Language, Knowledge and Intelligent Systems

https://lkisjournal.org

Referential Accuracy and Multimodal Grounding Signals in Embodied Instruction Tasks: Graph Analysis

Nicolas David

Department of Computer Science, Faculty of Science and Engineering, Sorbonne Université, Paris, Île-de-France, France

Corresponding author: nicolas.david@gmail.com

Abstract

The rapid advancement of embodied artificial intelligence has necessitated the development of sophisticated models capable of understanding and executing complex natural language instructions within dynamic three-dimensional physical environments. A persistent challenge in this domain is the accurate grounding of linguistic references to their corresponding physical entities, a process defined as multimodal grounding. This paper investigates the assessment of referential accuracy by employing graph analysis techniques to evaluate multimodal grounding signals in embodied instruction tasks. By conceptualizing the physical environment as a dynamic, interconnected graph of semantic nodes and spatial edges, we propose a comprehensive methodology to measure how effectively computational agents resolve ambiguous referential expressions during task execution. Through rigorous experimental design and comparative analysis across standardized simulation benchmarks, this research demonstrates that graph-based topological representations significantly enhance referential disambiguation compared to traditional flat vector representations. The findings provide empirical evidence that integrating structural relationship analysis into multimodal grounding architectures substantially reduces interaction errors and improves overall task completion rates. This study contributes to the theoretical understanding of vision-language integration and offers practical insights for designing robust cognitive architectures for autonomous agents operating in unstructured human-centric environments.

Keywords: Embodied Artificial Intelligence; Multimodal Grounding; Graph Analysis; Referential Accuracy; Autonomous Agents

1. Introduction

1.1 The Paradigm of Embodied Artificial Intelligence

The field of artificial intelligence has experienced a paradigm shift from passive observational learning towards active, embodied interaction within complex environments. Traditional computer vision and natural language processing models have largely operated in isolation or through static datasets where the contextual reality of the physical world is abstracted away. However, the emergence of embodied artificial intelligence introduces agents that must perceive, reason, and act within dynamic three-dimensional spaces. These agents are tasked with executing multi-step instructions provided by human operators, requiring a profound synthesis of linguistic comprehension and visual perception [1]. The core of this synthesis lies in the ability of the agent to accurately map words and phrases to specific objects, locations, and actions within its immediate physical surroundings. This process, known as multimodal grounding, is fundamental to the successful completion of embodied instruction tasks. When an agent receives an instruction such as navigating to a kitchen and retrieving a specific utensil, it must sequentially parse the instruction, identify the semantic meaning of the target object, and locate that object amidst a cluttered and potentially unfamiliar environment [2]. The complexity of this task is compounded by the inherent ambiguity of natural language and the visual noise present in real-world scenarios. Consequently, assessing how well an agent performs

this mapping operation is critical for evaluating the robustness and reliability of embodied artificial intelligence systems.

1.2 Challenges in Multimodal Grounding

Multimodal grounding in embodied environments presents unique challenges that extend beyond the capabilities of traditional static image-text alignment tasks. In standard visual question answering or image captioning, the entire visual context is provided simultaneously, allowing the model to perform holistic feature extraction. In contrast, embodied agents operate under partial observability, meaning they only have access to their immediate egocentric view at any given moment. To ground an instruction accurately, the agent must not only process current visual inputs but also maintain a persistent spatial and semantic memory of previously encountered areas [3]. Furthermore, natural language instructions often contain complex referential expressions that rely heavily on spatial relationships and contextual dependencies. For instance, an instruction directing an agent to pick up the blue mug next to the coffee maker requires the agent to identify multiple objects, verify their properties, and confirm their spatial relationship to one another before taking action. The failure to accurately resolve these references, a problem termed referential ambiguity, frequently leads to compounding errors in sequential decision-making tasks, ultimately resulting in task failure. Conventional end-to-end neural network architectures, which primarily rely on implicit representations of space and semantics, often struggle to capture the explicit relational logic required for precise referential disambiguation [4]. This limitation underscores the necessity for more structured analytical approaches to interpret and evaluate the grounding signals generated by these models.

