특집논문 (Special Paper)
방송공학회논문지 제31권 제1호, 2026년 1월 (JBE Vol.31, No.1, January 2026)
https://doi.org/10.5909/JBE.2026.31.1.41
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

공간 미디어를 위한 언어 생성 기반의 3D 장면 이해

김 연 의[a], 양 진 우[a], 강 제 원[a]‡

Language Generation-based 3D Scene Understanding for Spatial Media

Yeoneui Kim[a], Jinoo Yang[a], and Jewon Kang[a]‡

요 약

로보틱스, 웨어러블 디바이스, 확장현실/가상현실(AR/VR)의 확산으로 3차원(3D) 공간을 배경으로 사용자와 상호작용하며 미디어 서비스를 제공하는 공간 미디어에 대한 관심이 높아지고 있다. 공간 미디어에서는 사용자의 위치와 행동이 주변 공간의 구조 정보와 통합되며, 인공지능 모델이 공간에 대한 설명, 질의응답, 콘텐츠 생성 등과 같은 서비스를 제공하는 새로운 미디어 패러다임을 제공하는 것이 가능하다. 본 논문은 이러한 공간 미디어 응용의 핵심적인 기술이 되는 언어 생성 기반 3D 장면 이해를 중심으로 관련 연구를 체계적으로 정리한다. 먼저 대표적 데이터셋과 3D 장면 이해에서의 언어 생성 태스크를 정의하고, 최신 방법들을 단일 과제 학습과 다중 과제 학습으로 분류하여 정리한다. 이어서 최신 모델들의 성능을 비교 및 평가하고, 마지막으로 기존 연구의 한계를 논의하며 향후 연구 방향을 제시한다.

Abstract

The rapid growth of robotics, wearable devices, and augmented reality/virtual reality (AR/VR) has garnered attention of spatial media, in which a user can naturally interact with their three-dimensional (3D) environments and media services. This paper provides a comprehensive review of recent research on language generation‐based 3D scene understanding, which is crucial for such agents. We first introduce representative datasets and define key language-generation tasks in 3D scene understanding. We then survey existing methods categorized into single-task and multi-task learning approaches. Next, we compare and evaluate state-of-the-art models on standard benchmarks. Finally, we discuss current limitations and outline promising directions for future research.

Keyword : Spatial Media, Embodied Agent, 3D Scene Understanding, Multi-modal, Language-generation

a) 이화여자대학교 전자전기공학과(Ewha Womans University)
‡ Corresponding Author : 강제원(Jewon Kang)
E-mail: jewonk@ewha.ac.kr
Tel: +82-2-3277-2347
ORCID: https://orcid.org/0000-0002-1637-9479

※ This work was supported by the National Research Foundation of Korea (NRF), South Korea grant funded by the Korea government (MSIT) (RS-2025-23524046) and Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (Project Number: RS-2024-00439534).
· Manuscript December 3, 2025; Revised January 8, 2026; Accepted January 9, 2026.

Ⅰ. 서 론

로보틱스, 웨어러블 디바이스, 확장현실/가상현실(AR/VR)의 확산과 함께 공간 미디어 기술에 대한 관심도 높아지고 있다. 공간 미디어는 사용자가 주변 3차원(3D) 공간 상에서 위치하고 이동하는 상황에서, 사용자와 주변 환경의 이해를 바탕으로 설명, 질의응답, 내비게이션, 콘텐츠 생성과 같은 능동적 서비스를 제공하는 새로운 패러다임을 포함한다. 이를 위해 인공지능 모델 및 에이전트는 사용자 및 주변 환경 모두와 상호작용할 수 있도록 가상적 또는 물리적 형태로 구현될 수 있다[1]. 공간 미디어 서비스를 제공하기 위해서는 단순한 공간의 인식을 넘어 3차원상의 복잡한 구조, 객체들의 속성과 공간적 관계를 종합적으로 파악하는 3D 장면 이해(3D Scene Understanding)가 필수적이다. 또한 사용자 - 에이전트 상호작용을 위해서는 언어 입력을 이해하고 적절한 응답을 생성하는 능력도 중요하다. 예를 들어 XR 환경에서의 실시간 장면 설명, 웨어러블 기반의 상황 인지형 어시스턴트, 로봇의 언어 기반 공간 조작 및 경로 계획 등 모두 3D 장면 이해와 언어 이해 기반 생성을 핵심 구성으로 한다.

최근 대형 언어 모델(LLM)과 멀티모달 모델의 발전으로 언어 기반 시각 이해가 크게 향상되었다. 대형 언어 모델의 풍부한 언어 지식과 추론 능력을 계승한 멀티모달 모델은 시각적 추론과 언어 이해 능력이 함께 향상되면서, 다양한 멀티모달 태스크를 수행할 수 있게 되었고, 언어 쿼리에 대해 더욱 세밀하고 정확한 응답 생성이 가능해졌다[2-6]. 그러나 기존 멀티모달 모델은 주로 2D 시각 데이터에 집중되어 왔기 때문에 2D와 달리 더 복잡한 3D 공간 구조와 2D 픽셀 표현 - 3D 표현 간 불일치로 인해 깊이·거리 등 공간적 특성을 정밀하게 이해하는 데 어려움이 존재한다. 이로 인해 실제 3D 환경에의 직접적 적용에는 한계가 있었다. 이러한 한계를 극복하기 위해, 3D 장면 표현을 처리하는 비전 모델과 언어 모델을 결합하여 3차원 공간에 대한 심층적 이해와 언어적 추론 능력을 동시에 향상시키기 위한 연구가 활발히 진행되고 있다. 본 논문은 사용자 입력에 따라 적절한 자연어 응답을 생성할 수 있는 언어 생성 기반 3D 장면 이해 기술을 중심으로 공간 미디어 관련 연구를 조사하고 분석하고자 한다.

Ⅱ. 언어 생성 기반 3D 장면 이해 태스크 정의

본 장에서는 3D 장면과 자연어의 결합이 3D 장면 이해

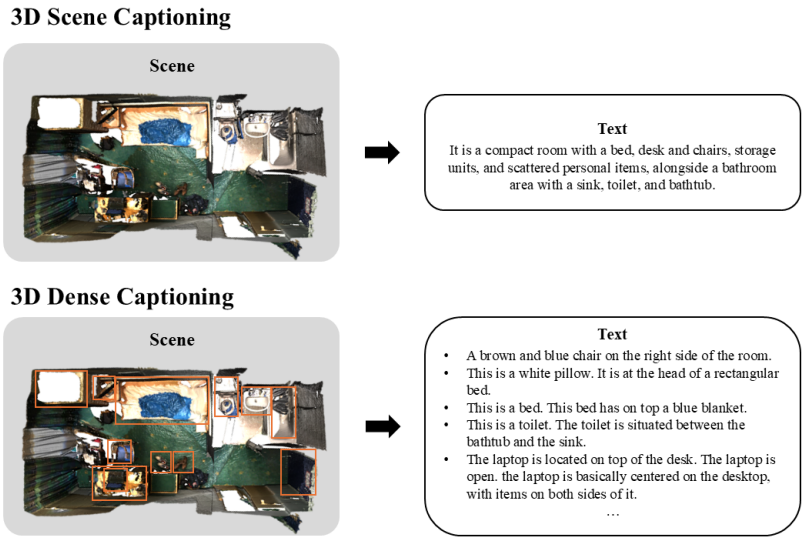


그림 1. 3D 장면 캡셔닝과 고밀도 캡셔닝 예시
Fig. 1. Illustration of 3D Scene Captioning and Dense Captioning Tasks

와 추론을 요구하는 다양한 과제 중 자연어 응답 생성을 목표로 하는 3D 질의응답과 3D 캡셔닝을 중심으로 문제 정의를 제시한다.

1. 언어 생성 기반 태스크

1.1 3D 캡셔닝(3D Captioning)

3D 캡셔닝은 3D 장면에 나타나는 물리적, 시각적 정보를 자연어로 설명하는 것을 목표로 한다. 이는 전체 장면에 대한 전역적 장면 정보와 주요 객체, 그리고 객체들 사이의 관계를 설명하여 장면 수준의 3D 장면 이해를 요구하는 3D 장면 캡셔닝(3D Scene Captioning)과, 장면 내 감지된 모든 객체에 대해 개별적인 자연어 설명을 생성하는 객체 수준의 3D 장면 이해를 요구하는 3D 고밀도 캡셔닝(3D Dense Captioning)으로 구분된다[7,8]. 그림 1은 ScanNet[11] 데이터셋을 사용한 각 캡셔닝 태스크의 입력과 출력에 대한 예시를 나타내며, 3D 고밀도 캡셔닝의 경우, 일부 캡션만 제시하였다.

