A Comprehensive Review on Multimodal Language Models and Their Applications in Robotics

张三

January 10, 2025

Abstract

This paper presents a comprehensive review of the latest advancements in multimodal language models and their applications in robotics. The review covers a wide range of topics, including embodied multimodal language models, robotic vision-language planning, embodied generalist agents, imitation learning and data generation, and advanced techniques in object detection. Specifically, it delves into the details of PaLM-E, an embodied multimodal language model that integrates real-world continuous sensor modalities into language models for enhanced robotic tasks. It also explores ManipVQA and ManipLLM, which aim to improve robotic manipulation through the infusion of robotic affordance and physically grounded information, and the leveraging of multimodal large language models, respectively. The review further discusses the significance of robotic vision-language planning with the introduction of ViLa, a novel approach that leverages vision-language models for long-horizon robotic planning. Additionally, it highlights the development of LEO, an embodied multi-modal generalist agent proficient in 3D world tasks, and various imitation learning and data generation systems such as MimicPlay, Vid2Robot, Universal Manipulation Interface (UMI), DexCap, HIRO Hand, RoboGen, and MimicGen, which collectively aim to enhance robot learning through human demonstrations and generative simulation. Lastly, the review touches upon advanced object detection techniques, focusing on the Adaptive Clustering Transformer (ACT) for efficient end-to-end object detection. This review aims to provide a thorough understanding of the current state and future directions of multimodal language models in robotics.

1 Introduction

The integration of multimodal language models (MLMs) into robotics represents a significant leap forward in the quest to bridge the gap between artificial intelligence and real-world applications. These models, which combine the prowess of large language models (LLMs) with the ability to process and understand continuous sensor data, are setting new benchmarks in the field of robotics. The essence of these advancements lies in their ability to ground language in the physical world, enabling robots to perform complex tasks with a level of understanding and adaptability previously unattainable. This review delves into the latest developments in embodied multimodal language models, robotic vision-language planning, embodied generalist agents, imitation learning, data generation techniques, and advanced object detection methods. Each of these areas contributes uniquely to the overarching goal of creating robots that can seamlessly interact with their environment, understand and execute tasks based on natural language instructions, and learn from human demonstrations in a scalable and efficient manner. The exploration begins with an examination of PaLM-E, an embodied multimodal language model that integrates real-world continuous sensor modalities into language models, thereby establishing a direct link between words and percepts. Following this, the review discusses ManipVQA and ManipLLM, which focus on injecting robotic affordance and physically grounded information into MLMs and leveraging the robust reasoning capabilities of MLMs for object-centric robotic manipulation, respectively. The discussion then shifts to Robotic Vision-Language Planning (ViLa), a novel approach that leverages vision-language models for long-horizon robotic planning, and LEO, an embodied multi-modal generalist agent that excels in perceiving, grounding, reasoning, planning, and acting in the 3D world. The review also covers MimicPlay and Vid2Robot, which explore long-horizon imitation learning by watching human play and end-to-end video-conditioned policy learning with cross-attention transformers, respectively. Additionally, the Universal Manipulation Interface (UMI), DexCap, and a wearable robotic hand for hand-over-hand imitation learning are examined for their contributions to scalable and portable data collection and policy learning frameworks. The exploration concludes with a look at RoboGen and MimicGen, which aim to unleash infinite data for automated robot learning via generative simulation and generate large-scale, rich datasets

from a small number of human demonstrations, respectively. Finally, the review touches upon advanced techniques in object detection, specifically the Adaptive Clustering Transformer (ACT), which proposes a novel variant of transformer to reduce computation cost for high-resolution input in object detection tasks. Through this comprehensive review, the aim is to provide a detailed overview of the current state of the art in multimodal language models and their applications in robotics, highlighting the challenges, opportunities, and future directions in this rapidly evolving field.

2 Embodied Multimodal Language Models

Embodied Multimodal Language Models represent a significant advancement in the field of robotics and artificial intelligence, integrating real-world continuous sensor modalities directly into language models. This integration establishes a crucial link between words and percepts, enabling these models to perform a wide range of embodied tasks with remarkable efficiency. The following subsections delve into the specifics of three such models: PaLM-E, ManipVQA, and ManipLLM, each contributing uniquely to the domain of embodied multimodal language models.

