Journal of KIIT. Vol. 24, No. 3, pp. 61-68, Mar. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.3.61

대형 언어 모델의 공간 추론을 위한 지도 이미지 및 텍스트 기반 지형 표현에 대한 비교

박종현*[1], 김성헌*[2], 김예찬*[3], 권철희**[1], 조규태**[2], 진정훈**[3], 김지아**[4], 강선종**[5], 전문구***

Comparison of Textual Terrain Representations and Map image for LLM Spatial Reasoning

Jonghyun Park*[1], Sungheon Kim*[2], Yechan Kim*[3], Cheolhee Kwon**[1], Kyutae Cho**[2], Junghun Jin**[3], Jia Kim**[4], Seonjong Kang**[5], and Moongu Jeon***

이 연구는 LIG NEX1 산학협력과제 지원으로 연구되었음

요 약

본 논문에서는 세 가지 텍스트 표현(레스터, 벡터, 무작위 포인트) 방식에 따른 LLM의 공간 추론 능력과, VLM의 지도 이미지 기반 추론 능력을 비교하였다. 동일 영역의 SVG와 고도 맵 데이터를 가공하여 입력으로 사용하였으며, 텍스트는 격자 기반 레스터, 다각형 기반 벡터, 좌표 기반 표현으로 변환하였다. 다양한 표현 방식에 대해 고도 추정 및 가시성 판별 등 세 가지 과제를 평가하였다. 그 결과, LLM에서는 고도 맵 기반 레스터 표현이 벡터 및 무작위 점 표현보다 전반적으로 우수한 성능을 보였다. 반면 VLM은 고도 맵의 픽셀 의미를 잘 활용하지 못하고, SVG 기반 데이터에서 더 좋은 성능을 나타냈다. 또한 VLM의 공간 추론에는 객체 표시와 기호가 포함된 Visual grounding이 중요한 요소로 확인되었다.

Abstract

This paper compares the spatial reasoning ability of large language models (LLMs) across three text representations (raster, vector, and random points) with that of vision-language models (VLMs) using map images. For experiments, SVG and elevation map data from the same region were processed and used as inputs. The text representations were converted into grid-based raster, polygon-based vector, and coordinate-based formats. Three tasks, including elevation estimation and line-of-sight (LOS) prediction, were evaluated across different data representations. The results show that, for LLMs, the elevation map-based raster representation consistently outperforms vector and random point representations. In contrast, VLMs struggle to effectively utilize pixel-wise semantic information in elevation maps and perform better with SVG-based data. Furthermore, visual grounding, including object annotations, symbols, and textual cues, is identified as a crucial factor for spatial reasoning in VLMs.

Keywords

large language models, LLMs, spatial reasoning, terrain data representation, raster format, vector format

* 광주과학기술원 전기전자컴퓨터공학과
- ORCID[1]: https://orcid.org/0009-0005-5404-0707
- ORCID[2]: https://orcid.org/0009-0001-5221-6036
- ORCID[3]: https://orcid.org/0000-0002-2438-3590
** LIG NEX1 AI 연구소 AI연구개발팀
- ORCID[1]: http://orcid.org/0000-0002-5811-1622
- ORCID[2]: http://orcid.org/0009-0008-4037-1486
- ORCID[3]: http://orcid.org/0009-0003-1737-9213
- ORCID[4]: http://orcid.org/0009-0003-9050-7023
- ORCID[5]: https://orcid.org/0009-0009-8839-6778
*** 광주과학기술원 전기전자컴퓨터공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-2775-7789

· Received: Feb. 24, 2026, Revised: Mar. 12, 2026, Accepted: Mar. 15, 2026
· Corresponding Author: Moongu Jeon
Dept. of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology
Tel.: +82-62-715-2406, Email: mgjeon@gist.ac.kr

