Transactions of KSAE, Vol. 34, No. 1, pp.119-131 (January, 2026)
Check for updates

pISSN 1225-6382 / eISSN 2234-0149
DOI http://dx.doi.org/10.7467/KSAE.2026.34.1.119

VLM 시대의 목표 객체 탐색을 위한 지식 융합 전략 서베이

서 보 건[1] · 김 지 선[1] · 손 준 우[2,3] · 박 명 옥[2] · 김 기 섭[*1]

대구경북과학기술원 로봇및기계전자공학과[1] · 대구경북과학기술원 HumanLab[2] · (주)소네트 자율주행연구팀[3]

Where to Fuse in the VLM Era: A Survey on Integrating Knowledge into Object Goal Navigation

Bokeon Suh[1] · Jiseon Kim[1] · Joonwoo Son[2,3] · Myoungouk Park[2] · Giseop Kim[*1]

[1]*Department of Robotics and Mechatronics Engineering, Daegu Gyeongbuk Institute of Science and Technology, Daegu 42988, Korea*
[2]*HumanLAB, Daegu Gyeongbuk Institute of Science and Technology, Daegu 42988, Korea*
[3]*Autonomous Driving R&D Team, SonnetAI, 243 Techno Jungang-daero, Dalseong-gun, Daegu 43017, Korea*
(Received 4 December 2025 / Revised 12 December 2025 / Accepted 12 December 2025)

Abstract : The rapid advancement of robotics and deep learning has increasingly accelerated the use of Embodied AI, where robots autonomously explore and reason in complex real-world environments. With the growing demand for domestic service robots, efficient navigation in unfamiliar settings has become even more crucial. Object Goal Navigation (OGN) is a fundamental task for this capability, requiring a robot to find and reach a user-specified object in an unknown environment. Solving OGN demands advanced perception, contextual reasoning, and effective exploration strategies. Recent Vision-Language Models (VLMs) and Large Language Models (LLMs) provide agents with external common knowledge and reasoning capabilities. This paper poses the critical question: “Where should VLM/LLM knowledge be fused into Object Goal Navigation?” We categorize knowledge integration into the three stages adapted from the Perception-Prediction-Planning paradigm to offer a structured survey of Object Goal Navigation approaches shaped by the VLM era. We conclude by discussing current dataset limitations and future directions, including further studies on socially interactive navigation and operation in mixed indoor - outdoor environments.

Key words : Indoor navigation(실내탐색), Vision language model(비전언어모델), Large language model(거대언어모델), Object goal navigation(목표객체탐색)

1. 서 론

최근 로봇 기술과 인공지능 분야의 발전은 로봇이 단순히 주어진 경로를 따라 이동하는 것을 넘어, 복잡한 실내 환경에서 스스로 판단하고 행동하는 Embodied AI의 시대를 열고 있다. 가사 보조와 노인 돌봄 등 다양한 임무를 수행하는 가정용 로봇에 관한 관심이 급증하면서 로봇이 우리 생활 공간에서 효율적으로 움직이는 능력의 중요성이 커졌다.

Object Goal Navigation(OGN)은 Fig. 1과 같이 가정용 로봇이 미지의 환경에서 사용자의 요구사항에 부합하는 목표 객체(예: ‘소파’)에 성공적으로 도달할 수 있는지 평가하는 과제이다. 로봇이 생활 환경에서 운용되기 위해서는 고차원적인 시각 정보 처리, 공간 추론, 환경의 맥락 이해 그리고 탐색 전략 확립 능력을 통합적으로 요구한다. OGN을 해결하기 위한 End-to-end 방법들은 주로 Visual feature를 활용해 강화 학습으로 정책 학습에 중점을 뒀다.[1,2] 한편, VLM(Vision Language Model)과 LLM(Large Language Model)의 외부 지식과 추론 능력을 적극적으로 활용하여 OGN을 해결하려는 시도가 증가하고 있다.

본 논문은 이러한 변화 속에서 “VLM/LLM의 지식을 어디에 주입할 것인가?”라는 질문을 던진다. 기존의 자

[*]Corresponding author, E-mail: gsk@dgist.ac.kr

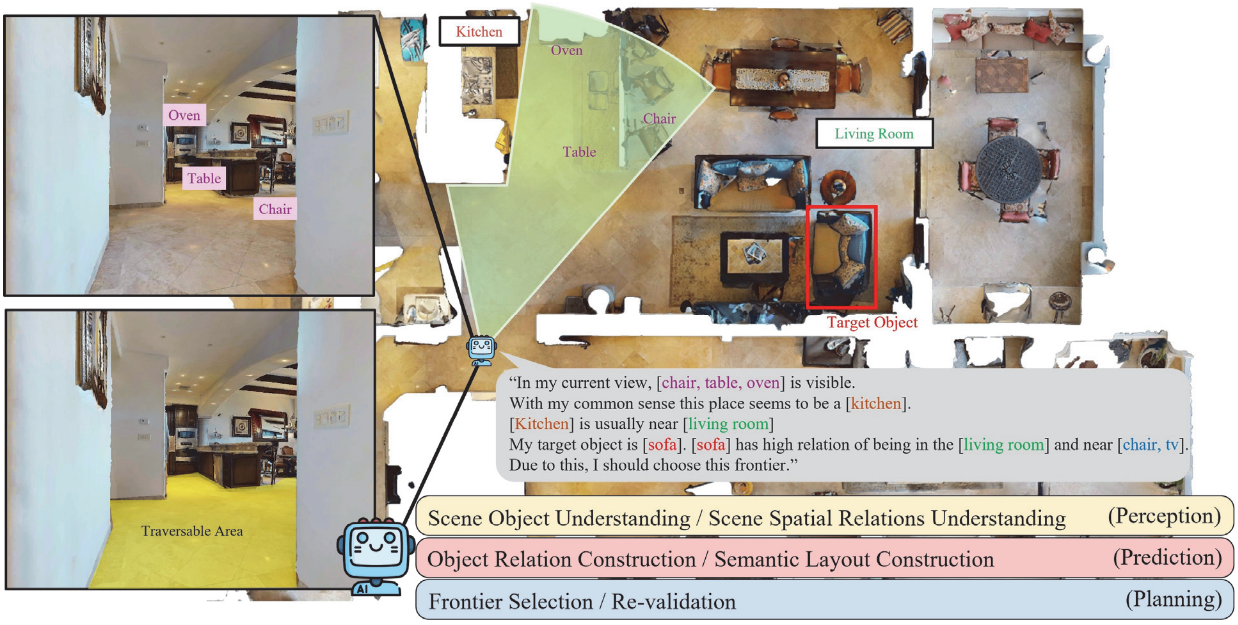

Fig. 1 The agent leverages VLM/LLM to understand the scene (Perception), predict the target object location based on context (Prediction) and selects the most promising frontier to explore (Planning)

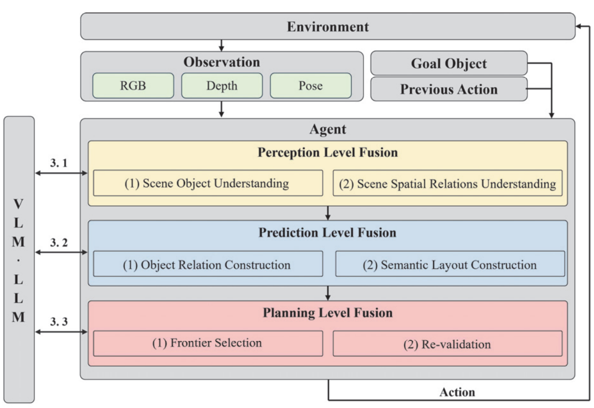

Fig. 2 The proposed framework for categorizing VLM/LLM fusion process into three levels. Classifying fusion strategies as Perception, Prediction and Planning

율주행의 Perception-Prediction-Planning[3-5] 구조를 빌려, OGN을 해결하기 위한 VLM/LLM 지식을 통합하는 단계를 체계적으로 분석하고자 한다. Fig. 2와 같이 VLM/LLM의 지식 융합 단계를 3.1) Perception-level, 3.2) Prediction-level, 3.3) Planning-level의 세 가지 단계로 분류하여, 각 융합 방식의 특징과 대표적인 연구들을 살펴본다.

궁극적으로, 본 논문은 OGN 분야의 최근 흐름을 조망하고 현재 OGN 연구의 주요 데이터셋이 가진 한계점에 대해 알아보고 나아가 사회적 규범 준수를 통해 사람과 함께 실세계에서 로봇이 운행될 수 있는 Socially interactive navigation과 실내외 혼합 환경 등 OGN의 확장 가능성의 소개 등 향후 연구 방향을 제시하고자 한다.