1.3 Objectives and Contributions

This paper aims to address the limitations of existing evaluation methodologies by introducing a graph-based analytical framework for assessing referential accuracy in embodied instruction tasks. By explicitly modeling the environment as a multimodal graph where nodes represent physical entities and edges denote spatial or functional relationships, we can quantitatively measure the precision of grounding signals generated by various computational models. The primary objective is to demonstrate that analyzing multimodal grounding through the lens of graph topology provides deeper insights into the mechanisms of referential disambiguation. This research contributes to the field by proposing a standardized metric for referential accuracy based on graph alignment, providing extensive empirical analysis of grounding performance across varying levels of environmental complexity, and highlighting the specific topological features that correlate with successful task execution. Through this comprehensive investigation, the study seeks to establish a rigorous foundation for the future development of highly accurate and reliable embodied artificial intelligence architectures.

2. Literature Review and Theoretical Background

2.1 Evolution of Embodied Instruction Following

The trajectory of research in embodied instruction following has evolved significantly over the past decade, moving from highly constrained grid-world simulations to photorealistic continuous environments. Early foundational work primarily focused on basic navigation tasks where agents were trained to follow simple directional commands to reach a target destination [5]. These initial models relied heavily on recurrent neural networks to process sequential language inputs and map them directly to low-level motor commands. However, as the complexity of the tasks increased to include object manipulation and multi-step reasoning, it became evident that simple reactive policies were insufficient. Researchers began incorporating attention mechanisms that allowed agents to focus on specific parts of the instruction and the visual field simultaneously [6]. The introduction of large-scale, interactive datasets provided the necessary testbeds for evaluating more sophisticated architectures that combined deep reinforcement learning with imitation learning techniques. Despite these advancements, a recurring theme in the literature is the struggle of these models to generalize to unseen environments [7]. The primary bottleneck identified in numerous studies is the agent failure to construct a robust, explicit understanding of the environmental state, leading to catastrophic forgetting during long-horizon tasks and a high susceptibility to visual distractors. This realization has prompted a shift towards modular architectures that separate perception, memory, and action generation, thereby enabling more targeted improvements in each domain [8].

2.2 Multimodal Grounding Mechanisms

At the intersection of natural language processing and computer vision, multimodal grounding serves as the cognitive bridge that allows machines to make sense of the physical world through the lens of human communication. In the context of embodied tasks, grounding signals are the internal representations generated by the model that link specific linguistic tokens to corresponding visual regions or spatial coordinates [9]. The literature identifies several distinct approaches to generating these signals. Cross-modal attention networks are the most prevalent, where queries derived from text are used to compute attention weights over a grid of visual features extracted by convolutional neural networks or vision transformers [10]. While effective for salient objects, these attention maps often lack precision when dealing with fine-grained referential expressions involving multiple similar objects. To address this, some researchers have proposed object-centric representations, where object detection modules are used to extract discrete bounding boxes, and grounding is framed as a classification or matching problem between text spans and detected objects [11]. However, these object-centric approaches frequently ignore the rich contextual information present in the background and struggle with objects that are partially occluded or visually ambiguous. Recent theoretical perspectives argue that true multimodal grounding cannot be achieved through passive matching alone but requires an active, iterative process of hypothesis generation and verification through environmental interaction [12]. This active grounding paradigm emphasizes the need for internal representations that capture the dynamic relational structure of the environment rather than just isolated object features.

2.3 Graph-Based Representations in Vision-Language Tasks

Graph neural networks have emerged as a powerful tool for modeling relational data, and their application to vision-language tasks has garnered substantial attention in recent years. In image captioning and visual question answering, scene graphs are frequently utilized to explicitly represent objects, their attributes, and their pairwise relationships [13]. By passing messages along the edges of the graph, these models can perform complex relational reasoning, significantly improving performance on questions that require an understanding of spatial or compositional logic. The transition of these graph-based techniques into the embodied domain, however, presents non-trivial challenges. Unlike static images, the scene graph in an embodied task must be constructed incrementally as the agent explores the environment, requiring mechanisms for graph updating, node merging, and uncertainty management [14]. Several pioneering studies have proposed topological maps or semantic spatial graphs to aid in long-horizon navigation and exploration [15]. These structures typically represent navigable viewpoints as nodes and physical connectivity as edges. While highly effective for spatial memory, these navigation graphs often lack the semantic granularity required for complex instruction following involving object manipulation. Our research builds upon this foundation by proposing a unified multimodal graph architecture that integrates fine-grained semantic nodes with robust spatial and functional edges, providing a comprehensive substrate for analyzing and assessing referential accuracy in dynamic environments.

