2026 하계세미나

Discriminative Target Specification for Visual Counting

2026. 07. 03.

Presented By
장순원

Outline

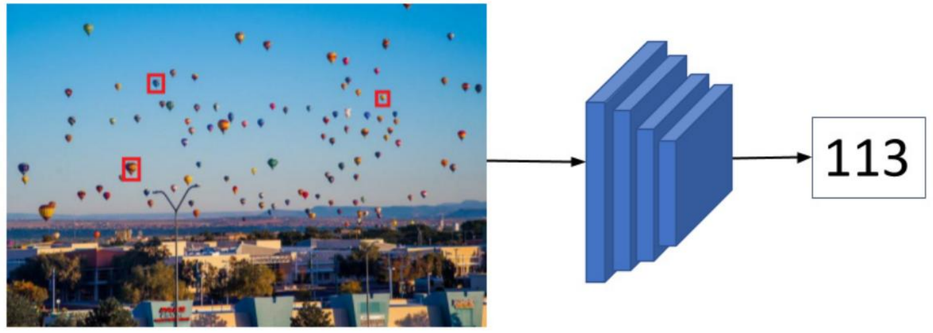
- Object Counting: Visual exemplar to Referring expression

- Decoupling What to Count and Where to See for Referring Expression Counting (W2-Net)
  - AAAI 2026

- CountGD++: Generalized Prompting for Open-World Counting
  - CVPR 2026

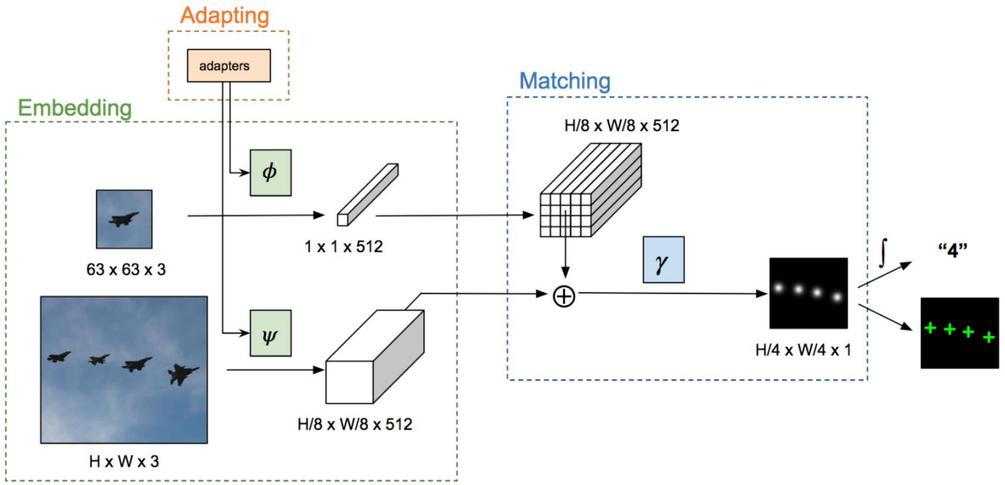
Object Counting: Visual exemplar to Referring expression

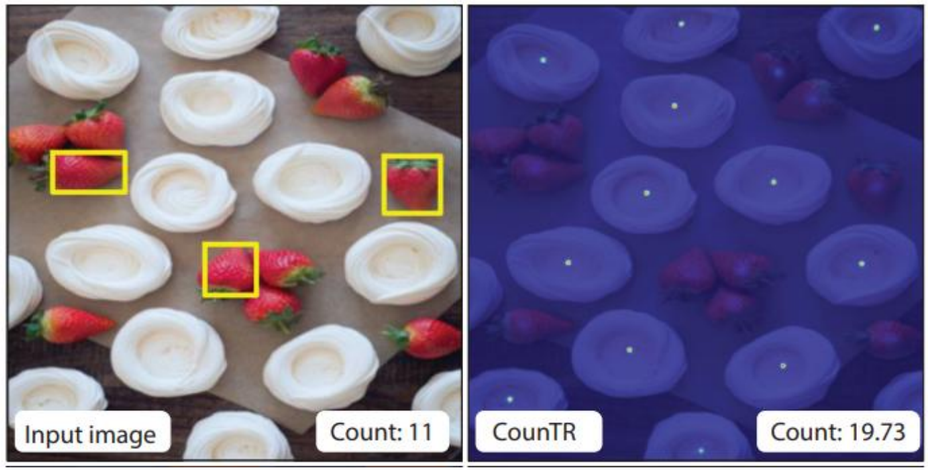
- Object Counting?
  - 이미지 내 target object의 위치를 대략적으로 식별하고, 전체 개수를 추정하는 task
    - Dense scene에 대해 중복 및 누락 없이 객체 수를 추정하는 것이 핵심
      - 정답 count와의 RMSE와 MAE가 주요 평가 지표로 활용됨
      - Precision / Recall / F1-score는 정답 point annotation과 일정 거리 내의 여부로 측정
    - 객체의 정확한 bounding box 예측을 평가하는 detection과 초점에 차이가 존재
  - 적용 사례
    - 상품 계수 및 불량품 탐지, 생산 부품 제고 관리, 불량률 모니터링 및 품질 향상
    - 이미지 생성 모델의 피드백 생성
  - Object detection과의 차이점
    - 평가 metric – 정확한 GT에서 벗어나더라도 계수에 집중
    - 객체의 해상도, Occlusion, 한 이미지 내에 존재하는 객체 수



