Journal of Korea Robotics Society (2026) 21(2):140-147
https://doi.org/10.7746/jkros.2026.21.2.140 ISSN: 1975-6291 / eISSN: 2287-3961

VLM을 이용한 해양로봇 자율주행 시스템 최적화 연구

Optimization of Autonomous Controls Systems using Vision-Language Model for Maritime Robotics

김 민 정[1] · 김 태 연[1] · 김 병 모[2] · 최 원 석†
Min-Jung Kim[1], Tae-Yeon Kim[1], Byung-Mo Kim[2], Woen-Sug Choi†

Abstract: In the recent developments, contrary to previous question-answering systems, VLMs (Vision-Language Models) are capable to analyze the semantic meaning with visual interpretation simutanously. Such semantic capabilities of the VLM has been exploited in the autonomous control systems. Requirements of techniques for applications are challenging due to high versatility of the scenes at marine environments. Compared to static scenarios, in ocean, the surface changes continuously due to waves effected by currents and winds. Therefore, the difficulty of localizations and eliminations of noise from sensory data causes wrong decision makings for autonomous systems. Controls based on those inaccurate estimates of the position and status of the maritine robots converge to incorrect navigations. The commonly practiced procedures of the autonomous sytem includes visual input, semantic analysis using VLM, decision inferences, and motor command excutions. In this paper, to overcome challenges of the marine environments, improvement of autonomous control performance by integrating perception and decision making procedures are developed. Therefore, the time delay and information loss have been decreased. The new structure has shown enhancement on reactivity and reliability of the decision makings in maritime navigations. In simulation, a comparative analysis was carried out based on ROS (Robot Operating System) and Gazebo between a non-integrated model (with separate image interpretation and generation of the commands) and the integrated one. The integrated system also had a faster response time and higher decision reliability across trials. However, as the number of commands increased, consistency decreased, resulting in a slight increase in the total driving distance. Nevertheless, the performance was smoother and more robust than that of the conventional system overall. Important items for future work are, for example, better stability behavior with frequent command updates and reduced computational demands that guarantee that VLM-driven systems can be used for reliable real-time navigation in challenging maritime environments.

Keywords: Vision-Language Model, Prompt Engineering, Autonomous Control System, ROS-Gazebo Framework, Maritime Robotics

1. 서 론

기존의 자율주행 시스템은 LiDAR, 카메라, IMU 등 다양한 센서를 융합한 SLAM[1] 기반 기술을 중심으로 발전해 왔다. 거리 계산 및 정밀한 위치 추적에 높은 성능을 나타내는 이러한 기술은 사전에 정의되지 않은 급격한 환경 변화, 다양한 센서의 상호작용에 유연하게 대응하기 어렵다는 한계를 지닌다. 특히, 최근 해양 로보틱스 분야에서 로봇 시스템 사용분야 및 작동환

Received : Aug. 25. 2025; Revised : Nov. 6. 2025; Accepted : Apr. 22. 2026
1. Master's Student, Department of Convergence Study on the Ocean Science and Technology, Korea Maritime and Ocean University, Busan, Korea (kmjeong000@gmail.com, yncity@naver.com)
2. Assistant Professor, Department of Ocean Engineering, Korea Maritime and Ocean University, Busan, Korea (bmkim@kmou.ac.kr)
† Assistant Professor, Corresponding author: Department of Ocean Engineering, Korea Maritime and Ocean University, Busan, Korea (wschoi@kmou.ac.kr)

경의 확대에 따라 자율운항 지능화와 상호작용의 유연성이 중요한 과제로 대두되고 있다. 이에 따라 실시간 시각 정보의 판단과 대응이 요구되는 복잡한 환경에서 전통적 제어 방식이 가진 제한적인 인식의 의사결정 한계를 넘어 언어와 시각을 통합적으로 해석할 수 있는 인공지능 모델의 활용이 각광받고 있다. 특히, 본 연구에서는 VLM (Vision-Language Model)[2]의 이미지 - 텍스트 간의 복합 추론 능력에 집중하였다.

LLM (Large-Language Model)에 멀티모달 기능을 결합하여 발전된 형태인 VLM은 자연어 기반의 질의응답 뿐만 아니라 장면의 의미를 이해하고 설명할 수 있는 차세대 인공지능 모델로 평가된다. 대규모 이미지와 데이터를 학습한 VLM은 기존의 딥러닝 기반 시각 인식보다 우수한 일반화 능력을 제공하며, 별도의 미세조정 (fine-tuning) 없이도 다양한 시각 인식 작업을 수행할 수 있는 제로샷(zero-shot) 예측이 가능하다[3]. 이를 통해 단순한 객체 인식을 넘어 복합적인 상황 인식과 판단 작업이 적합하며[4], 이러한 VLM의 특성은 자율주행 시스템과 같이 실시간으로 시각 정보의 해석이 요구되는 환경에 효과적으로 작용한다.