2.1 PaLM-E: An Embodied Multimodal Language Model

PaLM-E, as proposed by Driess et al. (2023), is a groundbreaking model that incorporates visual, continuous state estimation, and textual input encodings into a single large embodied multimodal model. This model is trained end-to-end, in conjunction with a pre-trained large language model, for multiple embodied tasks including sequential robotic manipulation planning, visual question answering, and captioning. The evaluations of PaLM-E demonstrate its ability to address a variety of embodied reasoning tasks across different observation modalities and embodiments. Notably, PaLM-E exhibits positive transfer, benefiting from diverse joint training across internet-scale language, vision, and visual-language domains. The largest model, PaLM-E-562B with 562 billion parameters, not only excels in robotics tasks but also achieves state-of-the-art performance on OK-VQA, retaining generalist language capabilities with increasing scale.

2.2 ManipVQA: Injecting Robotic Affordance and Physically Grounded Information

ManipVQA, introduced by Huang et al. (2024), addresses the limitations of conventional Multi-modal Large Language Models (MLLMs) in understanding robotics-specific knowledge crucial for manipulation tasks. By infusing MLLMs with manipulation-centric knowledge through a Visual Question-Answering (VQA) format, ManipVQA enhances the models' ability to comprehend tool detection, affordance recognition, and physical concepts. The framework leverages a unified VQA format and a fine-tuning strategy to integrate robotics-specific knowledge with the inherent vision-reasoning capabilities of MLLMs. Empirical evaluations in robotic simulators and across various vision task benchmarks demonstrate the robust performance of ManipVQA, showcasing its potential to significantly improve robotic manipulation tasks.

2.3 ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation

ManipLLM, developed by Li et al. (2024), introduces an innovative approach for robot manipulation by leveraging the robust reasoning capabilities of Multimodal Large Language Models (MLLMs). This model enhances the stability and generalization of manipulation through a fine-tuning paradigm that encompasses object category understanding, affordance prior reasoning, and object-centric pose prediction. ManipLLM utilizes an RGB image and text prompt to predict the end effector's pose in a chain of thoughts, followed by an active impedance adaptation policy for planning upcoming waypoints in a closed-loop manner. The model's performance in both simulator and real-world experiments highlights its promising capabilities in object-centric robotic manipulation, paving the way for more advanced and versatile robotic systems.

3 Robotic Vision-Language Planning

Recent advancements in robotics have underscored the importance of integrating vision and language for effective task planning and execution. The study titled "Look Before You Leap: Unveiling the Power of GPT-4V in

Robotic Vision-Language Planning" by Hu et al. (2023) introduces a novel approach that leverages the capabilities of vision-language models (VLMs) for long-horizon robotic planning. This approach, termed Robotic Vision-Language Planning (ViLa), aims to overcome the limitations of traditional large language models (LLMs) by directly integrating perceptual data into the reasoning and planning process. ViLa's methodology enables a profound understanding of commonsense knowledge in the visual world, including spatial layouts and object attributes, thereby facilitating flexible multimodal goal specification and the natural incorporation of visual feedback.

The significance of ViLa lies in its ability to generate a sequence of actionable steps for robots by leveraging the extensive knowledge embedded in VLMs. This is particularly crucial for tasks that require a deep understanding of the environment and the objects within it. The authors argue that a task planner should be an inherently grounded, unified multimodal system, capable of jointly reasoning with LLMs without the need for external affordance models. This perspective is supported by extensive evaluations conducted in both real-robot and simulated environments, which demonstrate ViLa's superiority over existing LLM-based planners. The results highlight ViLa's effectiveness in a wide array of open-world manipulation tasks, showcasing its potential to significantly enhance robotic capabilities in complex environments.

Moreover, the study by Hu et al. (2023) contributes to the broader discourse on the integration of vision and language in robotics by providing a comprehensive framework that bridges the gap between high-level task planning and low-level execution. The authors' approach not only advances the state-of-the-art in robotic planning but also opens new avenues for research in embodied AI, where the seamless integration of perception, reasoning, and action is paramount. The implications of this work extend beyond robotics, offering insights into the development of more general and adaptable AI systems capable of operating in dynamic and unstructured environments.