2. Object Goal Navigation

본격적으로 VLM/LLM 융합 전략을 논의하기에 앞서, 먼저 OGN 벤치마크에서 사용되는 실험 셋팅과 평가 지표를 살펴본다. 이어서 실험에 활용되는 주요 데이터셋을 소개하고, OGN과 구조적으로 공통적인 특성을 지닌 다른 태스크들을 살펴본다. 또한, VLM이나 LLM의 도움 없이 OGN 문제를 해결하고자 한 기존 연구들을 간략히 정리하며, 마지막으로 선행 서베이들과 비교하여 본 논문의 차별성을 강조한다.

2.1 Formal Definition of Object Goal Navigation

Object Goal Navigation(OGN)은 로봇이 사전에 관측하지 못한 실내 환경 내에서, 가장 효율적인 경로로 주어진 목표 객체를 탐색하고 접근하는 문제이다.[2,21,22] 각 시간 t에서 로봇은 $I^t_{rgb} \in R^{H\times W\times 3}$ RGB 이미지, $I^t_{depth} \in R^{H\times W}$ 깊이이미지, $p_t=(x,y,z,\theta)$ 위치 및 자세 등의 관측을 환경으로부터 얻는다.

로봇은 전진, 우회전, 좌회전, 정지와 같은 이산 행동 공간에서 다음 행동을 선택한다. 예를 들어, 전진은 0.25 m 전방 이동, 좌회전과 우회전은 각각 반시계와 시계 방향으로 30° 회전을 수행한다. 탐색을 통해 주어진 목표 객체를

발견하고 충분히 접근하면 정지를 선언한다. 정지가 선언된 시점에서 로봇의 위치와 실제 목표 객체 위치의 유클리드 거리 $d(p_t, g)$가 임계값 (예: $\epsilon = 0.1$ m) 이하일 때, 해당 에피소드는 성공으로 인정된다. 거리가 임계값 (예: $\epsilon > 0.1$ m) 이상이거나, 잘못된 물체를 찾은 경우, 주어진 Step 수 (예: 500 Step)를 넘어서게 되면 실패로 간주한다.

2.2 Evaluation Metrics

본 절에서는 Object Goal Navigation(OGN)에서 에이전트의 성능을 평가하기 위해 사용되는 정량적 지표들에 관해 설명하고자 한다. 그중에서도 가장 대표적인 SR (Success Rate), SPL(Success weighted by Path Length), Soft-SPL에 대해 알아본다.

2.2.1 SR(Success Rate)

에이전트가 에피소드 내에서 목표 객체로부터 거리 임계값 ϵ 이하에 도달한 상태에서 정지 행동을 수행했는지를 기준으로 성공 여부를 판단한다. 전체 에피소드 N 중 성공한 에피소드 수에 대한 비율로 식 (1)로 정의된다. 여기서 1_i는 i번째 에피소드의 성공 여부를 나타낸다.

$$\mathrm{SR} = \frac{1}{N}\sum_{i=1}^{N} 1_i \tag{1}$$

2.2.2 SPL(Success weighted by Path Length)

성공 여부와 경로 효율성을 함께 고려하는 지표로 이상적인 최단 경로 길이 l_i에 비해 실제 에이전트가 이동한 경로 길이 p_i가 얼마나 효율적인지를 고려한다. 성공한 경우에만 가중치가 부여되며, 식 (2)로 정의된다. 여기서 1_i는 i번째 에피소드의 성공 여부를 나타낸다.

$$\mathrm{SPL} = \frac{1}{N}\sum_{i=1}^{N} 1_i \cdot \frac{\ell_i}{\max(p_i, \ell_i)} \tag{2}$$

2.2.3 Soft-SPL

SPL은 에이전트가 정확한 거리 임계값 내에 도달했을 때만 간주한다는 단점을 가진다. 이를 보완하여 연속적인 거리 기반 가중치를 도입한 변형 메트릭이다. 식 (3)으로 정의가 되며 실패한 에피소드에도 부분 점수를 반영한다.

$$\text{Soft-SPL} = \frac{1}{N}\sum_{i=1}^{N} \frac{\ell_i}{\max(p_i, \ell_i)} \tag{3}$$

2.3 Dataset

OGN 연구는 대규모 실험과 재현성 확보를 위해 시뮬레이션에서 진행하는데, 대표적으로 Habitat,[6] AI2-THOR[7]에서 이뤄진다. 실제 공간을 디지털화한 여러 데이터셋이 표준 벤치마크로 사용되고 있으며, 본 절에서는 각 데이터셋의 특성과 사용 경향을 정리한다.

Table 1 Comparison of object goal navigation datasets

Metric	MP3D	Gibson (4+)	HM3D
Number of scenes	90	571 (106)	1,000
Floor area (m^2)	101.82 k	217.99 k (17.74 k)	365.42 k
Navigable area (m^2)	30.22 k	81.84 k (7.18 k)	112.50 k
Navigation complexity	17.09	14.25 (11.90)	13.31
Scene clutter	2.99	3.14 (3.04)	3.90

MP3D[8]는 90개의 실제 건물을 3D 스캔을 통해 고품질로 구축한 데이터셋이다. 사실적인 시각화 품질 덕분에 OGN 연구의 표준 데이터셋으로 가장 먼저 자리 잡았다. 복잡한 실제 주거 환경 구조, 다양한 조명 조건, 가구 배치 등으로 현재까지도 널리 활용되고 있다.

Gibson[9]은 570개의 실제 공간을 3D 모델로 재구성했다. MP3D가 시각적 사실성에 중심을 뒀다면 Gibson은 깨끗한 기하학적 정보를 제공하는 데 강점이 있다. 하지만 일부 스캔에서 벽이나 가구 표면에 구멍이 뚫려 있거나 얇은 의자 다리가 제대로 표현되지 않는 등 재구성 결함(Reconstruction artifacts)이 많다. 각 Scene의 품질을 1점 ~ 5점으로 평가하여 4점 이상을 받은 106개의 고품질 Scene만을 선별하여 사용하는 경우가 많다. 하지만 이 경우 Table 1과 같이 데이터셋의 규모가 MP3D보다 작아지는 한계가 있어 최근 연구에서는 이전 방법론과의 비교 목적으로 활용되고 있다.

HM3D[10]는 MP3D의 후속 데이터셋으로 현재 가장 널리 사용되고 있는 데이터셋이다. 1,000개의 주거 공간에 대한 고품질 3D 스캔을 포함하고 있으며 넓은 공간(Navigable area), 높은 품질로 Embodied AI의 학습과 일반화 성능을 검증하는 데 사용되고 있다.

Table 1에 제시된 수치에서 볼 수 있듯이, HM3D는 전체 환경 수, Floor area, navigable area 측면에서 다른 데이터셋에 비해 가장 넓으며, Scene clutter 수준도 가장 높다. 여기서 Scene clutter는 통행 가능한 영역 주변에 분포하는 실제 물체 표면의 면적 비율을 의미하며, 공간의 혼잡도를 반영한다. 반면 MP3D는 상대적으로 높은 탐색 복잡성(Navigation complexity)을 보이는데, 이는 두 지점 간 직선거리 대비 장애물들을 고려한 최단 경로가 얼마나 우회하는지를 나타냄으로써 환경의 구조적 복잡도를 의미한다. Gibson은 비교적 작은 규모로 간단한 실험에 적합하다.[10] 이처럼 각 데이터셋은 규모, 복잡도, 시각적 다양성에서 서로 다른 특성을 보인다.

2.4 The Common Core of Diverse Embodied AI Tasks

Object Goal Navigation(OGN)을 비롯한 다양한 비전-언어 기반 Embodied AI 태스크들은 서로 다른 입력과 목표 정의 방식을 사용하지만, 구조적으로는 “목표 지시와 시각적 관찰을 통해 공간 내에 의미 있는 위치에 도달하는 것”이란 공통된 문제 구조를 가진다. 본 절에서는 OGN과 유사한 구조적 특성을 가지는 대표적인 벤치마크들을 소개하고 어떻게 유사한 특성을 공유하는지 살펴본다.

GOAT[11]는 OGN를 포함하면서도 서로 다른 방식으로 목표가 주어지는 ImageNav, TextNav을 포함한 목표 지향 탐색 태스크(Goal oriented navigation task)를 하나의 벤치마크에서 종합적으로 평가한다. 결과적으로 OGN은 GOAT 내 하위 태스크로 포함되며 객체에 도달하는 것이 목적이지만, GOAT는 다른 태스크들도 종합적으로 평가하기에 더욱 일반화된 정책 학습과 크로스모달 추론 능력을 요구한다.

REVERIE[12]는 복잡한 자연어 지시를 따라 에이전트가 실내 환경을 탐색하고 목표 객체를 정확하게 지목해야 하는 지시 기반 시각 탐색 태스크다. 목표는 OGN처럼 명시적인 클래스 명이 아니라 문장 속 ‘텍스트 정보’에 암묵적으로 제공된다. 따라서, OGN과 같이 객체 중심의 도달 목표를 가지지만, REVERIE는 긴 지시문에 대한 문맥 이해 능력과 Fine-grained grounding 능력을 더 요구한다.