Object Counting: Visual exemplar to Referring expression

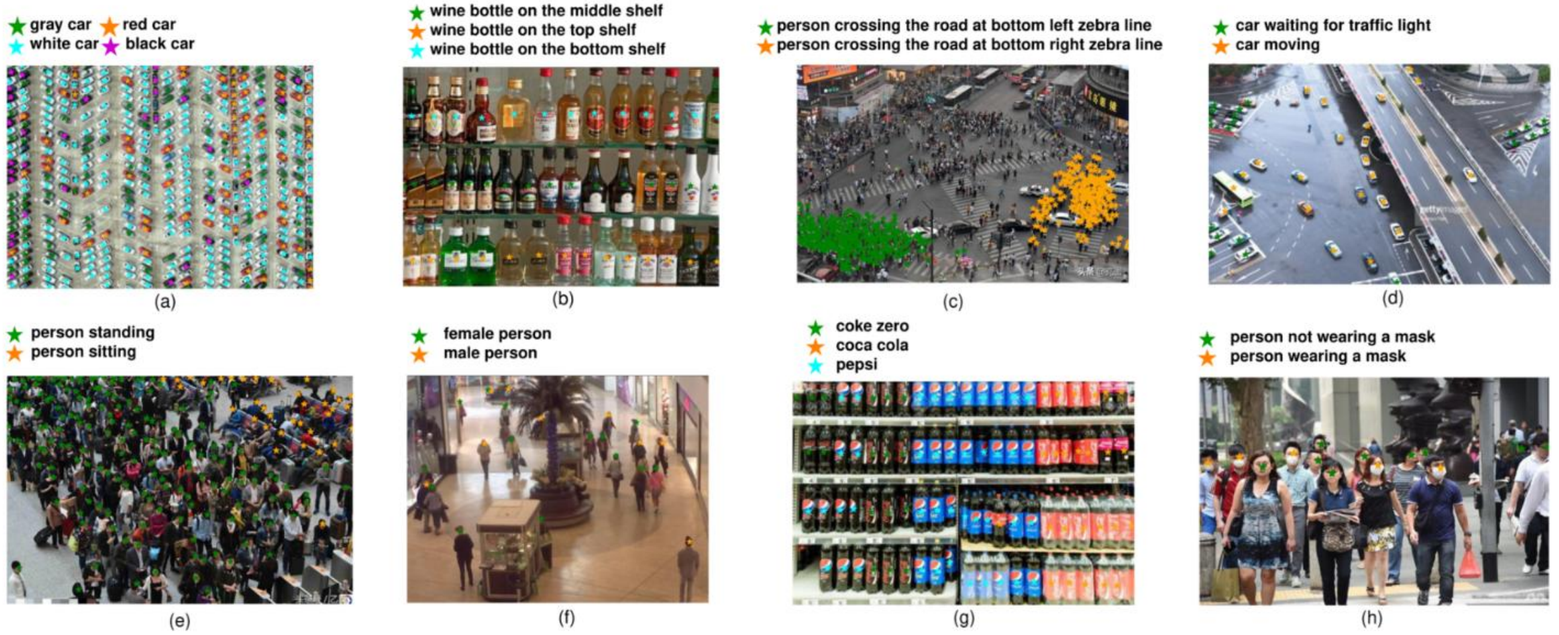
- Visual exemplar: few-shot counting
  - 학습에 등장하지 않은 open-set에서의 동작을 위해 exemplar image 정보를 통해 counting 수행
    - Counting을 원하는 객체에 대한 소수의 annotation 작업 필요
  - 예시 영역에 대한 feature를 생성하여 counting에 사용
  - Text prompt에 비해 정확도 면에서 유리하나, 추론을 위해서는 test image마다의 annotation이 필요