자율주행 제어와 관련된 기존의 연구[5,6]는 주로 정적이고 단순한 환경에서 진행되었으며, 이는 실제 해양 환경에서 나타나는 동적 특성을 반영하기에 역부족이다. 해양 표면은 파도, 조류, 바람 등에 의해 끊임없이 변화하기 때문에 정적 환경에 비해 센서 노이즈나 오차가 더 자주 발생한다. 특히 수면 반사나 굴절, 그림자, 파도 및 햇빛의 간섭과 같은 현상은 거리 기반 센서의 신호 왜곡과 손실의 요인이 되기도 한다[7,8]. 결국 거리 추정의 정확도를 저하해 잘못된 제어 판단으로 이어질 수 있기에 개선이 필요하다.

이처럼 복잡한 환경에서 가지는 기존 거리 기반 센서의 한계를 극복하기 위해, 본 연구는 시각 정보를 보다 고차원적으로 해석할 수 있는 VLM을 활용한 자율주행 제어 시스템을 제안하고자 한다. VLM의 시각 정보 처리 능력을 활용한 복잡한 해양 환경에서도 실시간 시각 정보 분석 능력이 개선된 정교한 상황 인식을 목표로 한다.

제안된 시스템은 이미지 입력, 시각적 분석, 명령 판단, 그리고 제어 실행의 과정을 수행하며, VLM은 이 모든 판단 과정에서 중심적인 역할을 담당한다. 이러한 VLM의 구현을 위해, 본 연구에서는 OpenAI의 GPT-4o[9]를 활용 하였다. GPT-4o는 2024년 5월에 출시된 최신 멀티모달 모델로, 자율주행 제어 시스템 내에서 이미지 해석과 명령 판단을 동시에 수행하는 데 적합한 구조를 지닌다.

자율주행 시스템은 ROS 2와 Gazebo 시뮬레이터를 이용해 구현했다. 전체 시스템은 ROS의 모듈형 노드 아키텍처 기반의 독립적인 모듈로 구분하고, 상호 통신이 가능하도록 설계되었다. 이와 같은 구조는 VLM이 이미지 이해 및 제어 명령을 실시간으로 처리하기 위한 시스템 확장과 유지 보수에 효과적[10]이다. 또한, Gazebo와 ROS 2를 연동하여 실제 해양 환경과 유사하게 구현해 시스템 신뢰도를 높이고, 이미지 센서 데이터 수신부터 명령 전달까지의 전 과정을 시뮬레이션 상에서 검증할 수 있도록 했다.

기존 연구 방식에서는 VLM을 이용한 이미지 분석과 명령 기능이 각각의 GPT-4o 호출을 필요로 했다. 그러나 이런 분리 구조에서는 복합적인 연산이 요구되어 처리 시간이 길어지고, 이미지 분석 결과를 전달하는 과정에서 일부 정보가 손실될 수 있어 잘못된 판단을 할 가능성도 존재하였다. 복잡한 환경이나 새로운 상황에서 대응이 어려워 실사용에 한계가 존재하기도 한다[11]. 이와 같은 한계를 극복하고자 본 연구에서는 이미지 분석과 명령을 하나의 노드에서 동시에 수행하는 통합형 구조를 착안하였다. 단일 노드 구조의 자율주행 시스템은 GPT 호출 횟수를 줄이고 연산 과정을 단순화 함으로써 기존 분리형 구조에서 나타났던 처리 지연을 완화하여 실시간성이 강화되고, 정보 손실 문제가 해소 되어 판단의 정확도가 높아지는 효과를 기대할 수 있다.

본 연구는 자율주행 제어 시스템의 이미지 분석과 명령 생성 기능을 하나의 노드로 통합함으로써, 기존의 분리형 구조에서 발생할 수 있는 처리 지연과 정보 손실 문제를 완화하고 시스템 성능 향상을 도모한다.

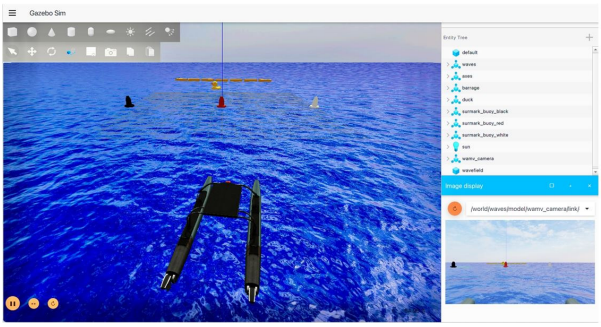
이를 위해 분리형 구조와 통합형 구조를 동일한 ROS-Gazebo 시뮬레이션 환경에서 구현하고, 응답 시간, 판단 정확도, 시스템 유지보수성 측면에서 성능을 비교 하였다. 두 구조 모두 GPT-4o를 기반으로 구축되었으며, 다중 모달 입력에 대한 명확한 응답을 유도하기 위해 프롬프트 엔지니어링[12] 기법을 적용하였다. 시스템 전체 효율성을 구조적 차이의 관점에서 평가함으로써, 향후 VLM을 이용한 실시간 자율주행 시스템 최적화의 방향성을 제시하고자 한다.

