Contextual Reasoning in Healthcare Social Robotics: A Preliminary Study Using Multimodal Language Models

Luca Pallonetto[1,*], Luigi D'Arco[1] and Silvia Rossi[1]

[1]Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Via Claudio 21, 80125, Naples, Italy

Abstract

Socially assistive robots in healthcare must interpret complex and ambiguous environments to behave safely and appropriately. This pilot study investigates the use of Multimodal Large Language Models (MLLMs) for context-aware scene understanding by combining visual and auditory inputs. We propose a modular pipeline integrating Moondream, a CLIP-based vision-language model, and CoNeTTE, an audio captioning model, to interpret static images and ambient sounds. The system was evaluated on two datasets: the Audiovisual Aerial Scene dataset and a custom synthetic hospital dataset with images from HIOD and audio generated via Stable-Audio 1.0. On the aerial dataset, multimodal input improved classification accuracy from 69.04% to 81.09% and F1-score from 65.15% to 80.22%, showing the benefit of audio in disambiguating visually similar scenes. In contrast, limited gains were observed on the hospital dataset due to weak image–audio alignment, highlighting challenges in synthetic healthcare data. The findings highlight the significant impact that MLLM-based perception can have on healthcare robotics, yet they also reveal present challenges with data quality, domain adaptation, and cross-modal grounding in practical applications. Future work will integrate the proposed perception layer into an actual robotic platform to evaluate its real-time context awareness and adaptive responses in dynamic settings. Ultimately, combining the multimodal perception layer with advanced planning, dialogue, and emotion recognition capabilities will be essential for developing socially intelligent robots capable of assisting both patients and healthcare professionals in a contextually aware manner.

Keywords

Social Assistive Robots & Multimodal Large Language Model & Scene Recognition & Healthcare

1. Introduction

The adoption of social robots in healthcare settings is steadily increasing, with applications ranging from patient monitoring to therapeutic support and logistical assistance [1]. However, successful integration into these environments demands a deep understanding of the surrounding context. Unlike industrial robots operating in structured and predictable environments, social robots deployed in hospitals, elderly care facilities, and rehabilitation centers must function in dynamic, unstructured spaces filled with social cues, physical obstacles, and safety-critical scenarios [2]. These settings are often ambiguous and continuously changing, requiring robots to interpret not only spatial information but also the evolving human activities, emotional tone, and environmental constraints [3]. For instance, in a hospital setting, an empathetic demeanor may be appropriate in a waiting room when interacting with people, whereas efficient and streamlined movement may be prioritized in a hallway. In contrast, discretion and minimal disruption are essential in more sensitive areas such as operating rooms. These scenarios require not only spatial awareness but also necessitate a sophisticated level of contextual understanding that goes beyond basic perception. The potential of context-aware robotics lies in its ability to deliver not only enhanced operational robustness but also more intuitive, trustworthy, and human-aligned behavior. This includes promoting safer interactions, improving task effectiveness, and ensuring that robotic actions are interpretable and appropriate within a healthcare scenario.

To enable effective context awareness, a fundamental capability is the ability for the robot to interpret and understand its surrounding environment. Early approaches to scene understanding primarily relied

Workshop on Social Robotics for Human-Centered Assistive and Rehabilitation AI (a Fit4MedRob event) - ICSR 2025

*Corresponding author.

luca.pallonetto@unina.it (L. Pallonetto); luigi.darco@unina.it (L. D'Arco); silvia.rossi@unina.it (S. Rossi)

0009-0004-0773-0564 (L. Pallonetto); 0000-0001-7179-8281 (L. D'Arco); 0000-0002-3379-1756 (S. Rossi)

