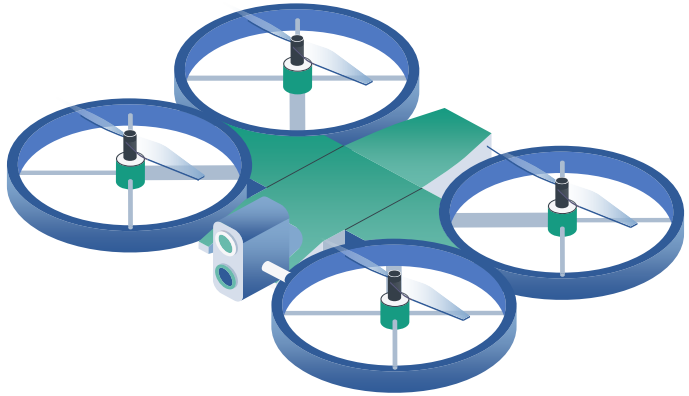
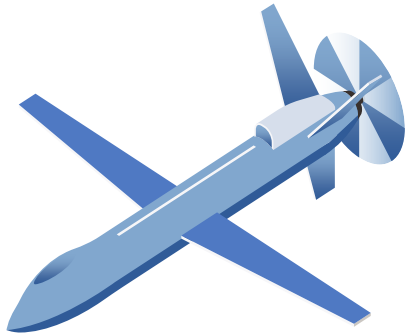


산학연계프로젝트 설명서

개방형 어휘 활용한 제로샷 객체 탐지
(Zero-shot object detection using open vocabulary)



지도 교수: 장한얼 교수님

팀 원:
컴퓨터공학과 20201735 박우진
컴퓨터공학과 20222019 김다빈
컴퓨터공학과 20232013 김수안
컴퓨터공학과 20232014 김연주
정보통신공학과 20238024 박진형
컴퓨터공학과 20231203 엄기원

국립한밭대학교

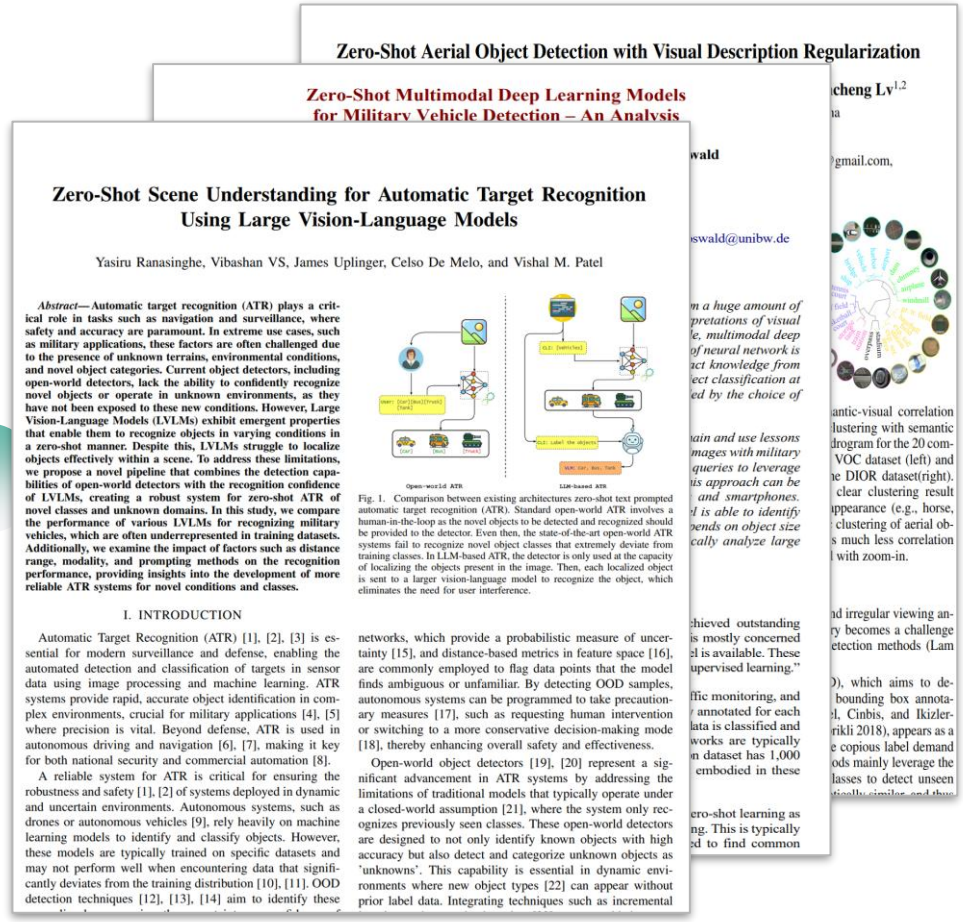
CONTENTS

01

연구 목표

01 연구 목표

국방 분야에서의 Zero-Shot Object Detection 관심 증가



Zero-Shot Aerial Object Detection with Visual Description Regularization

Zero-Shot Multimodal Deep Learning Models for Military Vehicle Detection – An Analysis

Zero-Shot Scene Understanding for Automatic Target Recognition Using Large Vision-Language Models

Yasiru Ranasinghe, Vibashan VS, James Uplinger, Celso De Melo, and Vishal M. Patel

