Human-Inspired Context-Selective Multimodal Memory for Social Robots

Hangyeol Kang
Department of Computer Science, University of Geneva
Geneva, Switzerland
hangyeol.kang@unige.ch

Slava Voloshynovskiy
Department of Computer Science, University of Geneva
Geneva, Switzerland
Svyatoslav.Voloshynovskyy@unige.ch

Nadia Magnenat Thalmann
MIRALab, University of Geneva
Geneva, Switzerland
nadia.thalmann@unige.ch

ABSTRACT

Memory is fundamental to social interaction, enabling humans to recall meaningful past experiences and adapt their behavior accordingly based on the context. However, most current social robots and embodied agents rely on non-selective, text-based memory, limiting their ability to support personalized, context-aware interactions. Drawing inspiration from cognitive neuroscience, we propose a context-selective, multimodal memory architecture for social robots that captures and retrieves both textual and visual episodic traces, prioritizing moments characterized by high emotional salience or scene novelty. By associating these memories with individual users, our system enables socially personalized recall and more natural, grounded dialogue. We evaluate the selective storage mechanism using a curated dataset of social scenarios, achieving a Spearman correlation of 0.506, surpassing human consistency ($\rho = 0.415$) and outperforming existing image memorability models. In multimodal retrieval experiments, our fusion approach improves Recall@1 by up to 13% over unimodal text or image retrieval. Runtime evaluations confirm that the system maintains real-time performance. Qualitative analyses further demonstrate that the proposed framework produces richer and more socially relevant responses than baseline models. This work advances memory design for social robots by bridging human-inspired selectivity and multimodal retrieval to enhance long-term, personalized human-robot interaction.

KEYWORDS

Human-Robot Interaction; Social robots; Multimodal memory; Selective memory storage; Cognitive architecture

ACM Reference Format:
Hangyeol Kang, Slava Voloshynovskiy, and Nadia Magnenat Thalmann. 2026. Human-Inspired Context-Selective Multimodal Memory for Social Robots. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/JQIR9876

1 INTRODUCTION

Memory plays a central role in human cognition and social interaction. The ability to selectively store and recall past experiences enables people to sustain meaningful relationships, learn from previous encounters, and adapt flexibly to changing contexts [50]. For artificial agents, particularly embodied robots and assistive AI systems, equipping them with memory capabilities is essential for fostering natural, effective, and personalized interactions with humans [27].

In real-world environments, agents operate within a continuous stream of multimodal sensory information while engaging in complex social exchanges. To act as credible social partners, robots and AI need to remember not only factual events but also how and when those were meaningful to specific individuals. This requires memory systems that go beyond non-selective or text-based storage, enabling multimodal retention and retrieval of contextually meaningful experiences. For embodied agents such as social robots, contextually selective memory grounded in both perceptual and social cues is critical for developing user-aligned behavior.

Despite recent advances in conversational AI and embodied agents, most existing systems still rely on non-selective, text-based memory [49]. Although some frameworks have introduced visual memory modules, they often remain object-centric, interval-based, or non-selective—storing scene snapshots or generic visual information at fixed intervals regardless of contextual importance [17, 34]. Many also convert visual memories into textual descriptions, constraining the robot's ability to answer detailed or flexible queries when relevant cues were never explicitly transcribed [33]. This lack of contextual selectivity limits the depth of social interaction that such agents can achieve, as object- or scene-centric memory alone cannot capture what makes an experience personally or socially meaningful. Few systems prioritize what is memorable from a human perspective, neglecting factors such as emotional salience, novelty, and user-centered relevance [43]. As a result, they often miss opportunities to support richer, more personalized, and emotionally resonant interactions with humans.

Insights from cognitive neuroscience demonstrate that human memory is inherently selective: people do not store every moment of experience, but instead encode memories according to intrinsic and extrinsic cues [53]. Intrinsic cues, such as the perceptual complexity or distinctiveness of a scene, and extrinsic factors, such as emotional intensity, novelty, or personal relevance, are known to strongly influence which events are remembered and which are forgotten [4, 26, 28]. This selectivity allows people to prioritize significant, emotionally resonant, or socially important experiences, an ability that is essential for adaptive social functioning and efficient use of cognitive resources [53].

For social robots and embodied AI, adopting a similar selective approach is essential. Real-world environments are dynamic and information-rich, and storing every sensory input or conversational

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/JQIR9876

exchange is neither scalable nor meaningful. Memory systems for social agents, therefore, need to integrate both selectivity and multimodality, capturing not only textual or visual information but also when an experience holds emotional or contextual significance for the user. Without such mechanisms, artificial agents risk being overwhelmed by irrelevant details and failing to form genuine, user-aligned social connections. Consequently, there is a clear need for frameworks that integrate selective and multimodal memory storage and retrieval inspired by how humans prioritize, store, and recall their most memorable experiences.

In this work, we introduce SUMMER (Selectivity Unified Multimodal Memory for Embodied Robots), an end-to-end framework that equips social robots with context-selective and multimodal memory capabilities. Grounded in principles of human memory selectivity, the framework stores and retrieves information that is not only factual but also socially and emotionally meaningful. SUMMER operates without additional model training or fine-tuning, enabling efficient integration into diverse robotic platforms. Its lightweight and modular design ensures practical deployment in real-world social and embodied agents.

To systematically evaluate the framework, we conduct four complementary studies. First, we curate a pilot dataset of social interaction scenarios to analyze how emotional salience, novelty, and scene complexity influence selective memory storage. Second, we benchmark multimodal retrieval on public datasets (Flickr8k [13], Flickr30k [57], and MS COCO [32]), comparing fusion-based retrieval against unimodal text and image baselines. Third, we assess runtime performance using the pilot dataset to verify the framework's suitability for real-time social interaction. Finally, we perform qualitative evaluations on the same dataset, comparing responses generated by the proposed framework with those of a baseline vision–language model to illustrate its ability to produce richer and more socially grounded outputs. Together, these evaluations demonstrate the framework's capacity to capture socially meaningful experiences, retrieve contextually relevant information, and operate efficiently in dynamic, human-centered environments.

Our main technical contributions are as follows:

- We present an end-to-end framework for multimodal memory storage and retrieval in social robots, integrating selective capture guided by contextual cues.
- We design a train-free multimodal retrieval mechanism that enables robots to respond to broader, human-like memory queries by leveraging both textual and visual modalities.
- We conduct comprehensive quantitative and qualitative evaluations using both public benchmarks and a new human-annotated pilot dataset focused on socially relevant, user-centered memories.

