[2026 winter seminar]

Agentic Query Refinement for Open-Vocabulary Object Detection

Pusan National University
Computer Vision & Signal Processing Lab, Dept. of Electronic Engineering

Presented By
정다운

CVSP Lab
Computer Vision & Signal Processing

Index

1) Wang, Ao, et al. "Yoloe: Real-time seeing anything." arXiv preprint arXiv:2503.07465 (2025).
2) Chen, Yuming, et al. "Re-Aligning Language to Visual Objects with an Agentic Workflow." arXiv preprint arXiv:2503.23508 (2025).

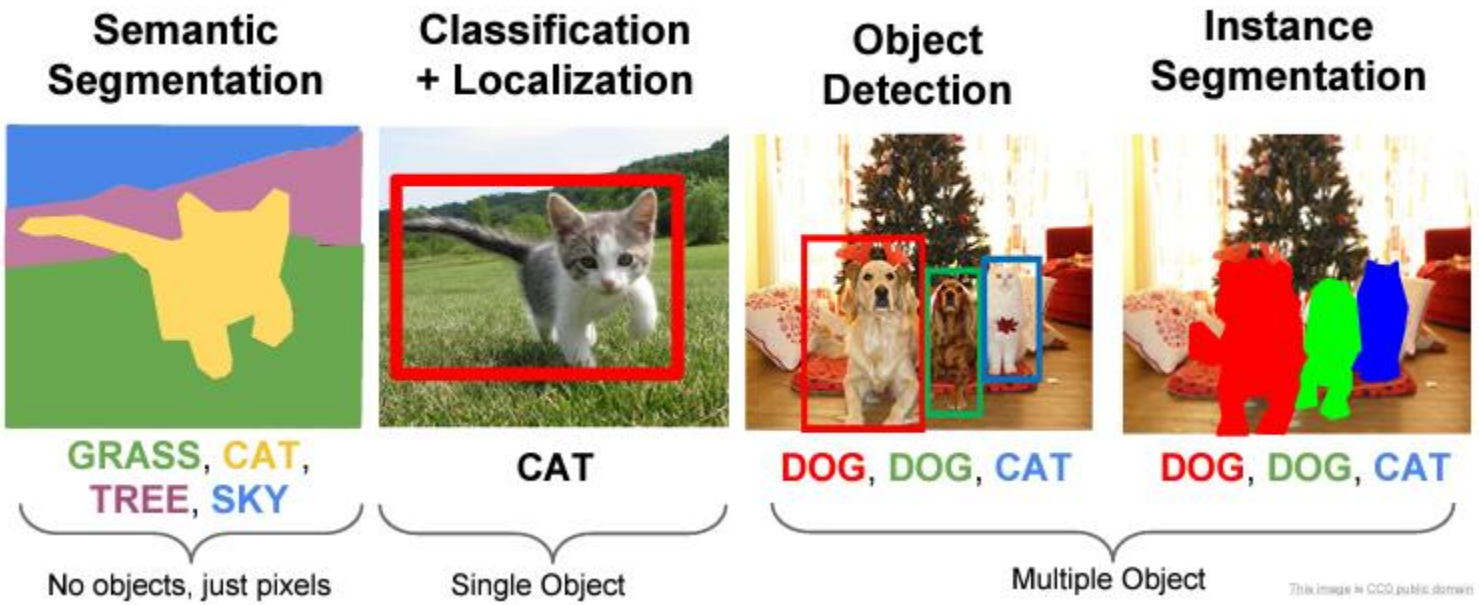
- Purpose of today's seminar
- 논문 소개
  - YOLOE: Real-Time Seeing Anything [ICCV 2025] [1)]
  - Re-Aligning Language to Visual Objects with an Agentic Workflow [ICLR 2025] [2)]
- Conclusion

Background

- Purpose of today's Seminar
  - Open-Vocabulary Object Detection
    - Closed-set Object Detection과 Open-Vocabulary Object Detection의 차이 이해
    - Open-Vocabulary Object Detection이 가지는 특징 파악
    - YOLOE가 속도를 올리기 위해 제안한 방법 이해
  - Agentic flow
    - Agentic system을 언제, 왜 쓰는지 알기
    - Agentic flow의 일반적인 구조(Planning- Execution – reflection) 이해

Background

- Object Detection
  - Computer vision분야의 대표적인 Task 중 하나
  - 이미지 내에서 물체를 분류하는 Classification과 그 물체가 어디에 있는지 Bounding box를 통해 위치를 나타내는 Localization을 함께 해결하는 분야



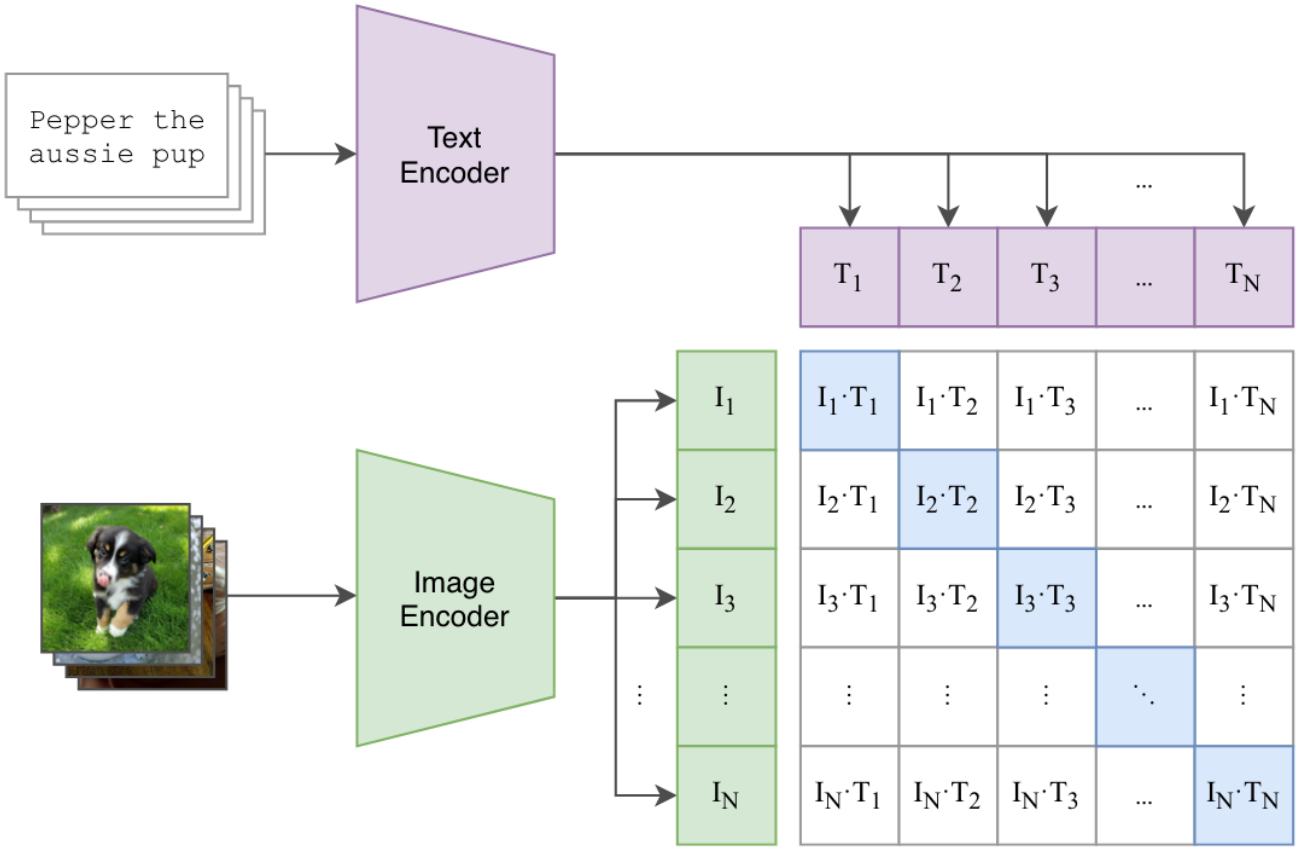
Background

- Open-Vocabulary Object Detection(OVOD)
  - Closed-set Object Detection
    - 학습 시에 정의된 고정 클래스만 탐지
    - 클래스 = 정수 ID (0, 1, 2 , …, n)
    - 각 bbox마다 C차원의 logit이 출력되고 softmax적용하여 최종 class 결정
  - Open-Vocabulary Object Detection
    - 자연어 텍스트 쿼리로 객체 정의
      - 학습 중 없었던 novel객체도 탐지 가능
    - 클래스 = text embedding
    - Region과 text feature간의 cosine similarity로 score 계산

CVSP Lab
Computer Vision & Signal Processing

Background

- Open-Vocabulary Object Detection(OVOD)
  - CLIP model



CVSP Lab
Computer Vision & Signal Processing

Background

- Open-Vocabulary Object Detection(OVOD)

- 학습 중에 없었던 **새로운 객체**도 탐지 가능
- 사전에 정의된 Class에 의존하지 않고 **자연어 Query기반으로 객체 인식**
- 클래스를 추가할 때 **재학습이 필요 없어** 확장성이 높음

실제 상황에서 Class를 미리 정의할 수 없는 경우에 적합

- **매번 새로운 물건이나 상황이 등장하는 경우 → 매번 모델 재학습 불가**
- **사용자가 원하는 물건을 맞춤으로 인식해야 하는 경우 → "red cup", "speaker under the TV"**







CVSP Lab
Computer Vision & Signal Processing

1) Wang, Ao, et al. "Yoloe: Real-time seeing anything." arXiv preprint arXiv:2503.07465 (2025).



YOLOE: Real-Time Seeing Anything[1)]

ICCV 2025

Introduction

- Motivation
  - Text prompt-based detection
    - Text encoder 매번 실행, image-text 연산 → 계산량 많음
  - Visual prompt-based detection
    - Transformer 기반, Visual encoder 필요함 → 계산량 많음
  - Prompt-free detection
    - Large Language Model(LLM) 의존, 매우 느림