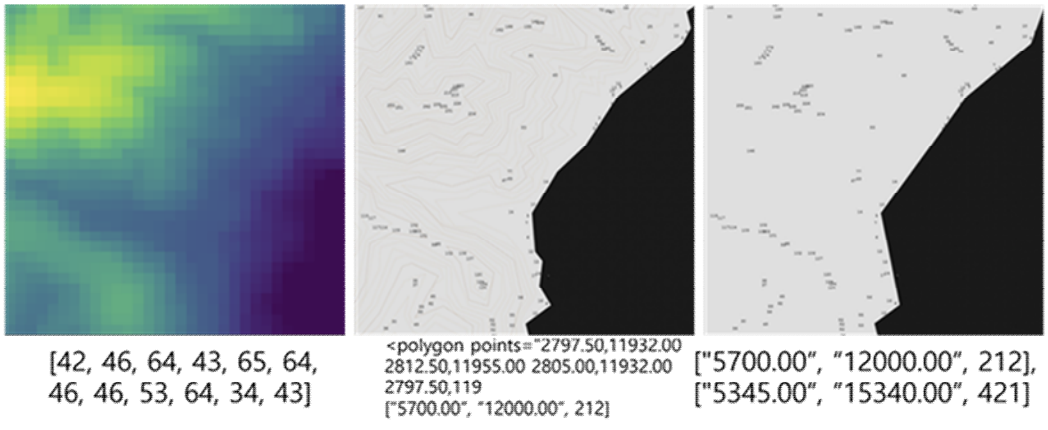
Ⅰ. 서 론

최신 대형 언어 모델(LLM, Large Language Model)이 등장함에 따라 자연어 이해, 수학적 계산, 코드 생성 등 다양한 추론 능력이 벤치마크를 통해 비교되고 있으며, 그 결과 여러 분야에서 연구 및 활용이 활발히 이루어지고 있다. 공간 추론 능력에 대한 연구 또한 마찬가지로, LLM의 WKT (Well-Known Text) 형식 데이터를 활용한 객체 관계 추론[1], 비전 언어 모델(VLM, Vision Language Model)의 지도 이미지 내 공간 추론 능력 평가[2], LLM을 활용한 게임 기반 벤치마크[3] 등 다양한 방향에서 진행되고 있다. 특히 비전 언어 모델의 지도 이미지 공간 추론 능력과 관련한 연구들은 공통적으로 인간 평가자에 비해 성능이 크게 부족함을 지적하고 있으며, 텍스트로 표현 가능한 경우에는 이를 텍스트화하여 실험을 수행하고 있다.

이러한 지도에 대한 표현 방식에 이미지와 텍스트, 그리고 텍스트 간에도 다양한 표현 방식이 있고 개별 표현에 대한 연구는 존재하지만 다양한 표현 방식에 대한 종합적인 비교 연구는 아직 없다.

따라서 본 논문에서는 동일한 두가지 데이터(고도맵, SVG(Scalable Vector Graphics) 형식의 데이터)를 LLM의 지형의 텍스트 표현 방식과 VLM의 이미지에서의 공간 추론 능력 비교를 수행한다. 텍스트 기반 지형 표현은 세 가지 유형으로 구분하였다. 첫 번째는 일정한 간격의 격자에 따라 값을 고도값을 순차적으로 나열하는 그림 1(a)와 같은 레스터 표현 방식이며, 두 번째는 그림 1(b)로 다각형과 무작위로 지정된 고도와 좌표로 기술된 벡터 표현, 세 번째는 그림 1(c)와 같이 벡터 표현 방식에서 다각형을 제외하고 좌표와 고도 값으로만 표시하는 무작위 점 표현 방식이다. 레스터 표현 방식은 공간 전체를 균일한 단위로 분할하여 밀집된 형태의 정보를 제공하는 반면, 벡터 표현 방식은 같이 의미 단위의 경계와 구조를 중심으로 희소한 형태로 공간을 기술한다. 무작위 점 표현 방식의 경우에는 벡터 표현 방식을 인식할 때 얼마나 점 표현 방식에 의존하는 지 비교하기위해 다각형을 제외하였다. 이미지의 경우, VLM의 특성상 이미지 픽셀의 값에 대한 토큰 변환이나 프롬프팅을 이용한 위치 참조 및 이미지 간의 정렬에 어려움[4]이 있어 텍스트 및 기호 참조가 있는 이미지를 사용하였다[5][6].

실험은 다음 세 가지로 구성된다. ① LLM의 레스터 표현과 벡터 표현 방식, 고도 포인트만 사용하였을때 공간 추론 능력 비교, ② 관심 영역 외 데이터를 추가적인 컨텍스트로 포함하여 LLM이 벡터 표현 방식에서 관심 영역을 정확히 추출하는지에 대한 평가 ③ 동일한 데이터 셋으로 VLM이 이미지를 통해 공간 추론 진행이다.