2 RELATED WORK

2.1 Memory Systems in Social Robots

Memory enables social robots to sustain coherent interactions, personalize behavior, and adapt to human partners over time. Early approaches focused on text-based conversational memory, using context windows or prompt-based retrieval to maintain dialogue coherence in large language models (LLMs) [38]. Frameworks such as MemoryBank [60] and Memory Sandbox [18] extend this concept with long-term conversational traces, autonomous recall, and user-level inspection. However, text-centric memories often lose perceptual richness when multimodal inputs are reduced to language [20]. To enable more structured reasoning, symbolic or knowledge-based systems explicitly represent events and facts using graphs or databases. Examples include ChatDB, which integrates SQL-based retrieval for multi-hop reasoning [15], and ROSA, which dynamically adapts robot behaviors based on stored contextual knowledge [45].

Recent studies emphasize multimodal memory, integrating visual, auditory, and behavioral data to bridge perception and language [6]. The Nadine robot combines LLM reasoning with multimodal perception to retrieve user-specific episodic memories and simulate emotional states [21], while LLM-Brain unifies perception, planning, and memory through interconnected multimodal models [35]. Other systems leverage paralinguistic and behavioral cues to infer personality traits or improve interaction quality in specific contexts such as sign language communication [31], as well as socially assistive settings that rely on contextual user information [1], underscoring the importance of perceptual grounding and social sensitivity in memory design [20].

Despite these advances, most frameworks still store information indiscriminately or at fixed intervals, lacking selectivity based on emotional salience, novelty, or social relevance [14]. Text-dominant designs further omit non-verbal and contextual cues critical for nuanced social understanding [52]. Few systems explicitly connect memory to emotional or social context, limiting robots' capacity to interpret atmosphere, respect norms, and adapt to user personality [21]. These challenges highlight the need for more human-aligned mechanisms for selective and contextually grounded memory formation.

2.2 Image Memorability and Selective Memory

Beyond robotic architectures, research on image memorability offers complementary insights into how selective encoding emerges in human cognition. Large-scale studies such as LaMem [23], SUN Memorability [19], and FIGRIM [5] demonstrate that memorability is an intrinsic and consistent property of images across observers. Early work showed that low-level features like color or contrast are poor predictors, whereas semantic attributes—such as the presence of people or faces and overall scene structure—strongly influence recall [19]. Emotional salience and contextual distinctiveness further enhance memorability, suggesting that recall depends on meaningful, socially grounded features [4].

Deep learning models such as MemNet [23], AMNet [8], and ViTMem [12] now approach human-level prediction accuracy, yet remain centered on static, object- or scene-focused datasets and overlook social or interactive dynamics. From a cognitive perspective, memory is a selective process shaped by affective and contextual cues rather than a passive record [37]. Emotion and novelty determine which experiences are encoded and consolidated [29, 51], while social significance further governs what is remembered [39]. Robotic memory systems, however, typically focus solely on visual salience [25], limiting their ability to capture the interactive and social nature of real-world memory.

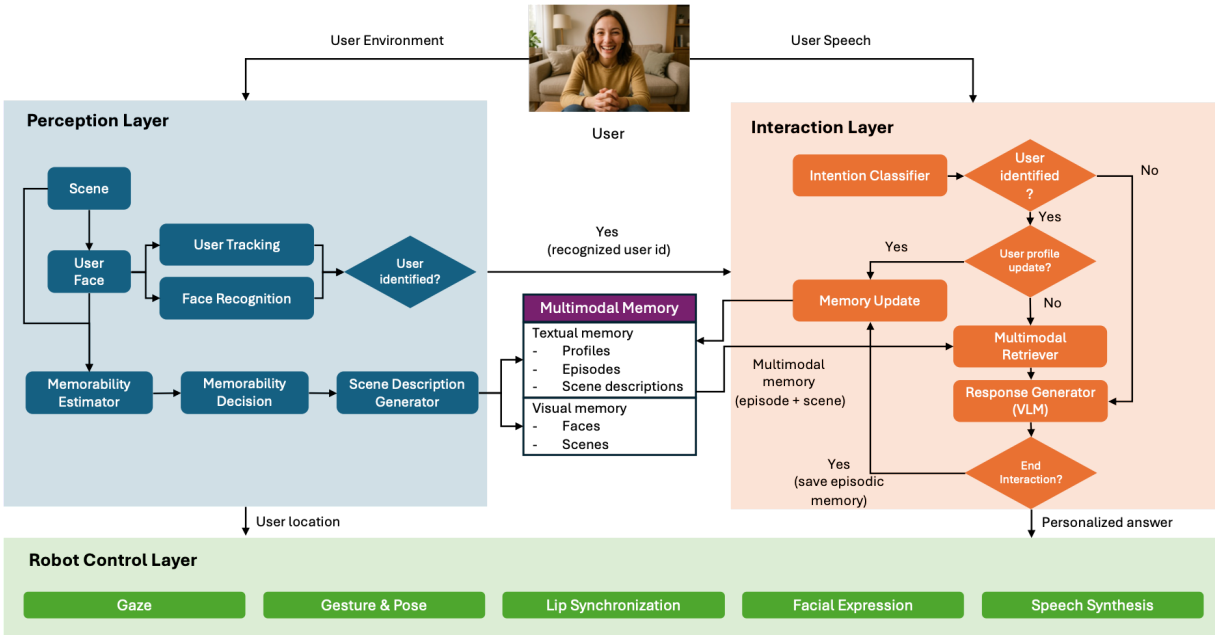

Figure 1: Overview of the SUMMER architecture for selective multimodal memory in social robots. The perception layer analyzes the user's scene and face to estimate memorability; memorable scenes are stored along with generated scene descriptions. The interaction layer manages and retrieves contextually significant multimodal memories, enabling socially relevant and personalized responses.

3 METHOD

3.1 Overview of the SUMMER Framework

The SUMMER framework is structured into three main layers: the Perception Layer, the Interaction Layer, and the Control Layer, as depicted in Figure 1.

Perception Layer. This layer analyzes the user's environment and facial expressions to estimate the memorability of each encountered scene. When a scene meets predefined selectivity criteria, it is encoded and stored in the multimodal memory database along with a textual scene description automatically generated by a lightweight vision-language model. These descriptions later facilitate efficient and contextually relevant retrieval during user interactions. The detailed mechanism for selective memory storage is discussed in Section 3.2.

Interaction Layer. The interaction layer manages context-aware retrieval and episodic memory updates. Leveraging the reasoning capabilities of LLMs, the system first classifies the user's intention based on the incoming query, for example, updating a user profile, ending a conversation, or continuing the dialogue. Memory-related modules (memory update and memory retrieval) are activated once the user is identified. Upon identification, user-specific information is retrieved from the multimodal memory database, enabling the response generation module to produce more personalized and contextually grounded replies. Details on intention classification, user identification, and multimodal retrieval processes are provided in Section 3.3 and 3.4.

