

VLM 기반 객체중심 6DoF 3D Visual Grounding

VLM-based Object-centric 6DoF 3D Visual Grounding

이 성 엽¹ · 신 현 수² · 조 규 범³ · 배 지 훈⁴ · 이 성 온[†]Seongyeop Lee¹, Hyunsoo Shin², Gyubeom Jo³, Ji-Hun Bae⁴, Sungon Lee[†]

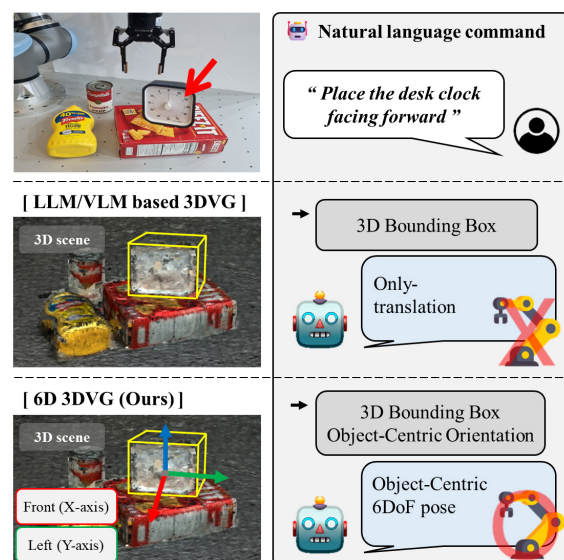
Abstract: 3D Visual Grounding aims to localize target objects in 3D scenes from natural language descriptions. Existing methods mainly provide object location information and do not explicitly capture object orientation, limiting their applicability to robotics. To address this limitation, we propose a VLM-based object-centric 6DoF 3D Visual Grounding framework that generates object-centric 6DoF poses from RGB multi-view images and text prompts. The framework reconstructs a 3D scene and camera viewpoints, extracts an object point cloud through mask-based multi-view projection with erosion and voting, and defines an object-centric coordinate frame by combining OBB-based geometric cues with semantic direction inference from a VLM enhanced via Egocentric Instruction Tuning. Experiments on seven objects show that the proposed method generates reliable object-centric 6DoF poses for objects with clear semantic directions.

Keywords: Vision-Language Models, 3D Visual Grounding, Object-Centric 6DoF Pose

1. 서 론

최근 대규모 언어 모델(Large Language Models, LLM)^[1]과 비전-언어 모델(Vision-Language Models, VLM)^[2]의 발전은 로봇이 시각적 정보로부터 단순히 환경을 인식하는 수준을 넘어, 인간의 자연어 명령을 해석하고 객체의 의미론적 특징을 이해하는 것을 가능하게 하였다.

3D Visual Grounding (3DVG)은 자연어로 기술된 문장을 3차원 장면 내 특정 영역 또는 객체와 연결하여 그 위치를 파악하는 기술로, 다양한 로봇틱스 분야에 응용되고 있다^[3,4]. 기존



[Fig. 1] Example of natural language-based robotic manipulation. Existing 3DVG methods, which lack object pose information, fail to execute the command. In contrast, our method provides object 6DoF poses that embed semantic directions of the object

연구들은 주로 포인트 클라우드-텍스트 쌍으로 구성된 데이터 셋 기반 모델 학습에 의존한다. 이러한 접근은 학습 데이터 확보의 어려움과 사전 정의된 어휘의 한계로 인해, 개방 어휘(open-vocabulary) 환경에서 일반화 성능이 제한적이다^[5-8]. 이러한

Received : Sep. 5. 2025; Revised : Apr. 28. 2026; Accepted : May. 18. 2026

※ This research was financially supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under the "Scale-up Technology Commercialization Program" (P0026386), supervised by the Korea Institute for Advancement of Technology (KIAT), and by the Technology Development Program (00507790) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

1. Researcher, Industry-Academic Cooperation Foundation, Hanyang University ERICA, Ansan, Korea (ehfdclslrn22@hanyang.ac.kr)

2. Research Assistant Professor, Research Institute of AI Convergence, Hanyang University, Ansan, Korea (shs2316@hanyang.ac.kr)

3. MS Student, Department of Robotics, Hanyang University ERICA, Ansan, Korea (whrb106@hanyang.ac.kr)

4. Professor, Co-Corresponding author: Department of Robotics, Hanyang University ERICA, Ansan, Korea (jihunbae@hanyang.ac.kr)

† Professor, Corresponding author: Department of Robotics, Hanyang University ERICA, Ansan, Korea (sungon@hanyang.ac.kr)

문제를 해결하기 위해 최근 LLM과 VLM을 활용하는 접근들이 제안되었고, 이를 통해 제로샷(zero-shot) 3DVG을 달성하였다^[9-12]. 그러나 위치 정보만으로는 3DVG의 로보틱스 분야 확장성과 응용성에 제약이 있다. 특히 자연어 기반 로봇 작업에는 객체의 고유 방향을 고려한 작업이 요구될 수 있다. 예를 들어, [Fig. 1]과 같이 “탁상 시계를 정면으로 놓아”라는 명령은 로봇이 시계의 위치뿐만 아니라, 시계의 정면 축을 인식하여 사용자를 향하도록 배치할 수 있어야 한다. 따라서 로봇 작업의 확장성과 응용성을 위해서는 객체 중심의 자세 정보가 필수적이다.

