karlekar, Sugiri Pranata, Akira Kinose, Koki Oguri, Felix Wick, and Yang You. 2024. Rap: Retrieval-augmented planning with contextual memory for multimodal llm agents. *arXiv preprint arXiv:2402.03610*.

Chan Kim, Keonwoo Kim, Mintaek Oh, Hanbi Baek, Jiyang Lee, Donghwi Jung, Soojin Woo, Younkyung Woo, John Tucker, Roya Firoozi, and 1 others. 2025. E2map: Experience-and-emotion map for self-reflective robot navigation with language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12811–12817. IEEE.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, and 1 others. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, and 1 others. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Myungkyu Koo, Daewon Choi, Taeyoung Kim, Kyungmin Lee, Changyeon Kim, Younggyo Seo, and Jinwoo Shin. 2025. Hamlet: Switch your vision-language-action model into a history-aware policy. *arXiv preprint arXiv:2510.00695*.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 104–120. Springer.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412.

Yuxuan Kuang, Hai Lin, and Meng Jiang. 2024. Openfmnav: Towards open-set zero-shot object navigation via vision-language foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 338–351.

Taeyoon Kwon, Dongwook Choi, Hyojun Kim, Sunghwan Kim, Seungjun Moon, Beong-woo Kwak, Kuan-Hao Huang, and Jinyoung Yeo. 2025. Embodied agents meet personalization: Investigating challenges and solutions through the lens of memory utilization. *arXiv preprint arXiv:2505.16348*.

Mingcong Lei, Honghao Cai, Yuyuan Yang, Yimou Wu, Jinke Ren, Zezhou Cui, Liangchen Tan, Junkun Hong, Gehan Hu, Shuangyu Zhu, and 1 others. 2025. Robomemory: A brain-inspired multi-memory agentic framework for interactive environmental learning in physical embodied systems. *arXiv preprint arXiv:2508.01415*.

Hao Li, Shuai Yang, Yilun Chen, Xinyi Chen, Xiaoda Yang, Yang Tian, Hanqing Wang, Tai Wang, Dahua Lin, Feng Zhao, and 1 others. 2025a. Cronusvla: Towards efficient and robust manipulation via multi-frame vision-language-action modeling. *arXiv preprint arXiv:2506.19816*.

Haoxuan Li, Sixu Yan, Yuhan Li, and Xinggang Wang. 2025b. Towards fast, memory-based and data-efficient vision-language policy. *arXiv preprint arXiv:2503.10322*.

Ji Li, Jing Xia, Mingyi Li, and Shiyan Hu. 2026. Improving mllms in embodied exploration and question answering with human-inspired memory modeling.

Runhao Li, Wenkai Guo, Zhenyu Wu, Changyuan Wang, Haoyuan Deng, Zhenyu Weng, Yap-Peng Tan, and Ziwei Wang. 2025c. Map-vla: Memory-augmented prompting for vision-language-action model in robotic manipulation. *arXiv preprint arXiv:2511.09516*.

Min Lin, Xiwen Liang, Bingqian Lin, Liu Jingzhi, Zijian Jiao, Kehan Li, Yuhan Ma, Yuecheng Liu, Shen Zhao, Yuzheng Zhuang, and 1 others. 2025a. Echovla: Robotic vision-language-action model with synergistic declarative memory for mobile manipulation. *arXiv preprint arXiv:2511.18112*.

Minghui Lin, Pengxiang Ding, Shu Wang, Zifeng Zhuang, Yang Liu, Xinyang Tong, Wenxuan Song, Shangke Lyu, Siteng Huang, and Donglin Wang. 2025b. Hif-vla: Hindsight, insight and foresight through motion representation for vision-language-action models. *arXiv preprint arXiv:2512.09928*.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. 2023. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791.

Chenghao Liu, Zhimu Zhou, Jiachen Zhang, Minghao Zhang, Songfang Huang, and Huiling Duan. 2026. Msnav: Zero-shot vision-and-language navigation with dynamic memory and llm spatial reasoning. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 20112–20116. IEEE.

