Jornadas de Automática

Memoria Episódica Jerárquica para Razonamiento Espacial en Robots Asistenciales

Alicia Mora*, Lucía Lishan, Gonzalo Espinoza, Luis Moreno, Ramón Barber

RoboticsLab, Departamento de Ingeniería de Sistemas y Automática, Universidad Carlos III de Madrid, Av. de la Universidad, 30, 28911 Leganés, España.

To cite this article: Mora, A., Lishan, L., Espinoza, G., Moreno, L., Barber, R. 2026. Hierarchical Episodic Memory for Spatial Reasoning in Assistive Robots. Jornadas de Automática, 47.
https://doi.org/10.17979/ja-cea.2026.47.13708

Resumen

Conseguir una interpretación semántica robusta de entornos domésticos representa uno de los principales desafíos en robótica de servicio, especialmente con alta densidad de objetos. Esta tarea se ve dificultada por la disposición estructural del entorno, donde las barreras arquitectónicas fragmentan la percepción y complican la organización lógica de la memoria. En este trabajo, presentamos un sistema de memoria episódica jerárquica diseñado para vincular el razonamiento de los modelos de lenguaje (LLM) a la geometría espacial. A diferencia de los métodos basados en proximidad, nuestra propuesta segmenta el espacio en zonas lógicas definidas por el campo de visión efectivo, obligando al sistema a respetar la arquitectura física. El sistema implementa una búsqueda mediante anclas semánticas y un algoritmo de desambiguación espacial para el conteo preciso de entidades. Los resultados muestran que esta arquitectura permite realizar tareas de búsqueda, conteo y explicación con alta fiabilidad, mitigando las alucinaciones del LLM mediante una validación visual proactiva que garantiza un razonamiento situado y veraz.

Palabras clave: Ingeniería de sistemas cognitivos, Robótica inteligente, Computación centrada en el ser humano, Percepción y sensorización, Robots móviles.

Hierarchical Episodic Memory for Spatial Reasoning in Assistive Robots

Abstract

Achieving a robust semantic interpretation of domestic environments represents one of the main challenges in service robotics, particularly in settings with a high density of objects. This task is made more difficult by the structural layout of the environment, where architectural barriers fragment perception and complicate the logical organisation of memory. In this work, we present a hierarchical episodic memory system designed to link the reasoning of language models (LLMs) to spatial geometry. Unlike proximity-based methods, our proposal segments the space into logical zones defined by the effective field of view, forcing the system to respect the physical architecture. The system implements a search using semantic anchors and a spatial disambiguation algorithm for the accurate counting of entities. The results show that this architecture enables search, counting and explanation tasks to be performed with high reliability, mitigating LLM hallucinations through proactive visual validation that ensures situated and accurate reasoning.

Keywords: Cognitive systems engineering, Intelligent robotics, Human-centered computing, Perception and sensing, Mobile robots

*Autor para correspondencia: almorav@ing.uc3m.es

1. Introducción

La interpretación semántica de un entorno por parte de un robot requiere integrar observaciones en un modelo coherente. En entornos con desorden físico, esta tarea se ve dificultada por memorias robóticas convencionales que oscilan entre dos extremos: la rigidez de agrupaciones por habitaciones (Zhang et al., 2025), que ignora áreas funcionales, o una laxitud excesiva que omite cómo los elementos estructurales delimitan la visión (Mao et al., 2025). Al ignorar estas barreras físicas, se impide una categorización lógica, provocando oclusiones no resueltas o duplicados en el conteo. El problema se acentúa en sistemas que operan en un plano puramente simbólico: al confiar en descripciones pregeneradas, el razonamiento del modelo de lenguaje (LLM) queda a merced de una abstracción que omite detalles visuales granulares o relaciones espaciales sutiles no indexadas inicialmente (Zha et al., 2025). Esta desconexión entre la representación simbólica y la realidad física del entorno suele derivar en alucinaciones espaciales, donde el agente asume la existencia o ubicación de objetos basándose en probabilidades lingüísticas en lugar de en evidencias sensoriales.