(a) (b) (c)

그림 1. 본 논문의 실험에 사용한 텍스트 기반 지형 표현과 시각화
(a) 레스터 표현 방식 (b) 벡터 표현 방식 (c) 다각형 제외 벡터 표현 방식

Fig. 1. Terrain text-based representation used for the experiments and its corresponding visualiza
(a) Raster representation (b) Vector representation (c) Vector representation without polygon

실험 결과 각 픽셀이 의미를 가지고 있는 깊이(Depth) 이미지와 같은 고도 맵에서는 레스터 표현으로 LLM의 텍스트 입력으로 사용하는 것이 벡터 표현 보다 적절하였다. LLM은 벡터 표현에 대해서 다각형 표현에 대해 방해 요소로 인식하였다. 반면에 VLM의 경우에는 반대로 고도 맵을 통해 각 픽셀의 의미에 대해 제대로 토큰화하지 못하였고 SVG(벡터 기반의 지도)에 각종 기호와 텍스트로 실험 데이터를 이미지로 시각화 했을때 VLM은 높은 이해도를 보였다.

본 연구의 주요 기여는 다음과 같다. ① 동일한 지형 정보를 세가지 표현(레스터, 벡터, 무작위 포인트 표현)로 텍스트화 하여 LLM 및 이미지에 대한 VLM의 공간 추론 성능을 비교하였다. ② 텍스트 입력에 대한 컨텍스트 크기, 관심 영역 외 데이터의 포함 여부, 희소 및 밀집 정보 구조 차이에 따른 공간 추론 성능 변화를 차이를 비교하였다.

II. 관련 연구

이미지에 대한 공간 추론 능력 연구는 다수 진행되어 왔다. 초기 VLM은 공간 추론에서 이미지 정보보다 텍스트 정보에 크게 의존하는 경향을 보였으며, 노이즈 이미지를 입력한 경우와 큰 차이가 없는 성능을 나타냈다[7]. 최신 상업용 모델인 GPT-5에 이르러서야 성능이 크게 향상되었으나, 여전히 인간 평가자와 비교하면 부족한 수준이며 오픈소스 모델에서는 그 격차가 더욱 크게 나타난다[8]. 지도 데이터의 경우에도 자연 영상과 마찬가지로 인간에 비해 낮은 공간 추론 성능을 보이며, 텍스트 크기나 기호 색상의 채도와 같은 시각적 요소의 영향을 받는다[2].

이러한 한계의 주요 원인 중 하나는 VLM은 비전 트렌스포머 기반의 비전 인코더를 통해 이미지를 2차원 공간 토큰 형태로 처리하며, 각 토큰은 이미지 내 특정 위치 좌표에 대응한다. 따라서 프롬프트로 설명된 공간 정보를 이미지의 특정 영역과 정확히 정렬하기 어렵다[4]. 이러한 문제를 완화하기 위해서 최근에는 Visual grounding(공간 정보를 텍스트로 설명하는 대신 이미지 내부에 직접 표시하는 방식) 기반 접근 방식을 활용하고 있다[5][6].

LLM의 공간 추론 능력을 활용하려는 AI Agent 연구들 또한 이미지 기반 공간 추론 성능의 한계로 인해 텍스트 기반 레스터 표현을 주로 사용하고 있다[3][9]. 한편, 텍스트 기반 벡터 표현의 경우 공간 관계 추론에서 일정 수준의 성능을 보였으며[1], SVG(Scalable Vector Graphics) 형식의 데이터 시각화 차트에 관한 연구[10]나 SVG를 직접 생성하는 모델 연구[11]도 존재한다.

III. 실험 방법

평가는 정량적으로 계산 가능한 방법으로 구성되어있다. 고도 계산. 지형은 Arma 3 게임의 Altis 맵에서 추출한 SVG 파일로 실제 지형인 그리스의 '림노스' 섬을 배경으로 하며 총 30.7 × 30.7 km의 면적을 가지고 있지만 본 실험에서는 2 × 2, 4 × 4 km로 크롭하여 사용한다. 지상에만 객체 및 목표를 배치하고 최고 고도는 380 m이다.