본 연구에서는 객체 방향 추론 능력 향상을 위해 Egocentric Instruction Tuning^[13]을 적용한 VLM을 기반으로 객체의 의미론적 방향이 반영된 자세를 제공하는 6DoF 3DVG을 제안한다. 제안된 방법은 다중 시점 이미지와 특정 객체를 설명하는 텍스트를 입력으로 활용하여, 객체중심(Object-centric) 좌표계를 정의한다. 본 좌표계는 객체의 전면(Front), 후면(Back), 좌측(Left), 우측(Right)과 같은 의미론적 방향성을 각각 X-axis, -X-axis, Y-axis, -Y-axis에 매핑하여, 객체중심 6DoF 포즈를 표현할 수 있다. 6DoF 3DVG은 물체의 방향 및 형상 특성을 기준으로 7개의 물체를 선정하고 실험을 수행하여 성능을 평가한다. 본 논문의 구조는 다음과 같다.

2장에서는 관련 연구를 검토하여 본 연구의 배경 및 차별점을 제시하고, 3장에서는 다중 시점 이미지와 텍스트 기반의 객체중심 좌표계를 정의하여 객체중심 6DoF 포즈를 생성하는 프레임워크를 상세히 설명한다. 이어 4장에서는 실험 설계 및 구현 세부사항과 함께, 6DoF 포즈 추정 성능을 정량적으로 분석한다. 마지막으로, 5장에서는 연구 결과를 요약하고 본 접근 방식의 의의, 한계 및 향후 연구 방향을 논의한다.

2. 관련 연구

2.1 3D Visual Grounding (3DVG)

2.1.1 지도 학습 기반 접근

기존 3DVG 연구는 주로 포인트 클라우드-텍스트 쌍 데이터셋을 활용한 지도 학습에 의존하였다. 이들은 3D 검출^[5-7] 또는 객체 분할 모델^[8]로 후보 객체를 생성한 뒤, 텍스트 인코더로 주어진 설명과 의미론적 특징을 정합하여 최종적으로 3D 바운딩 박스를 예측한다. 이후 연구에서는 융합 기법을 개선하여 성능을 향상시켰다^[14,15]. 그러나 이러한 지도 학습 기반 접근은 데이터 확보의 어려움과 사전 정의된 어휘의 한계로 인해, 개방 어휘 환경에서 일반화 성능이 제한적이다.

2.1.2 LLM/VLM 기반 접근

3차원 장면 내 객체를 단순히 사전에 정의된 어휘에 의존하

[Table 1] Comparison of existing and proposed 3DVG methods

	Supervised Learning	LLM/VLM based	Proposed
Input	3D scene + text	3D scene + text	Multi-view images + text
Training	Supervised	Pre-trained	Pre-trained
Text	Fixed	Open vocabulary	Open vocabulary
Output	3D bounding box	3D bounding box	Object-centric 6DoF pose

지 않고, 방대한 텍스트 데이터로 학습된 LLM의 추론 능력을 활용하여 개방 어휘 표현을 해석하고, 이를 후보 객체와의 의미적 일치 여부를 판별하는 방식이 제안되었다^[9,10]. 그러나 LLM 단독으로는 이미지 인코더가 추출한 시각적 특징과 언어 인코더가 생성한 언어적 표현의 정렬 능력이 부족하여, 3차원 장면 내 개방 어휘 표현을 직접적으로 객체 위치에 매핑하는 데 한계가 있다. 이를 보완하기 위해 LLM-Grounder^[9]는 LLM을 추론에 이진트로 두고, CLIP 기반 멀티모달 3DVG모델인 OpenScene^[16]이나 LERF^[17]와 결합하는 방식을 제안하였다.

VLM은 대규모 이미지-텍스트 쌍을 기반으로 학습된 멀티모달 모델로, 시각적 특징과 언어적 표현을 공통 임베딩 공간에서 정렬하는 능력을 지닌다. 또한 높은 일반화 성능을 보이며, 이는 제로샷(Zero-shot) 3DVG에 특히 적합하다^[11,12]. LLM/VLM 기반 접근 역시 주로 객체의 위치 정보 파악에 초점을 맞추고 있으며, 객체의 방향에 대한 정보는 제공하지 못한다는 한계점이 있다. [Table 1]에서 기존 3DVG 접근법들과 제안된 방법을 자세히 비교하였다.

