HERMES: Habit- and Episode-aware Retrieval Memory for Embodied Systems

Jakub Dawid Szkudlarek[1], Marc Pollefeys[1,3], Hermann Blum[2], Zuria Bauer[1]
[1]ETH Zurich [2]University of Bonn [3]Microsoft

***Abstract*—Long-horizon human–robot interaction requires more than short-term task context: to be helpful over time, a robot must remember user-specific habits, preferences, and their updates. A natural baseline is to place the full interaction history into a large language model context window, but this becomes increasingly inefficient and unreliable as the histories grow. We present HERMES, a Habit- and Episode-aware Retrieval Memory for Embodied Systems, which extends scene-graph-based environment representations into a novel temporal scene graph memory that stores scene evolution together with user-specific preference information. To study this system, we design a synthetic benchmark on top of the TEACh dataset by injecting user preferences and temporal updates. We construct three task categories: preference question answering, preference-conditioned command resolution, and temporal preference update reasoning. We compare HERMES with a flat-context baseline over histories of 5 to 25 episodes. On average, HERMES improves overall strict accuracy from 0.398 to 0.612, with especially large gains on preference question answering (0.340 to 0.760), while matching the baseline on command resolution and temporal update reasoning. These results suggest that extending scene representations into structured temporal memory is a promising research direction for scalable preference-aware embodied assistance. Code is available at https://github.com/jacobinsilico/HERMES.**

***Index Terms*—human-robot interaction, embodied AI, episodic memory, scene graphs, large language models**

I. Introduction

Embodied assistants in homes and other human-centered environments are only practically useful if they achieve proactivity. They should remember how a user prefers objects arranged, how underspecified or ambiguous requests such as *"bring me a cup"* should be resolved, and whether those preferences change over time [1], [2]. Commands such as *"put the mug away"* or questions such as *"where are my reading glasses?"* cannot be answered reliably from the current scene alone, but require access to user-specific evidence accumulated across multiple interactions [3].

Scene graphs provide a natural structured representation of an embodied environment by encoding objects together with their semantic and spatial relations [4]–[6]. For a language-capable embodied agent, this kind of structure is appealing because it turns a scene into something that can be inspected semantically rather than only perceived visually. By representing a scene as a scene graph, one can build a history of scene states that is later queried by an LLM. However, doing this efficiently and at scale remains an open challenge.

A simple baseline is to place the full interaction history into the language model prompt. However, this flat-context strategy becomes increasingly expensive and unreliable as histories grow and relevant evidence is buried in long, noisy traces [7]. A scalable method should instead retrieve only the user-specific information needed for the current query, in line with recent work on retrieval-based reasoning for long-horizon language systems [8]–[10].

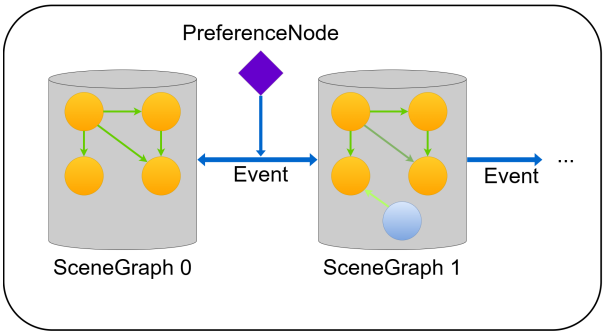

Fig. 1. **Overview of HERMES.** SceneGraph 0 represents the initial scene obtained from semantic scene parsing. When an event occurs (shown here by the appearance of the blue node) HERMES creates SceneGraph 1 and links the two scene graphs with an event edge encoding the scene change. If the event is preference-relevant, HERMES also derives a PreferenceNode. This preference is stored as user-specific memory.

We propose **HERMES**, a **H**abit- and **E**pisode-aware **R**etrieval **M**emory for **E**mbodied **S**ystems. HERMES builds on scene-graph-based representations [4] and extends them into a *temporal scene graph memory* whose nodes capture scene states across episodes, whose temporal edges encode scene changes and interaction outcomes, and whose higher-level memory entries store user preferences and habits. HERMES stores them in an explicit, queryable memory that can be accessed and updated via tools by a language-capable embodied agent [1], [11].