Abstract— **Automatic target recognition (ATR) plays a critical role in tasks such as navigation and surveillance, where safety and accuracy are paramount. In extreme use cases, such as military applications, these factors are often challenged due to the presence of unknown terrains, environmental conditions, and novel object categories. Current object detectors, including open-world detectors, lack the ability to confidently recognize novel objects or operate in unknown environments, as they have not been exposed to these new conditions. However, Large Vision-Language Models (LVLMs) exhibit emergent properties that enable them to recognize objects in varying conditions in a zero-shot manner. Despite this, LVLMs struggle to localize objects effectively within a scene. To address these limitations, we propose a novel pipeline that combines the detection capabilities of open-world detectors with the recognition confidence of LVLMs, creating a robust system for zero-shot ATR of novel classes and unknown domains. In this study, we compare the performance of various LVLMs for recognizing military vehicles, which are often underrepresented in training datasets. Additionally, we examine the impact of factors such as distance range, modality, and prompting methods on the recognition performance, providing insights into the development of more reliable ATR systems for novel conditions and classes.**

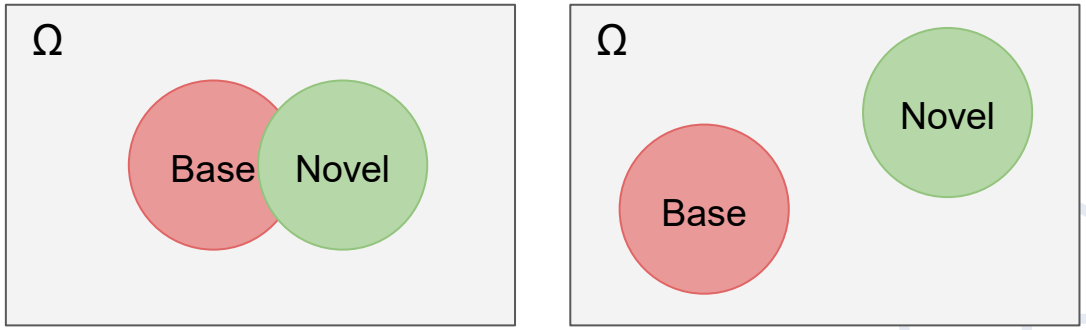
Fig. 1. Comparison between existing architectures zero-shot text prompted automatic target recognition (ATR). Standard open-world ATR involves a human-in-the-loop as the novel objects to be detected and recognized should be provided to the detector. Even then, the state-of-the-art open-world ATR systems fail to recognize novel object classes that extremely deviate from training classes. In LLM-based ATR, the detector is only used at the capacity of localizing the objects present in the image. Then, each localized object is sent to a larger vision-language model to recognize the object, which eliminates the need for user interference.

I. INTRODUCTION

Automatic Target Recognition (ATR) [1], [2], [3] is essential for modern surveillance and defense, enabling the automated detection and classification of targets in sensor data using image processing and machine learning. ATR systems provide rapid, accurate object identification in complex environments, crucial for military applications [4], [5] where precision is vital. Beyond defense, ATR is used in autonomous driving and navigation [6], [7], making it key for both national security and commercial automation [8].

A reliable system for ATR is critical for ensuring the robustness and safety [1], [2] of systems deployed in dynamic and uncertain environments. Autonomous systems, such as drones or autonomous vehicles [9], rely heavily on machine learning models to identify and classify objects. However, these models are typically trained on specific datasets and may not perform well when encountering data that significantly deviates from the training distribution [10], [11]. OOD detection techniques [12], [13], [14] aim to identify these networks, which provide a probabilistic measure of uncertainty [15], and distance-based metrics in feature space [16], are commonly employed to flag data points that the model finds ambiguous or unfamiliar. By detecting OOD samples, autonomous systems can be programmed to take precautionary measures [17], such as requesting human intervention or switching to a more conservative decision-making mode [18], thereby enhancing overall safety and effectiveness.

Open-world object detectors [19], [20] represent a significant advancement in ATR systems by addressing the limitations of traditional models that typically operate under a closed-world assumption [21], where the system only recognizes previously seen classes. These open-world detectors are designed to not only identify known objects with high accuracy but also detect and categorize unknown objects as 'unknowns'. This capability is essential in dynamic environments where new object types [22] can appear without prior label data. Integrating techniques such as incremental



