INCEPTION: Aligning Agent and Human Memory Through Consolidation for Always-on AI Wearables

Anonymous Author(s)

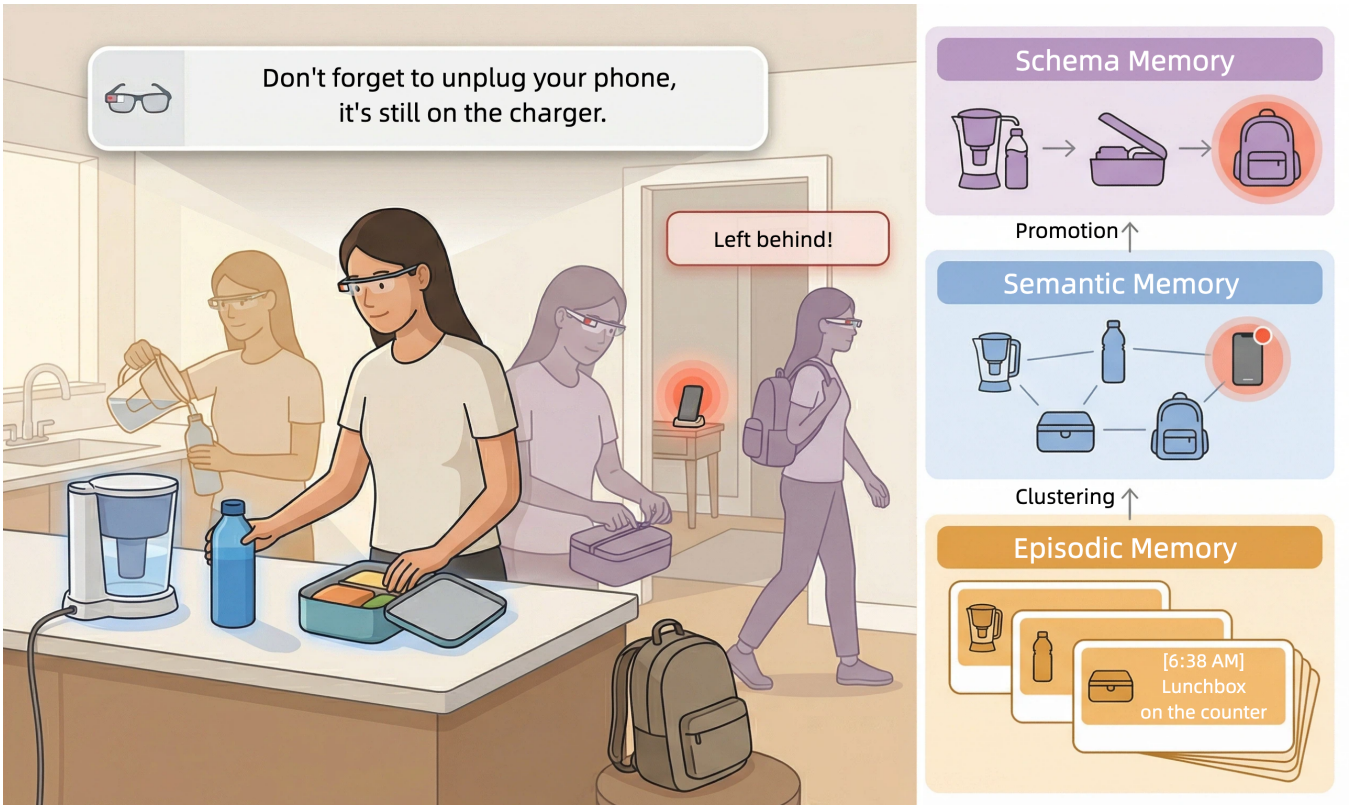

Figure 1: INCEPTION consolidates raw egocentric observations from always-on AI wearables into structured memory through three layers. Timestamped episodic traces of individual objects and events (bottom right) are compiled into semantic context profiles capturing co-occurrence patterns across situations (middle right), which are further promoted into behavioral schemas encoding the user's routines (top right). At runtime, the schema layer generates expectations which detects that a phone should be carried at departure but remains on the charger, and produces a grounded reminder (left).

Abstract

As AI wearables evolve from on-demand tools into everyday companions, interaction is shifting from discrete query-response turns to continuous micro-collaboration. Yet current wearable memory designs still treat interactions as isolated traces, making long-term support difficult under limited on-device storage, privacy-sensitive always-on capture, and lack of meaningful interaction pattern recognition. We introduce INCEPTION, a framework that structures memory for always-on AI wearables through three functional layers based on Complementary Learning Systems (CLS) theory: episodic memory for instance recognition, semantic memory for context recognition, and schema memory for routine recognition. INCEPTION operationalizes memory compression as two transitions, Episode-to-Semantic (E->S) and Semantic-to-Schema (S->K), which transform raw observations into reusable situational and behavioral knowledge. This framing enables top-down routine exception detection and bottom-up cross-context transfer, two complementary interaction modes that each traverse all three layers. INCEPTION provides a pathway toward realizing the alignment between human and always-on AI agents.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Keywords

Do, Not, Use, This, Code, Put, the, Correct, Terms, for, Your, Paper

ACM Reference Format:

Anonymous Author(s). 2018. INCEPTION: Aligning Agent and Human Memory Through Consolidation for Always-on AI Wearables. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Wearable assistants are now appearing in both research and products: from task guidance that answers in-situ questions [13], to consumer devices that support capture and assistance, such as Plaud NotePin [41] and Meta Ray-Ban glasses [36]. The interaction pattern between the user and the intelligent wearable system is progressively transitioning from occasional query-response exchanges to continuous micro-collaboration. Once the interaction becomes continuous, a memory system is necessary for the system to carry context, preferences, and routines over time.

This creates three problems for memory design. First is *storage*: always-on devices accumulate large volumes of first-person data, while on-device storage is limited, so simply retaining all traces is infeasible over long-term use [54]. Second is *privacy concerns*: A cloud-first alternative can potentially solve the capacity limitations, but always-on capture often records intimate personal contexts, people may be reluctant to share this data with a cloud service; otherwise, it becomes unclear who actually owns the resulting memory [4, 11, 16]. Third is *representation*: for continuous collaboration, solely retaining isolated traces is not sufficient. The assistant must also understand the situational context in which objects appear and recognize recurring behavioral routines, so that it can provide proactive support without requiring repeated explicit specification from the user [44]. Consider a user who wears smart glasses throughout their daily routine. The system should not only accumulate the traces that running shoes were seen near the door on weekday evening, but also recognize that the water bottle, running shoes, and door consistently co-occur in the early evening, then understand that this pattern reflects an evening jogging routine.

Most of the recent works are designed around query-response behavior, and they mainly focus on discrete snapshots and short-horizon episodic recall [7, 27, 30, 42, 64]. For instance, Satori [30] infers user intent and task state to provide immediate next-step guidance; Memoro [64] focuses on concise real-time memory augmentation. These contributions establish strong baselines for in-the-moment assistance. Yet as wearable assistants shift from session-based tools to always-on companions, the quality of assistance should improve with prolonged use: an agent that assists on day thirty should know more about the user than the one that assisted on day one. This requires the system to progressively compile interaction traces into a structured understanding of the user's contexts, preferences, and routines. The system should build a long-term memory that increasingly aligns with how the user themselves organizes their daily life. With this alignment, the agent becomes a collaborator that grows with the user over time.

In this paper, we argue that always-on assistance requires *memory consolidation*, the process by which fragmented experiences are progressively reorganized into structured knowledge [15]. Inspired by consolidation perspectives in cognitive science [28, 34, 57], INCEPTION defines three memory layers, each operating at a distinct level of recognition:

- *Episodic memory* (instance recognition): identifies and tracks specific objects and events — what was seen, where, and when. Example: "your blue water bottle was last seen near the fountain at 7:15 PM."
- *Semantic memory* (context recognition): understands situational contexts and what typically belongs in them — the relationships between objects, locations, times, and activities. Example: "you usually bring the water bottle with you when jogging after dinner."
- *Schema memory* (routine recognition): recognizes recurring behavioral patterns and routines that span across contexts. Example: "you jog every evening after dinner."

