Memento: Towards Proactive Visualization of Everyday Memories with Personal Wearable AR Assistant

Yoonsang Kim*
Stony Brook University

Yalong Yang†
Georgia Institute of Technology

Arie E. Kaufman‡
Stony Brook University

Figure 1: Concept illustration of Memento. The user's prior queries are stored in the form of "memories", along with their referent, spatiotemporal, and activity contexts. These memories are proactively situated onto daily environments upon alignment with the user's current interaction context, serving as a visual reminder and for learning.

ABSTRACT

We introduce Memento, a conversational AR assistant that permanently captures and memorizes users' verbal queries alongside their spatiotemporal and activity contexts. By storing these "memories," Memento discovers connections between users' recurring interests and the contexts that trigger them. Upon detection of similar or identical spatiotemporal activity, Memento proactively recalls user interests and delivers up-to-date responses through AR, seamlessly integrating the AR experience into their daily routine. Unlike prior work, each interaction in Memento is not a transient event, but part of a connected series of interactions with coherent long-term perspective, tailored to the user's broader multimodal (visual, spatial, temporal, and embodied) context. We conduct a preliminary evaluation through user feedback from participants of diverse expertise in immersive apps, and explore the value of a proactive context-aware AR assistant in everyday settings. We share our findings and challenges in designing a proactive, context-aware AR system.

Index Terms: Proactive Visualization, Context-aware, Spatial Computing, Wearable AI, Personal Assistant, Augmented Reality, Large Language Models.

1 INTRODUCTION

Recent advances in Large Language Models (LLMs) have significantly transformed user interfaces, enabling natural-language conversations that capture the semantics and nuances of user queries and generating rich, context-aware responses. Coupled with AR (Augmented Reality), LLM-powered conversational assistants not

*e-mail:yoonsakim@cs.stonybrook.edu
†e-mail:yalong.yang@gatech.edu
‡e-mail:ari@cs.stonybrook.edu

only deliver dynamic, real-time answers, but also ground digital content within the physical environment, creating interactions that empower intuitive engagement with real-world contexts.

Existing systems [9, 16, 38] have demonstrated the potential of combining AR, conversational interface, multi-sensory data, and LLMs to enable immersive, context-aware interactions. These approaches enrich user experiences by overlaying digital content onto everyday environments. However, these systems primarily focus on responding to immediate queries based on current or recent visual inputs, and short-term conversational history. While suitable for generic verbal queries, they fall short of long-term, personalized contexts. Also, the response to a question about an object can vary significantly since the meaning of the object can carry different value across users, and can shift according to spatial (*Where*), temporal (*When*), and task-driven (*What*) contexts. For example, in existing systems, querying about a toy doll would output the cost and the property of the product, regardless of its spatiotemporal context, while it may possess different personal meaning at home–"my kid's favorite toy"–compared to a store. Thus, a comprehensive understanding of the context behind a user's query is necessary. Another set of research explores personal "contextual memory banks" that assist users in recalling prior embodied interactions when queried [42, 43, 70]. Yet, there remain opportunities to expand the use cases beyond user-initiated (reactive) queries, towards proactive, and immersive, situated visualization of memories.

Moreover, we focus on the growing research interest in everyday smart wearable technologies (e.g., HMDs, glasses, earbuds, watches) that lifelog users' actions and surroundings [3, 19, 26, 70, 79, 80]. These works hint at how future personal device form factors can capture patterns and contexts of users' daily activities, and memorize them for remembrance or insight assistance.

To this end, we introduce **Memento**, a conversational AR assistant that aims to explore the usability and potential of proactive visualizations combined with contextual memories on a wearable AR form factor. Memento captures and memorizes a user's ver-

bal queries (Fig. 1–"What is the weather like today", "Tell me the breaking news of the day"), along with their spatiotemporal and activity contexts (Fig. 1–"House entrance", "at 7:50AM", while "Preparing to commute"). Utilizing this personal "contextual memory," Memento identifies the user's daily routine and proactively proposes visualizations of the user's interests learned from prior memories. This enables the users to bypass an explicit query amid repeated routines and serves as an informative visual reminder, as well as a new learning opportunity within everyday environments.

We pose '*proactivity*' as one potential interaction pattern of everyday wearable form factors–beyond user-requested unidirectional information flow–given their capability to memorize recurring behaviors and predict users' needs through multimodal AI-driven assistance [19, 47, 80]. By designing a prototype proactive context-aware system, Memento, and deploying it in various real-world scenarios, we share preliminary insights of everyday use of wearable AR. Our main contributions are:

- We design a **proactive AR assistant** prototype that situates visualizations of user's prior interests (e.g., weather, traffic) without explicit requests, serving as a smart visual reminder and live information provider in daily environments.

- We introduce the concept of **Referent-anchored Spatiotemporal Activity Memory (RSAM)** representation, which anchors user interactions to everyday objects along with their spatial, temporal, and activity context. RSAM enables Memento to surface long-term personal interests and recurring intentions, supported by a lightweight retrieval mechanism for recalling contextually relevant memories.

- We share **insights on the applicability of proactive, context-aware AR assistant** in real-world scenarios via a simplified prototype and report insights into designing such a system.

2 RELATED WORK

2.1 AI and AR Assistance in Routine Daily Life

Recent systems such as MaRginalia [62], XR-Objects [16], and Mirai [19] showed that linking digital content to physical referents and past interactions helps users continue tasks without re-finding information. With emerging wearable XR devices, XR is increasingly positioned as an everyday interface that maintains continuity between physical environments and digital memory [5, 6, 7, 51, 74]. Memoro [80] intelligently completes sentences and assists the user through audio using prior interactions.

The daily use of portable devices is highly routine [1, 15, 31, 66, 72], yet, current assistants require users to repeat similar queries daily and are reactively triggered. Proactive systems resurface information when contexts recur, enabling continuation of interactions without restarting. Satori [41] anticipates needs from task-based contexts, and Sensible Agent [37] proactively recognizes the needs of users for unobtrusive use of XR in public and minimizes interaction burden.

As XR transitions toward everyday use, proactively resurfacing personal context without repeated requests aligns with how users revisit routines, spaces, and objects. This motivates treating past interactions as personal "memories" linked to spatial, temporal, and activity cues, and resurfacing them whenever similar routines return, integrating digital recall into daily behavior.