Object Counting: Visual exemplar to Referring expression

- Text instruction: zero-shot counting / referring expression counting
  - Visual exemplar 없이 VLM을 통해 class name 만으로 counting 수행
  - Referring Expression Counting
    - 같은 class 내에서도 expression에 따라 subclass를 구분하여 count
    - 더 높은 수준의 시각-언어 정렬 및 정밀한 인식 필요
    - Counting datasets에 annotation을 추가한 REC-8K 데이터셋을 통해 학습 및 평가

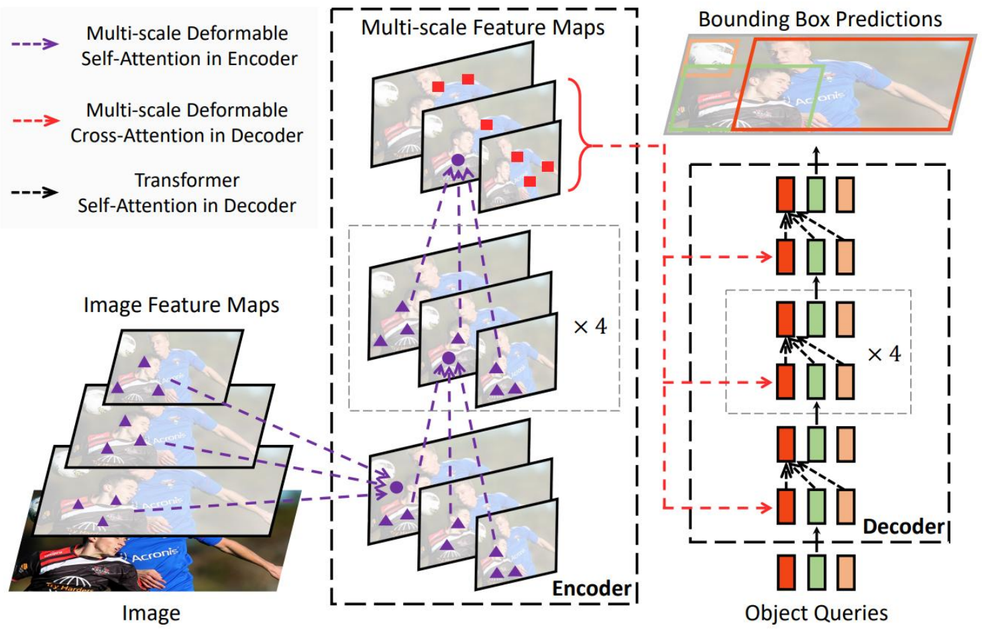


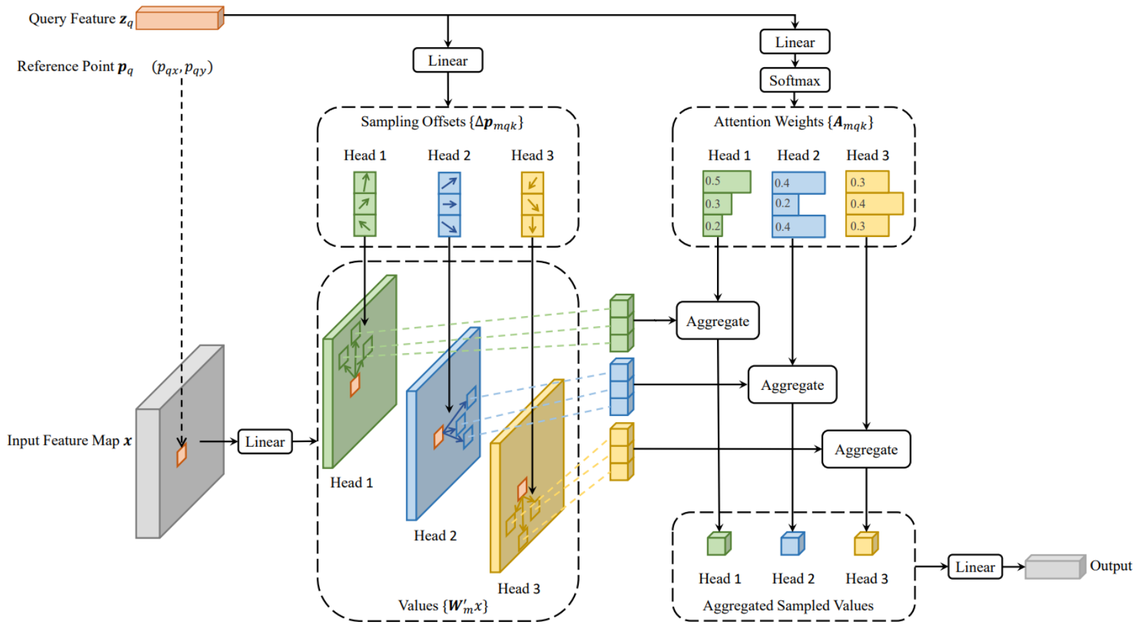
- **Decoupling What to Count and Where to See for Referring Expression Counting [AAAI 2026]**

- **CountGD++: Generalized Prompting for Open-World Counting [CVPR 2026]**

W2-Net

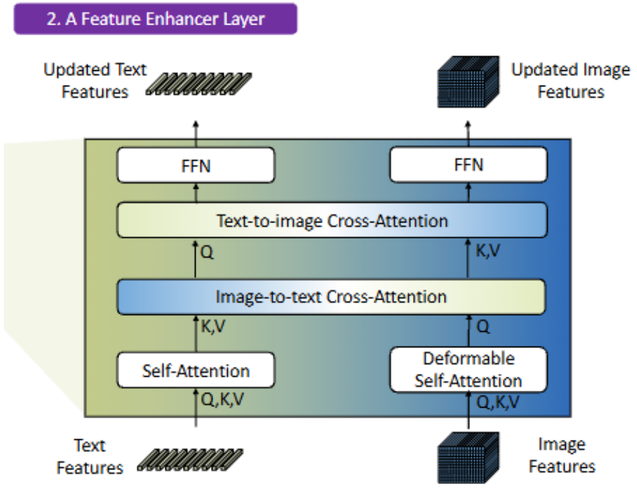
- Background – GroundingDINO
  - DINO (DETR with Improved deNoising anchor boxes) 기반
  - Deformable attention
    - 기준점 주변의 소수 sampling point만 선택하여 해당 위치의 feature만 attention에 사용