2. 시스템 설계 및 구현

연구 목적에 따라 ROS 2와 Gazebo를 활용한 VLM 기반 자율주행 제어 시스템을 설계 및 구현하였다. 본 장에서는 이미지 입력부터 판단, 명령 생성, 추력 제어까지의 과정을 포함한 자율주행 제어 시스템의 시뮬레이션 환경 구성과 시스템 아키텍처, 그리고 노드 통합 방식에 대해 자세히 설명한다.

2.1 시뮬레이션 환경

[Fig. 1]은 실험이 수행된 시뮬레이션 환경을 이미지로 보여준다. WAM-V 로봇 전방에는 서로 다른 색상의 부표 세 개, 목



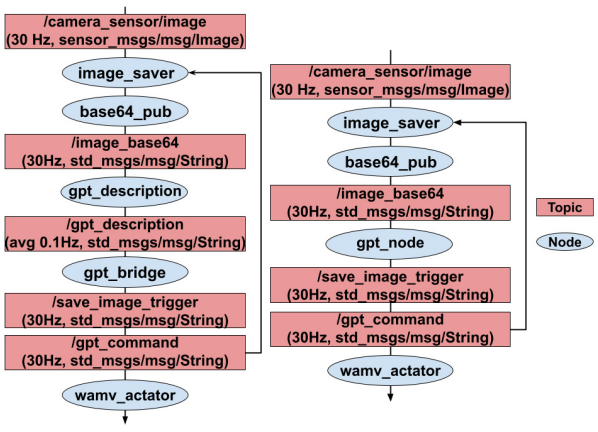
[Fig. 1] Gazebo simulation screen

표물인 고무 오리, 그리고 목표물과 동일한 색상의 긴 부표가 일렬로 배치되어 있다. 시뮬레이션은 Apple M2 (8-core CPU, 10-core GPU, 16GB unified memory)가 탑재된 환경에서 ROS 2 Jazzy와 Gazebo Harmonic(v8.9)을 이용해 구축되었다. 분리형 구조에서는 두 번의 요청에서 각각 평균 1500 tokens 및 374 tokens를 사용하였고, 통합형 구조는 단일 요청으로 약 1423 tokens를 사용하였다.

모든 노드는 Python 환경에서 ROS 2 Python 라이브러리(rclpy)를 사용하여 작성되었으며, 노드 간 메시지 교환을 통해 센서 데이터 수집과 제어 명령 전달이 실시간으로 이루어진다. Gazebo 시뮬레이터는 실제 환경과 유사한 물리적 특성을 반영하여 센서 데이터를 생성하고, ROS 2와 연동하여 자율주행 시스템의 동작을 검증하는 데 사용되었다.

2.2 전체 시스템 구성 및 노드 구조

[Fig. 2]는 전체 노드 간 통신 구조를 도식화한 것이다. 왼쪽은 이미지 설명과 명령 판단을 분리한 구조이고, 오른쪽은 이를 하나의 노드로 통합하여 처리하는 구조이다. 각 노드 간의 통신



[Fig. 2] Communication architecture of the autonomous control system: separated architecture (left) and integrated architecture (right)

은 ROS 2 토픽을 통해 이루어지며, 토픽명, 메시지 타입, 발행 주기가 명시되어 있다.

먼저, 시뮬레이터 상의 해양 로봇에 장착된 카메라 센서가 주변 환경을 촬영한 이미지를 토픽으로 발행한다. 이 토픽을 구독한 노드는 수신된 이미지를 로컬 디렉토리에 실시간으로 저장하며, (image_saver) 저장된 이미지는 GPT 모델이 수신 가능한 base64 형식으로 인코딩한다. (base64_pub) GPT모델은 해당 토픽을 구독하여 이미지를 해석하고, 주행에 필요한 프로펠러 추력 명령을 생성한다. 동시에 다음 이미지 저장을 위한 트리거를 생성한다. 명령에 따라 양쪽 프로펠러의 추력이 발행되면 드론은 이동하며 (wamv_actuator) 전체 과정은 하나의 루프로 구성되어 반복적으로 수행된다.

두 구조의 핵심적인 차이는 GPT 모델의 역할 분리에 있다. 분리형 구조는 이미지 분석(gpt_description)과 명령 판단(gpt_bridge)을 별도의 노드로 분리하여 GPT를 두 차례 호출한다. 이 방식은 이미지 설명과 판단 과정이 명확히 구분되므로 각 단계의 출력 결과를 독립적으로 확인하고 분석할 수 있다. 따라서 디버깅이 용이하고, GPT의 시각적 해석 정확도를 정량적으로 평가할 수 있다는 이점이 존재한다. 그러나, 프롬프트를 단계 분리형 구조로 구성함에 따라 개별 프롬프트의 길이가 길어지고, 복합적인 연산이 요구된다. GPT를 두 번 호출하기 때문에 평균 응답 시간도 증가하여 결국 시스템의 실시간성이 저하될 수 있다. 또한, 전달하는 이미지 분석 정보가 한정되기 때문에 주행에 필요한 맥락이 손실될 가능성이 있다.