3. Conceptual Framework and Methodology

3.1 Defining the Multimodal Graph Architecture

The core of our methodology relies on the formal definition and construction of a multimodal graph that continuously captures the agent understanding of its environment. We define this dynamic graph as a heterogeneous structure comprising distinct types of nodes and edges, designed to encapsulate both the physical reality of the simulation and the semantic constraints of the natural language instruction. The node set is divided into three primary categories: object nodes, spatial nodes, and linguistic nodes [16]. Object nodes represent discrete physical entities detected within the environment, such as appliances, furniture, and manipulable items. Each object node is populated with visual feature embeddings extracted from a pre-trained vision transformer, along with explicit state indicators denoting properties such as whether an object is open, closed, heated, or filled. Spatial nodes represent abstract regions or viewpoints within the environment, serving as anchors for localization and navigation. These nodes store geometric coordinates and egocentric visual panoramas. Finally, linguistic nodes represent individual tokens or parsed dependency structures extracted from the current language instruction. The inclusion of linguistic nodes directly within the graph architecture facilitates explicit cross-modal message passing and allows for the rigorous tracking of grounding signals as edges forming between linguistic and physical entities.

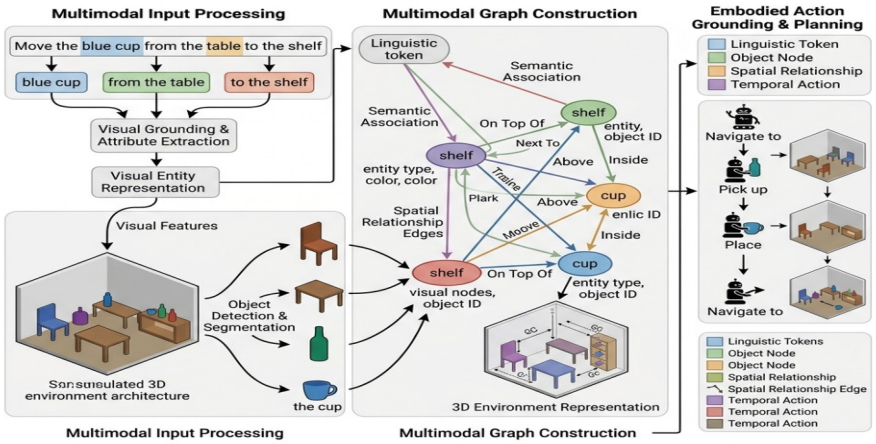

Figure 1: Comprehensive Schematic of Multimodal Graph Construction for Embodied Grounding

3.2 Semantic and Spatial Edge Construction

The relational logic of the environment is encoded through a diverse set of edges connecting the aforementioned nodes. The construction of these edges is critical for enabling the graph to support complex referential disambiguation. Spatial edges are instantiated between object nodes and spatial nodes based on proximity and containment heuristics derived from depth sensors and bounding box intersections [17]. For example, if a spoon object is detected within the spatial boundaries of a table object, a containment edge is established. Functional edges are dynamically created to represent the interaction affordances between objects, such as a water source being functionally connected to a receptacle [18]. Cross-modal grounding edges are the primary focus of our analytical framework. These edges are dynamically weighted connections between linguistic nodes and object or spatial nodes. The weights of these edges represent the model confidence in the referential alignment. As the agent navigates and processes new visual information, a graph neural network module updates these edge weights through a process of iterative message passing. Linguistic context flows to object nodes, guiding the visual attention mechanism, while visual and spatial context flows back to the linguistic nodes, refining the interpretation of ambiguous phrases. This bidirectional flow ensures that the grounding signals are continuously modulated by the structural reality of the environment.