< Visual prompt 기반 detection, segmentation 예시 >

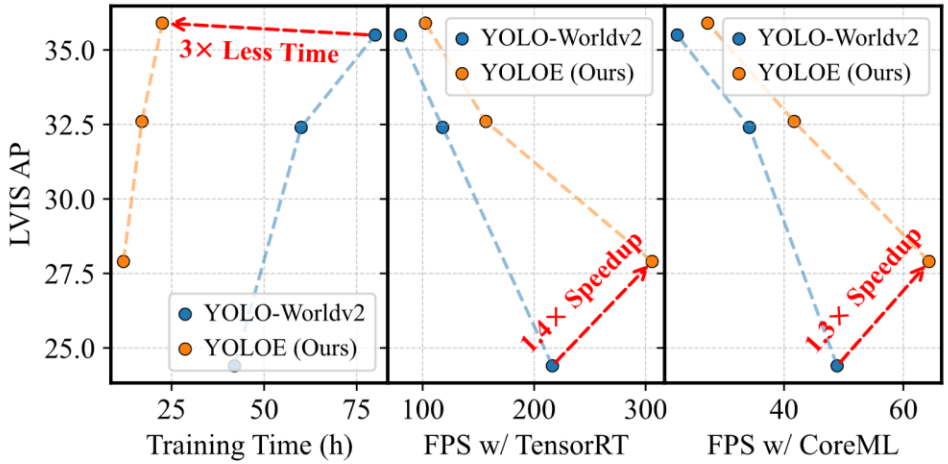


< Prompt-free detection, segmentation 예시 >

CVSP Lab
Computer Vision & Signal Processing

Introduction

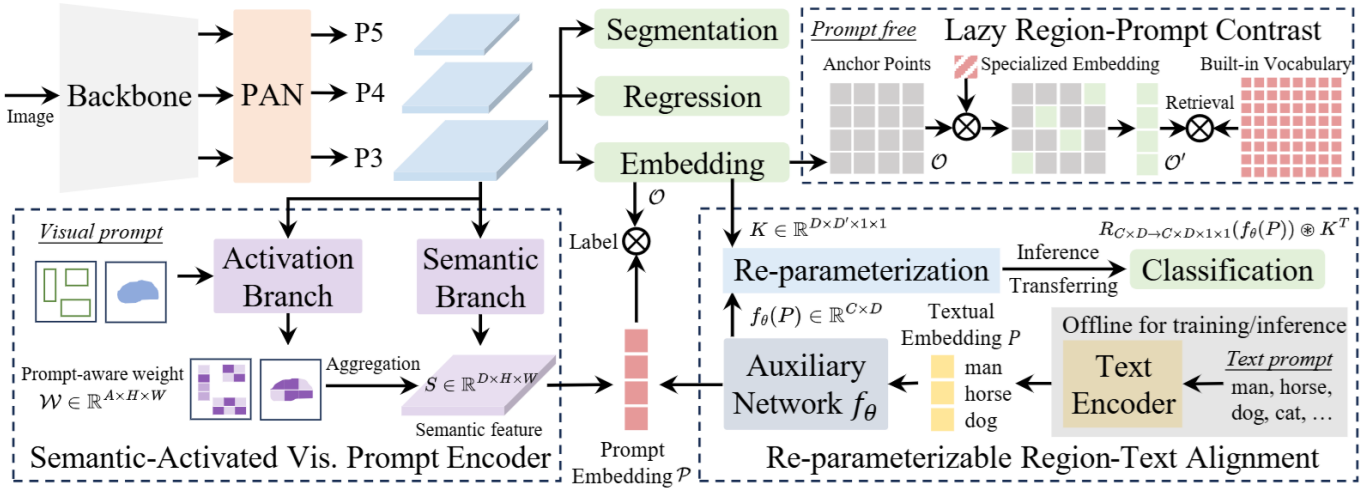
- Contribution
  - 빠르고 통합된 Open-set detector
    - 3가지 prompt기반 detection을 한 모델에서 가능하게 함
    - 기존 통합모델(DINO-X)은 사이즈가 커 real-time 불가
  - 각 prompt 방식별 경량화 방법 제안
    - Text: RepRTA (inference overhead 0)
    - Visual: SAVPE (minimal complexity)
    - Prompt-free: LRPC (no LLM dependency)



CVSP Lab
Computer Vision & Signal Processing

Method

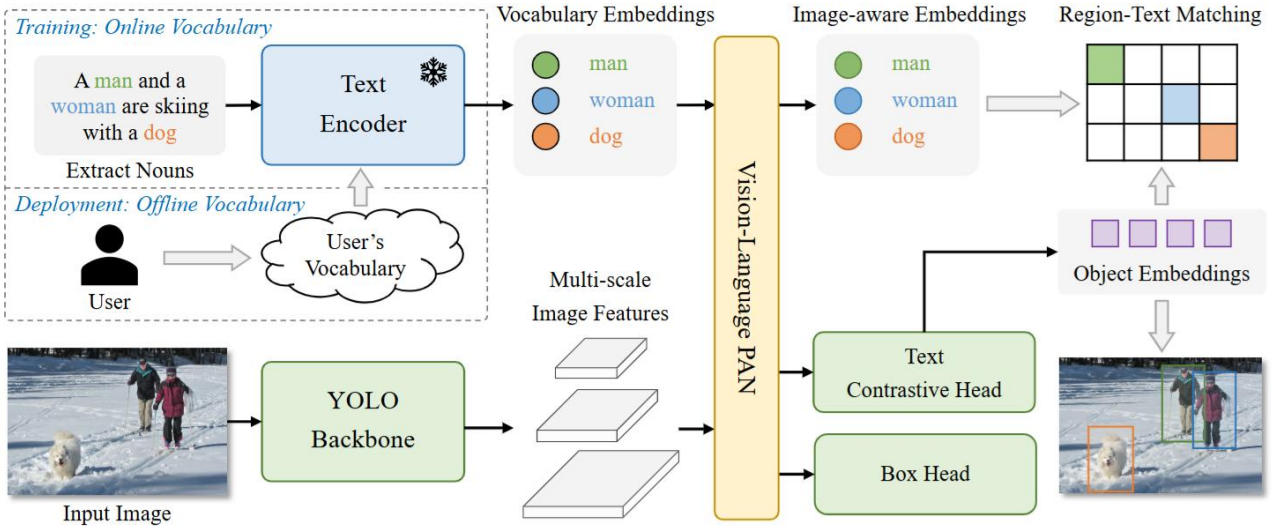
- The overview of YOLOE
  - 세 가지 prompt에 대한 방법을 각각 최적화 하여 하나의 모델로 통합
    - Text prompt: Region-Text Alignment (RepRTA)
    - Visual prompt: Semantic-Activated Visual Prompt Encoder (SAVPE)
    - Lazy Region-Prompt Contrast(LRPC)



CVSP Lab
Computer Vision & Signal Processing

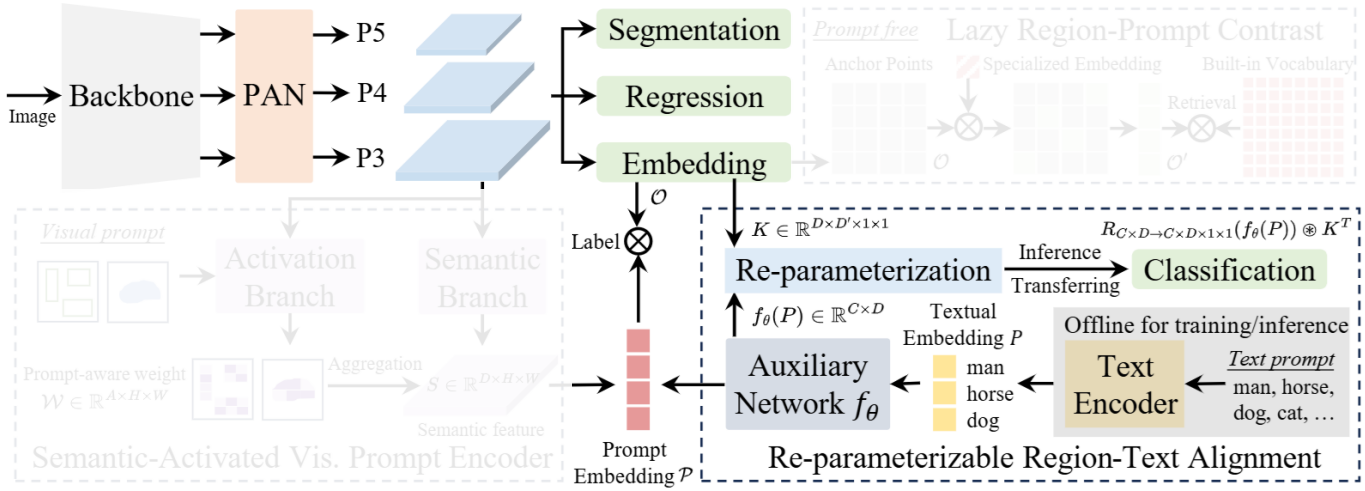
Method

- 기존 Text prompt based OVOD
  - 기존 text-prompt 기반 OVOD는 text prompt를 inference입력으로 사용하였음
    - Text prompt → Text feature
    - Image → region feature
  - Region feature와 text feature 간 similarity를 계산하여 Classification logit 계산
    - Inference 단계에서도 text feature가 필요
    - Text feature가 많아질수록 연산량이 증가함



Method

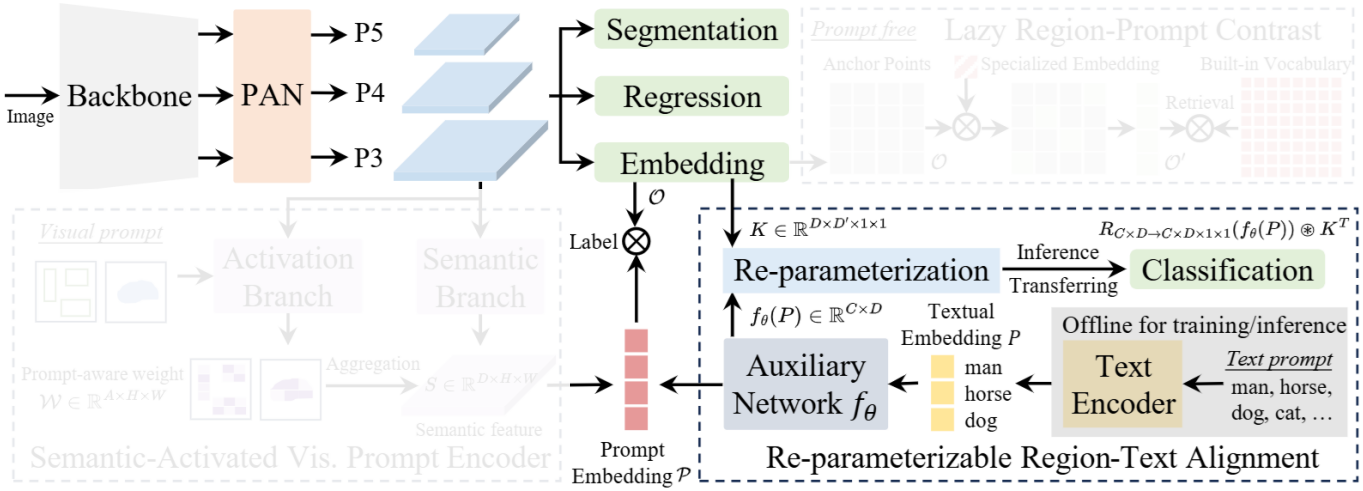
- Text prompt based OVOD
  - Re-parameterizable region-text alignment(RepRTA)
    - 기존엔 inference할 때 image-text feature similarity 연산 수행
      - Text query가 많아질 수록 계산량이 커지는 구조
    - YOLOE의 Core Idea
      - inference 때 계산하지 말고 training 때 계산한 걸 학습해서 모델 파라미터 안에 저장하자



Method

C: Text prompt의 개수
D: Embedding dimension

- Text prompt based OVOD
  - Re-parameterizable region-text alignment(RepRTA)
    - Training
      - 먼저 CLIP text encoder를 통해 pretrained textual embedding $P = TextEncoder(T)$ 뽑기
      - Light auxiliary network f_θ를 도입
        - ✓Training과정에서 anchor point의 object embedding과 contrast하여 visual-semantic alignment개선
        - ✓향상된 Textual embedding인 $\hat{P} = f_\theta(P) \in \mathbb{R}^{C \times D}$ 생성