The framework uses two compression transitions:

- *E->S* (Episode-to-Semantic): aggregates instance-level observations into contextual knowledge — learning what objects, locations, and actions co-occur within recognizable situations.
- *S->K* (Semantic-to-Schema): compiles contextual regularities into stable behavioral routines that generalize across settings.

This layered recognition structure enables two complementary interaction modes, each traversing all three layers:

(1) *Routine exception detection* (Top-down): a recognized routine (*Schema*) generates expectations about what should be present in the current context (*Semantic*), which are checked against actual observations (*Episodic*). For example, the system knows the user jogs after dinner (*Schema*), expects the water bottle to be present during jogging (*Semantic*), notices it is absent (*Episodic*), and proactively reminds the user that the bottle was last seen near the fountain.

(2) *Cross-context transfer* (Bottom-up): a current observation (*Episodic*) is matched to a known situational context (*Semantic*), which activates a routine from a different setting (*Schema*). For example, the system recognizes the user's cast-iron pan at a campsite (*Episodic*), understands this pan typically appears in dinner-cooking contexts (*Semantic*), and suggests the user could cook steak as they usually do at home (*Schema*).

To bridge framework and practice, we instantiate INCEPTION in an AI assistant system. We evaluate INCEPTION through two complementary methods: an automated ablation study using LLM-as-judge scoring across multiple stochastic runs, and a within-subjects user study (N=12) comparing INCEPTION against an episodic baseline. Our results reveal a functional boundary in memory-augmented

assistance: consolidation uniquely enables inference tasks that require reasoning from compiled routines where neither raw episodic retrieval nor partial consolidation can succeed. This boundary emerges consistently across both the automated ablation study and the user study.

2 Related Work

Our work draws on three research threads: proactive wearable assistance, egocentric memory augmentation, and cognitive theories of memory consolidation. Below we review each area and position Inception's contribution relative to the current state of the art.

2.1 Intelligent Wearable Assistants

Intelligent wearable assistants are rapidly diversifying across form factors, from Mixed Reality headsets [2, 35] and smart glasses [36] to pendant and clip-on wearables [24, 41]. Despite this diversity, all of these systems face a shared design tension: effective proactive assistance requires balancing the expected value of intervening against the cost of disrupting the user [26], and user tolerance for such initiative is highly context-dependent [37].

Recent research systems have explored specific facets of this design problem across form factors . ProMemAssist [42] models the user's working memory to estimate cognitive availability, achieving selective delivery from a large pool of candidate messages in a within-subject evaluation. Memoro [64] takes a different angle on minimizing conversational disruption by using its queryless, concise-response strategy to reduce output length while maintaining conversational quality. Persistent Assistant [7] and Sensible Agent [29] emphasize context-adaptive interaction loops that adapt to the user's current social and physical context. Satori [30] forecasts user task actions by modeling intention through a belief-desire-intention framework and presents proactive guidance.

These contributions collectively advance the timing and content to deliver across a range of device categories and interaction contexts. However, each system's decisions are grounded in the present moment, with no mechanism to carry forward what the system has learned from one interaction to enrich the next.

2.2 Egocentric and Object-Centric Memory Augmentation

In the specific area of memory augmentation, several systems have demonstrated that language-encoded and context-augmented retrieval can support personal memory question-answering. AiGet leverages sensing data from AR glasses to construct a knowledge library of the user's daily life and provide proactive interaction [6]. Encode-Store-Retrieve converts egocentric video into language-encoded memory units stored in a vector database, reporting strong performance on episodic recall benchmarks [47]. OmniQuery extends this by integrating contextual information across interconnected memories using metadate of media files [31]. Memento combines proactive resurfacing with context recognition in AR, triggering recalled memories in everyday settings [27]. Unobtrusive Physical AI [23] augments everyday objects with context-aware intelligence and robotic motion.

As these systems aim to become always-on companions, they must also decide what knowledge to accumulate, compress, and carry forward across interactions to have richer meaning [38, 59]. Dasgupta and Gershman frame memory itself as a computational resource with explicit storage and retrieval costs [10], implying that the quality of proactive assistance depends more on the structure of what it has learned to remember for more meaningful interactions with the users. Inception complements this foundation with explicit consolidation stages that progressively restructure raw observations into higher-level representations.

2.3 Cognitive Foundations of Memory Consolidation

Inception's architecture draws on cognitive and neuroscience theories of how biological memory organizes, compresses, and restructures experience. We review the theoretical threads that motivate our design.

Complementary learning systems. CLS theory shows that biological memory separates rapid encoding of specific episodes from slower extraction of statistical regularities [28, 34]. Consolidation research further shows that memory traces are actively transformed, not passively stored [15, 45, 57], and that this transformation improves future prediction [49]. These findings motivate Inception's separation into episodic, semantic, and schematic layers with selective promotion between them.

Episodic memory and event-specific encoding. Instance-level memories are organized by spatiotemporal context: what was experienced is encoded with where and when it occurred, and this binding determines both retention and later access [8, 43, 47, 52, 53]. This motivates Inception's episodic layer as an instance recognition system that preserves observations bound to their spatiotemporal metadata.

Personal semantics and scene-object binding. Knowledge about one's own life is more abstract than episodic recollection yet more self-specific than generic world knowledge [44]. Scene grammar research shows that object memory is structured by relational and spatial context, making violations of expected object-scene relations perceptually salient [5, 14, 25]. These findings motivate Inception's relational features in semantic profiles and its exception-detection mechanisms.

Schema theory. Schemas are knowledge structures built from repeated experience that both accelerate encoding of expected events and flag deviations — including triggering reconsolidation when prediction errors are sustained [3, 18–20, 48, 51]. This dual role directly motivates Inception's schema layer and its top-down exception detection: compiled routines generate expectations about what should be present, and mismatches trigger proactive assistance.

While these principles have been studied individually, few interactive systems have operationalized them as a unified memory architecture for always-on wearable assistance. Inception translates progressive memory consolidation into a systems-and-interaction design.

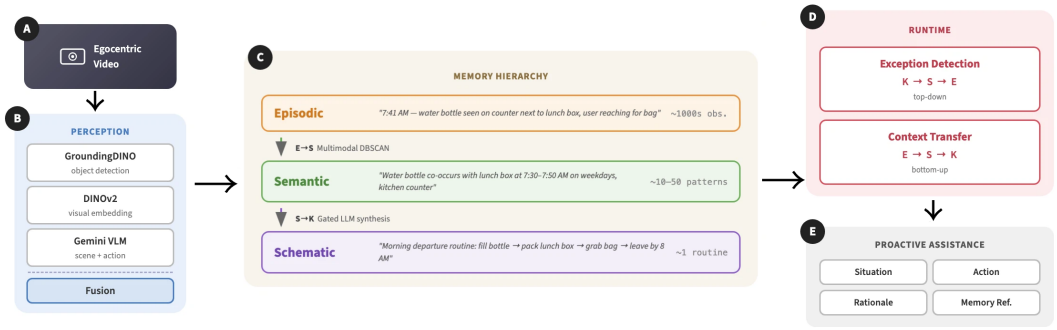

Figure 2: Inception Workflow. (A): Egocentric video feed. (B): A multi-modal perception pipeline detects objects, computes visual embeddings, and describes scenes, fusing results into a structured context. (C): A three-tier memory hierarchy consolidates observations: raw episodic observations are clustered into semantic co-occurrence patterns, which are synthesized into stable schematic routines. (D): At runtime, the engine traverses the hierarchy in two modes: top-down (K→S→E) to detect missing items by matching context against routines, and bottom-up (E→S→K) to transfer learned routines to new environments. (E): Structured proactive assistance is delivered to the user.