정답 레이블은 LLM이 입력 데이터로 사용하는 SVG 파일로 계산된 것이 아닌 1m/pixel 공간 해상도를 가진 고도 맵으로 제작되었다. 따라서 고도라는 단일 항목만이 고려되는 과제고 LLM이 등고선만을 이용해서 정확한 답을 추론하는 것은 불가능하다.

텍스트 표현에 대한 LLM의 공간 추론 능력 실험은 RTX A5000 4개를 사용한 서버에 openai의 오픈소스 모델인 gpt-oss-120b 모델로 진행, 지도 이미지에 대한 VLM의 공간 추론 능력 실험은 RTX 3090 8개를 사용한 서버에서 최신 오픈소스 VLM인 InterVL3.5-36B 모델[12]로 실험을 진행하였다. 실험 매개변수는 샘플링 온도는 0.0, 최대 토큰 생성 수는 8192, 시간 제한은 200초로 설정하였다.

3.1 질문 유형 및 평가 방법

질문 유형은 총 3가지로 ① 특정 좌표의 지형 고도 추정, ② 두 지점 간 직선 경로상의 최대 고도 계산, ③ 에이전트 간 가시성(LOS, Line-of-Sight) 판별로 구성되어 있으며 문제 수는 각각 90, 60, 60 문제이다. 에이전트 및 좌표 선택에 있어 무작위 방식으로 선택되었고 2 km, 4 km 총 두가지 영역에

대한 질문 세트로 구성이 되었다. 동일한 크기의 지도 영역에 대해서는 같은 질문 세트가 사용되었다. 각 문제에 대한 평가는 LLM의 환각 문제로 이상치가 너무 큰 문제가 있어 이상치를 제거하지 않고 중앙값을 표기하는 방법을 사용한다. 따라서 값이 큰 오류가 전체 데이터의 과반을 차지하는 경우에는 Median 임에도 오류값이 크게 표시 될 수 있다. ①, ②은 중앙값의 오차를 나타내는 Median Absolute Error(MedAE)를 사용, ③는 정확도(Accuracy)를 사용하였다. 시간 제한 및 최대 토큰 수 초과, 잘못된 형식의 답변으로 인한 오류는 평가 지표 계산에 사용하지 않고 실험 결과의 에러 비율로 표기하였다.

3.2 공간 데이터의 텍스트 변환

레스터 표현 방식은 컨텍스트 크기와 환각 문제로 인해 고도맵의 2, 4 km의 영역에 대해 64m 단위로 maxpooling 하여 LLM 입력으로 사용하였다. 따라서 컨텍스트에 입력되는 값의 수는 각각 1024, 4096 개이다. 레스터 표현의 경우에는 벡터 표현과는 다르게 행렬에 대한 좌표 표기와 각 픽셀이 표현하는 공간 해상도의 크기를 프롬프트로 함께 전달하는 것이 필수 임으로 그림 2와 같이 프롬프트를 추가하였다.