In conclusion, the exploration of Robotic Vision-Language Planning, as exemplified by the work of Hu et al. (2023), represents a significant step forward in the quest to imbue robots with the ability to perform complex tasks in the real world. By leveraging the power of VLMs, this approach offers a promising pathway towards the realization of more intelligent, au-

tonomous, and versatile robotic systems. The continued evolution of this field is expected to yield further innovations that will enhance the interaction between robots and their environments, ultimately leading to more effective and efficient robotic solutions for a wide range of applications.

4 Embodied Generalist Agents

Embodied Generalist Agents represent a significant leap forward in the field of robotics and artificial intelligence, aiming to bridge the gap between digital intelligence and physical interaction within the 3D world. These agents are designed to perceive, reason, plan, and act within complex environments, leveraging multimodal inputs to achieve a wide range of tasks. The development of such agents is driven by the need for more versatile and adaptable robotic systems that can operate in diverse and unstructured environments, from household settings to industrial applications.

4.1 An Embodied Generalist Agent in 3D World

The concept of an Embodied Generalist Agent in the 3D world, as introduced by Huang et al. (2023), marks a pivotal advancement in the quest for general-purpose robotic systems. LEO, the agent developed by the authors, is a testament to the potential of integrating large language models (LLMs) with 3D perception and action capabilities. LEO's architecture is built upon a unified task interface, model architecture, and objective, trained in two stages: 3D vision-language (VL) alignment and 3D vision-language-action (VLA) instruction tuning.

LEO's training regimen involves large-scale datasets that encompass a variety of object-level and scene-level tasks, necessitating a deep understanding of and interaction with the 3D world. The authors have meticulously designed an LLM-assisted pipeline to generate high-quality 3D VL data, ensuring that LEO is equipped with the necessary knowledge to perform tasks ranging from 3D captioning and question answering to embodied reasoning, navigation, and manipulation.

The significance of LEO's development lies in its ability to perform across a wide spectrum of tasks without the need for task-specific training. This is achieved through the agent's capacity for positive transfer learning, where knowledge gained from one task can be applied to another, enhancing

the agent's adaptability and efficiency. LEO's proficiency in 3D tasks not only demonstrates the feasibility of creating generalist agents but also highlights the importance of 3D perception and interaction in achieving general intelligence.

Moreover, LEO's success underscores the potential of leveraging LLMs for robotics, suggesting a future where robotic systems can understand and execute complex instructions with minimal human intervention. The agent's ability to navigate and manipulate objects in the 3D world with a high degree of autonomy opens up new possibilities for applications in areas such as autonomous vehicles, robotic assistants, and smart manufacturing.

In conclusion, the development of Embodied Generalist Agents like LEO represents a significant milestone in robotics and AI. By integrating 3D perception and action with the reasoning capabilities of LLMs, these agents offer a glimpse into a future where robots can seamlessly interact with the physical world, performing a wide range of tasks with human-like proficiency. The work of Huang et al. (2023) not only advances the field but also sets a new benchmark for what is achievable in the realm of general-purpose robotics.

5 Imitation Learning and Data Generation

Imitation learning and data generation have emerged as pivotal methodologies in the field of robotics, enabling robots to acquire complex manipulation skills through the observation of human actions and the synthesis of large-scale datasets. This section delves into several innovative approaches that leverage these methodologies to enhance robotic capabilities.

5.1 MimicPlay: Long-Horizon Imitation Learning by Watching Human Play

MimicPlay introduces a hierarchical learning framework that capitalizes on human play data to facilitate the learning of long-horizon tasks by robots. By extracting latent plans from freely interacting human videos, MimicPlay guides low-level visuomotor control, significantly reducing the data requirement for learning complex tasks. This approach has demonstrated superior performance in task success rate, generalization ability,

and robustness to disturbances across 14 long-horizon manipulation tasks [6].