반면, 통합형 구조는 GPT를 한 번만 호출하여 이미지 분석과 판단을 단일 프롬프트에서 처리한다. 그 결과, 응답 지연이 줄어 시스템 실시간성이 개선되고, 처리 단계가 축소됨으로써 맥락 보존과 정보 손실의 최소화라는 구조적 효과를 얻는다. 이로 인해 모델이 상황을 보다 정확하게 해석하고 추론하여 주행 정확도가 향상될 것이다.

2.3 프롬프트 엔지니어링

본 연구에서는 VLM이 시각 정보를 정확히 인식하고 상황 판단을 수행할 수 있도록 프롬프트 엔지니어링 기법을 적용하였다. 두 구조 모두 카메라의 위치 및 시야각, 목표 객체와 장애물 정보 등을 프롬프트에 포함하며, GPT의 입출력 형식을 JSON으로 제한하여 응답 효율성과 처리 일관성을 높였다.

[Table 1]과 [Table 2]는 각각 분리형 시스템에서 사용된 이미지 분석 노드와 명령 판단 노드 프롬프트의 주요 구문이다.

이미지 분석 노드에서는 객체 식별 명령을 통해 탐지 누락을 최소화하고, 각도와 거리 등을 정량적으로 추정하도록 설계되었다. 이어서, 명령 판단 노드는 분석 결과를 받아 전후좌우 네 개의 명령 중 하나를 선택하는 역할을 담당하며, 조건을 기반으

[Table 1] Prompt used in the separated structure

gpt_description Node
Identify all major objects. For each, provide: - type: Must be either 'obstacle' or 'rubber_duck - color: dominant color - angle: horizontal angle of the object's closest visible point relative to image center. FOV is -42.5° (left) to +42.5° (right) - distance (0-10): estimate based on **lowest visible pixel of the object Distance estimation guidelines: - Bottom edge near engines (bottom of image): close (0-3) - Bottom edge near top of image: far (7-10)

[Table 2] Prompt used in the separated structure

gpt_bridge Node
1. Find the rubber_duck. If not found, go to Step 4. 2. Identify all obstacles closer (lower 'distance' value) than the duck. If such obstacles exist, treat as blocking. 3. If obstacles block the duck: - If obstacle is left of center, command = 'right' - If obstacle is right of center, command = 'left' - If obstacle is centered and very close (distance ≤ 3), randomly pick left or right 4. If no blocking obstacles: - duck angle > 0 → 'right' - duck angle < 0 → 'left' - duck angle = 0 → 'forward' - duck's distance ≤ 3 → 'stop' 5. If no duck present and all obstacles are far (distance > 5): initiate target search by turning ('left' or 'right')

로 명령 추론을 유도하고, 현재 환경 상태를 통해 행동을 결정하는 상태-행동 매핑 구조를 갖는다. 이와 같은 분리형 구조의 제어 로직은 If then 기반의 Heuristic 방식[13]에 해당한다. 이는 명시적 규칙 기반 제어 로직으로, 사람이 사전에 정해놓은 규칙에 따라 조건을 평가해 즉시 결과를 반환한다. 단계적 분석 없이 조건에서 결과로 직결되는 제어 로직이다.

[Table 3]는 통합형 구조 프롬프트의 주요 구문이다. 해당 프롬프트는 단일 GPT 호출로 시각 인식과 주행 판단을 동시에 수행하는 end-to-end 방식으로, GPT-4o의 비전-언어 멀티모달 능력을 활용하여 입력 이미지로부터 직접적으로 이동명령을 추론한다. 이는 단순한 질의응답을 넘어서 복잡한 연역적 사고(Chain of Thought)가 유도되도록 설계되었다. CoT 방식은 단계적 사고를 유도하며 언어 모델은 상황을 이해하고 추론 후 행동을 스스로 결정한다.

프롬프트 내에서는 GPT-4o에 자율주행 드론의 판단 AI 역할을 부여하고, step-by-step 방식의 추론을 유도함으로써 논리적 사고 흐름을 끌어낸다. 클래스 구분을 기반으로 한 시각 인식을 요청하고, 시각 단서에 기반한 간접적 거리 추론 방식과

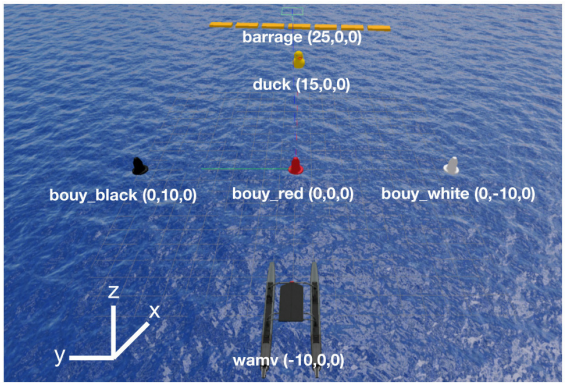
[Table 3] Prompt used in the integrated sturcture

gpt_node Node
Reason step-by-step: 1. Identify the yellow duck and any obstacles 2. Estimate distance using the lowest visible pixel of each object (closer = lower). - Objects near the gray engines at the image bottom are very close. 3. If the duck is visible and no closer obstacle blocks the path: - Move toward the duck's direction: `left`, `right`, or `forward` depending on its position in the image. - Avoid going `forward` if the duck is far off-center (left/right edge) — turn instead to keep it centered. 4. If the duck is not visible → `left` or 'forward' to search. 5. If a closer obstacle blocks the path → turn `left` or `right` to avoid it. 6. If the duck appears very close (bottom edge of duck near engine) → issue `stop`. Ignore all UI overlays, labels, or non-physical graphics.

