Perception to Action with Vision-Language-Action Models for Fast and Reliable Decision Making in Dynamic Environments

Yeryeong Cho, Jaeyoung Choe, and Joongheon Kim
Department of Electrical and Computer Engineering, Korea University, Seoul, Republic of Korea
E-mails: {joyena0909,joongheon}@korea.ac.kr

***Abstract*—Executing natural-language instructions in dynamic environments requires robust integration of perception, planning, and control. Conventional vision–language models (VLMs) provide open-vocabulary recognition but often leave a perception–action gap and fail to ensure safety under uncertainty. To address this challenge, we propose a Vision-Language Action Model (VLAM) that directly maps visual observations and instructions to actions through adaptive perception and uncertainty-aware control. The algorithm leverages contrastive language-image pretraining (CLIP)-based vision–language similarity to score candidate actions while incorporating rule-based safety priors as a fallback mechanism when confidence is low or collision risk is detected. This design narrows the perception–action gap, maintains semantic grounding, and guarantees stable behavior. We evaluate VLAM in a dynamic multi-agent environment with moving obstacles. Experimental results demonstrate that the proposed method achieves higher task success and reduced inter-agent conflicts compared to baseline strategies.**

***Index Terms*—Vision-Language Model (VLM), Vision-Language Action Model (VLAM), Adaptive Perception, Dynamic Environment**

I. Introduction

Acting on natural language in dynamic environments requires an integrated loop of perception, planning, and control [1]. Moving obstacles, multi-object interactions, and goal switching make static perception insufficient [2]. Therefore, real-time and safe decisions are difficult without decision-time updates [3]. This paper considers instruction-conditioned control in which language and visual observations evolve while the policy must output actions that satisfy safety constraints. At first, adaptive perception updates relations and risks at every timestep using vision–language similarity to maintain decision-time grounding [4]. Second, an uncertainty-aware hybrid policy falls back to a rule-based safety prior when similarity confidence is low or collision risk is detected [5]. Therefore, this architecture yields conservative yet robust behavior in crowded, dynamic scenes [6]. The proposed vision-language action model (VLAM) directly maps visual observations and language commands to actions. This algorithm narrows the perception–action gap while remaining stable under abrupt instruction changes such as goal switching [7]. Furthermore, the model integrates open-vocabulary scene grounding with uncertainty gating and a safety fallback so that action selection adapts as the environment evolves [8]. Therefore, the contributions of this paper are as follows:

- We propose a vision–language action architecture that maps observations and instructions to actions in dynamic scenes. The implementation uses discrete grid control, stable matching, and risk-aware scoring.
- An uncertainty-aware hybrid strategy that combines vision–language similarity with a rule safety prior.
- A realistic evaluation is conducted. Success and conflicts jointly assess safety and efficiency.

These contributions highlight the central goal of this paper: to demonstrate that integrating adaptive perception with uncertainty-aware control leads to safer and more reliable instruction following. The following sections present the theoretical background, algorithmic design, and experimental evaluation that substantiate these claims.

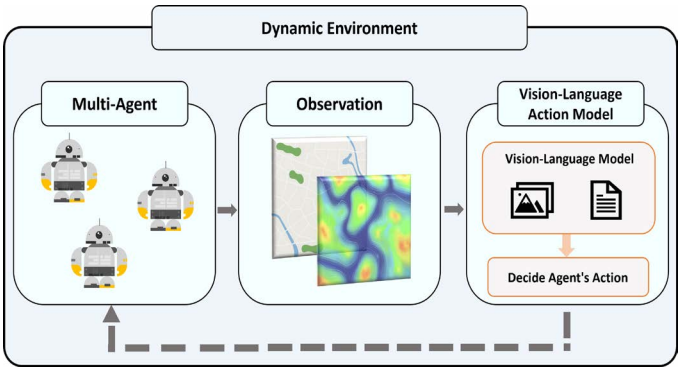

Fig. 1: Proposed VLAM architecture.