Control Layer. The control layer translates high-level interaction outcomes into embodied robot behaviors, such as gaze direction, gestures, and spoken responses, grounded in the context provided by the perception and interaction layers. While not the primary focus of this work, this layer enables the social robot to deliver personalized and socially intelligent responses.

3.2 Selective Memory Storage

Human memory is inherently selective, shaped by both intrinsic and extrinsic factors that determine which experiences are encoded and retained [56]. Inspired by this cognitive principle, our framework implements a context-selective storage mechanism driven by two complementary cues: emotional salience and scene novelty. Although we initially explored scene complexity as a potential intrinsic cue, empirical results showed that it did not contribute meaningfully to the selection of memorable moments in socially interactive scenarios. We therefore exclude it from the final system, focusing instead on the two cues that most robustly predict socially relevant memorability.

Emotion Estimation. Emotional salience plays a central role in memory formation, particularly in social contexts [51]. To capture this signal, we integrate a face detection and emotion recognition module that analyzes the most prominent face in each frame. (assumed to correspond to the primary user). For the detected face, we estimate both the discrete emotion category $\mathcal{E}$ = neutral, happy,

sad, surprise, fear, disgust, anger, contempt and the corresponding intensity score p_k for each emotion $k \in \mathcal{E}$.

Recognizing that not all emotions contribute equally to memorability [23], we adopt emotion-specific intensity thresholds $T_{e,k}$ that are empirically optimized through nested cross-validation. This approach allows the system to weight emotional categories according to their distinct contributions to memorability, rather than applying a uniform threshold across all emotions. It also accounts for cognitive findings that negative or high-arousal emotions tend to facilitate memory encoding more strongly than positive or low-arousal states [23, 36].

For each emotion k, the emotion-specific salience score s_k is defined as the normalized activation above its threshold:

$$s_k = \max\left(0, \frac{p_k - T_{e,k}}{1 - T_{e,k}}\right) \tag{1}$$

and the overall emotional salience of a frame is then determined by the maximum salience across all emotion categories:

$$e = \max_k \{s_k \mid k \in \mathcal{E}\} \tag{2}$$

A frame satisfies the emotional salience condition if $e > 0$, i.e., if the intensity of at least one emotion exceeds its category-specific threshold. This formulation ensures that even a non-dominant but highly salient emotion can trigger memory encoding. The resulting optimized thresholds and their implications for selective formation are analyzed in detail in Section 4.

Novelty Estimation. Novelty serves as a complementary intrinsic cue, reflecting the distinctiveness of the current scene relative to previously encountered scenes. To quantify this, we represent each scene as an embedding vector f_{scene} extracted from a vision encoder. We then compute the cosine distance between the current scene and all previously stored scenes f_{scene_i}, and use the minimum distance as the novelty score:

$$n = \min_i \ \text{dist}(f_{\text{scene}}, f_{\text{scene}_i}) \tag{3}$$

where $\text{dist}(\cdot,\cdot)$ denotes cosine distance. The first scene for each user is always stored as a reference, providing a meaningful baseline for subsequent comparisons.

A new scene is considered novel if its distance from all stored scenes exceeds a novelty threshold T_n. This threshold-based decision rule ensures that only sufficiently distinct and non-redundant events are encoded into memory, preventing the database from being populated with repetitive or trivial scenes. In this way, the novelty mechanism complements the emotional salience module by capturing contextually meaningful moments that differ substantially from prior observations.

Selective Memory Capture. A scene is marked as memorable if it meets either the emotional salience condition or the novelty condition. For each selected moment, a lightweight vision-language model generates a short textual caption describing the event. Both the raw image and its caption, as well as their corresponding visual and textual embeddings, are stored in the multimodal memory database. These stored traces support contextually grounded retrieval and enable the system to reference past experiences in future interactions:

$$\text{Memorable} = \begin{cases} 1, & \text{if } e > 0 \text{ or } n > T_n \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Algorithm 1 Multimodal Hybrid Memory Retrieval in SUMMER

Input: Query q; precomputed episode features $\{v_i^{\text{text}}\}$, scene features $\{v_j^{\text{mm}}\}$, and scene description features $\{d_j^{\text{text}}\}$; episode timestamps $\{t_i^e\}$; scene timestamps $\{t_j^s\}$
Parameter: Text encoder E_{text}, multimodal encoder E_{mm}, scene similarity weight $\alpha \in [0,1]$, small $\varepsilon > 0$
Output: Retrieved episode e^* and scene s^*

Encode query: $v_q^{\text{text}} = E_{\text{text}}(q), \quad v_q^{\text{mm}} = E_{\text{mm}}(q)$
Compute episode similarities: $S_i^{\text{ep}} = \text{cosine}(v_q^{\text{text}}, v_i^{\text{text}}) \quad \text{for all } i$
Compute scene similarities:
- $S_j^{\text{img}} = \text{cosine}(v_q^{\text{mm}}, v_j^{\text{mm}})$
- $S_j^{\text{desc}} = \text{cosine}(v_q^{\text{text}}, d_j^{\text{text}})$
- $S_j^{\text{scene}} = \alpha S_j^{\text{img}} + (1-\alpha) S_j^{\text{desc}}$

Z-score normalize within each modality:
$\widehat{S}_i^{\text{ep}} = \dfrac{S_i^{\text{ep}} - \mu_{\text{ep}}}{\sigma_{\text{ep}} + \varepsilon} \quad \text{for all } i$ (means/SDs over $\{S_i^{\text{ep}}\}$)
$\widehat{S}_j^{\text{scene}} = \dfrac{S_j^{\text{scene}} - \mu_{\text{scene}}}{\sigma_{\text{scene}} + \varepsilon} \quad \text{for all } j$ (means/SDs over $\{S_j^{\text{scene}}\}$)
$i^* = \arg\max_i \widehat{S}_i^{\text{ep}}, \quad j^* = \arg\max_j \widehat{S}_j^{\text{scene}}$
if $\widehat{S}_{i^*}^{\text{ep}} > \widehat{S}_{j^*}^{\text{scene}}$ **then**
  $e^* \leftarrow i^*; \quad t^* \leftarrow t_{i^*}^e$
  $s^* \leftarrow \arg\min_j \left| t_j^s - t^* \right|$
else
  $s^* \leftarrow j^*; \quad t^* \leftarrow t_{j^*}^s$
  $e^* \leftarrow \arg\min_i \left| t_i^e - t^* \right|$
end if
return e^*, s^*

3.3 Multimodal Hybrid Memory Retrieval

To support flexible and socially intelligent interaction, the SUMMER framework incorporates a hybrid retrieval mechanism that leverages both textual (conversation-based) and visual (scene-based) memories. This design enables the system to address a wide range of user queries, including references to past events, visual details, or social contexts.