En este trabajo, proponemos una arquitectura de memoria episódica jerárquica que ancla el razonamiento del LLM a la geometría real, obligándolo a operar sobre una estructura de datos que respeta las barreras físicas. El núcleo del sistema es una segmentación basada en el campo de visión efectivo, garantizando que la lógica interna del robot se alinee con la arquitectura del entorno. Para potenciar la recuperación, introducimos el razonamiento por anclas semánticas, permitiendo inferir ubicaciones mediante relaciones funcionales. Esta capacidad permite que el agente asocie elementos mediante el conocimiento del contexto, facilitando la búsqueda de objetivos que, aunque ocultos en el momento de la consulta, guardan una dependencia con los elementos ya indexados en la jerarquía. Finalmente, el sistema incorpora un algoritmo de desambiguación espacial que utiliza evidencia física para resolver conflictos en el modelo de lenguaje y asegurar la integridad en el conteo de entidades. Un esquema de esta propuesta se observa en la Fig. 1.

Las principales aportaciones del trabajo son las siguientes:

- Segmentación por campo de visión efectivo: Algoritmo de *clustering* que delimita zonas semánticas integrando restricciones de oclusión arquitectónica y superando la rigidez de las divisiones por estancias al identificar áreas basadas en la visibilidad real.
- Inferencia mediante anclas semánticas: Estrategia de búsqueda que utiliza el conocimiento funcional para localizar objetos no registrados explícitamente.
- Desambiguación espacial de entidades: Método de análisis multivista que utiliza coordenadas y solape visual para evitar duplicados y garantizar la integridad en el conteo.

Esta combinación metodológica, junto con un ciclo de verificación visual proactiva, permite al sistema ejecutar tareas de búsqueda y conteo con alta robustez, impidiendo que la abstracción lingüística del LLM se desvíe de la realidad física del entorno asistencial.

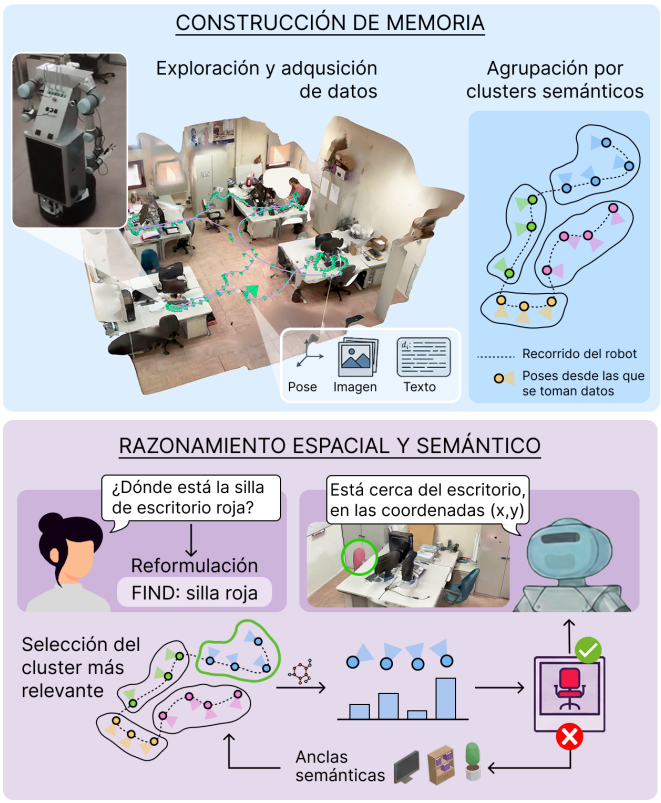

Figura 1: Arquitectura de memoria episódica jerárquica y razonamiento espacial. Durante la exploración, el robot captura *waypoints* (pose, imagen y texto) organizados en *clusters* semánticos. Las consultas en lenguaje natural se clasifican en FIND, EXPLAIN o COUNT, resolviéndose mediante similitud de *embeddings* y validación visual.