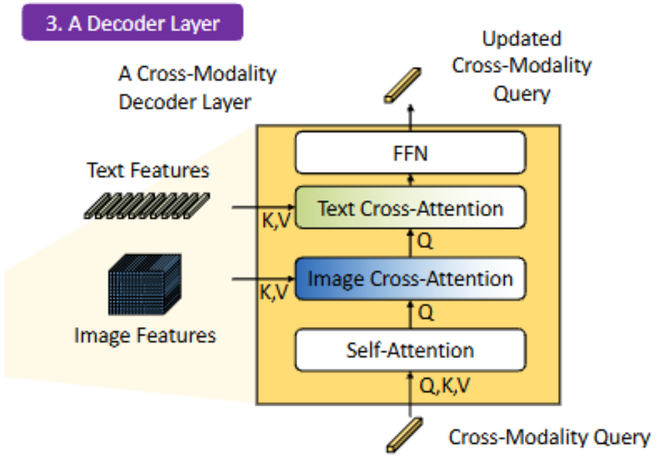




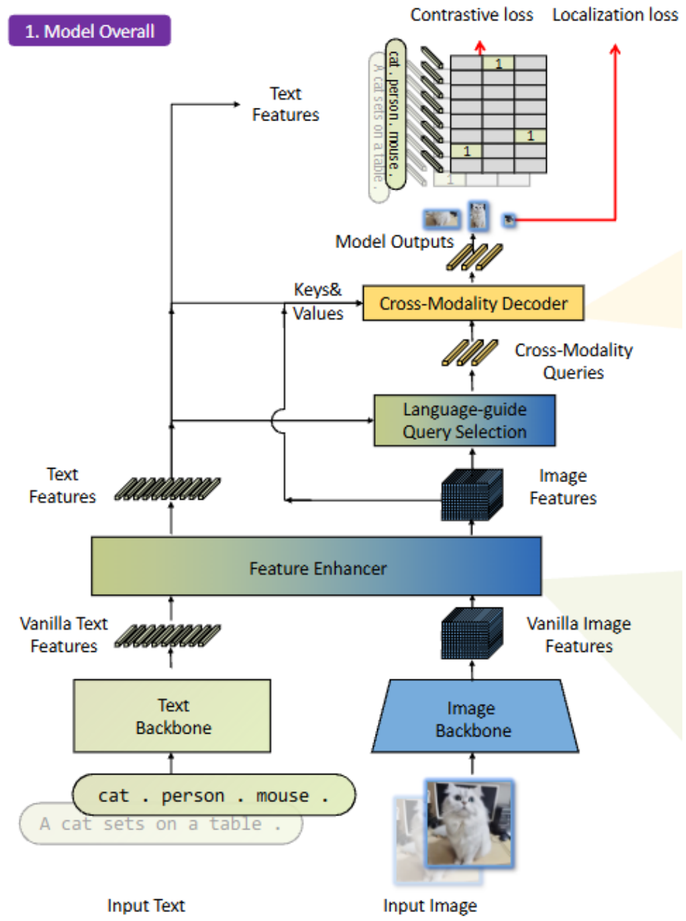
W2-Net

- Background – GroundingDINO
  - IDEA-Research의 End-to-End Object detection model
    - Swin-T image backbone을 통해 multi-scale feature 추출
    - 각각을 cross-attention을 통한 feature enhancement 후 decoder 통과
    - 900개 query를 선정하여 최종적으로 총 900개의 후보 bounding box 생성
      - Text feature와의 similarity score에서 thresholding을 통해 최종 후보 선택



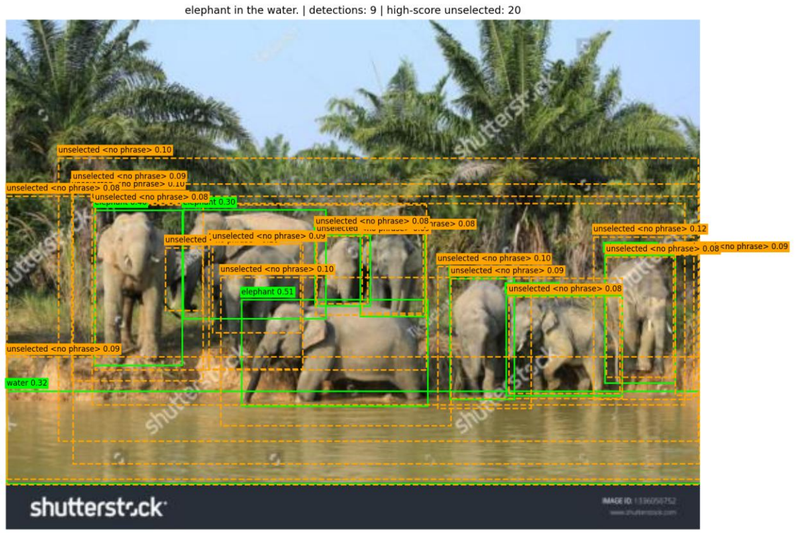


reference + feature



W2-Net

- Background – GroundingDINO
  - Hungarian matching
    - Query-based detection/counting model은 고정된 개수의 prediction query를 출력
    - 하지만 GT object 개수는 이미지마다 다름
    - 따라서 어떤 prediction을 어떤 GT object에 대응시킬지 결정해야 함
    - Prediction–GT pair의 cost matrix를 만들고, 전체 cost가 최소가 되는 1:1 matching을 선택하는 알고리즘