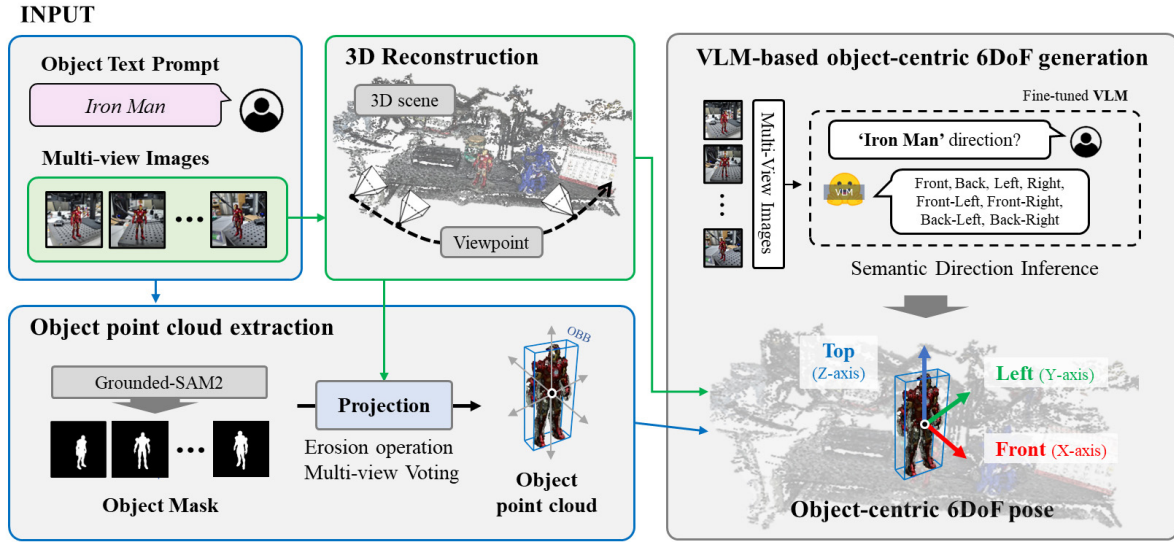
2.2 VLM 기반 객체 방향 추론

2.2.1 객체 방향 추론 능력의 한계

현재 VLM은 이미지에서 객체의 방향을 정확하게 해석하는데 어려움을 겪고 있으며^[13,18], 이는 자연어 기반 로봇 작업에서의 확장성과 응용성을 제약한다. 이러한 한계는 학습 데이터의 방향 주석 불일치에서 비롯된다^[16]. MS-COCO^[19], LAION-5B^[20]와 같은 대규모 이미지-텍스트 데이터셋에서는 객체 방향이 일부는 관찰자의 시점에 기반하고, 또 일부는 객체 자체의 시점에 기반하는 등 객체 방향 주석이 일관적이지 않다. 이러한 불일치로 인해 VLM은 객체 방향을 일관성 있게 학습하지 못하고, 결과적으로 사용자 의도와 어긋나는 출력을 생성할 수 있다.

2.2.2 Egocentric Instruction Tuning

최근 객체 방향 추론 능력을 향상시키기 위한 다양한 방법들이 연구되고 있다^[21,22]. 특히 Egocentric Instruction Tuning^[13]은



[Fig. 2] Overall pipeline of the proposed method. From an object text prompt and multi-view images, the method reconstructs the 3D scene and viewpoints, extracts the object point cloud through mask-based projection and multi-view voting, and finally generates an object-centric 6DoF pose by combining VLM-based semantic direction inference with geometric shape information

자기중심적 관점에 기반한 객체 방향 주석 데이터셋을 기반으로 VLM의 객체 방향 추론 능력을 사용자의 자기중심적 관점에 맞추도록 조정한다. 이는 객체 방향의 기준점을 일관적으로 일치시켜, 성능 저하 없이 강건한 방향 추론 능력을 제공한다. 본 연구에서 [13]은 객체의 의미론적 방향 추론 성능을 향상시키기 위한 VLM 튜닝 기법으로 활용된다.

3. 방 법

본 연구는 다중 시점 이미지와 특정 객체를 설명하는 텍스트만으로 3차원 장면에서 객체의 위치 정보와 의미론적 방향이 반영된 정보를 포함하는 6DoF 포즈를 기하학적 접근을 통해 생성한다. 전체 프레임워크의 개요는 [Fig. 2]에 나와있으며, 세 가지 핵심 모듈로 구성된다.

- 3차원 장면 및 카메라 시점 재구성
- 객체 포인트 클라우드 추출
- VLM 기반 객체중심 6DoF 생성

3.1 3차원 장면 및 카메라 시점 재구성

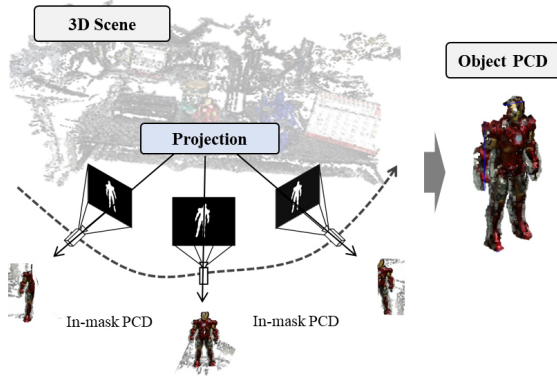
다중 시점 이미지를 기반으로 3차원 장면과 카메라 시점을 재구성하기 위하여 COLMAP^[23]을 활용하였다. SfM (Structure-from-Motion)에서 다중 시점 이미지 간의 특징점을 추출하고 대응점 매칭을 통해 카메라 파라미터를 추정한다. 이를 바탕으로 초기 3차원 포인트 클라우드를 생성한다. 이후 MVS (Multi-View Stereo)에서 보다 조밀한 대응점을 계산하여 고해상도의

포인트 클라우드를 재구성한다. 추정된 카메라 파라미터와 재구성된 3차원 장면은 3.2.2에서 3차원 객체를 추출하기 위해 활용되고, 카메라 시점은 3.3절에서 각 시점과 시점에 해당하는 이미지에서 VLM이 추론한 객체의 방향을 매칭한 시점-방향 집합을 생성하는 데 활용된다.