Siyuan Liu, Jiawei Du, Sicheng Xiang, Zibo Wang, and Dingsheng Luo. 2024. Long-horizon embodied planning with implicit logical inference and hallucination mitigation. *arXiv preprint arXiv:2409.15658*.

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2025. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*.

Guanxing Lu, Ziwei Wang, Changliu Liu, Jiwen Lu, and Yansong Tang. 2023. Thinkbot: Embodied instruction following with thought chain reasoning. *arXiv preprint arXiv:2312.07062*.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, and 1 others. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. 2022. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334.

Yuan Meng, Zhenguo Sun, Max Fest, Xukun Li, Zhenshan Bing, and Alois Knoll. 2025. Growing with your embodied agent: A human-in-the-loop lifelong code generation framework for long-horizon manipulation skills. *arXiv preprint arXiv:2509.18597*.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

Dekang Qi, Shuang Zeng, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, and Mu Xu. 2026. Mernav: A highly generalizable memory-execute-review framework for zero-shot object goal navigation. *arXiv preprint arXiv:2602.05467*.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9982–9991.

Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. 2025. Vln-r1: Vision-language navigation via reinforcement fine-tuning. *arXiv preprint arXiv:2506.17221*.

Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. 2023. Hop+: History-enhanced and order-aware pre-training for vision-and-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8524–8537.

Kaixian Qu, Guowei Lan, René Zurbrügg, Changan Chen, Christopher E Mower, Haitham Bou-Ammar, and Marco Hutter. 2026. A pragmatist robot: Learning to plan tasks by experiencing the real world. *IEEE Robotics and Automation Letters*.

Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, and 1 others. 2021. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*.

Allen Z Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. 2024. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*.

Koya Sakamoto, Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. 2024. Map-based modular approach for zero-shot embodied question answering. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10013–10019. IEEE.

Gabriel Sarch, Lawrence Jang, Michael J Tarr, William W Cohen, Kenneth Marino, and Katerina Fragkiadaki. 2024a. Vlm agents generate their own memories: Distilling experience into embodied programs of thought. *Advances in Neural Information Processing Systems*, 37:75942–75985.

Gabriel Sarch, Sahil Somani, Raghav Kapoor, Michael J Tarr, and Katerina Fragkiadaki. 2024b. Helper-x: A unified instructable embodied agent to tackle four interactive vision-language domains with memory-augmented language models. *arXiv preprint arXiv:2404.19065*.

Gabriel Sarch, Yue Wu, Michael Tarr, and Katerina Fragkiadaki. 2023. Open-ended instructable embodied agents with memory-augmented large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3468–3500.

Saumya Saxena, Blake Buchanan, Chris Paxton, Peiqi Liu, Bingqing Chen, Narunas Vaskevicius, Luigi Palmieri, Jonathan Francis, and Oliver Kroemer. 2024. Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering. *arXiv preprint arXiv:2412.14480*.

Dhruv Shah, Michael Robert Equi, Błażej Osiński, Fei Xia, Brian Ichter, and Sergey Levine. 2023. Navigation with large language models: Semantic guesswork as a heuristic for planning. In *Conference on Robot Learning*, pages 2683–2699. PMLR.

Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. 2025. Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2508.19236*.

Sangwoo Shin, Seunghyun Kim, Youngsoo Jang, Moontae Lee, and Honguk Woo. 2024. Semantic skill grounding for embodied instruction-following in cross-domain environments. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3354–3376.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020a. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2020b. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*.

Shahram Najam Syed, Yatharth Ahuja, Arthur Jakobsson, and Jeff Ichnowski. 2025. Expres-vla: Specializing vision-language-action models through experience replay and retrieval. *arXiv preprint arXiv:2511.06202*.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, and 1 others. 2021. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems*, 34:251–266.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR.

Georgios Tziafas and Hamidreza Kasaei. 2024. Lifelong robot library learning: Bootstrapping composable and generalizable skills for embodied control with language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 515–522. IEEE.