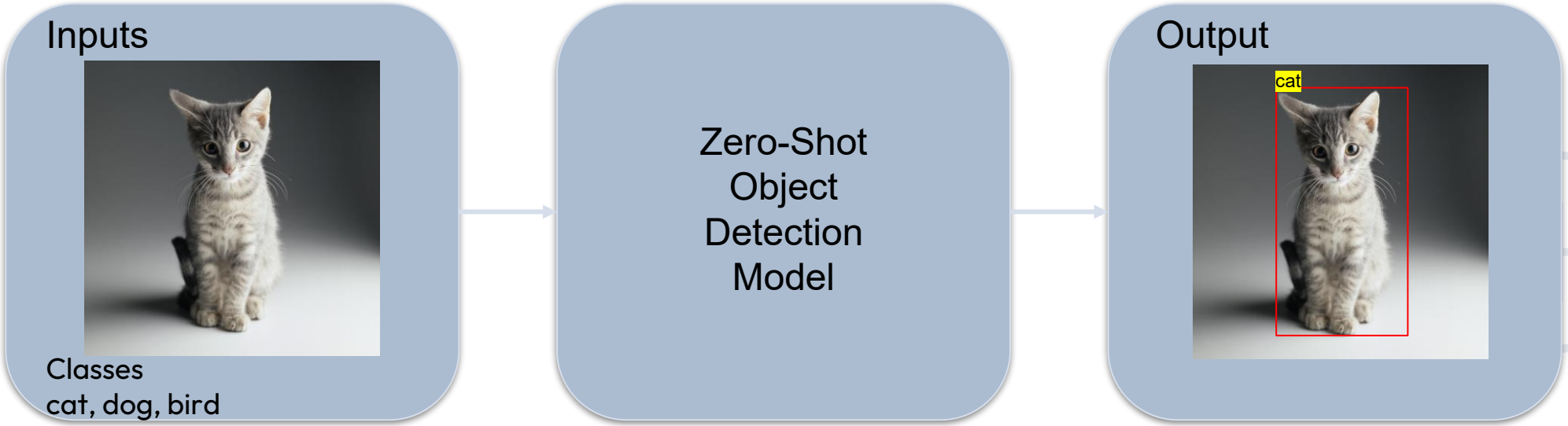
Object Detection

Zero-Shot
Object Detection

01 연구 목표

Zero-Shot Object Detection

- 학습 시 전혀 **본 적이 없는 객체**도 **텍스트 설명만으로 탐지** 가능
- 입력으로 **이미지**와 **후보 클래스**를 목록으로 받아 객체가 탐지된 **bounding box**와 **label**을 출력

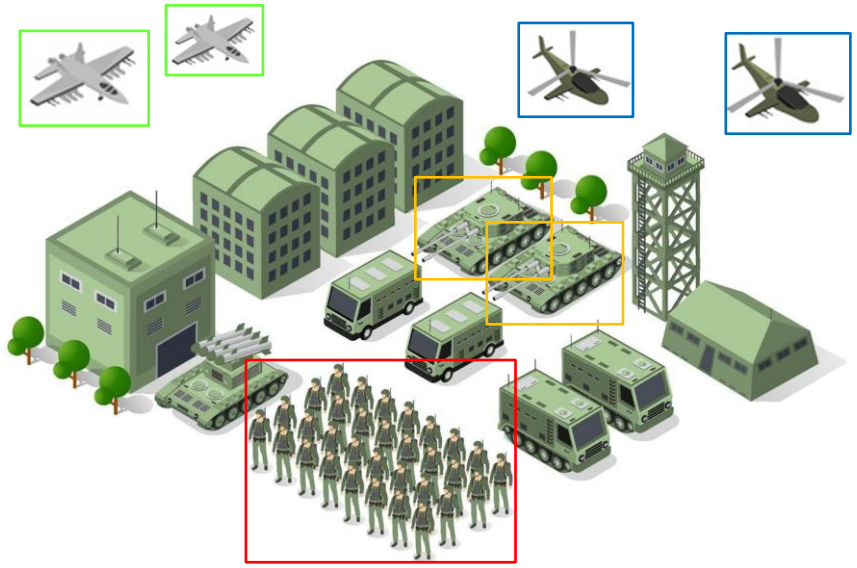


01 연구 목표

Zero-Shot Object Detection의 필요성

- **학습 데이터에 없는 신종 위협 탐지**
  - 자연어 설명으로 새로운 객체를 탐지 가능

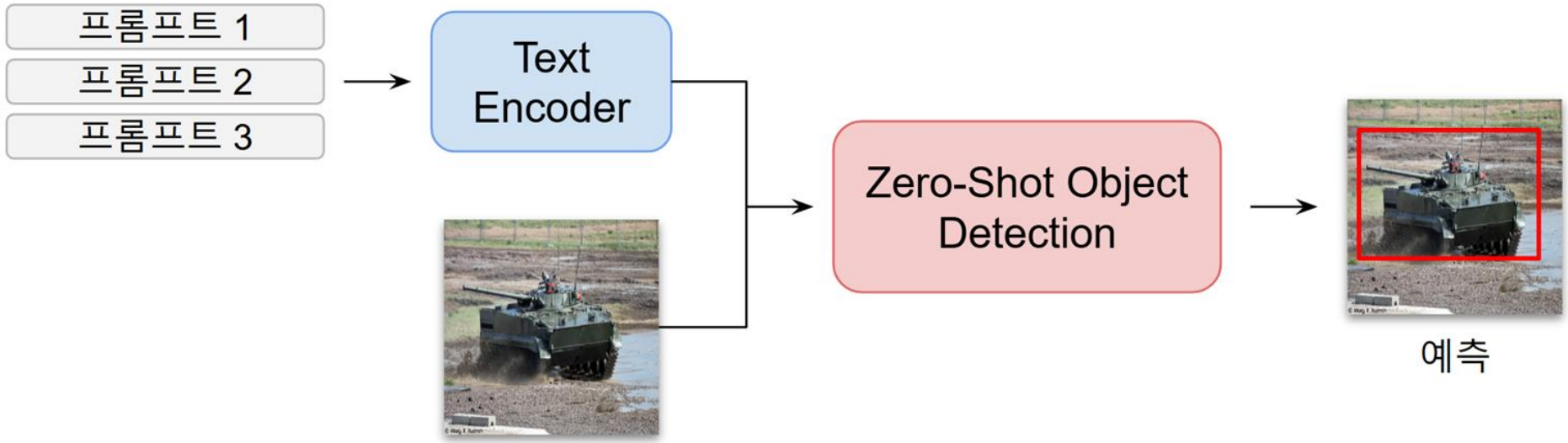
- **제로샷 탐지를 통한 전술적 우위 확보**
  - 사전 정보가 없는 상황에서도 탐지 가능
  - 특히 새로운 무기체계나 시설을 식별 가능