When a user query is received, the system encodes the input using both a dedicated text encoder and a multimodal encoder, generating embeddings tailored to each memory modality. For textual retrieval, the query embedding generated by the text encoder is compared via cosine similarity to all user-specific textual conversation embeddings stored in the database. For visual retrieval, two similarity measures are computed: (1) the similarity between the query embedding from the multimodal encoder and stored scene embeddings, and (2) the similarity between the query embedding from the text encoder and textual scene descriptions. These two scores are then combined as a weighted sum to produce an overall scene similarity score.

To ensure fair comparison between modalities, the top similarity scores for both conversation and scene memories are z-score normalized within their respective pools. The system then selects the memory item (conversation or scene) with the highest normalized similarity as the primary retrieval result. The timestamp associated with this selected item is then used to retrieve the closest

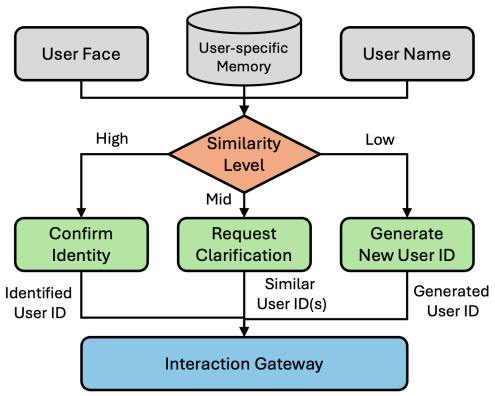

Figure 2: Overview of the user identification process

corresponding item from the other modality, thereby constructing a coherent and context-rich memory pair.

Finally, the retrieved textual (conversation) and visual (scene) memories are jointly processed by a vision-language model (VLM) to generate the final response. This integrated retrieval strategy enables the system to handle a wide spectrum of queries, whether users ask about past conversations, specific visual scenes, or require answers that bridge both modalities, thereby supporting more natural and human-like interactions.

3.4 Supporting Modules

To enable seamless operation of the memory framework in real-world social robot applications, we incorporate several supporting modules beyond our core contributions.

Intention Classifier. The dialogue flow begins with an intention classifier that leverages LLM reasoning to infer the purpose of a user's utterance, reducing computational overhead by activating only the necessary downstream modules. Each input is categorized as *ProfileUpdate*, *SessionEnd*, or *Continue*. *ProfileUpdate* is assigned when the user provides personal information (e.g., name, city, occupation, interests), triggering the memory update module. *SessionEnd* indicates dialogue termination, prompting storage of the current exchange as an episodic memory. All remaining queries are labeled *Continue*; if the user is identified, relevant user-specific memories are retrieved before response generation, otherwise the query is passed directly to the response generator. Identity verification is described in the following section.

User Identification. Reliable user identification underpins consistent personalization and memory retrieval. The system verifies identity via both name and face similarity before entering the dialogue flow (Figure 2). Name matching uses the Levenshtein ratio [30]: scores above 0.8 confirm identity, while scores between 0.6 and 0.8 trigger a brief clarification prompt. Facial verification compares embeddings from the buffalo_s model in InsightFace [2] using a threshold of 0.5.

If no match is found, a new user ID (e.g., 251008_0001) is created and initialized in memory. Once identification is confirmed or established, control passes to the interaction gateway, enabling retrieval of user-specific information for personalized responses.

To protect privacy, memory operations are activated only after users voluntarily provide their name, and all stored profiles or episodic records can be permanently deleted upon request, ensuring transparency and user trust.

4 EXPERIMENTS

4.1 Experimental Setup

All experiments are conducted on 2×GeForce RTX 4090 GPUs using the PyTorch deep learning framework. The baseline VLM for response generation is Mistral-small3.2 [40]. For emotion detection, we employ OpenFace3.0 [16]. Novelty is assessed with CLIP-ViT-L14 [44], and scene complexity is estimated with ICNet [9]. Text features are extracted with the jina-embeddings-v4 model [11], while visual encoding utilizes SigLip-so400m-P14 [58]. Scene descriptions for memorable frames are generated by Moondream2 [41]. All feature embeddings and multimodal memory storage are managed using ChromaDB as the vector database.

4.2 Selective Memory Storage Evaluation

Curated Pilot Dataset. To specifically assess our selective memory storage module, focusing on the memorability of social interactions and contextually significant moments, we constructed a custom pilot dataset. This dataset contains 81 images depicting a wide range of social scenarios, systematically generated using Sora, OpenAI's text-to-video generative model [3]. Images were designed to vary in user demographics, emotional expressions, and background contexts to reflect real-world diversity. For ground truth annotation, we recruited 25 human raters, who each evaluated all images for memorability on a 1-9 Likert scale. To minimize order and recency effects, images were shown in randomized order, each for 1 second. This duration follows standard memorability protocols used in prior work, which measure rapid scene encoding rather than detailed inspection [19, 23]. This dataset was also employed in our runtime analysis and for qualitative evaluation of our SUMMER framework.

We deliberately adopted a synthetic dataset because no existing resource captures the socially grounded and interaction-centric scenes needed to assess selective memory in social robots. Synthetic generation enabled precise control over key factors such as emotional salience, novelty, and scene complexity, which are difficult to manipulate systematically in real-world data. Although synthetic images may differ from natural distributions, their use is an established practice across fields, including medical research [7], visual question answering [24], and autonomous driving [47], where they enable controlled evaluation and reproducibility. This pilot dataset was also used in our runtime analysis and qualitative evaluation of the SUMMER framework.

Human Consistency. To contextualize the performance of our model, we estimated human consistency by computing the average Spearman correlation [48] between each participant's ratings and the mean rating of all other annotators. This yielded a human consistency score of $\rho = 0.4152$, which reflects the reliability of human judgments on our dataset and serves as a practical reference point

for model evaluation. Although our model achieves a higher correlation with the aggregated human ratings than this value, this does not imply that it surpasses human memory performance. Instead, it indicates that the human consistency score represents a baseline indicator of inter-annotator agreement rather than a strict upper bound.

Baselines. We compare our framework against heuristic baselines and state-of-the-art image memorability models. The heuristic baselines include random selection (uniform scores in [0,1]) and interval-based selection, which marks frames as memorable at fixed temporal intervals (e.g., every fifth or tenth frame). Both were evaluated using the same 5-fold stratified outer cross-validation with 20 repeats as our main experiments. We also evaluate two representative memorability models, ResMem [42] and ViTMem [12], which predict intrinsic memorability from visual features. For each test image, we compute predicted scores and measure Spearman correlation with human ratings using the same protocol. Together, these baselines provide heuristic and intrinsic memorability references for comparison with our approach.