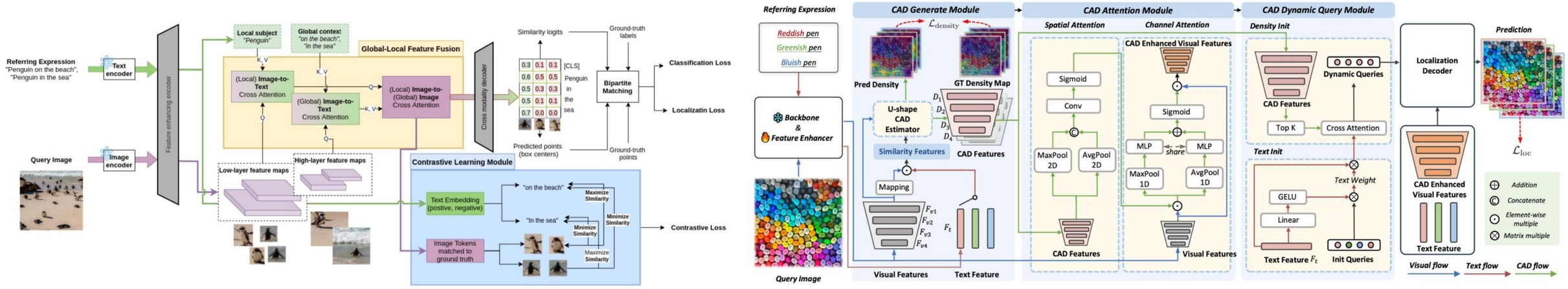


$$C(i,j) = \lambda_{cls} C_{cls}(i,j) + \lambda_{loc} C_{loc}(i,j) + \lambda_{giou} C_{giou}(i,j)$$

	j_1	j_2	j_3	
i_1	**0**	**5**	**6**	**3**
i_2	4	12	7	
i_3	4	8	5	

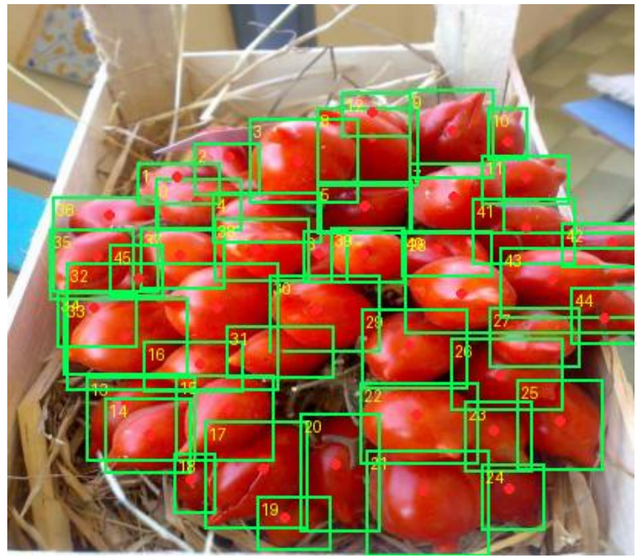
W2-Net

- Background – GroundingDINO
  - Counting에서의 활용
    - GroundingDINO 기반의 모델이 가장 높은 성능을 보임
      - Point annotation에 맞춰 bounding box 출력을 단일 point 출력으로 변형
    - Feature enhancement 이후 counting을 위한 추가 모듈을 통과 시킨 후 decoder로 넘기는 구조가 자주 채택됨
      - Local/Global feature fusion
      - Attribute density attention module