1.2 3D 질의응답(3D Question Answering, 3D QA)

3D 질의응답은 3D 장면에 대한 심층적인 이해와 추론 능력을 바탕으로, 장면 내 공간적 관계와 객체의 속성 등 다양한 질문에 대해 자연어로 적절한 답변을 생성하는 과제이다. 본 과제는 전역적 관점에서 전체 장면을 대상으로 주어진 질문에 응답하는 3D 시각적 질의응답(3D Visual Question Answering, 3D VQA)[9]과, 언어로 제공되는 상황 설명과 질문을 입력으로 받아 1인칭 시점의 답변을 생성하는 3D 상황적 질의응답(3D Situated Question Answering, 3D SQA)[10]으로 구분된다. 3D 상황적 질의응답은 답변 생성과 함께 선택적으로 에이전트의 위치, 시야, 방향 정보를 예측하여 출력한다. 그림 2는 ScanNet[11] 데이터셋을 사용한 각 태스크의 입력과 출력 예시를 나타낸다. 그림 1과 그림 2에서 볼 수 있듯이, 3D 캡셔닝은 질문 입력 없이 장면 정보만을 입력으로 사용하는 반면, 3D 질의응답은 질문 또는 질문과 상황적 설명을 함께 입력으로 받는 구조를 따른다.

Ⅲ. 3D 멀티모달 데이터셋

본 장에서는 3D 장면 이해 및 추론에 사용되는 다양한 3D 멀티모달 데이터셋의 구성과 특성을 정리하였다. 표 1은 언어 생성 기반 3D 장면 이해에 사용되는 주요 데이터셋

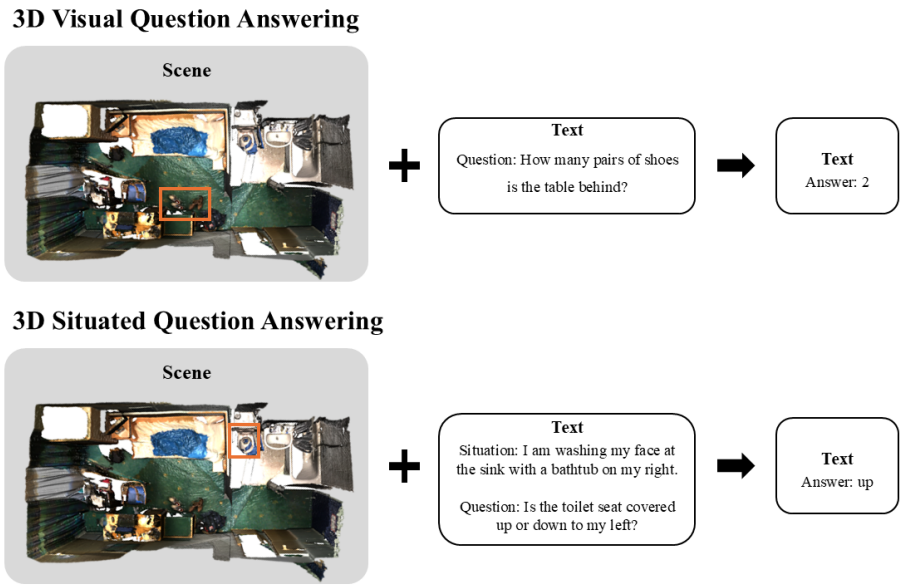


그림 2. 3D 시각적 질의응답과 상화적 질의응답 예시
Fig. 2. Illustration of 3D Visual Question Answering and Situated Question Answering

표 1. 3D 멀티모달 데이터셋
Table 1. 3D Multi-modal Datasets

Datatset	Source	Scene	Objects	Data Pairs	Task
Scan2Cap[8]	ScanNet[11]	800	11,046	51,583	3D dense captioning
Nr3D[12]	ScanNet[11]	707	4,664	32,919	3D object grounding
ScanQA[9]	ScanNet[11]	800	-	41,363Q, 58,191A	VQA
SQA3D[10]	ScanNet[11]	650	-	20,400Q, 33,400A	Situated VQA
LEO[13]	ScanNet[11], 3RScan[14], Objaverse[15]	3000	-	83,000	Multi-task (Scene captioning, 3D object captioning 3D QA, Embodied Navigation etc.)
MSQA[16]	ScanNet[11], 3RScan[14], ARKitScenes[17]	-	-	251,000	Situated VQA

들의 구성과 태스크 범위를 비교하여 나타내었다.

ScanNet 기반 데이터셋으로는 3D 고밀도 캡셔닝을 위한 Scan2Cap과 3D 객체 참조 태스크를 다루는 Nr3D가 있으며, 이들은 객체 수준의 언어 표현 학습에 초점을 두고 주로 3D 고밀도 캡셔닝 및 객체 참조 태스크의 벤치마크 데이터셋으로 활용된다. 반면 ScanQA와 SQA3D는 질문 - 응답 형식의 데이터셋으로, 전자는 장면 전반에 대한 3D 시각적 질의응답을 다루며, 후자는 에이전트 중심의 상황 인지와 공간 추론을 요구하는 3D 상황적 질의응답 태스크를 포함한다.

MSQA(Multi-modal Situated Question Answering)는 기존의 3D 질의응답 데이터셋이 주로 텍스트 기반의 상황 설명에 의존하여 시각 정보를 질의 해석을 위한 입력으로 명시적으로 활용하지 못하는 한계를 보완하고자 제안되었다. 실제 실내 3D 장면을 기반으로 텍스트, 이미지, 3D 포인트 클라우드를 결합한 인터리브 입력 형식을 도입하고, LLM 기반 질문 - 응답 생성을 통해 약 251,000개의 대규모 situated QA 쌍을 구축함으로써, 보다 정밀한 상황 인지와 복합적인 3D 추론 능력의 학습 및 평가를 가능하게 한다.

한편, 다양한 3D 태스크를 포괄하는 대규모 다중 태스크 데이터셋도 제안되었다. LEO는 ScanNet, 3RScan, Objaverse 등 다양한 소스 데이터셋으로부터 수집된 3D - 텍스트 쌍을 기반으로 객체(object), 장면에서의 객체(object-in-scene), 장면(scene) 수준의 다중 표현을 제공하며, 캡셔닝, 질의응답, 내비게이션 등 광범위한 태스크를 지원한다.

Ⅳ. 언어 생성 기반 3D 장면 이해의 최신 연구 동향

본 장에서는 언어 생성 기반 3D 장면 이해의 최신 연구 동향을 다룬다. 먼저 포인트 클라우드와 복셀 그리드처럼 3D 기하학적 형태와 규모 정보를 명시적으로 표현하는 방식, 그리고 다중 시점 이미지와 카메라 정보를 활용하는 다중 시점 기반 표현 등 주요 3D 장면 표현 방법을 살펴본다. 이어서 언어 생성 기반 3D 장면 이해 연구를 단일 과제 학습과 다중 과제 학습 기법으로 분류하여 체계적으로 설명하고자 한다.

1. 3D 장면 표현

1.1 포인트 클라우드(Point Cloud)

포인트 클라우드는 3D 장면이나 물체의 기본 표면을 3D 공간의 점 집합으로 나타낸다. 이러한 집합은 xyz 좌표 위치를 저장하며, 선택적으로 색상, 강도, 표면 법선과 같은 정보를 포함할 수 있다. 포인트 클라우드는 가볍고 직관적이지만, 무질서하고 구조화되지 않은 특성으로 인해 딥러닝을 활용할 때 어려움이 있다[18,19].

