Towards Task Planning for Proactive Safety in Home Service Robots: A Scene Graph-Augmented LLM Approach

Sena Ishii, Ankit A. Ravankar and Yasuhisa Hirata

***Abstract*— Service robots deployed in homes must do more than detect potential hazards—they must decide what to do about them. Prior work has demonstrated that large language and vision-language models (LLMs and VLMs) can infer household accident risks from a single image, but it largely stops at recognition and does not ground this recognition in what the robot can *physically do* in its own environment. We propose a safety task planning that utilizes a pre-built semantic map of the home, represents it as a compact hierarchical scene graph augmented with a *statemap*—a layer encoding abstract room-level states such as "children playing" or "cluttered"—and feeds it to an LLM together with an onboard RGB observation. The LLM selects one of four grounded actions (MOVE, PUSH, NO ACTION, ALERT ONLY) together with the target object and, when applicable, a destination landmark drawn from the scene graph. We evaluate the pipeline on 15 simulated scenarios in Isaac Sim, comparing conditions with and without the scene graph, and on two real-robot scenarios with a mobile manipulator. Our results show that risk reasoning is robust regardless of the scene graph, while the scene graph improves plan executability by grounding destinations in the robot's actual environment. A case study on statemap-augmented planning further demonstrates that providing abstract environmental states can redirect the LLM's destination selection when safety implications are direct—suggesting that scene-state awareness is a promising lever for safer autonomous task planning.**

***Index Terms*— Service robotics, Home safety, Vision-language models, scene graph, LLM task planning, semantic reasoning**

I. Introduction

Service robots are beginning to enter human-centered environments such as homes, offices, and care facilities [1]–[5]. As this transition progresses, the set of competences the robot must possess broadens well beyond navigation and manipulation of pre-specified objects. A service robot sharing space with a human family encounters incidental hazards continuously: a knife left on the table after cooking, a cup resting on the edge of a desk, clothing draped near a space heater. Detecting such hazards early—*before* they cause an accident—is widely recognized as a valuable capability for household robots that aim to reduce the burden on caregivers and improve everyday safety. In our previous work [6], [7], we proposed a risk prediction method based on semantic relationships between objects derived from a statistical database, but it cannot account for the physical state of individual objects. Recent advances in LLMs and VLMs have made the *recognition* side of this problem tractable [8]–[10], establishing that foundation models carry sufficient commonsense knowledge to reason about household safety.

Yet an important step is consistently left to the reader: given that a hazard has been recognized, **what should the robot actually do next?** Most existing pipelines either stop at risk prediction, delegate the decision to a human operator, or assume that hazard-specific policies are pre-programmed.

All authors are with the Department of Robotics, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan. Email: {s.ishii, ankit, hirata}@srd.mech.tohoku.ac.jp

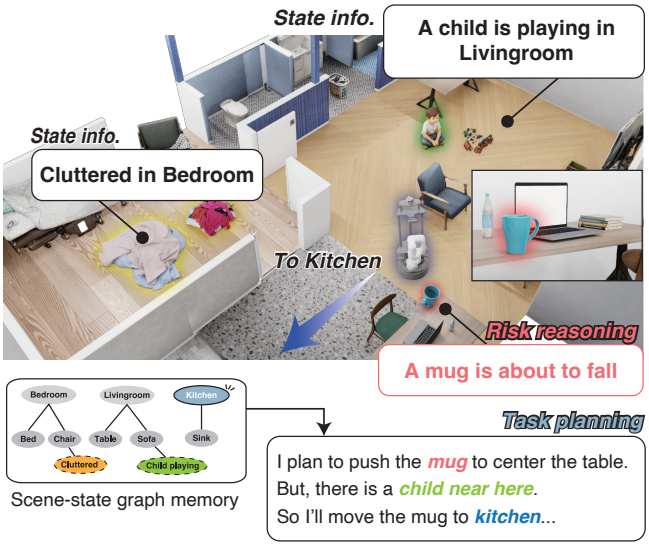


Fig. 1: Overview of the proposed system. The robot interprets the environment through a hierarchical scene-state graph, which integrates physical structures with situational states (e.g., "cluttered" or "child playing"). Based on this graph, the LLM-based planner performs risk reasoning and selects the most appropriate task and destination (e.g., moving a risky object to the kitchen to avoid a nearby child).