3 Designing Inception

This section describes the design and implementation of Inception. We first present the design goals that motivated architectural decisions, then describe how raw egocentric observations are compiled into higher-level memory representations.

3.1 Design Goals

The following design goals address the central challenge of aligning agent memory with the user's long-term behavior.

DG1: Functionally partitioned memory. Episodic traces, situational profiles, and behavioral routines serve distinct functions and degrade when conflated. The architecture must partition memory by function, with evidence-thresholded promotion between layers.

DG2: Progressive abstraction. An assistant that retains only raw traces has the same capacity on day thirty as on day one. The system requires an explicit compilation path from instance observations to contextual knowledge to behavioral routines.

DG3: Exception-sensitive consolidation. Not all regularities should be promoted. When recent observations deviate from a compiled pattern, promotion should be blocked until the pattern re-stabilizes.

DG4: Context-portable abstraction. Cross-context transfer requires compiled knowledge that generalizes beyond surface features. Schemas must encode functional roles and co-occurrence structure so that recognizing a familiar item in a new setting can activate associated routines.

3.2 System Architecture

We instantiate these design goals as a staged perceive-compile-interact architecture (Figure 2) with four components: (1) a continuous episodic perception pipeline that continuously captures object identity, location, action, and contextual metadata from the user's egocentric perspective (DG5); (2) a periodic semantic compilation stage (E→S) that periodically clusters co-occurring episodic observations into situational context profiles (DG2, DG3); (3) a periodic schema compilation stage (S→K) that promotes stable contextual regularities into behavioral routines through conservative multi-criterion gating (DG1, DG3, DG4); and (4) a runtime interaction engine that traverses all three layers in either exception detection or cross-context transfer.

The system follows a hybrid deterministic-plus-LLM pattern. Recent works on guidance and knowledge ecosystem for LLMs [9, 60, 61] demonstrate that structured, deterministic pipelines can be effectively augmented with language model capabilities. In Inception, deterministic components handle gating, scoring, distance computation, and state management. LLM components contribute natural-language pattern naming, routine synthesis, and contextual summarization. Deterministic fallbacks ensure the pipeline remains functional when LLM outputs are absent or malformed.

3.2.1 Tier 1: Episodic Perception Pipeline. The episodic layer must continuously encode observations with enough richness to support both runtime retrieval and downstream E→S compilation. This creates a few tensions: rich scene understanding requires expensive inference from vision language models (VLMs), but always-on capture demands lightweight processing. To address these tensions, three strategies are adopted to balance perceptual richness with processing efficiency.

Dual-layer perception: Inception uses two processing layers at different cadences. DINOv2 ViT-S/14 [39], a vision encoder runs on every sampled frame, producing compact embeddings suitable for fast similarity comparison. A VLM (Gemini-2.5-flash) [21] generates rich natural-language scene descriptions, but only when the visual encoder indicates a scene transition so that it can avoid per-frame VLM latency costs.

Item discovery: The system discovers tracked items through a VLM-first dual-modality pipeline. The VLM analyzes each scene-changed frame using an open-ended prompt and returns identified objects and actions without a predefined object list. These identified objects then drive GroundingDINO [32] localization for precise bounding boxes. Detections are confirmed only when both modalities agree: the VLM identifies an object by name, and the visual encoder localizes and encodes it. Confirmed detections are matched

against a running item registry; new items are created automatically on first observation, and their canonical embeddings are updated as running averages across re-observations.

Each episodic observation stores: object identity, visual embedding, VLM description, scene context, action, location, timestamp, and confidence, which can provide both structured similarity for compilation and natural-language grounding for interaction.

3.2.2 Tier 2: Episode-to-Semantic Compilation (E→S). The E→S transition discovers which observations regularly co-occur in recognizable situations, transforming raw episodes into contextual knowledge. Consider a user observed over three mornings: each day around 6:45 AM, the system separately detects a water bottle near the door, running shoes by the entrance, and the user packing a bag. Individually, these are isolated sightings. The E→S stage asks: Which observations belong to the same situational context?

Multimodal distance clustering: A composite distance between pairs of episodic observations across four channels is computed:

$$d(e_i, e_j) = \frac{\sum_{c \in C} w_c \cdot d_c(e_i, e_j)}{\sum_{c \in C} w_c}$$

where $C \subseteq \{time, object, background, action\}$ is the set of available channels for both observations, and d_c is the cosine distance in each channel: d_{time} captures temporal proximity via cyclical time-of-day embeddings, d_{object} compares DINOv2 visual embeddings for object identity, $d_{\text{background}}$ compares scene context embeddings, and d_{action} compares action descriptions. We weight object identity most heavily, as co-occurrence of the same items is the strongest signal for recognizing a recurring situation in our target scenarios. Normalizing by available weights ensures graceful degradation when an observation lacks a particular modality.

In the morning routine example, the water bottle's repeated appearance near the door at similar times, alongside the same running shoes, produces small distances across all four channels — clustering these observations together. A one-time co-occurrence with an umbrella on a rainy morning would not cluster, because its temporal and object-identity distances to the jogging observations are large.

Density-based grouping with support thresholds: A density-based clustering ($\epsilon = 0.45$, $minPts = 3$) is applied over the resulting distance matrix. Only groups meeting at least 3 observations are promoted. Observations that do not reach cluster density remain as episodic evidence.

Semantic profiles: Each promoted cluster produces a situational context profile containing temporal distributions, location distributions, typical co-occurring objects, typical actions, support count, and confidence. In our evaluation (Section 4), the E→S stage compressed 73 episodic observations into 5 semantic patterns — a 14.6:1 reduction that retains the contextual structure needed for downstream schema compilation.

3.2.3 Tier 3: Semantic-to-Schema Compilation (S→K). Not every semantic regularity should become a routine. The S→K stage determines which semantic patterns have enough evidence and consistency to warrant promotion to behavioral routines.

Conservative two-stage promotion: First, a deterministic gate evaluates each candidate pattern against multiple criteria simultaneously: sufficient total observations (≥ 5), sufficient active days (≥ 2), high mean confidence (≥ 0.55), low exception ratio (≤ 0.7), and sufficient importance (≥ 0.35) (DG3). A pattern that is frequent but inconsistent, or consistent but too recent, does not pass.

Second, patterns that pass the gate undergo routine synthesis. An LLM generates a natural-language routine summary and a role label (e.g., *primary hydration item*, *general supporting item*) describing the item's function within the routine. This functional labeling enables cross-context transfer (DG4): when the system recognizes a familiar item in a new environment, it retrieves the associated routine through shared item identity and co-occurrence structure rather than requiring an exact environmental match.

Adaptive importance with decaying reinforcement: A pattern not observed for weeks should fade in influence, while one that continues to recur should strengthen. Semantic pattern importance is modeled as a weighted combination of recency decay, saturating reinforcement from repeated observations, and evidence quality:

$$I(t) = \frac{w_r \cdot R(t) + w_{rf} \cdot S(n) \cdot R(t) + w_u \cdot u}{w_r + w_{rf} + w_u} \cdot Q(c)$$

where $R(t) = e^{-\lambda \Delta t}$ is a recency term ($\lambda = 0.08$/day) that decays exponentially with time since last reinforcement, $S(n) = 1 - e^{-\alpha n}$ is a saturating reinforcement term over n observations, $u \in [0, 1]$ is user-assigned utility, and $Q(c)$ scales the score by evidence quality. The weights w_r (recency), w_{rf} (reinforcement), and w_u (user utility) are set to 0.6, 0.3, and 0.1 respectively, reflecting the design priority that recent patterns should dominate the system's behavior. The key design property is that reinforcement is multiplied by recency, so a pattern frequently observed months ago does not retain high importance without fresh evidence, mirroring the biological principle that consolidated memories require periodic reactivation to maintain accessibility [15].