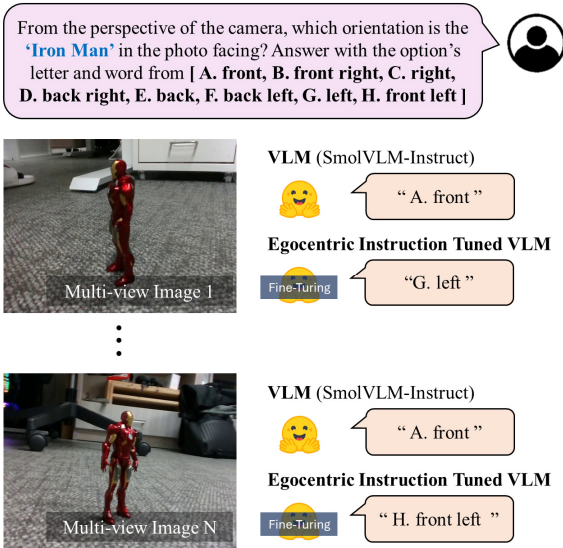
3.2 객체 포인트 클라우드 추출

3.2.1 객체 마스크 이미지 생성

Grounded-SAM2^[24]는 Grounding DINO^[25]를 개방 어휘 기반 객체 검출기로 사용하고, 객체 세그멘테이션 모델인 SAM2 (Segment Anything Model 2)^[26]와 결합한 멀티모달 파이프라인이다. Grounded-SAM2를 사용하여 객체에 대한 마스크 이미지를 생성한다. 생성된 마스크는 객체 경계의 픽셀 단위 분할을 제공하지만, 오차가 발생할 수 있다. 마스크가 과도하게 크게 생성된 경우에는 객체 외부의 불필요한 영역이 포함되어 객체가 아닌 포인트가 함께 추출될 수 있다. 반대로, 마스크가 지나치게 작게 생성된 경우에는 객체의 일부 포인트가 손실되어 불완전한 객체 포인트 클라우드가 형성된다. 두 오류 모두 객체 인식의 정확성을 저하시킬 수 있으나, 특히 비객체 포인트가 포함되는 경우는 후속 좌표계 정합 과정에서 잘못된 방향 정보를 유발할 수 있어 더 치명적이다. 따라서 마스크 이미지에 침식 (erosion) 연산만을 적용하여 마스크 경계를 보수적으로 축소한다. 이러한 연산을 통해 비객체 포인트가 포함되는 위험을 최소화하고, 일부 객체 포인트 손실이 발생하더라도 전체 좌표계 정합 및 방향 추론 과정에서의 안정성을 보장한다.



[Fig. 3] Object point cloud extraction by projecting the reconstructed 3D scene onto multi-view object masks

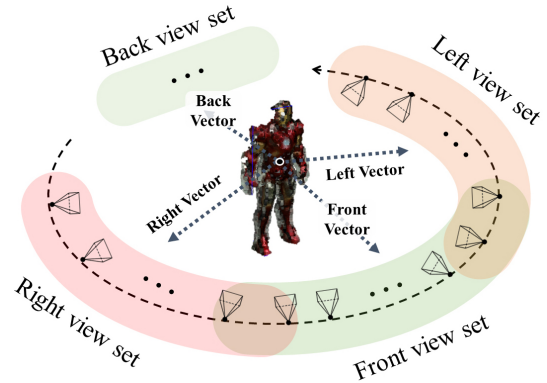


[Fig. 4] Comparison of semantic direction understanding between the base VLM and the fine-tuned VLM

3.2.2 다중 시점 투영 기반 객체 포인트 클라우드 추출

3.2.1에서 생성된 객체 마스크 이미지와 3.1에서 재구성한 3차원 장면을 활용하여 3차원 객체를 추출한다. 3.1에서 추정된 카메라 파라미터를 이용하여 포인트 X_i 를 카메라 시점 j 에 해당하는 마스크 이미지에 투영한다. $v_{i,j} \in \{0, 1\}$ 는 포인트 X_i 가 카메라 시점 j 의 객체 마스크 내부에 투영되는지를 나타내는 변수이며, 마스크 내부에 투영되면 1, 그렇지 않으면 0으로 정의한다. 이때, 신뢰성 있는 객체 포인트 클라우드를 추출하기 위하여, 소수의 시점에서 객체 포인트로 판단된 포인트들을 제거하는 투영 기반 다중 시점 투표 전략을 제안한다. 포인트가 절반(0.5) 이상의 마스크 이미지에서 투영된 경우에만 최종 객체 포인트로 판별한다. 판별 함수는 아래와 같다.

$$X_i \in P_{obj} \Leftrightarrow \frac{1}{N} \sum_{j=1}^N v_{i,j} \geq 0.5 \quad (1)$$



[Fig. 5] Visualization of the view-direction sets and vectors

투표 기반 필터링은 국소적으로 검출된 불확실한 객체 포인트들을 효과적으로 제거하고, 여러 시점에서 일관되게 검출된 객체 포인트만을 남김으로써 [Fig. 3]와 같이 신뢰성 있는 객체 포인트 클라우드를 추출한다.

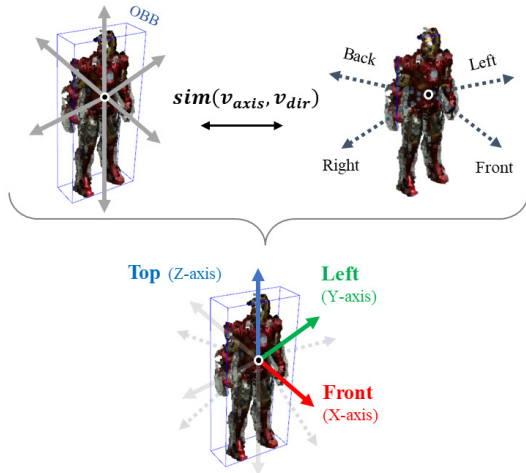
3.3 VLM 기반 객체 중심 6DoF 생성

3.3.1 VLM 기반 객체 의미론적 방향 추론

객체 방향 추론 능력 향상을 위해 Egocentric Instruction Tuning을 적용한 VLM을 기반으로, 다중 시점 이미지에서 객체의 의미론적 방향을 8개의 방향 클래스(Front, Back, Left, Right, Front-Left, Front-Right, Back-Left, Back-Right) 중 하나로 추론한다[Fig. 4]. 이후 3.1에서 재구성한 카메라 시점과 각 시점에 해당하는 객체의 방향을 매칭하여 방향별 시점-방향 집합을 구성한다. 이때 각 시점-방향 집합의 중심을 계산한 뒤, 특정 시점이 현재 속한 집합의 중심보다 다른 시점-방향 집합의 중심에 더 가까운 경우, 해당 시점은 VLM의 방향 추론 오류로 판단하여 최근접 시점-방향 집합의 방향으로 재할당한다. 이러한 과정을 통해 이상치가 보정된 최종 시점-방향 집합을 생성한다. 이후 객체 포인트 클라우드의 중심에서 각 시점-방향 집합의 중심을 향하는 벡터를 생성하여 시점-방향 벡터를 정의한다. 이와 같은 시점-방향 벡터 생성 절차를 [Algorithm 1]에 요약하였다.