In practice, a domestic robot must choose, for each observed hazard, an action that is (i) physically executable in its home, (ii) targeted at the right object, and (iii) calibrated in ambition: sometimes a small nudge is enough, sometimes the robot should not act at all and instead alert a human. To bridge this gap, we propose a proactive safety task planning that tightly couples LLM-based hazard reasoning with a grounded action-selection procedure that is aware of both the spatial layout of the home and the semantic state of its regions.

Our approach (Fig. 1) builds on the observation that a compact, hierarchical representation of the home is sufficient for an LLM to reason about actions. We represent the home as a four-level scene graph (home root, rooms, landmarks, and optional sub-landmarks attached to landmarks—e.g., a toybox next to the TV). The scene graph can optionally be augmented with a *statemap* layer that annotates rooms or landmarks with abstract states such as "cluttered" or "child is playing" (Section III). At inference time, the scene graph together with the robot's current room and an RGB observation are fed to a reasoning VLM (GPT-4.1-mini), which selects one action from a fixed vocabulary—MOVE, PUSH, NO ACTION, or ALERT ONLY—and fills in

the required fields. The vocabulary is deliberately small: it captures the four qualitatively distinct responses a mobile manipulator may produce, and forces the model to defer to a human (ALERT ONLY) when it cannot safely intervene.

We evaluate the proposed pipeline on 15 simulated scenarios in Isaac Sim spanning six hazard categories, comparing conditions with and without the scene graph, and reproduce two representative scenarios on a physical mobile manipulator. Our results show that risk reasoning is unaffected by the scene graph, while task planning—especially destination selection—benefits from grounded landmark knowledge (Section V).

Contributions. The contributions of this paper are:

1) A proactive home-safety pipeline that couples VLM-based hazard reasoning with a grounded, four-option action vocabulary, enabling the robot to *act* on its predictions rather than merely report them.
2) A scene-graph representation with a statemap layer designed to be populated from the robot's own observations
3) A comparative evaluation on 15 simulated scenarios and two real-robot scenarios, showing that the scene graph is critical for grounded task planning, along with a case study demonstrating the potential of abstract state annotations to influence destination selection.

II. Related Work

A. Foundation Models for Household Hazard Reasoning

Recent works leverage LLMs and VLMs to reason about household hazards: Brunke et al. [9] constrain manipulator trajectories based on semantic scene understanding; Wu et al. [10] identify safe grasp affordances; Mullen et al. [8] feed spatial object relations to an LLM for risk prediction; and Choi et al. [11] augment navigation maps with VLM-generated context. Broader LLM-planned manipulation works [12]–[14] close the recognition-to-action loop but are not specialized to safety and do not treat inaction as a first-class option. Our work proposes an end-to-end pipeline that grounds hazard reasoning in executable robot actions.

B. Scene Graphs for LLM Planning

While existing works [15], [16]apply hierarchical scene graphs to general navigation, our statemap layer extends this representation with abstract room-level states, making situational context directly consumable by downstream LLM prompts.

III. Methodology

A. System Overview

The robot is assumed to operate in a home whose spatial layout has been mapped prior to deployment [16]–[18]. We assume that a hierarchical map—consisting of rooms and landmarks (static furniture)—established by Chikhalikar et al. [19] is available.

Our pipeline operates in two stages. First, the pre-built scene graph is optionally augmented with a *statemap* layer that encodes abstract room or landmark states. Second, at inference time, the scene graph, the robot's current room, and an RGB observation are passed to a reasoning VLM, which identifies any hazard and selects one of four actions: MOVE, PUSH, NO ACTION, or ALERT ONLY.

B. Scene Graph Construction

The home is represented as a four-level hierarchy: home root, rooms, landmarks (static furniture), and sub-landmarks (e.g., laundry basket). Levels 1–3 are derived from object-detection-based pipelines [19]; sub-landmarks are registered manually. The hierarchy is optionally augmented with a statemap layer that attaches category-valued tags (e.g., "cluttered," "children playing") to nodes. In this study, statemap annotations are predefined; automated construction is left for future work.