Eugene Vorobiov, Ammar Jaleel Mahmood, Salim Rezvani, and Robin Chhabra. 2025. Arrc: Advanced reasoning robot control-knowledge-driven autonomous manipulation using retrieval-augmented generation. *arXiv preprint arXiv:2510.05547*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Qianji Wang, Yongkui Liu, Zilu Zhu, Sichao Liu, Zinan Wang, Bin Zi, Lihui Wang, and Lin Zhang. 2026. An mvun-dp skill library-driven embodied ai robotic assembly method with llms. *Robotics and Computer-Integrated Manufacturing*, 100:103263.

Xiaofan Wang, Xingyu Gao, Jianlong Fu, Zuolei Li, Dean Fortier, Galen Mullins, Andrey Kolobov, and Baining Guo. 2025a. Lola: Long horizon latent action learning for general robot manipulation. *arXiv preprint arXiv:2512.20166*.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. 2023b. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 15625–15636.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023c. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*.

Zixuan Wang, Bo Yu, Junzhe Zhao, Wenhao Sun, Sai Hou, Shuai Liang, Xing Hu, Yinhe Han, and Yiming Gan. 2025b. Karma: Augmenting embodied ai agents with long-and-short term memory systems. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE.

Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. 2019. Embodied question answering in photorealistic environments with point cloud perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6659–6668.

Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. 2024. Voronav: Voronoi-based zero-shot object navigation with large language model. *arXiv preprint arXiv:2401.02695*.

Shaokai Wu, Yanbiao Ji, Qiuchang Li, Zhiyi Zhang, Qichen He, Wenyuan Xie, Guodong Zhang, Bayram Bayramli, Yue Ding, and Hongtao Lu. 2025. Dejavu: Towards experience feedback learning for embodied intelligence. *arXiv preprint arXiv:2510.10181*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. 2018. Gibson env: Real-world perception for embodied agents. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 9068–9079.

Fujing Xie, Sören Schwertfeger, and Hermann Blum. 2026. osmag-llm: Zero-shot open-vocabulary object navigation via semantic maps and large language models reasoning. *IEEE Robotics and Automation Letters*, 11(3):2426–2433.

Quanting Xie, So Yeon Min, Pengliang Ji, Yue Yang, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Ruslan Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. 2024. Embodied-rag: General non-parametric embodied memory for retrieval and generation. *arXiv preprint arXiv:2409.18313*.

Siyu Xu, Zijian Wang, Yunke Wang, Chenghao Xia, Tao Huang, and Chang Xu. 2025a. Affordance field intervention: Enabling vlas to escape memory traps in robotic manipulation. *arXiv preprint arXiv:2512.07472*.

Weiye Xu, Min Wang, Wengang Zhou, and Houqiang Li. 2024. P-rag: Progressive retrieval augmented generation for planning on embodied everyday task. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6969–6978.

Yunzhe Xu, Yiyuan Pan, and Zhe Liu. 2026. Dream to recall: Imagination-guided experience retrieval for memory-persistent vision-and-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yunzhe Xu, Yiyuan Pan, Zhe Liu, and Hesheng Wang. 2025b. Flame: Learning to navigate with multimodal llm in urban environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9005–9013.

Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. 2023. Foundation models for decision making: Problems, methods, and opportunities. *arXiv preprint arXiv:2303.04129*.

Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. 2024. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems*, 37:5285–5307.

Minjong Yoo, Jinwoo Jang, Wei-Jin Park, and Honguk Woo. 2024. Exploratory retrieval-augmented planning for continual embodied instruction following. *Advances in Neural Information Processing Systems*, 37:67034–67060.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. 2020. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR.

Junpeng Yue, Xinrun Xu, Börje F Karlsson, and Zongqing Lu. 2025. Mllm as retriever: Interactively learning multimodal retrieval for embodied agents. In *International Conference on Learning Representations*, volume 2025, pages 31551–31580.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, Xing Wei, and Ning Guo. 2025. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. *arXiv preprint arXiv:2509.22548*.