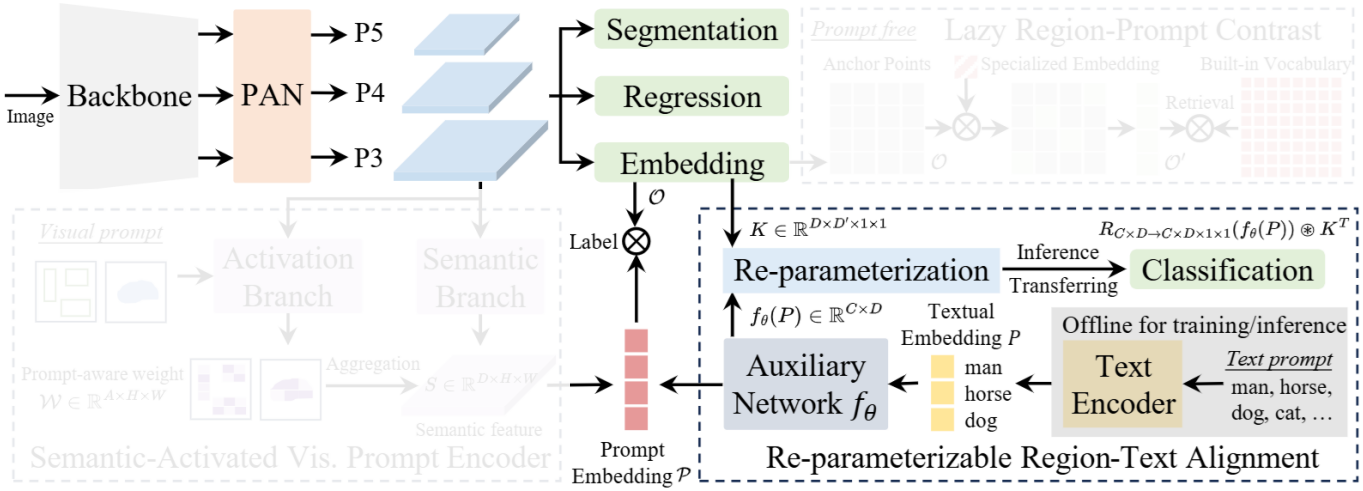
CVSP Lab
Computer Vision & Signal Processing

Method

- Text prompt based OVOD
  - Re-parameterizable region-text alignment(RepRTA)
    - Training

$$Label = R_{D\times H\times W \to HW\times D}(I \circledast K) \cdot \left(f_\theta(P)\right)^T$$

    - Training 중 YOLO region embedding $(I \circledast K)$ 과 보정된 text embedding $f_\theta(P)$ 간 dot product로 classification 수행
    - 이 과정에서 Text-region alignment가 학습됨



Method

- Text prompt based OVOD
  - Re-parameterizable region-text alignment(RepRTA)
    - Training

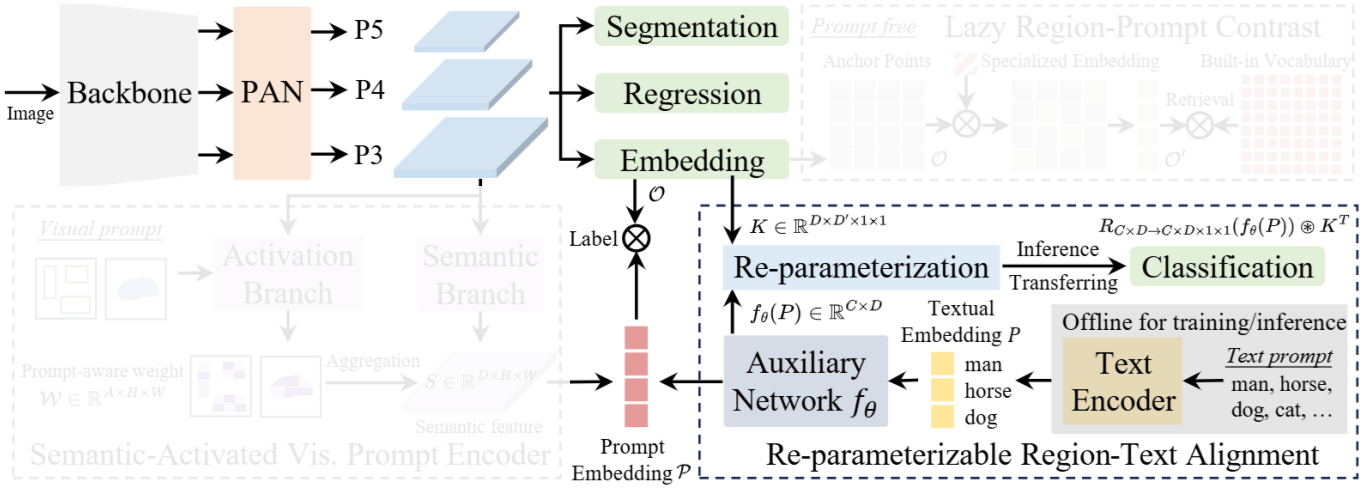
$$Label = R_{D\times H\times W \to HW\times D}(I \circledast K) \cdot \left(f_\theta(P)\right)^T$$

✓Object embedding 생성

- $I \circledast K$

✓Text-region contrast: object embedding과 text embedding의 alignment를 계산

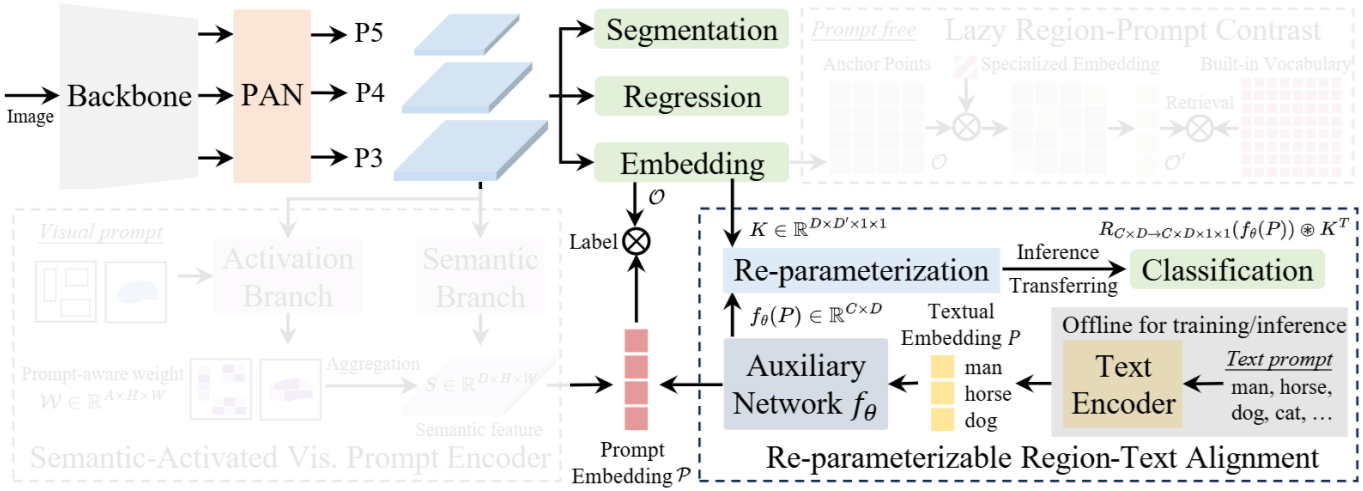
- $(R(I \circledast K)) \cdot \left(f_\theta(P)\right)^T$



CVSP Lab
Computer Vision & Signal Processing

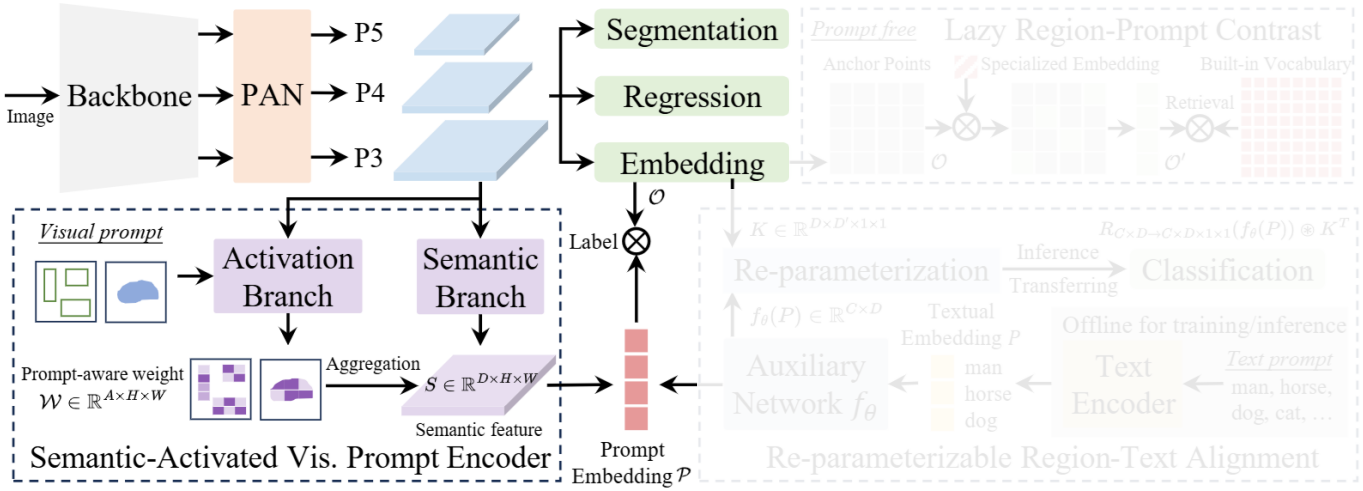
Method

- Text prompt based OVOD
  - Re-parameterizable region-text alignment(RepRTA)
    - Inference: auxiliary network를 제거하고 새로운 kernel로 합침
      - $K' = R_{C\times D\to C\times D\times 1\times 1}(f_\theta(P)) \circledast K^T$
      - 학습된 Text embedding $f_\theta(P)$
      - 1×1 conv kernel 형태로 reshape함
      - 기존 kernel K와 convolution하여 새로운 kernel 생성
    - 최종 Inference: Label=$I \circledast K'$



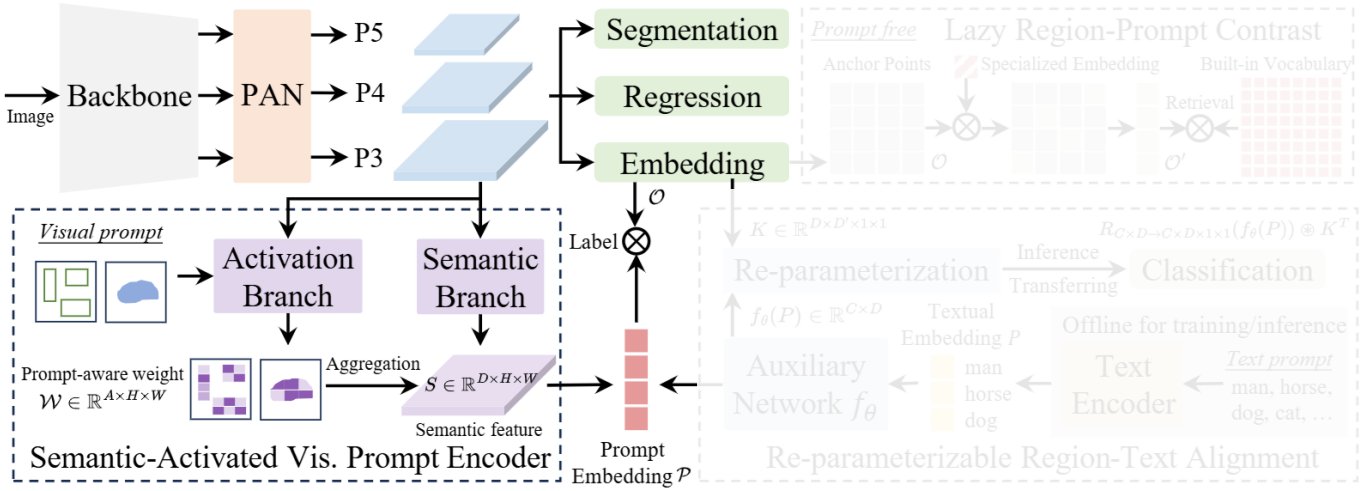
Method

- Visual prompt based OVOD
  - Semantic-activated visual prompt encoder (SAVPE)
    - Visual cue를 효율적으로 처리하기 위한 모듈 SAVPE 제안
    - SAVPE는 분리된 두 개의 lightweight branch로 구성
      - Semantic branch
        - ✓visual cue를 직접 결합하지 않고도 D 채널의 prompt-agnostic semantic feature를 출력한다
      - Activation branch
        - ✓훨씬 적은 채널 수에서 visual cue와 image feature를 상호작용시켜 grouped prompt-aware weight를 생성