Exception pressure: Reconsolidation-style gating [48] is operated as an exception ratio over recent observations: when the proportion of observations that deviate from a pattern's expected distribution exceeds a threshold, S→K promotion is blocked. For instance, if the user stops jogging for a week and the water bottle no longer appears near the door in the evenings, the exception ratio rises and the system withholds promotion, preventing it from compiling a routine that the user's recent behavior contradicts.

3.3 Runtime Interaction

The runtime engine reads from pre-compiled memory layers and traverses all three in either direction: top-down for exception detection and bottom-up for cross-context transfer.

Exception detection (top-down): When the user enters the kitchen on day four for their morning routine, the schema tier generates expectations: all four items should be present. The semantic tier provides each item's typical location and temporal window. The episodic tier checks current observations against these expectations and finds that the water filter is absent. The system traverses K → S → E and produces a proactive reminder: "Your water filter was last seen near the sink yesterday evening", grounding the alert in compiled routine knowledge rather than raw retrieval.

Cross-context transfer (bottom-up): Now consider a different trigger: the user uses the dorm kitchen, where their lunchbox is recognized via visual embedding similarity (episodic). The semantic tier

retrieves the water bottle's contextual profile — it typically appears in the morning routine contexts alongside specific co-occurring items. The schema tier activates the associated cooking routine in this new setting and suggests the user could prepare a familiar lunch with available equipment. The system traverses E → S → K, and this transfer is possible because the schema encodes functional roles and co-occurrence structure rather than surface-level environmental features (DG4).

In both modes, LLM prompts are constructed from structured memory fields (pattern summaries, co-occurrence statistics, temporal distributions, episodic details). This ensures responses are grounded in compiled knowledge while remaining traceable to source evidence.

4 System Evaluation

We evaluate Inception through two complementary methods: (1) an automated ablation study using LLM-as-judge scoring across multiple stochastic runs, which isolates the contribution of each memory layer, and (2) a within-subjects user study (N=12) in which participants rate and compare system outputs. The automated evaluation provides statistical power through repeated trials; the user study validates that the measured differences are perceptible and meaningful to real users.

4.1 Evaluation Scenarios

The evaluation targets two interaction patterns: *exception detection*, where the system identifies deviations from an established routine, and *cross-context transfer*, where the system applies a known routine in a new environment. Ground-truth item presence is used to isolate the contribution of the memory consolidation architecture.

We designed eight scenarios organized into two trial plots: four targeting exception detection (P1E1–P1E4) and four targeting cross-context transfer (P2T1–P2T4). Each scenario has two phases. A *setup phase* establishes the user's routine through repeated observation in the video clips across 3 days, making objects and their roles visually clear through repetition. A *trigger phase* presents a moment requiring system intervention. We recorded all scenarios as first-person video with an actor performing everyday tasks using a Meta Rayban [36].

Plot 1 — Home Kitchen Morning (exception detection). The setup consists of three pre-recorded egocentric videos of a morning kitchen routine. The system ingests these videos through the perception pipeline, discovering 4 tracked items — water filter, water bottle, lunch box, and phone, then it records 73 episodic observations across 3 simulated mornings. The trigger phase presents 4 exception moments: 3 *absent* trials where one item is missing from the scene (water filter, water bottle, lunch box), and 1 *left-behind* trial where the user is leaving the house but the phone remains visible on the counter.

Plot 2 — Dorm Kitchen (cross-context transfer). The system's memory comes entirely from Plot 1 — no new observations are ingested. The user uses the dorm kitchen, where the same items appear in an unfamiliar environment. The trigger phase presents 4 transfer trials mirroring Plot 1: 3 absent items and 1 left-behind phone. The system must transfer its compiled routines from the home kitchen to this novel setting.

In total, we evaluate on 8 trials (4 exception + 4 transfer) covering two detection modes (absent and left-behind) and two interaction paths (top-down schema-driven and bottom-up transfer).

4.2 Automated Ablation Study

To isolate the contribution of each memory layer, we conduct a systematic ablation that removes consolidation stages one at a time. We compare three conditions (full system, semantic-only, and flat episodic baseline) across all eight trial scenarios, running each condition multiple times to account for LLM output stochasticity. An independent LLM judge scores each output on correctness, specificity, and false alarm rate.

4.2.1 Conditions. We compare three ablation conditions, each producing the same structured four-slot output (situation understanding, suggested action, rationale, memory reference):

1. **E+S+K (full system)**: All three memory layers. The runtime traverses K → S → E for exception detection (top-down) and E → S → K for transfer (bottom-up).
2. **E+S (no schema)**: Episodic and semantic layers only. Multimodal clustering discovers co-occurrence patterns, but no schema promotion occurs. The runtime matches semantic patterns to the current context.
3. **E-only (baseline)**: Flat episodic retrieval with LLM reasoning. No consolidation layers. Retrieves recent observations for all known items and passes them to the same LLM for free-form reasoning, but the baseline is not artificially constrained and may infer patterns from raw data.

All conditions use the same LLM (Gemini 2.5 Flash [21]), the same episodic observation data, and the same item set. The only variable is which consolidation layers are active. This ensures differences in output quality are attributable to the memory architecture.

4.2.2 Method. As the output of LLM is often stochastic, a single run cannot establish whether performance differences are reliable. To address this, we run each condition 10 times per trial with linearly spaced temperatures from 0.0 to 0.7 (capturing the range from deterministic to moderately creative generation), producing **8 trials × 3 conditions × 10 rounds = 240 system outputs**.

LLM-as-Judge Scoring. Manual evaluation of 240 outputs across three scoring dimensions would require substantial annotator effort and introduce inter-rater variability. Recent work has demonstrated that strong LLMs can serve as scalable and reliable evaluators of generated text, achieving agreement with human judges that matches inter-annotator agreement levels [62]. This LLM-as-judge paradigm has been adopted in HCI research for tasks ranging from large-scale screenshot coding [22] to interactive prompt evaluation [12], with validation studies consistently reporting substantial agreement with human raters [46, 63]. Following this approach, each output is scored by an independent LLM judge (Gemini 2.5 Flash [21], temperature 0) on three dimensions:

- **Correctness** (binary): Does the output identify the ground-truth missing or left-behind item by name?
- **Specificity** (1–5 Likert): Does the output provide contextual detail—item location, routine relevance, temporal context—beyond simply naming the item?

- **False alarm** (binary): Does the output incorrectly flag items that are not the ground-truth target?

Correctness and false alarm are objective criteria with unambiguous ground truth, minimizing the subjectivity concerns that can affect LLM judges on open-ended tasks [50]. Specificity is the only subjective dimension; we use a structured rubric with anchored examples at each scale point to reduce rating variance.

We report per-condition means with 95% confidence intervals and test for significance using the Friedman test (non-parametric repeated measures) with Wilcoxon signed-rank post-hoc comparisons and Bonferroni correction ($\alpha = 0.05/3 = 0.0167$).

4.3 User Study

We then evaluate whether the three-tier memory architecture produces better proactive assistance than a flat retrieval baseline. Twelve participants watched pre-recorded scenario videos depicting daily routines, then rated the assistance outputs produced by each system.

4.3.1 System Outputs. For each trigger moment, we generated two textual assistance outputs: one from Inception and one from the baseline. Inception draws on all three memory tiers as described. The baseline stores all observations as a flat memory stream with simple summarization. When a trigger moment occurs, the baseline retrieves the most relevant entries via embedding similarity and generates assistance from the retrieved context without distinguishing between episodic, semantic, and schematic information. Both outputs were normalized for style and length to prevent participants from distinguishing conditions based on surface features.