2. Estado del arte

2.1. *Modelos de representación de conocimiento en robots*

Tradicionalmente, el conocimiento del entorno se ha basado en modelos geométricos fijos, como el trabajo presentado en Funk et al. (2021) en el que utilizan representaciones volumétricas o el trabajo de Wang and Tian (2022) en el que representan objetos en el espacio mediante sus coordenadas 2D. Otros enfoques añaden grafos de escena en los que definan relaciones espaciales (Liao et al., 2020) o modelos semánticos con anotaciones de objetos (Martins et al., 2020). Sin embargo, la tendencia actual se dirige hacia mapas mejorados mediante modelos de lenguaje (*Large Language Models*, LLMs), que fusionan reconstrucciones físicas con características semánticas para alinear imágenes y descripciones naturales (Huang et al., 2023).

Más recientemente, la literatura ha evolucionado hacia arquitecturas de agentes cognitivos donde el LLM actúa como el núcleo central de decisión (Shaji et al., 2026). Para gestionar la escalabilidad, se han propuesto sistemas de memoria multinivel que separan el conocimiento de trabajo del episódico (Ali et al., 2024), permitiendo al robot recuperar estados previos tras interrupciones en tareas complejas. Otros enfoques introducen controladores de memoria activa para filtrar observaciones redundantes en tiempo real (Dorbala and Manocha, 2026), o sistemas híbridos que permiten al robot verbalizar sus experiencias pasadas de forma narrativa (Plewnia and Asfour, 2025).

A diferencia de estos métodos, que a menudo dependen de descripciones textuales estáticas propensas a omitir detalles o de estructuras de memoria con una alta abstracción sin considerar la estructura del espacio físico, nuestra propuesta organiza la experiencia en una jerarquía basada en el campo de visión efectivo. Además, nuestro enfoque implementa una validación visual y una búsqueda por anclas semánticas. Esto garantiza que el razonamiento no quede limitado al plano simbólico, permitiendo corregir inconsistencias del modelo de lenguaje mediante la reintroducción de evidencia física.

2.2. Razonamiento espacial mediante LLMs

Los modelos de lenguaje han exhibido capacidades emergentes para traducir instrucciones abstractas en acciones ejecutables. Sin embargo, la competencia en razonamiento espacial es subóptima en los modelos multimodales en comparación con su dominio lingüístico. En este contexto, investigaciones recientes confirman que el razonamiento relacional y la transformación de perspectivas constituyen el principal cuello de botella para alcanzar una inteligencia espacial equiparable a la humana (Yang et al., 2025). Se ha observado que, aunque los modelos pueden procesar información visual, tienden a generar una serie de modelos locales desconectados en lugar de una representación global consistente.

Para estructurar este razonamiento, propuestas como ConceptGraphs (Gu et al., 2024) utilizan grafos de escena 3D de vocabulario abierto para inferir relaciones funcionales, aunque dependen de la calidad de los subtítulos generados y pueden omitir objetos pequeños o delgados críticos para la planificación. Por su parte, sistemas basados en RAG tradicional para robótica, como Embodied-RAG (Xie et al., 2024), introducen estructuras jerárquicas para gestionar la redundancia de las observaciones, pero presentan dificultades en tareas que requieren un conteo preciso de objetos a pequeña escala al no considerar la consistencia multivista en su proceso de *clustering*. Finalmente, aproximaciones como Meta-Memory (Mao et al., 2025) intentan mitigar las alucinaciones mediante la integración de mapas cognitivos y la verificación visual de hipótesis, pero su arquitectura se ve limitada por la dependencia de anclas semánticas únicas y la incapacidad de gestionar entornos dinámicos.

A diferencia de estos métodos, nuestro enfoque utiliza una jerarquía basada en el campo de visión efectivo y un ciclo de validación proactiva que reintroduce evidencia física granular, superando la limitación de confiar exclusivamente en representaciones simbólicas o descriptivas que suelen omitir atributos visuales clave para la desambiguación espacial.

3. Arquitectura de Memoria Jerárquica

La arquitectura propuesta organiza la experiencia del agente en una estructura de tres niveles de abstracción, permitiendo pasar de representaciones visuales locales a una comprensión semántica global del entorno.