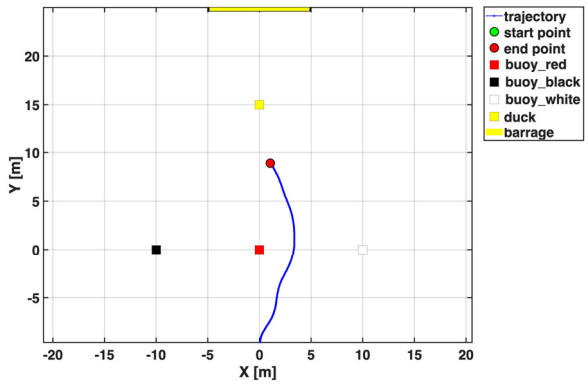
같은 단계별 판단 기준도 함께 제시한다. 특히, 통합형 프롬프트의 "Reason step-by-step"이라는 구문은 CoT 유도 프롬프트의 전형적인 트리거로[14] reasoning 지시어를 포함하고 있으며, 프롬프트는 명시적으로 추론 단계를 요구하고 있다. 언어 모델이 입력된 이미지를 보고 스스로 생각하는 과정을 거쳐 최종 행동을 내리도록 구성되어있기 때문에 통합형 구조는 CoT 방식이라고 볼 수 있다.

3. 실험 및 결과

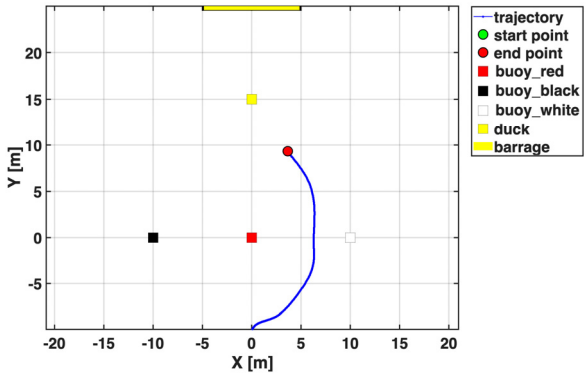
제안한 자율주행 시스템의 성능을 평가하기 위해 분리형 구조와 통합형 구조에 대한 실험을 동일한 시나리오에서 각각 50회씩 수행하였다. 두 구조의 응답 속도, 명령 횟수, 성공률을 비교하였고, 실험은 [Fig. 3]과 같은 환경에서 진행되었다. 물체들은 드론으로부터 각각 10m, 25m, 35m 떨어진 위치에 배치하였고, 세 개의 원통형 부표는 좌우 10m 간격으로 배열되었다.



[Fig. 3] Experimental environment



[Fig. 4] Trajectory visualization of the separated architecture



[Fig. 5] Trajectory visualization of the integrated architecture

실험 목표는 드론이 세 개의 원통형 부표와 가장 후방의 부표선에 충돌하지 않고 고무 오리에 도달하는 것이며, 드론이 부표와 충돌하거나 고무 오리를 찾지 못한다고 판단되는 경우에는 해당 실험을 실패로 기록하였다.

[Fig. 4]와 [Fig. 5]는 시뮬레이션에서 기록된 wamv_camera 모델의 pose(x,y) 데이터를 시각화한 것으로, 각각 분리형과 통합형의 이동 궤적을 보여준다. 로봇 중심 좌표계를 기준으로 주행 경로를 기록하여, 실제 이동 패턴과 목표 접근 과정을 시각적으로 확인할 수 있다.

실험 결과의 정량적 비교는 [Table 4]에 정리하였다.

[Table 4] Experimental Results

Category	Separated	Integrated
Pass	13 (26%)	22 (44%)
Fail	37	28
All	50	50
Time to Target (s)	133	180
Avg. Command Count	14.3	30.7
Avg. Response Time (s)	9.24	5.87
Response Time Variance	3.74	4.25
Response Time Std. Dev (s)	1.94	2.06

3.1 실험 결과 분석

3.1.1 성공률 비교

분리형 구조는 26%, 통합형 구조는 44%의 성공률을 보여 통합형 구조가 비교적 안정적인 주행 성능을 가진다. 이는 통합형 구조가 인식과 판단을 단일 콘텍스트에서 처리하여 정보 손실이 줄어든 결과로 해석할 수 있다.

3.1.2 목표 도달 시간 비교

분리형 구조의 목표 도달 시간은 평균 133초, 통합형 구조는 평균 180초 소요되었다. 통합형의 응답 속도가 더 빨랐음에도 목표 도달 시간이 분리형에 비해 오래 걸리는 이유는 평균 명령 횟수가 2배 이상 많기 때문이라 분석할 수 있다.