on handcrafted features and rule-based systems. These traditional methods, while foundational, often struggled with generalization and scalability across diverse environments [4]. Initially, scene recognition relied on global attribute descriptors, which aimed to mimic human visual perception using low-level features. Techniques such as GIST [5], CENTRIST [6], and LDBP [7] captured holistic characteristics of scenes, but lacked robustness to variations in viewpoint, lighting, and object occlusion. To address these limitations, researchers introduced local patch-based representations, leveraging descriptors like SIFT [8], and SURF [9] to extract finer-grained features. These were often aggregated using methods like Bag-of-Visual-Words (BoVW), improving recognition accuracy by capturing local structure and texture. However, these methods still required substantial manual tuning and could not dynamically adapt to novel scenes. Subsequent methods introduced spatial layout pattern learning and discriminative region detection, aiming to model scene composition more flexibly and emphasize key visual regions. While these approaches offered improvements, they still fell short of integrating semantic and contextual information in a unified framework. More recently, transformer-based architectures have emerged as powerful tools for multimodal perception in robotics [10]. These models excel at fusing visual and linguistic data, enabling richer contextual understanding. In particular, Vision-Language Models (VLMs) integrate image features with textual knowledge, allowing robots not only to recognize scenes but also to reason about them semantically. Further steps forward have been taken with the release of Multimodal Large Language Models (MLLMs) [11]. These models, trained on large corpora of data and types, such as images, text, and audio, can not only describe a scene, but also reason about it, inferring affordances, intentions, and risks from abstract cues. MLLM are able to interpret a scene through a human-like reasoning lens, overcoming the problem of partial information or occluded images.

This study explores the feasibility of exploiting MLLMs as a foundation for scene understanding and contextual reasoning, with the long-term goal of enabling adaptive and context-aware behaviors in social robots. By evaluating the reasoning capabilities of MLLMs, we aim to determine their potential as high-level perception modules to support intelligent and socially appropriate robotic behaviors in dynamic environments. In this preliminary stage, we focus on static scenes containing partial or ambiguous visual information, such as images depicting only a fragment of a room, isolated medical equipment, or occluded spaces. This setup allowed the investigation of the effectiveness of MLLMs in inferring contextual meaning, environmental affordances, and potential human activities from incomplete or implicit visual cues. Initially, the models were assessed using only images; subsequently, audio information was integrated to improve situational interpretation. This form of multimodal reasoning enables a transition to goal-aware, context-sensitive behavioral modulation, where a robot's actions are guided not only by spatial awareness but also by inferred intent, emotion, urgency, and social appropriateness. For example, understanding whether a scene suggests a quiet waiting room, a critical medical event, or a routine interaction can influence a robot's decision to speak, move, assist, or remain passive. The final aim will be to expand on this foundation by incorporating richer sensory inputs and evaluating real-time performance in human-robot interaction scenarios.

2. Methodology

To explore the potential of multimodal perception in assistive robotics for healthcare, we developed a structured evaluation pipeline utilizing a MLLM to assess its ability to interpret and reason about scene context. This preliminary study explores whether integrating visual and auditory inputs can improve a model's ability to interpret complex, ambiguous, or safety-critical scenes, conditions that social robots frequently encounter in real-world clinical settings. To simulate realistic sensory inputs, we focus on static visual scenes paired with ambient audio recordings, evaluating how well MLLMs can extract meaningful semantic and contextual understanding from this multimodal data.

The developed pipeline integrates Moondream[1], a lightweight yet expressive vision-language model designed for real-time applications. Moondream is built on the CLIP (Contrastive Language-Image Pre-training) encoder architecture [12], which aligns visual and textual representations in a shared

[1]https://moondream.ai/

embedding space. Unlike larger multimodal transformers, Moondream is optimized for low-latency inference and on-device execution, making it particularly suitable for robotic platforms with limited computational resources. In our setup, Moondream is tasked with producing semantic-level descriptions of visual scenes based on single static images, capturing both object-level content and contextual cues. To incorporate auditory information, the pipeline also integrates CoNeTTE [13], a neural captioning model that generates descriptive text from audio recordings. CoNeTTE uses a conformer-based encoder to extract temporal and spectral features from raw audio and decodes them into natural language descriptions via a transformer-based language decoder. This model is capable of identifying both environmental sounds (e.g., alarms, footsteps, conversations) and their semantic implications, offering a high-level linguistic summary of the audio context. In our implementation, the audio caption produced by CoNeTTE is prepended to the visual prompt before being passed to Moondream, effectively creating a multimodal composite input that allows the system to reason over combined visual and auditory cues.