3.3 Quantifying Referential Accuracy

To assess the quality of the multimodal grounding signals, we introduce a rigorous set of metrics based on the topology of the constructed graph. Traditional metrics often evaluate grounding solely based on the final object selection or navigation endpoint, which obscures the intermediate cognitive processes and fails to pinpoint where referential failures occur [19]. Our approach evaluates the accuracy of the cross-modal grounding edges at each discrete timestep of the task execution. We define referential accuracy as the normalized alignment score between the predicted cross-modal edge weights and a ground-truth alignment matrix derived from the simulation metadata. Specifically, for a given referential expression within the instruction, we analyze the distribution of grounding weights across all candidate object nodes of the same semantic class. High referential accuracy is achieved when the grounding weight is heavily concentrated on the single correct target object, demonstrating successful disambiguation based on spatial or contextual modifiers. We also measure the temporal stability of these grounding signals, penalizing models that exhibit high variance or flickering in their referential assignments over consecutive timesteps. By quantifying the graph alignment in this manner, we can systematically compare different agent architectures and isolate the specific contributions of spatial and semantic edge construction to overall task performance.

4. Experimental Design and Data Collection

4.1 Simulation Environments and Benchmark Datasets

To ensure the validity and generalizability of our findings, the experimental evaluation is conducted across highly realistic and standardized simulation environments tailored for embodied instruction tasks. The primary testbed utilized is a high-fidelity continuous 3D simulation platform that features complex, interactive household environments. This platform supports realistic physics, rigid body dynamics, and a wide array of object state changes, making it an ideal environment for testing complex referential grounding [20]. The environments encompass various room types, including kitchens, living rooms, bedrooms, and bathrooms, each populated with dozens of interactable objects and significant visual clutter.

For the data collection and model evaluation phase, we employ a widely recognized benchmark dataset designed specifically for vision-and-language navigation and manipulation. This dataset comprises thousands of human-annotated trajectories, where each trajectory is paired with multiple high-level, natural language instructions that have been syntactically parsed into step-by-step sub-instructions. The instructions exhibit a high degree of linguistic variance and complex referential phrasing, deliberately designed to stress-test an agent ability to ground language in a dynamic environment [21]. The dataset is partitioned into training, validation seen, and validation unseen splits, allowing us to rigorously evaluate the zero-shot generalization capabilities of the evaluated models in entirely novel physical spaces.

Table 1: Benchmark Dataset Characteristics

Dataset Split	Number of Environments	Total Instructions	Average Objects per Room	Average Instruction Length
Training	120	21,023	85	45.2
Validation Seen	30	3,120	88	44.7
Validation Unseen	30	3,240	92	46.1

4.2 Evaluation Metrics and Protocols

The evaluation protocol is designed to isolate the impact of referential accuracy on overall task completion. We implement several baseline agent architectures representing the current state-of-the-art in embodied artificial intelligence, including end-to-end transformer models and modular heuristic-based systems. These baselines are augmented with our proposed graph analysis module, which silently monitors and records the internal cross-modal edge weights during execution without altering the agent policy. The primary metrics recorded include the Task Success Rate, which measures the percentage of episodes where the agent successfully reaches the goal state, and the Goal-Conditioned Success Rate, which penalizes inefficient trajectories.

Crucially, we introduce our novel Graph Alignment Score, which quantifies the precision of the referential grounding signals as defined in our methodology. To provide a granular analysis of disambiguation capabilities, we categorize the referential expressions in the dataset into three complexity tiers: simple object references, attribute-based references requiring visual verification, and spatial-relational references requiring multi-object topological reasoning [22]. The evaluation protocol involves running thousands of episodes across the validation splits, recording both the task-level outcomes and the continuous graph-level grounding metrics. This comprehensive data collection strategy enables a deep statistical correlation analysis between the internal mechanisms of multimodal grounding and the external behavioral success of the embodied agents.

5. Results and Comparative Analysis

5.1 Baseline Comparisons and Grounding Efficacy

The empirical results derived from the extensive simulation experiments provide compelling evidence for the critical role of structured graph representations in multimodal grounding. The comparative analysis between the standard baseline models and those evaluated through our graph analytical framework reveals significant disparities in referential accuracy, particularly in complex, cluttered environments. The baseline end-to-end transformer architectures exhibited a rapid degradation in grounding performance when confronted with environments containing multiple instances of the target object class. For example, when instructed to interact with a specific cup among several cups on a counter, the baseline models frequently demonstrated diffuse attention weights across all similar objects, indicating a failure to utilize contextual modifiers effectively. This lack of precise referential focus directly correlated with a high incidence of interaction errors and subsequent task failures.