3.1. Representación a Nivel de Waypoint

El primer nivel de la jerarquía se compone de una secuencia de *waypoints* $W = \{w_1, w_2, \ldots, w_n\}$ capturados durante el desplazamiento del robot. Para cada nodo w_i, se almacena la pose del robot $P_i = (x, y, \theta)$ y la información visual I_i capturada por su cámara.

Con el fin de optimizar el almacenamiento y el procesamiento posterior, la captura se realiza con una frecuencia temporal de f fotogramas establecida experimentalmente. Posteriormente, cada imagen I_i es procesada por un Modelo de Lenguaje Multimodal (MLLM), específicamente GPT-5 (OpenAI, 2025), para extraer una descripción textual sintética d_i. Esta descripción captura las entidades y relaciones espaciales más relevantes presentes en el campo de visión del robot en el instante i.

3.2. Generación de Zonas Lógicas

El segundo nivel consiste en la agrupación de los *waypoints* en *clusters* que representen áreas funcionales del entorno. Nuestro algoritmo integra tres métricas:

1. **Similitud Semántica:** Se calcula la proximidad en el espacio de *embeddings* entre las descripciones de texto d_i. Para dos descripciones d_i y d_j, la proximidad se define mediante la similitud del coseno:
$$S_{cos}(d_i, d_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{||\mathbf{e}_i||||\mathbf{e}_j||} \quad (1)$$
donde $\mathbf{e}$ representa el vector característico generado por un modelo de codificación de texto (Reimers et al., 2021).
2. **Distancia Estructural:** Para respetar las restricciones físicas del entorno (paredes y obstáculos), la distancia entre dos puntos no se calcula de forma euclídea, sino mediante el algoritmo de Dijkstra (Dijkstra, 1959) sobre un mapa de ocupación. Esto garantiza que dos puntos separados por una pared no pertenezcan al mismo *cluster*, calculando el camino más corto libre de obstáculos $D_{path}(P_i, P_j)$.
3. **Orientación:** Se incorpora una métrica de distancia geodésica angular (Vincenty, 1975) para asegurar que los puntos agrupados compartan una dirección de observación común hacia la zona de interés, evitando la inclusión de vistas opuestas que pertenezcan a áreas funcionales distintas.

Para la generación de las zonas lógicas, se emplea un algoritmo de *Hierarchical Agglomerative Clustering* (HAC) (Ward, 1963). El proceso se inicia calculando una matriz de distancias híbridas, $\mathbf{D}_{hybrid}$, que resulta de la ponderación unificada de las tres métricas anteriores (semántica, estructural y orientación). Sobre esta matriz se aplica el método de vinculación completa (*complete linkage*), definido como:

$$d(C_u, C_v) = \max(\{\mathbf{D}_{hybrid}[i, j] : i \in C_u, j \in C_v\}) \quad (2)$$

donde $d(C_u, C_v)$ representa la distancia entre los *clusters* C_u y C_v, calculada como la máxima distancia entre cualquier par de *waypoints* i y j pertenecientes a dichos grupos.

Finalmente, se utiliza un criterio de asignación basado en un umbral de distancia $t_{threshold}$ para extraer los *clusters* finales C_k de la jerarquía. Una vez definidos estos grupos, el sistema utiliza el MLLM para procesar el conjunto de descripciones $\{d_i \in C_k\}$ y generar un resumen de zona R_k que encapsula la funcionalidad y contenido predominante del área identificada.

3.3. Abstracción del Entorno Global

El nivel superior de la jerarquía recoge la definición global del entorno. El sistema concatena los resúmenes de las zonas lógicas $\{R_1, R_2, \ldots, R_m\}$ y los utiliza como contexto para que el modelo fundacional genere la descripción. Esta descripción permite clasificar el contexto general (por ejemplo, distinguiendo entre entornos domésticos, oficinas o comercios) y comprender la distribución macroscópica de las utilidades disponibles. Esta estructura jerárquica facilita en etapas posteriores la recuperación de información por consultas.