The contributions of this paper are:

- We introduce HERMES, a temporal scene graph memory for preference-aware embodied interaction.
- We construct a long-horizon benchmark on top of the TEACh dataset with injected preferences and temporal updates, and evaluate preference QA, command resolution, and update reasoning.
- We compare HERMES against flat-context prompting and show higher strict accuracy together with substantially lower token and latency costs.

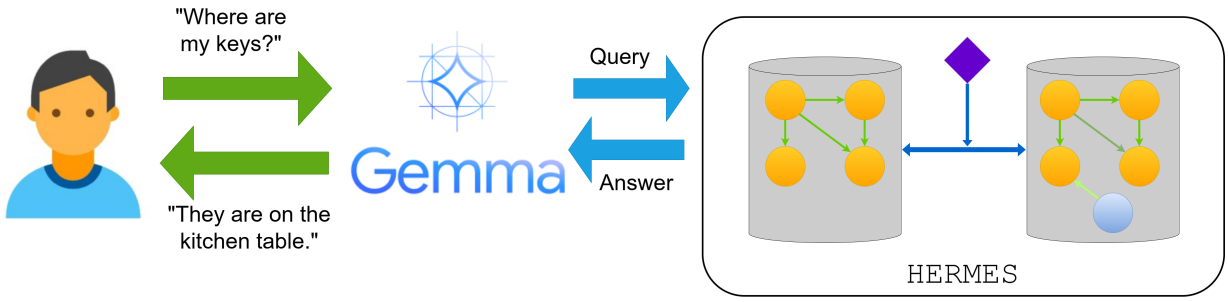

Fig. 2. **Example interaction with HERMES.** A user issues a natural-language query to the language-capable controller, which retrieves relevant information from HERMES before generating a response. The agent grounds its answer in the temporal scene graph memory, enabling access to user-specific information stored across past interactions.

II. Related Work

Scene graphs and episodic memory. Scene-graph-based representations are widely used in robotics because they provide an interpretable model of objects, their attributes, and their semantic or spatial relations, supporting downstream tasks such as open-vocabulary retrieval and manipulation, functional relation discovery from interaction, and tracking scene changes over time [12]–[14]. This structure also connects naturally to *episodic memory*, i.e., memory for events situated in time and place [15], since explicit scene-based representations remain interpretable, support symbolic querying, and interface well with higher-level reasoning systems [16]. Recent work has explored this direction more directly: H-EMV supports episodic recall through a hierarchical memory organization and language-based querying [17], while ReMEmbR focuses on spatio-temporal memory for long-horizon navigation [18]. Unlike prior episodic or navigation-oriented memory systems, HERMES is designed specifically for queryable preference-state retrieval over scene evolution for personalized embodied assistance.

Personalization and long-horizon reasoning. Quality assistance in domestic robotics depends on remembering user-specific preferences, conventions, and prior corrections over time [1], [2], [11]. This raises two linked questions: *how should such information be stored*, and *how should it be accessed when needed*? The first motivates memory-based personalization, while the second connects to the distinction between flat long-context prompting and retrieval-based reasoning. Directly prompting with the full interaction history becomes increasingly inefficient and less reliable as histories grow [7], whereas retrieval-based methods offer a more scalable way to select only query-relevant information [8]–[10]. HERMES sits at this intersection by combining explicit preference-aware memory with targeted retrieval over extended interaction histories.

III. Methods

HERMES assumes an upstream perception module that parses the environment into a semantic scene representation. It then maintains a structured memory over scene states, interaction events, and user-specific preferences, which a language-capable embodied agent can access through retrieval and update tools. These tools are callable functions exposed to the agent, which selects an appropriate tool and provides arguments derived from the user query. Object retrieval is grounded through lexical matching against stored labels rather than embedding-based retrieval, which keeps the current implementation simple and deterministic but makes it sensitive to synonyms and paraphrases. Figure 1 shows the memory structure, and Fig. 2 illustrates an example interaction.