2.2 Context-aware AI Assistant with LLM

LLMs have transformed user interfaces by enabling conversational, context-aware assistance. They are increasingly integrated into data analysis and visualization workflows [12, 13, 24, 71], as well as immersive applications, where LLMs function as intelligent companions [14, 76]. GazePointAR [38] demonstrates how combining gaze, gestures, and conversational history can disambiguate user
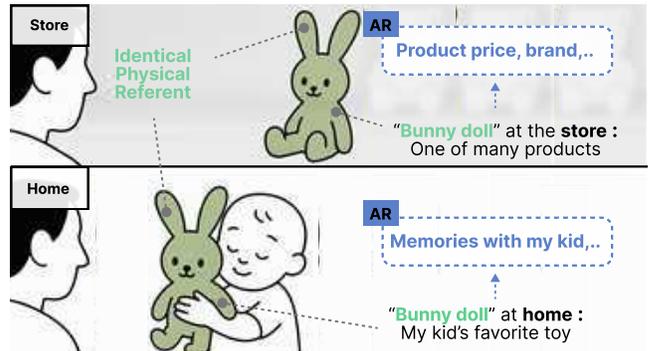

Figure 2: The view of situated AR in Memento. A physical referent derives its meaning from its spatiotemporal context (*Where* and *When* of the object). AR content anchored to it serves as an augmentation layer that extends the physically visible information of the referent, within the same context. The referent acts as the bridge between the physical and virtual worlds.

queries through richer contextual understanding. Similarly, World-Scribe [9] addresses accessibility by generating textual scene descriptions for visually impaired users.

These works illustrate how LLMs enhance AR by grounding multimodal inputs into context-aware responses. Memento builds on this by shifting from reactive assistance to proactive recall of prior user queries, enabled through memory lifelogging [25, 70] and context matching.

2.3 Capture and Retrieval of Memories

In situated visualization, the relationship between the user and the physical referents they interact with carries significant contextual value. In this work, we define each interaction–capturing spatial, temporal, verbal, visual, and activity context–as a "memory" [6, 30, 44, 67, 68]. Effectively storing and retrieving these memories requires robust data management. Traditional techniques such as R-trees and Oct/Quad-trees have been adopted to handle large-scale datasets [35, 61, 73], and are efficient for geo-tagged data, but struggle with unstructured queries or those requiring semantic context fusion [45, 75]. Recent research leverages vector embeddings from LLMs to represent multimodal data, enabling flexible retrieval through natural language and allowing systems to synthesize semantically coherent responses [4, 10, 28, 29, 40].

We build on this by integrating spatiotemporal indexing with embedding-based search, supporting real-time memory retrieval that balances structural filtering and semantic understanding.

2.4 User Activity and Context Modeling

Modeling user activity and context is essential for delivering intelligent and timely AR assistance. Prior research leverages multimodal data, including gaze, gesture, speech, and spatial context, to infer user intent [19, 26, 47]. Kim et al. [32] demonstrate that combining user actions, dialogue, and spatiotemporal information enables LLMs to extrapolate intent. Memoro [80] introduces query-less assistance by detecting conversation topics and augmenting users' remembrance through audio feedback. OmniActions [42] utilizes multi-sensory inputs to anticipate user actions.

Building on these insights, we capture user activities with their personal context to leverage them as a "personal spatial memory". The memories act as triggers for surfacing information when the context is semantically aligned–each spatiotemporal-activity-referent context holds disparate personal memories (Fig. 2).

3 MEMENTO: PROACTIVE MEMORY-AWARE ASSISTANT
3.1 Design Considerations

Designing an AR system for everyday use requires principles that support convenience, seamlessness, and robustness across routine
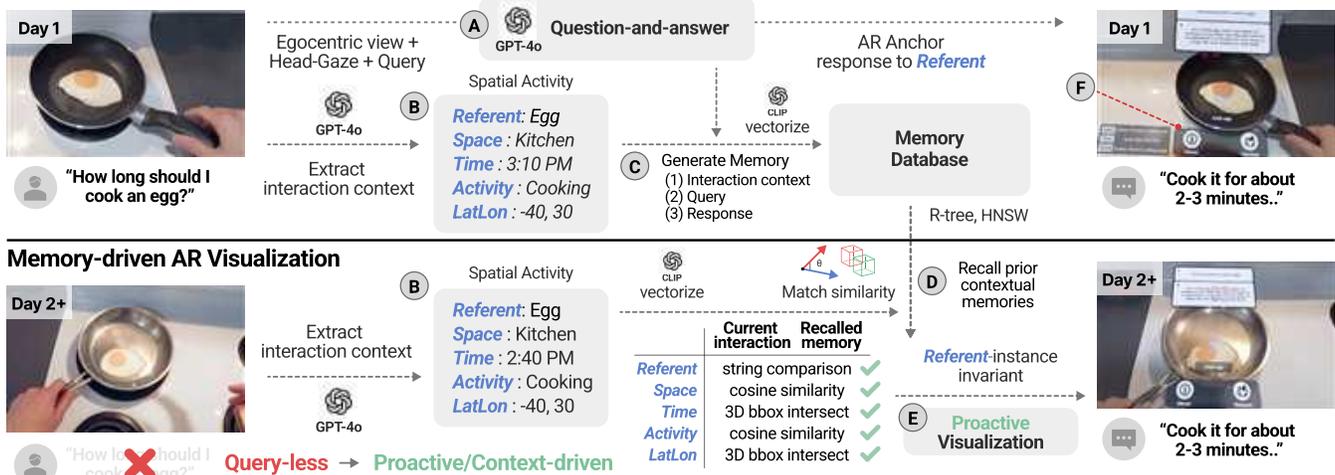
Figure 3: Memento pipeline overview: With a verbal query, the visual, head-gaze, and the transcribed textual query from the user are sent to Memento. Upon completion of response generation, the answer is situated onto the referent that the query is initiated to, and is permanently stored along with its spatial activity context as a form of a memory. After the initial query, the memory is proactively recalled when the user's similar contextual spatial activity, and situated onto the referent of the user's prior interest, without an explicit query from the user. (A) Personal Assistant, (B) Spatial Activity Recognizer, (C) Memory Generator, (D) Memory Retriever, (E) Proactive Visualizer, (F) Adjustment Interface

contexts. Prior work in ubiquitous and immersive computing highlights the need for systems that move beyond user-initiated interactions toward pervasive assistance [5, 22]. It shows that proactive support can reduce cognitive effort and improve continuity across repeated behaviors [2, 47, 54]. Wearable AR devices further enable embodied, egocentric interaction and situated visualization directly within the user's viewpoint [36, 79], while practically deploying them in-the-wild requires open-world scene understanding capable of identifying objects beyond predefined data class distribution [11, 46, 60]. These insights motivate our design considerations:

Proactive assistance. To reduce repeated explicit queries for routine tasks [2, 47], we employ semi-proactive resurfacing. Assistance is triggered by recurring user behaviors while allowing dismissal to avoid information overload [48, 56, 80]. A context-aware memory bank stores interaction histories and resurfaces them under matched spatiotemporal and activity cues, with an adjustable interface providing users agency over repetition cycles [54].