C. LLM-based Reasoning

At inference time, the LLM receives:

- the scene graph (optionally with statemap annotations);
- the robot's current state and position in the environment(room), obtained via its localization module;
- a single RGB image of the current scene in the field of view of the robot.

The system prompt defines an explicit action vocabulary of four options, together with a preference order (NO ACTION $\rightarrow$ PUSH $\rightarrow$ MOVE $\rightarrow$ ALERT ONLY) so that the model escalates to ALERT ONLY only when the robot cannot safely handle the situation.

The prompt distinguishes **hazard objects** (e.g., heaters) from portable **target objects** (e.g., clothes). This prevents the robot from moving fixed or heavy appliances, addressing a failure mode observed in preliminary experiments.

We use GPT-4.1-mini as the reasoning VLM. The output is a single JSON object describing the action and its parameters.

D. Action Execution

A downstream navigator dispatches each action: MOVE triggers navigation to the chosen landmark followed by a pick-and-place request; PUSH emits a directional push; ALERT ONLY publishes a notification; NO ACTION does nothing. For real-robot execution, the target object is localized in 3D using Grounded-SAM-2 [20] with depth back-projection.

IV. Experiments

We evaluate this method to address this research question: how the scene graph affects *task planning*—specifically, the grounding and appropriateness of generated actions and destinations. We additionally present a qualitative case study on how room-level *statemap* annotations influence destination selection.

A. Simulated Scenarios and Ablation Setup

We constructed 15 scenarios in NVIDIA Isaac Sim across six hazard categories: sharp objects (knife, scissors, broken glass), piled clothing (shirt, towel, blanket), unstable placement (mug or bottle at a table edge), heat-source proximity (shirt near a portable heater), scattered toys, and normal safe scenes. Each scenario specifies a scene image, the robot's current room, and a ground-truth (GT) plan comprising the expected action, target object, and an acceptable set of destination landmarks or push directions.

GT was obtained by pooling independent judgments from three annotators, each shown the scene image alone and asked: "Do you perceive any safety risk? If so, how would you handle it?" Annotators chose from three actions (MOVE/PUSH/NO ACTION) and specified a target object

and destination. To ensure grounding, annotators were familiarized with the simulated environment beforehand; they were provided with a top-down map and a walkthrough video showing the layout and furniture placement in the Isaac Sim scene. We determined the GT action via majority consensus, defining the acceptable-destination set as any landmark proposed by the annotators.

B. Evaluation Setup

We compare two primary conditions that differ in the context provided in the LLM prompt:

- **No Scene Graph (No-SG)**: The prompt contains only the action vocabulary and the current RGB observation.
- **Scene Graph (SG)**: The prompt additionally includes the home's scene graph, consisting of rooms and their associated landmarks/sub-landmarks.

Both conditions share the same system prompt, utilize GPT-4.1-mini as the reasoning engine, and receive the identical RGB input. We compare these predictions against the human-annotated ground truth while performing an ablation analysis to examine how the presence of the scene graph influences the output quality.

To address our research questions, we evaluate the performance using the following metrics:

a) Risk Reasoning:

1) **Accuracy**: The proportion of scenarios where the predicted risk aligns with human judgments, measuring how closely the model's intuition mimics human safety perception.
2) **Recall**: Among the scenarios where annotators identified a risk, the fraction that the model also detected as risky. A recall of 100% means the model never missed a hazard flagged by humans.

b) Task Planning: For task planning, we evaluate a subset of 11 scenarios where both the model and annotators reached a consensus that a risk was present. We compare the SG and No-SG conditions across two dimensions:

1) **Target Object Accuracy**: The alignment between the model-selected target object and the GT for MOVE or PUSH actions.
2) **Destination & Push-Direction Accuracy**: The success rate in selecting a valid destination landmark (for MOVE) or an appropriate movement direction (for PUSH) relative to the GT set.

C. Case Study: Statemap Integration

Beyond the basic scene graph, we evaluate a condition where the representation is augmented with a *statemap*. In this context, a "state" refers to abstract, room-level or landmark-level conditions—such as "cluttered," "children playing in the room," or "occupied by resident in the room". These semantic tags are attached as attributes within the scene graph.

