LLM 및 VLM 기반 매니퓰레이터 제어 정책 연구 동향

A Survey of Manipulator Control Policies Based on LLM and VLM

이재찬[1] · 김태주[2] · 최유경 †

Jaechan Lee[1], Taejoo Kim[2], Yukyung Choi †

Abstract: This paper reviews recent research trends in robotic manipulation control policy generation using large language models (LLMs) and vision-language models (VLMs). It categorizes three major approaches proposed to address the limitations of traditional learning-based control methods. First, generating hierarchical control policies based on motion primitives through LLMs; second, the integrated use of visual and linguistic information through VLMs; and third, enhancing performance by incorporating human feedback. The aim of this paper is to provide a comprehensive overview of these approaches and to introduce recent advancements in robotic control systems.

Keywords: Robotic Manipulation Control Policy Generation, Large Language Models, Vision-Language Models

1. 서 론

GPT[1]와 같은 대규모 언어 모델(LLM)[2]과 CLIP[3]과 같은 비전-언어 모델(VLM)[4]의 능력은 최근 인공지능 연구 전반에 걸쳐 괄목할 만한 영향력을 행사하고 있다. 특히 로보틱스 분야에서는 이를 활용한 제어 정책 생성 연구[5-6]가 활발하다. VLM은 시각적 인지와 언어적 이해를 통합해 로봇이 환경을 보다 효율적으로 이해하도록 돕고, LLM은 이를 바탕으로 고도화된 제어 명령과 작업 계획을 생성한다. 이러한 접근은 로봇이 복잡한 환경에 적응하고 유연하게 작업을 수행할 수 있도록 돕는다.

기존 학습 기반 로봇 제어 방식은 일반화가 어렵고, 새로운 환경에 적응하기 힘들다는 한계가 있다. 강화학습[7-9]은 방대한 시도와 오류를 필요로 하고, 환경 변화에 민감하다. 모방학습[10-12]도 학습 데이터와 특정 시나리오에 의존해 새로운 상황에서 성능이 제한된다. 또한, 이러한 방식들은 높은 데이터 수집 비용과 학습 시간이 요구된다.

LLM은 이러한 문제를 해결할 수 있는 가능성을 제시한다. 추가학습 없이도 LLM의 잠재적 지식을 활용해 원하는 출력을 유도하며 비용과 시간 효율성을 보이는 프롬프팅 기반의 접근법[13]이 그 근거가 된다. 특히 복잡한 추론 문제에서 논리 구조 분해를 통해 단계적 추론을 가능하게 하는 기법[14-15]의 연구가 대두되면서 LLM의 추론 능력은 한 차례 더 고도화된다.

이에 더해 VLM의 사용은 로봇의 시각-언어적 이해 능력을 크게 확장한다. 즉 로봇은 단순한 객체 인식을 넘어 장면 전체의 구조와 의미를 파악하고, 상황에 적합한 행동을 생성할 수 있다[16-18]. 이는 특히 복잡한 작업 환경에서 로봇이 적절히 대응하고 자율적으로 작업을 수행하는 데 중요한 역할을 한다.

이러한 배경을 바탕으로, 본 논문에서는 LLM과 VLM을 활용한 로봇 제어 정책 생성의 최신 연구 동향을 다음과 같이 세 가지 범주로 정리하여 소개하고자 한다. 첫째, LLM을 통해 동작 단위 기반의 계층적인 제어 정책을 생성하는 방식, 둘째, VLM을 통해 시각 및 언어 정보를 통합하는 방식, 셋째, 인간 피드백을 결합해 정책을 강화하는 방식이다.

2. 매니퓰레이션 제어 정책 생성 연구 동향

매니퓰레이션 제어 정책 생성 분야에서는 최근 LLM과 VLM을 기반으로 한 다양한 접근들이 빠르게

※ This work was supported by the Technology Innovation Program(RS-2024-00442513, Development of robotic manipulation task learning based on Foundation model to understand and reason about task situations) funded by the Ministry of Trade Industry & Energy(MOTIE, Korea)

1. Undergraduate Student, Department of Intelligent Mechatronics Engineering in Sejong University, Seoul, Korea (jclee@rcv.sejong.ac.kr)
2. Graduate Student, Department of Intelligent Mechatronics Engineering in Sejong University, Seoul, Korea (tjkim@rcv.sejong.ac.kr)
† Associate Professor, Corresponding author: Department of Intelligent Mechatronics Engineering in Sejong University, Seoul, Korea (ykchoi@rcv.sejong.ac.kr)

발전하고 있다. 이로써 강화 학습 및 지도 학습 기반의 기존 제어 방식이 가진 방대한 양의 학습 데이터 제약을 극복하고, 더 유연하고 상황에 적응적인 매니퓰레이션을 가능하게 한다. 본 절에서는 매니퓰레이터 제어 정책 생성에 관한 최신 연구들을 세 가지 주요 접근 방식으로 분류하여 설명한다.