3.1.3 명령 횟수 비교

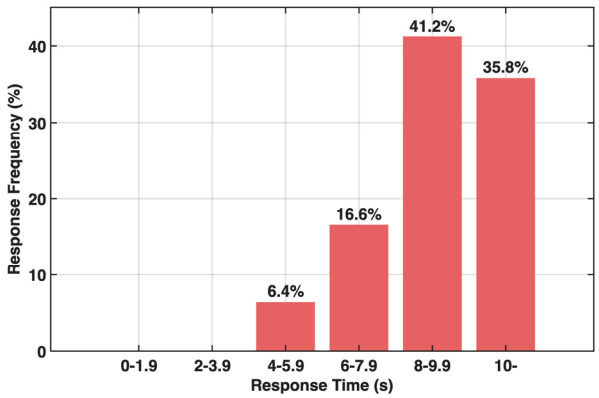
분리형 구조의 경우 목표에 도달하기 위해 생성된 평균 명령 횟수는 14.3회였다. 통합형 구조는 30.7회로, 분리형보다 약 2.1배 더 많은 명령을 생성했다.

3.1.4 응답 속도 비교

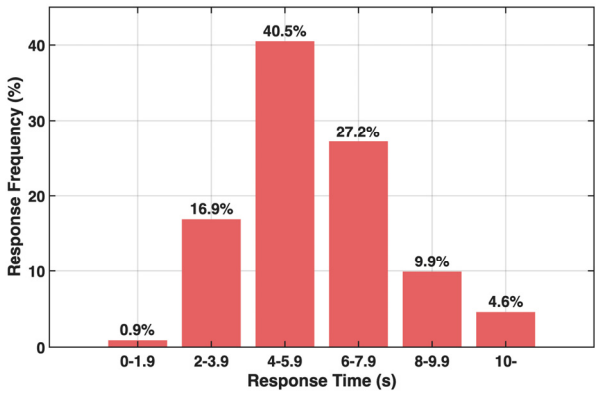
평균 응답 시간은 분리형이 9.24초, 통합형은 36.5% 더 빠른 5.87초로 측정되었다. [Fig. 4]와 [Fig. 5]에서 확인할 수 있듯이, 통합형 구조의 주행 궤적은 보다 부드러운 곡선을 그리고 있으며, 이는 제어 명령 생성 주기가 짧고 실시간 반응성이 높음을 보여준다. 또한, 구간별 응답 시간 분포를 비교한 결과, 두 구조 간의 차이가 더욱 명확하게 나타났다.

목표 도달에 성공했을 때의 명령 횟수를 합산하여 응답 빈도를 계산한 그래프이다. [Fig. 6]의 분리형 구조에서는 전체 응답의 77%가 8초 이상 소요되어 전반적으로 지연되는 경향이 나타난다. 반면, [Fig. 7]의 통합형 구조에서는 8초 미만 응답이 85.5%를 차지한다. 또한, 4~5.9초 구간에서 최대 빈도를 보인다. 즉, 명령 횟수가 두 배 이상 많음에도 불구하고 통합형 구조에서 상대적으로 빠르고 집중적인 응답 분포를 확인할 수 있었다. 이러한 분포는 통합형 구조가 제어 시스템의 실시간성 측면에서 우수함을 시사한다.

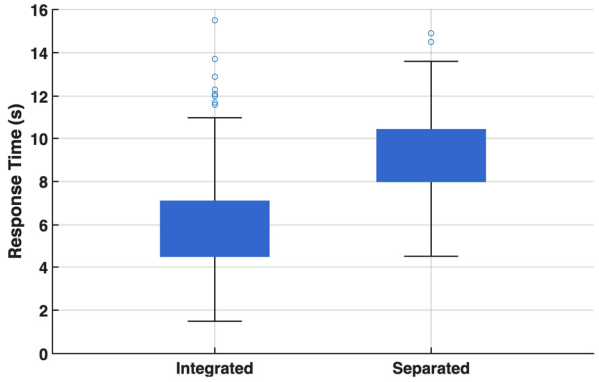
이러한 응답 시간의 차이는 상자 수염(Box-Whisker) 그래프를 통해서도 명확하게 확인된다. [Fig. 8]에서 분리형 구조의 중앙값은 약 10초로 높고 사분위수 범위가 매우 좁아, 응답 속도가 전반적으로 느리고 일정한 경향을 보인다. 반면 통합형 구조는 중앙값이 약 6초로 낮고 사분위수 범위가 넓으며 일부 이상값이 존재한다. 이는 통합형 구조가 전반적으로 빠른 응답 특성을 가지면서도 상황에 따라 변동성을 가짐을 보여준다.