벡터 표현에 대한 실험은 원본의 관심영역을 자르고 지형 다각형을 RDP(Ramer – Douglas – Peucker)를 이용해 단순화하여 사용하였다. 표 1을 통해 단순화에 따른 각 데이터의 전체 포인트 수를 확인할 수 있다. 표 1의 RDP ε 값이 클수록 다각형의 단순화 정도가 크다. 단순화 과정에서 줄어든 다각형, 점의 갯수는 표 1에서 확인할 수 있다. 다만 다각형 단순화 시, 경계에서 다수의 선으로 분해되기도 한다. 표 1에서 ε 값이 64 일때가 ε 값이 32인 경우보다 폴리곤은 더 많지만 폴리곤당 포인트 수는 더 적어서 결과적으로 총 포인트 수가 더 적은 것을 확인할 수 있다. 그리고 등고선외에도 무작위 선택된 고도값 점 들이 포함되어있는데 이 포인트 값은 줄이지 않았다. 표 1의 맵 역영이 같은 경우 RDP의 ε 값이 다른 경우에도 고도 포인트 수는 같다. LLM이 등고선이 아닌 고도 포인트만으로 계산하는지를 평가하기 위해 추가적인 실험을 진행한다. 추가적인 실험으로는 ① 등고선을 유지 경계만 남기고 삭제시킨 상태(Vector representation without polygon)에서도 비교, ② 10m 단위의 등고선외에도 50m 단위의 등고선과도 비교, ③ 관심 영역(2 × 2 km) 이외의 지도 영역(4 × 4 km)이 포함된 경우에 대한 비교도 진행한다.

```
return (
    f"[MISSION]\n{mission_yup}\n\n"
    f"[TERRAIN_HEIGHT_GRID]\n"
    f"- COORD_SYSTEM: cartesian-like (x→right, y→down)\n"
    f"- HEIGHT_UNITS: meters\n"
    f"- ENCODING: elevation_m = pixel_u16/{SCALE} - {OFFSET}\n"
    f"- GRID_ORIGIN_XY: ({ox},{oy})\n"
    f"- STEP_PX: {step_px}\n"
    f"- STEP_M: {step_m:.2f}\n"
    f"- SHAPE_HW: ({H},{W})\n"
    f"- VALUE_AT(i,j): elevation at (x=ox + i*STEP_PX, y=oy + j*STEP_PX)\n"
    f"- VALUES_ROW_MAJOR:\n"
    f"{grid_text}\n\n"
    f"[TASK]\n"
    f"- Use the grid values directly.\n"
    f"- Do not assume missing data.\n"
    f"- Output strictly in the required answer format.\n"
    f"{note}"
)
```

그림 2. 레스터 표현 데이터 이해를 위해 추가된 프롬프트
Fig. 2. Additional prompt for understanding raster-represented data

표 1. 텍스트 기반 벡터 표현 데이터 셋 내 입력 컨텍스트 평균 다각형 및 점의 갯수
Table 1. Number of polygons and points in the text-based vector represen representation dataset

Parameter			Average numbers			
Map extent (km)	RDP ε	Contour interval	Number of polygons	Average points per polygon	Height points	Total points
2,048	32	10	73.31	6.50	129.78	627.46
	64	10	109.48	3.98	129.78	580.37
4,096	32	10	218.44	4.03	393.26	1975.55
	64	10	300.80	4.57	393.26	1797.16
	64	50	29.44	6.54	393.26	615.27

3.3 VLM 입력 지도 이미지

VLM의 입력으로 지도 이미지를 사용해 동일한 데이터 셋에 대해 비교 실험을 추가로 진행한다. 실험에 사용한 이미지는 그림 3과 같이 지도 격자선과 관심 객체들에 대한 Visual grounding이 포함하였다. 지도 격자선과 Visual grounding이 포함되지않은 이미지에 대해서도 실험을 진행하였으나 좌표 정렬이 되지않아 실험 결과에서 제외하였다. 2 × 2 km 영역의 이미지를 2048 × 2048 해상도의 이미지 입력으로 사용하였다. Visual grounding의 예시로 그림 3을 보면 가시선 계산 문제에서 두 객체 사이를 노란색 실선으로 표시했다.

IV. 실험 결과

4.1 텍스트 표현 방식에 따른 비교

표 2를 보면 표현 방식의 차이로 레스터 표현 (a), (b) 방식이 벡터 표현 (c), (d), (e), (f), (g)이나 고도 포인트 표현 (i), (j)에 비해 압도적인 우위를 보였다. 특히 LOS binary 문제의 경우에 벡터 표현 방식의 경우에는 정확도가 최고 0.58일 정도로 무작위 추정 수준의 결과를 보여줬지만 레스터 표현은 64 공간 해상도로 축소되어 세부 지형 정보가 손실되었음에도 0.94로 높은 정확도를 보였다.