4. Razonamiento Espacial y Semántico

Una vez estructurada la memoria episódica, el sistema emplea un motor de razonamiento diseñado para transformar la información jerárquica en respuestas ante consultas en lenguaje natural. Este proceso clasifica la intención del usuario en tres tareas fundamentales: localización (*FIND*), conteo (*COUNT*) y explicación (*EXPLAIN*).

4.1. Localización Jerárquica y Validación Visual

Para la tarea *FIND*, el sistema emplea una estrategia de búsqueda en dos etapas impulsada por *embeddings* y verificación visual.

En primer lugar, el sistema utiliza el MLLM para extraer el objeto objetivo de la consulta original, eliminando ruido semántico. Se genera un *embedding* de dicho objetivo y se calcula la similitud del coseno contra los resúmenes de cada zona lógica R_k usando la ecuación 1. Una vez identificado el *cluster* más relevante, se comparan los *embeddings* del objetivo con las descripciones individuales de los *waypoints* d_i dentro de dicho *cluster*. A partir de este ranking de similitud, se seleccionan los 5 puntos de observación más relevantes para proceder con la fase de validación visual. En esta etapa, el sistema recupera las imágenes originales I_i correspondientes a estos candidatos y solicita al MLLM una verificación. Si la verificación se clasifica como exitosa, se elige el *waypoint* correspondiente como la solución a la solicitud.

Si la búsqueda directa no arroja resultados confirmados, el sistema activa una fase de sentido común en la que el MLLM genera una lista de anclas semánticas (objetos o muebles usualmente próximos al objetivo). Para ello, se envía una petición estructurada al modelo solicitando la identificación de cinco objetos domésticos o elementos de mobiliario con una alta probabilidad de co-ocurrencia espacial en el entorno con el objeto que no se ha encontrado. A diferencia de otros trabajos como Mao et al. (2025), estas anclas no se limitan a un único elemento en el espacio. El sistema expande la búsqueda calculando la similitud máxima entre los *embeddings* de estas anclas y el resto de la memoria episódica, permitiendo localizar objetos por asociación contextual. Al igual que en la etapa anterior, las anclas con mayor similitud semántica son verificadas directamente sobre la imagen original para comprobar que el objeto está efectivamente presente en la imagen.

4.2. Conteo Desambiguado mediante Análisis Multivista

Para la tarea *COUNT*, el proceso se inicia localizando la zona lógica y el *waypoint* con mayor probabilidad de contener el objeto objetivo siguiendo el esquema de búsqueda por *embeddings*. Una vez identificada una vista candidata, el sistema expande el contexto recuperando todas las imágenes pertenecientes al mismo *cluster* cuyo campo de visión (*Field of View*) presente un solape significativo con la zona de interés.

Estas imágenes se envían al MLLM junto con sus respectivas poses P_i (específicamente la orientación θ) para proporcionar un marco de referencia geométrico. El modelo utiliza este contexto multivista para razonar sobre posibles oclusiones y desambiguar instancias repetidas detectadas desde diferentes ángulos. Este enfoque garantiza que el conteo final sea robusto, evitando duplicidades y corrigiendo las limitaciones informativas de las descripciones textuales aisladas.

4.3. Generación de Explicaciones

La resolución de la tarea *EXPLAIN* aprovecha la estructura jerárquica de la memoria para generar respuestas que combinan dos niveles de abstracción. El razonador integra la información macroscópica del mapa global con el detalle microscópico de los *waypoints* locales para situar al objeto en su contexto funcional.

Para ello, el sistema concatena el resumen de la zona lógica R_k con las descripciones detalladas d_i de los nodos más relevantes del *cluster*. Esta dualidad permite que el robot proporcione descripciones espacialmente ricas, indicando no solo la presencia de un objeto, sino su relación con el entorno inmediato (por ejemplo, su ubicación sobre un mueble específico o su proximidad a otros elementos). Para validar la tarea, la respuesta del modelo debe identificar correctamente la zona lógica global y mencionar los elementos más relevantes, omitiendo entidades alucinadas ajenas al *cluster*.

5. Resultados Experimentales

En esta sección se presenta la evaluación cuantitativa y cualitativa del sistema propuesto, con el objetivo de validar su capacidad para modelar la memoria episódica jerárquica y resolver consultas abstractas y complejas en escenarios reales. A través de este análisis, se identifican y discuten las principales propiedades de desambiguación visual y sus limitaciones actuales de reformulación de lenguaje.