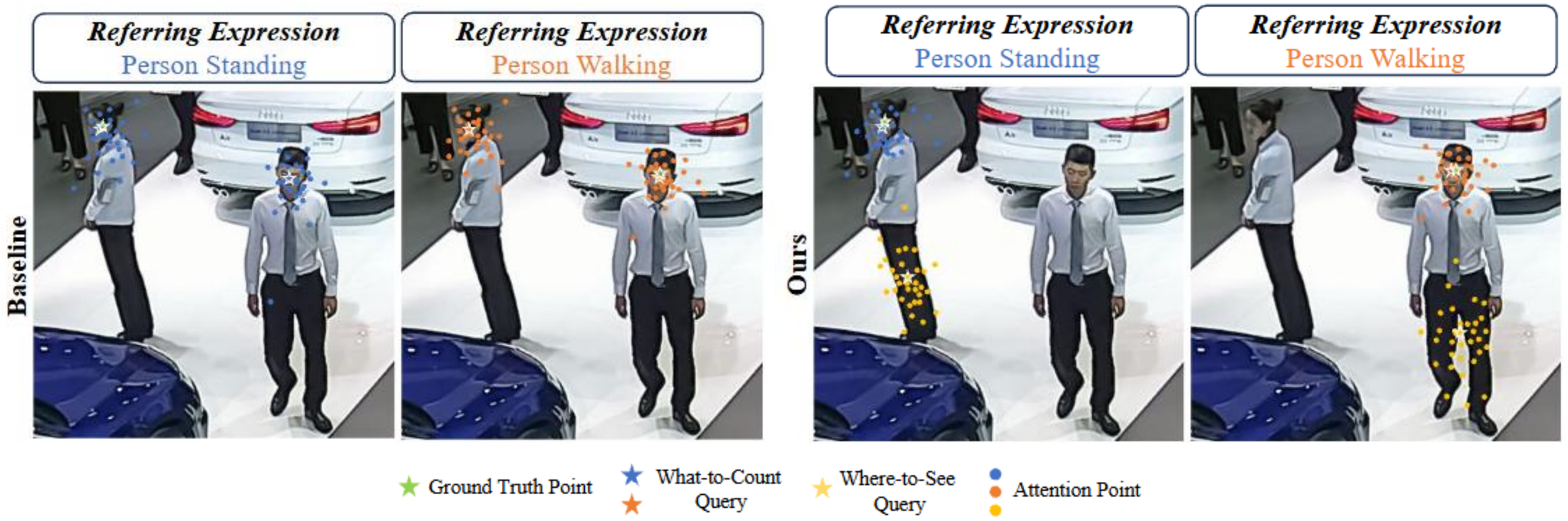
W2-Net

- REC-8K dataset 문제점
  - 정답 객체에 대해 단일 point annotation 제공
    - Referring expression에 대한 판단을 하기에는 적절하지 않은 위치일 수 있음
      - Ex) person의 head 영역을 보고 standing/walking 여부를 판단하기 어려움
    - Lack of attribute guidance / Over-emphasize class-level feature