[Fig. 6] Response time of the separated structure



[Fig. 7] Response time of the integrated structure



[Fig. 8] Box-Whisker plot of response times for the separated and integrated structures

3.2 결과 요약

실험 결과에 따르면, 본 연구에서 제안한 통합형 구조는 분리형보다 1.5배 빠른 응답 속도를 보였다. 명령 횟수는 2.1배 증가했다. 따라서 통합형 구조가 분리형에 비해 전반적으로 우수한 성능을 보이며, 그 결과, 성공률도 약 1.7배 상승한 것을 알 수 있다. 특히, 통합형은 단일 프롬프트에서 판단과 명령 생성을 동시에 수행함으로써 응답 지연과 정보 손실을 줄였고, 이로 인해 더욱 세밀하고 연속적인 제어 루프를 실현할 수 있었던 것으로 분석된다.

이러한 성능 차이는 이미지 인식 결과를 활용하는 방식에서 기인한 것으로 판단된다. 두 구조 모두 동일한 VLM을 사용하므로 인식 성능의 절대적인 차이는 크지 않다. 그러나 분리형 구조는 수치화된 분석 결과만을 근거로 판단하기 때문에 정보 손실이 발생할 가능성이 높지만, 통합형은 모델이 이미지를 직접 보고 추론하여 판단함으로써 풍부한 맥락 정보를 활용할 수 있다. 이러한 특성이 응답 지연 감소와 성공률 향상의 주요 원인으로 작용했을 것으로 해석된다.

다만, 두 구조 모두 이전 상황에 대한 기억 없이 단일 이미지만을 보고 판단과 결정을 내리고 있다. 맥락에 대한 이해 없이 단편적으로 판단할 경우, 유사한 이미지에 대해 상반되는 명령을 출력하는 등 판단 일관성이 떨어지는 경향이 확인되었다. 그에 따라 불필요한 움직임이 발생하여 결국 전체 주행 거리가 증가하였다. 이러한 특징은 특히 응답 시간이 짧고 명령 횟수가 많은 통합형 구조에서 두드러지게 나타나났다.

이러한 점에서 전체 주행 소요 시간을 비교하면, 분리형 구조의 경우 약 133초, 통합형은 약 180초가 요구된다. 즉, 개별 응답 속도는 빨라졌음에도 불구하고 주행 제어의 세분화로 인해 실제 목표 도달 시간은 오히려 증가할 수 있음을 보여준다.

4. 결 론

본 연구에서는 구조 설계에 따른 주행 성능을 비교함으로써 시각-언어 모델을 접목한 자율주행 시스템의 성능 향상 효과를 분석하였다. 프롬프트 엔지니어링을 통해 단계 분리형 구조의 두 프롬프트를 하나의 논리 판단 구조 프롬프트로 통합함으로써, 지연 시간 단축 및 주행 정확도 향상을 실험적으로 검증하였다.

실험 결과를 통해 자율주행 시스템의 성능을 향상하는 방법으로 시스템의 구조 변화가 효과적임이 확인되었다. 분리되어 있던 두 기능을 통합함으로써 언어 모델의 인식과 판단이 단일 콘텍스트에서 처리되며 맥락이 보존되고 정보 손실이 최소화되는 구조적 효과를 얻을 수 있었다. 구조적 통합과 더불어 단계적 사고를 유도하는 방식의 프롬프트 설계는 GPT가 합리적으로 생각하도록 유도해 객체 인식과 주행 판단의 정확도를 높였고, 그에 따라 목표 도달률도 증가하였다. 또한, 프롬프트의 병합으로 인해 자연스럽게 GPT 호출 횟수가 감소하고 복잡한 연산이 줄어들었다. 그 결과 지연 시간이 절감되고 실시간성이 향상되었다.

응답 시간의 단축과 주행 정확도 향상에 초점을 맞추어 실험을 진행했고, 두 목표 모두에서 기대한 결과를 확인할 수 있었다. 그러나, 실험을 진행하며 이번 연구의 분석 범위에 포함되

지 않았던 문제점을 발견하였다. 명령 횟수의 증가에 따른 전체 주행 시간의 변화는 GPT의 단편적인 장면 판단에 의한 것으로, 주행의 일관성을 저해하는 요인이 된다. 이는 더 짧은 주기로 판단을 내리는 통합형 구조에서 명확하게 확인할 수 있었고, 시스템이 실제 물리 환경에 적용될 경우 연료 소비량 증가와 같은 악영향을 일으킬 수 있다. 특히 본 연구는 시뮬레이션 환경을 기반으로 수행되었기 때문에 실제 환경에서는 파도, 풍속, 센서 노이즈 등 다양한 외부 요인으로 인해 성능 특성이 달라질 수 있다. 따라서 향후 연구에서는 실제 해역 실험을 통해 환경 요인이 시스템 성능에 미치는 영향을 분석하고, 다양한 환경 조건에서도 높은 판단 일관성을 유지하는 안정적인 자율주행 시스템 설계 방안을 모색할 필요가 있다.

또한, 실험은 GPT-4o를 기반으로 수행하였기 때문에 결과가 모델 특성에 부분적으로 종속될 가능성이 있다. 모델 의존성을 완화하고 알고리즘의 일반화 가능성을 평가하기 위해 후속 연구에서는 LLaVA 1.6, LLaVA-NeXT, CogVLM 등 대표적인 VLM 모델을 포함한 다양한 모델을 적용하여 프롬프트 구조에 따른 성능을 비교 및 검증하고자 한다.