01 연구 목표

연구 목표: 항공영상 기반의 제로샷 객체 탐지 연구

- **시각-언어 모델(VLM) 기반** 군용 객체 탐지 성능 향상을 위한 **프롬프트 튜닝** 실험
- 군용 **데이터 부족 문제**를 **보완**하기 위한 **데이터 합성 기법** 제안
- **제로샷 객체 탐지 성능 개선**을 위한 방법 제안



제로샷 객체 탐지 파이프라인

02

연구 수행 내용

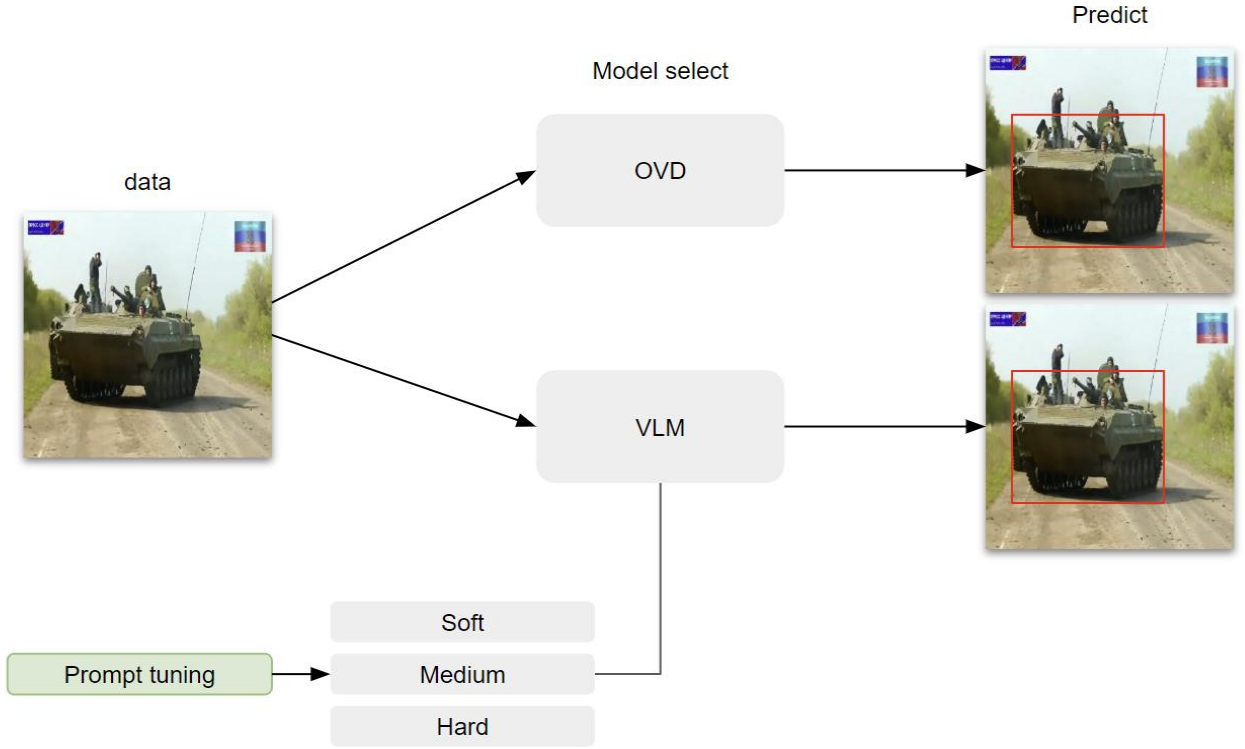
02 연구 수행 내용

1. 진행 사항

- **시각-언어 모델(VLM) 기반** 군용 객체 탐지 성능 향상을 위한 **프롬프트 튜닝** 실험

3단계 난이도 프롬프트 설계

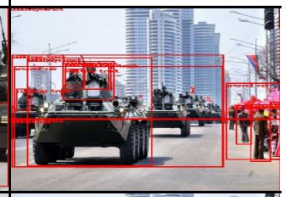
- Soft: 탐지 기준 완화, 유사 객체까지 포함
- Medium: 주요 특징 포함 시 탐지,
  불확실한 객체는 제외
- Hard: **명확한 시각적 특징**(위장, 장갑 등)
  있는 경우만 탐지

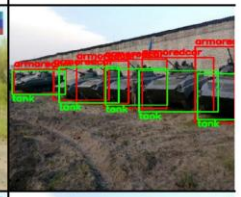
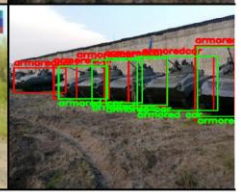