Model Output. For each image in the evaluation set, our model produced a continuous memorability score based on the weighted combination of emotional salience, scene novelty, and visual complexity. Emotion probabilities were obtained using the OpenFace model [16] applied to the largest detected face in the scene, and visual complexity was estimated with the pretrained model described in the Experimental Setup.

Novelty estimation was adapted for the evaluation phase to reflect comparative distinctiveness better. Unlike the deployment setting, where novelty is defined as the embedding distance between the current scene and previously stored ones, we employed a repeated burn-in approach to mitigate inflated scores for early images. In each run, the first $k = 5$ images were excluded from novelty computation. For every subsequent image, we computed the minimum cosine distance to all preceding images in that run. This run was repeated 1000 times with randomly shuffled image sequences, and the final novelty score for each image was obtained by averaging across all repetitions. These novelty scores were then used in combination with a learned novelty threshold T_n to determine whether a scene was sufficiently distinct.

The overall memorability score was computed as:

$$\text{MemScore} = w_e \cdot S_e + w_n \cdot S_n + w_c \cdot S_c \quad (5)$$

where S_e, S_n, and S_c are the emotion, novelty, and complexity scores, respectively, and w_e, w_n, w_c are their respective weights.

We evaluated the model using repeated stratified nested cross-validation (5 outer folds, 3 inner folds, 20 repeats). For each weight configuration, the inner loop searched for the optimal per-emotion thresholds and novelty threshold that maximized Spearman correlation with human ratings. These optimal parameters were then applied to the outer test folds to generate memorability predictions, which were compared against human scores using fold-level Spearman correlations and Fisher-combined p-values [10].

Results. Table 1 summarizes the predictive performance of all evaluated models in terms of Spearman correlation with human memorability ratings and their associated statistical significance. Heuristic approaches such as random and interval-based selection exhibited near-zero or weak correlations, indicating that memorability cannot be reliably predicted through simple time-based or random sampling strategies. Image memorability models originally trained on large-scale datasets, including ResMem [42] and ViTMem [12], also failed to generalize to our socially grounded interaction scenarios, showing no significant correlation with human judgments.

Table 1: Spearman correlation (ρ), standard deviation, and significance level (p_{Fisher}) for human consistency, baseline models, and SUMMER variants. For SUMMER, the triplet denotes (w_{emotion}, w_{novelty}, $w_{\text{complexity}}$) in the memorability score. Significance: ** $p < 10^{-100}$, * $p < 10^{-10}$, ns = not significant. Bold indicates the best-performing configuration.

Category	Approach	Mean ρ	Std ρ	p_{Fisher}
Human	Consistency	0.415	—	—
Heuristic	Random Selection	-0.000	0.269	ns
	Interval (n=5)	-0.052	0.226	ns
	Interval (n=10)	0.013	0.230	ns
Image Mem.	ResMem [42]	-0.055	0.249	ns
	ViTMem [12]	0.046	0.235	ns
SUMMER	**(0.5, 0.5, 0.0)**	**0.506**	0.166	**
	(1.0, 0.0, 0.0)	0.474	0.170	**
	(0.5, 0.3, 0.2)	0.447	0.170	**
	(0.5, 0.0, 0.5)	0.319	0.188	*
	(0.0, 1.0, 0.0)	-0.014	0.201	ns
	(0.0, 0.0, 1.0)	-0.037	0.249	ns

Although scene complexity was initially considered as an intrinsic cue, it showed little predictive value ($\rho = -0.037$). Novelty alone also performed poorly ($\rho = -0.014$). In contrast, combining emotional salience with novelty improved alignment with human ratings ($\rho = 0.506$) compared to emotion alone ($\rho = 0.474$), while adding complexity reduced performance ($\rho = 0.319$). These results indicate that scene complexity is not a useful predictor of memorability in social-interaction contexts.

To verify that this performance drop was not due to implementation issues, we additionally evaluated ResMem [42], ViTMem [12], and SUMMER on the LaMem benchmark (16,810 test images). ResMem achieved ρ=0.8145, ViTMem ρ=0.7583, and SUMMER ρ=0.2062 (pooled Spearman), consistent with each model's intended domain. This confirms correct integration and indicates that the low performance on our pilot dataset arises from domain mismatch rather than experimental error.

In contrast, all variants of the proposed SUMMER framework demonstrated substantial improvements over baseline approaches. The best-performing configuration, which combined emotion and novelty cues with weights (0.5, 0.5, 0.0), achieved a mean Spearman correlation of 0.506, significantly surpassing both heuristic and pretrained memorability baselines. Several other configurations also consistently exceeded the human consistency level, highlighting the robustness of the proposed memory encoding strategy.

Table 2: Recall@K (%) for *Image*, *Text*, and *Fusion* (inline best α) across datasets. Bold indicates the best per dataset.

Image Encoder	Text Encoder	Mode	Flickr8k			Flickr30k			MS COCO		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP ViT-L/14 [44]	mE5-large [54]	Text	49.2	76.8	85.6	54.6	80.2	86.6	28.8	52.2	63.3
		Image	61.6	86.3	92.8	64.7	87.1	92.1	36.5	61.1	71.1
		Fusion (0.7)	**68.4**	**90.5**	**95.4**	**72.2**	**90.8**	**95.2**	**42.1**	**66.9**	**76.5**
	mGTE-base [59]	Text	51.9	78.8	87.2	55.6	79.8	86.1	30.9	54.7	65.2
		Image	61.6	86.3	92.8	64.7	87.1	92.1	36.5	61.1	71.1
		Fusion (0.7)	**69.0**	**90.8**	**95.8**	**72.3**	**91.2**	**95.3**	**42.6**	**67.2**	**77.0**
	jina-embeddings-v4 [11]	Text	56.2	82.0	89.6	59.6	82.6	88.5	36.0	61.3	71.9
		Image	61.6	86.3	92.8	64.7	87.1	92.1	36.5	61.1	71.1
		Fusion (0.7)	**68.9**	**91.3**	**95.9**	**72.6**	**91.1**	**95.3**	**43.9**	**69.0**	**78.0**
MetaCLIP-B/32-400M [55]	mE5-large [54]	Text	49.2	76.8	85.6	54.6	80.2	86.6	28.8	52.2	63.3
		Image	58.1	84.1	91.2	62.3	85.5	91.4	35.9	61.8	72.2
		Fusion (0.7)	**65.6**	**88.9**	**94.4**	**70.9**	**90.9**	**94.6**	**41.9**	**67.7**	**77.4**
	mGTE-base [59]	Text	51.9	78.8	87.2	55.6	79.8	86.1	30.9	54.7	65.2
		Image	58.1	84.1	91.2	62.3	85.5	91.4	35.9	61.8	72.2
		Fusion (0.7)	**66.1**	**89.6**	**95.1**	**71.7**	**90.7**	**94.7**	**42.3**	**68.2**	**77.6**
	jina-embeddings-v4 [11]	Text	56.2	82.0	89.6	59.6	82.6	88.5	36.0	61.3	71.9
		Image	58.1	84.1	91.2	62.3	85.5	91.4	35.9	61.8	72.2
		Fusion (0.5)	**67.2**	**90.1**	**95.1**	**72.0**	**90.7**	**94.3**	**44.6**	**70.0**	**79.5**
SigLIP-so400m-384 [58]	mE5-large [54]	Text	49.2	76.8	85.6	54.6	80.2	86.6	28.8	52.2	63.3
		Image	66.1	86.9	92.5	71.7	89.2	93.3	41.8	65.4	74.5
		Fusion (0.7)	**70.9**	**90.4**	**95.2**	**76.1**	**92.5**	**95.5**	**45.6**	**69.3**	**78.0**
	mGTE-base [59]	Text	51.9	78.8	87.2	55.6	79.8	86.1	30.9	54.7	65.2
		Image	66.1	86.9	92.5	71.7	89.2	93.3	41.8	65.4	74.5
		Fusion (0.7)	**71.6**	**90.7**	**95.4**	**76.4**	**92.4**	**95.8**	**46.1**	**70.2**	**78.9**
	jina-embeddings-v4 [11]	Text	56.2	82.0	89.6	59.6	82.6	88.5	36.0	61.3	71.9
		Image	66.1	86.9	92.5	71.7	89.2	93.3	41.8	65.4	74.5
		Fusion (0.7)	**71.8**	**90.9**	**95.2**	**75.8**	**92.1**	**95.6**	**47.3**	**71.2**	**79.9**