Method

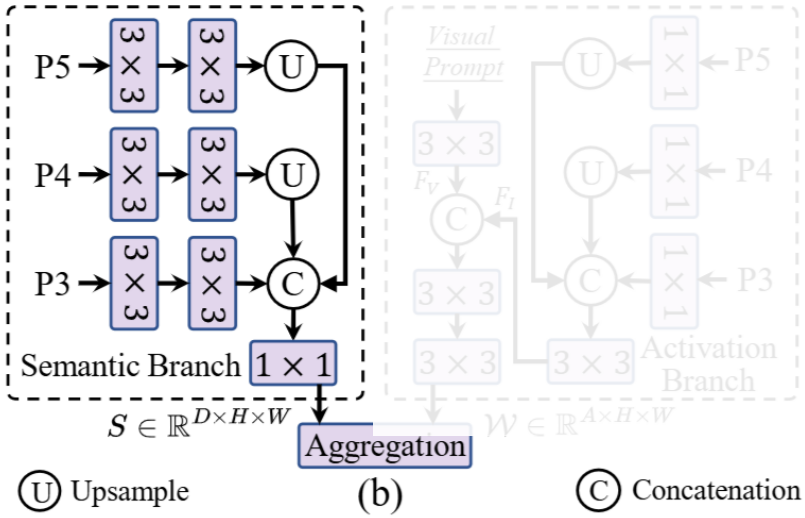
- Visual prompt based OVOD
  - Semantic-activated visual prompt encoder (SAVPE)
    - 기존 방법
      - Visual prompt, prompt feature를 cross attention/fusion 하여 연산량 많음
    - SAVPE 구조
      - Semantic branch: Image의 multi-scale feature가 들어가 semantic feature가 됨
      - Activation branch: semantic feature의 중요도(weight)를 조절해 줌



CVSP Lab
Computer Vision & Signal Processing

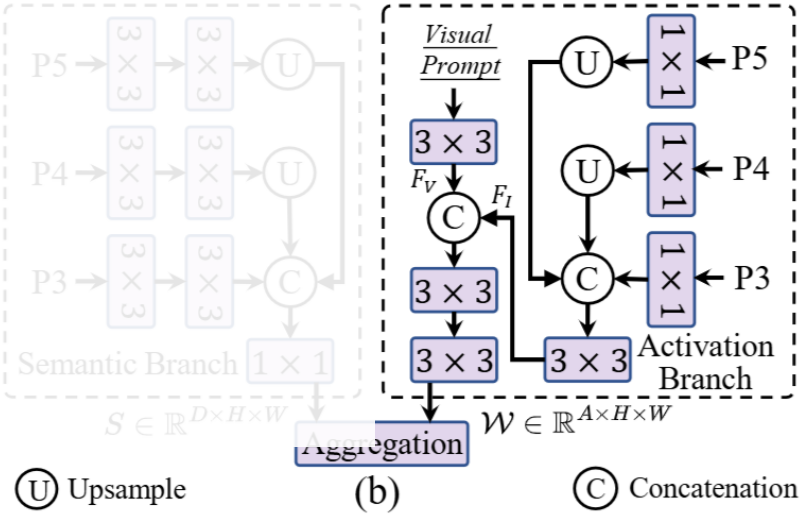
Method

- Visual prompt based OVOD
  - Semantic-activated visual prompt encoder (SAVPE)
    - Semantic Branch
      - Prompt와 무관하게 Image 자체의 semantic feature
        - ✓PAN에서부터 추출된 multi-scale feature에 대해 각각 두 개의 conv 를 적용
        - ✓Upsampling을 수행한 뒤 feature들을 concatenation하고 projection하여 $S \in \mathbb{R}^{D\times H\times W}$ 생성
        - ✓Projection 하여 semantic feature output



Method

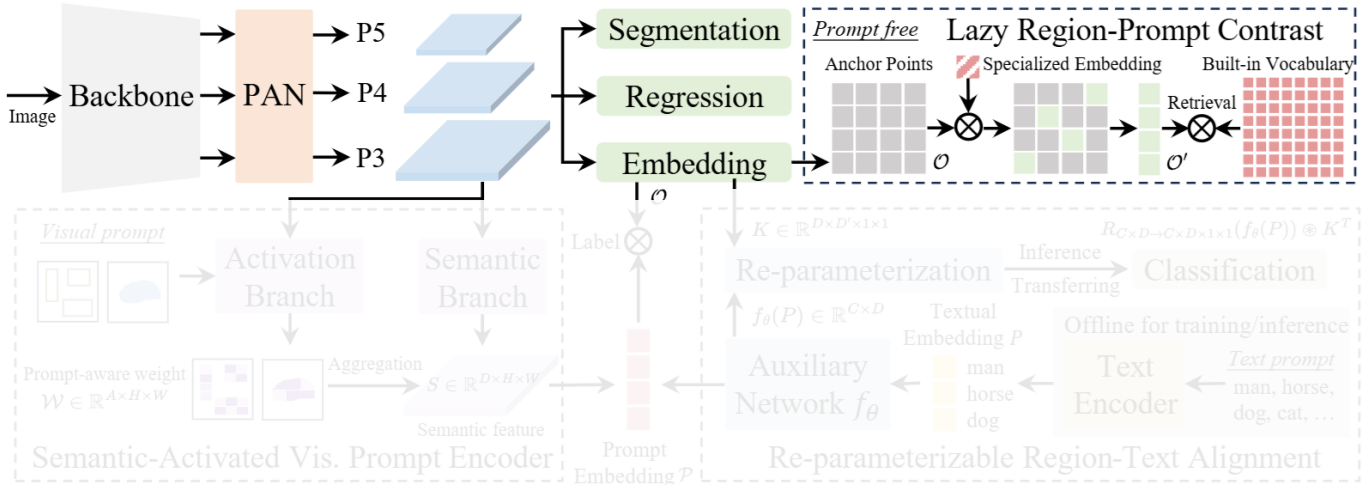
- Visual prompt based OVOD
  - Semantic-activated visual prompt encoder (SAVPE)
    - Activation Branch
      - Visual prompt를 찾고자 하는 영역에는 1, 그 외 영역에는 0을 갖게 masking
      - 이 mask를 downsampling한 후에 conv를 적용하여 prompt feature 생성
        - ✓$F_V \in \mathbb{R}^{A\times H\times W}$
      - Image feature에 conv를 적용해서 $F_I \in \mathbb{R}^{A\times H\times W}$ 생성
      - 둘을 concatenate하여 $\mathcal{W} \in \mathbb{R}^{A\times H\times W}$ 한 후 prompt-aware weight를 예측
      - prompt가 해당 영역 내에서 softmax 정규화



CVSP Lab
Computer Vision & Signal Processing

Method

- Prompt-free detection
  - Lazy region-prompt contrast(LRP)
    - 기존 방식
      - LLM을 활용해 object category를 생성 → Inference가 오래 걸림, 연산량 큼
    - LRP
      - 객체인지 아닌지만 판단하는 anchor만 추림
      - Built-in vocabulary에서 가장 잘 맞는 단어를 검색(retrieval)



Experiments

- Quantitative Results
  - Zero-shot Detection on LVIS

Model	Prompt Type	Params	Training Data	Training Time	FPS T4 / iPhone	AP	AP_r	AP_c	AP_f
GLIP-T [32]	T	232M	OG,Cap4M	1337.6h	- / -	26.0	20.8	21.4	31.0
GLIPv2-T [69]	T	232M	OG,Cap4M	-	- / -	29.0	-	-	-
GDINO-T [37]	T	172M	OG,Cap4M	-	- / -	27.4	18.1	23.3	32.7
DetCLIP-T [64]	T	155M	OG	250.0h	- / -	34.4	26.9	33.9	36.3
G-1.5 Edge [50]	T	-	G-20M	-	- / -	33.5	28.0	34.3	33.9
T-Rex2 [20]	V	-	O365,OI,HT CH,SA-1B	-	- / -	37.4	29.9	33.9	41.8
YWorldv2-S [5]	T	13M	OG	41.7h	216.4 / 48.9	24.4	17.1	22.5	27.3
YWorldv2-M [5]	T	29M	OG	60.0h	117.9 / 34.2	32.4	28.4	29.6	35.5
YWorldv2-L [5]	T	48M	OG	80.0h	80.0 / 22.1	35.5	25.6	34.6	38.1
YOLOE-v8-S	T / V	12M / 13M	OG	**12.0h**	**305.8 / 64.3**	**27.9** / 26.2	**22.3** / 21.3	**27.8** / 27.7	**29.0** / 25.7
YOLOE-v8-M	T / V	27M / 30M	OG	**17.0h**	**156.7 / 41.7**	**32.6** / 31.0	26.9 / **27.0**	**31.9** / 31.7	**34.4** / 31.1
YOLOE-v8-L	T / V	45M / 50M	OG	**22.5h**	**102.5 / 27.2**	**35.9** / 34.2	**33.2 / 33.2**	**34.8** / 34.6	**37.3** / 34.1
YOLOE-11-S	T / V	10M / 12M	OG	**13.0h**	**301.2 / 73.3**	**27.5** / 26.3	21.4 / **22.5**	26.8 / **27.1**	**29.3** / 26.4
YOLOE-11-M	T / V	21M / 27M	OG	**18.5h**	**168.3 / 39.2**	**33.0** / 31.4	26.9 / **27.1**	**32.5** / 31.9	**34.5** / 31.7
YOLOE-11-L	T / V	26M / 32M	OG	**23.5h**	**130.5 / 35.1**	**35.2** / 33.7	**29.1** / 28.1	**35.0** / 34.6	**36.5** / 33.8

CVSP Lab
Computer Vision & Signal Processing

Experiments

- Quantitative Results
  - Prompt-free evaluation on LVIS

Model	Backbone	Params	AP	AP_r	AP_c	AP_f	FPS
GenerateU [33]	Swin-T	297M	26.8	20.0	24.9	29.8	0.48
GenerateU [33]	Swin-L	467M	27.9	22.3	25.2	31.4	0.40
YOLOE-v8-S	YOLOv8-S	13M	**21.0**	19.1	21.3	21.0	**95.8**
YOLOE-v8-M	YOLOv8-M	29M	**24.7**	22.2	24.5	25.3	**45.9**
YOLOE-v8-L	YOLOv8-L	47M	**27.2**	23.5	27.0	28.0	**25.3**
YOLOE-11-S	YOLO11-S	11M	**20.6**	18.4	20.2	21.3	**93.0**
YOLOE-11-M	YOLO11-M	24M	**25.5**	21.6	25.5	26.1	**42.5**
YOLOE-11-L	YOLO11-L	29M	**26.3**	22.7	25.8	27.5	**34.9**

CVSP Lab
Computer Vision & Signal Processing

Conclusion

- Text Query Refinement
  - Open-Vocabulary Detection의 특성
    - 객체를 사전 정의된 Class로 탐지하는 것이 아닌 자연어 Query로 인식
    - 동일한 객체라도 Query 표현에 따라 탐지 여부와 Confidence가 달라짐
    - Open-Vocabulary Detection의 성능은 모델뿐 아니라 Query 설계에 의존
  - Query Refinement의 필요성
    - 실제 환경에서는 다양한 외형의 객체가 등장
    - 사용자가 설정한 Query는 실제 이미지의 특징을 표현하지 못하거나 일반적인 표현으로 성능이 낮을 수 있음
    - 실제 이미지를 기반으로 Query를 반복적으로 Refinement할 필요가 있음

2) Chen, Yuming, et al. "Re-Aligning Language to Visual Objects with an Agentic Workflow." arXiv preprint arXiv:2503.23508 (2025).

