LLM-VLM 기반 실시간 상황인지-행동 통합 제어 서비스 로봇 아키텍처

김지은, 조연후, 델가도 라이마리우스*
명지대학교 전자공학과

jieun0910@mju.ac.kr, dusgn8289@mju.ac.kr, *raim.delgado@mju.ac.kr

Real-Time LLM-VLM Architecture for Situation Awareness and Action Control in Service Robots

Jieun Kim, Yeonhu Cho, Raimarius Delgado*
Department of Electronics Engineering, Myongji University

요 약

본 논문은 고령화와 의료 인력 부족으로 복잡성이 증가하는 의료 환경에서 단일 작업에 한정된 기존 서비스 로봇의 한계를 극복하기 위한 실시간 컨텍스트 인지 기반 작업 스케줄링 프레임워크를 제안한다. 기존의 LLM 기반 작업 계획 및 VLM- LLM 융합 접근법은 인지 성능을 향상시켰으나, 실시간 피드백과 상황 변화에 따른 유연한 다중 도메인 전환을 충분히 지원하지 못하였다. 본 연구에서는 자연어 명령, 로봇 상태 등 다중 모달 인지 정보를 통합하는 LLM- VLM 파이프라인과 우선순위 기반 도메인 전환 구조를 설계하였다. 제안하는 프레임워크는 시스템의 안정성을 유지하면서 응급 상황에 대한 대응성을 향상시키고, 다양한 의료 서비스를 단일 로봇 플랫폼에서 효율적으로 수행할 수 있음을 보인다.

Ⅰ. 서 론

최근 헬스케어 환경은 고령화와 인력 부족 문제에 직면함에 따라 서비스 로봇 기술의 필요성이 그 어느 때보다 강조되고 있다. 기존의 서비스 로봇은 사전에 정의된 단일 작업 수행 및 각 목적에 따른 별도의 로봇 하드웨어와 소프트웨어를 요구하는 한계점을 가진다. 이는 개발 시간 및 유지보수의 부담을 가중시키며 다수의 로봇이 한정된 공간을 비효율적으로 점유하는 문제를 발생시킨다. 이러한 환경에서 서비스 로봇이 실제 임무를 수행하기 위해서는 상황과 맥락을 실시간으로 해석하고 적절한 작업을 선택하도록 지능형 스케줄링 방법론이 필수적이다.

이를 위하여 Kim et al. [1] 은 대규모 언어모델 (LLM)이 로봇 작업 순서를 결정하도록 사용자의 자연어 지시를 이해하고 이를 로봇 시스템에 연동하여 작업을 실행하도록 한다. 그러나 LLM을 이용해 언어적 지시만으로 상황을 인식하는 데는 한계가 있다. 이에 Shirai et al. [2]는 VLM과 LLM을 결합하여 시각적으로 확장된 상황인식 기반 로봇 작업 계획 방법을 제시하였다. 다만 실시간 피드백의 부재로 인한 동적 상황 대응 한계와 다중 도메인 전환 메커니즘이 마련되어 있지 않다는 문제가 존재한다.

따라서 본 논문에서는 실시간 상황인식 기반 태스크 스케줄링 프레임워크를 통한 다중 도메인의 동적 전환을 목표로 한다. 이를 위해 본 연구는 다음과 같은 접근을 시도한다. 자연어 명령과 로봇 상태를 통합하고 VLM, LLM 기반 상황인식 파이프라인을 구축한다. 또한 우선순위 기반 도메인 전환을 이용한 안정적인 아키텍처를 설계하여 개별 SW 모듈의 단일 시스템 통합을 진행하였다.