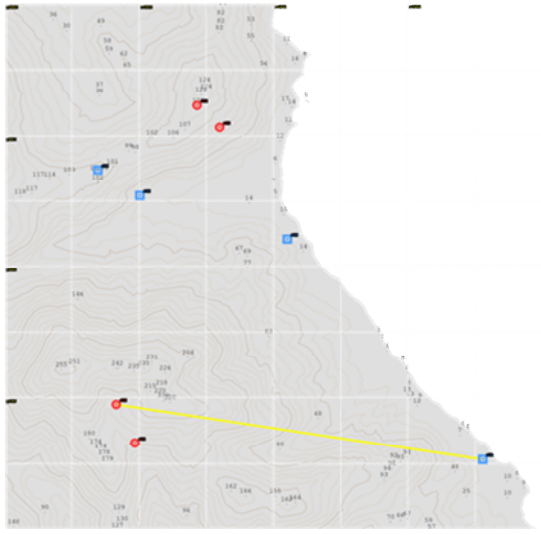


그림 3. Visual Groung이 포함된 지도 이미지
Fig. 3. Example of Visual Grounding on a map image

표 2. 텍스트 기반 지형 표현에 대한 LLM 공간 추론 능력 실험 결과
Table 2. Experimental results on LLM spatial reasoning with text-based terrain representation

	Rester representation			Metrics			error
	Map extent (km)	Max pooling		Point elevation (MedAE)	Max elevation on path (MedAE)	LOS binary (Acc)	error ratio
(a)	2,048	64	-	5.06	**5.45**	0.94	0.30
(b)	4,096	64	-	**4.23**	5.50	0.76	0.35
	Vector representation						
	Map extent (km)	RDP ε	Contour interval (m)				
(c)	2,048	32	10	25.88	23.65	0.40	**0.05**
(d)		64	10	8.23	20.34	0.34	0.12
(e)	4,096	32	10	8117.66	75.56	0.39	0.11
(f)		64	10	103.41	62.96	0.48	0.14
(g)			50	13.28	42.2	0.45	0.18
(h)	4,096 (roi: 2,048)	64	10	19.60	20.10	0.29	0.12
	Vector representation without polygon						
	Map extent (km)	RDP ε					
(i)	2,048	64	-	8.2	21.05	0.58	0.07
(j)	4,096	64	-	10.94	45.65	0.46	0.10
	Image						
	Map extent (km)	RDP ε					
(k)	2,024	64		12.94	17.28	**1.0**	0.22
(l)	2,024	0		10.96	16.43	**1.0**	0.21

벡터 표현 방식과 포인트 표현 방식간의 성능 차이를 비교하면 오히려 다각형의 등고선 표현 방식이 정보의 양은 더욱 많지만 성능에 악역향을 미쳤다는 것을 알 수 있다. 본 연구진은 벡터 표현에서 고도 점의 갯수를 하나만 사용한 추가 실험을 진행하였지만 LLM이 동일한 답변만 계속 하는 문제가 있어 LLM이 벡터 표현 계산에 등고선을 보지않고 랜덤으로 샘플링된 고도 및 좌표 만으로 충분했다는 것을 알 수 있다. 벡터 방식에서 오히려 많은 등고선 (등고선 수: (j) < (g) < (f))입력 컨텍스트 길이가 너무 긴 경우에는 표 2의 (e)와 같이 최대 고도가 380m 임에도 환각 증세가 생긴다. 고도 값이 아닌 좌표 값을 읽고 답변 했을 가능성이 있다.

4.2 관심 영역 외 입력에 따른 성능 비교

같은 컨텍스트 크기를 가지지만 객체가 배치된 관심 영역만 다른 표 2의 (e), (g)를 보면 입력 컨텍스트 길이보다는 관심 영역 사이의 계산에 신경 써야 할 값들이 얼마나 더 많은가가 중요하다. 다만 Point elevation 같은 경우에는 (e), (g)가 차이를 보이면 안되지만 큰 차이가 났다. 이에 대해 추가적인 실험이 필요해 보인다.

4.3 이미지와 텍스트 표현 간의 성능 비교

텍스트와 동일한 질문에 대해서 가시성 계산 문제는 100% 정확도를 보였으며 고도 값 추정의 경우에는 텍스트 표현보다 소폭 성능이 낮았다.

다만 실험에 사용된 이미지는 Visual grounding이 포함된 그림 3과 같은 이미지이다. 고도 맵에 대한 데이터와 객체 참조를 프롬프트에서 진행한 실험 결과는 점수가 무작위 추론에 가까워 포함하지 않았다. 먼저 고도맵 이미지의 경우 각 픽셀이 고도 정보를 포함하고 있지만 실험에 사용한 InternVL3.5 [12] 와 같이 CLIP[13] 기반의 VLM은 이미지 내 객체나 장면의 의미적 특징을 인식하는데 강점을 가지지만 연속적인 수치 정보나 기하학적 구조를 해석하는 능력이 떨어진다. 깊이(Depth) 이미지에서의 추론 능력이 떨어지는 기존 연구와 동일하다[5]. Visual grounding이 포함 되었지만 이미지를 격자를 통한 좌표 표현이 없는 경우 프롬프트에서 제공된 좌표 대응 정보에 대해서 연결하지 못하였다. 따라서 지도내의 관심 객체에 대한 위치를 이미지위에 함께 표기하는 Visual grounding이 필수적이다.