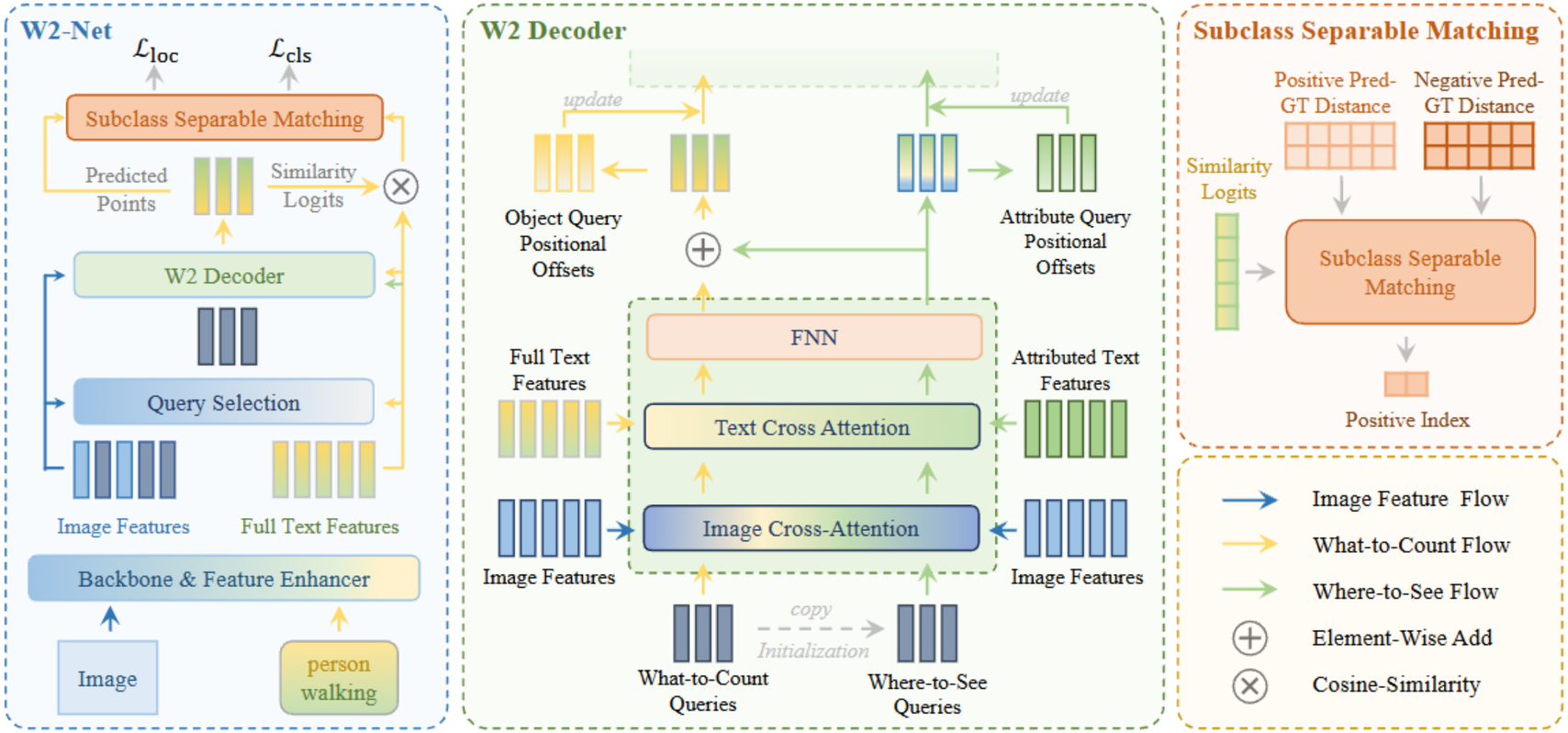
W2-Net

- Introduction
  - Decouple problem into “what to count” and “where to see”
  - Subclass Separable Matching을 통한 class내 separability 강화