02 연구 수행 내용

1. 진행 사항

- 모델별 / 프롬프트 단계별 객체 탐지 결과

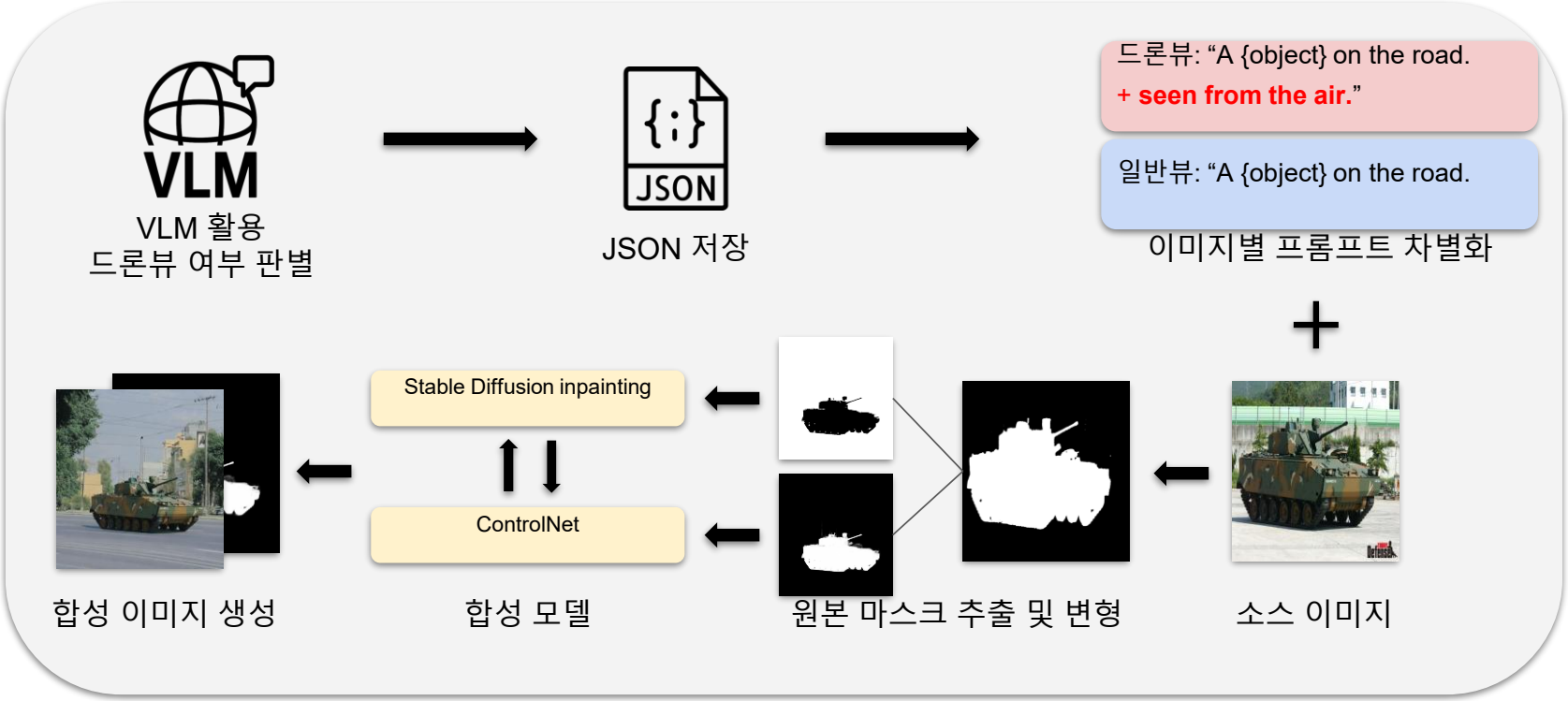
모델	이미지1	이미지2
YOLO-World		
Grounding DINO		
Florence-2		
Qwen_S		
Qwen_M		
Qwen_H		