Visual Question Answering(VQA)[13]는 정적 이미지에 기반으로 한 시각적 질의응답 태스크이며, Embodied Question Answering(EQA)[14]은 실내 환경에서 주어진 질문에 답하기 위해 탐색을 수행해야 한다. EQA는 로봇이 능동적으로 환경을 이동하며 정보를 획득하고 그에 따라 질문의 답을 생성한다. 이는 OGN처럼 ‘탐색 기반 시각적 정보 수집’을 포함하는 점에서 유사하나 최종 출력이 공간적 도달이 아닌 언어적 응답이라는 점에서 차이가 있다.

이처럼 GOAT, REVERIE, EQA 등의 태스크는 각각 다른 입력 및 출력 구조를 가지지만, 다음과 같은 공통 구조를 OGN과 공유한다.

(1) 목표 지시가 존재한다 - 명시적(Class label), 예시적(Image), 또는 서술적(Natural language)의 형태로 주어지는 목표가 존재하며, 이는 다음 행동 선택의 기준이 된다.

(2) 탐색 기반의 시각적 추론(Reasoning)이 필요하다 - 정답을 얻기 위해 환경 내에서 시각 정보를 능동적으로 수집하고 판단하는 절차가 포함된다.

(3) 멀티모달 입력을 처리한다 - RGB, depth, text 또는 Pose 정보 등 다양한 Modality가 통합되어 입력된다.

따라서 위의 태스크들은 독립적으로 보이지만, 근본적으로는 ‘언어・시각・공간 정보의 통합과 목표 지향적 의사결정’이라는 공통된 문제 구조를 해결하고 있으며 OGN은 정보 통합을 바탕으로 효율적인 탐색에 비중을 둔 과제라고 할 수 있다.

2.5 Before the Advance of VLM/LLM

VLM/LLM이 활용되기 전의 OGN 연구는 주로 시각 관측을 활용하여 정책 학습을 통해 해결하려는 시도를 해왔다. RGB와 Depth를 기반으로 강화학습[1,2]과 모방학습[15]을 통해 정책을 End-to-end로 학습하는 방식이 주류를 이뤘다. 대표적으로, Habitat-Web[15]는 인간 시연을 활용하여 모방학습을 통해 탐색 전략을 구성하였지만, VLM/LLM의 언어적 추론이나 일반적인 상식들은 포함되지 않았다.

그러나 End-to-end 학습 기반의 접근은 정책 안정성과 실세계 환경에서의 일반화 측면에서 제약이 많았다. 이에 따라 점차 지각(Perception), 매핑(Mapping), 계획(Planning) 으로 분리하여 처리하는 모듈형(Modular)[16-20] 방식으로 전환됐다. 모듈형 방식은 RGB와 Depth를 바탕으로 정밀한 2D 맵을 생성한다. 이 지도를 바탕으로 목표 객체의 존재 가능성이 높은 위치를 예측하거나 탐색 우선 순위를 결정하는 방식을 구성한다.

SemExp[16]는 객체가 존재할 가능성이 큰 지역을 선택적으로 확장하는 탐색 정책을 제안했다. 그러나 이 접근은 Semantic label의 정확성에 의존하는 구조였기 때문에 객체 인식 오차가 탐색 실패로 이어지는 경우가 잦았다.

이를 보완하기 위해 목표 객체에 대한 시각 정보가 불완전할 경우 Depth 정보를 바탕으로 3D 점군 데이터로 복원하여 객체 식별 결과를 보정하여 정확도를 높이는 방식도 제안되었다.[17] SSCNav[18]는 보이지 않는 영역의 Semantic 정보를 추정하고 Confidence prediction을 통해 신뢰도를 평가하는 방식을 택했다. 또한, PONI[19]는 잠재함수를 통해 각 프론티어 별로 목표 객체의 존재 잠재성을 바탕으로 탐색 전략을 제안했다. 한편, PEANUT[20]은 목표 객체의 잠재 위치를 셀 단위로 추론하였지만, 여전히 VLM/LLM의 추론과 외부 상식은 활용되지 않았다.

요약하자면, VLM/LLM 이전에는 OGN은 시각 기반 지도 생성과 탐색 정책의 학습을 중심으로 발전해 왔으며 복잡한 의미론적 추론이나 고수준 지식 주입은 고려되지 않았다. 이는 이후 OGN이 VLM/LLM의 융합을 통해 더 추론 지향적이고 지식 기반으로 변화하게 되었다.

2.6 Survey of Surveys

본 절에서는 OGN에 대해 자세히 서술되어 있는 기존 서베이 논문들을 소개하고, 본 논문의 차별성을 정리한다.

우선, Ieong과 Tang[21]은 OGN 외에도 앞서 언급된 ImageNav, TextNav와 더불어 PointNav, Multi OGN, Sound goal navigation 등 다양한 Goal oriented task를 포괄적으로 다룬다. 이 서베이는 각 태스크에서 활용한 Latent space를 기반으로 분류하여 소개했다. 서베이의 OGN 태스크 경우에는 Map, graph, implicit representation, linguistic, CLIP embedding, BLIP embedding, diffusion으로 분류했다. 그러나 VLM/LLM의 역할과 지식 주입 단계에 대한 논의는 부재하며 본 논문과 같이 Perception-Prediction-Planning의 어느 단계에서 통합되는지에 대한 분석이 부족하다.

반면, Sun 등[22]은 구체적으로 OGN에 초점을 맞춰 End-to-end 방식, Modular 방식, Zero-shot 방식으로 논문들을 분류하여 서술하고 있다. 각 방법론에 대한 설명은 구체적이지만 VLM/LLM 기반 지식 주입 위치에 대한 관점은 없었다.

위와 같이, 기존 서베이들은 VLM/LLM의 지식 활용을 구조적으로 분류하고 이를 Perception-Prediction-Planning 관점에서 비교하는 시도는 찾을 수 없었다. 그런데도 OGN 연구에서는 VLM/LLM의 지식을 통합하는 개별 사례들이 다수 존재하여, ‘어디에 지식을 융합할 것인가?’라는 질문을 중심으로 자율주행의 대표적 구조인 Perception-Prediction-Planning 기반으로 OGN의 논문들을 살펴봄으로써 VLM/LLM의 역할에 대해 논의하고자 한다.

3. Where to Fuse?

로봇이 처음 보는 환경에서 사용자가 지정한 목표 객체를 효율적으로 탐색하는 태스크인 OGN에서 VLM/LLM을 활용하는 연구는 꾸준히 증가하고 있다. 하지만 VLM/LLM의 지식을 OGN의 어느 단계에서 주입하는지에 대한 분석은 미비하여 본 논문에서는 VLM/LLM이 활용되는 시점과 역할에 따라 Perception, Prediction, Planning의 구조로 논의한다.

3.1 Perception Level Fusion

Perception 단계는 로봇이 현재의 관측 장면에 대한 시각적, 공간적 정보를 이해하는 단계이다. ‘지금 앞에 무엇이 보이는가?’라는 질문에 답하는 과정으로 VLM의 멀티모달리티를 활용하여 현재 시점의 객체 및 공간 정보를 변환하는데 핵심적인 역할을 한다. 이 정보는 이후, Prediction과 Planning 단계에서 사전지식으로 활용된다. VLM을 통해 장면에 있는 객별 객체를 이해하는 Scene object understanding과 객체 간의 공간적인 배치와 장면 내의 공간적 특성을 인식하는 Scene spatial relations understanding에 대해 다룬다.

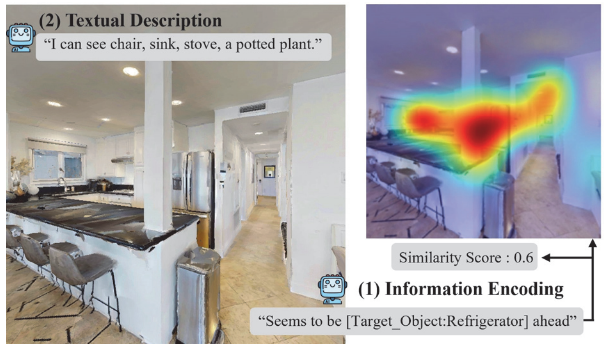

Fig. 3 Two VLM approaches for scene object understanding at the perception level. (1) Information Encoding, VLM encodes information between the target object label and image, often quantifies a similarity between the visual input and the target object. (2) Textural Description directly generates a natural language description of objects visible in the current view

3.1.1 Scene Object Understanding

현재 시점에서 관측 가능한 객체들을 정확하게 습득하는 것을 목표로 한다. VLM을 통해서 ‘현재 어떤 객체들이 보이는가?’라는 질문에 답하며 기존의 OGN이 사전에 정의된 객체 집합인 Closed-set에 묶여 있었던 것과 다르게 VLM을 활용함으로써 Open-set 탐지가 가능해진다. Fig. 3과 같이 VLM을 활용한 방식은 크게 두 가지로 나눌 수 있다.