Bowen Zhang, Yuhang Jiang, Lingxiang Hu, Dun Li, and Qianqian Hu. 2025a. Cl-rag: A closed-loop multimodal retrieval-augmented generation architecture for robust human-robot control interaction. In *2025 WRC Symposium on Advanced Robotics and Automation (WRC SARA)*, pages 177–184. IEEE.

Heng Zhang, Wei-Hsing Huang, Gokhan Solak, and Arash Ajoudani. 2025b. Omnivic: A self-improving variable impedance controller with vision-language in-context learning for safe robotic manipulation. *arXiv preprint arXiv:2510.17150*.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and 1 others. 2026. 4d-vla: Spatiotemporal vision-language-action pretraining with cross-scene calibration. *Advances in Neural Information Processing Systems*, 38:33914–33937.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. 2024a. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*.

Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. 2025c. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13032–13056.

Lingfeng Zhang, Yuecheng Liu, Zhanguang Zhang, Matin Aghaei, Yaochen Hu, Hongjian Gu, Mohammad Ali Alomrani, David Gamaliel Arcos Bravo, Raika Karimi, Atia Hamidizadeh, and 1 others. 2025d. Mem2ego: Empowering vision-language models with global-to-ego memory for long-horizon embodied navigation. *arXiv preprint arXiv:2502.14254*.

Mike Zhang, Kaixian Qu, Vaishakh Patil, Cesar Cadena, and Marco Hutter. 2024b. Tag map: A text-based map for spatial reasoning and navigation with large language models. *arXiv preprint arXiv:2409.15451*.

Weichen Zhang, Chen Gao, Shiquan Yu, Ruiying Peng, Baining Zhao, Qian Zhang, Jinqiang Cui, Xinlei Chen, and Yong Li. 2025e. Citynavagent: Aerial vision-and-language navigation with hierarchical semantic planning and global memory. In *Proceedings*

of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31292–31309.

Yi Zhang, Qiang Zhang, Xiaozhu Ju, Zhaoyang Liu, Jilei Mao, Jingkai Sun, Jintao Wu, Shixiong Gao, Shihan Cai, Zhiyuan Qin, and 1 others. 2025f. Embodiedvsr: Dynamic scene graph-guided chain-of-thought reasoning for visual spatial tasks. *arXiv preprint arXiv:2503.11089*.

Guoyang Zhao, Yudong Li, Weiqing Qi, Kai Zhang, Bonan Liu, Kai Chen, Haoang Li, and Jun Ma. 2025a. Slam-free visual navigation with hierarchical vision-language perception and coarse-to-fine semantic topological planning. *arXiv preprint arXiv:2509.20739*.

Wei Zhao, Pengxiang Ding, Zhang Min, Zhefei Gong, Shuanghao Bai, Han Zhao, and Donglin Wang. 2025b. Vlas: Vision-language-action model with speech instructions for customized robot manipulation. In *International conference on learning representations*, volume 2025, pages 51676–51693.

Qi Zheng, Daqing Liu, Chaoyue Wang, Jing Zhang, Dadong Wang, and Dacheng Tao. 2025. Esceme: Vision-and-language navigation with episodic scene memory. *International Journal of Computer Vision*, 133(1):254–274.

Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. 2023. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *International Conference on Machine Learning*, pages 42829–42842. PMLR.

Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. 2021. Soon: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12689–12699.

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, and 1 others. 2023. 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*.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.

Appendix Roadmap

The appendix answers five reader-facing questions about methodology, evidence, and differentiation; it is not an independent literature review. Appendix A reports survey statistics; Appendix B records the corpus construction protocol; Appendix C summarizes benchmark adoption; Appendix D provides the extended design-dimension analysis that the main text compresses into per-role summaries; and Appendix E differentiates this survey from related surveys. The tables are intended as claim-support infrastructure: they record what design choice a paper instantiates and which main-text pattern it supports.