3.3.2 방향 정합 및 좌표계 정의

OBB (Oriented Bounding Box)는 주성분 분석(Principal Component Analysis, PCA)으로 주축 벡터를 찾고, 이를 기준으로 최소 부피의 경계 상자를 형성한다. 이러한 OBB의 좌표축은 객체의 의미론적 방향을 반영하지 않지만, 객체 형상에 정렬된 축 벡터를 제공하기 때문에 객체중심 좌표계의 초기 기준 축으로 활용할 수 있다. 먼저, 객체 포인트 클라우드의 중심을 기준으로 방향 별 시점 중심과 연결되는 시점-방향 벡터를 생성한다[Fig. 5]. 이는 객체를 특정 방향으로 관측한 시점들을 대표



[Fig. 6] Generation of an object-centric coordinate frame by aligning the OBB axes with the view-direction vectors

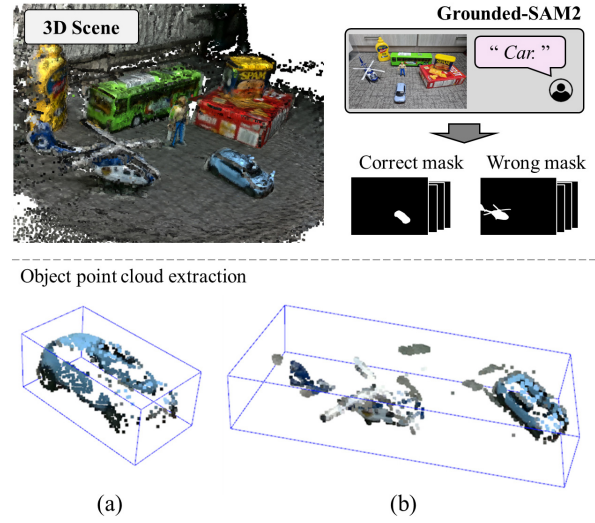
[Algorithm 1] View-direction vector generation

Input: multi-view images I , camera poses C
Output: view-direction vectors V
for each image I_i do
 $d_i \leftarrow$ infer semantic direction using VLM
assign d_i to camera pose C_i
end for
 $G \leftarrow$ construct view-direction sets from $\{(C_i, d_i)\}$
for each viewpoint C_i do
if C_i is closer to another view-direction set center then
 $d_i \leftarrow$ direction label of the nearest set
end if
end for
 $V \leftarrow$ generate view-direction vectors from G
return V

[Algorithm 2] Object-centric 6DoF pose generation

Input: object point cloud P_{obj} , view-direction vectors V
Output: object-centric 6DoF pose T_{obj}
 $B \leftarrow$ extract OBB axis vectors from P_{obj}
 $A \leftarrow$ align V with B using cosine similarity
 $T_{obj} \leftarrow$ construct object-centric 6DoF pose from A
return T_{obj}

하는 시점으로, 의미론적 방향 정보가 반영된 벡터로 해석된다. 이후 [Fig. 6]와 같이 시점-방향 벡터를 OBB의 축벡터와 코사인 유사도를 기반으로 정합한다. 최종적으로 객체의 의미론적 방향 정보인 전면(Front), 후면(Back), 좌측(Left), 우측(Right)과 형상 정보인 X-axis, -X-axis, Y-axis, -Y-axis가 매핑되어, 객체중심 좌표계를 정의하고, 이를 바탕으로 3차원 장면에서 객체의 6DoF 포즈를 생성하였다. 전체 과정은 [Algorithm 2]에 요약하였다.



[Fig. 7] Comparison of object point clouds under the proposed method. (a) shows the result with erosion operation and voting strategy applied, whereas (b) shows the result without them

4. 실험 및 결과

4.1 실험 설정

본 연구에서는 제안한 6DoF 3DVG의 성능 평가를 위해, 방향 및 형상 특성을 기준으로 세 가지 물체 유형을 정의하고 총 7종의 물체를 선정하였다. 각 물체마다 30 장의 다중 시점 이미지(640x480)를 생성하였으며, 이를 기반으로 실험을 수행하였다. 각 유형과 선정된 물체는 아래와 같다.

Type 1) 방향성이 명확하고, 형상이 단순한 물체

: Bus

Type 2) 방향성이 명확하지만, 형상이 복잡한 물체

: Car, Man, Mustard, Elephant, Helicopter

Type 3) 방향성이 모호한 물체

: Basketball

4.2 객체 포인트 클라우드 추출

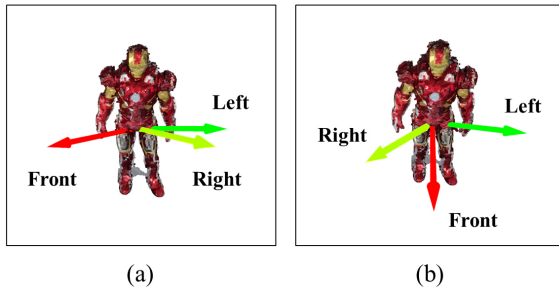
먼저 제안한 침식 연산과 다중 시점 투표 방식이 객체 포인트 클라우드 추출에 미치는 영향을 평가하였다. [Fig. 7]은 각 조건에서 생성된 객체 포인트 클라우드를 시각적으로 비교한 결과이다. 마스크 이미지 침식과 투영 기반 다중 시점 투표 전략을 적용한 경우(a)에는 여러 시점에서 일관되게 관측된 포인트만이 유지되어 경계 노이즈와 비객체 포인트가 효과적으로 억제되었으며, 보다 안정적인 객체 포인트 클라우드가 생성됨을 확인하였다. 반면, 아무 처리도 적용하지 않은 경우(b)에는