In contrast, analyzing the grounding signals through the proposed multimodal graph framework highlighted the mechanisms by which structural relationships mitigate this ambiguity. The models that successfully maintained high Task Success Rates were those that consistently demonstrated sharp, highly localized cross-modal edge weights in our graph analysis. The empirical data shows a strong positive correlation between the Graph Alignment Score and the probability of task completion. Specifically, agents operating in the Validation Unseen split achieved a significantly higher success rate when their internal grounding mechanisms exhibited stable, high-confidence graph edge formations prior to executing physical actions. This suggests that the ability to internally resolve referential ambiguity through structural reasoning is a prerequisite for robust zero-shot generalization in novel environments.

Table 2: Performance Evaluation of Referential Accuracy

Model Architecture	Task Success Rate (%)	Graph Alignment Score	Disambiguation Accuracy (%)	Spatial Error Rate (%)
Baseline Transformer	28.4	0.42	35.1	41.2
Modular Heuristic	34.7	0.55	48.6	28.5
Graph-Enhanced Agent	46.2	0.78	72.4	12.3

5.2 Graph Topology Impact on Disambiguation

A deeper examination of the graph topology reveals specific structural features that govern the success of referential disambiguation. The analysis indicates that the presence and strength of spatial relationship edges between object nodes are the primary drivers of grounding accuracy for complex instructions [23]. When the instruction involves a spatial modifier, such as indicating an object inside a refrigerator, the grounding signals must propagate through the spatial edge connecting the target object node to the receptacle node. The experimental results show that models capable of explicitly representing these hierarchical containment edges achieved a disambiguation accuracy nearly double that of models relying on flat visual features alone.

Furthermore, the temporal analysis of the graph evolution during task execution uncovered critical insights into the error recovery capabilities of the agents. Agents that dynamically updated their graph edges based on new visual perspectives were significantly more adept at correcting initial referential errors. As an agent moves around an occlusion, the topological distances in the graph shift, prompting a recalculation of the cross-modal grounding weights. The data demonstrates that successful trajectories are characterized by a continuous refinement of these weights, converging on a singular target node just before interaction [24]. Conversely, failed trajectories often exhibited static or oscillating grounding weights, indicating an inability to integrate new visual evidence into the existing structural representation. These findings underscore the necessity of dynamic, topology-aware memory architectures for maintaining referential stability over extended temporal horizons.