A Survey Statistics

Figure 3 shows the distribution of surveyed papers across the four functional roles. Planning accounts for the largest share, followed by Action and Perception, while Learning remains underexplored. Figure 4 presents the temporal distribution, revealing rapid recent growth across all categories.

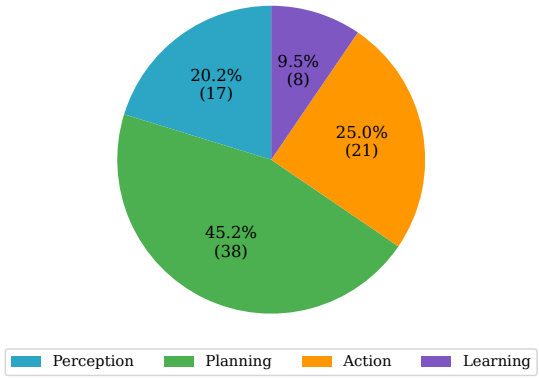

Figure 3: Distribution of surveyed papers by functional role.

B Survey Protocol

This appendix makes the survey's scope decisions auditable. It records how papers were selected and how each cited work was coded for the functional-role analysis.

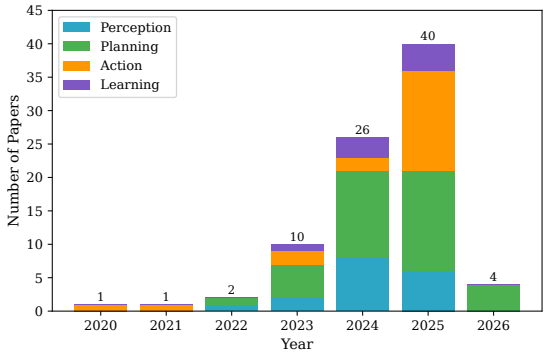

Figure 4: Number of surveyed papers by publication year and functional role.

Protocol field	Specification
Search scope	ACL Anthology, arXiv, IEEE Xplore, Springer, and major AI/robotics venues (CVPR, ICCV, ECCV, NeurIPS, ICML, ICLR, AAAI, CoRL, RSS, IROS, ICRA), 2020–early 2026.
Seed queries	Embodied memory; spatial memory embodied agent; scene graph navigation; experience retrieval embodied; skill library robot; memory-augmented VLA; lifelong embodied learning; VLN memory; retrieval-augmented embodied planning.
Expansion	Snowballing from seed papers and related surveys (Duan et al., 2022; Wang et al., 2024; Xi et al., 2025; Liu et al., 2025), plus forward/backward citation tracking.
Inclusion	(1) 3D or physical environment; (2) explicit memory persisting across time steps; (3) integration with an FM (LLM, VLM, or VLA). All three required.
Exclusion	Pure text agents; classical SLAM without FM; single-step perception; multi-agent coordination (unless memory-focused); benchmark-only papers.
Coding schema	Functional role; sub-category; what is stored; write/update policy; FM interface modality; capacity management (if any).

Scope note. This survey targets vision-language embodied agents; audio, tactile, and multi-agent memory systems are excluded. The functional-role

taxonomy is one of several possible organizations; boundary cases are resolved by the *primary question* the memory answers within the agent loop (§2).

C Benchmark Usage Across Surveyed Works

Table 2 summarizes the benchmarks and evaluation environments adopted by the surveyed works, grouped by task family. We report the number of papers that evaluate on each benchmark to indicate community adoption.

Table 2: Benchmarks and evaluation environments used by surveyed works, grouped by task family.