[Table 2] Object orientation prediction accuracy on EgoOrientBench (ImageNet, D_3 , DomainNet, PACS, and OmniObject3D) using the SmolVLM-Instruct(2.2B) backbone

Method	IN	D_3	DN	PA	3D	Avg.
Base	14.8	21.1	20.8	16.0	13.9	16.0
Fine-tuned	64.7	37.5	71.7	69.8	42.4	56.3

[Table 3] Average semantic direction inference accuracy of the base and fine-tuned VLMs on the experimental objects, excluding Type 3 objects with ambiguous direction

Method	Semantic Direction Inference Acc. (%)
Base	22.78
Fine-tuned	75.56



[Fig. 8] Visual comparison of the view-direction vectors generated from the semantic direction predictions of the base VLM (a) and fine-tuned VLM (b)

부정확한 마스크로 인해 비객체 포인트와 경계 부근의 노이즈가 다수 포함되어 객체 형상이 명확하게 추출되지 않았다.

4.3 의미론적 방향 추론

VLM의 객체 의미론적 방향 추론 성능이 최종 객체 중심 좌표계 생성에 미치는 영향을 평가하였다. 본 연구에서는 VLM으로 SmolVLM-Instruct^[27]를 사용하였다. 객체 방향 추론 능력을 향상시키기 위해, Egocentric Instruction Data로 QLoRA 기반 파인튜닝을 수행하였다. 그 결과 [Table 2]와 같이 ImageNet, D_3 , DomainNet, PACS, OmniObject3D 데이터에 객체 방향에 대한 주석이 있는 EgoOrientBench^[13]에서 base VLM 대비 객체 방향 정확도가 40.3% 향상되었다.

실험 물체에 대한 의미론적 방향 추론 정확도는 [Table 3]와 같이 평균 75.56%를 보인 반면, base VLM은 평균 22.78%의 정확도를 보였다. 정답 여부는 사람이 각 객체에 대해 VLM의 방향 추론 결과를 검토하여 판단하였다. [Fig. 8]은 이러한 추론 결과로 생성된 시점-방향 벡터를 시각적으로 비교한 것이다. base VLM의 경우(a) 대부분의 이미지에서 객체의 방향을 불만

정하게 추론하였으며, 이로부터 생성된 시점-방향 벡터 또한 정답과 크게 어긋나는 모습을 보였다. 반면 fine-tuned VLM의 경우(b) 객체의 의미론적 방향을 보다 일관되게 추론하였고, 이를 기반으로 생성된 방향 벡터 역시 객체에 대해 더 타당한 방향성을 나타냈다. 여기서 본 평가의 목적은 파인튜닝 자체의 성능 향상을 강조하는 데 있는 것이 아니라, 향상된 방향 추론 성능이 최종 객체 중심 포즈 생성에 필요한 수준의 신뢰도를 제공할 수 있는지를 확인하는 데 있다. 즉, 일부 방향 추론 오차가 존재 하더라도, 제안 방법이 최종적으로 안정적인 객체 중심 좌표계를 생성할 수 있음을 확인하였다.

4.4 객체 중심 6DoF 생성

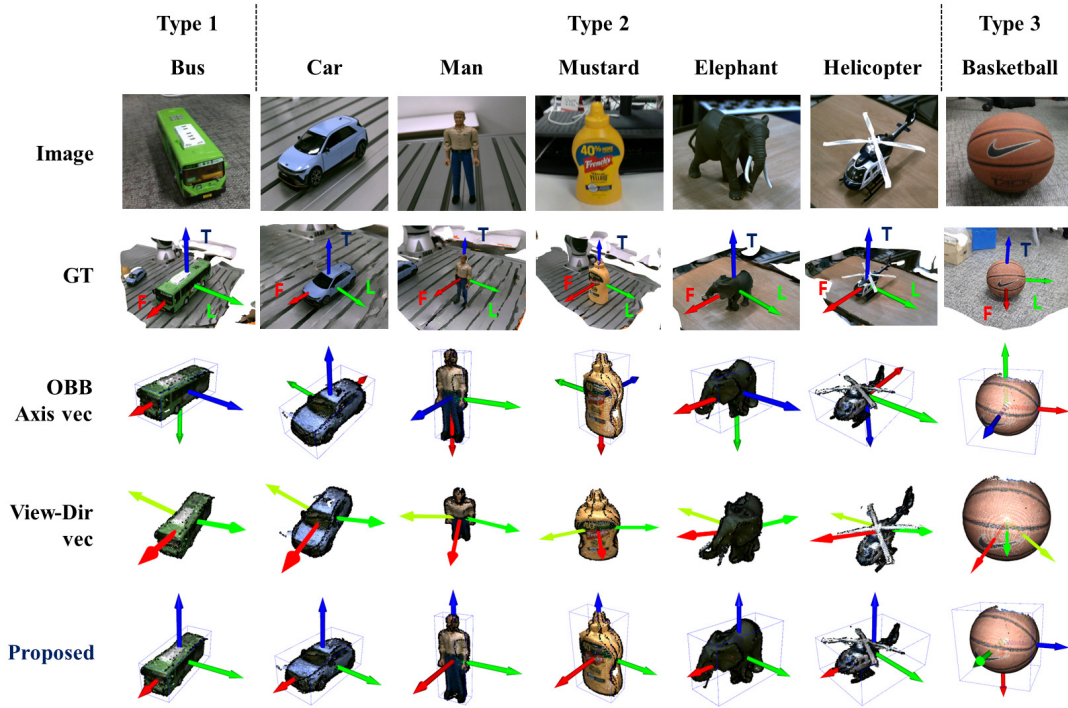
제안 방법이 객체 중심 6DoF 포즈를 얼마나 정확하게 추정하는지를 평가하였다. 선정된 물체에 대해 사람이 직접 생성한 포인트 클라우드와 객체 중심 6DoF 포즈를 정답(ground truth)으로 정의한다. 이를 기준으로 제안하는 6DoF 3DVG이 생성한 객체 중심 좌표계의 회전행렬 R_{est} 에 대한 회전오차(2)와 평행 이동 벡터 t_{est} 에 대한 평행이동 오차(3)를 평가한다.