The experimental pipeline was tested under two configurations:

- *Visual-only condition*: the system is presented with a static image and tasked with classifying the scene based solely on visual cues;
- *Visual + Audio condition*: the system is presented with a static image and 5 seconds of audio recording.

This modular setup allows us to systematically assess how multimodal inputs contribute to contextual reasoning, laying the groundwork for future, real-time integration into socially intelligent robotic platforms.

2.1. Evaluation Datasets

To validate the proposed architecture, we initially employed the Audiovisual Aerial Scene Recognition Dataset [14], a publicly available collection of 5,075 paired images and environmental audio clips depicting various ambiguous outdoor scenes. For this study, a total of 2996 image-audio pairs have been chosen across 9 scene categories characterized by the presence of partial information in order to test the feasibility of the pipeline. The categories included: airport, beach, bridge, farmland, forest, grassland, and harbour. This dataset is not directly aligned with healthcare scenarios, but it provides a valuable benchmark to test the pipeline's robustness in disambiguating semantically distinct contexts, which is a relevant capability for real-world robotic perception.

Building on this preliminary phase, the methodology was extended to the healthcare domain, where multimodal datasets suitable for robotics applications remain scarce and underexplored. To address this gap, we developed a custom synthetic dataset designed to simulate realistic indoor hospital scenarios. Visual data were manually selected from the Hospital Indoor Object Detection (HIOD) dataset [15], focusing on scenes with partial occlusions, limited visible cues, or ambiguous content. A total of 160 images were selected from this dataset and included common healthcare spaces, including waiting areas, patient rooms, operating rooms, and hospital corridors. Furthermore, this dataset contains various visual cues such as partially occluded equipment and hallways without identifiable signage. These images were particularly relevant as they often depicted only partial views of the environments mentioned before, simulating the constrained and task-oriented perspective a robot might have while performing specific actions within the scene.

Since the HIOD dataset does not include audio data, we synthetically generated ambient sounds to simulate auditory context. As shown in Fig. 1, for each image, a semantic caption summarizing the scene was generated using the BLIP (Bootstrapping Language-Image Pretraining) model [16], which extracts high-level visual descriptions by aligning image content with natural language. Additionally, two categorical labels were manually assigned to each image to reflect the likely room type and its function, aiding in downstream contextual interpretation. For each label, a prompt template was constructed by combining it with the image captions, and passed through LLAMA 3.1 to generate a natural language audio scene description. This textual description was subsequently used as input for Stable Audio 1.0[2],

[2]https://stability.ai

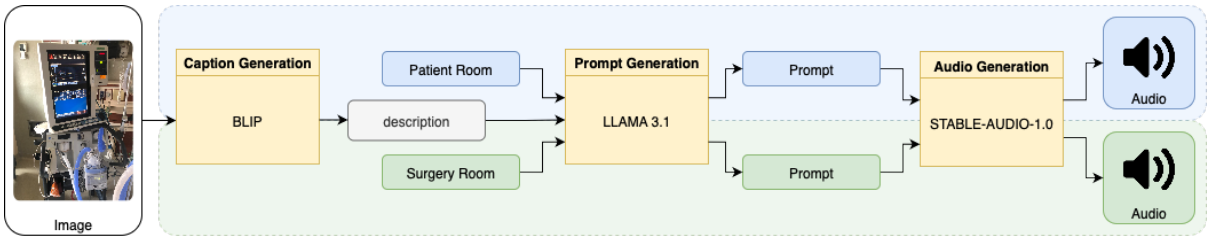

Figure 1: Flow of the dataset generation. Images selected from the Hospital Indoor Object Detection (HIOD) dataset [15].

Table 1
Performance of the proposal pipeline on the Audiovisual Aerial Scene Dataset.

Input Configuration	Accuracy (%)	F1 Score (%)
Visual Only	69.04	65.15
Visual + Audio	81.09	80.22

a state-of-the-art diffusion-based generative model capable of producing high-fidelity 5-second audio clips. These clips reflect typical ambient sounds expected in each room type (e.g., beeping monitors in patient rooms, footsteps and murmurs in waiting areas, equipment sounds in operating rooms).