Egocentric visualization. Wearable AR serves as the primary platform to align digital content with users' perspectives and enable natural interaction (e.g., hand, gaze) in everyday environments. Compared to handheld AR, head-mounted displays more reliably capture multimodal egocentric inputs [79], supporting situated visualization without restricting physical interaction.

Open-world scene analysis. Everyday AR requires recognizing objects users reference, even when they are unseen during training. To address this, Memento leverages open-vocabulary object detection (YOLO-World [11]) and multimodal LLMs [20, 58] to identify objects from visual and textual descriptions, supporting assistance in unstructured, real-world environments.

While our design lays the groundwork for a full-fledged proactive assistant, in this work, we leverage our prototype system for simplified use case scenarios. Memento acts as a context-driven reminder that resurfaces relevant information when routines repeat (beyond simple time-based notifications) serving as a step towards long-term proactive, context-aware AR support.

3.2 Prototype System Design

Implementation Settings. At the time of our prototype development, raw access to a camera sensor (with RGB color) with AR content display, was limited on glasses form factors (e.g., Meta Ray Ban's), while these form factors are more suitable for daily use of AR. Therefore, we focus on the exploratory use-cases of daily AR assistance rather than form factor comparisons, and deploy Memento on a Meta Quest3 HMD (with Passthrough AR mode), and collect and discuss user feedback for preliminary insight building. Memento is developed with Unity AR Foundation. The HMD is connected to a custom remote Python server, which offloads compute-heavy processes including LLM interactions. In addition, the HMD is network-paired with an auxiliary device (smartphone) to provide accurate GPS location of the user. The storing and recalling of memories are crucial internal operations of Memento. We leverage two multimodal LLM models: GPT-4o [58] and Gemini 2.0-Flash [20], for generating memories, textual descriptions of scenes/activities, and referent localization – and CLIP embedding model [63] to vectorize memories, and deterministically compare retrieved memories.

3.2.1 Personal Conversational Assistant

The Assistant module supports the other five modules of Memento. As shown in Fig. 3, the assistant acts as the starting point that enables a user to interact with a referent in the physical environment. We track the user's head-gaze and pinch gesture to identify the user's referent of interest. At any point during an AR session, the user can perform the head-gaze-and-pinch to inquire about a topic from the assistant and receive an answer. Upon the detection of a pinch gesture, we capture the user's view as well as a verbal query, which is transcribed into text with the Whisper model [18], and transmitted to the LLM, connected to our server. We also place a visual marker on the 2D view capture to provide visual guidance to the referent of the user's query [38]. The marker is generated by projecting the 3D position of the referent derived from the gaze, onto the 2D capture. Moreover, to respond to the user's queries with live information (beyond the static knowledge-base of LLMs), we integrate the Google Custom Search API [21] into the assistant.

3.2.2 Spatial Activity Analysis

Timely recognizing the user's current space, and identifying the activity the user is engaged in, is pivotal to Memento. The appropriate inference of this information enables proactive visualizations of context-relevant AR content. We lifelog the user's first-person view (via the external camera of the HMD), latitude, longitude,
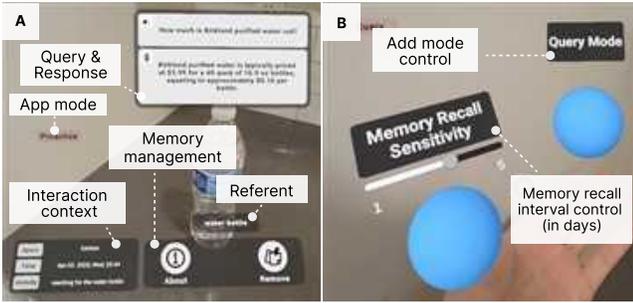
Figure 4: User interfaces of Memento. (A) Example visualization of RSAM situated onto a 'Water bottle' referent. (B) Context adjustment interface provides a switch between Query (New query) and Proactive (Prior query/Memory recall) mode, and adjust the interval of memory recalls (How often memory is recalled).

and timestamp, in a continuous manner, and infer the user's current scene and activity. This spatial activity analysis maintains an up-to-date context of the user's behaviors in the real-world. The interaction contexts include, the user's immediate space, activity, time of activity, the referent of the activity, and the geographical location of the user.

A textual description may lack sufficient details to fully characterize the user's spatial context. For example, two locations, both described as "Kitchen" may be in two geographically distinct locations (e.g., kitchens of different households). Thus, we first use GPT-4o to represent the user's current spatial activity in text, and associate geographical coordinates (latitude, longitude, accuracy: margin-of-error) with the Spatial Activity Profile. This combination of spatial activity contexts (latitude, longitude, visual-text descriptor, activity) enables Memento to distinguish spaces of the real-world spaces, while preserving the fine-grained semantics of each space and activity. The descriptor which is a single vector representation, is generated by prompting GPT-4o to output textual description of the scene given an image capture of the user's view, and converted to CLIP-encoded embeddings.

3.2.3 Personal Memory Generation

The referent is at the heart of the memory that bridges AR visualization with the physical realm (Fig. 2). Upon initial query interaction with a physical referent via Personal assistant, the query and its metadata are recorded in the form of a memory. It holds the moment of the verbal query, the response, the source of the response, and the context of the spatial activity. We term this, **R**eferent-anchored **S**patiotemporal **A**ctivity **M**emory (RSAM).

To generate an RSAM, we first identify the referent in question, and retrieve the most recent Spatial Activity Profile. Then, we bring the user's verbal query, its response, and the response source together, and form a memory. We store a new RSAM in a data structure that combines R-tree [23] with a graph-based hierarchical K-approximate nearest neighbor search algorithm known as HNSW [52]. This hybrid structure enables fast insertion and retrieval of RSAM, suitable for an interactive AR assistance system. The R-tree provides coarse partitioning of data based on latitude, longitude, and timestamp keys, while fast retrieval of the vector-represented spatial activity context captures the semantics of the user's scene and activity via HNSW.

3.2.4 Context-aware Memory Retrieval

Memory recall begins by accessing the most recent Spatial Activity Profile. It contains the user's current geographical location, scene information, activity, and timestamp. By filtering the existing RSAMs by geographic coordinates and time first, and hierarchically retrieving vectorized spatial activity, Memento enables fast top-K context similarity retrieval in a deterministic way.

In an effort to reduce false positive memory suggestions upon the retrieval of RSAMs, we equally weigh the significance of each

dimension in the Spatial Activity Profile. We sequentially verify the semantic alignment between each dimension, and the user's current context. The proximity of spatiotemporal context is assessed as the first step, then we verify the existence of the RSAM referents in the user's view, and finally fine-grained spatial activity semantics matching. Only when all dimensions are satisfied, an RSAM is suggested to the user.