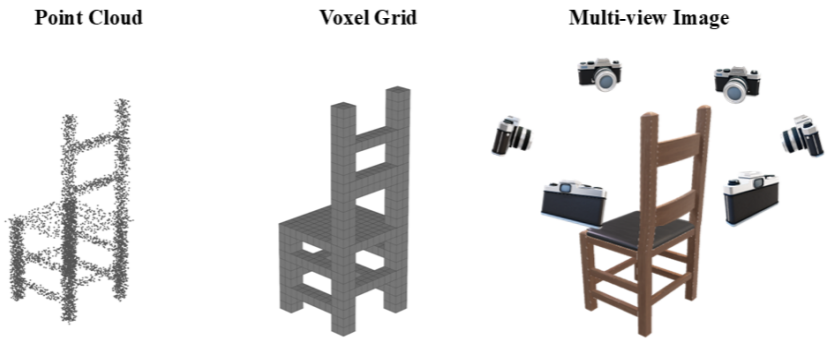


그림 3. 3D 장면 표현의 시각화
Fig. 3. Visualization of 3D scene representations

1.2 복셀 그리드(Voxel Grid)

연속적이고 불규칙한 기하학적 데이터(예: 포인트 클라우드)를 이산 격자로 변환해 얻은 구조화된 표현이다. 3차원 격자의 각 셀을 복셀(voxel)이라 하며, 각 복셀에는 색상·밀도·의미(semantic) 등의 속성을 저장할 수 있어 정밀한 볼류메트릭 표현이 가능하다[19,20]. 구조화된 격자를 사용하므로 개념과 구현이 단순하다는 장점이 있으나, 장면의 점유 영역뿐 아니라 비점유 영역까지 동일하게 표현하므로 메모리 효율이 떨어진다는 문제점이 있다[21].

1.3 다중 시점 이미지(Multi-view image)

다중 시점 이미지는 서로 다른 위치와 방향에서 촬영된 영상으로부터 3D 객체나 장면을 표현한다. 이미지 기반이므로 풍부한 의미적 특성을 추출할 수 있으며, 2D 멀티모달 모델을 직접 적용해 베이스 모델의 언어 이해 및 추론 능력을 활용할 수 있다. 또한 다중 시점 이미지와 카메라 파라미터를 이용해 장면을 Neural Radiance Field(NeRF)[22]나 3D Gaussian Splatting(3DGS)[23]과 같은 3D 표현으로 재구성할 수 있다. NeRF[22]는 3D 장면을 신경망에 인코딩하여, 3D 공간 좌표와 시야 방향을 입력으로 각 지점의 색상과 밀도를 예측하는 방식이다. 반면 3DGS[23]는 장면을 가우시안들의 집합으로 명시적으로 표현하며, 각 가우시안은 공간적 위치, 크기, 방향, 색상 등의 매개변수를 포함해 장면의 기하와 외관을 효율적으로 근사한다. 이러한 명시적 표현으로 인해 3DGS[23]는 NeRF[22]에 비해 학습 및 렌더링이 빠르고, 실시간 뷰 합성이 가능하다는 장점이 있다.

2. 단일 태스크 학습 기법(Single-task Learning)

2.1 3D 고밀도 캡셔닝(3D Dense Captioning)

Scan2Cap[8]은 3D 고밀도 캡셔닝의 선도 연구로서, 장면 내에서 검출된 객체와 객체 간 관계를 그래프 구조로 표현하여 학습한 뒤, GRU 기반의 맥락 인지형 어텐션 캡셔닝 모듈을 통해 객체별 설명을 생성한다. 이후 제안된 MORE[24]는 Scan2Cap[8]을 확장하여 다차수(multi-order) 관계를 점진적으로 그래프에 인코딩함으로써, 보다 복잡하고 풍부한 객체 간 공간적 상호작용을 정교하게 포착한다. 한편, Scan2Cap[8]과 MORE[24]가 그래프를 통해 객체 관계를 모델링한 것과 달리, SpaCap3D[25]는 트랜스포머 기반 인코더-디코더 구조를 도입하여 객체를 토큰으로 표현하고, 토큰 간 공간 관계 학습과 객체 중심 디코더를 통해 공간성이 강화된 객체 설명을 생성한다. 또한 X-Trans2Cap[26]은 관계 모델링 대신 2D-3D 정보 융합에 초점을 맞추어, 텍스처나 색상 등 풍부한 2D 정보를 3D 포인트 클라우드에 효과적으로 전달하기 위한 교차 모달 교사-학생(teacher-student) 프레임워크를 제안한다. 여기서 교사 모델과 학생 모델의 트랜스포머 디코더 피처를 정렬하고, 생성된 객체 설명 간의 교차 엔트로피(Cross-Entropy) 손실을 통해 학습을 수행한다.

앞선 연구들은 별도의 객체 검출 모델을 사용하여 후보 객체를 먼저 예측한 뒤, 그 후보에 대한 설명을 생성하는 방식은 모델의 검출 성능에 대한 의존성이 높아 오류 누적의 문제가 발생할 수 있다. 이를 해결하기 위해 Vote2Cap-

DETR[27] 및 Vote2Cap-DETR++[28]는 객체 위치 예측과 설명 생성을 동시에 처리하는 트랜스포머 기반의 end-to-end 프레임워크를 제안한다. 두 모델은 투표 기반으로 생성된 후보 객체 쿼리를 통해 트랜스포머 디코더가 두 태스크에 대한 예측을 동시에 수행한다. 기존 Vote2Cap-DETR[27]은 두 태스크가 요구하는 표현 수준이 상이함에도 불구하고 단일 객체 쿼리를 공유하여 처리함으로써 태스크 특화 표현 학습에 한계가 있었다. 이를 보완하기 위해 Vote2Cap-DETR++[28]는 쿼리를 위치 예측 쿼리와 설명 쿼리로 분리하여 각 태스크에 특화된 표현 학습을 가능하게 한다.

2.2 3D 질의응답(3D Question Answering, 3D QA)

ScanQA[9]는 트랜스포머 기반 프레임워크에서 3D 포인트 클라우드 표현을 언어 임베딩을 융합하고, 통합 특징을 객체 분류, 정답 분류, 객체 위치 추정 모듈에 각각 전달하여 3D VQA를 수행한다. SQA3D[10]는 3D SQA를 위해 ScanQA[9]의 구조에서 에이전트의 위치와 방향을 예측하는 보조 작업(auxiliary tasks)을 추가하여 확장하였다. 두 모델 모두 후보 답변들 사이에서 선택하는 정답 분류 접근을 사용한다. 한편, ScanQA[9]와 SQA3D[10]처럼 3D 포인트 클라우드 특징에 기반한 방식과 달리 2D 정보와 사전 학습된 2D VLM의 지식을 활용한 접근 방식도 제안되었다.

BridgeQA[29]는 3D 데이터 규모와 다양성의 부족으로 3D-VQA의 일반화가 제한되는 문제를 해결하고자 2D와 3D 정보를 통합하는 방식을 도입한다. 질문에 연관된 핵심 2D 뷰를 선별하고, 선정된 핵심 이미지와 3D 포인트 클라우드를 트윈 트랜스포머(Twin-Transformer)로 각각 처리하여, 사전 학습된 2D VLM 지식을 보존하면서 2D와 3D의 시각적 맥락을 융합한다. 마지막 단계에서 사전 학습된 VLM의 언어 디코더를 활용해 자유 형식의 답변을 생성함으로써 향상된 언어 능력을 달성한다.

SplatTalk[30]는 포인트 클라우드와 같은 별도의 3D 감독 신호 없이 다중 시점 이미지만을 활용하는 자가 지도 학습(self-supervised) 3D 가우시안 프레임워크를 제안한다. 2D VLM 토큰을 가우시안 표현의 잠재 특징으로 압축해 장면의 기하학적 구조와 의미 정보를 동시에 인코딩하고, 이를 LLM과 정렬된 시각 토큰으로 변환하여 zero-shot VQA를 가능하게 한다. 추론 단계에서는 엔트로피 기반 토큰 샘플링으로 정보량이 높은 가우시안을 선택해, 추가 파인튜닝 없이도 효율적인 성능을 달성한다.

cdViews[31]는 다중 시점 RGB 이미지만을 활용한다는 점에서는 SplatTalk[30]와 유사하지만, 3D 재구성 없이 단순히 다중 시점 2D 이미지와 2D VLM만을 사용한다. 3D 데이터가 희소한 상황에서 3D 모달리티와 언어의 직접 정렬은 어렵고 불안정하며, 이미 언어와 잘 정렬된 2D VLM의 특징을 활용하더라도 대규모 데이터와 복잡한 설계가 요구되는 한계가 있다. 이를 극복하기 위해 cdViews[31]는 3D-언어 직접 정렬을 회피하고, 질문 관련성, 정보 충분성, 뷰 다양성을 동시에 만족하는 뷰 선별만으로 zero-shot VQA를 수행하는 단순하면서도 효율적인 접근을 제안한다.