2.1 LLM을 활용한 단위 동작 기반 계층적 제어 정책

첫 번째 접근 방식은 CoT(Chain of Thought)[14] 기법 기반의 코드 생성형 프롬프팅 방법론[15,19]을 활용한다. 이 방식은 LLM에게 작업 지시, 상황 규제, 그리고 제어 코드의 예제 프롬프트를 제공하여, 이를 바탕으로 새로운 작업에 적합한 제어 코드를 생성하도록 유도[6, 20]한다.

즉, 고수준의 복잡한 동작 계획을 계층적으로 하위 단계로 분해하고, 분해된 세부 동작 계획을 사전 정의된 기본 단위 동작 API와 통합하여 제어 코드 정책을 수립하는 CaP(Code as Policies) 방식[6, 20]을 사용하는 것이다. 이를 통해 로봇은 작업 명령에 따른 필요한 동작을 수행한다.

해당 접근법은 LLM의 추론 능력을 활용해 생성된 제어 정책 코드가 시뮬레이션 및 실제 로봇 환경에서 실행 가능함을 입증했다. 이를 통해 대규모 데이터를 사용한 사전학습 없이도 로봇이 적응력 있는 작업 동작 수행이 가능함을 보인다.

2.2 VLM을 활용한 시각-언어간 맥락 매칭

두 번째 접근 방식은 첫 번째 제어 정책 방식의 계승을 전제로 한다. VLM을 통해 시각 정보와 언어 정보를 통합하여 장면 중심 혹은 객체 중심의 의미론적 작업 추론을 수행하는 기법이다.

기존 LLM 기반 제어 정책에서 생성된 작업 계획 및 이해에 필요한 언어 정보를 VLM을 활용해 시각적 정보와 결합하고, 이를 작업 추론의 근거로 함께 사용하며 실제 세계에 대한 시각-언어간 맥락 매칭(Visual Grounding)을 수행한다. 이는 사전 정의된 단위 동작 의존에 의해 생기는 작업 추론 및 계획 병목 현상을 해결한다. 본 논문에서는 해당 접근 방식을 장면 중심적 추론과 객체 중심적 추론 방법론으로 나눠 분류를 하고자 한다.

먼저, 장면 중심적 추론 방법론은 장면 내의 공간적, 시각적 특징을 활용해 복잡한 행동을 계획하고 실행한다. 특히, VoxPoser[21]는 LLM과 VLM을 결합하여 객체의 기능적 속성인 affordance와 장면에 대한 행동 규제를 바탕으로 3D 가치 지도를 구성하고, 주어진 명령어에 따라 3D 공간에서 작업 경로 판단 기반의 액션 정책을 수립한다.

객체 중심적 추론 방법론은 지시에 따른 객체의 affordance를 결정하고, 이에 따른 제어 정책 생성을 목적으로 한다. MOKA[22]는 visual mask를 통해 장면 내 객체를 인지하고, 마킹 포인트를 시각 프롬프트로 VLM에게 제공함으로써 객체의 affordance를 기반으로 합리적인 동작 추론을 가능하게 한다. OVAL-Prompt[23]는 장면 내 미학습 객체를 인지하고 작업 명령 내의 동작과 해당 객체 간의 affordance를 VLM을 통해 이해한 후 동작 제어 정책을 다루며, 로봇과 객체 간의 효과적인 상호작용을 가능하게 한다.

2.3 인간 피드백을 결합한 정책 신뢰성 강화

세 번째 접근 방식은 LLM 및 VLM을 활용한 앞선 두 방식을 기반으로, 인간의 피드백을 결합하여 정책의 신뢰성을 강화[20, 24-26]한다.

InnerMonologue[24]는 실제 로봇환경에서 작업의 중간 단계에서 실패할 때, 수동적/능동적 장면 묘사 피드백과 작업 성공 판단 피드백 등의 다양한 피드백 소스를 통해 closed-loop feedback 방식을 통해 미학습된 복잡한 작업 세팅에서도 계획 재구성이 가능하게 한다. HELPER[25]는 언어-프로그램 쌍을 외부 메모리로써 저장하고, 메모리 검색 증강을 통해 자유 형식의 인간-로봇 대화를 동작 프로그램으로 해석 후 행동한다. 작업 성공 시 이에 대한 피드백으로 작업 계획에 대한 범주를 더욱 확장할 수 있으며, 이는 사용자 개인화된 확장이 가능하다는 장점이 있다. HRC[26]는 인간의 원격 조작 시연을 피드백으로 도입한다. 기존 사전 정의된 기본 단위 동작 라이브러리와 인간에 의해 동적으로 재구성된 동작 함수를 통합하여 긴 호흡의 복잡한 작업에 대응한다.

이렇게 인간의 직관적 피드백을 통해 로봇의 작업 실패를 바로잡는 방식은, 로봇이 학습 데이터만으로는 얻기 어려운 맥락적 이해와 상식을 전이하는 데 큰 도움을 준다. 더불어 로봇이 실행 가능한 행동 계획을 지향하도록 정책 코드를 반복 개선하며, 넓은 행동 범주를 포괄하는 작업 수행능력을 갖추게 한다. 이를 통해 로봇의 강건성과 신뢰성을 한층 강화할 수 있다.