Ⅱ. 시스템 아키텍처

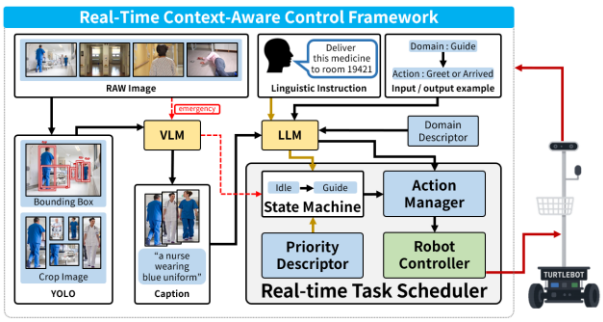
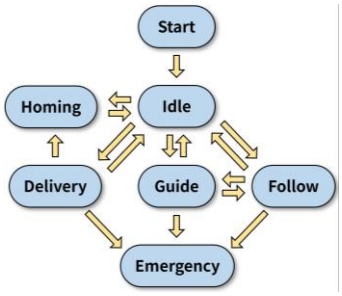


그림 1. 시스템 구조

본 시스템 아키텍처의 핵심 구조인 실시간 태스크 스케줄러는 우선 자연어와 우선순위 기술서를 기반으로 도메인을 결정한다. 이후 도메인이 결정되면 RAW 이미지에서 YOLO를 통해 객체를 탐지하고, 탐지된 객체를 VLM의 입력으로 넣어 현재 환경을 파악하고 이후 LLM을 통해 Action Manager에게 정보를 전달하여 최종 동작을 결정한다. 그 후 Robot Controller를 통해 실제 제어 신호를 로봇에게 전달한다. 이때, Emergency 도메인은 가장 높은 우선순위를 가지기 때문에 현재 도메인이나 동작 상태와 상관없이 즉시 상태 기계로 즉각 전달되어 해당 도메인으로 전환된다. 언어적 명령을 통해 고수준 판단인 도메인이 결정되고, 응급 상황 모니터링과 이미지 캡셔닝 (Captioning) 결과를 통해 인식된 action 단위의 저수준 동작을 제어하는 계층적 구조가 보다 안정적인 인식 및 판단을 결정한다.



순위	이벤트
1	Emergency
2	Idle
3	Follow Guide Delivery
4	Homing
5	Start

그림 2. 우선순위 기반 작업 스케쥴링 및 상태 기계

상황인식 모듈은 사용자의 자연어 명령을 해석하고, 이를 시스템의 도메인 지정이 가능한 형태의 명령어로 변환하며 VLM 기반 상황 인지를 바탕으로 초기 상태를 생성하는 기능을 함께 수행한다. 도메인은 로봇이 수행할 수 있는 행동 영역을 의미하며, 이후 실현 가능한 action을 결정하는 데 기여한다. 언어적 명령이 입력되었을 때 대규모 언어 모델이 로봇의 현재 상태가 고려된 문맥 기반 해석을 실행하여 자연어 명령의 모호성 및 생성 이벤트와 규칙의 출동을 방지하고 VLM의 텍스트 설명, 도메인 기술서를 활용해 초기상태를 구성한다. 위 모듈은 행동을 결정짓는 것이 아닌 자연어, VLM 초기 상태를 생성하는 것에 한정하며 최종 동작에 대한 핵심적인 판단 근거로 사용된다.

실시간 태스크 스케줄러는 도메인 우선순위 판단부터 동작 결정 및 제어 신호 생성까지의 실행 흐름을 계층적으로 구성하며, 구조는 [Fig. 2]에 제시되어 있다. 시스템은 이벤트 수집 후 응급 상황 여부와 규칙 기반 조건을 검증하여 상태 기계로 전달할 이벤트를 결정한다. VLM 기반 응급 상황 감지는 가장 높은 우선순위를 가지며, 탐지 즉시 모든 사용자 요청보다 우선 처리되어 시스템을 응급 상태로 전환한다. 응급 상황이 없을 경우에는 현재 상태에서 허용되는 이벤트만을 선택하여 중복되거나 맥락과 충돌하는 입력을 제거함으로써 LLM 출력의 비결정성이 제어 흐름에 미치는 영향을 최소화한다. 상태가 확정되면 Action Manager가 대응 동작을 결정하고, Robot Controller는 이를 ROS2 기반 제어 신호로 변환하여 실제 로봇 동작으로 실행한다.