3. 다중 태스크 학습 기법(Multi-task Learning)

3.1 사전 학습 기법

3D-VLP[32]와 3D-VisTA[33]는 통합된 프레임워크에서 3D 장면 특징과 언어 표현을 트랜스포머로 융합하여, 장면–언어 정렬을 위한 사전 학습 방식을 도입한다. 사전 학습 후에는 다양한 태스크 헤드를 추가하고 미세조정(fine-tuning)을 통해 3D 시각 그라운딩(3D Visual Grounding), 3D 고밀도 캡셔닝, 3D 질의응답과 같이 다양한 3D 관련 태스크를 수행하도록 한다. 3D-VLP[32]는 문맥 인지형 공간–의미 정렬(context-aware spatial–semantic alignment)과 상호 마스킹 모델링(mutual masked modeling)을 도입해 멀티모달 특징의 의미 정렬을 강화하고 두 모달리티가 상호 보완적으로 작동하도록 설계하였다. 한편 3D-VisTA[33]는 마스크드 언어 모델링(Masked Language Modeling)과 마스크드 객체 모델링(Masked Object Modeling) 전략을 적용하여 3D 장면 특징과 언어 특징을 공동 임베딩 공간에서 효율적으로 융합한다.

3D-LLM[34]은 3D 포인트 클라우드 장면을 다중 시점 이미지로 렌더링한 뒤, 각 이미지로부터 언어와 정렬된 고밀도 2D 시각 피처를 추출한다. 추출된 다중 시점 2D 피처들을 융합하여 동일한 특징 공간 상의 3D 피처로 재구성한다. 이를 통해 사전 학습된 2D VLM을 백본으로 활용한 효율적인 3D-LLM 학습이 가능해지며, LLM을 활용하기 때문에 3D 캡셔닝과 3D 질의응답을 포함하여 대화, 작업 분해,

내비게이션 등 다양한 언어 생성 기반 3D 태스크를 수행할 수 있다.

Chat-3D v2[35]는 각 객체에 대해 속성 인지(attribute-aware) 및 관계 인지(relation-aware) 토큰을 학습하고, 이를 객체의 고유 식별자와 결합함으로써 LLM이 특정 객체를 명시적으로 참조하고 더 심도 있는 3D 공간 정보를 이해하도록 한다. 이를 통해 기존 VLM 기반 접근에 비해 LLM의 3D 공간 이해 능력을 강화하여, 3D-언어가 정렬된 모델에서 파인튜닝을 활용해 다양한 3D 장면 이해 과제에서 일관된 성능 향상을 달성한다.

Scene-LLM[36]은 1인칭 시점 정보와 장면 수준의 3D 정보를 통합하여 LLM에 직접 결합하는 3D 비전 - 언어 모델로, 3D 시각 정보와 LLM의 효과적인 정렬을 통해 3D 장면 이해 및 추론을 강화한다. 구체적으로, 포인트 - 복셀 기반의 하이브리드 3D 표현을 텍스트 임베딩 공간에 투영하여 LLM과 정렬하고, 이후 3D 프레임 및 장면 - 언어 데이터로 LLM과 프로젝션 레이어를 함께 미세 조정하여 사용자 명령을 정확하게 수행하도록 최적화한다. 이를 통해 다양한 언어 생성 태스크를 하나의 프레임워크로 처리한다.

3.2 합동 학습 기법(Joint Training)

LEO[13]는 1인칭 시점(egocentric) 기반 3D 이미지, 3D 포인트 클라우드, 텍스트를 입력으로 받아, 이들을 토큰화하고 통합된 시퀀스 형식으로 변환한 후, 이를 자기회귀적(autoregressive) 방식으로 처리하는 범용 프레임워크이다. 3D 캡셔닝, 질의응답, 대화, 작업 계획 등 다양한 언어 생성 기반 태스크를 단일 통합 모델 아키텍처를 통해 수행하며, 객체 및 장면 수준 특징 정렬을 언어 표현과 정렬함으로써 뛰어난 zero-shot 성능을 달성한다.

Chat-Scene[37]은 객체 식별자(object identifiers)를 도입하여 3D 장면 내 객체에 대한 효율적인 참조와 그라운딩을 가능하게 하며, 이를 활용해 다양한 3D 장면 - 언어 태스크를 하나의 통합된 질의응답 형식으로 변환하여 태스크 특화 헤드 없이 합동 학습만으로 수행한다. 또한 사전 학습된 파운데이션 모델(foundation model)로부터 추출한 객체 중심 멀티모달 표현 시퀀스를 사용하여 3D 장면을 효과적으로 표현한다.

LL3DA[38]는 포인트 클라우드, 텍스트 지시, 그리고 클릭·바운딩 박스와 같은 시각적 상호작용을 입력으로 받아 이를 통합하고, 자기회귀적 방식으로 다양한 언어 생성 기반 태스크를 단일 프레임워크로 수행하는 범용 3D 비전 - 언어 모델이다. 사전 학습된 3D 인코더와 LLM은 고정하고 멀티모달 트랜스포머(Q-Former)와 프로젝션 모듈만을 instruction tuning으로 미세 조정하여 3D 장면 표현과 언어 공간을 정렬한다. 이를 통해 태스크 특화 헤드 없이도 다양한 3D 태스크에서 효과적인 zero-shot 성능을 달성한다.

Inst3D-LMM[39]은 3D 포인트 클라우드와 다중 시점 RGB 이미지에서 추출한 기하·의미 정보를 인스턴스 단위로 융합하여 LLM에 입력하는 인스턴스 인지형(instance-aware) 3D 멀티모달 모델이다. 다중 시점 2D 의미론적 open-vocabulary 정보를 3D 기하학적 특징에 주입함으로써 객체의 세부 속성과 객체 간 공간 관계를 정밀하게 반영한 인스턴스 수준 토큰을 생성하며, 이를 기반으로 3D 캡셔닝, 질의응답, 시각적 그라운딩 등 다양한 태스크를 단일 지시(instruction) 형식으로 공동 학습(joint instruction tuning)한다.

Ⅴ. 성능 평가 지표 및 최신 연구 성능 비교

본 장에서는 주요 성능 평가 지표와 대표 벤치마크에서의 최신 연구 성능을 비교·분석한다.

1. 주요 성능 평가 지표

언어 생성 기반 태스크의 성능 평가는 생성된 텍스트의 품질과 3D 객체 위치 정확도를 종합적으로 평가하기 위해 다양한 평가 지표를 사용한다. 대표적으로 생성된 텍스트의 품질은 Exact Match(EM), BLEU, ROUGE, METEOR, 그리고 CIDEr를 사용하고, 3D 고밀도 캡셔닝과 같이 객체 검출의 정확성도 고려하는 경우에는 생성된 텍스트의 품질과 검출된 객체의 정확성을 통합적으로 평가하기 위해 m@kIoU 지표를 사용하며, 표 2에서 각 평가 지표의 목적, 적용 도메인, 그리고 장단점을 요약하여 정리하였고, 그림 4에서는 각 평가 지표 결과에 대한 정성적 평가를 나타내기 위해 Scan2Cap(상단)과 ScanQA(하단) 데이터셋의 예시를 나타내었다.

