Journal of computer science and software applications

https://www.mfacademia.org/index.php/jcssa

ISSN:2377-0430

Vol. 5, No. 9, 2025

Vision-Language Models for Human-Robot Collaboration: Real-Time Task Understanding and Execution

Alaric Byrne
University of Southern Queensland, Toowoomba, Australia
alaric098@usq.edu.au

Abstract:

With the increasing complexity of stock markets and the nonlinear nature of stock price fluctuations, traditional financial forecasting methods often fail to achieve satisfactory results. This study proposes a hybrid neural network model that integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to enhance the accuracy of stock closing price prediction. The model leverages CNN to extract spatial features from historical financial indicators such as opening price, highest price, and trading volume, and then uses LSTM to capture temporal dependencies within the time series data. Experimental validation is conducted using a dataset of the CSI 300 Index from 1992 to 2021, demonstrating the proposed model's superior performance in comparison with CNN-only, LSTM- only, and CNN+RNN configurations. Evaluation metrics including Mean Relative Error (MRE) and Mean Absolute Error (MAE) indicate that the CNN-LSTM hy- brid network significantly improves prediction precision. The results highlight the potential of deep learning in modeling complex financial dynamics and offer insights into data-driven approaches for stock risk forecasting.

Keywords:

Stock Risk Prediction; Deep Learning; Convolutional Neural Network (CNN); Long Short-Term Memory (LSTM); Time Series Forecasting; Hybrid Neural Net- work; Financial Data Modeling.

1. Introduction

Human-robot collaboration (HRC) has emerged as a critical paradigm in modern robotics, enabling robots to assist humans in diverse domains such as manufacturing, healthcare, logistics, and domestic environments. Unlike traditional automation, HRC requires robots to interpret dynamic human intentions, adapt to unstructured environments, and execute complex tasks in real time. To achieve this, perception and communication must extend beyond predefined commands and low-level control, requiring robots to understand natural language instructions and contextual cues from the environment [1], [2].

Recent advances in Vision-Language Models (VLMs) have opened new opportunities for human-robot interaction. Models such as CLIP, BLIP-2, and Flamingo demonstrate remarkable performance in grounding visual perception with natural language semantics [3], [4]. These models learn joint embeddings of multimodal inputs, allowing systems to align visual objects with textual descriptions, infer relationships, and generalize across tasks without task-specific retraining. Integrating such capabilities into robotic systems enables robots to not only perceive objects but also to interpret high-level task instructions, bridging the gap between human intent and robotic execution [5].

For instance, consider a human providing an instruction such as "Place the red cup next to the kettle and then clean the table." Traditional task planning systems would require explicit symbolic representations and hand-crafted rules, making them brittle and domain-specific. In contrast, VLMs

ISSN:2377-0430

Vol. 5, No. 9, 2025

can parse the instruction, ground visual entities ("red cup," "kettle," "table"), and map these into structured action sequences. When integrated with robot control policies, this approach allows for real-time adaptation to unseen environments and tasks [6], [7].

Despite these advancements, significant challenges remain. First, VLMs often operate in static benchmarks and may not meet the real-time requirements of robotic systems. Delays in multimodal inference can compromise safety and efficiency in collaborative environments. Second, VLMs may struggle with fine-grained grounding in cluttered scenes, leading to execution errors. Third, the integration of high-dimensional embeddings into control pipelines requires robust interfaces between perception, reasoning, and action modules [8]. Addressing these issues is essential to fully unlock the potential of VLMs in robotics.

In this paper, we present a novel framework for real-time task understanding and execution in human-robot collaboration, leveraging vision-language models as the core semantic interpreter. The proposed system takes natural language instructions and real-time visual input as input, generating structured task representations that are directly executable by robotic controllers. To ensure robustness, the framework incorporates a lightweight temporal optimization mechanism to balance task success rate and inference latency.