4.3.2 Participants and Procedure. We recruited 12 participants (9 male, 2 female, 1 non-binary; 10 aged 18–24, 2 aged 25–34; 9 East/Southeast Asian). Participants were compensated with 15 Canadian dollars per hour and the study lasted approximately 30 minutes. All participants were at least 18 years old and provided informed consent independently.

Table 1: Participant Demographics

ID	Age	Gender	Race/Ethnicity	AI	Wear.
P1	18–24	W	E/SE Asian	Med	Low
P2	25–34	M	E/SE Asian	High	High
P3	18–24	M	South Asian	High	Med
P4	18–24	W	E/SE Asian	Med	Low
P5	18–24	M	E/SE Asian	High	Low
P6	18–24	M	E/SE Asian	High	Med
P7	18–24	M	E/SE Asian	Med	Med
P8	18–24	W	E/SE Asian	Med	Med
P9	18–24	M	E/SE Asian	Med	Med
P10	25–34	M	White	Med	Low
P11	18–24	M	E/SE Asian	High	Med
P12	18–24	M	White	Med	Med

W = Woman, M = Man, Med = Moderate

AI = AI Familiarity, Wear. = Wearable Familiarity

The study followed a within-subjects design. Participants watched scenario videos and evaluated the textual outputs each system produced. For each scenario, both systems' outputs were presented

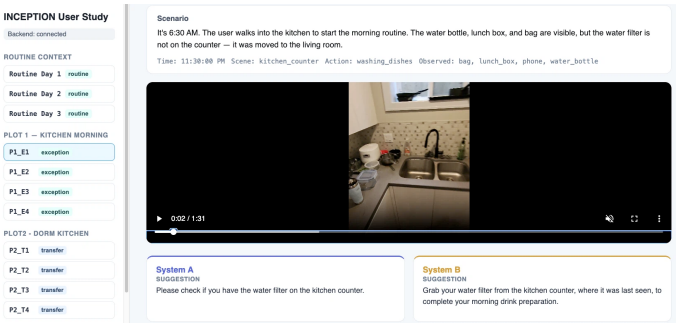

Figure 3: The interface for the participants to review the plot videos and the suggestions by both the full system and the flat baseline.

simultaneously and unlabeled, enabling direct comparison without order effects (Figure 3).

At each trigger moment, participants rated each system's output on a 7-point Likert scale (1 = poor, 7 = excellent). For exception trials, the prompt asked: "How well did the system detect and explain the deviation?" For transfer trials: "How well did the system apply known routines?" Participants also made a forced-choice selection of which system's response was better. After all scenarios, participants rated each system on comprehension ("I could understand what the system was doing") and adoption intent ("I would be willing to use a system like this in real life") on 7-point scales, and responded to four open-ended questions addressing preference rationale, information repetitiveness, handling of unexpected situations, and trust.

4.3.3 Analysis. Forced-Choice Preference: Forced choice is the most robust measure in the study because it forces a direct comparison and eliminates scale-use bias. We report pooled proportions with exact (Clopper-Pearson) 95% confidence intervals and binomial p-values as descriptive context. Because the 96 comparisons (12 participants × 8 scenarios) are not independent, the primary inferential test treats the participant as the unit of analysis: for each participant, we compute the proportion of choices favoring the baseline out of four trials per type, then test the deviation from 0.5 with a Wilcoxon signed-rank test. We report Cohen's g for proportions and matched-pairs rank-biserial r for Wilcoxon tests.

Likert Ratings by Trial Type: For each participant, we compute the median of their four ratings per system per trial type. The paired difference (baseline − Inception) across 12 participants is tested with a Wilcoxon signed-rank test, separately for exception and transfer trials. Comparing the effect sizes across trial types reveals whether any system advantage is specific to one capability.

Comprehension and Adoption Intent: We compare the two single-item measures using Wilcoxon signed-rank tests on paired differences. These analyses are exploratory given limited power at $N = 12$.

Qualitative Analysis: In addition to the quantitative measures, participants responded to two open-ended questions: one asking them to provide rationale for their system preference, and one asking whether they would trust such a system in their daily life. Two researchers reviewed these responses to identify recurring

patterns and organized them into descriptive categories. Representative quotes are included in the results to contextualize and supplement the quantitative findings.

Statistical Approach: All paired comparisons use the Wilcoxon signed-rank test, which makes no distributional assumptions and is appropriate for ordinal data at small N. The exception/transfer split is pre-planned and reflects the study's trial structure; p-values are reported without correction for multiple comparisons. Effect sizes are reported throughout.

5 Results

We present results in two parts. The automated ablation study establishes where architectural differences manifest across 240 system outputs, and the user study then examines whether these differences are perceptible and meaningful to participants, and surfaces qualitative tensions around tone and trust that the automated evaluation cannot capture.

5.1 Automated Ablation Results

We first report aggregate performance across all 80 outputs per condition to establish overall differences between memory configurations, then examine per-trial correctness to identify the specific scenarios that drive these differences.

5.1.1 Aggregate Results. Table 2 summarizes the aggregate results across 80 scored outputs per condition (8 trials × 10 rounds).

Metric	E+S+K (Full)	E+S (No Schema)	E-only (Baseline)
Correctness	**0.93** [0.87, 0.98]	0.56 [0.45, 0.67]	0.75 [0.65, 0.85]
Specificity (1–5)	**4.78** [4.62, 4.93]	3.33 [2.90, 3.75]	3.89 [3.53, 4.25]
False Alarm Rate	**0.01**	0.00	0.25

Table 2: Aggregate scores (mean [95% CI]) across 80 outputs per condition. Bold indicates best performance.

All pairwise differences are statistically significant. The Friedman test reveals significant main effects for correctness ($\chi^2(2) = 33.21$, $p < .001$), specificity ($\chi^2(2) = 30.33$, $p < .001$), and false alarm rate ($\chi^2(2) = 36.29$, $p < .001$). Wilcoxon post-hoc comparisons confirm significance for all pairs after Bonferroni correction: E+S+K vs. E-only (correctness: $p = .006$; specificity: $p < .001$), E+S+K vs. E+S (all $p < .001$), and E-only vs. E+S (correctness: $p < .001$; specificity: $p = .003$).

5.1.2 Per-Trial Analysis. Table 3 reports correctness rates (correct/10 rounds) for each trial, revealing where the architectural differences manifest.

Three patterns emerge:

Pattern 1: Simple absent-item detection is solvable by all conditions. For Plot 1 absent trials (p1_e1–p1_e3), all three conditions achieve 10/10 correctness. When an item is simply missing from a familiar context and the system has episodic observations of its typical presence.

Pattern 2: Left-behind detection requires schematic knowledge. Both left-behind trials (p1_e4, p2_t4) show a binary separation: E+S+K achieves 10/10 while both E-only and E+S score 0/10. Left-behind detection requires knowing that a visible item *should be carried* during departure, a knowledge that only exists in the schema layer's routine order encoding. Without compiled routines, neither raw retrieval nor semantic patterns can distinguish "item is present (normal)" from "item is present but should be taken (exception)."

Trial	Mode	Missing Item	E+S+K	E-only	E+S
p1_e1	absent	water_filter	10/10	10/10	10/10
p1_e2	absent	water_bottle	10/10	10/10	10/10
p1_e3	absent	lunch_box	10/10	10/10	10/10
p1_e4	left_behind	phone	**10/10**	0/10	0/10
p2_t1	absent	water_filter	9/10	10/10	5/10
p2_t2	absent	water_bottle	7/10	10/10	0/10
p2_t3	absent	lunch_box	8/10	10/10	10/10
p2_t4	left_behind	phone	**10/10**	0/10	0/10

Table 3: Per-trial correctness across 10 rounds. Left-behind trials (p1_e4, p2_t4) are only solvable by E+S+K.