5.2 Vid2Robot: End-to-end Video-conditioned Policy Learning with Cross-Attention Transformers

Vid2Robot represents a groundbreaking end-to-end video-conditioned policy that translates human demonstration videos into robot actions. Utilizing cross-attention transformer layers, Vid2Robot aligns human and robot action representations, enabling robots to perform tasks as demonstrated in the videos. This method has shown over 20

5.3 Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots

The Universal Manipulation Interface (UMI) framework revolutionizes robot teaching by enabling skill transfer from in-the-wild human demonstrations to deployable robot policies. UMI's innovative design allows for the collection of rich, portable data and the learning of hardware-agnostic policies, facilitating zero-shot generalization to novel environments and objects. This approach unlocks new capabilities in dynamic, bimanual, precise, and long-horizon robotic behaviors [8].

5.4 DexCap: Scalable and Portable Mocap Data Collection System for Dexterous Manipulation

DexCap addresses the challenges of portability and complexity in hand motion capture systems for dexterous manipulation. By offering precise, occlusion-resistant tracking and integrating a novel imitation learning algorithm, DexCap enables robots to replicate human actions with remarkable accuracy. This system paves the way for effective learning from in-the-wild mocap data, advancing the pursuit of human-level robot dexterity [9].

5.5 A Wearable Robotic Hand for Hand-over-Hand Imitation Learning

The HIRO Hand, a wearable robotic hand, introduces a novel solution for expert data collection in dexterous manipulation. By enabling operators

to use their tactile feedback, the HIRO Hand ensures accurate imitation of expert actions, overcoming the limitations of traditional data gloves. This innovation facilitates the implementation of dexterous operations, enhancing the robot's grasping and in-hand manipulation abilities [10].

5.6 RoboGen: Towards Unleashing Infinite Data for Automated Robot Learning via Generative Simulation

RoboGen proposes a generative robotic agent that automates the learning of diverse robotic skills through generative simulation. By leveraging foundation and generative models, RoboGen generates diversified tasks, scenes, and training supervisions, scaling up robotic skill learning with minimal human supervision. This approach exemplifies the potential of generative schemes in robotics, offering an endless stream of skill demonstrations across diverse tasks and environments [11].

5.7 MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations

MimicGen introduces a system for synthesizing large-scale, rich datasets from a minimal number of human demonstrations. By adapting demonstrations to new contexts, MimicGen generates extensive datasets that enable effective robot training through imitation learning. This system demonstrates the effectiveness and utility of generated data, presenting a powerful and economical approach to scaling up robot learning [12].

6 Advanced Techniques in Object Detection

6.1 End-to-End Object Detection with Adaptive Clustering Transformer

The field of object detection has witnessed significant advancements with the introduction of transformer-based models, notably the End-to-End Object Detection with Transformer (DETR) approach. DETR revolutionized object detection by eliminating the need for hand-designed components like non-maximum suppression and anchor generation, instead relying on a set-based global loss that forces unique predictions via bipartite

matching. However, the computational demands of DETR, particularly for high-resolution inputs, pose a substantial challenge, necessitating innovative solutions to reduce complexity without compromising accuracy.

In response to these challenges, the Adaptive Clustering Transformer (ACT) was proposed as a novel variant of the transformer architecture aimed at reducing the computational cost associated with high-resolution inputs. ACT introduces an adaptive clustering mechanism that leverages Locality Sensitive Hashing (LSH) to cluster query features dynamically. This approach approximates the query-key interaction through prototype-key interactions, effectively reducing the quadratic complexity of self-attention mechanisms to a more manageable linear complexity. The key innovation of ACT lies in its ability to maintain high accuracy levels while significantly decreasing the computational load, making it a viable drop-in replacement for traditional self-attention modules without necessitating additional training.

The implications of ACT for object detection are profound. By enabling more efficient processing of high-resolution images, ACT facilitates the deployment of transformer-based object detection models in resource-constrained environments, such as mobile devices and embedded systems. Furthermore, the adaptability of ACT to various tasks beyond object detection, including image classification and segmentation, underscores its versatility and potential for broader applications in computer vision.

Empirical evaluations of ACT have demonstrated its effectiveness in achieving a favorable balance between accuracy and computational efficiency. The reduction in FLOPs (floating-point operations) without a corresponding drop in performance metrics such as mean Average Precision (mAP) highlights the practical benefits of ACT for real-world applications. Moreover, the open-source availability of ACT's codebase encourages further research and experimentation, paving the way for future innovations in transformer-based models.