Re-Aligning Language to Visual Objects with an Agentic Workflow[2)]

ICLR 2025

Background

- Vision Language Model(VLM)
  - 이미지와 텍스트를 동시에 이해하고 연결하는 모델
    - 대표적으로 CLIP, ALIGN, BLIP, CoCa 등이 있음
  - Core Idea
    - 시각적 정보(image feature)와 언어적 정보(text feature)를 embedding space에서 정렬
    - 이를 통해 zero-shot image classification, retrieval, captioning 등 다양한 멀티모달 task 수행 가능
  - 구조
    - Vision Encoder: CNN, Vision Transformer(ViT) 등으로 image feature 추출
    - Text Encoder: Transformer 기반 언어모델로 텍스트 embedding 추출
    - Contrastive learning 등을 활용해 시각-언어 feature alignment

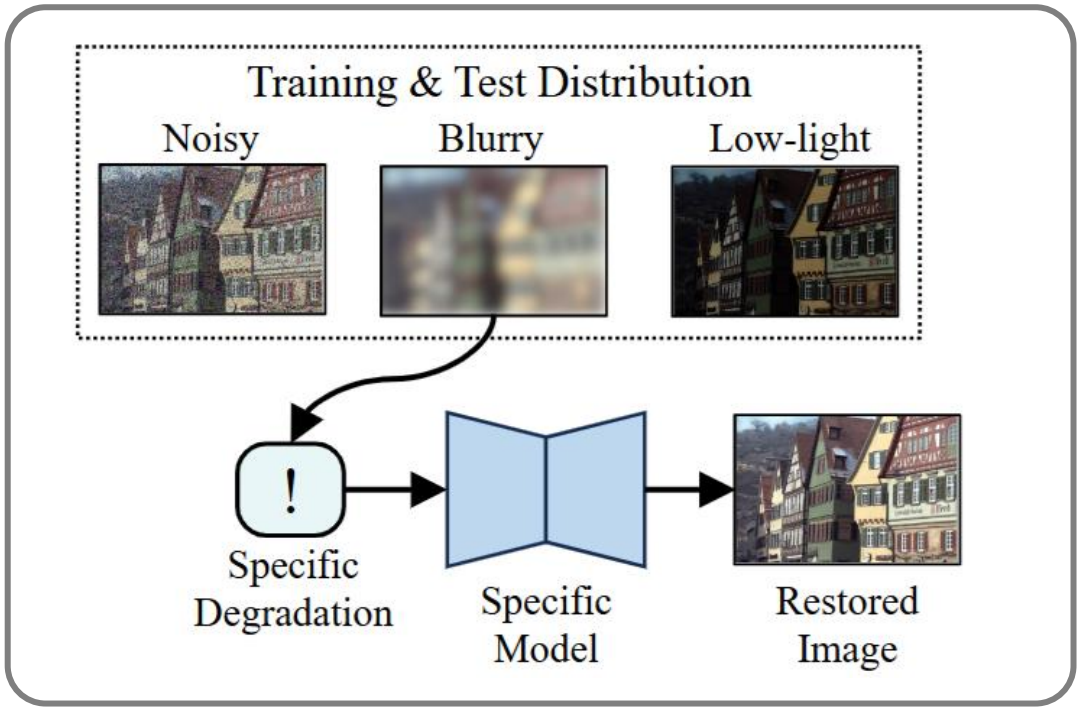
CVSP Lab
Computer Vision & Signal Processing

Background

- Agentic flow
  - LLM/VLM을 통해 순환적으로 여러 step을 수행하는 구조
    - 상태를 판단(Reasoning)
    - 다음 행동 결정(Planning)
    - 도구 사용(Tool use)
    - 결과를 평가·피드백 (Reflection)

Background

- Agentic flow 적용 예시
  - RestoreAgent: Autonomous Image Restoration Agent via Multimodal Large Language Models



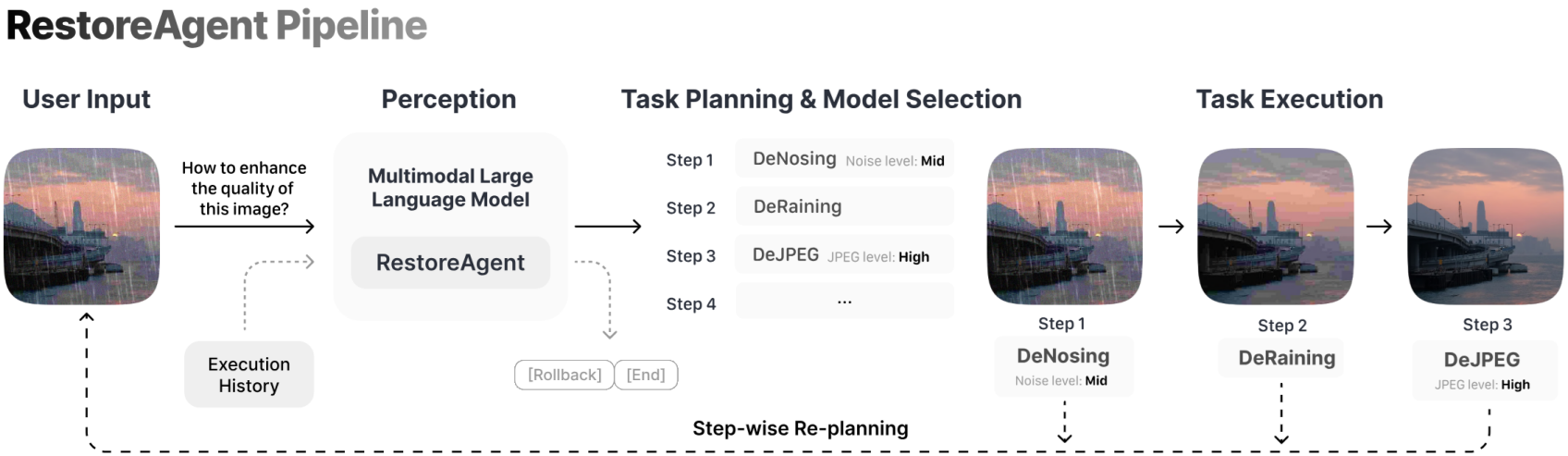
<**Classic** Image Restoration Model>

각 Degradation 유형별 특화된 모델이 있음

CVSP Lab
Computer Vision & Signal Processing

Background

- Agentic flow 적용 예시
  - RestoreAgent: Autonomous Image Restoration Agent via Multimodal Large Language Models



Background

- Agentic flow를 사용하면 좋은 경우

한번에 해결 불가능	중간결과를 보면서 판단을 반복적으로 수정해야 하는 경우
상황 인식이 필요한 경우	문제 상황이 여러 case → Case별 최적화된 tool
	여러 문제 상황이 겹쳐 복수의 tool을 선택·조합해야 하는 경우
	Tool 조합 및 순서 최적화가 필요한 경우

CVSP Lab
Computer Vision & Signal Processing

Introduction

- Motivation
  - OVOD 모델 학습을 위해 다양한 표현을 가진 데이터셋이 필요
  - VLM을 활용하여 다양한 expressions를 가진 training data pair 생성
    - Language Query가 target object에 대한 설명을 다양하게 할 수록 기존 연구들의 detection performance가 향상 됐음
  - VLM의 hallucinations로 인해 target object를 잘 설명하지 못함
    - → 데이터 규모 확장만으로는 Object Detection 성능 향상을 보장하지 않음

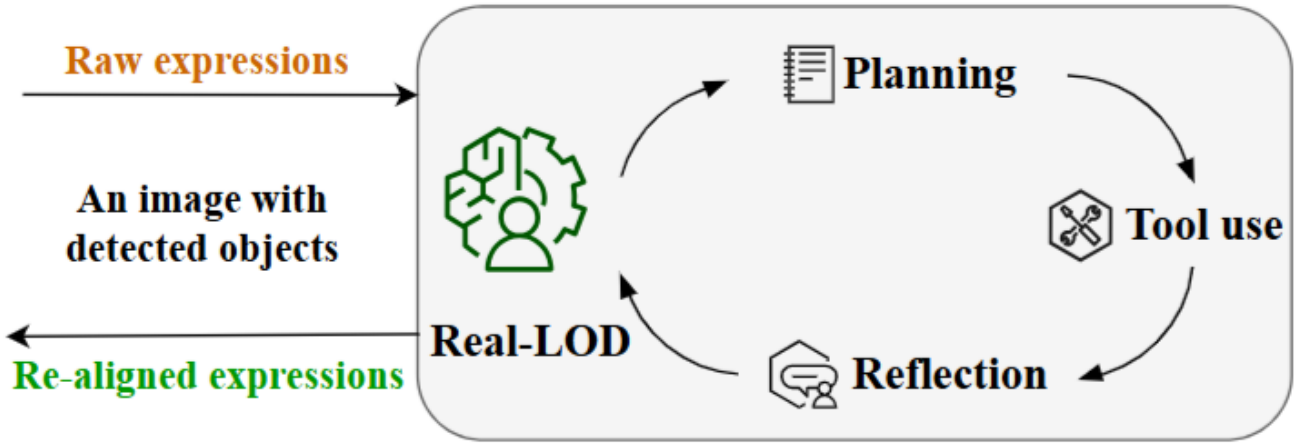


(a) VLM for a small object (b) VLM with extended crop (c) VLM with general prompt (d) VLM with specified prompt

CVSP Lab
Computer Vision & Signal Processing

Introduction

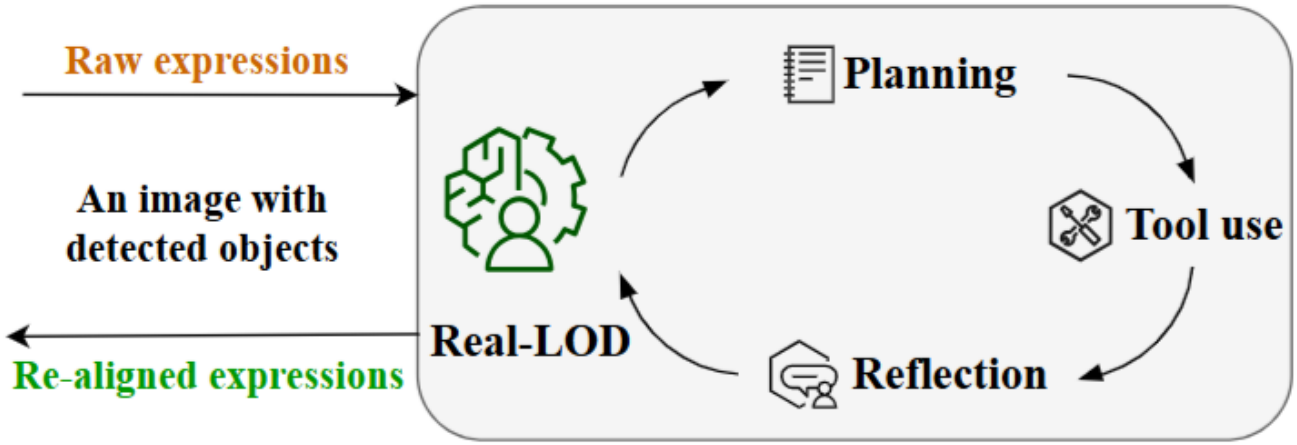
- Proposed Method
  - Real-LOD(Agentic flow)
    - 목적: Language expressions를 visual object에 맞게 Re-Align 하는 것
      - 데이터셋을 정제하는 것
    - Input: image + detected object + raw language expression
    - Output: object aligned language expression