II. Background and Motivation

Vision–language models (VLMs) have enabled open-vocabulary recognition and instruction following by aligning image and text embeddings at scale [9]. These models allow task grounding without task-specific retraining, which is appealing for robotics and embodied agents operating outside curated datasets [10]. However, a perception-to-action gap often persists [11]. Many systems compute static similarity scores or one-shot plans from a single prompt, then rely on open-loop execution [12]. When the environment is dynamic, this

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2025-00561377). *(Corresponding author: Joongheon Kim)*

creates latency and brittleness [13]. Furthermore, the scene state changes between perception and actuation, degrading task grounding and safety [14]. Safety and efficiency further deteriorate under uncertainty [15]. Visual occlusions, rapidly moving obstacles, and multi-agent interactions create partial observability and distribution shift [16]. In these conditions, models output overconfident yet unsafe actions, unless there is a mechanism to recognize low confidence and to defer to conservative priors [17]. Therefore, the proposed algorithm integrates contrastive language-image pretraining (CLIP)-based vision–language similarity to achieve stable instruction following in dynamic environments.

III. Algorithm

The algorithm maps instructions and visual observations to actions. It operates in a closed loop at each timestep. Furthermore, it maintains task grounding and safety by scoring candidate actions based on vision–language similarity and simple priors [18]. Both the visual and text encoders are provided by CLIP. CLIP is particularly suitable in this context because it is trained on large-scale image–text pairs and thus supports open-vocabulary grounding [19]. This capability enables the agent to generalize to unseen instructions and visual variations without task-specific retraining [20]. Each agent evaluates five candidate actions: up, down, left, right, and stay. For each action, an image patch is cropped around the candidate position, and its embedding is compared with label embeddings using similarity. The value of a candidate action is defined as

$$V(a) = s_{\text{img}}(l_{\text{target}}) - \lambda_o \, s_{\text{img}}(l_{\text{obstacle}}) - \lambda_d \, d(a). \quad (1)$$

In 1, s_{img} is the cosine similarity between the patch and a text label, l_{target} is the current task label, l_{obstacle} is the obstacle label. Furthermore, $d(a)$ is the Manhattan distance from the candidate position to the target, and λ is a weighting coefficient that scales the penalty terms for obstacle similarity and distance cost. When confidence is below a threshold or collision risk is detected, the policy switches to a rule-based safety controller. The rule controller enforces collision avoidance, goal-directed progress, and movement stability. When confidence is below a threshold or collision risk is detected, the policy switches to a rule-based safety controller. The combination of CLIP-based perception and rule-based fallback provides two complementary benefits. CLIP enables open-vocabulary grounding of instructions, which allows the agent to adapt flexibly to different goals and objects [21]. At the same time, rule-based control ensures adherence to stringent safety constraints, guaranteeing that the system does not produce unsafe or unstable actions even when perceptual confidence is low [22]. This hybrid design reduces variance in performance and provides a safety margin against unpredictable environmental changes. Through the integration of semantic grounding and structured priors, the algorithm achieves both generalization and stability.

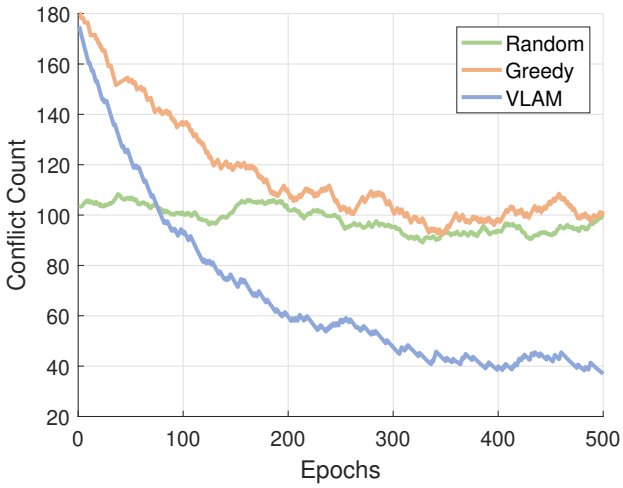

Fig. 2: Comparison of conflict count.