4.3 Multimodal Retrieval Evaluation

Datasets. We evaluated the proposed multimodal retrieval mechanism on three standard benchmarks: Flickr8k [13], Flickr30k [57], and MS COCO [32]. For each dataset, we used the standard Karpathy test splits [22] and generated a concise textual description for each test image using a lightweight vision-language model.

Baselines. We compared our method against two unimodal retrieval settings: (1) Textual retrieval, which measures cosine similarity between the query embedding and the textual scene descriptions, and (2) Visual retrieval, which computes similarity between the query embedding and the image embeddings. These baselines isolate the contribution of each modality and serve as reference points for evaluating the proposed fusion method.

Multimodal Fusion. To leverage the complementary strengths of both modalities, we combine normalized similarity scores from the visual and textual representations using a weighted fusion scheme. The final score is computed as

$$\text{sim}_{\text{final}} = \alpha \cdot \text{sim}_{\text{img\_norm}} + (1 - \alpha) \cdot \text{sim}_{\text{text\_norm}} \quad (6)$$

where α controls the relative contribution of each modality. We varied α from 0.0 to 1.0 in increments of 0.1 and reported the best-performing value for each image-text encoder combination. Higher α values emphasize visual similarity, while lower values prioritize textual alignment.

Results. Table 2 summarizes Recall@K performance across datasets. The proposed multimodal fusion consistently outperforms both unimodal baselines, with best results typically obtained around $\alpha = 0.7$. This pattern holds across different image and text encoder pairs, confirming that even a simple, train-free fusion can effectively exploit the complementary strengths of visual and textual representations for robust multimodal retrieval.

4.4 Runtime Analysis

Efficient response time is essential for maintaining natural interaction in social robots. In human–computer interaction, systems are generally expected to respond within two seconds to preserve conversational flow and prevent user frustration [46]. We evaluated the runtime performance of the baseline and SUMMER system over 100 trials, reporting the mean and standard deviation. All tests used 768×512 pixel inputs. The baseline employed a VLM without memory retrieval or scene analysis, while SUMMER incorporated multimodal memory operations.

As shown in Table 3, SUMMER achieved an average generation time of 0.87 ± 0.16 seconds per response, remaining well below the two-second limit. The additional 0.4 seconds over the baseline correspond to retrieval and image processing, enabling richer and more contextually grounded replies. The perception layer required 0.37 ± 0.66 seconds per memorable scene (Table 4), which is acceptable since memorable events occur intermittently due to the novelty criterion. Overall, SUMMER maintains real-time performance while substantially enhancing contextual and social relevance in robot interactions.

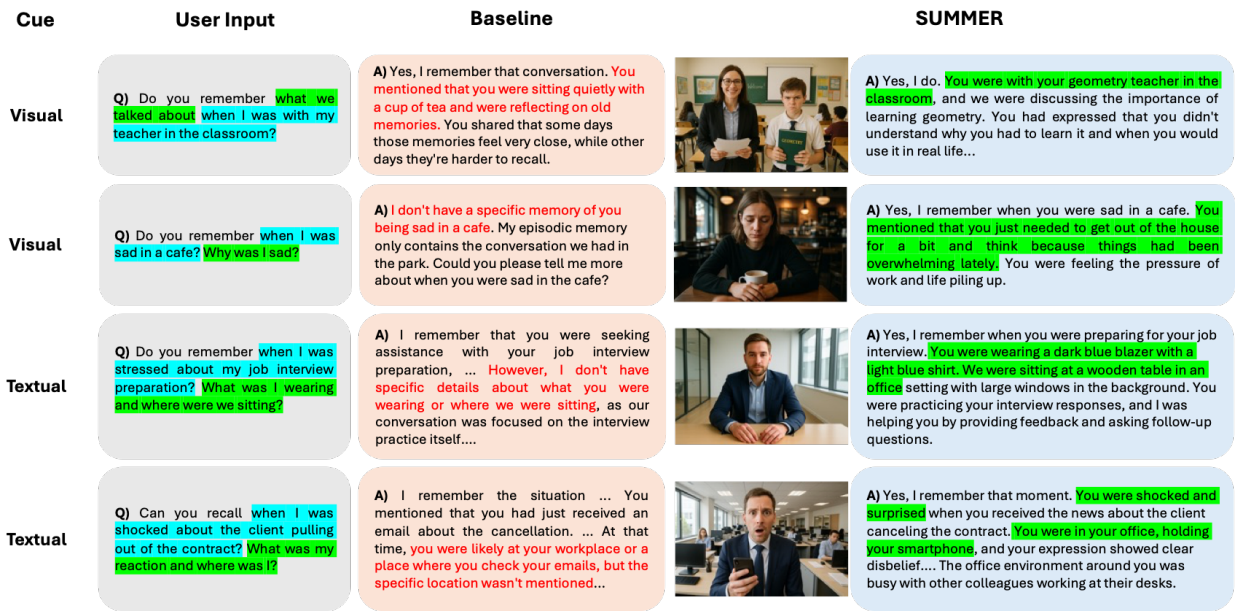

Figure 3: Qualitative comparison of response generated by the baseline model (left) and the SUMMER-augmented system (right) across diverse user queries. Each sample shows the query, both responses, and the retrieved image (for SUMMER), along with the retrieved cue type.