Task Family	Benchmark	#Papers
Vision-Language Navigation	R2R (Anderson et al., 2018)	13
	REVERIE (Qi et al., 2020)	9
	RxR (Ku et al., 2020)	8
	R2R-CE (Krantz et al., 2020)	6
	CVDN (Thomason et al., 2020)	4
	R4R (Jain et al., 2019)	2
	SOON (Zhu et al., 2021)	2
	VLN-CE (Krantz et al., 2020)	1
Object-Goal Navigation	HM3D ObjectNav (Ramakrishnan et al., 2021)	13
	MP3D ObjectNav (Chang et al., 2017)	5
	Gibson ObjectNav (Xia et al., 2018)	2
	RoboTHOR ObjectNav (Deitke et al., 2020)	2
Embodied Question Answering	OpenEQA (Majumdar et al., 2024)	4
	MP3D-EQA (Wijmans et al., 2019)	1
Robotic Manipulation	LIBERO (Liu et al., 2023)	15
	CALVIN (Mees et al., 2022)	5
	MetaWorld (Yu et al., 2020)	3
	RLBench (James et al., 2020)	3
	Franka Kitchen (Gupta et al., 2019)	2
Open-World Exploration	Minecraft (Fan et al., 2022)	3
Rearrangement	TidyHouse (Szot et al., 2021)	1
Instruction Following	ALFRED (Shridhar et al., 2020a)	6
	AI2-THOR (Kolve et al., 2017)	3
	TEACh (Padmakumar et al., 2022)	3
	ALFWorld (Shridhar et al., 2020b)	2
	DialFRED (Gao et al., 2022)	1
	CARLA (Dosovitskiy et al., 2017)	1
Other	ScanNet (Dai et al., 2017)	1
	Matterport3D (Chang et al., 2017)	1

D Extended Design-Dimension Analysis

This appendix answers the design-depth question: what specific choices do surveyed works make along the three shared analysis dimensions—*what* is stored, *how* it is maintained, and *how* it interfaces with the foundation model? The main text compresses these into per-role summaries; the tables below record the full evidence. Each table is organized by the design choice, with representative papers and the trade-off each choice entails.

D.1 Memory for Perception

Table 3 compares the three perceptual memory families on six design dimensions.

Table 4 records specific design choices within metric-semantic maps, the most prevalent family.

Table 3: Comparison of perceptual memory representations across six design dimensions.

Dimension	Metric Map	Scene Graph	Topological
Spatial precision	High (cell-level)	Medium (bbox)	Low (place-level)
Relation modeling	None	Explicit edges	Connectivity only
Storage scaling	$\propto$ explored area	$\propto$ entities	$\propto$ places visited
FM interface	Render / embed / serialize	Serialize / RAG	Text-native
Dynamic handling	Cell overwrite	Entity-level tracking	Mostly static
Cross-episode reuse	Difficult	Moderate	Natural

Table 4: Design choices for metric-semantic maps along three dimensions.

Dimension	Design choice	Representative works	Trade-off
What to store	Fixed-category labels	ESC (Zhou et al., 2023), Map-EQA (Sakamoto et al., 2024)	Low cost; cannot handle novel objects
	Open-vocab embeddings	VLMaps (Huang et al., 2023), NLMap (Chen et al., 2023)	Arbitrary queries; higher storage
	Dense visual features	3DLLM-Mem (Hu et al., 2026), EchoVLA (Lin et al., 2025a)	Rich grounding; requires cross-attention
	Language annotations	TagMap (Zhang et al., 2024b), MapNav (Zhang et al., 2025c)	VLM-readable; low storage ($\sim$0.17 MB)
How to update	Cumulative projection	VLMaps (Huang et al., 2023), ESC (Zhou et al., 2023)	Simple; redundant storage
	Novelty-based filtering	3DSPMR (Cai et al., 2025)	16–31% reduction; needs threshold
	Discrepancy-driven	EchoVLA (Lin et al., 2025a)	Adapts to change; higher compute
FM interface	Visual rendering	MapNav (Zhang et al., 2025c), Mem2Ego (Zhang et al., 2025d)	Interpretable; resolution-limited
	Text serialization	OpenFMNav (Kuang et al., 2024), LFG (Shah et al., 2023)	LLM-native; lossy conversion
	Implicit attention	JanusVLN (Zeng et al., 2025), 3DLLM-Mem (Hu et al., 2026)	Tight integration; opaque

Table 5: Design choices for planning memory across three sub-categories.