In conclusion, the Adaptive Clustering Transformer represents a significant step forward in the quest for more efficient and scalable object detection models. By addressing the computational bottlenecks associated with high-resolution inputs, ACT not only enhances the feasibility of deploying transformer-based models in diverse settings but also contributes to the ongoing evolution of computer vision technologies. As the field continues to

advance, the principles underlying ACT may inspire further breakthroughs, driving the development of even more sophisticated and efficient models for object detection and beyond.

7 Conclusion

The exploration of multimodal language models and their applications in robotics has unveiled a transformative potential for the field. The integration of language, vision, and action within a unified framework, as demonstrated by PaLM-E, ManipVQA, and ManipLLM, has significantly advanced the capabilities of robots in understanding and executing complex tasks. These models have shown that by incorporating real-world continuous sensor modalities and grounding language in physical contexts, robots can achieve a level of general inference previously unattainable. The success of these models in tasks ranging from sequential robotic manipulation planning to visual question answering underscores the importance of multimodal integration in achieving embodied intelligence.

Robotic Vision-Language Planning, as exemplified by the ViLa approach, further extends the capabilities of robots by enabling them to perform long-horizon planning tasks with a profound understanding of commonsense knowledge in the visual world. This approach, leveraging the power of GPT-4V, highlights the potential of vision-language models in generating actionable steps for robots, thereby bridging the gap between high-level task understanding and low-level execution.

The development of embodied generalist agents, such as LEO, represents a significant leap towards achieving general intelligence in robots. By excelling in perceiving, grounding, reasoning, planning, and acting in the 3D world, LEO demonstrates the feasibility of creating agents capable of performing a wide spectrum of tasks with minimal task-specific training. This advancement not only enhances the versatility of robots but also paves the way for their application in diverse real-world scenarios.

Imitation learning and data generation techniques, including MimicPlay, Vid2Robot, Universal Manipulation Interface, DexCap, and MimicGen, have revolutionized the way robots acquire new skills. By leveraging human demonstrations and generative simulation, these approaches have significantly reduced the data requirements for training robots, making it

feasible to scale up robot learning. The ability to generate diverse and rich datasets from a small number of human demonstrations, as demonstrated by MimicGen, is particularly noteworthy for its potential to accelerate the development of robotic skills.

Finally, the advancements in object detection, exemplified by the Adaptive Clustering Transformer, have contributed to the efficiency and accuracy of robotic perception. By reducing the computational cost of high-resolution input processing, this technique enables robots to perform end-to-end object detection more effectively, further enhancing their ability to interact with the environment.

In conclusion, the integration of multimodal language models, robotic vision-language planning, embodied generalist agents, imitation learning, and advanced object detection techniques has significantly advanced the field of robotics. These developments not only enhance the capabilities of robots in performing complex tasks but also open up new avenues for research and application. As the field continues to evolve, the potential for creating more intelligent, versatile, and efficient robots is boundless, promising a future where robots can seamlessly integrate into and enhance our daily lives.

References

[1] Driess, D., Xia, F., Sajjadi, M. S. M., et al. (2023). Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

[2] Huang, S., Ponomarenko, I., Jiang, Z., et al. (2024). Manipvqa: Injecting robotic affordance and physically grounded information into multi-modal large language models. *arXiv preprint arXiv:2403.11289*.

[3] Li, X., Zhang, M., Geng, Y., et al. (2024). Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18061-18070.

[4] Hu, Y., Lin, F., Zhang, T., et al. (2023). Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*.

[5] Huang, J., Yong, S., Ma, X., et al. (2023). An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*.

[6] Wang, C., Fan, L., Sun, J., et al. (2023). Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*.

[7] Jain, V., Attarian, M., Joshi, N. J., et al. (2024). Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers. *arXiv preprint arXiv:2403.12943*.

[8] Chi, C., Xu, Z., Pan, C., et al. (2024). Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*.

[9] Wang, C., Shi, H., Wang, W., et al. (2024). Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*.

[10] Wei, D., Xu, H. (2024). A wearable robotic hand for hand-over-hand imitation learning. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 18113-18119.

[11] Wang, Y., Xian, Z., Chen, F., et al. (2023). Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*.

[12] Mandlekar, A., Nasiriany, S., Wen, B., et al. (2023). Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*.

[13] Zheng, M., Gao, P., Zhang, R., et al. (2020). End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*.