표 2. 주요 성능 평가 지표
Table 2. Evaluation Metrics

Measure	Aim	Domain	Strengths	Weaknesses
Exact Match (EM)	precision	QA	Suitable for exact answer matching	Not suitable for open-ended response evaluation
BLEU[40]	n-gram precision	Machine translation	Computationally efficient	Sensitive to wording and sentence length
ROUGE[41]	n-gram recall	Text summarization	Computationally simple	Limited handling of paraphrases and synonyms
METEOR[42]	uni-gram precision, uni-gram recall	Machine translation	Consider stems and synonyms, Strong human correlation	Relies on external linguistic resources
CIDEr[43]	TF-IDF	Visual captioning	Strong human correlation	Biased toward consensus wording across references
m@kIoU[8]	IoU	3D dense captioning	Evaluates both localization and captioning	over-dependence on the threshold k

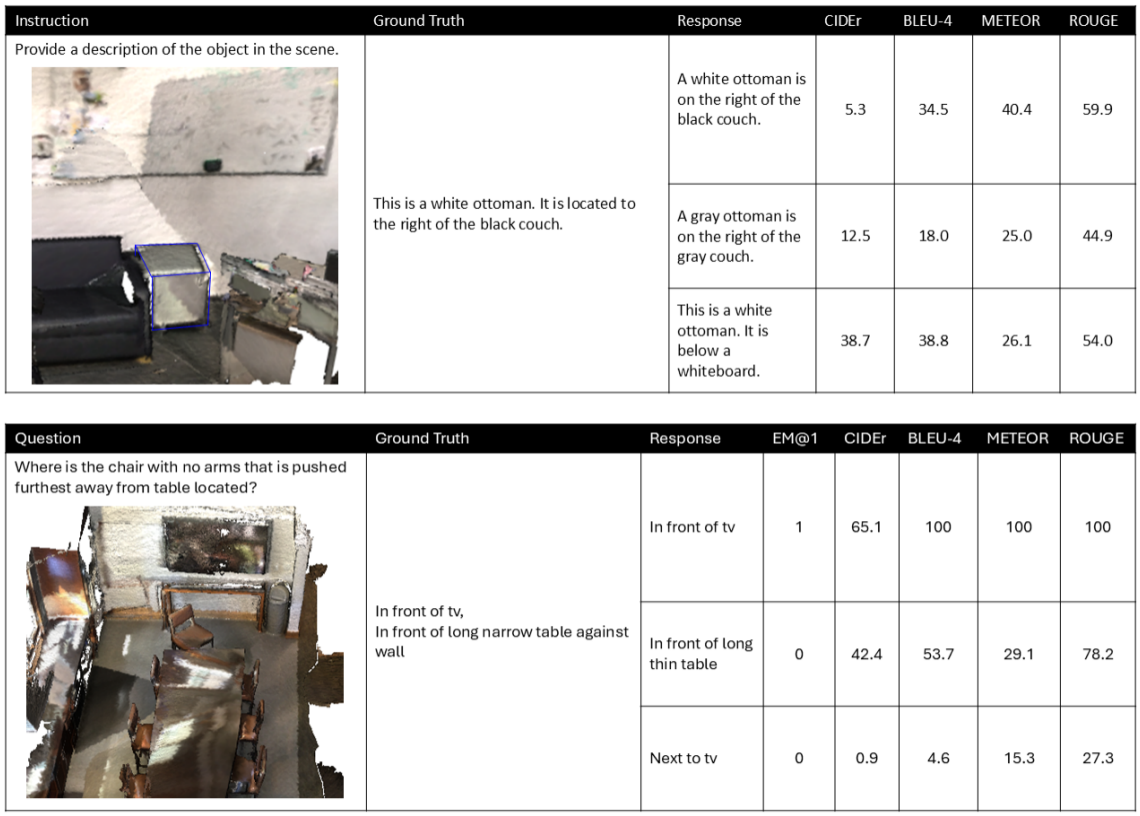


그림 4. Scan2Cap 및 ScanQA 데이터셋 기반 3D 장면 이해 태스크의 정성적 예시. 상단은 3D 고밀도 캡셔닝(Scan2Cap), 하단은 3D 질의응답(ScanQA) 결과를 보여주며, 각 응답에 대한 정량적 성능 평가 지표를 함께 제시한다.
Fig. 4. Qualitative examples of 3D scene understanding tasks based on the Scan2Cap and ScanQA datasets. The top illustrates dense 3D captioning results (Scan2Cap), and the bottom shows 3D question answering results (ScanQA), along with quantitative evaluation metrics for each response.

Exact Match(EM)는 정답과의 완전 일치 여부를 기준으로 하는 정밀도 기반 지표로 3D 질의응답 태스크 성능 평가에서 사용되며, 주로 EM@1 또는 EM@10 점수를 사용한다. EM@K는 예측 신뢰도 상위 K개의 답변 중 참조 답변과 정확하게 일치하는 답변이 하나라도 존재하는 비율을 의미하고, 동의어나 의역 등을 허용하지 않고 정확하게 일치하는 경우만 허용한다. 따라서 그림 4의 하단 예시에서 확인할 수 있듯이, “In front of tv”라는 응답은 정답에 포함된 핵심 객체와 정확히 일치하여 EM@1 점수 1을 획득한 반면, “In front of brown table” 또는 “Next to tv”과 같이 부분적으로 관련되거나 의미적으로 유사하지만, 정답과의 완전 일치가 아니기 때문에 모두 0점으로 처리된다. 이처럼 EM 점수는 폐쇄형 답변 평가에 적합하지만, 개방형 응답에는 한계가 있다.

BLEU[40]와 ROUGE[41]는 각각 n-그램 정밀도와 재현율을 기반으로 생성된 문장과 참조 문장 간의 유사성을 측정하는 평가 지표이다. 그림 4의 하단 예시와 같이, EM 점수에서는 0으로 계산된 응답도 BLEU와 ROUGE는 문장 간 n-그램 중첩을 고려함으로써 표현 수준의 유사성을 반영한 것을 확인할 수 있다. 표 2에서 나타난 바와 같이, 이러한 n-그램 기반 지표들은 비교적 간단하고 계산 효율적이라는 장점이 있으나, 표면적 표현 유사성에 중점을 두기 때문에 문장의 표현 방식이나 길이에 민감하며, 동의어나 의역에 대한 처리가 제한적이라는 한계를 가진다. 그림 4의 상단 예시를 보면, 의미적 유사도는 “A white ottoman is on the right of the black couch.”가 정답과 더 유사하지만, “This is a white ottoman. It is below a whiteboard.”는 문장 표현 중복으로 인해 높은 BLEU 점수를 가진다. 이는 n-그램 기반 지표들이 표현 유사성에는 민감하지만 의미적 정확성을 충분히 반영하지 못함을 보여준다.

METEOR는 이러한 한계를 완화하기 위해 어간 일치 및 동의어 정보를 함께 고려함으로써 의미적 유사성을 보다 유연하게 반영한다. 단일-그램(uni-gram) 정밀도(precision)와 재현율(recall) 모두 고려하되 재현율에 더 중점을 둔다. 동의어나 어간 일치 등을 함께 고려해 의미적 유사성, 어순, 그리고 문장 구조를 반영하므로 사람 평가와의 상관관계가 상대적으로 더 높다. 이는 그림 4의 상단 예시처럼 의미적 유사도가 높은 응답의 METEOR 점수가 더 높게 기록된 것을 통해 확인할 수 있다. 그러나 표 2에서 나타난 바와 같이, 외부 언어 자원에 의존한다는 점에서 평가 환경에 제약이 존재한다.

CIDEr[43]는 시각적 캡셔닝에 사용되는 평가 지표로, 생성된 캡션과 정답 간의 문장 수준에서 Term Frequency-Inverse Document Frequency(TF-IDF)를 통해 코사인 유사도를 계산한다. 이때 각 n-그램은 코퍼스 전반에서의 등장 빈도에 따라 가중치가 부여되며, 다수의 참조 문장들에서 빈도가 높은 구문일수록 상대적으로 낮은 가중치를 갖는다. 이러한 특성으로 인해 CIDEr는 사람 평가와 높은 상관관계를 보이는 장점이 있으나, 표에서 요약된 바와 같이, 코퍼스 내 다수 참조 문장의 합의된 표현(consensus wording)에 편향되는 경향을 가진다. 그 결과, 의미적으로 유사해도 표현 방식이 참조 문장과 상이한 경우에는 낮은 점수를 받을 수 있으며, 반대로 표면적 문장 구조가 유사한 경우 의미적 오류가 존재하더라도 상대적으로 높은 점수를 얻을 수 있다. 그림 4의 상단 예시와 같이, “A white ottoman is on the right of the black couch.”가 의미적 유사도는 가장 높지만 표면적 문장 구조가 유사한 “This is a white ottoman. It is below a whiteboard.”의 CIDEr 점수가 가장 높다.

마지막으로, m@kIoU[8]는 3D 고밀도 캡셔닝 태스크에서 객체 위치 정확도와 캡셔닝 품질을 함께 평가하기 위해 사용되며, 예측된 바운딩 박스와 실제 바운딩 박스 간의 IoU가 k 이상인 경우에만 캡셔닝 결과를 유효한 예측으로 간주한다. 이러한 방식은 위치 추정과 언어 생성 성능을 동시에 고려할 수 있다는 장점이 있으나, IoU 임계값 k의 설정에 따라 평가 결과가 민감하게 변할 수 있다는 한계를 가진다.