5.1. Configuración y escenarios de evaluación

El sistema ha sido evaluado en cinco entornos de interior diferenciados: dos despachos y tres viviendas con múltiples estancias, ambos con variabilidad de objetos y desorden físico.

Para la adquisición de datos, se utilizó una cámara Intel RealSense (Keselman et al., 2017). En los entornos de oficina, la cámara se integró en la plataforma robótica móvil ADAM (Mora et al., 2024), mientras que en los entornos domésticos se empleó una configuración portátil. Para garantizar la homogeneidad del dataset, en ambos casos se utilizó un algoritmo de Visual SLAM para la generación del mapa de ocupación y la localización precisa de los *waypoints* desde los que se capturaron las imágenes. Esta configuración híbrida permite validar la eficacia del sistema con independencia de la plataforma robótica utilizada, demostrando su versatilidad en diferentes condiciones de captura.

Tabla 1: Comparativa de rendimiento entre modelos frente a consultas episódicas

Modelo	SR FIND	SR COUNT	SR EXPLAIN	Tiempo (s)	Llamadas	Tokens Entrada	Tokens Salida
EmbodiedRAG	0.53	0.43	0.67	44.70	4	4396	2645
MetaMemory	0.63	0.53	0.70	**16.30**	**2**	**690**	**673**
Propuesto	**0.77**	**0.73**	**0.87**	18.03	4	1045	1199

5.2. Taxonomía de consultas y conjunto de datos

Para evaluar la robustez del razonamiento episódico, se ha diseñado un conjunto de treinta consultas en lenguaje natural, categorizadas según el tipo de proceso de recuperación requerido: *FIND*, *COUNT* o *EXPLAIN*. Con el fin de garantizar la comparabilidad de los resultados, se aplica un conjunto de plantillas de consulta idénticas en todos los escenarios, adaptando únicamente el objeto objetivo a la realidad de cada entorno (p. ej., "¿Dónde está el microondas?" en una vivienda frente a "¿Dónde está la impresora?" en un despacho).

Para establecer el Ground Truth, cada entorno fue inspeccionado por un anotador humano que registró manualmente la ubicación, el conteo real de objetos y una descripción detallada de referencia para cada consulta.

5.3. Baselines y métricas de evaluación

Para validar la eficacia de nuestra propuesta, el sistema se compara frente a dos enfoques del estado del arte: *EmbodiedRAG* (Xie et al., 2024) y *MetaMemory* (Mao et al., 2025). El primero, *EmbodiedRAG*, basa su recuperación exclusivamente en descripciones textuales y no utiliza la información visual en ninguna etapa del razonamiento, reconociendo explícitamente como limitación su incapacidad para considerar la consistencia multivista. Por otro lado, *MetaMemory* propone el uso de imágenes para la verificación de hipótesis, pero su alcance se limita estrictamente a la tarea de localización (*FIND*), sin ofrecer soluciones para el conteo o la explicación contextual. Además, para la búsqueda de objetos, *MetaMemory* depende de una única ancla semántica y una exploración local a su alrededor; si esta selección inicial es imprecisa, el sistema carece de mecanismos de recuperación robustos, lo que compromete su tasa de éxito.

Con el fin de cuantificar el rendimiento y la eficiencia, se han definido las siguientes métricas:

- **Tasa de Éxito (Success Rate, SR):** Porcentaje de respuestas que coinciden con el *ground truth* en las tres categorías (*FIND*, *COUNT* y *EXPLAIN*).
- **Eficiencia Temporal:** Tiempo medio de ejecución (s) por consulta.
- **Carga Computacional:** Número medio de llamadas a la API de LLM y consumo de tokens (Entrada/Salida) por consulta.