CVSP Lab
Computer Vision & Signal Processing

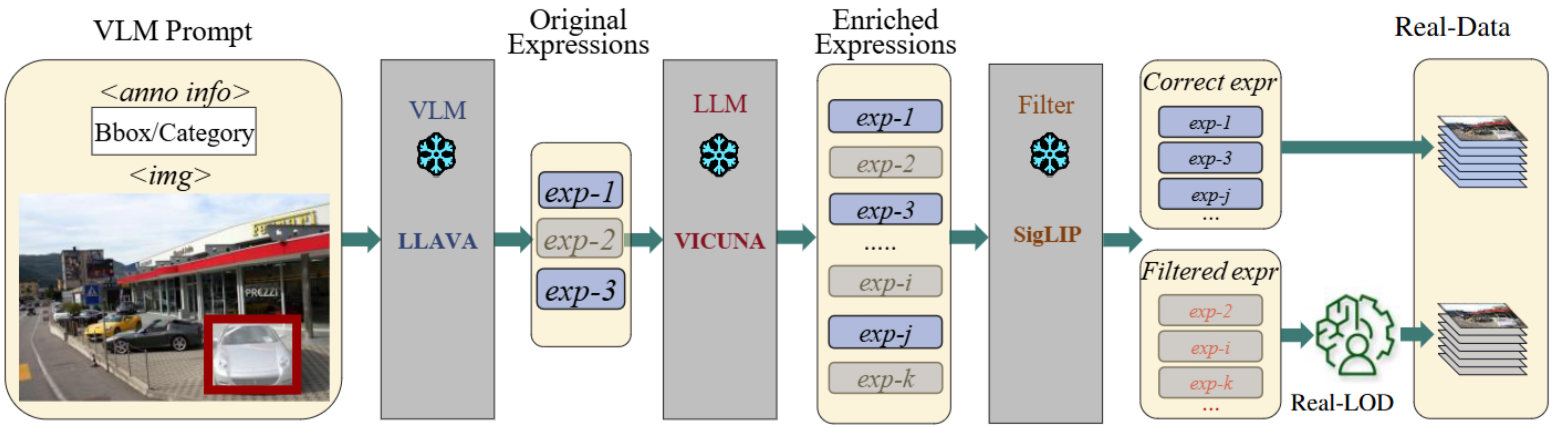
Introduction

- Proposed Method
  - Real-LOD(Agentic flow)
    - Agent가 Detection을 수행한다는 것 X
    - Detection model을 학습시킬 데이터셋을 정제하는 과정을 Agent화 한 것



Method

- Language expression generations
  - VLM model로 이미지에 포함된 object들에 대해 language expressions생성
    - 188k images에 포함된 336.5k objects들에 대해 673k language expressions 생성
    - LLM을 통해 동의어를 생성해 1346.1k개의 language expression 생성
  - Matching score가 낮은 data pair를 Re-align대상으로 삼음
    - SigLIP score를 통해 0.5 미만인 pair는 alignment에 문제 가능성 있음
    - Real-LOD Agent가 처리할 수 있도록 함

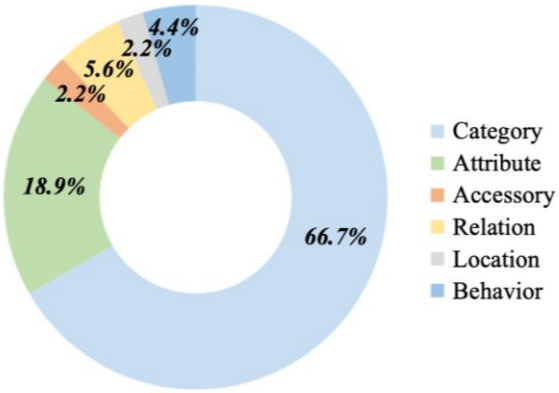


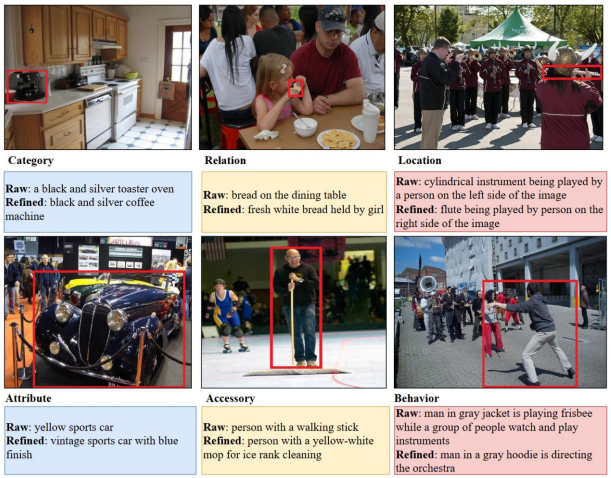
< Real-LOD: Expressions generation pipeline >

CVSP Lab
Computer Vision & Signal Processing

Method

- Analysis of language and visual objects
  - Real-LOD가 처리할 expressions들이 misalign되는 유형 분석
    - 1) Category: target object 가 아닌 다른 object를 설명하는 경우
    - 2) Attribute: color, shape, texture 등 target object의 attribute를 잘못 설명하는 경우
    - 3) Accessory: target object의 accessory를 잘못 설명하는 경우
    - 4) Location: image내 target object의 상대적 위치를 잘못 설명하는 경우
    - 5) Relation: 주변과의 object relation을 잘못 설명하는 경우
    - 6) Behavior: object 또는 human의 behavior를 잘못 설명하는 경우

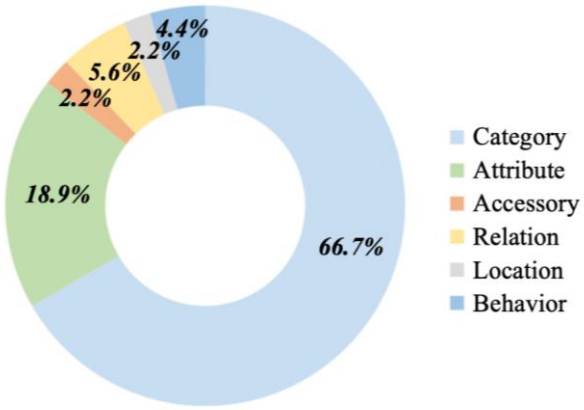




CVSP Lab
Computer Vision & Signal Processing

Method

- Analysis of language and visual objects
  - Real-LOD가 처리할 expressions들이 misalign되는 유형 분석





Category

Raw: a black and silver toaster oven
Refined: black and silver coffee machine



Relation

Raw: bread on the dining table
Refined: fresh white bread held by girl



Location

Raw: cylindrical instrument being played by a person on the left side of the image
Refined: flute being played by person on the right side of the image



Attribute

Raw: yellow sports car
Refined: vintage sports car with blue finish



Accessory

Raw: person with a walking stick
Refined: person with a yellow-white mop for ice rank cleaning



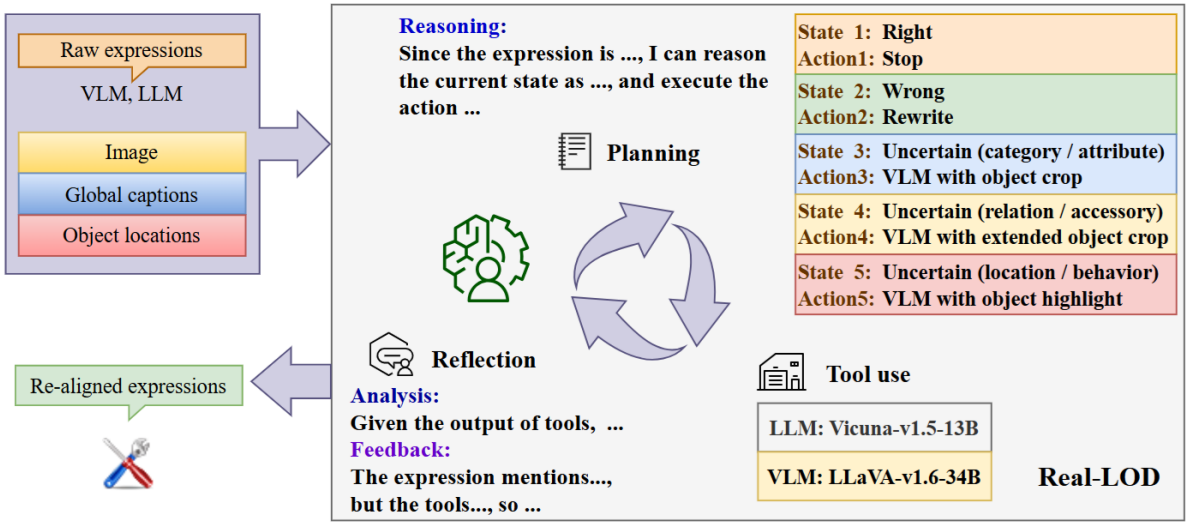
Behavior

Raw: man in gray jacket is playing frisbee while a group of people watch and play instruments
Refined: man in a gray hoodie is directing the orchestra