The motivation behind constructing this dataset was to enable experimentation under conditions that closely mirror real-world deployment scenarios for assistive robots in hospitals. In such settings, a robot must continuously interpret environmental cues and adjust its behavior accordingly.

3. Results

To evaluate the effectiveness of the proposed multimodal pipeline, a comparative analysis of the model's performance was conducted under two experimental conditions: using only visual input (visual-only) and using both visual and auditory inputs (visual + audio). Performance was assessed using standard scene classification metrics, including Accuracy and F1 score.

3.1. Audiovisual Aerial Scene Dataset

The performance of the proposed pipeline on the Audiovisual aerial scene dataset is presented in Table 1. In the visual-only experiments, where the model was prompted with a single image, classification performance reached an Accuracy of 69.04% and an F1-score of 65.15%. Similarity among the images with semantically distinct but visually overlapping environments posed challenges to the model in the recognition phase. Such limitations suggest that relying exclusively on visual input may be insufficient for robust scene interpretation, especially in real-world scenarios requiring nuanced contextual understanding. In the visual-audio experiments, the images were accompanied by the audio description. The pipeline's recognition performance increased to 81.09% and 80.22%, for Accuracy and F1-score, respectively. The inclusion of audio-derived context helped reduce confusion in key categories. For example, environmental sounds such as crowd chatter, transportation noises (e.g., engines, horns, or train signals), and various ambient sounds produced by everyday objects served as strong semantic information that complemented the image content, enabling the model to differentiate between similar-looking scenes more effectively.

These findings validate the benefit of multimodal integration for scene recognition tasks, demonstrating that ambient audio, even when transformed into linguistic input, can provide complementary cues. This improvement is particularly relevant in assistive robotics contexts, where misinterpretation of a setting could lead to inappropriate or unsafe behavior.

Table 2
Performance of multimodal models on a the hospital room classification task, comparing visual-only input with combined visual and audio inputs.

		Visual Only		**Visual + Audio**	
Model	Num of parameters	Accuracy	F1-Score	Accuracy	F1-Score
gemma3	4b	33.77	29.19	33.12	27.21
gemma3	12b	37.01	35.88	46.1	43.55
llava	7b	38.31	31.91	37.01	30.17
moondream	2b	35.95	29.88	35.29	32.50
qwen2.5	7b	38.96	36.53	38.96	40.41

3.2. Hospital Rooms Dataset

To evaluate the performance of the proposed pipeline in real-world healthcare settings, a synthetic dataset was created with ambiguous images and generated audios.

The pipeline, with moondream core, achieved an accuracy of 35.95% and F1 score of 29.88% in the visual-only experiment, and 35.29% and 32.50% in the visual-audio experiment. Following a manual inspection of the model outputs and a thorough analysis of the synthetic hospital dataset, it became evident that the data contained a high degree of semantic ambiguity. Many of the visual scenes lacked distinctive cues, and the alignment between images and their corresponding audio descriptions was often weak or non-informative, particularly when compared to the more coherent and contextually rich audiovisual dataset used in the earlier phase of the study.

To verify whether performance limitations were due to model constraints or dataset quality, several models were tested under both visual-only and visual+audio conditions, including gemma, llava, and qwen2.5. The results are reported in Table 2. While some models exhibited slightly improvements when prompts with audio description, the overall classification performance across models remained relatively low. This consistently poor Accuracy and F1-score across architectures supports the conclusion that the primary bottleneck lies in the dataset itself, rather than in the models or pipeline design. Even when selecting models with different number of parameters (gemma3 with 4b and 12b parameters) the overall performance increased slightly. These findings reinforce the importance of using high-quality, semantically aligned multimodal data, particularly in sensitive domains like healthcare, where clarity and precision are crucial.