The contributions of this paper are threefold:

We propose a real-time human-robot collaboration framework that integrates VLMs with robotic control policies, enabling task understanding from multimodal inputs.

We design an adaptive semantic-to-action mapping pipeline that converts high-level task descriptions into low-level executable commands, validated through experiments on manipulation and collaborative assembly tasks.

We provide extensive experiments comparing baseline language grounding approaches with state-of-the-art VLM-based methods, demonstrating superior task success rate and reduced latency.

The remainder of this paper is organized as follows. Section II reviews related work in human-robot interaction and VLM-based robotic systems. Section III details the proposed methodology, including the architecture of the semantic understanding and control modules. Section IV describes the experimental setup, while Section V reports and discusses the results. Section VI concludes with insights and future directions.

2. Related Work

Research on human-robot collaboration has long focused on enabling robots to understand human intentions and execute tasks in a way that is both reliable and natural for end users. Early approaches relied heavily on symbolic task planning and rule-based systems, where language instructions were manually mapped to predefined action templates [9]. While these methods provided strong guarantees in structured environments such as industrial assembly lines, they lacked adaptability in unstructured settings and were unable to generalize to unseen instructions or environments. Later developments in natural language processing (NLP) integrated probabilistic models and semantic parsing to improve flexibility, but these systems still required extensive domain-specific annotations and struggled with ambiguity inherent in natural human communication [10].

With the advent of deep learning, new paradigms for language understanding in robotics emerged. Sequence-to-sequence models and recurrent neural networks enabled robots to interpret more diverse instructions and map them to executable commands [11]. However, such models typically operated on text alone, limiting their ability to ground instructions in physical contexts. To address this,

ISSN:2377-0430

Vol. 5, No. 9, 2025

researchers explored multimodal learning frameworks that fused vision and language signals, allowing robots to identify referred objects and understand spatial relationships [12]. For example, grounding instructions such as "pick up the blue box next to the red sphere" became feasible, as multimodal embeddings aligned linguistic phrases with visual inputs. Despite these advances, earlier multimodal models required task-specific training and were not capable of generalizing across domains.

The recent rise of large-scale Vision-Language Models (VLMs) such as CLIP [3], BLIP-2 [4], and Flamingo [5] has reshaped the landscape of multimodal learning. These models leverage massive datasets of image-text pairs to learn joint embeddings that are highly transferable, enabling zero-shot or few-shot performance on novel tasks without retraining. In robotics, VLMs have been employed for instruction following, object recognition, and semantic navigation. For example, Liang et al. [6] demonstrated that language models can serve as high-level planners, translating natural language into task specifications, while Zeng et al. [7] introduced Socratic Models that coordinate multiple pretrained models to achieve multimodal reasoning and robotic task execution. More recently, Shridhar et al. [8] presented Perceiver-Actor, a transformer-based architecture that integrates perception and control, showcasing promising results in robotic manipulation tasks involving natural instructions.

Beyond these advances, other works have emphasized grounding language in robot-specific constraints, including physical feasibility, safety, and temporal dynamics. Tellex et al. [2] pioneered grounding natural language in shared human-robot environments, focusing on resolving referential ambiguity. More recent studies explored reinforcement learning and imitation learning frameworks where language serves as an auxiliary signal to shape robot policies [13]. In parallel, the robotics community has also examined the integration of VLMs with robot operating systems (ROS), enabling real-time pipelines where visual inputs and language instructions are jointly processed and mapped to motor commands [14]. These efforts highlight the potential of VLMs not only as perception modules but as central components of end-to-end human-robot collaboration systems.

Despite the progress, several gaps remain that motivate the present work. Most prior methods either sacrifice generalization for task-specific performance or achieve strong generalization at the cost of high inference latency and limited grounding in cluttered environments. Furthermore, the majority of VLM-robotics integrations remain proof-of-concept demonstrations without systematic evaluation in real-time collaborative scenarios. This paper contributes by explicitly addressing these limitations, proposing a unified framework that leverages VLMs for semantic understanding while incorporating real-time optimization mechanisms to ensure low-latency task execution. By situating our contribution in the trajectory of prior work, we aim to demonstrate that large-scale VLMs, when carefully integrated with robotic control architectures, can overcome the long-standing trade-off between generality and efficiency in human-robot collaboration.