모델	이미지1	이미지2
Qwen_S		
Qwen_M		
Qwen_H		

02 연구 수행 내용

1. 진행 사항

- 군용 차량 **데이터 부족 문제**를 보완하기 위한 **데이터 합성 기법** 제안



02 연구 수행 내용

1. 진행 사항

- 합성 결과

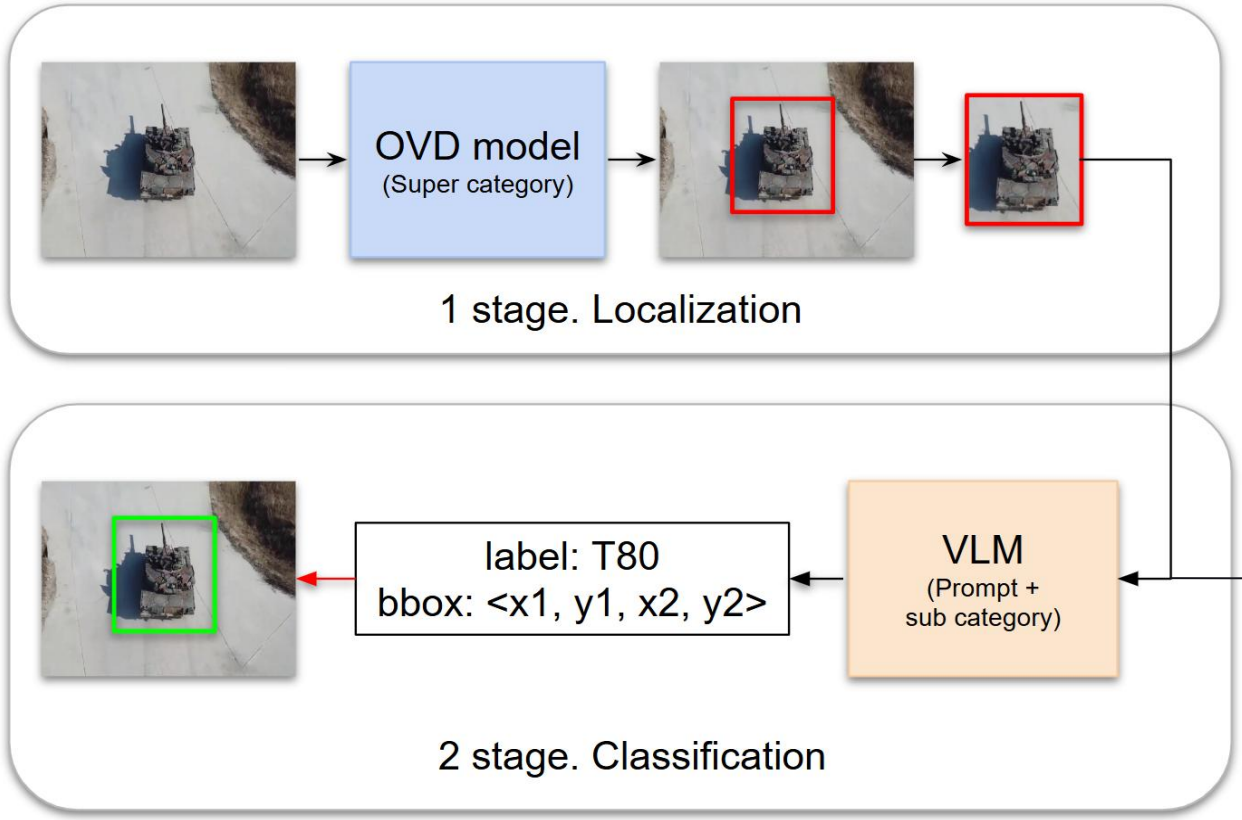
합성 기법별 정성적 비교

제안 기법의 합성 결과

02 연구 수행 내용

1. 진행 사항

- **Two-stage 전략(Grounding DINO + Qwen 2.5 VL)**



wiki Info: 각 class 외형, 설명

"For reference, here are brief descriptions for each class:"

"- **K2 Black Panther**: A South Korean main battle tank designed for high-speed maneuver warfare."

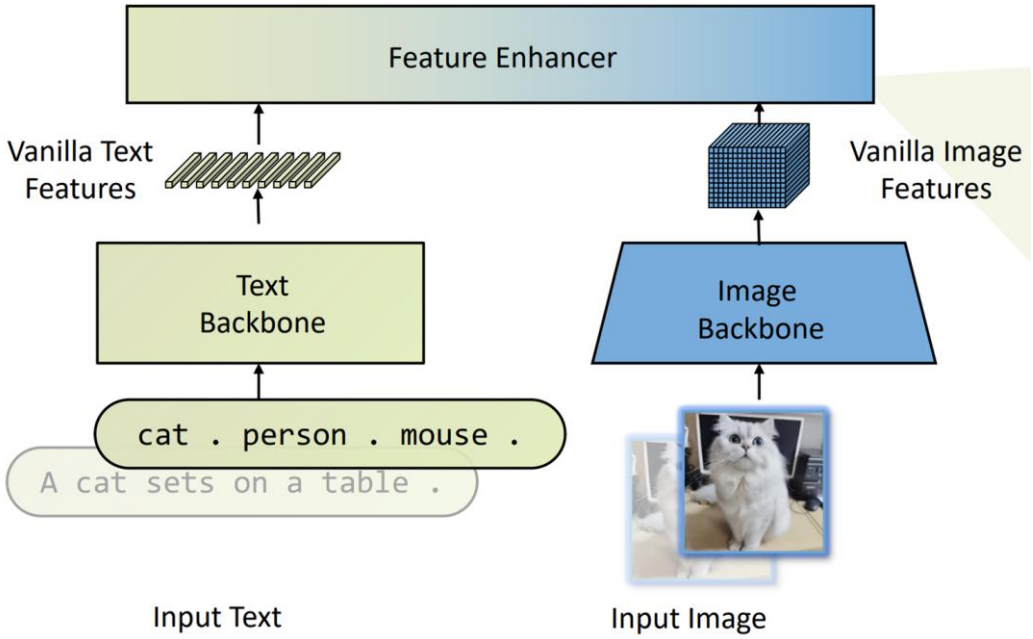
"- **M113 armored personnel carrier**: A widely used fully tracked APC known for its versatility in rough terrain." ....