We qualitatively evaluate how providing these abstract environmental states affects task planning compared to the SG-only condition. By applying these state annotations to existing scenarios, we analyze whether the robot can adjust its destination selection or action choice based on the contextual safety of the environment (e.g., avoiding placing an object in a "cluttered" area).

TABLE I: Comparison of No-SG and SG conditions. Risk reasoning: all 15 scenarios. Task planning: 11 "risk" scenarios (destination accuracy over the MOVE subset).

	Risk Reasoning		**Task Planning**	
Cond.	**Acc. (%)**	**Recall (%)**	**Act. (%)**	**Dest. (%)**
No-SG	73.3	100	100	45.5
SG	73.3	100	100	**54.5**

D. Real-Robot Scenarios

On a mobile manipulator, we reproduce two representative scenarios from the simulated set: **R1** (clothing scattered on the bedroom floor) and **R2** (cup on the edge of a table). For each trial, the robot's localization feeds the current room into the prompt and the LLM-selected action is dispatched through the navigator. We qualitatively evaluate whether the system runs end-to-end across three stages: plan generation, target 3D localization, and navigation.

V. RESULTS

Table I summarises the results across both evaluation dimensions. Risk reasoning was identical under both conditions, confirming that the scene graph does not affect hazard detection. Task planning accuracy, however, improved with the scene graph, with destination accuracy increasing from 45.5% to 54.5%.

A. Risk Reasoning

Both conditions achieved identical risk-detection accuracy (73.3%) and perfect recall (100%), confirming that **the scene graph does not affect risk reasoning**—the VLM's pretrained scene understanding already suffices for hazard detection. The four safe scenes produced safety-conservative false positives, a preferable failure mode for a home service robot.

B. Task Planning

No-SG. Without the scene graph, the LLM generates plausible but unverifiable destinations such as "shelf" or "bedside table"—furniture that does not exist in the configured simulation environment. The robot would need to verify each destination before execution, reducing operational reliability.

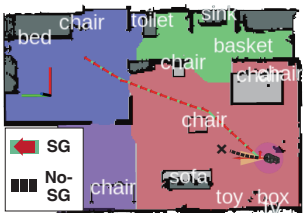
SG. With the scene graph, the LLM selects destinations exclusively from the provided landmark set, producing immediately executable plans. A representative scenario is "Toys on the bedroom floor": **No-SG** proposed "move to a shelf in the living room" (unverifiable), whereas **SG** selected `toybox`—a semantically appropriate and grounded destination.

Comparison with human annotations. Annotators and the LLM occasionally diverged in destination choice: annotators placed a knife into a toybox (prioritising immediate removal from reach), while the LLM moved it to the Kitchen (a more appropriate long-term location). Annotators also described actions such as "push the bottle toward the centre *while avoiding the laptop and books*," explicitly considering collateral effects on nearby objects—a nuance the LLM did not capture. Zero-shot awareness of surrounding-object interactions remains a desirable capability for household robots.

(a) Robot executing the pushing action.

(b) Object segmentation of the target mug.

Fig. 2: Physical manipulation for the cup scenario. (a) The robot executes a pushing action to mitigate the hazard of the mug at the table edge. (b) Real-time segmentation of the target mug by Grounded-SAM-2, used to derive the 3D contact point for the push.



(a) Planned path to the bedroom

(b) Segmentation of the clothes

Fig. 3: Navigation task for the clothing scenario. (a) The SG trajectory leads to `chair` in the Bedroom; the No-SG destination (×) indicates `sofa` armrest (unregistered). The red and blue areas represent the living room and bedroom, respectively. (b) Real-time segmentation of the target object by Grounded-SAM-2.

C. Case Study: Statemap Influence

We investigated whether room-level state annotations (*statemap*) can influence the LLM's destination choice by manually injecting statemap entries into the prompt. In one scenario, the LLM successfully shifted the destination away from the Livingroom after being informed that children were playing there, reasoning that "the cup could be knocked over, especially with children playing"—consistent with the findings of Mullen et al. [8]. In contrast, the strong prior associating knives with kitchens overrode the statemap signal. These results suggest that **statemap annotations can influence task planning when the safety implications are direct and severe**, but remain insufficient to override deeply ingrained object–location associations.