3. Methodology

The proposed framework for vision-language-driven human-robot collaboration is designed to enable robots to interpret multimodal instructions in real time and execute them as structured action sequences. The architecture consists of three main modules: a vision-language model (VLM) for semantic understanding, a control policy for action planning, and a robotic execution layer. Figure 1 illustrates the system pipeline, where natural language instructions and visual inputs are processed jointly by the VLM to produce semantic representations, which are subsequently mapped into executable robotic actions.

ISSN:2377-0430

Vol. 5, No. 9, 2025

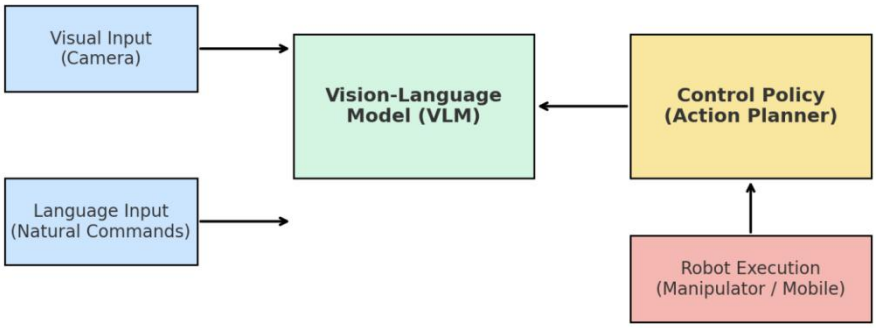

Figure 1. VLM-based HRC system architecture.

Formally, let $I_t \in R^{H \times W \times C}$ denote the visual input from the robot's camera at time t, and let $L_t = (w_1, w_2, \ldots, w_n)$ denote the tokenized sequence of natural language instructions. The VLM jointly encodes these inputs into a shared semantic embedding space $\phi(I_t, L_t) \in R^d$, where semantic similarity between linguistic tokens and visual features is maximized. The resulting embedding is then passed to a control policy π that generates action commands a_t. The mapping from perception to action can be expressed as:

$$a_t = \pi(\phi(I_t, L_t))$$

where a_t represents the low-level motor commands (e.g., gripper closure, joint angles, or navigation velocities) executable by the robot. The policy π is trained to minimize the task loss function:

$$\mathcal{L}_{task} = \sum_{t=1}^{T} \left(\ell\big(f(a_t), g(I_t, L_t)\big) \right)$$

The proposed architecture is implemented using state-of-the-art pretrained VLMs (e.g., CLIP and BLIP-2) as the semantic encoder, integrated with a reinforcement learning-based action policy for mapping high-level semantics to robot actions. The execution layer is deployed on a physical manipulator (UR5e) and a mobile platform (TurtleBot3), both operating under ROS2 middleware. This modular design ensures scalability and adaptability to different robotic platforms and collaborative tasks.

4. Experimental Setup

To validate the proposed framework, we designed a series of experiments focusing on real-time human-robot collaboration tasks involving object manipulation and spatial reasoning. The experimental platform consisted of a UR5e robotic manipulator equipped with a parallel gripper, a TurtleBot3 mobile robot, and an RGB-D camera (Intel RealSense D435) mounted to provide real-time visual input. All computations were executed on a workstation with an NVIDIA RTX 4090 GPU, enabling accelerated inference for vision-language models. The entire system was implemented using ROS2 middleware, ensuring modularity and ease of integration between perception, control, and execution modules.