02 연구 수행 내용

1. 진행 사항

- **Grounding-DINO**

	제목	저널/학술대회 명	연도
1	Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection	ECCV	2024



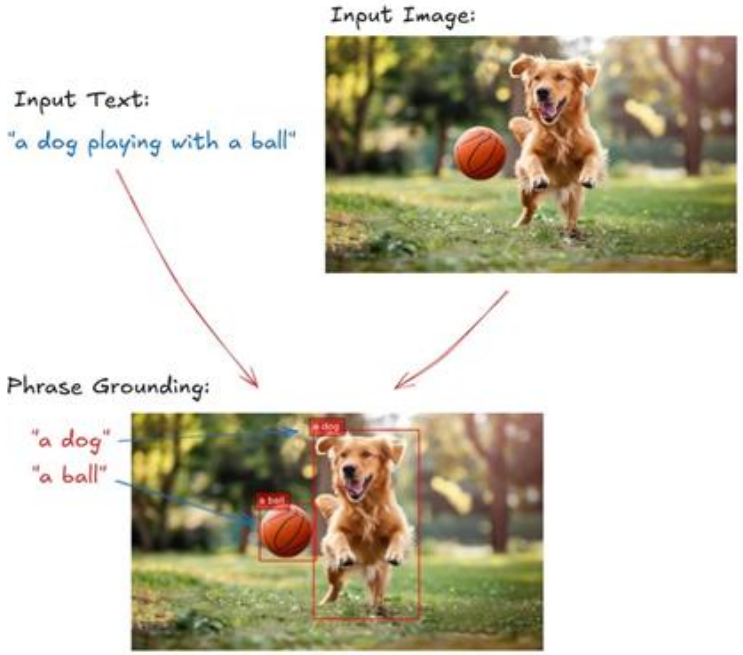
모델	military_object
YOLO-World	**0.299**
Grounding-DINO	**0.745**

단일 모델 mAP 성능

02 연구 수행 내용

1. 진행 사항

- **Florence-2**
  - **<CAPTION_TO_PHRASE_GROUNDING> Tag**를 사용하여 탐지 진행
  - Caption: “T80, K2, BMP-3, K200, Military Truck”
    - mAP 0.1 미만으로 세부 클래스 판별 능력이 낮음



Caption to Phrase Grounding

caption to phrase grounding task requires additional text input, i.e. caption.

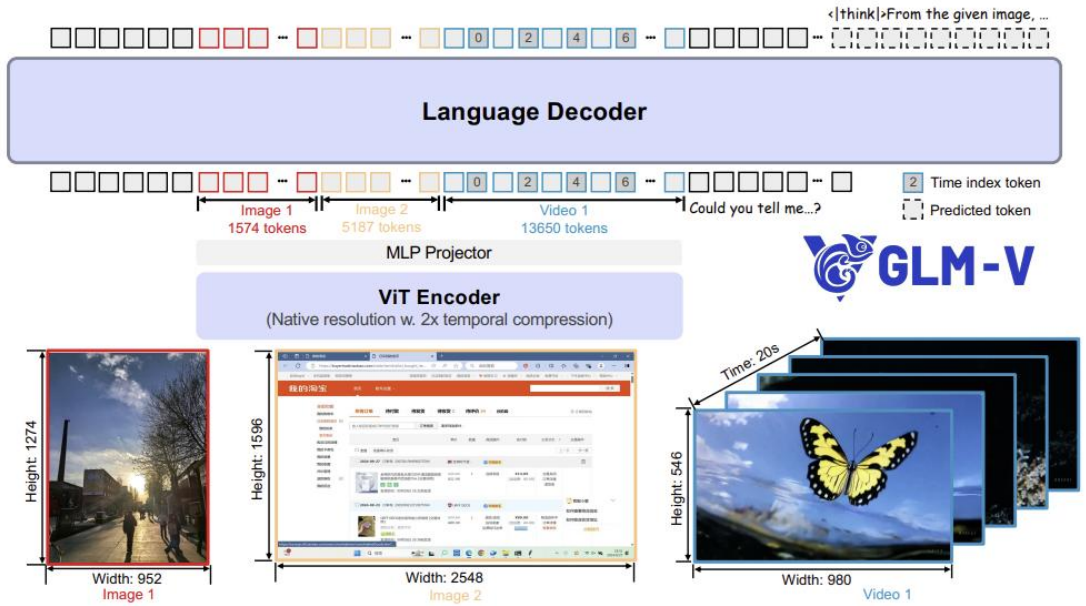
Caption to phrase grounding results format: {'<CAPTION_TO_PHRASE_GROUNDING>': {'bboxes': [[x1, y1, x2, y2], ...], 'labels': ['', '', ...]}}

```
task_prompt = "<CAPTION_TO_PHRASE_GROUNDING>"
results = run_example(task_prompt, text_input="A green car parked in
```

02 연구 수행 내용

1. 진행 사항

- **GLM-4.1-Thinking**
  - “T80, K2, BMP-3, K200, Military Truck”의 **외형적 정보를 프롬프트**로 입력
    - 모든 이미지를 K2로 예측
    - mAP 0.1 미만으로 세부 클래스 판별 능력이 낮음