2. 최신 연구 성능 비교

본 절에서는 태스크별 최신 연구의 성능을 비교하고 결과를 분석한다. 분석 범위는 정량 지표가 정의된 3D 고밀도 캡셔닝(3D Dense Captioning)과 3D 질의응답(3D QA)으로 한정한다. 3D 고밀도 캡셔닝(3D Dense Captioning)의 경우 Scan2Cap와 Nr3D 벤치마크 데이터에서의 결과를, 3D QA의 경우 ScanQA와 SQA3D 데이터셋의 결과를 제시한다. 각 결과는 모델의 태스크 구성 방식에 따라 STL(단일)과 MTL(다중)로 구분해 기술하며, 입력 3D 장면 표현 유형을

함께 보고한다.

2.1 3D 고밀도 캡셔닝(3D Dense Captioning)

표 3과 표 4는 각각 Scan2Cap과 Nr3D 검증(validation) 데이터셋에서의 3D 고밀도 캡셔닝 성능 비교 결과를 보여준다. 두 벤치마크 모두에서 LLM 기반 다중 태스크 학습 모델이 기존 단일 태스크 모델보다 우수한 성능을 보인다. 먼저 표 3의 Scan2Cap 결과를 보면, LLM 중심의 다중 태스크 모델이 대부분의 지표에서 상위 성능을 기록했다. 특히 인스턴스 인식 정보와 객체 식별 정보 등 객체 수준 특징을 LLM과 정렬하는 3D-언어 정렬 기반 접근(LEO, Chat-Scene, Inst3D-LMM)이 큰 성능 향상을 보였으며, Inst3D-LMM은 CIDEr@0.5, LEO는 ROUGE@0.5와 METEOR@0.5에서 최고 성능을 달성했다. 이는 객체 수준 표현의 명시적 모델링과 LLM의 언어 생성 능력 활용이 고밀도 캡셔닝 성능 향상에 핵심적임을 시사한다.

표 3. Scan2Cap validation 데이터셋에서의 성능 평가
Table 3. Experimental results on the Scan2Cap validation set

Method	Task Setting	LLM	Scan2Cap (val)			
			C@0.5	B-4@0.5	M@0.5	R@0.5
Scan2Cap	STL		35.2	22.4	21.4	43.6
Scan2Cap (w/ 2 D)[8]			39.1	23.3	22.0	44.8
MORE	STL		39.0	23.0	21.7	44.3
MORE (w/ 2 D)[2-4]			40.9	22.9	21.7	44.4
X-Trans2Cap	STL		41.5	23.8	21.9	45.0
X-Trans2Cap (w/2 D)[2-6]			43.9	25.1	22.5	45.3
SpaCap3D	STL		42.8	25.4	22.8	45.7
SpaCap3D (w/ 2 D)[2-5]			44.0	25.3	22.3	45.4
Vote2Cap-DETR	STL		73.8	38.2	26.6	54.7
Vote2Cap-DETR (w/2 D)[2-7]			70.6	35.7	25.5	52.3
Vote2Cap-DETR+ +	STL		78.2	**39.7**	26.9	55.5
Vote2Cap-DETR+ + (w/ 2 D)[2-8]			74.4	37.2	26.2	53.3
3D-VLP	MTL		50.0	31.9	24.5	51.2
3D-VLP (w/ 2 D)[32]			54.9	32.3	24.8	51.5
3D-VisTA[33]	MTL		66.9	34.0	27.1	54.3
LEO[13]	MTL	V	72.4	38.2	27.9	58.1
Chat-3D v2[35]	MTL	V	63.9	31.8	22.3	50.2
LL3DA[38]	MTL	V	65.2	36.8	26.0	55.1
Chat-Scene[37]	MTL	V	77.1	36.3	-	-
Inst3D-LMM[39]	MTL	V	**79.7**	38.3	27.5	57.2

표 4. Nr3D validation 데이터셋에서의 성능 평가
Table 4. Experimental results on the Nr3D validation set

Method	Task Setting	LLM	Nr3D (val)			
			C@0.5	B-4@0.5	M@0.5	R@0.5
X-Trans2Cap	STL		31.0	18.7	22.2	49.9
X-Trans2Cap (w/ 2 D)[2-6]			33.6	19.3	22.3	50.0
SpaCap3D	STL		31.4	19.0	22.2	49.8
SpaCap3D (w/ 2 D)[2-5]			33.7	19.9	22.6	50.5
Vote2Cap-DETR (w/ 2 D)[2-7]	STL		45.5	26.9	25.4	54.8
Vote2Cap-DETR+ + (w/ 2 D)[2-8]	STL		47.6	28.4	25.6	54.8
LL3DA[38]	MTL	V	**51.2**	**28.8**	**25.9**	**56.6**

한편, 객체 검출과 설명 생성을 통합한 end-to-end 방식인 Vote2Cap-DETR++는 전통적인 'Detect-then-Describe' 접근보다 전반적으로 높은 성능을 보였으며, 3D 입력만으로도 높은 BLEU-4@0.5를 기록해 위치 예측과 캡셔닝의 합동 최적화 효과를 입증한다.

Nr3D 데이터셋에서도 유사한 경향이 관찰되었고, LLM을 활용한 다중 태스크 모델(LL3DA)이 CIDEr, BLEU, METEOR 등 모든 지표에서 상대적으로 높은 성능을 보였다.

표 5. ScanQA validation 데이터셋에서의 성능 평가
Table 5. Experimental results on the ScanQA validation set

Method	Task Setting	LLM	ScanQA (val)				
			EM@1	BLEU-4	ROUGE	METEOR	CIDEr
ScanQA[9]	STL		21.1	10.1	33.3	13.1	64.9
BridgeQA[29]	STL		27.0	-	-	-	-
SplatTalk[30]	STL		17.1	-	32.7	14.2	61.7
3D-VLP[32]	MTL		21.7	11.2	34.5	13.5	67.0
3D-VisTA[33]	MTL		22.4	10.4	35.7	13.9	69.6
3D-LLM[34]	MTL	V	20.5		35.7	14.5	69.4
Chat-3D v2[35]	MTL	V	21.1	7.3	40.1	16.1	77.1
Scene-LLM[36]	MTL	V	27.2	-	40.0	16.6	80.0
LEO[13]	MTL	V	24.5	13.2	**49.2**	**20.0**	**101.4**
LL3DA[38]	MTL	V	-	-	37.3	15.9	76.8
Chat-Scene[37]	MTL	V	21.6	14.3	41.6	18.0	87.7
Inst3D-LLM[39]	MTL	V	24.6	**14.9**	42.6	18.4	88.6

표 6. ScanQA test 데이터셋에서의 성능 평가
Table 6. Experimental results on the ScanQA test set with objects

Method	Task Setting	LLM	ScanQA (test with object)				
			EM@1	BLEU-4	ROUGE	METEOR	CIDEr
ScanQA[9]	STL		23.5	12.0	34.3	13.6	67.3
BridgeQA[29]	STL		31.3	**24.1**	43.3	**16.5**	83.8
cdViews[31]	STL	V	**35.0**	-	**49.7**	-	**102.8**
3D-VLP[32]	MTL		24.6	11.2	36.0	14.2	70.2
3D-VisTA[33]	MTL		27.0	16.0	38.6	15.2	76.6
3D-LLM[34]	MTL	V	19.1	11.6	35.3	14.9	69.6
LL3DA[38]	MTL	V	-	13.5	37.3	15.9	76.8

표 7. SQA3D test 데이터셋에서의 성능 평가
Table 7. Experimental results on the SQA3D test set

Method	Task Setting	LLM	SQA3D (test)						
			What	Is	How	Can	Which	Other	Avg.
ScanQA[9]	STL		28.6	65.0	**47.3**	66.3	43.9	42.9	45.3
SQA3D[10]	STL		34.5	66.1	42.4	69.5	43.0	46.4	47.2
SplatTalk[30]	STL	V	-	-	-	-	-	-	26.1
cdViews[31]	STL	V	-	-	-	-	-	-	**56.9**
3D-VisTA[33]	MTL		34.8	63.3	45.4	69.8	47.2	48.1	48.5
3D-LLM[34]	MTL	V	35.0	66.0	47.0	69.0	**48.0**	46.0	48.1
Scene-LLM[36]	MTL	V	40.9	**69.1**	45.0	70.8	47.2	52.3	54.2
LEO[13]	MTL	V	**46.8**	64.1	47.0	60.8	44.2	**54.3**	52.9
Chat-Scene[37]	MTL	V	-	-	-	-	-	-	54.6

2.2 3D 질의응답(3D Question Answering, 3D QA)

표 5와 표 6에서는 ScanQA 데이터셋에서의 3D 시각적 질의응답 성능, 그리고 표 7에서는 SQA3D 데이터셋에서의 3D 상황적 질의응답 성능 결과를 나타낸다. 3D 고밀도 캡셔닝 결과와 유사하게 3D 정보를 LLM에 직접 통합해 추론을 강화한 방식(Scene-LLM, LEO, Inst3D-LMM)이 전반적으로 상위 성능을 기록하며, 이는 언어 생성과 추론 과정에서의 일관된 성능 우위와 함께, 다양한 3D 언어 태스크에 대한 우수한 일반화 능력을 보여준다.