The core representation is a lightweight two-level memory. At the first level, HERMES stores *scene graphs* representing the environment at particular moments. At the second level, it links these scene graphs over time into a *temporal scene graph memory*. Temporal edges, which we call *Events*, encode changes induced by actions or interactions, such as object movements, container state changes, and user corrections.

On top of this temporal structure, HERMES maintains preference- and habit-aware memory entries, which we call *PreferenceNodes*. These nodes are created through language-based registration: when the user expresses preference- or habit-related information during interaction, the agent can invoke a preference-update tool that stores it as a new *PreferenceNode*. This happens within the same tool-calling interaction loop used to answer the user, and therefore does not require a separate preference-extraction stage or an additional standalone LLM call. Each entry stores the target entity or concept, the preferred value or behavior, optional contextual qualifiers, temporal status, and links to supporting evidence in the interaction history.

HERMES is accessed through tools rather than exposed directly to the user. When the user asks a question or issues a command, the agent decides whether memory access is needed and, if so, invokes the appropriate tool. At a high level, HERMES provides *memory query tools* for retrieving structured information from the temporal memory, such as `query_memory_by_key`, which reconstructs the history of an object from a textual key, and `query_preferences`, which retrieves active user preferences from the persistent preference layer. It also provides *memory update tools*, such as `add_new_event` for registering new interactions into the episodic log and `register_user_preference` for

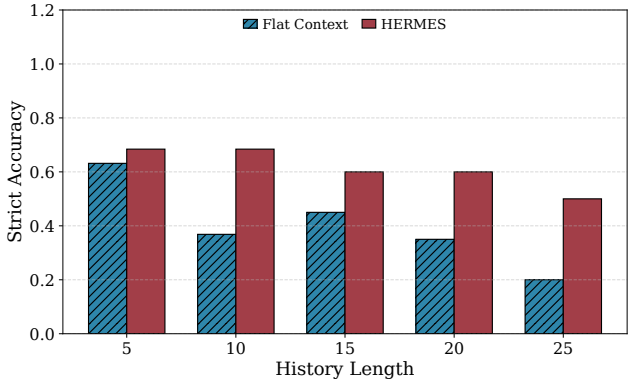

Fig. 3. **Strict accuracy across history lengths.** HERMES outperforms flat-context prompting under strict evaluation at all history lengths.

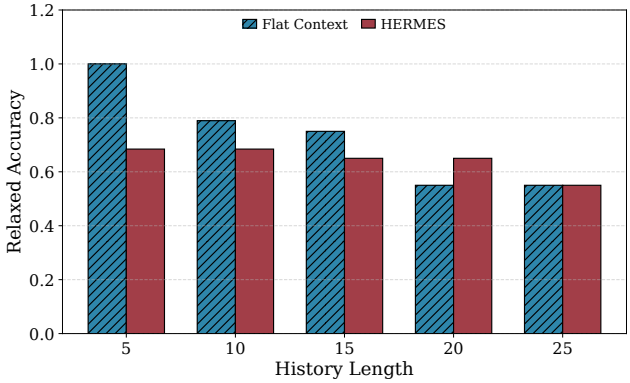

Fig. 4. **Relaxed accuracy across history lengths.** Flat-context prompting is stronger at shorter and mid-range histories, while HERMES catches up at longer horizons.

storing newly expressed user preferences and superseding outdated ones under the same topic.

For evaluation, we construct a synthetic long-horizon benchmark on top of the TEACh dataset of embodied interaction episodes [3], following the general history synthesis strategy from the related literature [17]. We parse 100 TEACh episodes and assemble 25 unique multi-episode histories of length 5 to 25 episodes from distinct source traces. We then inject user preferences and temporal updates using controlled natural-language templates, yielding 98 samples across three tasks: preference QA, where the agent recalls a stored user preference; command resolution, where it disambiguates an underspecified instruction using that preference; and update reasoning, where it must apply the most recent preference after a temporal change.