(1) VLM을 Information encoding 용도로 활용

(2) VLM을 Textual description 용도로 활용

Information encoding은 VLM의 Image-text alignment 능력을 활용하여 이미지와 텍스트의 관계를 점수화 혹은 임베딩을 통해 정보를 압축시키는 인코더로 활용된다. CoW[23]는 CLIP의 유사도 점수를 Grad-CAM 기반을 통해 목표 객체의 위치를 추정한다. 한편 ZSON[24]은 CLIP 임베딩을 활용하여 정책 학습에 활용하는데 학습 과정에서는 목표 객체의 Label 대신 목표 객체의 Image의 CLIP embedding을 통해 학습하고, 실제 OGN 실험에서는 Image embedding 대신 Label의 Text embedding으로 변경하여 사용한다. ESC[25]는 GLIP을 이용해 관측 이미지 내의 객체들을 알아낸다. GAMap[26]는 GPT4로부터 ‘목표 객체의 기능적 속성을 설명하라’와 ‘목표 객체의 형상적

특징을 말하라'의 결과를 Text embedding으로 활용하여 현재 관측 이미지를 CLIP으로 유사도를 2D 맵에 반영하는 방식을 제안하였다.

한편, BLIP-2를 통해 이미지와 목표 텍스트 간 코사인 유사도를 활용한 예시도 많다. 대표적으로 VLFM[27]은 유사도값을 2D맵에 직접 반영하였으며, ApexNav[28]는 이를 확장해 LLM을 통해 목표 객체와 외관이 유사하여 Detector가 오탐지할 수 있는 객체 정보를 전달함으로써 오탐지를 줄이는 데 활용하였다. SemNav[29]는 BLIP-2 대신 GPT4를 동일한 목적을 달성하였다. UIAP[30]는 VLFM[27]에서 사용한 하나의 Prompt가 아닌 LLM을 통해 비슷한 여러 Prompt를 각각 BLIP-2로 유사도 점수를 산출하고, 그 분산을 불확실성 지표로 도입하였다. LLM-ZSON[31]에서는 연속적인 5장의 관측 이미지를 LLaVA로 요약하게 하고 이를 목표 객체와 유사도를 탐색에 활용한다. WMNav[32] 360° 파노라마 사진을 Gemini 이용하여, 각 시야각에서 새롭게 얻을 수 있는 정보량을 점수화하도록 하였다.

Textural description은 VLM을 통해 직접 이미지를 자연어로 설명하게 함으로써 객체 정보를 확보한다. PixNav[33]는 LLama-adapter를 통해 관측 이미지 내의 객체를 직접 인식하는 방식을 취했다. 유사하게, OpenFMNav[34]는 GPT4를 활용해 같은 결과를 얻었다. CL-CoTNav[35]는 Qwen으로 현재 관측 이미지에서 목표 객체와 연관성이 있을 하위 목표 객체(Subgoal object)들을 중점으로 인식한다. VoroNav[32]는 보르노이(Voronoi)의 형태로 2D 맵을 구성하는데, WMNav[32]과 유사하게 하나의 노드에 달성하면 360° 회전을 통해 노드를 업데이트하고 각 노드에서 관측 이미지를 BLIP을 활용하여 관측 이미지를 설명하도록 한다.

이처럼 Scene object understanding 단계는 VLM을 활용하여 객체 탐지를 넘어 목표 객체와의 관계를 정량적으로 평가하거나 관찰 장면을 자연어 서술함으로써 이후 단계의 사전지식으로 활용된다.

3.1.2 Scene Spatial Relations Understanding

이는 '무엇이, 어디에, 어떻게 놓여 있는가?'라는 질문을 통해 객체 간의 물리적인 배치와 공간적 구조를 이해하는 단계이다.

PixNav[33]와 CL-CoTNav[35]는 질의응답(QA)의 형태로 객체 간의 공간적인 배치에 대한 정보를 얻는다. SG-Nav[37]는 그래프(Graph)를 통해 객체 간의 관계정보를 저장하는데 LLM의 상식으로 객체 간의 일반적 관계를 먼저 가정하게 한 후, VLM을 통해 현재 장면에서 그 관계가 실재하는지 시각적으로 검증하여 장면 그래프(Scene graph)를 구축한다.

이런 Spatial relation은 정적인 객체를 넘어 로봇의 통행 가능성(Traversability) 판단으로 확장할 수 있다. 다수의 방식이 정교한 2D맵으로 통행 가능성을 판단했던 것과 달리 DyNaVLM[38]은 VLM을 이용하여 직관적으로 평가하도록 한다. Fig. 7(a) 같이 현재 RGB 이미지 상에 주행 가능한 후보 지점들을 직접 투영하고, VLM 에게 '이 경로로 이동했을 때, 이동 중에 충돌이 발생할 가능성이 있는가?'를 질의한다. VLM은 시각적 이해 능력을 바탕으로 충돌 위험이 있는 지점들을 식별하고 해당 선택지는 후보에서 배제한다. 이는 2D 맵의 변환 과정 없이 시각을 통한 공간 이해와 행동 계획을 직접 연결하는 방식이다.

나아가, Spatial relation understanding의 대상은 정적 객체에서 동적 객체인 사람으로까지 확장되어 Social navigation 문제에 적용될 수 있다. Social-LLaVa[39]와 같은 연구는 VLM을 통해 장면 속 사람들의 위치, 자세, 시선 방향과 같은 사회적 단서들을 파악한다. 이를 바탕으로 물리적 충돌을 피하고 사람들의 개인 공간을 존중하며 통행할 수 있는 경로를 계획한다.

결론적으로 객체의 상대적 배치, 에이전트의 주행 가능성, 그리고 동적 객체와의 상호작용을 종합적으로 다루는 Scene spatial relations understanding은 OGN의 경로 계획 효율성을 높이는 핵심 요소다. 더 나아가, 이러한 공간적 이해는 단순한 탐색을 넘어 에이전트의 의사결정에 사회적 · 상황적 맥락을 반영하는 기반이 되며, 복잡한 실세계 환경에서의 적응적이고 안전한 탐색 전략 수립에 기여한다.

3.2 Prediction Level Fusion

이전 Perception 단계에서 얻은 단편적인 정보를 바탕으로, 직접 관찰되지 않은 의미론적인 정보들을 추론하는 과정이다. '이전까지의 관측을 통해 얻은 객체들은 어떤 관계가 있는가?', '이 공간은 어떤 장소인가?'라는 질문에 중점으로 답한다. 이는 본격적으로 LLM의 방대한 상식과 추론 능력을 활용한다. 객체 간의 일반적인 관계를 바탕으로 보이지 않는 목표 객체의 위치를 예측하는 Object relation construction과 관찰된 객체들을 통해 공간의 종류를 추론하는 Semantic layout construction에 대해 논의한다.

3.2.1 Object Relation Construction

'목표 객체 A는 어떤 객체 B 인근에 있을 가능성이 크다.'라는 질문과 같이 객체 간의 일반적인 관계를 알면 효율적인 탐색이 가능해진다. 예를 들어, Fig. 4에서 목표 객체가 '소파'라면 로봇은 '침대'나 '변기'보다는 'TV'를

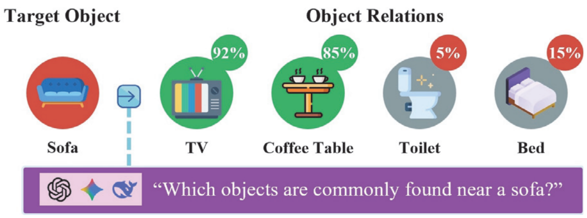

Fig. 4 An LLM is prompted to output common object co-occurrences. Given the target 'sofa', 'TV' and 'coffee table' are identified as highly related objects while 'toilet' and 'bed' are quantified as less related objects to 'sofa'

관측한 구역을 우선으로 탐색하는 것이 일반적이다. 이처럼 LLM이 가진 일반 상식을 통해 관측된 객체 간의 연관성을 평가하고 불필요한 탐색을 줄이는 데 결정적인 역할을 한다.

가장 직관적인 방식으로는 현재 관측 정보와 목표 객체 간의 관계를 LLM에 직접 질의하는 것이다. 이 접근법은 간단한 질문부터 점차 정량적인 확률을 도출하는 방향으로 변했다. ESC[25]는 관측 객체들과 목표 객체를 LLM에 제시하며 '두 객체가 함께 존재할 가능성이 있는가?'를 통해 관계를 파악한다. L3MVN[40]은 한 걸음 나아가 '관측 객체 1, 관측 객체 2, 관측 객체 3이 있는 프론티어(Frontier)에 목표 객체가 있다'와 같은 문장의 완성도를 LLM이 평가하도록 하여 간접적으로 객체 간의 관계를 평가한다.