한편, 명시적인 3D 입력 없이 다중 시점 이미지와 사전 학습된 2D 대형 비전‐언어 모델(LVLM)의 지식을 효과적으로 활용한 cdViews 역시 경쟁력 있는 성능을 보였다. 특히 ScanQA test 데이터셋에서는 EM@1, ROUGE, CIDEr 지표에서, SQA3D 데이터셋에서는 EM@1에서 가장 높은 성능을 달성하였다. 이러한 결과는 객체의 정밀한 위치 추정이 중요한 3D 고밀도 캡셔닝과 달리, 3D 질의응답 태스크에서는 질의에 필요한 핵심 정보만을 선별적으로 제공하는 것만으로도 정확한 답변 생성이 가능함을 시사한다.

Ⅵ. 결론 및 향후 연구 방향

본 논문은 언어 생성 기반 3D 장면 이해의 데이터셋, 태스크 정의, 그리고 최신 연구 동향을 종합적으로 정리하고, 각 방법의 성능과 특성을 비교·분석하였다. LLM과 VLM의 발전으로 언어 이해와 추론 능력이 향상되면서 보다 구체적이고 정확한 자연어 응답을 생성할 수 있게 되었으나, 3D 장면과 언어 주석이 결합된 고품질의 대규모 데이터의 부족과 제한된 장면 다양성으로 인해 여전히 3D 공간으로의 일반화에는 취약하다. 또한 2D/3D 인지(perception) 모델에 대한 의존성으로 인해 장면 구조와 객체 간 관계가 복잡해질수록 성능이 급격히 저하된다. 아울러 대부분의 LLM이 2D 데이터 기반으로 사전 학습되었기 때문에 3D와 언어의 정렬이 어렵고, 이로 인해 3차원적 지능과 추론에는 근본적 한계가 존재한다. 더 나아가 실제 3D 공간의 복잡도에 비해 3D 표현 단계에서는 정보 손실과 연산 효율을 위한 단순화가 불가피하여 활용 가능한 정보가 제한되며, 대규모 3D 장면으로 확장할수록 처리해야 할 정보량과 연산 비용이 급격히 증가한다는 문제가 있다.

이러한 한계를 해소하기 위해서는, 실제 데이터 수집만으로 확보하기 어려운 규모와 다양성을 보완할 수 있도록 시뮬레이션 환경에서 다양한 장면을 포함하는 대규모 합성 데이터를 구축하고 LLM의 능력을 활용해 복잡한 맥락을 반영한 풍부한 언어 주석을 자동 생성할 필요가 있다. 시뮬레이션 환경에서는 객체 위치, 바운딩 박스 등 정확한 기하 정보와 주석의 품질과 일관성이 보장되기 때문에 노이즈가 적다는 장점도 있다. 아울러 3D 데이터를 처리할 때 사전 학습된 2D VLM에 의존해 활용하기보다는 3D 장면 표현을 직접적으로 LLM과 통합하여 3차원 공간에서의 추론을 강화하는 연구가 필수적이다. 이러한 접근은 3D 이해와 언어 정렬을 동시에 강화하여 일반화 성능을 높이고, 결과적으로 LLM의 3D 태스크에서의 심층적 추론 능력 향상을 기대할 수 있다.

참 고 문 헌 (References)

[1] P. Fung, Y. Bachrach, A. Celikyilmaz, K. Chaudhuri, D. Chen, W. Chung, E. Dupoux, H. Gong, H. Jégou, A. Lazaric et al., “Embodied ai agents: Modeling the world,” arXiv preprint arXiv:2506.22355, 2025, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2506.22355

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual Instruction Tuning,” Advances in Neural Information Processing Systems, Vol. 36, pp. 34892‐34916, Dec. 2023. https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf

[3] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigpt-4: Enhancing vision-language understanding with advanced large language models,” arXiv preprint arXiv:2304.10592, 2023, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2304.10592

[4] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, “Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond,” arXiv preprint arXiv:2308.12966, 2023, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2308.12966

[5] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu et al., “Llava-onevision: Easy visual task transfer,” arXiv preprint arXiv:2408.03326, 2024, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2408.03326

[6] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved Baselines with Visual Instruction Tuning,” Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition, pp. 26296–26306, June 2024. https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html

[7] X. Ma, B. Smart, Y. Bhalgat, S. Chen, X. Li, J. Ding, J. Gu, D. Z. Chen, S. Peng, J. Bian et al., "When LLMs step into the 3D World: A Survey and Meta-Analysis of 3D Tasks via Multi-modal Large Language Models," arXiv preprint arXiv:2405.10255, 2024, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2405.10255

[8] Z. Chen, A. Gholami, M. Nießner, and A. X. Chang, "Scan2Cap: Context-Aware Dense Captioning in RGB-D Scans," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3193-3203, June 2021. https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Scan2Cap_Context-Aware_Dense_Captioning_in_RGB-D_Scans_CVPR_2021_paper.html

[9] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, "ScanQA: 3D Question Answering for Spatial Scene Understanding," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19129-19139, June 2022. https://openaccess.thecvf.com/content/CVPR2022/html/Azuma_ScanQA_3D_Question_Answering_for_Spatial_Scene_Understanding_CVPR_2022_paper.html

[10] X. Ma, S. Yong, Z. Zheng, Q. Li, Y. Liang, S. Zhu, and S. Huang, "SQA3D: Situated Question Answering in 3D Scenes," arXiv preprint arXiv:2210.07474, 2022, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2210.07474

[11] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5828–5839, July 2017. https://openaccess.thecvf.com/content_cvpr_2017/html/Dai_ScanNet_Richly-Annotated_3D_CVPR_2017_paper.html

[12] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes," European Conference on Computer Vision, Springer, pp. 422–440, Nov. 2020.
doi: https://doi.org/10.1007/978-3-030-58452-8_25

[13] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S. Zhu, B. Jia, and S. Huang, "An Embodied Generalist Agent in 3D World," arXiv preprint arXiv:2311.12871, 2023, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2311.12871

[14] J. Wald, A. Avetisyan, N. Navab, F. Tombari, and M. Nießner, "RIO: 3D Object Instance Re-Localization in Changing Indoor Environments," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7658-7667, Oct. 2019. https://openaccess.thecvf.com/content_ICCV_2019/html/Wald_RIO_3D_Object_Instance_Re-Localization_in_Changing_Indoor_Environments_ICCV_2019_paper.html

[15] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A Universe of Annotated 3D Objects," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13142–13153, June 2023. https://openaccess.thecvf.com/content/CVPR2023/html/Deitke_Objaverse_A_Universe_of_Annotated_3D_Objects_CVPR_2023_paper.html

[16] X. Linghu, J. Huang, X. Niu, X. S. Ma, B. Jia, and S. Huang, "Multi-modal Situated Reasoning in 3D Scenes," Advances in Neural Information Processing Systems, Vol. 37, pp. 140903–140936, Dec. 2024.
doi: https://doi.org/10.52202/079017-4473

[17] G. Baruch, Z. Chen, A. Dehghan, T. Dimry, Y. Feigin, P. Fu, T. Gebauer, B. Joffe, D. Kurz, A. Schwartz et al., "ARKitScenes: A Diverse Real-World Dataset For 3D Indoor Scene Understanding Using Mobile RGB-D Data," arXiv preprint arXiv:2111.08897, 2021, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2111.08897

[18] Z. Wang, "3D Representation Methods: A Survey," arXiv preprint arXiv:2410.06475, 2024, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2410.06475