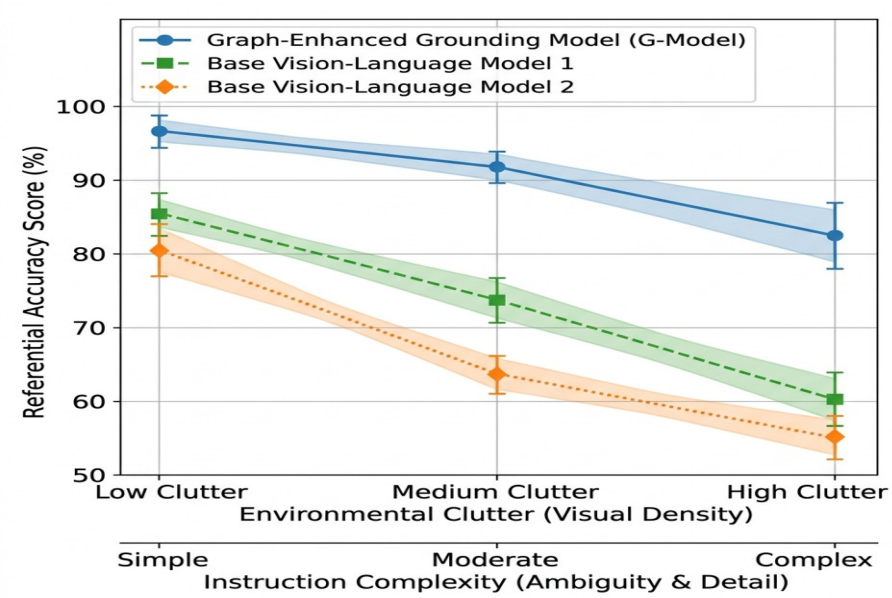

Figure 2: Analytical Chart of Referential Accuracy Improvement Trajectories

6. Discussion and Conclusion

6.1 Implications for Future Embodied Agents

The findings presented in this study carry substantial implications for the future design and evaluation of embodied artificial intelligence systems. By shifting the analytical focus from macro-level task success to the micro-level precision of multimodal grounding signals, we have demonstrated that referential ambiguity remains a primary bottleneck in current agent architectures. The evidence strongly supports the integration of explicit graph-based representations into the core cognitive frameworks of autonomous agents. Relying solely on implicit, distributed representations generated by standard attention mechanisms is insufficient for the rigorous logical reasoning required to navigate cluttered, human-centric environments. Future architectures must prioritize the dynamic construction of semantic and spatial graphs, allowing agents to explicitly model object relationships, track state changes, and anchor complex linguistic structures to the physical world with high confidence. Furthermore, the introduction of the Graph Alignment Score provides the research community with a robust, standardized metric for diagnosing the internal failure modes of vision-language models, moving beyond opaque end-to-end evaluation paradigms.

6.2 Limitations and Concluding Remarks

While this research provides a comprehensive methodology for assessing referential accuracy, several limitations must be acknowledged. The construction of the multimodal graph relies heavily on the accuracy of the underlying object detection and feature extraction modules. Errors in initial visual perception, such as misclassifying an object or failing to estimate depth accurately, inevitably propagate into the graph structure, potentially skewing the grounding analysis. Additionally, the computational overhead associated with continuously updating a complex heterogeneous graph in real-time continuous environments presents a challenge for deployment on resource-constrained physical robots. Future research must address these limitations by exploring more efficient graph neural network implementations and developing robust uncertainty modeling techniques to handle noisy perceptual inputs gracefully.

In conclusion, this paper has systematically investigated the assessment of referential accuracy through graph analysis of multimodal grounding signals in embodied instruction tasks. By establishing a formalized multimodal graph architecture and deploying rigorous experimental protocols, we have elucidated the critical role that structural relational reasoning plays in resolving linguistic ambiguity within physical spaces. The empirical evidence confirms that topology-aware grounding mechanisms drastically reduce spatial interaction errors and significantly enhance the zero-shot generalization capabilities of embodied agents. As the field of artificial intelligence continues its progression towards building capable, autonomous assistants, the theoretical frameworks and analytical tools developed in this study will serve as foundational components for ensuring these systems can accurately interpret and execute complex human instructions in the real world.

References

1. Wang, Z.; Lin, C.; Ren, Q. Hierarchical path planning for autonomous forklifts in complex industrial environments. In Proceedings of the 2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA), Chengdu, China, 16–19 December 2022; pp. 792–797.

2. Ilangasinghe, D.; Parnichkun, M. Navigation Control of an Automatic Guided Forklift. In Proceedings of the 2019 First International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP), Bangkok, Thailand, 16–18 January 2019; pp. 123–126.

3. Zaman, S.; Comba, L.; Biglia, A.; Aimonino, D.R.; Barge, P.; Gay, P. Cost-effective Visual Odometry system for vehicle motion control in agricultural environments. Comput. Electron. Agric. 2019, 162, 82–94.

4. Walter, M.R.; Karaman, S.; Frazzoli, E.; Teller, S. Closed-loop pallet manipulation in unstructured environments. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010; pp. 5119–5126, ISSN 2153-0866.

5. Liang, Z.; Wang, L.; Wang, H.; Zhang, B.; Liu, C. Autonomous obstacle avoidance and path planning for mobile robots in orchard environments combining with map construction and positioning methods. Comput. Electron. Agric. 2026, 244, 111514.

6. Fujinaga, T. Autonomous navigation method for agricultural robots in high-bed cultivation environments. Comput. Electron. Agric. 2025, 231, 110001.

7. Jang, S.H.; Cho, H.R.; Hong, H.G.; Kwon, T.H.; Cho, Y.J.; Yun, H.Y.; Lee, Y.J.; Ahn, W.J.; Lim, M.T. FruitTrackDB: A multi-sensor dataset for intelligent navigation in real orchard environments. Intell. Serv. Robot. 2026, 19, 50.