Dimension	Design choice	Representative works	Trade-off
Experience Retrieval			
What to store	Raw trajectories	RAP (Kagaya et al., 2024), P-RAG (Xu et al., 2024)	Rich context; noisy, expensive retrieval
	NL summaries	ReMEmbR (Anwar et al., 2025), PragmaBot (Qu et al., 2026)	Compact; loses situational detail
	Executable plans	MTP (Kagaya et al., 2025a), HELPER-X (Sarch et al., 2024b)	Direct reuse; assumes stable APIs
Retrieval method	Embedding similarity	MTP (Kagaya et al., 2025a), PragmaBot (Qu et al., 2026)	Simple; surface-level matching
	Preference-trained	MART (Yue et al., 2025)	"What helps" > "what matches"; needs training
	Hierarchical traversal	Embodied-RAG (Xie et al., 2024), Mind Palace (Gianting et al., 2025)	Scalable; requires tree construction
Skill Libraries			
Skill level	Code-level	Voyager (Wang et al., 2023a), LRLL (Tziafas and Kasaei, 2024)	Precise; environment-specific APIs
	Language-level	SemGro (Shin et al., 2024), GITM (Zhu et al., 2023)	Transferable; not directly executable
	Policy-level	MVUN-DP (Wang et al., 2026)	Continuous control; resists composition
Semantic Priors			
Knowledge source	FM internal priors	LFG (Shah et al., 2023), ESC (Zhou et al., 2023)	Zero-shot; training-data bias
	External KB	ARRC (Vorobiov et al., 2025), CL-RAG (Zhang et al., 2025a)	Controllable; static, needs curation

D.2 Memory for Planning

D.3 Memory for Action

Navigation vs. manipulation. Navigation operates over discrete actions at $\sim$1–5 Hz with episodes spanning hundreds of steps, driving predominantly *spatial* memory (maps, topological graphs, explored-area masks). Manipulation demands continuous actions at 10–50 Hz with temporally dense

Table 6: Design choices for action-level memory along the history encoding and control integration axes.

Dimension	Design choice	Representative works	Trade-off
History Encoding			
Encoding strategy	Recurrent compression	VLN-BERT (Hong et al., 2021)	$O(1)$ cost; information bottleneck
	Full-sequence attention	HAMT (Chen et al., 2021)	Full access; linear growth
	Selective sampling	VLN-R1 (Qi et al., 2025), NaVid (Zhang et al., 2024a)	Balanced; heuristic frame selection
Control Integration			
FM interface	Attention over memory bank	3DLLM-Mem (Hu et al., 2026), MemoryVLA (Shi et al., 2025)	Rich fusion; quadratic cost
	Lightweight modulation	HAMLET (Koo et al., 2025), HiF-VLA (Lin et al., 2025b)	$\sim$1$\times$ latency; compressed signal
	KV-cache retention	JanusVLN (Zeng et al., 2025), CronusVLA (Li et al., 2025a)	No recomputation; fixed window
	Prompt-level RAG	HELPER (Sarch et al., 2023), MAP-VLA (Li et al., 2025c)	No gradient; context-window bound
Plug-in correction	Affordance-based rollback	AFI (Xu et al., 2025a)	Escapes memory traps; needs 3D field
	Residual action correction	Dejavu (Wu et al., 2025)	Backbone-agnostic; additive only

episodes, driving predominantly *temporal* memory (compressed features, motion tokens, moment-token summaries). This divergence makes the action space a first-class design variable for memory architecture selection.

D.4 Memory for Learning

Table 7: Design choices for learning memory across the reflection–replay–lifelong pipeline.