한편, 관계를 직접 묻는 대신에 목표 객체와 함께 있을 법한 객체 목록을 생성하고 이를 탐색의 단서로 활용하는 방식도 있다. BeliefMapNav[41]과 CL-CoTNav[35]는 목표 객체와 연관성이 높은 객체 목록을 LLM으로 하여금 생성하게 한다. 이 목록에 있는 객체를 발견하면 해당 구역에 집중하여 탐색하도록 유도한다. 이와 유사하게, CogNav[42]는 목표 객체와 연관된 객체를 발견했을 때 해당 방향으로 탐색을 유도하는 방향 지표로 사용한다.

아직 보지 못한 영역을 예측하여 탐색 효율을 극대화하는 연구도 있다. SGM[43]은 객체 간의 연관성인 일반 상식을 활용하여 미관측 영역의 2D 맵을 MAE[44]를 통해 생성함으로써 선제 탐색을 가능하게 한다.

LLM을 통해 물체 간의 일반적인 관계정보를 사전에 인지함으로써 로봇이 집안 내의 탐색 공간을 줄인다. 이는 불필요한 탐색을 줄이게 되고, 나아가 효율적으로 탐색을 이어나갈 수 있다.

3.2.2 Semantic Layout Construction

현재까지의 단편적인 관측을 바탕으로 '이 공간은 어

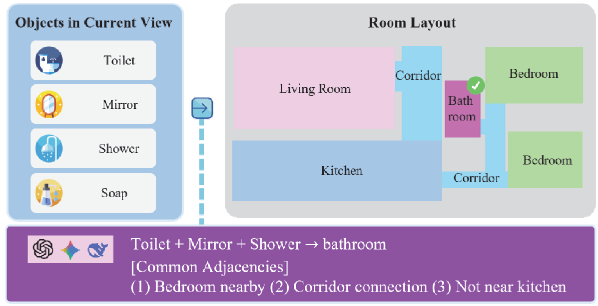

Fig. 5 Based on the objects in the current view such as 'toilet', 'mirror' and 'shower'. LLM infers the room type as 'bathroom' and predicts the layout of unobserved areas by reasoning about common room layout adjacencies

떤 장소인가?'를 판단한다. 예를 들어, Fig. 5와 같이 현재 시야에서 세면대와 변기를 보았다면, 해당 공간이 화장실일 확률이 높다고 추론할 수 있다. 나아가 '화장실은 보통 침실이나 복도에 인접하며 주방 근처는 아니다'와 같은 공간 관계에 대한 정보도 추론할 수 있다. 이처럼 전체적인 집의 구조를 예측은 불필요한 탐색을 줄이고 다음 행동을 결정하는데 중요한 단서가 된다.

ESC[25]와 PixNav[33]는 관측 객체 정보를 바탕으로 현재 어떤 방에 있는지를 LLM이 직접 도출한다. E2BA[45]는 나아가, 각 프론티어(Frontier) 에서 관측된 객체 정보를 바탕으로 해당 방의 종류를 추론하고 목표 객체의 연관성도 함께 고려한다.

CL-CoTNav[35]는 목표 객체가 있을 공간을 예측하고, 실제 관측 공간이 예측과 일치하는지를 검증하며 탐색한다. 반면에, CogNav[42]는 로봇의 탐색 단계를 5가지로 분리하였는데, 목표 객체와 연관성이 높은 공간을 발견하면 그 공간을 우선적으로 탐험하도록 유도한다.

그 외에도, 현재 위치를 넘어 전체적인 구조를 이해하는 방법으로 TopV-Nav[46]는 2D 맵을 Top-view에서 이해하며, BeliefMapNav[41]은 현재 방 뿐만 아니라 인접한 방이 무엇인지까지 추론하였다. DAR[47]은 미관측 영역을 생성하여 활용한다는 점에서 SGM[43]과 유사하다. 하지만, SGM[43]은 객체 간의 연관성을 활용하였고 DAR[47]은 LLM을 통해 각 프론티어가 어떤 방인지 분류와 그 방에 있을 만한 객체를 제시하여 확산 모델로 하여금 해당 방에 있을 법한 객체에 초점을 맞춰 2D 맵을 채워가도록 유도한다.

3.3 Planning Level Fusion

이전 단계에서 얻은 정보들을 종합하여 '이제 무엇을 해야하는가?'라는 질문에 대한 최적의 행동을 수립하는

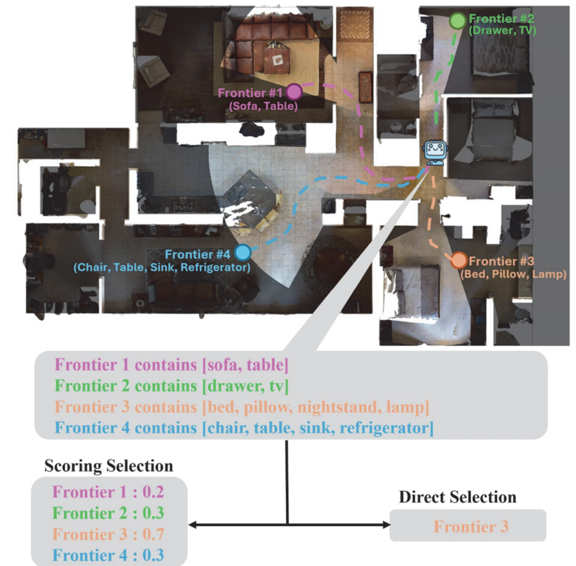

Fig. 6 Two distinct strategies for Frontier Selection. Scoring Selection, LLM assigns a quantitative score to each frontier based on the semantic relevance of the objects it contains. Direct Selection LLM acts as a high-level decision maker directly choosing the next best frontier on all available context

단계이다. 본 절에서는 여러 프론티어(Frontier) 중에서 하나를 선택하는 Frontier selection과 로봇이 현재의 결정을 재검토하는 Re-validation으로 나누어 설명한다.

3.3.1 Frontier Selection

탐색 과정에서 로봇은 아직 가보지 않은 미지의 영역 경계선, 즉 여러 개의 프론티어(Frontier)에 직면하게 된다. 이때 VLM/LLM은 이전 단계에서 얻은 정보를 바탕으로 목표 객체를 발견할 가능성이 가장 크다고 판단되는 프론티어를 선정하는 고수준의 의사결정자 임무를 수행한다. VLM/LLM에게 정보를 제공하는 방법에 따라 자연어로 제공하는 Linguistic method와 이미지로 제공하는 Visual method로 나눌 수 있다.

Linguistic method는 Fig. 6과 같이 현재까지 축적된 정보를 프롬프트의 형태로 VLM/LLM에 전달하여 최적의 프론티어를 선정하는 방식이다. 각 프론티어의 가치를 정량적인 점수로 변환하여 평가하는 ‘Scoring selection’은 가장 높은 점수를 받은 곳을 다음 목적지로 정하고 탐색한다. OpenFMNav,[34] L3MVN,[40] TriHelper[48]들은 각 프론티어에서 관측된 객체 목록을 LLM에 전달하여 점수를 산출하는 방식을 택했다. VoroNav[36]은 객체 목록뿐만 아니라 탐색을 통해 얻은 방의 종류, 객체 간의 위치정보를 제공하여 점수화에 함께 반영하도록 했다. 한편,

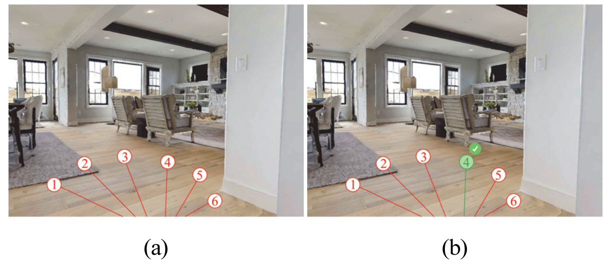

Fig. 7 Visual method for frontier selection. (a) Candidate waypoints are projected onto the egocentric view, (b) Candidate waypoints are directly evaluated and selected for the next waypoint

SG-Nav[37]는 점수화 과정에서 각 프론티어까지의 실제 거리를 함께 고려하여 의미적으로 유망하면서도 효율적으로 도달 가능한 프론티어를 선택하도록 한다. TopV-Nav[46]는 주변 객체 정보가 텍스트로 저장된 2D 맵을 VLM에 입력하여 각 위치의 목표 객체 존재 확률을 계산하고, 이를 가우시안 분포 기반으로 통합하여 탐색할 프론티어의 우선순위를 정한다.