Ⅲ. 결과

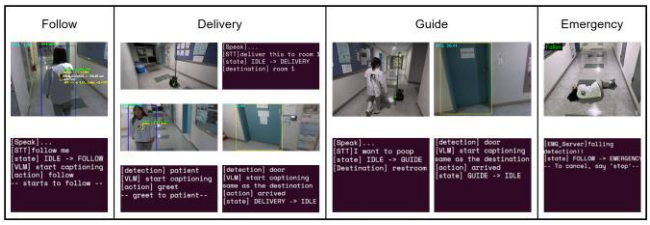


그림 3. 실환경에서의 상태전환 검증

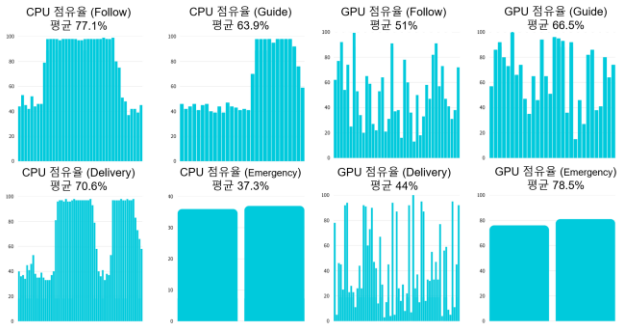


그림 4. 작업 수행 중 시스템 CPU 및 GPU 점유율

아키텍처의 실효성 검증을 위해 Turtlebot3 기반의 이동 로봇 플랫폼을 활용하였으며, 실험은 가상의 의료 환경을 가장한 실 환경에서 네 가지 핵심 도메인(Follow, Emergency, Guide, Delivery)을 포함한 통합 시나리오를 구성하여 수행하였다. Follow, Emergency, Guide, Delivery 순으로 진행되며 각 단계는 상황 인지, 도메인 전환을 검증하기 위한 독립적 동작 조건으로 구성된다.

표 1. 혼동행렬

State	Follow	Guide	Delivery	Emg.
Follow	98	8	0	11
Guide	4	92	14	2
Delivery	0	0	86	1
Emg.	0	0	0	86

도메인 별로 100개의 언어적 명령을 사용하여 로봇의 상태 전이 능력의 정량적 성능 평가를 수행하였으며 정확도는 위 혼동행렬과 같다. Delivery 도메인의 오분류(14%)와 Emergency 도메인의 오분류 (11%)는 이동에 대한 명령 요소 및 목적성에 대한 언어적 모호성을 원인으로 추정한다.

Ⅳ. 결론

본 연구는 헬스케어 환경의 복합적 요구에 대응하기 위해 LLM과 VLM의 지능형 추론을 통합하고, 이를 우선순위 기반 태스크 스케줄링 프레임워크로 제어하는 로봇 아키텍처를 제안하였다. 실험 결과, 자연어 명령 해석의 90% 정확도와 VLM 기반 응급 상황 선점 메커니즘의 성공적인 작동을 확인하며, 향후 연구는 도메인 간 의미 경계 강화를 위한 도메인 특화 미세조정과 실시간 반응성 확보를 위한 모델 경량화 및 하드웨어 가속 설계에 집중하여 시스템을 고도화할 계획이다.

ACKNOWLEDGMENT

본 과제(결과물)는 2025년도 교육부 및 경기도의 재원으로 경기RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다. (20YY-RISE-09-A15)

참 고 문 헌

[1] K. Kim, J. Windle, M. Christian, T. Windle, E. Ryherd, P.-C. Huang, A. Robinson, and R. Chapman, “Framework for Integrating Large Language Models with a Robotic Health Attendant for Adaptive Task Execution in Patient Care,” Appl. Sci., vol. 14, no. 21, p. 9922, 2024

[2] K. Shirai, C. C. Beltran-Hernandez, M. Hamaya, A. Hashimoto, S. Tanaka, K. Kawaharazuka, K. Tanaka, Y. Ushiku, and S. Mori, “Vision-Language Interpreter for Robot Task Planning,” arXiv preprint arXiv:2311.00967v2, Feb. 2024.