[19] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep Learning for 3D Point Clouds: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 12, pp. 4338-4364, Dec. 2021.
doi: https://doi.org/10.1109/TPAMI.2020.3005434

[20] C. R. Qi, "Deep Learning on 3D Data," 3D Imaging, Analysis and Applications, Springer, pp. 513-566, Sep. 2020.
doi: https://doi.org/10.1007/978-3-030-44070-1_11

[21] E. Ahmed, A. Saint, A. E. R. Shabayek, K. Cherenkova, R. Das, G. Gusev, D. Aouada, and B. Ottersten, "A survey on Deep Learning Advances on Different 3D Data Representations," arXiv preprint arXiv:1808.01462, 2018, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.1808.01462

[22] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: representing scenes as neural radiance fields for view synthesis," Communications of the ACM, Vol. 65, No. 1, pp. 99–106, Dec. 2021.
doi: https://doi.org/10.1145/3503250

[23] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering." ACM Transactions on Graphics, Vol. 42, No. 4, July 2023.
doi: https://doi.org/10.1145/3592433

[24] Y. Jiao, S. Chen, Z. Jie, J. Chen, L. Ma, and Y.-G. Jiang, "MORE: Multi-Order RElation Mining for Dense Captioning in 3D Scenes," European Conference on Computer Vision, Springer, pp. 528–545, Nov. 2022.
doi: https://doi.org/10.1007/978-3-031-19833-5_31

[25] H. Wang, C. Zhang, J. Yu, and W. Cai, "Spatiality-guided Transformer for 3D Dense Captioning on Point Clouds," arXiv preprint arXiv:2204.10688, 2022, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2204.10688

[26] Z. Yuan, X. Yan, Y. Liao, Y. Guo, G. Li, S. Cui, and Z. Li, "X-Trans2Cap: Cross-Modal Knowledge Transfer Using Transformer for 3D Dense Captioning," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8563–8573, June

2022. https://openaccess.thecvf.com/content/CVPR2022/html/Yuan_X-Trans2Cap_Cross-Modal_Knowledge_Transfer_Using_Transformer_for_3D_Dense_Captioning_CVPR_2022_paper.html

[27] S. Chen, H. Zhu, X. Chen, Y. Lei, G. Yu, and T. Chen, "End-to-End 3D Dense Captioning With Vote2Cap-DETR," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11124 – 11133, June 2023. https://openaccess.thecvf.com/content/CVPR2023/html/Chen_End-to-End_3D_Dense_Captioning_With_Vote2Cap-DETR_CVPR_2023_paper.html

[28] S. Chen, H. Zhu, M. Li, X. Chen, P. Guo, Y. Lei, G. Yu, T. Li, and T. Chen, "Vote2Cap-DETR++: Decoupling Localization and Describing for End-to-End 3D Dense Captioning," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 46, No. 11, pp. 7331 – 7347, Nov. 2024.
doi: https://doi.org/10.1109/TPAMI.2024.3387838

[29] W. Mo and Y. Liu, "Bridging the Gap between 2D and 3D Visual Question Answering: A Fusion Approach for 3D VQA," Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, No. 5, pp. 4261 – 4268, Mar. 2024.
doi: https://doi.org/10.1609/aaai.v38i5.28222

[30] A. Thai, S. Peng, K. Genova, L. Guibas, and T. Funkhouser, "SplatTalk: 3D VQA with Gaussian Splatting," arXiv preprint arXiv:2503.06271, 2025, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2503.06271

[31] F. Wang, S. Yu, J. Wu, J. Tang, H. Zhang, and Q. Sun, "3D Question Answering via only 2D Vision-Language Models," arXiv preprint arXiv:2505.22143, 2025, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2505.22143

[32] Z. Jin, M. Hayat, Y. Yang, Y. Guo, and Y. Lei, "Context-Aware Alignment and Mutual Masking for 3D-Language Pre-Training," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10984 – 10994, June 2023. https://openaccess.thecvf.com/content/CVPR2023/html/Jin_Context-Aware_Alignment_and_Mutual_Masking_for_3D-Language_Pre-Training_CVPR_2023_paper.html

[33] Z. Zhu, X. Ma, Y. Chen, Z. Deng, S. Huang, and Q. Li, "3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2911 – 2921, Oct. 2023. https://openaccess.thecvf.com/content/ICCV2023/html/Zhu_3D-VisTA_Pre-trained_Transformer_for_3D_Vision_and_Text_Alignment_ICCV_2023_paper.html

[34] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3D-LLM: Injecting the 3D World into Large Language Models," Advances in Neural Information Processing Systems, Vol. 36, pp. 20482 – 20494, Dec. 2023. https://proceedings.neurips.cc/paper_files/paper/2023/file/413885e70482b95dcbeeddc1daf39177-Paper-Conference.pdf

[35] H. Huang, Z. Wang, R. Huang, L. Liu, X. Cheng, Y. Zhao, T. Jin, and Z. Zhao, "Chat-3D v2: Bridging 3D Scene and Large Language Models with Object Identifiers," CoRR, 2023, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2312.08168

[36] R. Fu, J. Liu, X. Chen, Y. Nie, and W. Xiong, "Scene-LLM: Extending Language Model for 3D Visual Understanding and Reasoning," arXiv preprint arXiv:2403.11401, 2024, (accessed Dec. 3, 2025).
doi: https://doi.org/10.48550/arXiv.2403.11401

[37] H. Huang, Y. Chen, Z. Wang, R. Huang, R. Xu, T. Wang, L. Liu, X. Cheng, Y. Zhao, J. Pang et al., "Chat-Scene: Bridging 3D Scene and Large Language Models with Object Identifiers," Advances in Neural Information Processing Systems, Vol. 37, pp. 113991-114017, Dec. 2024.
doi: https://doi.org/10.52202/079017-3620

[38] S. Chen, X. Chen, C. Zhang, M. Li, G. Yu, H. Fei, H. Zhu, J. Fan, and T. Chen, "LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding Reasoning and Planning," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26428 – 26438, June 2024. https://openaccess.thecvf.com/content/CVPR2024/html/Chen_LL3DA_Visual_Interactive_Instruction_Tuning_for_Omni-3D_Understanding_Reasoning_and_CVPR_2024_paper.html

[39] H. Yu, W. Li, S. Wang, J. Chen, and J. Zhu, "Inst3D-LMM: Instance-Aware 3D Scene Understanding with Multi-modal Instruction Tuning," Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 14147 – 14157, June 2025. https://openaccess.thecvf.com/content/CVPR2025/html/Yu_Inst3D-LMM_Instance-Aware_3D_Scene_Understanding_with_Multi-modal_Instruction_Tuning_CVPR_2025_paper.html

[40] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311 – 318, July 2002.
doi: https://doi.org/10.3115/1073083.1073135

[41] C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," Text Summarization Branches Out, pp. 74 – 81, July 2004. https://aclanthology.org/W04-1013/

[42] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65 – 72, June 2005. https://aclanthology.org/W05-0909/

[43] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-Based Image Description Evaluation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566 – 4575, June 2015. https://openaccess.thecvf.com/content_cvpr_2015/html/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.html

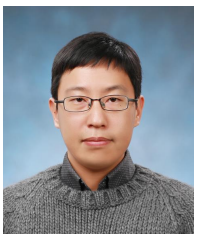
저 자 소 개

김 연 의

- 2024년 2월 : 이화여자대학교 전자전기공학과 졸업(공학학사)
- 2024년 3월 ~ 현재 : 이화여자대학교 전자전기공학과 석사과정
- ORCID : https://orcid.org/0009-0004-3735-0247
- 주관심분야 : 딥러닝, 3D 장면 이해 및 추론

양 진 우

- 2025년 8월 : 이화여자대학교 전자전기공학과 졸업(공학학사)
- 2025년 9월 ~ 현재 : 이화여자대학교 전자전기공학과 석사과정
- ORCID : https://orcid.org/0009-0004-9849-2716
- 주관심분야 : 3D 영상 이해, 딥러닝



강 제 원

- 2012년 : 미국 University of Southern California 박사
- 2012년 ~ 2014년 : 미국 퀄컴사 Senior Engineer
- 2014년 ~ 현재 : 이화여자대학교 교수
- ORCID : https://orcid.org/0000-0002-1637-9479
- 주관심분야 : 멀티모달 AI, 3D 영상 이해, 비디오 압축