Table 3: Runtime of modules in the interaction layer. 'Generation' refers to response generation, 'Retrieval' to multimodal retrieval, and 'Total Time' includes all modules in the interaction layer.

System	Module	Time (s)	Total Time (s)
Baseline	Generation	0.43 ± 0.21	0.44 ± 0.00
SUMMER	Retrieval	0.03 ± 0.00	
	Generation	0.69 ± 0.16	0.86 ± 0.16

Table 4: Runtime of modules in the perception layer. 'Total Time' indicates the cumulative runtime for the entire perception pipeline.

Module	Time (s)	Total Time (s)
Memorability Decision	0.08 ± 0.01	
Scene Description Generator	0.25 ± 0.03	
Multimodal Encoding	0.05 ± 0.00	0.37 ± 0.03

4.5 Qualitative Results

To illustrate SUMMER's advantages, we present qualitative comparisons between baseline response generation (textual memory only) and SUMMER's multimodal approach (Figure 3). When retrieval requires visual memory, such as recalling specific scenes or emotional contexts, the baseline often fails, retrieving unrelated episodes or producing incorrect answers. In contrast, SUMMER accurately recalls relevant visual memory and generates contextually grounded responses. For text-based queries, the baseline may recover the correct episode but cannot provide visually grounded details, frequently guessing or omitting visual information. By integrating both textual and visual memory, SUMMER consistently delivers more informative and socially relevant responses across both query types.

5 CONCLUSION

In this work, we introduced SUMMER, a train-free framework for selective multimodal memory storage and retrieval for social robots. Inspired by human memory selectivity, our system enables robots to capture and recall socially and emotionally significant moments, supporting more natural and context-aware interactions. Benchmarking on public and curated datasets shows that our selective memory storage mechanism, particularly the use of emotional and novelty cues, aligns closely with human memorability. Furthermore, our multimodal retrieval module consistently outperforms standard text-only and image-only baselines. For future work, we aim to expand SUMMER's selectivity beyond emotional salience, novelty, and scene complexity by incorporating additional factors identified in memorability research [4], such as social relevance, interpersonal relationships, and personalized significance.

ACKNOWLEDGMENTS

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them. INDUX-R project DOI 10.3030/101135556. Funded from the Swiss State Secretariat for Education, Research and Innovation (SERI).

REFERENCES

[1] Shivendra Agrawal, Suresh Nayak, Ashutosh Naik, and Bradley Hayes. 2024. ShelfHelp: Empowering humans to perform vision-independent manipulation tasks with a socially assistive robotic cane. *arXiv preprint arXiv:2405.20501* (2024).

[2] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. 2021. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1445–1449.

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. 2024. Video generation models as world simulators. *OpenAI Blog* 1, 8 (2024), 1.

[4] Zoya Bylinskii, Lore Goetschalckx, Anelise Newman, and Aude Oliva. 2021. Memorability: An Image-Computable Measure of Information Utility. https://doi.org/10.48550/arXiv.2104.00805 arXiv:2104.00805 [cs]

[5] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. 2015. Intrinsic and Extrinsic Effects on Image Memorability. *Vision Research* 116 (Nov. 2015), 165–178. https://doi.org/10.1016/j.visres.2015.03.005

[6] John A. Duncan, Farshid Alambeigi, and Mitchell W. Pryor. 2024. A Survey of Multimodal Perception Methods for Human–Robot Interaction in Social Environments. *ACM Transactions on Human-Robot Interaction* 13, 4 (Dec. 2024), 1–50. https://doi.org/10.1145/3657030

[7] Khaled El Emam, Lucy Mosquera, and Jason Bass. 2020. Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *Journal of medical Internet research* 22, 11 (2020), e23139.

[8] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. AMNet: Memorability Estimation with Attention. https://doi.org/10.48550/arXiv.1804.03115 arXiv:1804.03115 [cs]

[9] Tinglei Feng, Yingjie Zhai, Jufeng Yang, Jie Liang, Deng-Ping Fan, Jing Zhang, Ling Shao, and Dacheng Tao. 2022. Ic9600: A benchmark dataset for automatic image complexity assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 7 (2022), 8577–8593.

[10] Ronald Aylmer Fisher. 1970. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*. Springer, 66–70.

[11] Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, et al. 2025. jina-embeddings-v4: Universal Embeddings for Multimodal Multilingual Retrieval. *arXiv preprint arXiv:2506.18902* (2025).

[12] Thomas Hagen and Thomas Espeseth. 2023. Image Memorability Prediction with Vision Transformers. https://doi.org/10.48550/arXiv.2301.08647 arXiv:2301.08647 [cs]

[13] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.

[14] Yuki Hou, Haruki Tamoto, and Homei Miyashita. 2024. "My Agent Understands Me Better": Integrating Dynamic Human-like Memory Recall and Consolidation in LLM-Based Agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–7. https://doi.org/10.1145/3613905.3650839

[15] Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. 2023. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901* (2023).

[16] Jiewen Hu, Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2025. OpenFace 3.0: A Lightweight Multitask System for Comprehensive Facial Behavior Analysis. *arXiv preprint arXiv:2506.02891* (2025).

[17] Yifei Huang, Jilan Xu, Baoqi Pei, Yuping He, Guo Chen, Lijin Yang, Xinyuan Chen, Yaohui Wang, Zheng Nie, Jinyao Liu, et al. 2024. Vinci: A real-time embodied smart assistant based on egocentric vision-language model. *arXiv preprint arXiv:2412.21080* (2024).

[18] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen MacNeil. 2023. Memory Sandbox: Transparent and Interactive Memory Management for Conversational Agents. https://doi.org/10.48550/arXiv.2308.01542 arXiv:2308.01542 [cs]

[19] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. 2011. Understanding the Intrinsic Memorability of Images. (2011).

[20] Ruben Janssens and Tony Belpaeme. 2025. Towards Multimodal Social Conversations with Robots: Using Vision-Language Models. https://doi.org/10.48550/arXiv.2507.19196 arXiv:2507.19196 [cs]

[21] Hangyeol Kang, Maher Ben Moussa, and Nadia Magnenat-Thalmann. 2024. Nadine: An LLM-driven Intelligent Social Robot with Affective Capabilities and Human-like Memory. https://doi.org/10.48550/arXiv.2405.20189 arXiv:2405.20189 [cs]

[22] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[23] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and Predicting Image Memorability at a Large Scale. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Santiago, Chile. https://doi.org/10.1109/iccv.2015.275

[24] Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. 2025. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *Comput. Surveys* 57, 10 (2025), 1–35.