V. 결 론

본 논문의 실험으로 텍스트 기반의 공간 추론 문제에서는 고도 맵 기반의 레스터 표현 방식이 SVG를 가공한 벡터 표현 방식보다 좋은 성능을 냈다. 벡터 표현 방식의 다각형은 오히려 LLM의 추론 성능에 방해가 되었다. 반면에 VLM의 지도 이미지 인식의 경우에는 고도 맵 기반의 데이터에서 픽셀에 대한 고유 정보와 값과 좌표 정보가 손실 되기 때문에 등고선과 고도 포인트가 텍스트로 함께 포함된 SVG 기반의 데이터에서 이미지에서의 성능이 더 좋았다. Vision Transformer 기반 VLM의 시각 임베딩은 종종 세밀한 공간 구조나 기하학적 패턴을 유지하지 못한다는 기존의 연구와 같은 결과이다.

따라서 각 픽셀이 고유의 값을 가지는 데이터(고도맵, 깊이 이미지)에 대해서는 레스터 표현의 형태로 텍스트 입력을 하는 것이 낫고 벡터 표현의 지도와 같은 경우에는 관심 객체와 영역, 수치에 대해 명확히 이미지 위에 표기한 이후 VLM의 이미지 입력으로 사용해야한다.

References

[1] Y. Ji, S. Gao, Y. Nie, I. Majić, and K. Janowicz, "Foundation models for geospatial reasoning: Assessing the capabilities of large language models in understanding geometries and topological spatial relations", International Journal of Geographical Information Science, Vol. 39, No. 9, pp. 1866-1903, Sep. 2025. https://doi.org/10.1080/13658816.2025.2511227.

[2] J. Pyo, et al., "FRIEDA: Benchmarking multi-step cartographic reasoning in vision-language models", arXiv preprint, pp. 1-38, Dec. 2025. https://doi.org/10.48550/arXiv.2512.08016.

[3] M. U. Nasir and J. Togelius,

"GameTraversalBenchmark: Evaluating planning abilities of large language models through traversing 2D game maps", Advances in Neural Information Processing Systems (NeurIPS), Vancouver, Canada, Vol. 37, pp. 31813-31827, Dec. 2024. https://doi.org/10.52202/079017-1000.

[4] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Driess, P. Florence, D. Sadigh, L. J. Guibas, and F. Xia, "SpatialVLM: Endowing vision-language models with spatial reasoning capabilities", Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 14455-14464, Jun. 2024. https://doi.org/10.1109/CVPR52733.2024.01370.

[5] J. Qi, J. Liu, H. Tang, and Z. Zhu, "Beyond semantics: Rediscovering spatial awareness in vision-language models", arXiv preprint, pp. 1-26, Mar. 2025. https://doi.org/10.48550/arXiv.2503.17349.

[6] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu, "SpatialRGPT: Grounded spatial reasoning in vision-language models", Advances in Neural Information Processing Systems (NeurIPS), Vancouver, Canada, Vol. 37, pp. 135062-135093, Dec. 2024. https://doi.org/10.52202/079017-4293.

[7] J. Wang, Y. Ming, Z. Shi, V. Vineet, X. Wang, Y. Li, and N. Joshi, "Is a picture worth a thousand words? Delving into spatial reasoning for vision-language models", Advances in Neural Information Processing Systems (NeurIPS), Vancouver, Canada, Vol. 37, pp. 75392-75421, Dec. 2024. https://doi.org/10.52202/079017-2400.

[8] Z. Cai, et al., "Has GPT-5 achieved spatial intelligence? An empirical study", arXiv preprint, pp. 01-29, Aug. 2025. https://doi.org/10.48550/arXiv.2508.13142.