본 연구는 실험을 통해 VLM기반 자율제어 시스템의 구조 변화에 따른 시스템 성능 향상을 증명하였다. 이는 기존의 거리 기반 센서 시스템과 차별화된 시각 기반 자율주행 시스템의 새로운 가능성을 보여준다. 향후 멀티모달 인공지능을 활용한 로봇 및 드론 제어 분야에서 기초 자료가 될 것이다. 특히 구조 변화에 따른 제어 성능의 차이를 분석한다는 점에서 연구적 의의가 있다.

References

[1] C. Cadena et al., “Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age,” *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309-1332, Dec., 2016, DOI: 10.1109/TRO.2016.2624754.

[2] 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, DOI: 10.1109/TPAMI.2024.3369699.

[3] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” *in Proceedings of the 40th International Conference on Machine Learning*, Honolulu, HI, USA, PMLR, vol. 202, pp. 19730-19742, 2023, [Online], https://proceedings.mlr.press/v202/li23q.html.

[4] T. Wang, J. Fan, P. Zheng, R. Yan, and L. Wang, “Vision-Language Model-Based Human-Guided Mobile Robot Navigation in an Unstructured Environment for Human-Centric Smart Manufacturing,” *Engineering*, in press, Jul., 2025, DOI: 10.1016/j.eng.2025.04.028.

[5] W. Lee, T. Kim, and W. Choi, “Autonomous Navigation Command Generation System for Underwater Robots Using LLM,” *Journal of Institute of Control, Robotics and Systems*, vol. 30, no. 10, pp. 1191-1196, Oct., 2024, DOI: 10.5302/j.icros.2024.24.0142.

[6] J. Lee, E. Jo, H. Jeong, and H. Kim, “Development of Real-Time Control Software for Autonomous Mobile Robot,” *Journal of Institute of Control, Robotics and Systems*, vol. 17, no. 4, pp. 336-345, Apr., 2011, DOI: 10.5302/J.ICROS.2011.17.4.336.

[7] O. Korotkova and J. R. Yao, “Bi-static LIDAR systems operating in the presence of oceanic turbulence,” *Optics Communications*, vol. 460, Art. no. 125119, Apr., 2020, DOI: 10.1016/j.optcom.2019.125119.

[8] X. Liu, Y. Wang, F. Liu, and Y. Zhang, “Investigation on the utilization of millimeter-wave radars for ocean wave monitoring,” *Remote Sensing*, vol. 15, no. 23, Art. no. 5606, Dec., 2023, DOI: 10.3390/rs15235606.

[9] S. Shahriar et al., “Putting GPT-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency,” *Applied Sciences*, vol. 14, no. 17, Art. no. 7782, Sep., 2024, DOI: 10.3390/app14177782.

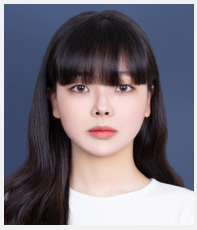
[10] M. Quigley et al., “ROS: an open-source Robot Operating System,” in *ICRA Workshop on Open Source Software*, Kobe, Japan, 2009, [Online], https://robotics.stanford.edu/~ang/papers/icraoss09-ROS.pdf.

[11] G. J. Nalepa, “A New Approach to the Rule-Based Systems Design and Implementation Process,” *Computer Science*, vol. 6, pp. 65-79, Jul., 2004, DOI: 10.7494/csci.2004.6.5.65.

[12] J. Wang et al., “Review of large vision models and visual prompt engineering,” *Meta-Radiology*, vol. 1, no. 3, Art. no. 100047, Nov., 2023, DOI: 10.1016/j.metrad.2023.100047.

[13] H. Liu, A. Gegov, and M. Cocea, “Rule-based systems: a granular computing perspective,” *Granular Computing*, vol. 1, no. 4, pp. 259-274, May, 2016, DOI: 10.1007/s41066-016-0021-6.

[14] J. Wei et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” in *Advances in Neural Information Processing Systems*, New Orleans, LA, USA, vol. 35, pp. 24824-24837, Jan., 2022, DOI: 10.52202/068431-1800.



김 민 정

2022 국립한국해양대학교 해양공학과(학사)

2026~현재 해양과학기술 전문대학원(석사)

관심분야: ROS-Gazebo 프레임워크, VLM

김 병 모

2009 국립한국해양대학교 해양공학과(학사)

2021 국립한국해양대학교 해양과학기술 전문대학원(박사)

2025~현재 국립한국해양대학교 해양공학과 조교수

관심분야: 해양공학

김 태 연

2018 국립한국해양대학교 해양공학과(학사)

2024~현재 해양과학기술 전문대학원(석사)

관심분야: ROS-Gazebo 프레임워크

최 원 석

2013 일본 요코하마국립대학교 기계재료 공학과(공학사)

2020 서울대학교 조선해양공학과(공학박사)

2022~현재 국립한국해양대학교 해양공학과 조교수

관심분야: ROS-Gazebo 프레임워크, ROS-LLM 통합, 해양 로보틱스