CVSP Lab
Computer Vision & Signal Processing

Method

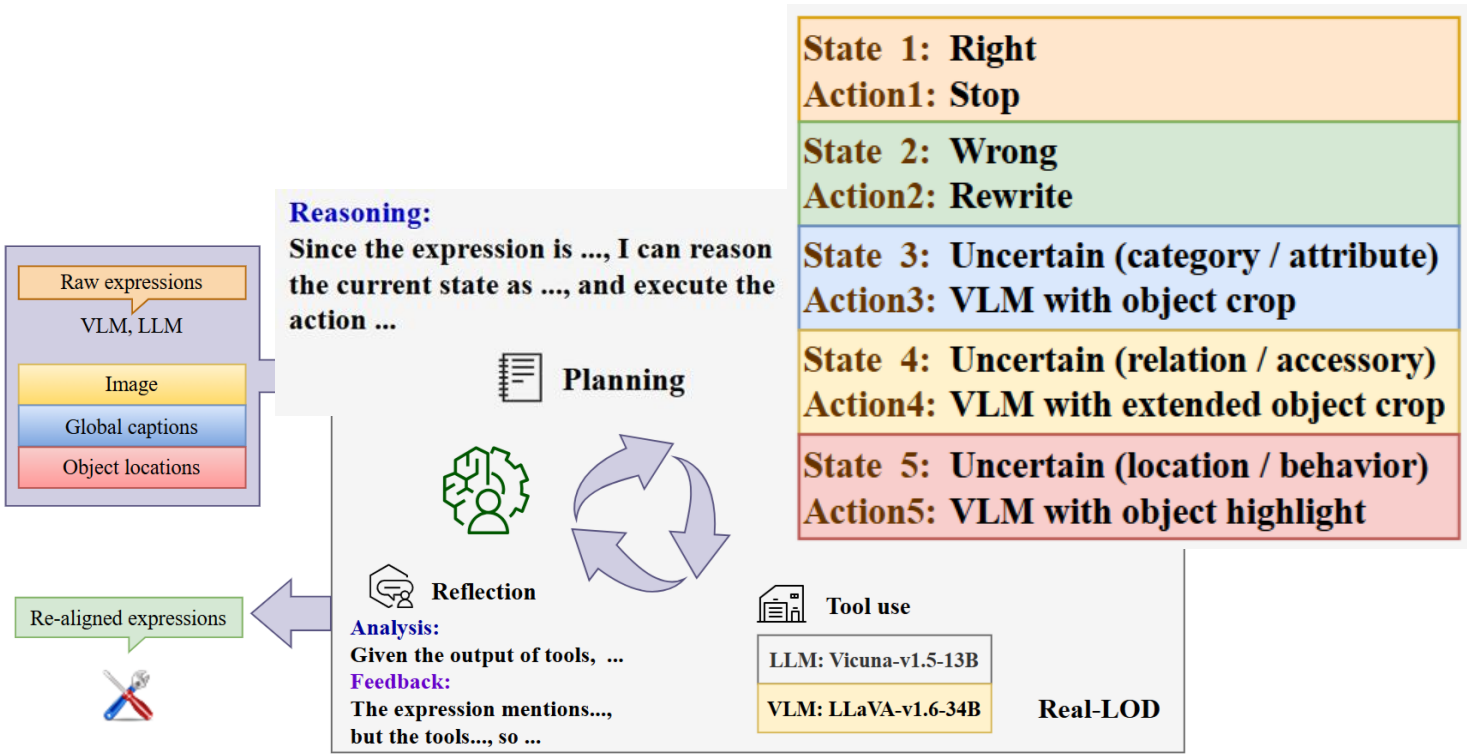
- Agentic workflows for language expression Re-Alignment
  - 생성한 object-expressions 쌍을 3단계를 거쳐 점진적으로 정제
    - Planning
      - 현재 expression이 target object랑 align이 잘 맞는지 state를 정의하고 각각에 맞는 action설정
    - Tool use
      - Agent가 plan에 맞게 VLM/LLM을 호출
    - Reflection
      - LLM기반의 Reflector: language expression이 object를 제대로 설명하는가? Feed back 생성



CVSP Lab
Computer Vision & Signal Processing

Method

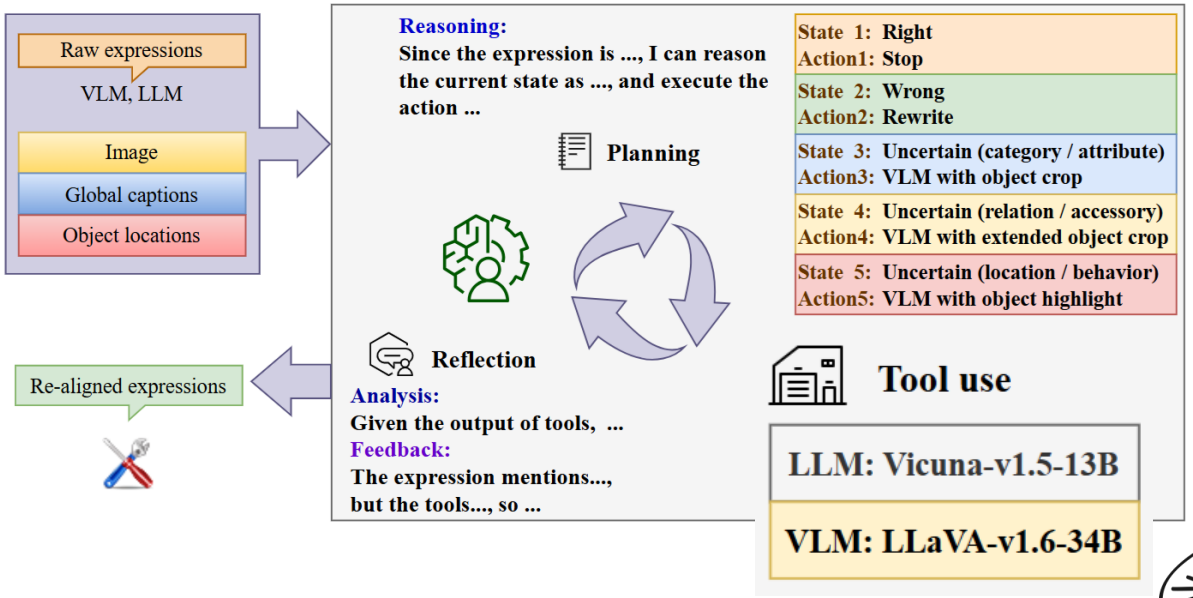
- Agentic workflows for language expression Re-Alignment
  - **Planning**: Identify the alignment state
    - 현재 Language expression이 target object와 얼마나 align되어 있는지 판단
    - Language expression-Object misalignment를 5가지 state로 정의해둠
      - Analysis of language and visual objects에서 관찰한 결과에 따라 state/action을 설정



CVSP Lab
Computer Vision & Signal Processing

Method

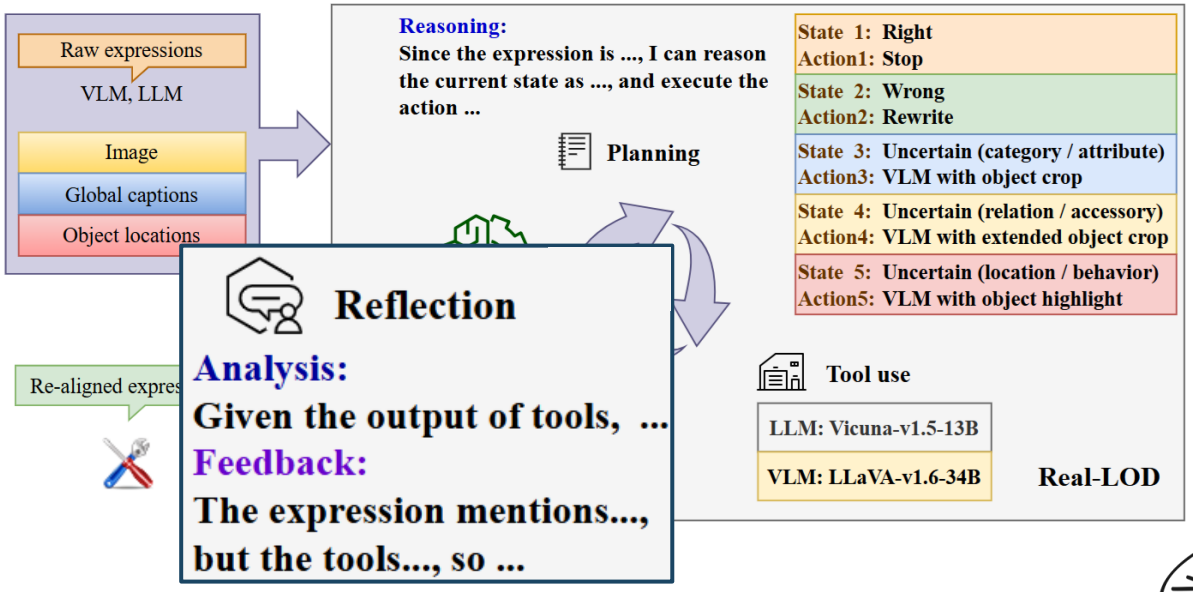
- Agentic workflows for language expression Re-Alignment
  - **Tool use** : Adaptive perception with VLM/LLM
    - Planning 결과에 따라서 Tool 사용
      - 여기서 tool은 이미 잘 학습된 VLM/LLM을 말함
    - Image editing 혹은 Prompt를 Agent가 직접 설계
      - Object crop: 객체 자체가 뭔 지 헷갈릴 때 target object 자체를 crop하여 VLM을 사용해 다시 인지
      - Extended object crop: 객체 주변과의 relation이 불확실해서 target object를 포함해 좀 더 넓은 region을 crop하여 VLM으로 재인지
      - Object highlight: 객체의 위치나 behavior에 대해 확실하지 않아서 red rectangle을 사용해 object를 highlight하여 VLM으로 재인지



CVSP Lab
Computer Vision & Signal Processing

Method

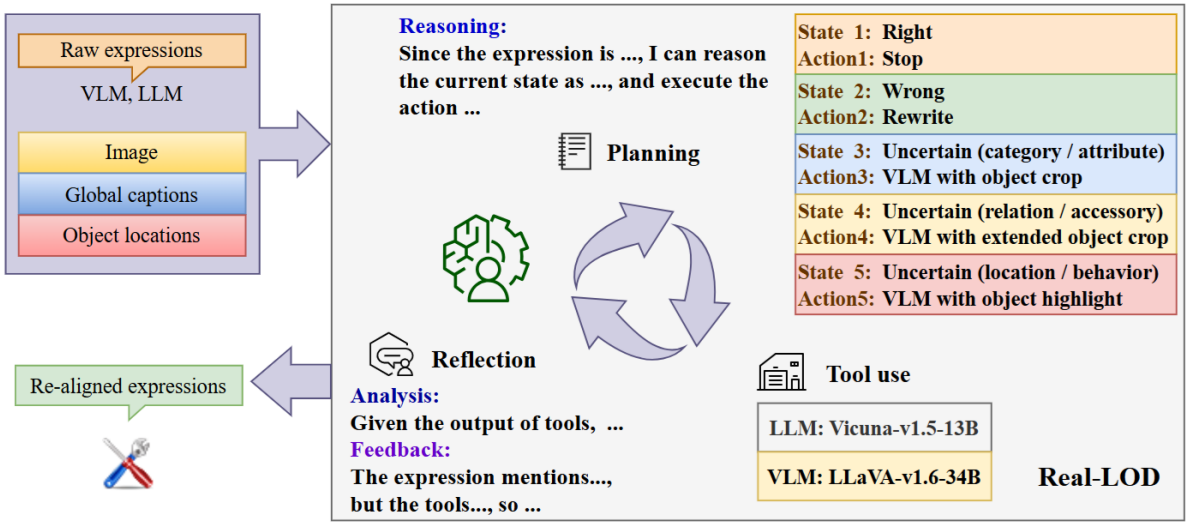
- Agentic workflows for language expression Re-Alignment
  - **Reflection**
    - Tool 사용 이후 Reflector(LLM)을 이용해 image caption을 함께 고려하면서 결과 분석
    - Reflector는 expression이 target object와 일치하는지 검증
    - 분석 결과는 다시 Agent에 전달되어 다음 planning을 지원



CVSP Lab
Computer Vision & Signal Processing

Method

- Agentic workflows for language expression Re-Alignment
  - Training data for Real-Agent
    - 현재 language expression이 어떤 state인지 판단하기 위한 Agent 학습
    - 학습 방식
      - ChatGLM-6B 모델 사용(text-only input)
      - LoRA기반 parameter-efficient fine-tuning
      - Object category/raw expression/reflector의 reasoning → state/action supervision



Experiments

- Quantitative Results
  - 실험 구성
    - Real-LOD가 Re-Alignment한 Data인 Language-Object data가 정말 Detection model의 성능을 끌어올리는가
    - Benchmark: OmniLabel benchmark
    - 모델 구조가 아니라 학습 데이터를 변경하여 적용
  - Task
    - Language-based Object Detection
    - 글이 주어졌을 때 그 문장이 가리키는 객체를 bbox로 찾는 task
      - 입력: 이미지 + 자연어 description
      - 출력: 해당 object의 bonding box