5.4. Resultados y análisis

Los resultados presentados en la Tabla 1 demuestran que el sistema propuesto supera a los *baselines* en términos de precisión semántica, logrando una Tasa de Éxito (SR) mayor en las tres categorías evaluadas. Aunque *MetaMemory* presenta el tiempo de ejecución más bajo con 16.30s, nuestra arquitectura mantiene una latencia altamente competitiva de 18.03s a cambio de ofrecer una precisión mucho mayor en todas las categorías de consulta. Es importante destacar que estas métricas de eficiencia representan la media global de los experimentos; en condiciones de búsqueda directa, el consumo de tokens y el tiempo de respuesta de nuestra propuesta son muy similares a los de *MetaMemory*, incluso menores. No obstante, el ligero incremento en los promedios totales se debe a los casos de alta complejidad donde el sistema debe activar el proceso de búsqueda mediante anclas semánticas. En estas situaciones, el número de llamadas a la API y el tiempo de búsqueda se disparan necesariamente para garantizar el éxito de la tarea, una capacidad de recuperación de la que carecen el resto de propuestas.

En el análisis cualitativo se observaron comportamientos que explican estas diferencias de rendimiento. En el caso de *EmbodiedRAG*, se identificó que, aunque el sistema selecciona en ocasiones el punto de vista adecuado para realizar el conteo, la carencia de información visual directa le impide procesar detalles granulares, lo que deriva en respuestas incompletas por falta de evidencia física. Esta limitación se extiende también a las tareas de búsqueda y explicación, donde el modelo falla al intentar localizar o describir objetos pequeños o atributos específicos que no fueron indexados explícitamente en las descripciones textuales iniciales. Por su parte, *MetaMemory* mostró una propensión a generar alucinaciones visuales, llegando a confirmar erróneamente la presencia de objetos inexistentes. Además, su metodología centrada en el análisis de un único frame limita su capacidad para captar la estructura global del entorno y las relaciones espaciales complejas.

En contraste, el sistema propuesto logra mitigar estas deficiencias mediante su estructura de memoria jerárquica y el ciclo de validación visual. Sin embargo, el método presenta una limitación específica en la etapa de reformulación de la petición del usuario. Se ha observado que, en ocasiones, el sistema omite información relevante del *prompt* original, como atributos específicos pero distintivos del objeto. Esta pérdida de información durante la interpretación inicial compromete la capacidad de desambiguación del agente ante la presencia de duplicados, pudiendo seleccionar una instancia incorrecta al carecer de los criterios necesarios para distinguir entre elementos similares en el entorno. Una comparativa cualitativa se muestra también en la Fig. 2.

6. Conclusiones

En este trabajo, hemos presentado una arquitectura de memoria episódica jerárquica que aborda uno de los retos en la robótica asistencial: la desconexión entre el razonamiento semántico de los modelos de lenguaje y la realidad física del entorno. Nuestra propuesta segmenta el conocimiento basándose en el campo de visión efectivo, garantizando que la organización de la memoria respete las barreras arquitectóni-

Figura 2: Resultados cualitativos comparativos en un escenario doméstico. Mientras que los métodos del estado del arte incurren en errores de conteo y localización, nuestra propuesta es capaz de desambiguar entidades mediante la utilización del campo de visión efectivo.

cas y las restricciones de oclusión reales. Los resultados demuestran que esta propuesta resuelve con éxito tareas de razonamiento complejas y mejora los resultados de propuestas actuales mediante el uso de anclas semánticas y desambiguación espacial.

El trabajo futuro se centrará en adaptar la jerarquía de memoria a entornos dinámicos temporales. Asimismo, se pretende refinar la interacción mediante un diálogo proactivo que permita al robot solicitar aclaraciones visuales ante ambigüedades espaciales persistentes. Por último, con el fin de mejorar la privacidad y reducir la dependencia de la conectividad y latencia de APIs externas, se evaluará el uso de redes ejecutadas localmente en hardware embarcado.

Agradecimientos

Este trabajo ha sido apoyado por el proyecto Advanced Mobile dual-arm Manipulator for Elderly People Attendance (AMME) (PID2022-139227OB-I00), financiado por el Ministerio de Ciencia e Innovación y por el iRoboCity2030-CM (TEC-2024/TEC-62) - Programas de Actividades I+D en tecnologías de la Comunidad de Madrid.