Pattern 3: Cross-context transfer degrades without schemas but is partially recoverable by baseline. In Plot 2, E+S+K shows some variance (7–9/10 on absent trials) because transfer requires mapping compiled routines to unfamiliar contexts. Unexpectedly, E-only (baseline) achieves 10/10 on all absent transfer trials: the LLM can reason from raw observations that "this item was always present before, so it should be here too" without needing explicit routines. However, E+S performs worst (0–5/10 on two trials), suggesting that partial consolidation without schema compilation can mislead the LLM with incomplete pattern information.

5.2 User Study Results

We report the user study results in four parts: forced-choice preference as the primary measure of perceived quality, Likert ratings for a complementary continuous assessment, comprehension and adoption intent as exploratory measures, and qualitative observations from open-ended responses.

5.2.1 Forced-Choice Preference. Across all 96 comparisons (12 participants × 8 scenarios), participants preferred Inception's output in 58 cases (60%). This overall split did not reach significance on a pooled binomial test ($p = .052$), though the pooled test is anti-conservative because comparisons within each participant are not independent. The more informative pattern emerges when exception and transfer trials are examined separately.

On exception trials, participants preferred Inception in 34 of 48 comparisons (71%). At the participant level, 7 of 12 favored Inception on a majority of exception trials, while only 1 favored the baseline; 4 participants split evenly. A Wilcoxon signed-rank test on per-participant deviations from chance confirmed a significant preference for Inception ($p = .039$, $r = -.86$). The effect was large.

On transfer trials, preferences were evenly split (24 vs. 24). Four participants favored Inception on a majority of transfer trials, three favored the baseline, and five were tied. The Wilcoxon test showed no departure from chance ($p = 1.000$, $r = -.04$).

The forced-choice results indicate that Inception's advantage over the baseline is specific to exception detection and does not extend to cross-context transfer.

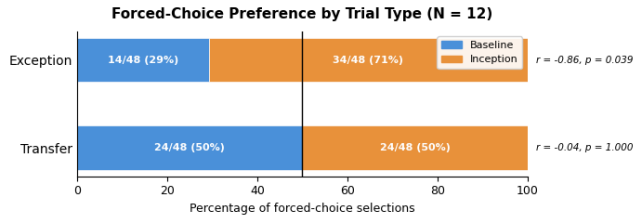

Figure 4: Forced-choice preference by trial type ($N = 12$).

5.2.2 Likert Ratings by Trial Type. Layer-specific ratings followed the same pattern as the forced-choice data, though the differences did not reach significance.

On exception trials, Inception received higher median ratings of perceived output quality than the baseline (5.0 vs. 4.2), corresponding to a medium effect ($p = .176$, $r = -.46$). On transfer trials, both systems received similar median ratings (Inception: 5.8, baseline: 6.0; $p = .625$, $r = -.20$).

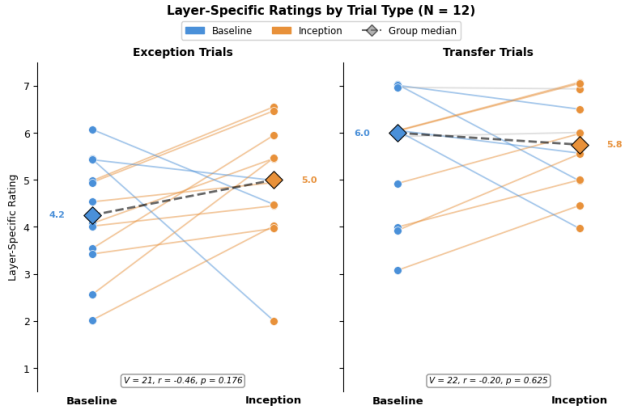

Figure 5: Layer-specific Likert ratings by trial type ($N = 12$). Each line connects one participant's median ratings.

5.2.3 Comprehension and Adoption Intent. Participants rated both systems similarly on comprehension. Median ratings for "I could understand what the system was doing" were 6.0 for both Inception and the baseline, with no significant difference ($p = .469$, $r = -.36$).

Willingness to use the system in real life also did not differ between conditions (Inception: Mdn = 4.5, baseline: Mdn = 5.0; $p = 1.000$, $r = -.02$). Inception's advantage in exception handling did not translate into greater perceived adoptability.

5.2.4 Qualitative Observations. Open-ended responses revealed several recurring patterns in how participants evaluated the two systems, broadly relating to the trade-off between specificity and autonomy, the role of accuracy in shaping preference, and concerns around trust and adoption.

Specificity versus autonomy. Participants frequently contrasted Inception's direct, detailed instructions with the baseline system's gentler, more suggestive tone. P1 noted that Inception's responses "were a lot more specific, like mentioning the location of the item and the specific task that was related to that item," which reduced the mental effort of remembering relevant details. However, P1 also acknowledged that the baseline system's softer wording—such as using "consider" rather than issuing direct commands—"made me feel more in control and less frustrated." P7 echoed this, preferring the baseline system because it "gave reasoning for its notifications" and "felt more like a reminder, which is what I would typically want for a morning routine," whereas Inception "seemed more demanding and felt less flexible to differences in routine." P12 similarly found the baseline system "more of a suggestive system than a demanding system." Conversely, P6 acknowledged that while Inception "was making a lot of assumptions" and could be "obtrusive," its focus on routine structure was ultimately more valuable: "the actual routine component is more valuable to have repeated to the user rather than the nuances which may change from day to day."

Accuracy as a threshold requirement. Several participants grounded their preference primarily in which system was more frequently correct. P3 noted that "[The baseline system] often gave me instructions that were irrelevant, while Inception gave me instructions relevant to the current routine." P10 preferred Inception because the baseline "misses the point of the scenario" in certain episodes, concluding that "Inception is less prone to failure." P12, however, reached the opposite conclusion, finding the baseline "correct more of the time and thus more convenient." Notably, P10 went further, stating they "would not trust either of these systems as both seem to be too inaccurate to be useful," suggesting that accuracy functioned as a baseline requirement regardless of other qualities.

Conditional trust and adoption barriers. When asked whether they would trust the system in daily life, most participants offered qualified rather than unconditional responses. P1 stated they "might trust it after trying it for a while or hearing about other people finding it useful" and expressed a desire to "tweak the responses so I can personalize them manually." P11 felt the system could be helpful "to an extent" but cautioned that "if it was not able to provide clear reasoning during uncertainty it would be more damaging than helpful." P3, by contrast, expressed clear enthusiasm, noting that the reassurance of knowing a system "can actively remind me not to forget" an important item provided peace of mind. A smaller number of participants raised more fundamental objections. P9 stated plainly, "it makes me feel dumb. I just don't like being told what to do." They also raised privacy concerns, describing the prospect of sharing personal data with such a system as "inviting a stranger into my life." P6 similarly questioned the premise, saying they would prefer to "enjoy doing my routine in peace" and challenging whether "this minute level of human cognitive load has any purpose in being replaced by agents."

5.2.5 Summary. The primary finding is that participants significantly preferred Inception's exception-detection responses over the baseline in forced-choice comparisons, with a large effect size ($r = -.86$, $p = .039$). This preference was specific to exception trials; transfer trials showed no systematic difference between systems. Likert ratings were directionally consistent with this pattern but did not reach significance, likely due to limited power at $N = 12$. Both systems were perceived as equally comprehensible and equally adoptable.

6 Discussion

The evaluation reveals that memory consolidation provides a categorical advantage for a specific class of assistance: tasks that require inferring what should happen based on compiled routine knowledge. We examine why this boundary emerges, what it implies for system design, and what tensions the qualitative findings surface around delivery tone, privacy, and trust.