명시적인 점수화 과정 없이 최적의 프론티어를 LLM이 직접 결정하여 탐색하는 ‘Direct selection’도 있다. E2BA[45]와 PixNav[33]는 축적된 정보를 LLM이 다음 프론티어를 직접 선정하도록 한다. LLM-ZSON[31]은 의미론적 가치 맵과 환경 구조가 담긴 맵을 함께 VLM에 제공하여 다음 프론티어를 선정한다.

Visual method는 Linguistic method 방식과 달리 VLM의 시각적 이해 능력을 활용하는데 VLMNav,[49] DyNaVLM,[38] WMNav[32]는 Fig. 7(a)와 같이 현재 로봇의 관측 이미지에 실행할 수 있는 선택지를 투영한 이미지를 VLM에 제시하여 Fig. 7(b)와 같이 가장 합리적인 선택지를 결정한다.

이처럼 VLM/LLM을 단순한 지식 제공으로 활용하는 것을 넘어 사전 지식을 활용하여 로봇의 다음 행동을 결정하도록 활용하고 있다.

3.3.2 Re-Validation

탐색 과정에서 VLM/LLM의 결과를 그대로 신뢰하기보다 결과를 재검토하고 탐색 전략을 동적으로 수정하여 가장 효율적으로 목표 객체를 탐색해야 하는 OGN 태스크에서 빠르게 목표 객체 도달이 가능하게 한다.

일부 연구는 Fig. 8(a)와 같이 관측 객체를 명시적으로 재검증하여 오탐지를 줄이는 데 집중한다. CogNav[42]는 탐색 단계를 5가지로 분리하여 실제 목표 객체와 일치하는지를 재확인한다. TriHelper[48] 역시 현재 시야 내에 목표 객체가 실제로 존재하는지를 이진 분류로 다시 확인함으로써 오탐지를 줄인다. SG-Nav[37]는 목표 객체를 한 번 포착

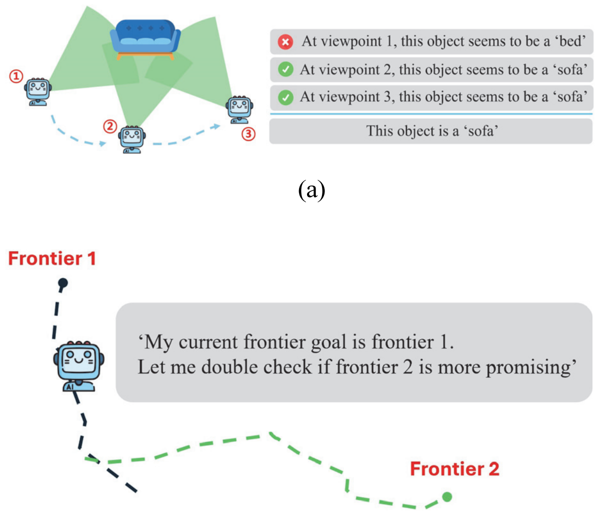

Fig. 8 Re-validation strategies at the planning-level for robust navigation. (a) The agent accumulates evidence from multiple viewpoints to confirm the identity of a potential target object, reducing false positives, (b) The agent continuously reassesses its chosen path in relation to other frontiers, thereby enabling dynamic backtracking or replanning

하였다고 바로 목적지로 택하는 것이 아닌 여러 시점에서 관찰하며 얻은 누적 신뢰도를 바탕으로 견고성을 높인다.

다른 한편으로는, Fig. 8(b)와 같이 현재 탐색 전략을 동적으로 수정하며 상황에 맞게 전환하는 연구도 있다. CL-CoTNav[35]는 Chain of Thought(CoT)를 통해 LLM의 계획을 바로 수행하지 않고, 해당 계획의 신뢰도를 스스로 평가하고 반영하는 자기 점검을 도입했다. 마지막으로, E2BA[45]는 탐색을 하며 기록해둔 과거의 이동 궤적과 객체들을 바탕으로 LLM으로 하여금 이전 위치로 돌아가는 백트래킹이 더 효율적인지 판단함으로써 동적으로 계획을 수정한다.

이처럼 로봇이 VLM/LLM을 통해 복잡하고 불확실한 환경에서 즉각적인 판단에만 의존하지 않고 유연하게 전략을 수정하는 것은 필수적이다. 이는 실제 환경에서 로봇의 탐색 실패를 VLM과 LLM으로 대처하는 사례이다.

4. Open Challenges and Research Opportunities

본 절에서는 OGN이 실제 세계 적용을 위해 해결해야 할 도전과제를 소개하고, 현재 데이터셋의 한계와 함께 Multi-floor navigation과 OGN with humans present 등 앞으로의 연구 기회에 대해 논의하고자 한다.

Fig. 9 An example of reconstruction artifacts (or defects) found in real-world scan datasets. The depth image (right) shows significant data loss also called 'holes' or 'black cracks'

4.1 Dataset Limitation and Opportunities

4.1.1 Issues in Existing Datasets

OGN에서 자주 활용되는 HM3D, MP3D와 같이 3D 스캔을 활용해 복원한 데이터셋은 실제 실내 환경과 같은 높은 사실성을 제공한다. 하지만 Fig. 9와 같이 3D 스캔 과정에서 발생하는 데이터 누락이나 기하학적 오류로 인해 일부 관측 이미지에서 음영 영역이 존재한다. 이런 결함은 주변 환경을 2D 맵으로 구축하는 접근법에서 부정확한 지도를 생성하게 하며 탐색 실패를 유발한다.

또한, 시뮬레이션에서 실험을 진행하기 위해서는 두 가지 데이터셋이 활용되는데 첫 번째는, HM3D 같은 3D 스캔 정보를 내포하고 있는 Scene dataset과 각각의 에피소드에서 로봇의 시작 위치, 목표 객체의 label과 위치 등의 정보를 내포하고 있는 Episode dataset이 있다. 하나의 환경(Scene) 내에 같은 Class의 객체가 여러 개 존재하더라도, 단 하나만을 정답으로 간주한다.

예를 들어, 하나의 환경(Scene)에 화분이 3개 있다면, 이 중에서 화분_1만이 정답으로 간주된다. 이로 인해 에이전트가 의미론적으로는 동일한 다른 객체에 도달하여도 실패로 처리된다. 이는 현실 세계의 성공 기준과 동떨어져 있다.

4.1.2 Emerging Dataset

위와 같은 한계를 극복하기 위한 데이터셋으로는 HSSD[50]가 있다. HSSD는 실제 환경과 같은 높은 품질의 시각적 정교함과 구조적 복잡성을 제공하며 데이터 생성과 확장이 쉽다. HSSD에서 사전 학습된 모델이 HM3D나 MP3D같은 3D 스캔 기반 데이터셋에서 보다 더 우수한 제로샷 성능을 보이는 결과도 보고되었다.[50] 이는 데이터의 다양성과 대규모 학습의 중요성을 시사하며 앞으로 HSSD의 활용이 점차 확대될 것으로 기대한다.

4.2 Expanding OGN Task

4.2.1 Beyond Single Floor

HM3D와 MP3D는 대부분 다층 주택의 스캔을 포함하고 있는데 몇몇 에피소드에서는 로봇의 시작 층과 목표 객체가 위치한 층이 다른 경우가 존재한다. 다수의 연구는 주변 환경을 투영한 2D 맵을 많이 활용했다. 결과적으로 계단을 활용하여 다른 층으로의 이동을 제한되었다. 이런 한계점에서 ASCENT[51)]와 MFNP[52)]는 계단 이동을 적극적으로 활용하여 다층에서의 OGN을 가능하도록 한다.

4.2.2 Object Goal Navigation with Humans Present

기존의 OGN 연구는 주로 사람이 없는 환경에서 가장 효율적으로 목표 객체에 도달하는 것에 초점을 두었다. 최근에는 Habicrowd,[53)] Social-HM3D[54)] 그리고 Social- MP3D[54)]와 같이 시뮬레이션 환경에 동적으로 움직이는 휴머노이드 아바타를 도입하여 실생활 환경에서의 연구도 이뤄지고 있다. 그러나 이런 시뮬레이션 기반의 OGN 연구들은 여전히 실세계에서 사람의 자연스러운 표현을 이해하거나 사람과의 상호작용을 고려하는 연구는 미비하다.

실세계에서 네비게이션 연구는 주로 Social compliant navigation에 대해 진행되었다.[55-57)] Social compliant navigation은 먼저 다가오는 사람에게 길을 양보하는 것과 같은 기본적인 사회적 규범을 준수하며 목적지에 도달하는 연구를 이른다. 이런 측면에서 OGN 연구와 Social compliant navigation은 사람을 움직이는 장애물로 판단하고 충돌을 회피하면서 사람에게 양보 후 이동하는 것에만 초점을 맞췄다.

하지만 로봇이 거주 환경이나 사람이 있는 환경에서 성공적으로 작업을 수행하기 위해서는 단순히 기본적인 사회적 규범을 준수하는 것을 넘어 Fig. 10과 같이 능동적으로 상황을 파악하고 적절한 도움을 제공할 수 있는 Socially interactive navigation 연구가 필수적이다.