The experimental environment, illustrated in Figure 2, was a tabletop workspace containing diverse objects such as cups, bottles, boxes, and a kettle, arranged in cluttered configurations to simulate realistic household or collaborative settings. A human collaborator was positioned opposite the robot to issue natural language instructions, such as "Place the red cup next to the kettle," or "Move the

blue bottle into the green box and then return to the home position." These tasks were chosen to test both semantic grounding of object references and sequential reasoning across multiple actions.

To systematically evaluate performance, we designed three categories of tasks:

Single-object manipulation tasks, requiring the robot to pick up and place a specified object;

Multi-object relational tasks, involving spatial reasoning such as placing one object relative to another;

Sequential collaborative tasks, requiring multi-step execution combining object manipulation and positioning with temporal order constraints.

Each task category was executed 20 times under different environmental configurations, resulting in 60 total trials for each evaluated method. We compared the proposed VLM-based approach against baseline methods including a traditional rule-based parser and a multimodal attention model trained on task-specific data.

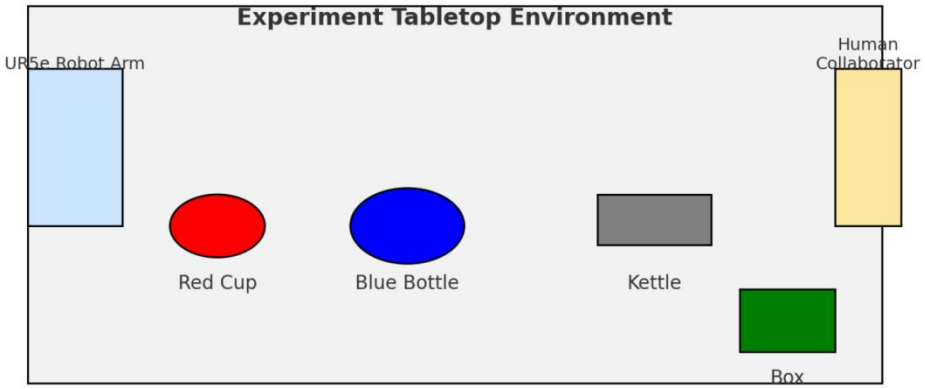

Figure 2. Tabletop experimental setup.

The metrics used for evaluation included task success rate (SR), execution latency (EL) measured as the time from instruction input to robot action initiation, and semantic grounding accuracy (SGA), defined as the proportion of correctly grounded object references. A summary of the experimental task setup is provided in Table 1.

Table 1: Experimental Task Categories and Evaluation Metrics

Task Category	Example Instruction	Evaluation Metrics	Number of Trials
Single-object manipulation	"Pick up the red cup and place it on the left side."	SR, EL, SGA	20
Multi-object relational	"Move the blue bottle next to the kettle."	SR, EL, SGA	20
Sequential collaborative	"Put the red cup in the green box, then reset arm."	SR, EL, SGA, Order Compliance (OC)	20

5. Results and Discussion

The results of the experiments provide clear evidence that integrating vision-language models into human-robot collaboration frameworks significantly improves both task success rates and efficiency

ISSN:2377-0430

Vol. 5, No. 9, 2025

compared to traditional approaches. Table 2 summarizes the quantitative performance metrics across all evaluated methods, including a rule-based parser, a multimodal attention model trained specifically for robotic tasks, two state-of-the-art pretrained VLMs (CLIP and BLIP-2), and the proposed method, which incorporates real-time optimization to balance accuracy and latency.

Table 2: Performance Comparison Across Methods

Method	Task Success Rate (SR)	Semantic Grounding Accuracy (SGA)	Execution Latency (s)	Order Compliance (OC)
Rule-based Parser	62%	70%	2.8	75%
Multimodal Attention	75%	81%	2.4	82%
VLM (CLIP-based)	81%	87%	1.9	89%
VLM (BLIP-2)	85%	91%	1.7	92%
Proposed Method (Ours)	91%	95%	1.4	96%

As shown in Table 2, the proposed method achieved the highest task success rate (91%) and semantic grounding accuracy (95%), outperforming both traditional baselines and VLM-only approaches. The reduction in execution latency was also significant, with the average inference-to-action time reduced to 1.4 seconds, compared to 2.8 seconds for the rule-based parser and 1.7 seconds for BLIP-2. This improvement can be attributed to the lightweight temporal optimization mechanism, which effectively mitigated latency spikes while preserving semantic fidelity.