$$e_{rot} = \cos^{-1}\left(\frac{\text{trace}(R_{est}R_{gt}^T) - 1}{2}\right) \quad (2)$$

$$e_{trans} = \|t_{est} - t_{gt}\|^2 \quad (3)$$

[Fig. 9]은 각 객체에 대해 정답 좌표계(GT), OBB 축, 시점-방향 벡터, 그리고 제안 방법으로 생성된 최종 객체 중심 좌표계를 시각적으로 비교한 결과이다. OBB 축은 객체의 형상과 주축 방향을 잘 반영하지만, 그 자체만으로는 정면이나 왼쪽과 같은 의미론적 방향 정보를 포함하지 않는다. 반대로 VLM로부터 얻은 방향 벡터는 의미론적 방향성은 가지지만, 객체의 전체 형상을 충분히 반영하지 못한다. 제안 방법은 이 두 정보를 결합함으로써, Front (X-axis), Back (-X-axis), Left (Y-axis), Right (-Y-axis)로 정의되는 의미론적 방향성과 객체의 기하학적 형상을 동시에 반영한 객체 중심 좌표계를 생성할 수 있음을 보여준다.

[Table 4]는 각 유형별 물체의 회전 오차 결과와 평행이동 오차를 나타낸다. Type 1, 2에서는 평균 10.73°, 13.84 mm 수준 오차의 좌표계를 생성하였으며, 특히 회전 오차 측면에서 90° 이상의 축 불일치가 발생하지 않았다. 이는 객체의 의미론적 방향 정보가 최종 좌표계의 축 정렬에 올바르게 반영되었음을 의미한다. 반면, 방향 구분이 모호한 Type 3에서는 회전 추정 성능이 크게 저하되었다. 이는 인간과 마찬가지로 물체의 고유한 방향성을 구별할 수 있는 시각적 단서나 물체의 특징이 부족할 경우, 신뢰성 있는 방향 추론이 어려운 것으로 해석된다.



[Fig. 9] Visual comparison of the ground-truth coordinate frame, OBB axes, view-direction vectors, and the final object-centric coordinate frame generated by the proposed method for each object

[Table 4] Rotation error (e_{rot}) and translation error (e_{trans}) for each object type. Type 1 and Type 2 show relatively consistent performance, whereas Type 3 exhibits a large rotation error due to the ambiguity of object orientation. The last column reports the average results excluding Type 3

		Type 1	Type 2					Type 3	Avg. w/o Type 3
		Bus	Car	Man	Mustard	Elephant	Helicopter	basketball	
e_{rot}	(deg)	12.31	3.37	7.86	7.02	12.46	21.39	169.72	10.73
e_{trans}	(mm)	17.88	9.51	8.37	15.66	16.94	14.72	14.18	13.84

5. 결 론

본 연구에서는 다중 시점 이미지와 특정 객체를 설명하는 텍스트라는 비교적 단순한 입력만으로, 의미론적 방향이 반영된 객체중심 좌표계를 정의하여 6DoF 포즈를 생성하는 새로운 6DoF 3D Visual Grounding을 제안하고 성능을 평가하였다. RGB-D 데이터에 의존하지 않고 RGB 다중 시점 이미지만을 사용하며, 별도의 CAD 모델이나 학습 과정 없이 기하학적 접근만으로 신뢰성 있는 객체중심 6DoF 포즈를 직접 생성할 수 있음을 보였다. 이를 통해 로봇과 사용자는 보다 자연스럽고 직관적으로 상호작용할 수 있으며, 로봇틱스 분야에서의 확장성과 응용성을 확인하였다.

그러나 본 방법은 객체의 형상 및 방향 단서에 일정 부분 의존한다는 한계를 가진다. 실험 결과에서도 방향성이 모호한 물체에서는 자세 추정 성능이 크게 저하되었으며, 이는 객체의 고

유한 방향성을 구별할 수 있는 시각적 단서가 부족할 경우 신뢰성 있는 방향 추론이 어려움을 보여준다. 향후에는 다중 시점 방향 예측의 일관성을 기반으로 객체 방향 추론의 불확실성을 정량화하는 메커니즘을 도입하여 로봇틱스 응용에서 안전성을 향상시키고, 다양한 객체에서 실제 pick-and-place 실험을 통해 제안 방법의 실환경 적용 가능성을 검증할 예정이다.

References

- [1] T. Brown et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877-1901, Dec. 2020, DOI: 10.18653/v1/2021.mrl-1.1.
- [2] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v(ision)," *arXiv preprint arXiv:2309.17421*, 2023, DOI: 10.48550/arXiv.2309.17421.