Referencias

Ali, H., Allgeuer, P., Mazzola, C., Belgiovine, G., Kaplan, B. C., Gajdošech, L., Wermter, S., 2024. Robots can multitask too: Integrating a memory architecture and llms for enhanced cross-task robot action generation. In: 2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids). IEEE, pp. 811–818.

Dijkstra, E. W., 1959. A note on two problems in connexion with graphs. Numerische Mathematik 1 (1), 269–271.

Dorbala, V. S., Manocha, D., 2026. Memctrl: Using mllms as active memory controllers on embodied agents. arXiv preprint arXiv:2601.20831.

Funk, N., Tarrio, J., Papatheodorou, S., Popović, M., Alcantarilla, P. F., Leutenegger, S., 2021. Multi-resolution 3d mapping with explicit free space representation for fast and accurate mobile robot motion planning. IEEE Robotics and Automation Letters 6 (2), 3553–3560.

Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K. M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al., 2024. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 5021–5028.

Huang, C., Mees, O., Zeng, A., Burgard, W., 2023. Visual language maps for robot navigation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 10608–10615.

Keselman, L., Iselin Woodfill, J., Grunnet-Jepsen, A., Bhowmik, A., 2017. Intel realsense stereoscopic depth cameras. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1–10.

Liao, Z., Zhang, Y., Luo, J., Yuan, W., 2020. Tsm: Topological scene map for representation in indoor environment understanding. IEEE Access 8, 185870–185884.

Mao, Y., Ye, H., Dong, W., Zhang, C., Zhang, H., 2025. Meta-memory: Retrieving and integrating semantic-spatial memories for robot spatial reasoning. arXiv preprint arXiv:2509.20754.

Martins, R., Bersan, D., Campos, M. F., Nascimento, E. R., 2020. Extending maps with semantic and contextual object information for robot navigation: a learning-based framework using visual and depth cues. Journal of Intelligent & Robotic Systems 99 (3), 555–569.

Mora, A., Prados, A., Mendez, A., Espinoza, G., Gonzalez, P., Lopez, B., Muñoz, V., Moreno, L., Garrido, S., Barber, R., 2024. Adam: a robotic companion for enhanced quality of life in aging populations. Frontiers in Neurorobotics 18, 1337608.

OpenAI, 2025. Presentamos gpt-5. Accedido el: 5 de mayo de 2026. URL: `https://openai.com/index/introducing-gpt-5/`

Plewnia, J., Asfour, T., 2025. Combining episodic memory and llms for the verbalization of robot experiences. In: 2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids). IEEE, pp. 531–538.

Reimers, N., et al., 2021. all-mpnet-base-v2: Sentence transformer model. `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`, accedido el: 5 de mayo de 2026.

Shaji, S., Huppertz, F., Mitrevski, A., Houben, S., 2026. From language to action: Can llm-based agents be used for embodied robot cognition? arXiv preprint arXiv:2603.03148.

Vincenty, T., 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of geocentric Hezagonal co-ordinates. Survey Review 23 (176), 88–93.

Wang, Z., Tian, G., 2022. Hybrid offline and online task planning for service robot using object-level semantic map and probabilistic inference. Information Sciences 593, 78–98.

Ward, Jr., J. H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58 (301), 236–244. DOI: `10.1080/01621459.1963.10500845`

Xie, Q., Min, S. Y., Ji, P., Yang, Y., Zhang, T., Xu, K., Bajaj, A., Salakhutdinov, R., Johnson-Roberson, M., Bisk, Y., 2024. Embodied-rag: General non-parametric embodied memory for retrieval and generation. arXiv preprint arXiv:2409.18313.

Yang, J., Yang, S., Gupta, A. W., Han, R., Fei-Fei, L., Xie, S., 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643.

Zha, J., Fan, Y., Yang, X., Gao, C., Chen, X., 2025. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. arXiv preprint arXiv:2504.05786.

Zhang, J., Wu, S., Ma, X., Schwertfeger, S., 2025. Generation of indoor open street maps for robot navigation from cad files. arXiv preprint arXiv:2507.00552.