3.2.5 Proactive Situated Information Display

Referent detection. Once the K-similar RSAMs–the past memories that align with the user's current spatiotemporal and activity context–return, the Proactive visualizer localizes the referents where the memories were situated previously. This situating of AR content is done by localizing the referents specified in RSAMs via an open-vocabulary detector. If present in the current view, the 2D center points are transmitted to the user's HMD to compute the 3D anchor points using the depth map acquired from the external camera of the HMD, at the time of the image capture.

Content visualization. A memory is AR-anchored onto the referent specified by an RSAM. As shown in Fig. 4-A, users are given the option to view the system reasoning on Spatial Activity contextual alignment. The Space, Time, and Activity of the matched information in a human-readable format is shown for explainability. The 'Remove' option allows permanent dismissal of unwanted (e.g., mismatched) memories or one-time queries that do not require repeating. Memento also proactively self-update the data of the contents providing newly available information on each repeat cycle. This makes Memento, beyond a simple immersive reminder, a fresh information-fetching agent.

Content interaction. To interact intuitively with the memories, we employ "head-gaze-and-pinch" and the design pattern of "Head-gaze and dwell" [55]. Users can naturally look towards (head-gaze) the memory of interest, and select it with the pinch gesture. It not only provides an embodied way of interacting with data, but also serves as a two-step safeguard to confirm the desired memory of interest before initiating the raycast selection.

3.2.6 User Preference and Proactivity Adjustment Interface

Fully autonomous information suggestion systems risk diminishing the user's sense of ownership and control [48]. As discussed in Sec. 3.2.5, Memento presents a user interface that indicates the spatial activity of the user's situated memory (space, time, activity), and the option to remove the memory (Fig. 4-A). Furthermore, the level of proactivity of AR visualizations can vary depending on personal preferences and tasks [54], thus, we provide an interface to adjust the memory recall sensitivity (Fig. 4-B) that controls the frequency of recall (How often the system recalls), in days.

4 EXAMPLE USE CASES AND APPLICATIONS

Memento leverages recurring patterns of daily activities to proactively surface relevant information without requiring repeated explicit queries. Rather than prompting the user to ask again, the system infers when past needs are likely to reappear based on recurring spatiotemporal routines and resurfaces the most contextually relevant memories through situated visualization on everyday objects. Below are three representative patterns of daily use, observed in our user study (denoted A1 to A3 across the sections and in Fig. 5).

Managing daily routines through recurring checks (A1). Users frequently request information tied to the start of their day, such as reviewing personal schedules or checking weather conditions before leaving home. Memento captures such explicit queries once, and upon detecting similar activity patterns at a later time (e.g., approaching a calendar in the bedroom or preparing to exit through the front door), we proactively surface updated reminders of upcoming events, weather, or commute conditions. By aligning past
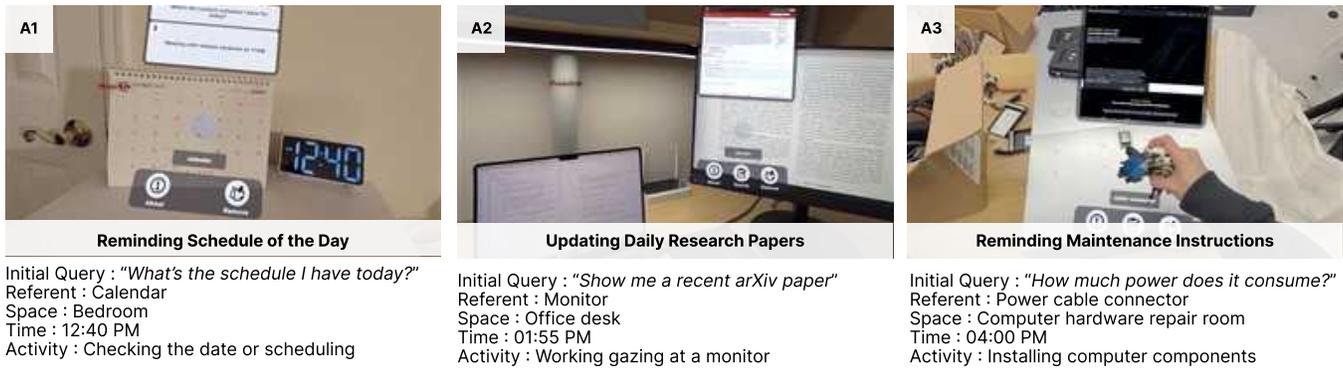
A1 — Reminding Schedule of the Day

Initial Query : *"What's the schedule I have today?"*
Referent : Calendar
Space : Bedroom
Time : 12:40 PM
Activity : Checking the date or scheduling

A2 — Updating Daily Research Papers

Initial Query : *"Show me a recent arXiv paper"*
Referent : Monitor
Space : Office desk
Time : 01:55 PM
Activity : Working gazing at a monitor

A3 — Reminding Maintenance Instructions

Initial Query : *"How much power does it consume?"*
Referent : Power cable connector
Space : Computer hardware repair room
Time : 04:00 PM
Activity : Installing computer components

Figure 5: The three real-world use-cases of Memento. Upon the contextual alignment of the user's current spatial activity (Referent, Space, Time, Activity), Memento proactively recalls a prior RSAM and situates via AR.

interests with present routine, the system enables users to bypass repeated queries and supports smooth daily planning.

Staying up-to-date in information-rich contexts (A2). For activities involving regular information monitoring, such as research and entertainment, users often revisit similar referents and spaces with similar intentions (e.g., sitting at a workstation to look up new papers, or taking a break in a coffee corner while catching up on sports news). Recognizing these consistent contextual cues, the system retrieves previously asked queries (e.g., "latest research papers") and presents fresh responses situated on familiar objects in those spaces. This pairing of personal querying history with recurring usage locations turns routine contexts into a lightweight information provider.

Recalling task-relevant instructions and reminders (A3). Some tasks benefit from resurfacing specific information at the moment of re-engagement (e.g., technical maintenance steps or health-related reminders). When users return to the same tools, locations, or objects, the system proactively resurfaces prior instructions anchored to those referents. This helps reduce the cognitive overhead of re-remembering previously searched details and provides a just-in-time memory cue at the point of action.

5 PRELIMINARY EVALUATION

To assess the practicality and potential of proactive, context-aware AR assistance, we conducted a two-part preliminary evaluation: (1) a technical analysis of Memento capabilities and performance, and (2) a "take-home" diary-based user study examining user acceptance, user experience, real-world usage patterns, and how the technology integrates into everyday routines.