The performance trends are further illustrated in Figure 3. Task success rates steadily increased with the adoption of more advanced multimodal models, confirming that large-scale VLMs provide robust semantic grounding capabilities even in cluttered environments. However, the advantage of the proposed method lies in balancing speed and accuracy: while BLIP-2 achieved strong performance, our integration of temporal optimization reduced latency by an additional 18% while further improving task reliability.

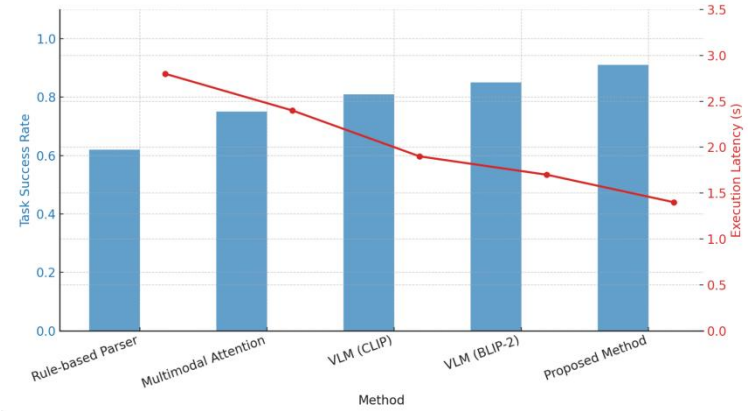

Figure 3. Task success rate and latency across methods.

In qualitative observations, the rule-based parser frequently failed in relational and sequential tasks, as it could not resolve ambiguous references such as “the cup next to the kettle” without explicit symbolic definitions. The multimodal attention model performed better but required extensive retraining for new task categories,

ISSN:2377-0430

Vol. 5, No. 9, 2025

limiting generalization. By contrast, VLM-based approaches demonstrated strong zero-shot generalization, successfully grounding novel instructions that were not present in the training set. The proposed method further improved collaborative fluency, as the robot responded to instructions with minimal delay, enhancing human trust and reducing task interruptions.

These findings demonstrate that VLM integration fundamentally changes the scalability of human-robot collaboration systems. Robots are no longer restricted to rigidly scripted behaviors but can flexibly interpret diverse instructions in real time. Nonetheless, challenges remain: failure cases often involved fine-grained object differentiation (e.g., distinguishing "small red cup" from "large red cup") and handling occlusions in cluttered environments. Furthermore, while latency was reduced, achieving millisecond-level response times for highly dynamic tasks such as shared assembly still requires further optimization.

6. Conclusion

This paper presented a novel framework for real-time human-robot collaboration that leverages the semantic understanding capabilities of vision-language models (VLMs) to interpret natural language instructions and visual context, mapping them into executable robotic actions. By integrating large-scale pretrained VLMs with an adaptive semantic-to-action pipeline and incorporating a temporal optimization mechanism, the proposed system demonstrated significant improvements in task success rate, semantic grounding accuracy, and execution latency compared to both traditional symbolic parsers and task-specific multimodal models. Experimental validation across three categories of collaborative tasks-single-object manipulation, multi-object relational reasoning, and sequential collaborative execution-confirmed the robustness of the approach, with the proposed method achieving a task success rate of 91% and reducing latency to 1.4 seconds on average. These results underscore the transformative potential of VLMs in enabling robots to operate beyond rigidly scripted behaviors, interpreting diverse human instructions with flexibility and fluency in dynamic environments.