Fig. 3: Comparison of success rate and step efficiency.

IV. Performance Evaluation

A. Evaluation Setup

The dynamic multi-agent environment consists of agents tasked with interpreting natural language instructions and adapting their behavior accordingly. The agents must transport objects to the designated goal while avoiding dynamic obstacles and inter-agent conflicts. Additionally, the scenario incorporates instruction switches that occur mid-episode, compelling agents to re-align their actions with newly assigned goals. This configuration offers a rigorous evaluation of both semantic grounding and safety under uncertainty. In addition, the proposed VLAM is compared against baseline strategies, including Random and Greedy algorithms. This setup allows a quantitative assessment of whether adaptive perception and safety-aware control improve over naive exploration and distance-based heuristics.

B. Evaluation Results

In Fig. 2, VLAM significantly lowers conflict counts compared to both the Random and Greedy algorithms. This reduction validates the effectiveness of uncertainty-aware gating, which switches to conservative fallback behaviors when

risks emerge. Furthermore, VLAM not only reduces the average number of conflicts but also minimizes variance across episodes. Consequently, it yields more consistent performance under dynamic conditions. Fig. 3 further indicates that VLAM improves both success rate and step efficiency. The Random algorithm shows a low success rate, while the Greedy algorithm maintains a moderate success rate initially but suffers from unstable increases in steps. However, VLAM consistently maintains high success rates with fewer steps. Therefore, it demonstrates that its task grounding and safety priors jointly support efficient and reliable decision making. Consequently, Fig. 2 and Fig. 3 clearly show that VLAM surpasses all baseline algorithms in both conflict avoidance and task success. These results suggest that VLAM generates behaviors that are not only safer but also more generalizable than the other algorithms.

V. Future Work and Conclusion

This paper proposes VLAM, which directly maps natural-language instructions to actions in dynamic multi-agent environments. By integrating adaptive perception with uncertainty-aware control, the proposed algorithm effectively narrows the perception–action gap. Furthermore, experimental results demonstrate that VLAM achieves both high success rates and low conflict counts to maintain efficiency and safety under dynamic conditions. Nevertheless, several limitations remain. The current evaluation is constrained to a discrete grid environment, and future work should extend VLAM to continuous control domains. Furthermore, the reliance on CLIP for similarity scoring has computational overhead that limits real-time applications. Future research should consider lightweight multimodal encoders or online adaptive embedding techniques. In conclusion, this paper demonstrates that VLAM improves perception–action integration and achieves a balance between safety and efficiency in dynamic environments. Future research directions include extending the approach to continuous domains, improving computational efficiency in multimodal representation. These developments will broaden the practical applicability of VLAM, and they enable real-time deployment in autonomous driving and robotic control.

References

[1] e. Anthony Brohan, "RT-2: vision-language-action models transfer web knowledge to robotic control," in *Proc. Conference on Robot Learning (CORL)*, Atlanta, USA, July 2023, pp. 2165–2183.

[2] P. Fiorini and Z. Shiller, "Motion planning in dynamic environments using velocity obstacles," *The International Journal of Robotics Research*, vol. 17, no. 7, pp. 760–772, July 1998.

[3] F. Chaumette and S. Hutchinson, "Visual servo control. i. basic approaches," *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, December 2006.

[4] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, A. Handa, T. Lin, G. Wetzstein, M. Y. Liu, and D. Xiang, "CoT-VLA: Visual chain-of-thought reasoning for vision-language-action models," in *Proc. IEEE Computer Vision and Pattern Recognition, (CVPR)*, Nashville, USA, June 2025, pp. 1702–1713.