[9] T. Anne, N. Syrkis, M. Elhosni, F. Turati, F. Legendre, A. Jaquier, and S. Risi, "Harnessing language for coordination: A framework and benchmark for LLM-driven multi-agent control", IEEE Transactions on Games, Vol. 17, No. 2, pp. 933-943, Jun. 2025. https://doi.org/10.1109/TG.2025.3564042.

[10] Z. Xu and E. Wall, "Exploring the capability of LLMs in performing low-level visual analytic tasks on SVG data visualizations," IEEE Visualization and Visual Analytics (VIS), St. Pete Beach, FL, USA, pp. 126-130, Oct. 2024. https://doi.org/10.48550/arXiv.2404.19097.

[11] K. Nishina and Y. Matsui, "SVGEditBench: A benchmark dataset for quantitative assessment of LLM's SVG editing capabilities", Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, USA, pp. 1-10, Jun. 2024.

[12] W. Wang, et al., "InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency", arXiv preprint, pp. 1-34, Aug. 2025. https://doi.org/10.48550/arXiv.2508.18265.

[14] 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", Proc. International Conference on Machine Learning (ICML), Online, Vol. 139, pp. 8748-8763, Jul. 2021. https://doi.org/10.48550/arXiv.2103.00020.

저자소개

박 종 현 (Jonghyun Park)

2021년 2월 : 수원대학교 정보통신공학과(학사)
2023년 8월 : 광주과학기술원 전기전자컴퓨터공학과(공학석사)
2023년 9월 ~ 현재 : 광주과학기술원 전기전자컴퓨터공학과 박사과정
관심분야 : AI, Robotics

김 성 헌 (Sungheon Kim)

2025년 2월 : 광주과학기술원
전기전자컴퓨터공학과(학사)
2025년 3월 ~ 현재 :
광주과학기술원
전기전자컴퓨터공학과 석사과정
관심분야 : AI, Deep Learning

김 예 찬 (Yechan Kim)

2021년 8월 : 광주과학기술원
전기전자컴퓨터공학부(공학석사)
2021년 9월 ~ 현재 :
광주과학기술원
전기전자컴퓨터공학부 박사과정
관심분야 : 인공지능, 컴퓨터 비전,
멀티모달 학습

권 철 희 (Chelhee Kwon)

1998년 2월 : 고려대학교
제어계측공학(공학사)
2000년 2월 : 고려대학교
제어계측공학(공학석사)
2000년 1월 ~ 현재 :
LIG 넥스원 연구위원
관심분야 : AI, Deep Learning

조 규 태 (Kyutae Cho)

2002년 2월 : 숭실대학교
전산학(공학사)
2004년 2월 : 한국과학기술원
전산학(공학석사)
2007년 2월 : 한국과학기술원
전산학(박사수료)
2007년 2월 ~ 현재 :
LIG 넥스원 연구위원
관심분야 : AI, Deep Learning

진 정 훈 (Junghun Jin)

2002년 2월 : 고려대학교
전자공학(공학사)
2004년 2월 : 고려대학교
영상정보처리(공학석사)
2004년 1월 ~ 현재 :
LIG 넥스원 수석연구원
관심분야 : AI, Deep Learning

김 지 아 (Jia Kim)

2007년 2월 : 이화여자대학교
컴퓨터공학(공학사)
2006년 12월 ~ 현재 :
LIG 넥스원 수석연구원
관심분야 : AI, Deep Learning

강 선 종 (Seonjong Kang)

2022년 2월 : 동국대학교
전자전기공학부(공학사)
2024년 2월 : 동국대학교
전자전기공학과(공학석사)
2024년 1월 ~ 현재 :
LIG 넥스원 선임연구원
관심분야 : AI, Deep Learning

전 문 구 (Moongu Jeon)

2001년 6월 : University of
Minnesota, Scientific
Computation(공학박사)
2001년 3월 ~ 2003년 2월 : UCSB,
USA, 박사후연구원
2003년 3월 ~ 2005년 8월 :
Institute for Biodiagnostics,
NRC, Canada 연구원
2005년 9월 ~ 현재 : 광주과학기술원 정보컴퓨팅대학
전기전자컴퓨터공학과 교수
관심분야 : 인공지능, 기계학습, 컴퓨터비전