5.1 Technical Evaluation

Identifying users' action contexts, and delivering timely and accurate situated AR content, are crucial for the practical adoption of proactive context-aware systems. In this section, we assess the context-aware capabilities of Memento and report our results.

Spatial activity context matching accuracy. We evaluate the accuracy of the Spatial Activity Recognizer in identifying spatial and activity contexts from egocentric views using text-based scene representations. Each image is converted into a textual description with GPT-4o, and CLIP is used to generate embeddings directly from these textual summaries. To assess classification performance, we compute centroid embeddings for each spatial and activity class using a sampled dataset of 1,000 egocentric images drawn from Ego4D. For each test sample, its embedding is compared against class centroids using cosine similarity, retrieving the closest match as the predicted label.

The recognizer achieves an F1 score of 0.82 for Scene context classification and 0.82 for Activity context classification (see Tab. 1). We find that this tendency continues until images contain motion blur or low lighting, where the accuracy drops to an F1 score

of 0.57 and 0.30, respectively. These results indicate that text-based embeddings can provide semantic distinction to identify recurring spatial activities for proactive resurfacing, even without fine-tuning or visual modality fusion.

Interaction referent recognition accuracy. We found that the open-vocabulary object detector, YOLO-World-XL, occasionally struggles to robustly localize a referent that the user is interacting with. While it outperforms DetCLIP-T in speed (52 vs 2FPS) with similar accuracy (34.4 vs 35.4 AP), false positives led to faulty retrieval of RSAM and inconsistent situating of proactive AR visualizations, impacting the user experience.

To address this, we use Gemini 2.0-Flash as a fallback when detection confidence is low (< 0.5). It is chosen for its spatial reasoning responsiveness ($\approx 3.4s$) with competitive accuracy among other multimodal LLMs, at the time of the prototype design [59].

5.2 User Evaluation and Feedback

We collected data from 25 sessions conducted over a 2-4 day period, resulting in a total of ≈ 435 aggregated minutes, and 196 memories (the total number of query-generated memories across sessions; including user-dismissed memories) across all participants. Although each session lasted for 17.4 minutes on average, and was relatively short in duration, participants were exposed to 135 memories via proactive recalls (68.9% of total memories).

Participants and procedure. We recruited 9 (P1-P9) self-volunteered participants (7 male, 2 female), aged 25-36 (μ=29.1, σ=3.7), from Computer Science, Mathematics, and Physics backgrounds. Prior AR familiarity was broadly distributed on a 5-point scale (1:Not familiar, 5:Familiar), (μ=2.8, σ=1.3), and gesture-based interaction experience showed a similar spread (μ=3.1, σ=1.6). All participants reported regular daily use of personal computing devices. Participants' identities were anonymized, and each provided informed consent prior to participation. This study was conducted under the IRB approval of Stony Brook University (1173920). The participants were instructed to complete a brief pre-study questionnaire and a 30-minute introduction to Memento. They were told to use Memento naturally over 2-4 days in their

Table 1: Performance comparison of CLIP variants–CNN-based (ResNet-101) and Transformer-based (ViT-B/32, ViT-L/14) architectures, for spatial Scene and Activity recognition tasks.

Task	CLIP Variants	Precision	Recall	F1-score
Scene	ViT-L-14	0.81	0.88	0.81
	ViT-B-32	0.80	0.87	0.80
	ResNet-101 (**Ours**)	0.82	0.89	**0.82**
Activity	ViT-L-14	0.81	0.85	0.80
	ViT-B-32	0.75	0.80	0.73
	ResNet-101 (**Ours**)	0.82	0.87	**0.82**

daily indoor environments, with each usage counted as a "session." After each session, participants submitted short diary entries describing context, perceived usefulness, and ease of integration, including a 5-point Likert scale and open-ended feedback. System logs of spatial context, interactions, and resurfaced memories were collected in parallel. After the multi-day use, participants completed a post-study survey comparing proactive and reactive support. Semi-structured interviews followed to explore how proactive cues affected routines and what limitations they encountered. To ensure consistency across participants, proactive resurfacing sensitivity was fixed at one resurfacing per 24 hours.

Repeated daily usages. Participants praised Memento for its proactive task reminding ability–"*I liked that I could place reminders around the house. Next to my shoes, I'll know I need to go out for a run*"(P3). However, one user reported referent misalignment: "*It mistook a square grocery bag to be a luggage and suggested me travel time*"(P3). A post-study interview revealed that the user moved close to the referent, enforcing our recognition model to infer with minimal visual context. In another case, a spatial activity was over-generalized: "*It said I'm gazing at an object. It's too generic*"(P2). We discuss candidate solutions in Sec. 6.

Acceptability in daily routine. Users were favorable towards Memento becoming part of their daily routine: "*Each time I use it, I get more comfortable with the interaction style. It feels like a solid foundation*"(P5). However, the requests to extend Memento, to a lightweight glasses-based AR were suggested across participants (P1, P3, P4, P6, P8); "*It'll be nice to have a VR app integrated into lifestyle through glasses that pops up information quick*"(P6). A non-tech user said : "*It is cool, but I wear a huge glasses, so I can't wear it for long*"(P3). Furthermore, even with a multi-day AR experience, users expressed the need for a longer duration, to observe any changes in their routines: "*I cannot say it changed my routines yet. It clearly helped me, but I'd need to try it on for longer. I might feel differently if it is a glasses I can wear casually, like the Rayban*"(P1); "*Very promising system. Looking forward to seeing how it adapts with more use*"(P5). We extend our discussion to Sec. 6.

Proactive vs Reactive assistance. Proactive context-aware AR in Memento was evaluated positively: "*I check weather and traffic every morning, this gave me the answers before I checked*"(P7); "*It reduces a step in your daily patterns*"(P6); "*Great for reminders. I usually use Google Calendar, but many times my phone is on silent or put away, I'm unable to follow through. Proactivity definitely helps more and put bunch of To-Dos around my house*"(P3). Another participant compared Memento to a traditional reactive assistant: "*I use Google Nest Mini to voice query and get responses, but Memento recall of previous queries is very useful*"(P1), and praised on its simplicity "*Biggest pros is that I don't need to think about setting up the conditions, just automatic and simple*"(P6).

6 FUTURE WORK AND DISCUSSION

Towards everyday AR form factors. Participants recognized the usefulness of proactive resurfacing but highlighted the mismatch between the system's value and its current hardware. As one noted, "*If it was like glasses, I could really see the benefits of AR outdoors. It's awkward to go out with this*" (P4). Everyday deployment likely depends on lightweight glasses-class devices such as Meta's Ray Ban glasses or XReal [78], where proactive information blending can occur with less social discomfort. Since the system already offloads computation to remote servers, cross-platform deployment is feasible. Future work will explore non-verbal triggers such as audio, gaze, and micro-gestures, building on research into subtle input for public AR use [49, 50], enabling proactive retrieval even when speech would be socially intrusive.