6.1 Converging Evidence Across Evaluation Methods

The automated ablation and user study reach the same conclusion through independent methods: schema-level knowledge is necessary for left-behind detection, and this architectural boundary is perceptible to users, who significantly preferred Inception's exception-detection outputs and explicitly identified the capability gap in open-ended responses. This convergence across an automated evaluation with statistical power (240 outputs) and a human evaluation with ecological validity strengthens the claim that the boundary is a property of the task structure, not an artifact of either evaluation method.

6.2 When Does Consolidation Matter?

The evaluation results suggest a functional boundary: memory consolidation is most valuable when assistance requires **inferring what should happen**. Absent-item detection is merely a recognition task where the system needs only to notice that something previously present is no longer there, which raw episodic traces support. While left-behind detection is an inference task where the system must reason that a currently visible item should be carried away based on a compiled knowledge that exists only in the schema layer. This distinction maps onto the cognitive science literature on recognition versus recall [53, 58]: recognition can operate on surface-level familiarity, while recall requires structured knowledge that supports generative inference.

This boundary also explains why cross-context transfer does not uniformly benefit from consolidation. Raw retrieval is sufficient when the inference required is "this item was always present, so it should be here too;" while only schema knowledge is able to reason when the inference required is "this item should be taken with you because your routine involves it at the next location."

6.3 The Tone Tension: Confidence Versus Autonomy

An unexpected qualitative finding is that schema-driven confidence creates a delivery tension. Participants who preferred the baseline cited its more suggestive tone, even when acknowledging Inception's greater accuracy. This parallels findings [1, 40, 42] where participants valued proactive assistance that felt like a reminder rather than a directive. The tension is inherent to schema-backed systems: compiled routines enable more specific suggestions, but that same specificity can feel obtrusive, especially when the system intervenes proactively in personal routines.

This suggests a design implication for memory-augmented wearable systems: the certainty of the underlying knowledge representation should not directly determine the assertiveness of delivery. A system that *knows* the user's routine should still *present* its suggestions with hedged language, offering the user the choice to accept or override [33, 40, 55].

6.4 Privacy and the Always-On Trade-off

Several participants expressed discomfort with the premise of continuous observation with no regard to system performance. It reflects a broader concern in always-on wearable intelligence between the richness of captured data and users' sense of autonomy [4, 16]. Inception's consolidation architecture offers a partial response: because episodic traces are progressively compressed into semantic and schematic summaries, the system does not need to retain raw observations indefinitely. The compression transitions (E→S, S→K) are privacy-preserving in principle, without preserving the perceptual details of individual moments. Participants' feedback points to a complementary design direction: surfacing the consolidation hierarchy as a user-facing transparency mechanism, allowing users to inspect, modify, or delete stored knowledge at each memory tier [17, 56].

7 Limitations and Future Work

Our findings should be interpreted in light of several design choices. The evaluation used recorded videos rather than live interaction, allowing us to isolate output quality in a controlled setting; a natural next step is to study how these results translate to in-situ assistance during real routines. Deploying in situ would also require replacing the ground-truth item presence used here with a real-time perception pipeline.

Moreover, our scenario set focuses on a single kitchen morning routine with a small item vocabulary, which leaves open the question of how the memory layers behave across more diverse routines and environments.

Finally, the tone tension surfaced in the qualitative findings points to a concrete design opportunity: future versions of Inception could modulate delivery assertiveness based on schema confidence, an extension that builds directly on the current architecture.

8 Conclusion

We presented Inception, a three-tier memory architecture for always-on AI wearables that aligns agent memory organization with the consolidation processes observed in human memory systems. By progressively compressing episodic observations into semantic co-occurrence patterns and schematic routines, the architecture supports two forms of proactive assistance: top-down exception detection and bottom-up cross-context transfer. Converging evidence from an automated ablation study and a within-subjects user study demonstrates that this alignment is most beneficial when assistance requires inference over compiled routines rather than recognition of familiar patterns. Qualitative findings further reveal a tension between the contextual specificity and the delivery assertiveness, suggesting tone modulation as a potential consideration for consolidation-based interactive memory systems.

References

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen,

et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[2] Apple. 2023. Introducing Apple Vision Pro: Apple's first spatial computer. https://www.apple.com/newsroom/2023/06/introducing-apple-vision-pro/ Apple Newsroom press release.

[3] Sam Audrain and Mary Pat McAndrews. 2022. Schemas provide a scaffold for neocortical integration of new memories over time. *Nature communications* 13, 1 (2022), 5795.

[4] Divyanshu Bhardwaj, Alexander Ponticello, Shreya Tomar, Adrian Dabrowski, and Katharina Krombholz. 2024. In focus, out of privacy: the wearer's perspective on the privacy dilemma of camera glasses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[5] Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. 1982. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology* 14, 2 (1982), 143–177.

[6] Runze Cai, Nuwan Janaka, Hyeongcheol Kim, Yang Chen, Shengdong Zhao, Yun Huang, and David Hsu. 2025. Aiget: Transforming everyday moments into hidden knowledge discovery with ai assistance on smart glasses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26.

[7] Hyunsung Cho, Jacqui Fashimpaur, Naveen Sendhilnathan, Jonathan Browder, David Lindlbauer, Tanya R Jonker, and Kashyap Todi. 2025. Persistent assistant: Seamless everyday AI interactions via intent grounding and multimodal feedback. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[8] David Clewett and Lila Davachi. 2017. The ebb and flow of experience determines the temporal structure of memory. *Current opinion in behavioral sciences* 17 (2017), 186–193.

[9] Valdemar Danry, Jean Ghislain Billa, Yasith Samaradivakara, Paul Pu Liang, and Pattie Maes. 2026. Mind Mapper: Modeling and Predicting Behavioral Patterns from Everyday Conversations with Wearable AI Systems and LLMs. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*. 2059–2083.

[10] Ishita Dasgupta and Samuel J Gershman. 2021. Memory as a computational resource. *Trends in cognitive sciences* 25, 3 (2021), 240–251.

[11] Nigel Davies, Adrian Friday, Sarah Clinch, Corina Sas, Marc Langheinrich, Geoff Ward, and Albrecht Schmidt. 2015. Security and Privacy Implications of Pervasive Memory Augmentation. *IEEE Pervasive Computing* 14, 1 (2015), 44–53. doi:10.1109/MPRV.2015.13

[12] Michael Desmond, Zahra Ashktorab, Qian Pan, Casey Dugan, and James M Johnson. 2024. EvaluLLM: LLM assisted evaluation of generative outputs. In *Companion proceedings of the 29th international conference on intelligent user interfaces*. 30–32.

[13] Mustafa Doga Dogan, Eric J Gonzalez, Karan Ahuja, Ruofei Du, Andrea Colaço, Johnny Lee, Mar Gonzalez-Franco, and David Kim. 2024. Augmented object intelligence with xr-objects. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[14] Dejan Draschkow and Melissa L-H Vo. 2017. Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific reports* 7, 1 (2017), 16471.

[15] Yadin Dudai, Avi Karni, and Jan Born. 2015. The consolidation and transformation of memory. *Neuron* 88, 1 (2015), 20–32.

[16] Passant Elagroudy, Mohamed Khamis, Florian Mathis, Diana Irmscher, Ekta Sood, Andreas Bulling, and Albrecht Schmidt. 2023. Impact of Privacy Protection Methods of Lifelogs on Remembered Memories. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–10.

[17] Passant Elagroudy, Rufat Rzayev, Tonja-Katrin Machulla, Huy Viet Le, Tilman Dingler, Lars Lischke, Sarah Clinch, Geoffrey Ward, and Albrecht Schmidt. 2025. Pixel Memories: Do Lifelog Summaries Fail to Enhance Memory but Offer Privacy-Aware Memory Assessments?. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[18] Marc TJ Exton-McGuinness, Jonathan LC Lee, and Amy C Reichelt. 2015. Updating memories—The role of prediction errors in memory reconsolidation. *Behavioural brain research* 278 (2015), 375–384.