IV. Experimental Setup

Baseline. We compare two alternative approaches: a flat-context baseline, which serializes the full synthesized interaction history into a single prompt, and HERMES, which accesses a structured temporal scene graph memory through callable tools at inference time. This comparison tests whether long-horizon user-specific memory is better handled through explicit structured retrieval or direct long-context prompting.

Model and inference. For both methods, we use Gemma-3-27B-IT with decoding temperature 0.2 and a maximum output length of 500 tokens. In the HERMES setting, memory access is exposed through a Chat Completions API with tool definitions, while the model is served through a Hugging Face inference backend. We additionally use the same model family as an automated judge, with judge temperature 0.0.

Evaluation protocol. Because outputs are free-form natural-language responses, we evaluate them with an LLM-as-a-judge pipeline [19] using task-specific instructions. We report both *strict* and *relaxed* correctness. Strict correctness requires the response to resolve to the correct target preference, location, or action without mismatch, while relaxed correctness allows semantically correct but less tightly formatted answers. We also report prompt token usage and inference latency. For preference retrieval and command-resolution tasks, location or destination mismatches are marked as incorrect.

V. Results and Discussion

A. Main Results

Tables I and II summarize results averaged over five history lengths. A key observation is that HERMES improves *exact* preference retrieval while being substantially more efficient than flat-context prompting. Under strict scoring, HERMES outperforms the flat-context baseline overall (0.612 vs. 0.398), with the largest gain on preference question answering (0.760 vs. 0.340). Command resolution and temporal update reasoning are matched under strict scoring. Under relaxed scoring, however, flat context remains stronger overall (0.724 vs. 0.643), suggesting that it often produces partially correct but less precise answers.

Under strict evaluation (Fig. 3), HERMES outperforms flat context at every evaluated history length, and the gap becomes especially pronounced at longer horizons. Under relaxed evaluation (Fig. 4), flat context is stronger at shorter and mid-range histories, but this advantage narrows as the history grows; HERMES surpasses flat context at 20 episodes and matches it at 25 episodes. This pattern suggests that HERMES favors more exact retrieval, whereas flat-context prompting often remains partially correct even when it is less precise.

Efficiency is one of the clearest advantages of HERMES. Figure 5 shows that HERMES uses far fewer prompt tokens at every history length, remaining in the low-thousands regime while flat context grows from roughly 6.3×10^4 to 6.7×10^5 tokens. This token efficiency also translates into lower end-to-end latency in Fig. 6, where HERMES remains faster than the flat-context baseline across the evaluated history lengths. The trend follows directly from the method design: HERMES retrieves task-relevant memory through explicit tool calls instead of serializing the full interaction history. However, latency should be interpreted more cautiously than token usage, since our measurements come from a remote inference setup and

TABLE I
STRICT EVALUATION RESULTS ACROSS THE SELECTED HISTORY LENGTHS. VALUES ARE AVERAGED OVER RUNS.

Method	Overall ↑	Pref-QA ↑	Cmd ↑	Upd. ↑	Avg. Tokens ↓
Flat Ctx.	0.398	0.340	0.400	0.522	262476
HERMES	**0.612**	**0.760**	0.400	0.522	**3892**

TABLE II
RELAXED EVALUATION RESULTS ACROSS THE SELECTED HISTORY LENGTHS. VALUES ARE AVERAGED OVER RUNS.

Method	Overall ↑	Pref-QA ↑	Cmd ↑	Upd. ↑	Avg. Tokens ↓
Flat Ctx.	**0.724**	**0.780**	**0.720**	0.609	262476
HERMES	0.643	0.760	0.400	**0.652**	**3892**

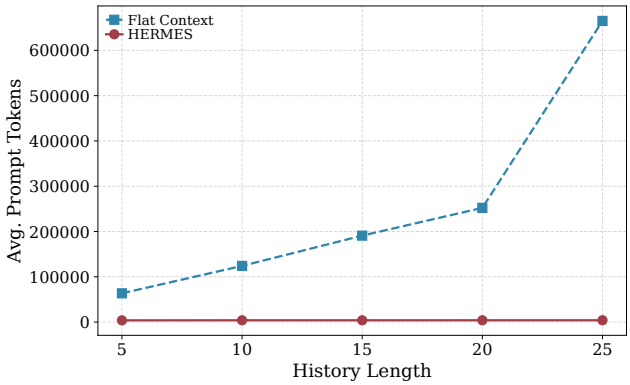

Fig. 5. **Prompt token usage across history lengths.** HERMES remains far more token-efficient than flat-context prompting as history length increases.