이러한 맥락에서, 로봇이 사람과의 능동적인 상호작용까지 고려한다면 로봇이 인간 사회 속에서 자연스럽게 공존하며 실생활 환경에서의 협력적 작업 수행 능력을 크게 향상하는데 이바지할 것으로 기대가 된다.

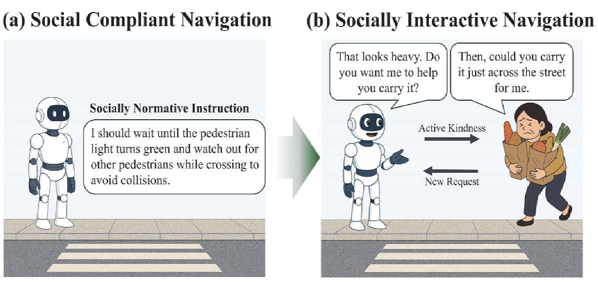

Fig. 10 Example of social compliant navigation and socially interactive navigation

4.2.3 Mixture of Indoor and Outdoor Environments

대부분의 OGN 벤치마크는 실내 거주 환경에 국한되어 있다. 병원, 물류 창고, 대학 캠퍼스와 같이 실내외를 넘나들어야 하는 실제 응용 시나리오와는 틈이 있다. 실내외 복합 환경에서의 네비게이션을 위해서는 실외에서는 GPS를 실내에서는 WiFi나 비전 기반 등 위치 추정의 정확도가 달라지며, 상황에 맞는 센서 융합 전략을 필요로 한다. 구조화된 실내와 달리 개방된 야외 공간은 탐색해야 할 영역이 방대하고 구조적 단서가 적어 새로운 전략을 요구한다. 또한, 같은 객체 label이 실내에서는 '화분'을 실외에서는 '나무'를 의미할 수 있는 것처럼 문맥에 맞는 의미 해석 능력이 필요하다.

5. 결 론

본 논문은 OGN에서 VLM과 LLM의 지식을 활용하는 전략을 자율 주행의 Perception-Prediction-Planning 구조를 바탕으로 체계적으로 분석했다. Perception 단계에서는 VLM을 활용하여 Open-set 객체 인식과 공간 관계 이해를 통해 실시간 관측 정보를 변환하는 방법을 살펴보았다. Prediction 단계에서는 일반 상식을 바탕으로 객체간 관계와 공간 구조를 예측하는 접근 그리고 Planning 단계에서는 중앙 의사 결정자로 활용하여 최적의 탐색 경로를 선택하고, 동적으로 수정하는 방법들을 소개하였다.

또한, 현재 OGN 벤치마크 평가를 위해 주로 사용 되는 데이터셋이 가지는 구조적 한계를 지적하고 이를 극복하기 위한 대안의 인조 데이터셋에 대하여도 소개했다. 나아가 OGN의 연구 범위를 확장할 미래 방향성으로는 다층 구조에서의 OGN, 사회적 규범을 준수하면서 사람과 공존하는 OGN 그리고 실내외 혼합 환경으로의 적용을 과제로 제시했다. 이는 OGN이 통제된 실내 환경의 '보물 찾기'를 넘어, 복잡한 실제 환경에서 인간과 상호작용하는 서비스 로봇 기술로 나아가기 위한 문제들이다.

결론적으로, Object Goal Navigation(OGN)에 있어서 VLM/LLM의 지식을 어느 단계에서 어떤 방식으로 융합하는지를 살펴보았다. 본 논문에서 제시된 과제들과 연구 방향을 통해, 더욱 지능적이고 실용적인 로봇 네비게이션 시스템이 실현되기를 기대한다.

후 기

본 연구는 산업통산자원부 한국산업기술기획평가원

자율주행기술개발혁신사업(과제명: 지정구역기반 Point to Point 이동 Lv.4 승합차급 자율주행 차량플랫폼 기술개발, 과제번호: 20014361)의 연구비 지원 및 과학기술정보통신부 대구경북과학기술원 기관고유사업(과제번호: 25-IT-03)의 연구비지원에 의해 수행되었습니다.

References

1) S. Zhang, X. Song, W. Li, Y. Bai, X. Yu and S. Jiang, "Layout-Based Causal Inference for Object Navigation," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10792–10802, 2023.
2) J. Ye, D. Batra, A. Das and E. Wijmans, "Auxiliary Tasks and Exploration Enable ObjectGoal Navigation," IEEE/CVF International Conference on Computer Vision, pp.16117–16126, 2021.
3) G. Min and J. Choi, "Development of SAE Level 3 Autonomous Driving Technology for Urban Environment Self-Driving," KSAE Spring Conference Proceedings, pp.454–455, 2020.
4) C. Park, H. Son, K. Min and Y. Shim, "Obstacle Perception and Prediction Method using Multi Sensors for Autonomous Vehicle," The Institute of Electronics and Information Engineers, 793–798, 2021.
5) J. B. Won, J. Seok, S. Lee and J. Yoo, "Software-in-the-Loop Simulation Implementation Methodology for Autonomous Vehicle System Development," Transactions of KSAE, Vol.33, No.11, pp.967–977, 2025.
6) M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh and D. Batra, "Habitat: A Platform for Embodied AI Research," IEEE/CVF International Conference on Computer Vision, pp. 9339–9347, 2019.
7) E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu, A. Kembhavi, A. Gupta and A. Farhadi, "AI2-THOR: An Interactive 3D Environment for Visual AI," arXiv preprint arXiv:1712.05474, 2017.
8) A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng and Y. Zhang, "Matterport3D: Learning from RGB-D Data in Indoor Environments," arXiv preprint arXiv:1709. 06158, 2017.
9) F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik and S. Savarese, "Gibson Env: Real-World Perception for Embodied Agents," IEEE Conference on Computer Vision and Pattern Recognition, pp.9068–9079, 2018.
10) S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang M. Savva, Y. Zhao and D. Batra, "Habitat- Matterport 3D Dataset (HM3D): 1000 Large-Scale 3D Environments for Embodied AI," arXiv preprint arXiv:2109.08238, 2021.
11) M. Khanna, R. Ramrakhya, G. Chhablani, S. Yenamandra, T. Gervet, M. Chang, Z. Kira, D. S. Chaplot, D. Batra and R. Mottaghi, "GOAT-Bench: A Benchmark for Multi-Modal Lifelong Navigation," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.16373–16383, 2024.
12) Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen and A. van den Hengel, "REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9982–9991, 2020.
13) S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick and D. Parikh, "VQA: Visual Question Answering," IEEE International Conference on Computer Vision, pp.2425–2433, 2015.
14) A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh and D. Batra, "Embodied Question Answering," IEEE Conference on Computer Vision and Pattern Recognition, pp.1–10, 2018.
15) R. Ramrakhya, E. Undersander, D. Batra and A. Das, "Habitat-Web: Learning Embodied Object-Search Strategies from Human Demonstrations at Scale," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.5173–5183, 2022.
16) D. S. Chaplot, D. P. Gandhi, A. Gupta and R. R. Salakhutdinov, "Object Goal Navigation Using Goal-Oriented Semantic Exploration," Advances in Neural Information Processing Systems, Vol.33, pp.4247–4258, 2020.
17) J. Zhang, L. Dai, F. Meng, Q. Fan, X. Chen, K. Xu and H. Wang, "3D-Aware Object Goal Navigation via Simultaneous Exploration and Identification," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.6672–6682, 2023.
18) Y. Liang, B. Chen and S. Song, "SSCNav: Confidence-Aware Semantic Scene Completion for Visual Semantic Navigation," IEEE International Conference on Robotics and Automation (ICRA), pp.13194–13200, 2021.
19) S. K. Ramakrishnan, D. S. Chaplot, Z. AlHalah, J.