[19] Vanessa E Ghosh and Asaf Gilboa. 2014. What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia* 53 (2014), 104–114.

[20] Asaf Gilboa and Hannah Marlatte. 2017. Neurobiology of schemas and schema-mediated memory. *Trends in cognitive sciences* 21, 8 (2017), 618–631.

[21] Google. 2026. *Gemini 2.5 Flash*. https://deepmind.google/technologies/gemini/

[22] Longjie Guo, Yue Fu, Xiran Lin, Xuhai Xu, Yung-Ju Chang, and Alexis Hiniker. 2025. What Social Media Use Do People Regret? An Analysis of 34K Smartphone Screenshots with Multimodal LLM. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[23] Violet Yinuo Han, Jesse T Gonzalez, Christina Yang, Zhiruo Wang, Scott E Hudson, and Alexandra Ion. 2025. Towards unobtrusive physical ai: Augmenting everyday objects with intelligence and robotic movement for proactive assistance. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

[24] Brian Heater. 2023. Humane's Ai Pin promises an 'ambient computing' future for $699 (plus $24 a month). *TechCrunch* (9 Nov. 2023). https://techcrunch.com/2023/11/09/humanes-ai-pin/

[25] Andrew Hollingworth. 2007. Object-position binding in visual memory for natural scenes and object arrays. *Journal of Experimental Psychology: Human Perception and Performance* 33, 1 (2007), 31.

[26] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[27] Yoonsang Kim, Yalong Yang, and Arie E Kaufman. 2026. Memento: Towards Proactive Visualization of Everyday Memories with Personal Wearable AR Assistant. *arXiv preprint arXiv:2601.17622* (2026).

[28] Dharshan Kumaran, Demis Hassabis, and James L McClelland. 2016. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in cognitive sciences* 20, 7 (2016), 512–534.

[29] Geonsun Lee, Min Xia, Nels Numan, Xun Qian, David Li, Yanhe Chen, Achin Kulshrestha, Ishan Chatterjee, Yinda Zhang, Dinesh Manocha, et al. 2025. Sensible agent: A framework for unobtrusive interaction with proactive ar agents. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[30] Chenyi Li, Guande Wu, Gromit Yeuk-Yin Chan, Dishita Gdi Turakhia, Sonia Castelo Quispe, Dong Li, Leslie Welch, Claudio Silva, and Jing Qian. 2025. Satori : Towards Proactive AR Assistant with Belief-Desire-Intention User Modeling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–24.

[31] Jiahao Nick Li, Zhuohao Zhang, and Jiaju Ma. 2025. Omniquery: Contextually augmenting captured multimodal memories to enable personal question answering. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[32] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*. Springer, 38–55.

[33] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[34] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review* 102, 3 (1995), 419.

[35] Meta. 2023. Meta Quest 3 Coming This Fall + Lower Prices for Quest 2. https://about.fb.com/news/2023/06/meta-quest-3-coming-this-fall/ Meta Newsroom announcement.

[36] Meta. 2025. Meta Ray-Ban Display: AI Glasses With an EMG Wristband. https://about.fb.com/news/2025/09/meta-ray-ban-display-ai-glasses-emg-wristband/ Originally published September 17, 2025; updated September 30, 2025. Accessed 2026-02-27.

[37] Christian Meurisch, Cristina A Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. 2020. Exploring user expectations of proactive AI systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.

[38] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Michael V Heinz, Ashmita Kunwar, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F Huckins, et al. 2024. MindScape study: integrating LLM and behavioral sensing for personalized AI-driven journaling experiences. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 8, 4 (2024), 1–44.

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[40] Yi-Hao Peng, Dingzeyu Li, Jeffrey P Bigham, and Amy Pavel. 2025. Morae: Proactively pausing ui agents for user choices. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[41] Plaud. [n. d.]. Plaud NotePin: Your Wearable AI Note Taker & Memory Capsule. https://www.plaud.ai/products/plaud-notepin Product page. Accessed 2026-02-27.

[42] Kevin Pu, Ting Zhang, Naveen Sendhilnathan, Sebastian Freitag, Raj Sodhi, and Tanya R Jonker. 2025. Promemassist: Exploring timely proactive assistance through working memory modeling in multi-modal wearable devices. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–19.

[43] Gabriel A Radvansky and Jeffrey M Zacks. 2017. Event boundaries in memory and cognition. *Current opinion in behavioral sciences* 17 (2017), 133–140.

[44] Louis Renoult, Patrick SR Davidson, Daniela J Palombo, Morris Moscovitch, and Brian Levine. 2012. Personal semantics: at the crossroads of semantic and episodic memory. *Trends in cognitive sciences* 16, 11 (2012), 550–558.

[45] Melanie J Sekeres, Gordon Winocur, and Morris Moscovitch. 2018. The hippocampus and related neocortical structures in memory transformation. *Neuroscience letters* 680 (2018), 39–53.
[46] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–14.
[47] Junxiao Shen, John J Dudley, and Per Ola Kristensson. 2024. Encode-store-retrieve: Augmenting human memory through language-encoded egocentric perception. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 923–931.
[48] Alyssa H Sinclair, Grace M Manalili, Iva K Brunec, R Alison Adcock, and Morgan D Barense. 2021. Prediction errors disrupt hippocampal representations and update episodic memories. *Proceedings of the National Academy of Sciences* 118, 51 (2021), e2117625118.
[49] Weinan Sun, Madhu Advani, Nelson Spruston, Andrew Saxe, and James E Fitzgerald. 2023. Organizing memories for generalization in complementary learning systems. *Nature neuroscience* 26, 8 (2023), 1438–1448.
[50] Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th international conference on intelligent user interfaces*. 952–966.
[51] Dorothy Tse, Tomonori Takeuchi, Masaki Kakeyama, Yasushi Kajii, Hiroyuki Okuno, Chiharu Tohyama, Haruhiko Bito, and Richard GM Morris. 2011. Schema-dependent gene activation and memory encoding in neocortex. *Science* 333, 6044 (2011), 891–895.
[52] Endel Tulving et al. 1972. Episodic and semantic memory. *Organization of memory* 1, 381-403 (1972), 1.
[53] Endel Tulving and David Murray. 1985. Elements of episodic memory.
[54] Peng Wang and Alan F. Smeaton. 2013. Using visual lifelogs to automatically characterize everyday activities. *Information Sciences* 230 (2013), 147–161. doi:10.1016/j.ins.2012.12.028
[55] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2021. Understanding user perceptions of proactive smart speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–28.
[56] Maximiliane Windl, Albrecht Schmidt, and Sebastian S Feger. 2023. Investigating tangible privacy-preserving mechanisms for future smart homes. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
[57] Gordon Winocur and Morris Moscovitch. 2011. Memory transformation and systems consolidation. *Journal of the International Neuropsychological Society* 17, 5 (2011), 766–780.
[58] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–16.
[59] Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems* 43, 6 (2025), 1–47.
[60] Ada Yi Zhao, Aditya Gunturu, Ellen Yi-Luen Do, and Ryo Suzuki. 2025. Guided reality: Generating visually-enriched ar task guidance with llms and vision models. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–15.
[61] Dora Zhao, Diyi Yang, and Michael S Bernstein. 2025. Knoll: Creating a knowledge ecosystem for large language models. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–23.
[62] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.
[63] Qingxiao Zheng, Minrui Chen, Pranav Sharma, Yiliu Tang, Mehul Oswal, Yiren Liu, and Yun Huang. 2025. EvAlignUX: Advancing UX Evaluation through LLM-Supported Metrics Exploration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.
[64] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using large language models to realize a concise interface for real-time memory augmentation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009