[25] Taewoon Kim, Michael Cochez, Vincent Francois-Lavet, Mark Neerincx, and Piek Vossen. 2024. A Machine With Human-Like Memory Systems. https://doi.org/10.48550/arXiv.2204.01611 arXiv:2204.01611 [cs]

[26] Max A Kramer, Martin N Hebart, Chris I Baker, and Wilma A Bainbridge. 2023. The features underlying the memorability of objects. *Science advances* 9, 17 (2023), eadd2981.

[27] Taeyoon Kwon, Dongwook Choi, Sunghwan Kim, Hyojun Kim, Seungjun Moon, Beong-woo Kwak, Kuan-Hao Huang, and Jinyoung Yeo. 2025. Embodied Agents Meet Personalization: Exploring Memory Utilization for Personalized Assistance. *arXiv preprint arXiv:2505.16348* (2025).

[28] Cameron Kyle-Davidson, Oscar Solis, Stephen Robinson, Ryan Tze Wang Tan, and Karla K Evans. 2025. Scene complexity and the detail trace of human long-term visual memory. *Vision Research* 227 (2025), 108525.

[29] Ryan T LaLumiere, James L McGaugh, and Christa K McIntyre. 2017. Emotional modulation of learning and memory: pharmacological implications. *Pharmacological reviews* 69, 3 (2017), 236–255.

[30] VI Lcvenshtcin. 1966. Binary coors capable or 'correcting deletions, insertions, and reversals. In *Soviet physics-doklady*, Vol. 10.

[31] Jie Li, Junpei Zhong, and Ning Wang. 2023. A Multimodal Human-Robot Sign Language Interaction Framework Applied in Social Robots. *Frontiers in Neuroscience* 17 (April 2023), 1168888. https://doi.org/10.3389/fnins.2023.1168888

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[33] Weihao Liu, Fangyu Lei, Tongxu Luo, Jiahe Lei, Shizhu He, Jun Zhao, and Kang Liu. 2023. MMHQA-ICL: Multimodal in-context learning for hybrid question answering over text, tables and images. *arXiv preprint arXiv:2309.04790* (2023).

[34] Yang Liu, Xinshuai Song, Kaixuan Jiang, Weixing Chen, Jingzhou Luo, Guanbin Li, and Liang Lin. 2024. Meia: Multimodal embodied perception and interaction in unknown environments. *arXiv preprint arXiv:2402.00290* (2024).

[35] Jinjie Mai, Jun Chen, Guocheng Qian, Mohamed Elhoseiny, Bernard Ghanem, et al. 2023. Llm as a robotic brain: Unifying egocentric memory and control. (2023).

[36] Artur Marchewka, Marek Wypych, Abnoos Moslehi, Monika Riegel, Jarosław M Michałowski, and Katarzyna Jednoróg. 2016. Arousal rather than basic emotions influence long-term recognition memory in humans. *Frontiers in behavioral neuroscience* 10 (2016), 198.

[37] Mara Mather. 2007. Emotional Arousal and Memory Binding: An Object-Based Framework. *Perspectives on Psychological Science* 2, 1 (March 2007), 33–52. https://doi.org/10.1111/j.1745-6916.2007.00028.x

[38] Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, et al. 2025. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334* (2025).

[39] Clinton Merck, Jeremy K Yamashiro, and William Hirst. 2020. Remembering the big game: Social identity and memory for media events. *Memory* 28, 6 (2020), 795–814.

[40] MistralAI. 2025. Mistral-Small-3.2-24B-Instruct-2506. https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506.

[41] MoondreamLabs. 2025. Moondream2-2B. https://https://huggingface.co/vikhyatk/moondream2.

[42] Coen D. Needell and Wilma A. Bainbridge. 2022. Embracing New Techniques in Deep Learning for Estimating Image Memorability. https://doi.org/10.48550/arXiv.2105.10598 arXiv:2105.10598 [cs]

[43] Fabian Peller-Konrad, Rainer Kartmann, Christian RG Dreher, Andre Meixner, Fabian Reister, Markus Grotz, and Tamim Asfour. 2023. A memory system of a robot cognitive architecture and its implementation in ArmarX. *Robotics and Autonomous Systems* 164 (2023), 104415.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[45] Gustavo Rezende Silva, Juliane Päßler, S. Lizeth Tapia Tarifa, Einar Broch Johnsen, and Carlos Hernández Corbato. 2025. ROSA: A Knowledge-Based Solution for Robot Self-Adaptation. *Frontiers in Robotics and AI* 12 (May 2025), 1531743. https://doi.org/10.3389/frobt.2025.1531743

[46] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2008. How quickly should communication robots respond?. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. 153–160.

[47] Zhihang Song, Zimin He, Xingyu Li, Qiming Ma, Ruibo Ming, Zhiqi Mao, Huaxin Pei, Lihui Peng, Jianming Hu, Danya Yao, et al. 2023. Synthetic datasets for autonomous driving: A survey. *IEEE Transactions on Intelligent Vehicles* 9, 1 (2023), 1847–1864.

[48] Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology* 15, 1 (1904), 72–101.

[49] Micol Spitale, Minja Axelsson, and Hatice Gunes. 2025. VITA: A Multi-Modal LLM-Based System for Longitudinal, Autonomous and Adaptive Robotic Mental Well-Being Coaching. *ACM Transactions on Human-Robot Interaction* 14, 2 (2025), 1–28.

[50] R Nathan Spreng. 2013. Examining the role of memory in social cognition. , 437 pages.

[51] CI Stewardson, MC Hunsche, V Wardell, DJ Palombo, and CM Kerns. 2022. Episodic memory through a social and emotional lens. Emotion. Advance online publication.

[52] Sydney Thompson, Kate Candon, and Marynel Vázquez. 2025. The Social Context of Human-Robot Interactions. https://doi.org/10.48550/arXiv.2508.13982 arXiv:2508.13982 [cs]

[53] Freek Van Ede and Anna C Nobre. 2023. Turning attention inside out: How working memory serves behavior. *Annual review of psychology* 74, 1 (2023), 137–165.

[54] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672* (2024).

[55] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023. Demystifying clip data. *arXiv preprint arXiv:2309.16671* (2023).

[56] Qianli Xu, Fen Fang, Ana Molino, Vigneshwaran Subbaraju, and Joo-Hwee Lim. 2021. Predicting event memorability from contextual visual semantics. *Advances in Neural Information Processing Systems* 34 (2021), 22431–22442.

[57] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics* 2 (2014), 67–78.

[58] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.

[59] Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669* (2024).

[60] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19724–19731.