Mixed-initiative context control and activity granularity. Accurate resurfacing depends on inferring when a resurfaced memory is contextually appropriate. AR content that is mis-triggered may negate the usefulness and utility of proactive visualization. While participants generally appreciated autonomous triggers, they wanted occasional corrective control: "*I sometimes wanted more control*"(P5). This suggests that proactive AR should adopt mixed-initiative mechanisms that remain simple, yet adjustable. Prior work on user-assistant collaboration and conditional triggers, such as IFTTT (If-This-Then-That) [27, 69]-like structures, may provide contexts for lightweight refinements. Improving recognition accuracy may also benefit from hierarchical activity representations as suggested by ongoing efforts in structured activity inference [17, 57, 77, 79], mitigating over-generalization and helping maintain alignment between resurfacing timing and user intent.

Continuity of context across realities. Participants responded positively when resurfaced information appeared, "*where it felt natural*," especially when memories were tied to objects already associated with an activity. While the current panel-based overlays provide invariance to layout changes, they lack deeper geometric coupling with the environment. Closer integration between AR content and spatial referents could strengthen relevance and improve interpretability and explainability, building on design considerations for spatially anchored AR interfaces [36]. Future work should explore referent-aware embeddings that balances spatial robustness with meaningful anchoring.

Need for longitudinal routine pattern evaluation. Our diary-based deployment captured short-term routine patterns but could not reveal how resurfacing should adapt as habits evolve. Participants occasionally treated the system as a temporary "study task," leading to usage that did not fully reflect everyday rhythms. Longer-term field studies would allow users to internalize proactive AR as part of daily life, similar to long-term wearable research [7]. Understanding when proactive reminders should scale back, pause, or revive requires extended observation of routine drift and habit change over weeks or months, rather than days.

Privacy boundaries and memory retention. Even though the system did not store raw imagery, participants raised concerns about what the system "remembers" and whether their home environment might be inferred from resurfaced cues: "*Do you save the captures? I don't want my room to be seen*"(P4). This aligns with longstanding concerns about visual privacy in lifelogging systems [5] and underscores the need for user control over memory retention and deletion. Prior work on selective access control and user agency [33, 34, 64, 65] suggests that proactive systems require transparency mechanisms and forgetting controls that allow resurfacing without unintended exposure of personal spaces or objects.

Systematic evaluation of proactive benefits. This work explored proactive resurfacing as a design space, but the cognitive and performance effects compared to reactive assistants remain underexamined. In our future work, we plan to investigate whether resurfaced memories reduce recall burden or task-switching overhead, and examine how multimodal cues such as visual, audio, and gaze combinations influence attention and action [8, 39, 53]. Understanding when proactivity assists, distracts, or overwhelms will refine how proactive AR integrates into everyday practice.

7 CONCLUSION

To explore the usability of proactive, context-aware assistance on a wearable AR device, we designed a foundational system, Memento, that captures and stores users' verbal interactions as Referent-anchored Spatiotemporal Activity Memories, and proactively resurfaces them based on recurring daily contexts. We introduced a hybrid data retrieval structure combining spatial indexing and semantic vector search, and implemented a simplified context-aware system by integrating multimodal sensing, LLM-based reasoning, and open-vocabulary referent detection. Through real-world use cases, we investigated the potential for everyday use of a proactive AR assistant, and discuss takeaways and future work.

REFERENCES

[1] M. B. Almourad, M. Hussein, E. Bataineh, and Z. Wattar. Exploring smartphone usage dynamics: Unveiling app-specific patterns and trends. *IJWI*, 22(2), 2024. 2

[2] S. Andolina, V. Orso, H. Schneider, K. Klouche, T. Ruotsalo, L. Gamberini, and G. Jacucci. Searchbot: Supporting voice conversations with proactive search. In *Proc. of ACM CSCW*, pp. 9–12, 2018. 3

[3] R. Arakawa, J. F. Lehman, and M. Goel. Prism-q&a: Step-aware voice assistant on a smartwatch enabled by multimodal procedure tracking and large language models. In *Proc. ACM IMWUT*, vol. 8, pp. 1–26, 2024. 1

[4] A. Asai, S. Min, Z. Zhong, and D. Chen. Retrieval-based language models and applications. In *Proc. of ACL*, pp. 41–46, 2023. 2

[5] R. T. Azuma. The road to ubiquitous consumer augmented reality systems. *HBET*, 1(1):26–32, 2019. 2, 3, 6

[6] W. Barfield. *Fundamentals of wearable computers and augmented reality*. CRC Press, 2015. 2

[7] N. Bressa, J. Vermeulen, and W. Willett. Data every day: Designing and living with personal situated visualizations. In *Proc. of ACM CHI*, pp. 1–18, 2022. 2, 6

[8] R. Cai, N. Janaka, H. Kim, Y. Chen, S. Zhao, Y. Huang, and D. Hsu. Aiget: Transforming everyday moments into hidden knowledge discovery with ai assistance on smart glasses. *arXiv:2501.16240*, 2025. 6

[9] R.-C. Chang, Y. Liu, and A. Guo. Worldscribe: Towards context-aware live visual descriptions. In *Proc. of ACM UIST*, pp. 1–18, 2024. 1, 2

[10] W. Chen, H. Hu, X. Chen, P. Verga, and W. Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proc. of EMNLP*, pp. 5558–5570, 2022. 2

[11] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proc. of IEEE/CVF CVPR*, 2024. 3

[12] K. Choe, C. Lee, S. Lee, J. Song, A. Cho, N. W. Kim, and J. Seo. Enhancing data literacy on-demand: LLMs as guides for novices in chart interpretation. *IEEE TVCG*, 2024. 2

[13] W. Cui, X. Zhang, Y. Wang, H. Huang, B. Chen, L. Fang, H. Zhang, J. Lou, and D. Zhang. Text-to-Viz: automatic generation of infographics from proportion-related natural language statements. *IEEE TVCG*, 26(1):906–916, 2020. 2

[14] F. De La Torre, C. M. Fang, H. Huang, A. Banburski-Fahey, J. Amores Fernandez, and J. Lanier. Llmr: Real-time prompting of interactive worlds using large language models. In *Proc. of CHI*, pp. 1–22, 2024. 2

[15] T. Deng, S. Kanthawala, J. Meng, W. Peng, A. Kononova, Q. Hao, Q. Zhang, and P. David. Measuring smartphone usage and task switching with log tracking and self-reports. *MMC*, 7(1):3–23, 2019. 2