D. Real-Robot Execution

We evaluate the real-robot deployment on two representative scenarios: a clothing-on-the-floor scenario and a cup-at-a-table-edge scenario.

Cup scenario. The LLM produced the expected PUSH action with a natural-language direction phrase. Grounded-SAM-2 segmented the cup from the onboard RGB image, depth-based mask projection yielded a 3D contact point, and the manipulator physically executed the push (Fig. 2).

Clothing scenario. This scenario illustrates how scene-graph grounding shapes destination selection (Fig. 3). Under the **No-SG** condition, the LLM selected the `sofa armrest` visible in the RGB image—a semantically plausible choice, but one limited to objects present in the current field of view. While `sofa` is registered in the map, its `armrest` is not, requiring additional exploration to localize the precise target. Under the full **SG** condition, the LLM instead selected `chair` in the Bedroom, reasoning over landmarks across the entire home rather than the immediate visual context, and the robot successfully navigated there and segmented the shirt on the floor. To further probe this behaviour, we ran an ablation in which the Bedroom branch was removed from the scene graph; the model then selected `laundry basket` as the destination. All three runs demonstrate that the model *reasons over the scene graph* rather than relying solely on visual context or fixed label-to-location mappings.

VI. Discussion

The scene graph plays distinct roles at different pipeline stages. Without it, the LLM inferred plausible but unverifiable destinations; with it, every destination corresponds to a real landmark, making plans immediately executable—a critical requirement for deployment. While the numerical improvement in destination accuracy is modest (45.5% to 54.5% over 11 scenarios), this reflects the small evaluation set rather than the practical benefit of grounded planning. We also note that the current ablation does not isolate the contribution of the graph structure from simply providing a valid destination list—a flat-list condition would clarify this, and we leave it to future work.

A notable gap remains between LLM and human reasoning: annotators spontaneously considered collateral effects on nearby objects (e.g., "avoid the laptop when pushing"), whereas the LLM did not. Addressing this zero-shot awareness of surrounding-object interactions is a valuable direction for future work. The case study showed that abstract state annotations can redirect destination selection when the safety implication is direct (mug × children playing), but fail to override deeply ingrained object–location priors (knife→kitchen), motivating explicit constraint propagation and automated statemap construction in future work.

VII. Conclusion

We presented a pipeline for proactive home safety in which a pre-built scene graph is provided to an LLM that selects a grounded action from a structured vocabulary. Evaluation of 15 scenarios revealed two key findings: (1) VLM-based risk reasoning is robust even without a scene graph, often surpassing human detection; (2) however, the scene graph is essential for task planning to ensure grounded, executable destinations rather than unverifiable ones. A case study further showed that room-level state annotations can influence destination selection when the safety implication is direct, demonstrating the potential of abstract environmental context for safer action generation. Real-robot trials confirmed end-to-end executability of the pipeline. Building on the finding that activity-based context (e.g., the presence of children) can meaningfully redirect the robot's actions, future work will develop an end-to-end pipeline that automatically incorporates diverse contextual information—from spatial clutter to human activity states [21] [22] that would allow the robot to dynamically reflect these factors in its autonomous task planning.

ACKNOWLEDGMENTS

This work was partially supported by JST Moonshot R&D [Grant Number JPMJMS2034] and JSPS Kakenhi [Grant Number JP24K07399].

REFERENCES

[1] G. Bardaro, A. Antonini, and E. Motta, "Robots for elderly care in the home: A landscape analysis and co-design toolkit," *International Journal of Social Robotics*, vol. 14, no. 3, pp. 657–681, 2022.

[2] M. Kim, S. Kim, S. Park, M.-T. Choi, M. Kim, and H. Gomaa, "Service robot for the elderly," *IEEE robotics & automation magazine*, vol. 16, no. 1, pp. 34–45, 2009.

[3] S. Haddadin and E. Croft, "Physical human–robot interaction," in *Springer handbook of robotics*, pp. 1835–1874, Springer, 2016.