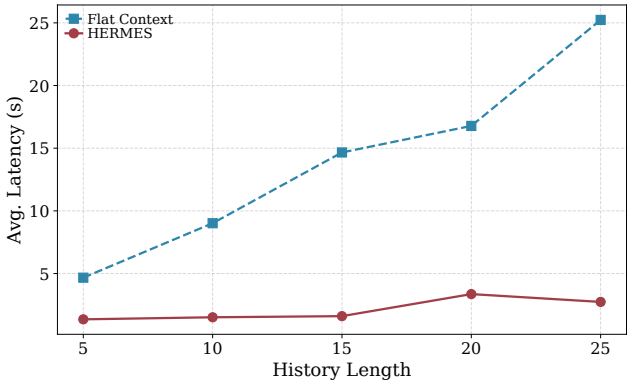

Fig. 6. **End-to-end latency across history lengths.** HERMES maintains lower latency than flat-context prompting as histories grow longer.

therefore include backend and network variability in addition to model computation.

The task breakdown helps explain these results. The clearest gain appears in preference question answering, which is the category most directly tied to retrieving user-specific evidence from long histories. Command resolution remains more challenging: under strict scoring the two methods are tied, while under relaxed scoring flat context remains stronger. Temporal update reasoning is competitive under strict scoring and favors HERMES under relaxed scoring, suggesting that the current memory design can support preference-state updates when the relevant evidence is successfully retrieved. Overall, the results indicate that HERMES is particularly effective when exact retrieval of user-specific preference information is critical, and it offers a significant efficiency advantage over flat-context prompting.

B. Limitations and Future Work

These results should be interpreted together with several limitations. The benchmark is synthetic: preferences and updates are injected into pooled TEACh-style histories rather than collected from naturally emerging long-term user trajectories. This gives a controlled testbed, but does not capture the ambiguity of real deployments, where scene changes, habits, and corrections must be detected from noisy observations. The current implementation also relies on lexical object/preference matching, one open-source LLM backend, and comparison mainly against a flat-context baseline; broader model sweeps and baselines such as long-horizon memory or scene-graph memory systems would better isolate the contribution of each design choice. The evaluation is also limited in scale, and larger benchmarks would be needed to test whether the observed trends hold across more diverse homes, users, and interaction patterns. Finally, latency is measured through remote serving and includes infrastructure noise.

Future work should evaluate HERMES on longer real-world interaction traces with organically emerging habits and corrections. Methodologically, command resolution and temporal preference handling could benefit from semantic retrieval, explicit state versioning, recency-aware conflict resolution, and a short-term/long-term memory split that consolidates stable habits over time. Another important direction is tighter integration with perception and action modules.

VI. CONCLUSIONS

We presented HERMES, a habit- and episode-aware retrieval memory for personalized long-horizon embodied interaction. HERMES extends scene-graph-based environment representations into a temporal scene graph memory that supports preference-aware retrieval over past interactions. By extending the TEACh dataset with injected user preferences and temporal updates, we constructed a benchmark for evaluating preference retrieval, command disambiguation, and update reasoning under increasing history length. Across histories of 5 to 25 episodes, HERMES achieved higher strict accuracy than flat-context prompting while using substantially fewer prompt tokens and lower latency. The clearest gains appear in preference question answering, while command resolution and temporal update reasoning remain competitive with the flat-context baseline under strict scoring. Overall, these results suggest that structured temporal memory is a promising and scalable research direction for preference-aware embodied assistants.