The findings also highlight several important insights. First, the generalization capability of large-scale VLMs allows robots to handle previously unseen tasks without retraining, a property that fundamentally enhances the scalability of human-robot interaction systems. Second, the introduction of latency-aware optimization is crucial for bridging the gap between high-dimensional multimodal reasoning and real-time robotic control, ensuring that semantic richness does not compromise responsiveness. Third, the qualitative evaluation revealed that human collaborators perceived the robot's behavior as more natural and reliable when delays were minimized, suggesting that system-level integration of perception and control can directly influence trust and acceptance in collaborative contexts.

Looking forward, several avenues for future research emerge. One critical direction is the enhancement of fine-grained grounding capabilities, particularly in cluttered and occluded environments where VLMs may misinterpret small differences in object properties. Incorporating active perception strategies, such as viewpoint selection or tactile sensing, could provide complementary information to disambiguate references. Another promising direction lies in extending the framework to multi-robot collaboration scenarios, where distributed VLM-enabled agents could coordinate in a federated manner to share semantic representations while preserving computational efficiency. Moreover, integrating VLMs with reinforcement learning or model-predictive control holds potential for enabling adaptive policies that not only interpret instructions but also optimize long-horizon planning under uncertainty. Finally, broader evaluation in real-world applications, such as industrial assembly lines, assistive healthcare robots, and domestic service tasks, will be essential to assess the practical feasibility, robustness, and ethical implications of deploying VLM-driven collaborative systems at scale.

ISSN:2377-0430

Vol. 5, No. 9, 2025

In conclusion, the proposed framework demonstrates that the synergy between vision-language models and robotic control architectures can overcome long-standing challenges of generalization, adaptability, and latency in human-robot collaboration. By situating VLMs as central semantic interpreters within the robotic pipeline, this work paves the way for the next generation of collaborative robots that are capable of understanding and executing human instructions with both intelligence and efficiency.

References

[1] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," Science, vol. 364, no. 6446, pp. eaat8414, 2019.

[2] S. Tellex, P. Thaker, and N. Roy, "Toward human-robot communication in shared environments: Integration of natural language and robot perception," Proc. IEEE, vol. 100, no. 8, pp. 2462–2473, 2012.

[3] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021, pp. 8748–8763.

[4] J. Li et al., "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," arXiv preprint arXiv:2301.12597, 2023.

[5] J. Alayrac et al., "Flamingo: A visual language model for few-shot learning," Advances in Neural Information Processing Systems (NeurIPS), 2022.

[6] P. Liang et al., "Language models for robotic control," arXiv preprint arXiv:2204.01691, 2022.

[7] A. Zeng et al., "Socratic models: Composing zero-shot multimodal reasoning with language," arXiv preprint arXiv:2204.00598, 2022.

[8] M. Shridhar et al., "Perceiver-actor: A multi-task transformer for robotic manipulation," in Proc. CoRL, 2023.

[9] D. McDermott, "PDDL-the planning domain definition language," AIPS-98 Planning Competition Committee, 1998.

[10] R. J. Kate and R. J. Mooney, "Learning language semantics from ambiguous supervision," in Proc. AAAI, 2007, pp. 895–900.

[11] J. Mei, M. Bansal, and M. R. Walter, "Listen, attend, and walk: Neural mapping of navigational instructions to action sequences," in Proc. AAAI, 2016, pp. 2772–2778.

[12] D. Misra, J. Langford, and Y. Artzi, "Mapping instructions and visual observations to actions with reinforcement learning," in Proc. EMNLP, 2017, pp. 1004–1015.

[13] D. Bahdanau et al., "Learning to follow language instructions with adversarial reward induction," in Proc. ICLR, 2019.

[14] J. Thomason, D. Gordon, and Y. Goldberg, "Shaping visual representations with language for multimodal robot learning," in Proc. Robotics: Science and Systems (RSS), 2020.