[4] A. A. Ravankar, S. A. Tafrishi, J. V. S. Luces, F. Seto, and Y. Hirata, "Care: Cooperation of ai robot enablers to create a vibrant society," *IEEE Robotics & Automation Magazine*, vol. 30, no. 1, pp. 8–23, 2022.

[5] A. Chikhalikar, A. A. Ravankar, J. V. S. Luces, S. A. Tafrishi, and Y. Hirata, "An object-oriented navigation strategy for service robots leveraging semantic information," in *2023 IEEE/SICE International Symposium on System Integration (SII)*, pp. 1–6, 2023.

[6] S. Ishii, A. Chikhalikar, A. A. Ravankar, J. V. S. Luces, and Y. Hirata, "Context-aware risk estimation in home environments: A probabilistic framework for service robots," in *2025 34th IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, IEEE, 2025.

[7] S. Ishii, A. A. Ravankar, J. V. S. Luces, and Y. Hirata, "Towards safer homes: Behavior-adaptive risk assessment for service robots using semantic context," in *Proceedings of the Workshop on Benefits of pErsonalization and behAvioral adaptation in assistive Robots (BEAR 2025), co-located with IEEE RO-MAN 2025*, CEUR Workshop Proceedings, CEUR-WS.org, 2025. To appear.

[8] J. F. Mullen, P. Goyal, R. Piramuthu, M. Johnston, D. Manocha, and R. Ghanadan, ""don't forget to put the milk back!" dataset for enabling embodied agents to detect anomalous situations," *IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 9087–9094, 2024.

[9] L. Brunke, Y. Zhang, R. Römer, J. Naimer, N. Staykov, S. Zhou, and A. P. Schoellig, "Semantically safe robot manipulation: From semantic scene understanding to motion safeguards," *IEEE Robotics and Automation Letters*, 2025.

[10] L. Wu, W. Wei, P. Yu, and J. Lan, "Open-vocabulary 3d affordance understanding via functional text enhancement and multilevel representation alignment," in *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7988–7997, 2025.

[11] D. Choi, H. Lee, S. Hwang, and Y. Oh, "Task-aware semantic map: Autonomous robot task assignment beyond commands," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13567–13573, IEEE, 2025.

[12] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[13] 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, pp. 1087–1102, 2023.

[14] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on robot learning*, pp. 287–318, PMLR, 2023.

[15] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10608–10615, IEEE, 2023.

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

[17] A. Chikhalikar, A. A. Ravankar, J. V. S. Luces, S. A. Tafrishi, and Y. Hirata, "An object-oriented navigation strategy for service robots leveraging semantic information," in *2023 IEEE/SICE International Symposium on System Integration (SII)*, pp. 1–6, 2023.

[18] J. Jiang, J. Zhou, Y. Yan, Y. Wang, L. Wang, and J. Li, "A hierarchical indoor spatial-semantic reasoning-based scene graph construction for elderly-centric safety warnings," *Results in Engineering*, vol. 26, p. 105397, 2025.

[19] A. Chikhalikar, A. A. Ravankar, J. Victorio Salazar Luces, and Y. Hirata, "Semantic-based multi-object search optimization in service robots using probabilistic and contextual priors," *IEEE Access*, vol. 12, pp. 113151–113164, 2024.

[20] IDEA-Research, "Grounded sam 2: Ground any object with segment anything 2." https://github.com/IDEA-Research/Grounded-SAM-2, 2024. Accessed: Jan. 2, 2026.

[21] R. J. Manríquez-Cisterna, P. Mishra, J. Peña-Queralta, M. Perez-Serrano, S. Garyfallidis, L. Kupper, M. Ejtehadi, A. Breuss, A. A. Ravankar, J. V. Salazar Luces, *et al.*, "Ros 4 healthcare: a framework for physiological human sensing for social, assistive, rehabilitation, and medical robotics," *Frontiers in Robotics and AI*, vol. 13, p. 1745197, 2026.

[22] A. Ravankar, A. Rawankar, and A. A. Ravankar, "Real-time monitoring of elderly people through computer vision," *Artificial Life and Robotics*, vol. 28, no. 3, pp. 496–501, 2023.