[16] M. D. Dogan, E. J. Gonzalez, A. Kumar, R. Du, A. Colaço, J. Lee, M. Gonzalez-Franco, and D. Kim. Augmented object intelligence with XR-Objects. In *Proc. of ACM UIST*, pp. 1–15, 2024. 1, 2

[17] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv:2404.16130*, 2024. 6

[18] A. Evgrashin. Whisper for unity. https://github.com/Macoron/whisper.unity/tree/master, 2024. Aug. 31. 2024. 3

[19] C. M. Fang, Y. Samaradivakara, P. Maes, and S. Nanayakkara. Mirai: A wearable proactive ai" inner-voice" for contextual nudging. *arXiv:2502.02370*, 2025. 1, 2

[20] Google. GoogleAI. Gemini models. https://ai.google.dev/gemini-api/docs/models/. Mar. 21. 2025]. 3

[21] Google. Programmable search engine. https://developers.google.com/custom-search/v1/overview, 2025. Mar. 23. 2025. 3

[22] J. Grubert, T. Langlotz, S. Zollmann, and H. Regenbrecht. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE TVCG*, 23(6):1706–1724, 2016. 3

[23] A. Guttman. R-trees: a dynamic index structure for spatial searching. *ACM SIGMOD*, 14(2):47–57, 1984. 4

[24] C. Han and K. E. Isaacs. A deixis-centered approach for documenting remote synchronous communication around data visualizations. *IEEE TVCG*, 31(1):930–940, 2024. 2

[25] M. Harvey, M. Langheinrich, and G. Ward. Remembering through lifelogging: A survey of human memory augmentation. *PMCJ*, 27:14–26, 2016. 2

[26] Y. Huang, J. Xu, B. Pei, Y. He, G. Chen, L. Yang, X. Chen, Y. Wang, Z. Nie, J. Liu, G. Fan, D. Lin, F. Fang, K. Li, C. Yuan, Y. Wang, Y. Qiao, and L. Wang. Vinci: A real-time embodied smart assistant based on egocentric vision-language model. *arXiv:2412.21080*, 2024. 1, 2

[27] IFTTT. What is IFTTT. https://ifttt.com/explore/new_to_ifttt, 2025. Apr. 4. 2025. 6

[28] S. Kashmira, J. L. Dantanarayana, J. Brodsky, A. Mahendra, Y. Kang, K. Flautner, L. Tang, and J. Mars. A graph-based approach for conversational ai-driven personal memory capture and retrieval in a real-world application. *arXiv:2412.05447*, 2024. 2

[29] M. S. U. Khan, M. Z. Afzal, and D. Stricker. SituationalLLM: proactive language models with scene awareness for dynamic, contextual task guidance. *arXiv:2406.13302*, 2024. 2

[30] O. Khan, Z. Ahmed, H. Nam, and K. Kim. Tangiblemoments: Embedding XR memories onto physical objects. In *Proc. of IEEE VRW*, pp. 1147–1153, 2025. 2

[31] D. Kim, J. Gluck, M. Hall, and Y. Agarwal. Real world longitudinal ios app usage study at scale. *arXiv:1912.12526*, 2019. 2

[32] Y. Kim, Z. Aamir, M. Singh, S. Boorboor, K. Mueller, and A. E. Kaufman. Explainable XR: understanding user behaviors of XR environments using LLM-assisted analytics framework. *IEEE TVCG*, 31(5):1–11, 2025. 2

[33] Y. Kim, S. Boorboor, A. Rahmati, and A. Kaufman. Design of privacy preservation system in augmented reality. In *Proc. of IEEE VizSec*, 2021. 6

[34] Y. Kim, S. Goutam, A. Rahmati, and A. Kaufman. Erebus: Access control for augmented reality systems. In *Proc. of USENIX Security*, pp. 929–946, 2023. 6

[35] N. Koutroumanis and C. Doulkeridis. Scalable spatio-temporal indexing and querying over a document-oriented nosql store. In *Proc. of EDBT*, pp. 611–622, 2021. 2

[36] B. Lee, M. Sedlmair, and D. Schmalstieg. Design patterns for situated visualization in augmented reality. *IEEE TVCG*, 30(1):1324–1335, 2023. 3, 6

[37] G. Lee, M. Xia, N. Numan, X. Qian, D. Li, Y. Chen, A. Kulshrestha, I. Chatterjee, Y. Zhang, D. Manocha, et al. Sensible agent: A framework for unobtrusive interaction with proactive ar agents. In *Proc. of ACM UIST*, pp. 1–22, 2025. 2

[38] J. Lee, J. Wang, E. Brown, L. Chu, S. S. Rodriguez, and J. E. Froehlich. GazePointAR: a context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In *Proc. of ACM CHI*, pp. 1–20, 2024. 1, 2, 3

[39] J. Lee, T. Wang, J. Fashimpaur, N. Sendhilnathan, and T. R. Jonker. Walkie-talkie: Exploring longitudinal natural gaze, llms, and vlms for query disambiguation in xr. In *Proc. of ACM CHI EA*, pp. 1–9, 2025. 6

[40] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 33:9459–9474, 2020. 2

[41] C. Li, G. Wu, G. Y.-Y. Chan, D. G. Turakhia, S. Castelo Quispe, D. Li, L. Welch, C. Silva, and J. Qian. Satori: Towards proactive ar assistant with belief-desire-intention user modeling. In *Proc. of ACM CHI*, pp. 1–24, 2025. 2

[42] J. N. Li, Y. Xu, T. Grossman, S. Santosa, and M. Li. OmniActions: predicting digital actions in response to real-world multimodal sensory inputs with LLMs. In *Proc. of ACM CHI*, pp. 1–22, 2024. 1, 2

[43] J. N. Li, Z. J. Zhang, and J. Ma. Omniquery: Contextually augmenting captured multimodal memory to enable personal question answering. In *Proc. of ACM CHI*, 2025. 1

[44] J. Liu, K. A. Satriadi, B. Ens, and T. Dwyer. Investigating the effects of physical landmarks on spatial memory for information visualisation in augmented reality. In *Proc. of IEEE ISMAR*, pp. 289–298, 2024. 2

[45] S. Liu, J. Xu, W. Tjangnaka, S. Semnani, C. Yu, and M. Lam. Suql: Conversational search over structured and unstructured data with large language models. In *Proc. of NAACL*, pp. 4535–4555, 2024. 2

[46] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proc. of ECCV*, pp. 38–55, 2024. 3

[47] X. B. Liu, S. Fang, W. Shi, C.-S. Wu, T. Igarashi, and X. Chen. Proactive conversational agents with inner thoughts. In *Proc. of ACM CHI*, 2025. 2, 3