Dimension	Design choice	Representative works	Trade-off
Self-Reflection			
Trigger policy	Event-driven	E2Map (Kim et al., 2025)	Immediate; misses gradual degradation
	Task-driven	DEPS (Wang et al., 2023c), SE-VLN (Dong et al., 2025)	Structured; delayed feedback
	Continuous monitoring	MerNav (Qi et al., 2026)	Catches drift; higher compute
Reflection depth	Factual recording	E2Map (Kim et al., 2025)	Precise; no causal insight
	Causal attribution	DEPS (Wang et al., 2023c)	Explanatory; FM accuracy-dependent
	Strategy revision	SE-VLN (Dong et al., 2025)	Full correction; error amplification risk
Experience Replay			
Adaptation pathway	Retrieval-based (frozen)	ICAL (Sarch et al., 2024a), MART (Yue et al., 2025)	No forgetting; bounded expressiveness
	Fine-tuning-based	CVLN (Jeong et al., 2024), ExpReS-VLA (Syed et al., 2025)	Deeper change; forgetting risk
Lifelong Memory			
What to accumulate	Raw interaction records	P-RAG (Xu et al., 2024), Dejavu (Wu et al., 2025)	Faithful; bulky, retrieval degrades
	NL summaries / rules	GITM (Zhu et al., 2023), Li et al. (Li et al., 2026)	Compact; loses context
	Executable code skills	Voyager (Wang et al., 2023a), LRLL (Tziafas and Kasaei, 2024)	Composable; assumes well-defined APIs
Capacity mgmt.	Consolidation-based	LRLL (Tziafas and Kasaei, 2024) ($\sim$8$\times$ compression)	Principled; complex pipeline
	Diversity-preserving dedup.	OmniVIC (Zhang et al., 2025b)	Simple; no semantic curation

The learning pipeline gap. The three stages are conceptually a pipeline but practically fragmented: reflective corrections are stored for retrieval but never replayed as training data; replay buffers are populated from task streams rather than reflective outputs; and lifelong skill libraries grow by verification gating without integrating reflective refinement. A key bottleneck is format incompatibility—reflections produce language traces, replay buffers store trajectory pairs, and lifelong stores accumulate executable code—making end-to-end pipeline integration an open challenge.

E Related Surveys and Differentiation

This appendix answers the differentiation question: what does this survey add beyond existing LLM-agent, embodied-AI, and foundation-model surveys? The organizing claim: existing surveys either catalogue agent architectures broadly or review embodied AI tasks comprehensively; none treats memory as the central organizing principle for vision-language embodied systems, analyzed through shared design dimensions across the full perception–planning–action–learning loop.

Table 8: Positioning relative to related surveys.

Survey	Year	Organizing principle	This survey's difference
Duan et al. (Duan et al., 2022)	2022	Embodied AI simulators and tasks	We organize by memory function, not task; analyze FM-memory interfaces absent from pre-FM surveys
Yang et al. (Yang et al., 2023)	2023	FM for decision making: problems, methods, opportunities	Broad FM scope; memory is one of many topics; no embodied-specific design analysis
Wang et al. (Wang et al., 2024)	2024	LLM-based autonomous agents: architecture and applications	Covers memory as one module; focuses on text agents; no spatial reasoning, real-time control, or closed-loop analysis
Xi et al. (Xi et al., 2025)	2025	LLM agents: capabilities and societal impact	Memory as short/long-term store for text agents; no embodied design dimensions or FM interface analysis
Liu et al. (Liu et al., 2025)	2025	Embodied AI: aligning cyber and physical spaces	Comprehensive embodied survey; treats memory as one component among perception, planning, and control

Novelty statement. This survey's contribution is not coverage breadth alone. It is the *functional-role memory framework*: the four-role taxonomy grounded in the agent loop (§2), the three shared design dimensions (what / how / interface) applied uniformly across roles (§3–§6), and the cross-functional gap analysis revealing capacity management and quality degradation as pervasive, under-addressed challenges (§7). Together, these elements explain *why* embodied memory is a first-class design problem that cuts across the perception–planning–action–learning loop, which existing surveys do not attempt.