3. 결 론

본 논문에서는 LLM과 VLM을 활용한 로봇 매니퓰레

이션 제어 정책 생성의 최신 연구 동향을 세 가지 접근 방식으로 분류하였다. 이는 점차 태동하며 융합하고 고도화되는 최신 연구들 간의 분류 체계를 확립하는 데 기여할 수 있을 것으로 기대된다.

향후 연구에서는 실제 환경에서의 검증을 통해 로봇 작업의 안정성과 일반화를 더욱 강화하는 것이 필요하다. 이는 로봇, 인간, 세계를 정합시킬 수 있는 중요한 목표가 될 것으로 예상된다.

References

[1] OpenAI, "GPT-4 Technical Report", *arXiv preprint arXiv:2303.08774*, 2023.

[2] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A Survey of Large Language Models", *arXiv preprint arXiv:2303.18223*, 2023.

[3] 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", *International Conference on Machine Learning (ICML)*, Vienna, Austria, pp. 8748-8763, 2021.

[4] 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, Aug, 2024.

[5] Y. Kim, D. Kim, J. Choi, J. Park, N. Oh, and D. Park, "A Survey on Integration of Large Language Models with Intelligent Robots", *Intelligent Service Robotics*, Vol. 17, No. 5, pp. 1091-1107, Aug, 2024.

[6] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as Policies: Language Model Programs for Embodied Control," *IEEE International Conference on Robotics and Automation (ICRA)*, London, United Kingdom, pp. 9493-9500, 2023.

[7] Y. J. Ma, W. Liang, G. Wang, D. A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-Level Reward Design via Coding Large Language Models.", *The Twelfth International Conference on Learning Representations (ICLR)*, Vienna, Austria, 2024.

[8] Ichter B, Brohan A, Chebotar Y, Finn C, Hausman K, et al. "Do as I can, not as I say: grounding language in robotic affordances", *Proceedings of The 6th Conference on Robot Learning*, PMLR, Vol. 205, pp. 287-318, 2023.

[9] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman, "MT-Opt: Continuous multi-task robotic reinforcement learning at scale," *arXiv preprint arXiv:2104.08212*, 2021.

[10] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "BC-Z: Zero-shot task generalization with robotic imitation learning," *in Conference on Robot Learning (CoRL)*, pp. 991–1002, 2021.

[11] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K. H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "RT-1: Robotics transformer for real-world control at scale," *Robotics: Science and Systems (RSS)*, 2023.

[12] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T. W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "RT-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners", *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33, pp.1877–1901, 2020.

[14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models", *Advances in neural information processing systems (NeurIPS)*, Vol. 35, pp. 24824–24837, 2022c.

[15] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen, "Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks", *Transactions on Machine Learning Research (TMLR)*,

2023e.

[16] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation", *arXiv* preprint arXiv:2104.13921, 2021.

[17] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "MDETR-modulated detection for end-to-end multi-modal understanding," *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[18] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby. "Simple open-vocabulary object detection with vision transformers", *arXiv preprint arXiv: 2205.06230*, 2022.

[19] M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, C. Sutton, and A. Odena, "Show Your Work: Scratchpads for Inter- mediate Computation with Language Models.", *arXiv preprint arXiv:2112.00114*, 2021.

[20] M. G. Arenas, T. Xiao, S. Singh, V. Jain, A. Z. Ren, Q. Vuong, J. Varley, A. Herzog, I. Leal, S. Kirmani, D. Sadigh, V. Sindhwani, K. Rao, J. Liang, and A. Zeng, "How to prompt your robot: A promptbook for manipulation skills with code as policies", *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4340-4348, 2024.

[21] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models." *arXiv preprint arXiv:2307.05973*, 2023.

[22] F. Liu, K. Fang, P. Abbeel, and S. Levine, "Moka: Open-vocabulary robotic manipulation through mark-based visual prompting." *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA*, 2024.

[23] E. Tong, A. Opipari, S. Lewis, Z. Zeng, and O. C. Jenkins, "OVAL-Prompt: Open-Vocabulary Affordance Localization for Robot Manipulation through LLM Affordance-Grounding." *arXiv preprint arXiv:2404.11000*, 2024.

[24] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, P. Sermanet, N. Brown, T. Jackson, L. Luu, S. Levine, K. Hausman, and B. Ichter, "Inner monologue: Embodied reasoning through planning with language models", *6th Annual Conference on Robot Learning*, 2022.

[25] G. Sarch, Y. Wu, M. J. Tarr, and K. Fragkiadaki, "Open-ended instructable embodied agents with memory-augmented large language models." *arXiv preprint arXiv:2310.15127*, 2023.

[26] H. Liu, Y. Zhu, K. Kato, A. Tsukahara, I. Kondo, T. Aoyama, and Y. Hasegawa, "Enhancing the llm-based robot manipulation through human-robot collaboration," *arXiv preprint arXiv:2406.14097*, 2024.