Malik and K. Grauman, “PONI: Potential Functions for ObjectGoal Navigation with Interaction-Free Learning,” IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.18890–18900, 2022.
20) A. J. Zhai and S. Wang, “Peanut: Predicting and Navigating to Unseen Targets,” IEEE/CVF International Conference on Computer Vision, pp.10926–10935, 2023.
21) I. -T. Ieong and H. Tang, “Multimodal Perception for Goal-Oriented Navigation: A Survey,” arXiv preprint arXiv:2504.15643, 2025.
22) J. Sun, J. Wu, Z. Ji and Y. -K. Lai, “A Survey of Object Goal Navigation,” IEEE Transactions on Automation Science and Engineering, Vol.22, pp.2292–2308, 2024.
23) S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt and S. Song, “Cows on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation,” IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.23171–23181, 2023.
24) A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman and D. Batra, “ZSON: Zero-Shot Object-Goal Navigation Using Multimodal Goal Embeddings,” Advances in Neural Information Processing Systems, Vol.35, pp.32340–32352, 2022.
25) K. Zhou, K. Zheng, C. Pryor, Y. Shen, H. Jin, L. Getoor and X. E. Wang, “ESC: Exploration with Soft Commonsense Constraints for Zero-Shot Object Navigation,” International Conference on Machine Learning, pp.42829–42842, PMLR, 2023.
26) H. Huang, Y. Hao, C. Wen, A. Tzes and Y. Fang, “Gamap: Zero-Shot Object Goal Navigation with Multi-Scale Geometric Affordance Guidance,” Advances in Neural Information Processing Systems, Vol.37, pp.39386–39408, 2024.
27) N. Yokoyama, S. Ha, D. Batra, J. Wang and B. Bucher, “VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation,” IEEE International Conference on Robotics and Automation (ICRA), pp.42–48, 2024.
28) M. Zhang, Y. Du, C. Wu, J. Zhou, Z. Qi, J. Ma and B. Zhou, “ApexNav: An Adaptive Exploration Strategy for Zero-Shot Object Navigation with Target-Centric Semantic Fusion,” arXiv preprint arXiv:2504.14478, 2025.
29) A. Debnath, G. J. Stein and J. Kosecka, “SemNav: A Model-Based Planner for Zero-Shot Object Goal Navigation Using Vision Foundation Models,” arXiv preprint arXiv:2506.03516, 2025.
30) U. Bajpai, J. Rückin, C. Stachniss and M. Popovic, “Uncertainty-Informed Active Perception for Open Vocabulary Object Goal Navigation,” arXiv preprint arXiv:2506.13367, 2025.
31) J. Shi, S. Yagi, S. Yamamori and J. Morimoto, “LLM-Guided Zero-Shot Visual Object Navigation with Building Semantic Map,” IEEE/SICE International Symposium on System Integration (SII), pp.1274–1279, IEEE, 2025.
32) D. Nie, X. Guo, Y. Duan, R. Zhang and L. Chen, “WMNav: Integrating Vision-Language Models into World Models for Object Goal Navigation,” arXiv preprint arXiv:2503.02247, 2025.
33) W. Cai, S. Huang, G. Cheng, Y. Long, P. Gao, C. Sun and H. Dong, “Bridging Zero-Shot Object Navigation and Foundation Models Through Pixel-Guided Navigation Skill,” IEEE International Conference on Robotics and Automation (ICRA), pp.5228–5234, 2024.
34) Y. Kuang, H. Lin and M. Jiang, “OpenFMNav: Towards Open-Set Zero-Shot Object Navigation via Vision-Language Foundation Models,” arXiv preprint arXiv:2402.10670, 2024.
35) Y. Cai, X. He, M. Wang, H. Guo, W. -Y. Yau and C. Lv, “CL-CoTNav: Closed-Loop Hierarchical Chain-of-Thought for Zero-Shot Object-Goal Navigation with Vision-Language Models,” arXiv preprint arXiv:2504.09000, 2025.
36) P. Wu, Y. Mu, B. Wu, Y. Hou, J. Ma, S. Zhang and C. Liu, “VoroNav: Voronoi-Based Zero-Shot Object Navigation with Large Language Model,” arXiv preprint arXiv:2401.02695, 2024.
37) H. Yin, X. Xu, Z. Wu, J. Zhou and J. Lu, “SGNav: Online 3D Scene Graph Prompting for LLM-Based Zero-Shot Object Navigation,” Advances in Neural Information Processing Systems, Vol.37, pp.5285–5307, 2024.
38) Z. Ji, H. Lin and Y. Gao, “DynaVLM: Zero-Shot Vision-Language Navigation System with Dynamic Viewpoints and Self-Refining Graph Memory,” arXiv preprint arXiv:2506.15096, 2025.
39) A. Payandeh, D. Song, M. Nazeri, J. Liang, P. Mukherjee, A. H. Raj, Y. Kong, D. Manocha and X. Xiao, “Social-LLaVA: Enhancing Robot Navigation Through Human-Language Reasoning in Social Spaces,” arXiv preprint arXiv:2501.09024, 2024.
40) B. Yu, H. Kasaei and M. Cao, “L3MVN: Leveraging Large Language Models for Visual Target Navigation,” IEEE/RSJ International Conference

on Intelligent Robots and Systems (IROS), pp.3554–3560, 2023.
41) Z. Zhou, Y. Hu, L. Zhang, Z. Li and S. Chen, “BeliefMapNav: 3D Voxel-Based Belief Map for Zero-Shot Object Navigation,” arXiv preprint arXiv:2506.06487, 2025.
42) Y. Cao, J. Zhang, Z. Yu, S. Liu, Z. Qin, Q. Zou, B. Du and K. Xu, “CogNav: Cognitive Process Modeling for Object Goal Navigation with LLMs,” arXiv preprint arXiv:2412.10439, 2024.
43) S. Zhang, X. Yu, X. Song, X. Wang and S. Jiang, “Imagine Before Go: Self-Supervised Generative Map for Object Goal Navigation,” IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.16414–16425, 2024.
44) K. He, X. Chen, S. Xie, Y. Li, P. Dollar and R. Girshick, “Masked Autoencoders Are Scalable Vision Learners,” IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.16000–16009, 2022.
45) Y. Shi, J. Liu, L. Sun and X. Zheng, “E2BA: Environment Exploration and Backtracking Agent for Visual Language Object Navigation,” IEEE Transactions on Circuits and Systems for Video Technology, 2025.
46) L. Zhong, C. Gao, Z. Ding, Y. Liao, H. Ma, S. Zhang, X. Zhou and S. Liu, “TopV-Nav: Unlocking the Top-View Spatial Reasoning Potential of MLLM for Zero-Shot Object Navigation,” arXiv preprint arXiv:2411.16425, 2024.
47) Y. Ji, K. Yun, Y. Liu, Z. Wang, B. Ma, Z. Xie and H. Liu, “Diffusion as Reasoning: Enhancing Object Navigation via Diffusion Model Conditioned on LLM-Based Object-Room Knowledge,” arXiv preprint arXiv:2410.21842, 2024.
48) L. Zhang, Q. Zhang, H. Wang, E. Xiao, Z. Jiang, H. Chen and R. Xu, “TriHelper: Zero-Shot Object Navigation with Dynamic Assistance,” IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.10035–10042, 2024.
49) D. Goetting, H. G. Singh and A. Loquercio, “End-to-End Navigation with Vision Language Models: Transforming Spatial Reasoning into Question-Answering,” arXiv preprint arXiv:2411.05755, 2024.
50) M. Khanna, Y. Mao, H. Jiang, S. Haresh, B. Shacklett, D. Batra, A. Clegg, E. Undersander, A. X. Chang and M. Savva, “Habitat Synthetic Scenes Dataset (HSSD200): An Analysis of 3D Scene Scale and Realism Tradeoffs for ObjectGoal Navigation,” IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.16384–16393, 2024.
51) Z. Gong, R. Li, T. Hu, R. Qiu, L. Kong, L. Zhang, Y. Ding, L. Zhang and J. Liang, “Stairway to Success: Zero-Shot Floor-Aware Object-Goal Navigation via LLM-Driven Coarse-to-Fine Exploration,” arXiv preprint arXiv:2505.23019, 2025.
52) L. Zhang, H. Wang, E. Xiao, X. Zhang, Q. Zhang, Z. Jiang and R. Xu, “Multi-Floor Zero-Shot Object Navigation Policy,” arXiv preprint arXiv:2409.10906, 2024.
53) V. Vuong, T. Nguyen, M. N. Vu, B. Huang, H. T. T. Binh, T. Vo and A. Nguyen, “HabiCrowd: A High Performance Simulator for Crowd-Aware Visual Navigation,” IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.5821–5827, 2024.
54) Z. Gong, T. Hu, R. Qiu and J. Liang, “From Cognition to Precognition: A Future-Aware Framework for Social Navigation,” arXiv preprint arXiv:2409.13244, 2024.
55) H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. Hart, J. Biswas and P. Stone, “Socially Compliant Navigation Dataset (SCAND): A Large-Scale Dataset of Demonstrations for Social Navigation,” IEEE Robotics and Automation Letters, Vol.7, No.4, pp.11807–11814, 2022.
56) D. M. Nguyen, M. Nazeri, A. Payandeh, A. Datar and X. Xiao, “Toward Human-Like Social Robot Navigation: A Large-Scale, Multi-Modal, Social Human Navigation Dataset,” IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.7442–7447, 2023.
57) M. J. Munje, C. Tang, S. Liu, Z. Hu, Y. Zhu, J. Cui, G. Warnell, J. Biswas and P. Stone, “Socialnav-sub: Benchmarking VLMs for Scene Understanding in Social Robot Navigation,” arXiv preprint arXiv:2509.08757, 2025.