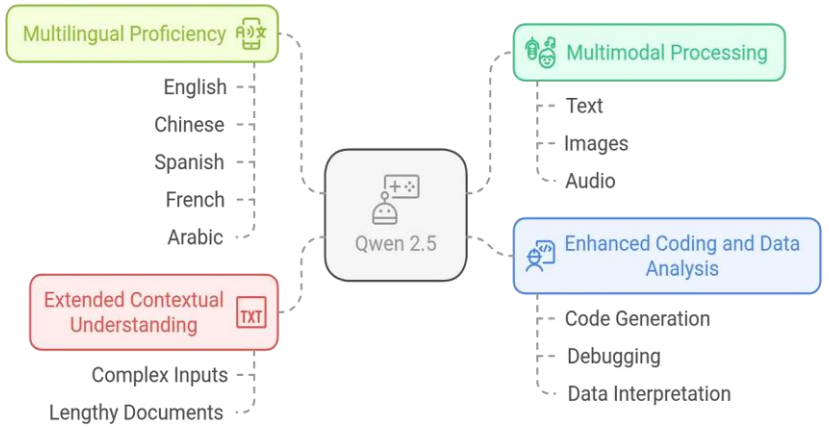
02 연구 수행 내용

1. 진행 사항

- **Qwen(Baseline)**

모델	프롬프트 입력 방법	T80	K2	BMP-3	K200	Military truck	mean
Qwen2.5-VL-Instruct(7B)	외형적 정보 기반	0.121	0.149	0.223	0.178	0.384	**0.264**

단일 모델 mAP 성능



02 연구 수행 내용

1. 진행 사항

- **Qwen(Baseline)**

tank_t80
상위 클래스: 전차(Main Battle Tank)
T-64를 기반으로 한 소련제 주력 전차이다.
가스터빈 엔진으로 구동되며, 무거운 장갑과 높은 기동성을 갖는다.

tank_k2
상위 클래스: 전차(Main Battle Tank)
대한민국에서 개발된 현대식 주력 전차이다.
고성능 사격 통제와 기동 능력을 갖춘다.

armoredcar_bmp3
상위 클래스: 보병전투차량(Infantry Fighting Vehicle, IFV)
소련/러시아에서 개발된 보병전투차량이다.
공격과 방어 임무를 모두 수행할 수 있다.

armoredcar_k200
상위 클래스: 장갑병력수송차량(Armored Personnel Carrier, APC)
대한민국에서 제작된 장갑병력수송차량이다.
병력 수송과 다양한 지원 임무에 활용된다.

military_truck
상위 클래스: 군용 차량(Military Vehicle)
물자와 인원을 험지에서 운송하기 위한 일반 군용 차량이다.
전차나 장갑전투차량은 포함하지 않는다.

02 연구 수행 내용

1. 진행 사항

- **Qwen(Prompt 1)**

모델	프롬프트 입력 방법	T80	K2	BMP-3	K200	Military truck	mean
Qwen2.5-VL-Instruct(7B)	외형적 정보 기반 + 드론뷰 명시	0.125	0.166	0.221	0.221	0.429	**0.284**

단일 모델 mAP 성능

02 연구 수행 내용

1. 진행 사항

- **Qwen(Prompt 1)**

공통 프롬프트

이미지는 드론 시점(drone-view)과 지상 시점(ground-level) 모두를 포함할 수 있다.

1. 진행 사항

- **Qwen(Prompt 2)**

모델	프롬프트 입력 방법	T80	K2	BMP-3	K200	Military truck	mean
Qwen2.5-VL-Instruct(7B)	외형적 구별 특징 선별	0.155	0.342	0.253	0.277	0.556	**0.368**

단일 모델 mAP 성능

02 연구 수행 내용

1. 진행 사항

- **Qwen(Prompt 2)**

tank_t80
T-64를 기반으로 한 소련제 주력 전차로, 오직 가스터빈 엔진으로 구동된다. 무거운 장갑에도 불구하고 높은 속도와 기동성이 뛰어나며, T-80B와 T-80U 같은 변형 모델들이 있다.

tank_k2
고급 사격 통제, 차체 내 서스펜션, 네트워크 중심 기능을 갖춘 최첨단 대한민국 전차이다. 120mm 활강포와 자동 장전 장치를 통해 빠른 표적 획득과 높은 발사 속도를 제공한다.

armoredcar_bmp3
100mm 대포, 30mm 자동포, 견고한 포탑 시스템을 갖춘 다목적 소련/러시아 보병전투차량이다. 공격적 역할과 방어적 역할 모두에서 뛰어난 성능을 발휘한다.

armoredcar_k200
경량형, 수륙양용, 모듈형 설계를 갖춘 현대적인 대한민국 장갑병력수송차량(K200 KIFV)이다. 병력 수송, 정찰, 지원과 같은 역할에 적응할 수 있다.

military_truck
주요 전차와 장갑 전투 차량을 포함하지 않고, 트럭 유형을 포함한, 험한 지형에서 물자와 인력을 운송하기 위해 내구성과 효율성을 갖춘 표준 군용 화물/지원 차량이다.

1. 진행 사항

- **Qwen(Prompt 2)**

일부 샘플의 객체를 **군집으로 반환**하는 결과가 존재

03

차후 계획

03 차후 계획

차후 계획

1. 추가적 Zero-shot Object Detection 방법론 탐색
2. 군용 차량을 위한 Two-stage Zero-shot Object Detection 기법 제안
3. Two-stage 모델 성능 개선 방안 탐색
4. 비교 논문 대비 State Of The Art(SOTA) 성능 달성

Q&A