4. Conclusion

This work explored the application of MLLMs to context-aware scene understanding for socially assistive robots in healthcare settings. We proposed and evaluated a modular perception pipeline capable of interpreting complex indoor scenes by combining visual and auditory inputs. Preliminary results indicate that the multimodal configuration improves performance, particularly when dealing with partially occluded or visually ambiguous scenes. Notably, the addition of audio information played a disambiguating role, helping the model to distinguish between visually similar settings by using contextual acoustic cues. However, the healthcare dataset did not demonstrate an improvement in performance, requiring additional investigation. Overall, equipping robots with the ability to interpret such differences through multimodal inputs and MLLMs could enable more appropriate, responsive, and socially aligned behaviors that can adapt to different scenarios.

Future efforts should focus on collecting real-world multimodal datasets within clinical settings, including rich audio environments and high-fidelity image data annotated for context, emotional tone, and functional zones. From a system integration perspective, the next step is to embed the proposed perception pipeline into a physical robotic platform, evaluating its real-time performance in a dynamic environment. Furthermore, future research will explore how to integrate the MLLMs capabilities of scene understanding with higher-level planning, dialogue, and emotion recognition capabilities with the final aim of building socially intelligent robots capable of assisting patients and professionals in a

context-sensitive manner.

Acknowledgments

This research has been supported by the European Union - Next Generation EU, Mission 4 Component 1, CUP E53D23016260001 PRIN 2022 PNRR ADVISOR, and under the complementary actions to the NRRP "Fit4MedRob - Fit for Medical Robotics" Grant (# PNC0000007).

Declaration on Generative AI

During this work, the authors used ChatGPT for grammar and spelling checks. All content was subsequently reviewed and edited by the authors, who take full responsibility for the final version.

References

[1] N. Lee, J. Kim, E. Kim, O. Kwon, The influence of politeness behavior on user compliance with social robots in a healthcare service setting, International Journal of Social Robotics 9 (2017) 727–743.

[2] D. L. Johanson, H. S. Ahn, E. Broadbent, Improving interactions with healthcare robots: a review of communication behaviours in social and healthcare contexts, International Journal of Social Robotics 13 (2021) 1835–1850.

[3] L. D'Arco, L. Raggioli, G. Randazzo, G. De Gasperis, A. Chella, S. Costantini, S. Rossi, Towards trustworthy and explainable socially assistive robots: A cognitive architecture for dietary guidance, in: 2025 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR), IEEE, 2025, pp. 1–6.

[4] L. Xie, F. Lee, L. Liu, K. Kotani, Q. Chen, Scene recognition: A comprehensive survey, Pattern Recognition 102 (2020) 107205.

[5] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, International journal of computer vision 42 (2001) 145–175.

[6] J. Wu, J. M. Rehg, Centrist: A visual descriptor for scene categorization, IEEE transactions on pattern analysis and machine intelligence 33 (2010) 1489–1501.

[7] X. Meng, Z. Wang, L. Wu, Building global image features for scene recognition, Pattern recognition 45 (2012) 373–380.

[8] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2004) 91–110.

[9] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9, Springer, 2006, pp. 404–417.

[10] X. Han, S. Chen, Z. Fu, Z. Feng, L. Fan, D. An, C. Wang, L. Guo, W. Meng, X. Zhang, et al., Multimodal fusion and vision-language models: A survey for robot vision, arXiv preprint arXiv:2504.02477 (2025).

[11] J. Wu, W. Gan, Z. Chen, S. Wan, P. S. Yu, Multimodal large language models: A survey, in: 2023 IEEE International Conference on Big Data (BigData), IEEE, 2023, pp. 2247–2256.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[13] Étienne Labbé, T. Pellegrini, J. Pinquier, Conette: An efficient audio captioning system leveraging multiple datasets with task embedding, 2023. URL: https://arxiv.org/pdf/2309.00454.pdf. arXiv:2309.00454.

[14] D. Hu, X. Li, L. Mou, P. Jin, D. Chen, L. Jing, X. Zhu, D. Dou, Audiovisual aerial scene recognition dataset, 2020. URL: https://doi.org/10.5281/zenodo.3828124. doi:10.5281/zenodo.3828124.

[15] D. Hu, S. Li, M. Wang, Object detection in hospital facilities: A comprehensive dataset and performance evaluation, Engineering Applications of Artificial Intelligence 123 (2023) 106223.

[16] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning, PMLR, 2022, pp. 12888–12900.