CVSP Lab
Computer Vision & Signal Processing

Experiments

- Quantitative Results

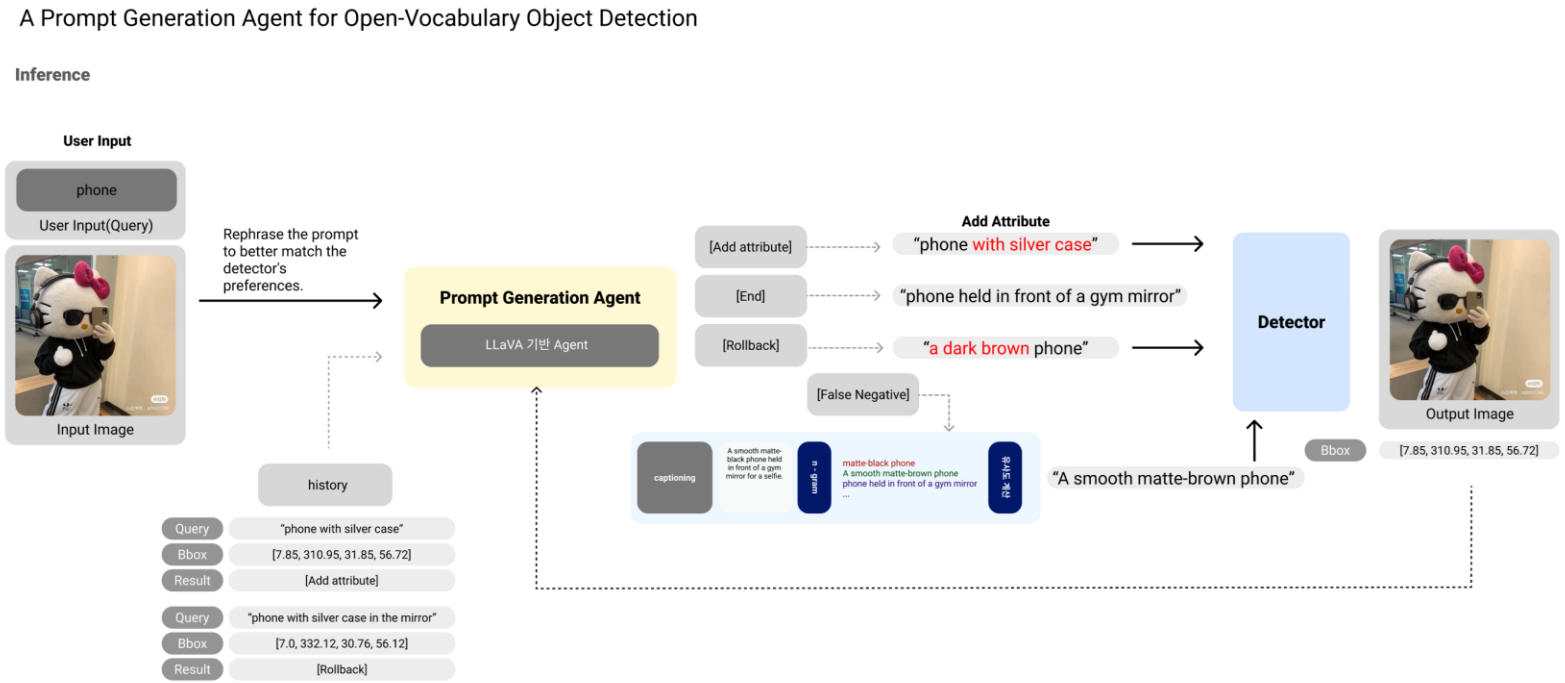
Table 1: State-of-the-art comparisons on the OmniLabel benchmark.

Subset	LOD method	Backbone	Source	#Img	AP-des	AP-des-pos	AP-des-S	AP-des-M	AP-des-L
COCO	MDETR (Kamath et al., 2021)	ENB3	COCO, VG, Flickr30K	0.3M	13.2	31.6	15.4	13.5	12.4
	GLIP (Li et al., 2022)	Swin-L	O365, OI, RefC/g/+, etc	17.5M	13.9	36.8	28.9	12.9	11.5
	mm-GDINO (Zhao et al., 2024)	Swin-B	GoldG, O365, COCO, etc	12M	15.2	47.0	29.3	14.9	15.1
	FIBER (Dou et al., 2022)	Swin-B	COCO, CC3M, SBU, etc	4M	14.3	38.8	31.3	12.7	16.1
	Real-Model	Swin-B	Real-Data	0.18M	**26.2**	**59.7**	**39.4**	**25.4**	**24.3**
O365	MDETR (Kamath et al., 2021)	ENB3	COCO, VG, Flickr30K	0.3M	3.2	5.9	3.0	3.2	2.7
	GLIP (Li et al., 2022)	Swin-L	O365, OI, RefC/g/+, etc	17.5M	24.0	35.2	44.5	20.5	11.8
	mm-GDINO (Zhao et al., 2024)	Swin-B	GoldG, O365, COCO, etc	12M	19.6	31.0	32.3	17.8	12.4
	FIBER (Dou et al., 2022)	Swin-B	COCO, CC3M, SBU, etc	4M	25.9	38.2	44.7	22.5	14.1
	Real-Model	Swin-B	Real-Data	0.18M	**36.0**	**52.1**	**55.7**	**32.3**	**23.7**
OI	MDETR (Kamath et al., 2021)	ENB3	COCO, VG, Flickr30K	0.3M	6.1	10.6	9.6	5.7	4.1
	GLIP (Li et al., 2022)	Swin-L	O365, OI, RefC/g/+, etc	17.5M	20.1	31.2	33.3	18.7	10.3
	mm-GDINO (Zhao et al., 2024)	Swin-B	GoldG, O365, COCO, etc	12M	23.2	34.5	32.3	23.8	16.9
	FIBER (Dou et al., 2022)	Swin-B	COCO, CC3M, SBU, etc	4M	20.1	30.9	34.1	18.5	10.5
	Real-Model	Swin-B	Real-Data	0.18M	**40.5**	**51.4**	**54.9**	**37.8**	**30.6**
ALL	MDETR (Kamath et al., 2021)	ENB3	COCO, VG, Flickr30K	0.3M	4.7	9.1	6.4	4.6	4.0
	GLIP (Li et al., 2022)	Swin-L	O365, OI, RefC/g/+, etc	17.5M	21.2	33.2	37.7	18.9	10.8
	mm-GDINO (Zhao et al., 2024)	Swin-B	GoldG, O365, COCO, etc	12M	20.8	33.1	31.9	19.8	14.1
	FIBER (Dou et al., 2022)	Swin-B	COCO, CC3M, SBU, etc	4M	22.3	34.8	38.6	19.5	12.4
	Real-Model	Swin-B	Real-Data	0.18M	**36.5**	**52.1**	**54.4**	**33.2**	**25.5**

CVSP Lab
Computer Vision & Signal Processing

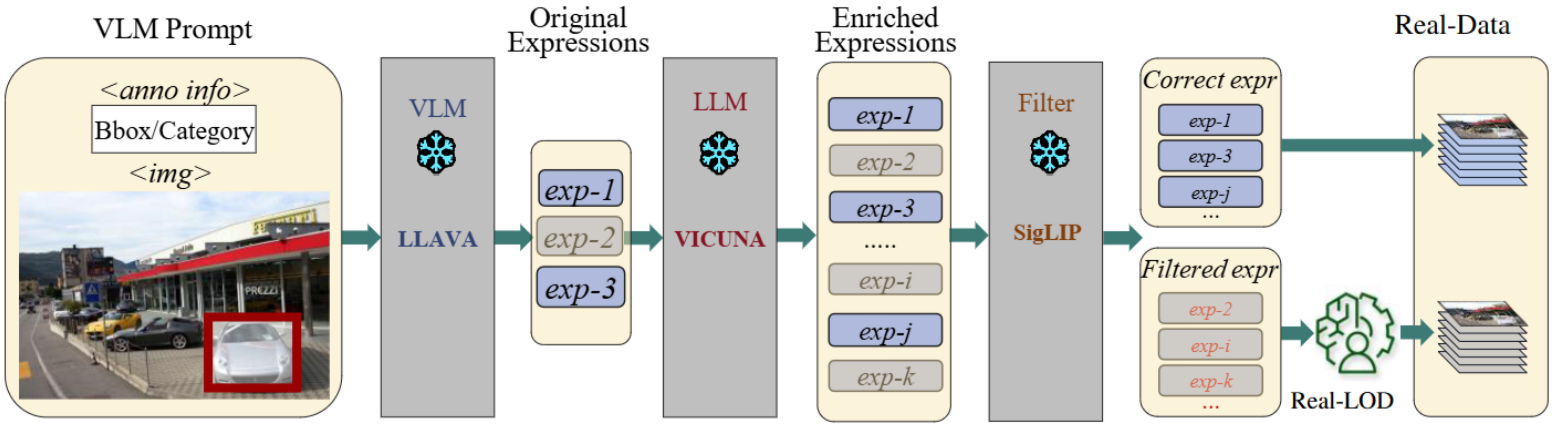
Conclusion

- OVOD Query Refinement를 위한 Agentic flow
  - LLM Agentic flow 기반 Text Query Refinement
    - 자유롭게 Query 변경이 가능한 OVOD의 특성을 활용
    - 탐지 결과를 피드백으로 활용하여 LLM이 text query를 점진적으로 Refinement
    - Query 변경 → 성능 변화 측정 → 재수정의 루프 자동화



Appendix

- Expressions generation pipeline
  - VLM(LLaVA)를 통해 descriptions를 생성
    - 각 object별로 준비된 프롬프트 중 두 개를 골라 입력
  - LLM(Vicuna)를 통해 동의어로 확장
    - 각 표현에 대해 두 번 반복
  - Paired data의 matching 점수를 확인
    - SigLIP을 통해 expression-image pairs의 score를 계산



< Real-LOD: Expressions generation pipeline >

CVSP Lab
Computer Vision & Signal Processing

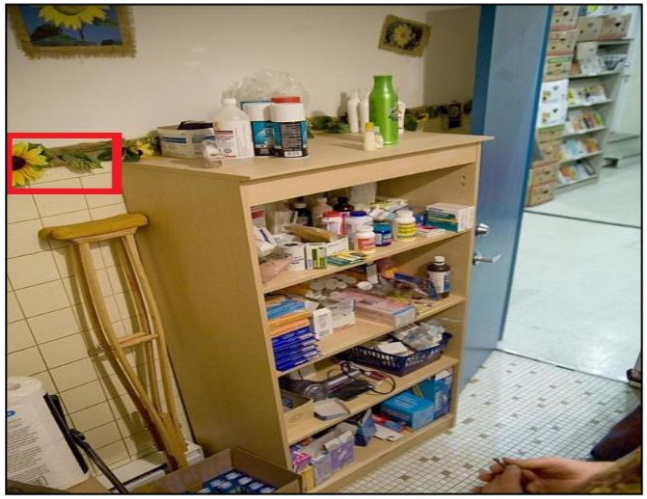
Appendix

- Re-Alignment examples of Real-LOD



Is '***man wearing a black hat***' correctly referring to the ***helmet*** in image? If not, please respond with a correct one.

No, the provided expression has errors in it. A correct one can be '***dark green helmet for motorcycle rider***'.



Is the expression '***yellow flower in a vase***' correctly describing the ***flower*** in image? If not, please provide with a correct one.

The expression has flaws in it. Here is a correct one: '***decorative paper sunflower on a wall***'.

Appendix

- Re-Alignment examples of Real-LOD



Is '***a woman wearing a white jacket***' correctly describing the ***person*** in image? If not, please respond with a correct one.

No, the provided expression is a wrong one. A correct one can be '***person standing near a blue tent wearing a green shirt and white pants***'.



Decide if the expression '***a man in a black shirt and blue jeans***' is correctly describing the ***person*** in image. If not, please provide with a correct one.

Yes. The provided expression is correct.