References

[1] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "Tidybot: personalized robot assistance with large language models," *Autonomous Robots*, vol. 47, no. 8, p. 1087–1102, Nov. 2023. [Online]. Available: http://dx.doi.org/10.1007/s10514-023-10139-z

[2] T. Kwon, D. Choi, H. Kim, S. Kim, S. Moon, B. woo Kwak, K.-H. Huang, and J. Yeo, "Embodied agents meet personalization: Investigating challenges and solutions through the lens of memory utilization," 2026. [Online]. Available: https://arxiv.org/abs/2505.16348

[3] A. Padmakumar, J. Thomason, A. Shrivastava, P. Lange, A. Narayan-Chen, S. Gella, R. Piramuthu, G. Tur, and D. Hakkani-Tur, "TEACh: Task-driven embodied agents that chat," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. [Online]. Available: https://arxiv.org/abs/2110.00534

[4] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3668–3678. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2015/html/Johnson_Image_Retrieval_Using_2015_CVPR_paper.html

[5] A. Rosinol, A. Violette, M. Abate, N. Hughes, Y. Chang, J. Shi, A. Gupta, and L. Carlone, "Kimera: From SLAM to spatial perception with 3D dynamic scene graphs," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1510–1546, 2021. [Online]. Available: https://arxiv.org/abs/2101.06894

[6] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3D scene graph construction and optimization," in *Robotics: Science and Systems (RSS)*, 2022. [Online]. Available: https://arxiv.org/abs/2201.13360

[7] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024. [Online]. Available: https://aclanthology.org/2024.tacl-1.9/

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.

[9] Q. Xie, S. Y. Min, P. Ji, Y. Yang, 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 preprint arXiv:2409.18313*, 2024. [Online]. Available: https://arxiv.org/abs/2409.18313

[10] Y. Zhu, Z. Ou, X. Mou, and J. Tang, "Retrieval-augmented embodied agents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/papers/Zhu_Retrieval-Augmented_Embodied_Agents_CVPR_2024_paper.pdf

[11] N. Abdo, C. Stachniss, L. Spinello, and W. Burgard, "Collaborative filtering for predicting user preferences for organizing objects," *arXiv preprint arXiv:1512.06362*, 2015. [Online]. Available: https://arxiv.org/abs/1512.06362

[12] O. Lemke, Z. Bauer, R. Zurbrügg, M. Pollefeys, F. Engelmann, and H. Blum, "Spot-Compose: A framework for open-vocabulary object retrieval and drawer manipulation in point clouds," in *2nd Workshop on Mobile Manipulation and Embodied Intelligence at ICRA 2024*, 2024. [Online]. Available: https://arxiv.org/abs/2404.12440

[13] T. Engelbracht, R. Zurbrügg, M. Pollefeys, H. Blum, and Z. Bauer, "Spotlight: Robotic scene understanding through interaction and affordance detection," in *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, 2025, pp. 1–8.

[14] T. Behrens, R. Zurbrügg, M. Pollefeys, Z. Bauer, and H. Blum, "Lost found: Tracking changes from egocentric observations in 3d dynamic scene graphs," *IEEE Robotics and Automation Letters*, vol. 10, no. 4, pp. 3739–3746, 2025.

[15] E. Tulving, "Episodic and semantic memory," in *Organization of Memory*, E. Tulving and W. Donaldson, Eds. Cambridge, MA: Academic Press, 1972, pp. 381–403.

[16] M. Beetz, D. Beßler, A. Haidu, M. Pomarlan, A. K. Bozcuoglu, and G. Bartels, "KnowRob 2.0 – a 2nd generation knowledge processing framework for cognition-enabled robotic agents," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 512–519. [Online]. Available: https://ai.uni-bremen.de/papers/beetz18knowrob.pdf

[17] L. Bärmann, C. DeChant, J. Plewnia, F. Peller-Konrad, D. Bauer, T. Asfour, and A. Waibel, "Episodic memory verbalization using hierarchical representations of life-long robot experience," in *Proceedings of the 2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, 2025.

[18] A. Anwar, J. Welsh, J. Biswas, S. Pouya, and Y. Chang, "Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 2838–2845.

[19] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Z. Lin, B. Zhang, L. Ni, W. Gao, Y. Wang, and J. Guo, "A survey on llm-as-a-judge," *The Innovation*, p. 101253, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666675825004564