- [3] Q. Feng, V. Ablavsky, and S. Sclaroff, "Cityflow- nl: Tracking and retrieval of vehicles at city scale by natural language descriptions," *arXiv preprint arXiv:2101.04741*, Jan., 2021, DOI: 10.48550/arXiv.2101.04741.
- [4] L. Kong et al., "Robo3d: Towards robust and reliable 3d perception against corruptions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 19994-20006, 2023, DOI: 10.1109/ICCV51070.2023.01830.
- [5] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in rgb-d scans using natural language," in *European conference on computer vision*, Glasgow, UK, pp. 202-221, 2020, DOI: 10.1007/978-3-030-58565-5_13.
- [6] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes," in *European conference on computer vision*, Glasgow, UK, pp. 422-440, 2020, DOI: 10.1007/978-3-030-58452-8_25.
- [7] A. Dai et al., "ScanNet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, USA, pp. 2432-2443, 2017, DOI: 10.1109/CVPR.2017.261.
- [8] Z. Zhu et al., "3d -Vista: Pre-trained transformer for 3d vision and text alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 2911-2921, 2023, DOI: 10.1109/ICCV51070.2023.00272.
- [9] J. Yang et al., "Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent," in *2024 IEEE International Conference on Robotics and Automation*, Yokohama, Japan, pp. 7694-7701, 2024, DOI: 10.1109/ICRA57147.2024.10610443.
- [10] Z. Yuan et al., "Visual programming for zero-shot open-vocabulary 3d visual grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 20623-20633, 2024, DOI: 10.1109/CVPR52733.2024.01949.
- [11] R. Xu et al., "Vlm-grounder: A vlm agent for zero-shot 3d visual grounding," in *Conference on Robot Learning*, vol. 270, pp. 3961-3985, 2025, DOI: 10.48550/arXiv.2410.13860.
- [12] R. Li, S. Li, L. Kong, X. Yang, and J. Liang, "Zero-shot 3d visual grounding from vision-language models," *arXiv preprint arXiv:2505.22429*, May 2025, DOI: 10.48550/arXiv.2505.22429.
- [13] J. H. Jung et al., "Is 'right' right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 14257-14267, 2025, DOI: 10.1109/CVPR52734.2025.01330.
- [14] A. Jain, N. Gkanatsios, I. Mediratta, and K. Fragkiadaki, "Bottom up top down detection transformers for language grounding in images and point clouds," in *European Conference on Computer Vision*, Tel Aviv, Israel, pp. 417-433, 2022, DOI: 10.1007/978-3-031-20059-5_24.
- [15] O. Unal, C. Sakaridis, S. Saha, and L. Van Gool, "Four ways to improve verbovisual fusion for dense 3d visual grounding," in *European Conference on Computer Vision*, Milan, Italy, pp. 196-213, 2024, DOI: 10.1007/978-3-031-73116-7_12.
- [16] S. Peng et al., "Openscene: 3d scene understanding with open vocabularies," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Vancouver, Canada, pp. 815-824, 2023, DOI: 10.1109/CVPR52729.2023.00085.
- [17] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *Proceedings of the IEEE/CVF international conference on computer vision*, Paris, France, pp. 19729-19739, 2023, DOI: 10.1109/ICCV51070.2023.01807.
- [18] S. Tong et al., "Eyes wide shut? exploring the visual shortcomings of multimodal llms," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, USA, pp. 9568-9578, 2024, DOI: 10.1109/CVPR52733.2024.00914.
- [19] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *European conference on computer vision*, Zurich, Switzerland, pp. 740-755, 2014, DOI: 10.1007/978-3-319-10602-1_48.
- [20] C. Schuhmann et al., "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in neural information processing systems*, vol. 35, pp. 25278-25294, 2022, DOI: 10.5555/3600270.3602103.
- [21] Y. Huang et al., "Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 22072-22086, 2024, DOI: 10.1109/CVPR52733.2024.02084.
- [22] Y. Mu et al., "Embodiedgpt: Vision-language pre-training via embodied chain of thought," *Advances in Neural Information Processing Systems*, vol. 36, pp. 25081-25094, 2023, DOI: 10.5555/3666122.3667212.
- [23] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, USA, pp. 4104-4113, 2016, DOI: 10.1109/CVPR.2016.445.
- [24] T. Ren et al., "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, Jan. 2024, DOI: 10.48550/arXiv.2401.14159.
- [25] S. Liu et al., "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*, Milan, Italy, pp. 38-55, 2024, DOI: 10.1007/978-3-031-72970-6_3.
- [26] N. Ravi et al., "Sam 2: Segment anything in images and videos," in *International Conference on Learning Representations*, *arXiv preprint arXiv:2408.00714*, 2024, DOI: 10.48550/arXiv.2408.00714.
- [27] A. Marafioti et al., "Smolvlm: Redefining small and efficient multimodal models," in *Second Conference on Language Modeling*, 2025, DOI: 10.48550/arXiv.2504.05299.



이 성 업

2024 대전대학교 정보통신공학과(학사)
2026 한양대학교 ERICA 인공지능융합학과
(석사)
2026~현재 한양대학교 ERICA 산학협력단
연구원

관심분야: VLM, 언어모델 기반 로봇 제어, 영상처리



배 지 훈

1999 명지대학교 전기공학과(학사)
2001 명지대학교 전기공학과(석사)
2004 Ritsumeikan University, Japan,
Robotics(박사)
2024~현재 한양대학교 ERICA 로봇공학과
교수

관심분야: 로봇핸드, 로봇암, 로봇매니퓰레이션



신 현 수

2015 한양대학교 ERICA 전자시스템공학과
(학사)
2017 한양대학교 ERICA 전자공학과(석사)
2023 한양대학교 ERICA 전자공학과(박사)
2026~현재 한양대학교 AI융합기술연구소

관심분야: 로봇 비전, 핸드아이 캘리브레이션, 센서 융합



이 성 온

1997 서울대학교 기계설계학과(학사)
1999 포항공과대학교 지능기계정보학(석사)
2008 동경대학교 기계공학과(박사)
2023~현재 한양대학교 ERICA 로봇공학과
교수

관심분야: 의료 영상, 영상기반제어, 딥러닝기반 물체인식



조 규 범

2026 한양대학교 ERICA 전자공학과(학사)
2026~현재 한양대학교 ERICA 로봇공학과
석사과정

관심분야: 로봇 비전, 딥러닝 기반 물체 인식