[48] X. B. Liu, V. Kirilyuk, X. Yuan, A. Olwal, P. Chi, X. A. Chen, and R. Du. Visual captions: augmenting verbal communication with on-the-fly visuals. In *Proc. of ACM CHI*, pp. 1–20, 2023. 3, 4

[49] F. Lu and D. A. Bowman. Evaluating the potential of glanceable ar interfaces for authentic everyday uses. In *IEEE VR*, pp. 768–777, 2021. 6

[50] F. Lu, L. Pavanatto, and D. A. Bowman. In-the-wild experiences with an interactive glanceable ar system for everyday use. In *Proc. of ACM SUI*, pp. 1–9, 2023. 6

[51] Z. Lv, N. Charron, P. Moulon, A. Gamino, C. Peng, C. Sweeney, E. Miller, H. Tang, J. Meissner, J. Dong, et al. Aria everyday activities dataset. *arXiv:2402.13349*, 2024. 2

[52] Y. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE TPAMI*, 42(4):824–836, 2018. 4

[53] R. E. Mayer. Multimedia learning. In *Psychology of learning and motivation*, vol. 41, pp. 85–139. Elsevier, 2002. 6

[54] C. Meurisch, C. A. Mihale-Wilson, A. Hawlitschek, F. Giger, F. Müller, O. Hinz, and M. Mühlhäuser. Exploring user expectations of proactive ai systems. *Proc. of ACM IMWUT*, 4(4):1–22, 2020. 3, 4

[55] Microsoft. Head-gaze and dwell. https://learn.microsoft.com/en-us/windows/mixed-reality/design/gaze-and-dwell-head. Mar. 23. 2025. 4

[56] C. Morris, V. Danry, and P. Maes. Wearable systems without experiential disruptions: exploring the impact of device feedback changes on explicit awareness, physiological synchrony, sense of agency, and device-body ownership. *FCS*, 5, 2023. 3

[57] L. Ning, L. Liu, J. Wu, N. Wu, D. Berlowitz, S. Prakash, B. Green, S. O'Banion, and J. Xie. User-llm: Efficient llm contextualization with user embeddings. In *Proc. of ACM WWW*, pp. 1219–1223, 2025. 6

[58] OpenAI. OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf. Mar. 28. 2024]. 3

[59] B. Pätzold, J. Nogga, and S. Behnke. Leveraging vision-language models for open-vocabulary instance segmentation and tracking. *arXiv:2503.16538*, 2025. 5

[60] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[61] J. Qi, G. Liu, C. S. Jensen, and L. Kulik. Effectively learning spatial indices. *Proc. VLDB Endow.*, 13(12):2341–2354, 2020. 2

[62] L. Qiu, E. S. Kim, S. Suh, L. Sidenmark, and T. Grossman. Marginalia: Enabling in-person lecture capturing and note-taking through mixed reality. In *Proc. of ACM CHI*, 2025. 2

[63] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of ICML*, pp. 8748–8763, 2021. 3

[64] S. Rajaram, M. Peralta, J. G. Johnson, and M. Nebeling. Exploring the design space of privacy-driven adaptation techniques for future augmented reality interfaces. In *Proc of ACM CHI*, pp. 1–19, 2025. 6

[65] S. Rajaram, F. Roesner, and M. Nebeling. Reframe: An augmented reality storyboarding tool for character-driven analysis of security & privacy concerns. In *Proc. of ACM UIST*, pp. 1–15, 2023. 6

[66] I. H. Sarker and K. Salah. Appspred: Predicting context-aware smartphone apps using random forest learning. *IoT*, 8:100106, 2019. 2

[67] K. A. Satriadi, A. Cunningham, R. T. Smith, T. Dwyer, A. Drogemuller, and B. H. Thomas. Proxsituated visualization: An extended model of situated visualization using proxies for physical referents. In *Proc. of ACM CHI*, pp. 1–20, 2023. 2

[68] K. A. Satriadi, B. Tag, and T. Dwyer. Context-dependent memory in situated visualization. *arXiv:2311.12288*, 2023. 2

[69] M. Shakeri, H. Park, I. Jeon, A. Sadeghi-Niaraki, and W. Woo. User behavior modeling for ar personalized recommendations in spatial transitions. *VR*, 27(4):3033–3050, 2023. 6

[70] J. Shen, J. J. Dudley, and P. O. Kristensson. Encode-store-retrieve: Augmenting human memory through language-encoded egocentric perception. In *Proc. of IEEE ISMAR*, pp. 923–931, 2024. 1, 2

[71] L. Shen, H. Li, Y. Wang, and H. Qu. From data to story: Towards automatic animated data video creation with LLM-based multi-agent systems. *IEEE TVCG*, 2024. 2

[72] F. A. Silva, A. C. Domingues, and T. R. B. Silva. Discovering mobile application usage patterns from a large-scale dataset. *ACM TKDD*, 12(5):1–36, 2018. 2

[73] R. Tian, H. Zhai, W. Zhang, F. Wang, and Y. Guan. A survey of spatio-temporal big data indexing methods in distributed environment. *IEEE J-STARS*, 15:4132–4155, 2022. 2

[74] T. T. M. Tran, S. Brown, O. Weidlich, S. Yoo, and C. Parker. Wearable ar in everyday contexts: Insights from a digital ethnography of youtube videos. In *Proc. of ACM CHI*, 2025. 2

[75] M. Urban and C. Binnig. Caesura: Language models as multi-modal query planners. In *Proc. of CIDR*, 2024. 2

[76] Z. Wang, M. Rao, S. Ye, W. Song, and F. Lu. Towards spatial computing: recent advances in multimodal natural interaction for XR headsets. *arXiv:2502.07598*, 2025. 2

[77] Q. Xie, S. Y. Min, T. Zhang, K. Xu, A. Bajaj, R. Salakhutdinov, M. Johnson-Roberson, and Y. Bisk. Embodied-rag: General non-parametric embodied memory for retrieval and generation. *arXiv:2409.18313*, 2024. 6

[78] XREAL. Xreal. https://www.xreal.com, 2025. Jul. 16. 2025. 6

[79] J. Yang, S. Liu, H. Guo, Y. Dong, X. Zhang, S. Zhang, P. Wang, Z. Zhou, B. Xie, Z. Wang, B. Ouyang, Z. Lin, M. Cominelli, Z. Cai, Y. Zhang, P. Zhang, F. Hong, J. Widmer, F. Gringoli, L. Yang, B. Li, and Z. Liu. Egolife: Towards egocentric life assistant. In *Proc. of IEEE/CVF CVPR*, 2025. 1, 3, 6

[80] W. D. Zulfikar, S. Chan, and P. Maes. Memoro: Using large language models to realize a concise interface for real-time memory augmentation. In *Proc. of ACM CHI*, pp. 1–18, 2024. 1, 2, 3