[5] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. H. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 7, pp. 2737–2752, October 2019.

[6] R. Sinha, E. Schmerling, and M. Pavone, "Closing the loop on runtime monitors with fallback-safe MPC," in *Proc. IEEE Conference on Decision and Control (CDC)*, Singapore, December 2023, pp. 6533–6540.

[7] J. Qian, S. Zhou, N. J. Ren, V. Chatrath, and A. P. Schoellig, "Closing the perception-action loop for semantically safe navigation in semi-static environments," in *Proc. International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, May 2024, pp. 11 641–11 648.

[8] e. Allen Z. Ren, "Robots that ask for help: Uncertainty alignment for large language model planners," in *Proc. Conference on Robot Learning (CoRL)*, Atlanta, USA, November 2023, pp. 661–682.

[9] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5625–5644, February 2024.

[10] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, S. Kirmani, B. Zitkovich, F. Xia, C. Finn, and K. Hausman, "Open-world object manipulation using pre-trained vision-language models," in *Proc. Conference on Robot Learning (CoRL)*, Atlanta, USA, November 2023, pp. 3397–3417.

[11] N. Kim, W. Na, D. S. Lakew, N.-N. Dao, and S. Cho, "DQN-based directional MAC protocol in wireless Ad Hoc network in internet of things," *IEEE Internet of Things Journal*, vol. 11, no. 7, pp. 12 918–12 928, December 2024.

[12] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar *et al.*, "RT-1: robotics transformer for real-world control at scale," in *Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023, pp. 1–22.

[13] D. Kappler, F. Meier, J. Issac, J. Mainprice, C. G. Cifuentes, M. Wuthrich, V. Berenz, S. Schaal, N. D. Ratliff, and J. Bohg, "Real-time perception meets reactive motion generation," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1864–1871, January 2018.

[14] T. G. Molnar, A. K. Kiss, A. D. Ames, and G. Orosz, "Safety-critical control with input delay in dynamic environment," *IEEE Transactions on Control Systems Technology*, vol. 31, no. 4, pp. 1507–1520, December 2023.

[15] M. C. Hoa, A. T. Trana, D. Leea, J. Paeka, W. Nohb, and S. Cho, "A DDPG-based energy efficient federated learning algorithm with SWIPT and MC-NOMA," *ICT Express*, vol. 10, no. 3, pp. 600–607, June 2024.

[16] T. T. H. Pham, W. Noh, and S. Cho, "Multi-agent reinforcement learning based optimal energy sensing threshold control in distributed cognitive radio networks with directional antenna," *ICT Express*, vol. 10, no. 3, pp. 472–478, June 2024.

[17] B. Lötjens, M. Everett, and J. P. How, "Safe reinforcement learning with model uncertainty estimates," in *Proc. International Conference on Intelligent Robots and Systems (IROS)*, Vancouver, Canada, October 2017, pp. 1343–1350.

[18] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, and D. Manocha, "VLM-Social-Nav: Socially aware robot navigation through scoring using vision-language models," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 508–515, January 2025.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. The International Conference on Machine Learning (ICML)*, Virtual, July 2021, pp. 8748–8763.

[20] H. Jiang and Z. Lu, "Visual grounding for object-level generalization in reinforcement learning," in *Proc. European Conference on Computer Vision (ECCV)*, Milan, Italy, September 2024, pp. 55–72.

[21] A. Khandelwal, L. Weihs, R. Mottaghi, and A. Kembhavi, "Simple but effective: CLIP embeddings for embodied AI," in *Proc. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, LA, USA, June 2022, pp. 14 809–14 818.

[22] U. P. S. Johann Thor Ingibergsson Mogensen, Dirk Kraft, "Declarative rule-based safety for robotic perception systems," *Journal of Software Engineering for Robotic*, vol. 8, no. 1, pp. 17–31, December 2017.