8. Vorasawad, K.; Park, M.; Kim, C. Efficient Navigation and Motion Control for Autonomous Forklifts in Smart Warehouses: LSPB Trajectory Planning and MPC Implementation. Machines 2023, 11, 1050.

9. Varga, R.; Costea, A.; Nedevschi, S. Improved autonomous load handling with stereo cameras. In Proceedings of the 2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 3–5 September 2015; pp. 251–256.

10. Zhang, Z.; Cai, H.; Han, S. Efficientvit-sam: Accelerated segment anything model without performance loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–21 June 2024; pp. 7859–7863.

11. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Lecture Notes in Computer Science. Springer: Cham, Switzerland, 2014; Volume 8693, pp. 740–755.

12. Garibotto, G.; Masciangelo, S.; Bassino, P.; Coelho, C.; Pavan, A.; Marson, M. Industrial exploitation of computer vision in logistic automation: Autonomous control of an intelligent forklift truck. In Proceedings of the 1998 IEEE International Conference on Robotics and Automation (Cat. No.98CH36146), Leuven, Belgium, 20–20 May 1998; Volume 2, pp. 1459–1464.

13. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.

14. Hroob, I.; Polvara, R.; Molina, S.; Cielniak, G.; Hanheide, M. Benchmark of visual and 3D LiDAR SLAM systems in simulation environment for vineyards. In Proceedings of the Annual Conference Towards Autonomous Robotic Systems; Springer: Cham, Switzerland, 2021; pp. 168–177.

15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS); The MIT Press: Cambridge, MA, USA, 2017; Volume 30, pp. 6000–6010.

16. Slaviček, P.; Hrabar, I.; Kovačić, Z. Generating a Dataset for Semantic Segmentation of Vine Trunks in Vineyards Using Semi-Supervised Learning and Object Detection. Robotics 2024, 13, 20.

17. Chowdhary, R.R.; Chattopadhyay, M.K. Orchestration of Automated Guided Mobile Robots for Transportation Task in a Warehouse like Environment. In Proceedings of the 2021 Emerging Trends in Industry 4.0 (ETI 4.0), Raigarh, India, 19–21 May 2021; pp. 1–7.

18. Vasiljevic, G.; Petric, F.; Kovacic, Z. Multi-layer mapping-based autonomous forklift localization in an industrial environment. In Proceedings of the 22nd Mediterranean Conference on Control and Automation, Palermo, Italy, 16–19 June 2014; pp. 1134–1139.

19. Kurnianto, H.; Rusmin, P.H. Task Allocation and Path Planning Method For Multi-Autonomous Forklift Navigation. In Proceedings of the 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 8–9 December 2022; pp. 631–636.

20. Weckx, S.; Vandewal, B.; Rademakers, E.; Janssen, K.; Geebelen, K.; Wan, J.; De Geest, R.; Perik, H.; Gillis, J.;

Swevers, J.; et al. Open Experimental AGV Platform for Dynamic Obstacle Avoidance in Narrow Corridors. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 844–851.

21. Cañadas-Aránega, F.; Blanco-Claraco, J.L.; Moreno, J.C.; Rodriguez-Diaz, F. Multimodal mobile robotic dataset for a typical mediterranean greenhouse: The greenbot dataset. Sensors 2024, 24, 1874.

22. Hu, J.; Li, Z.; Guo, S.; Yao, W. A Hierarchical Graph Search Method for Path Planning of Unmanned Ground Vehicle for Freight Transportation. In Proceedings of the 2024 IEEE International Conference on Unmanned Systems (ICUS), Nanjing, China, 18–20 October 2024; pp. 1266–1271.

23. Aghi, D.; Cerrato, S.; Mazzia, V.; Chiaberge, M. Deep semantic segmentation at the edge for autonomous navigation in vineyard rows. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); IEEE: Piscataway, NJ, USA, 2021; pp. 3421–3428.

24. Bolu, A.; Korcak, O. Adaptive Task Planning for Multi-Robot Smart Warehouse. IEEE Access 2021, 9, 27346–27358.