  - 유사한 feature를 가지는 같은 class 내에서 referring expression에 따라 subclass를 나누려면?



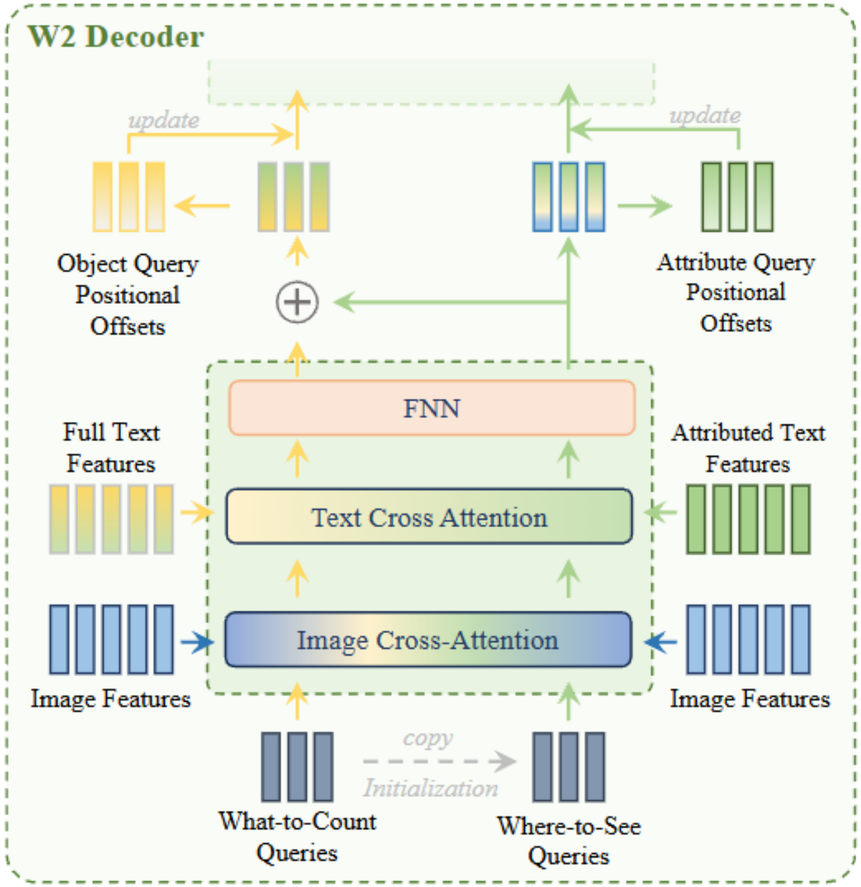
W2-Net

- Framework
  - GroundingDINO의 decoder를 새로 디자인
    - Box detection to point-based localization / W2 decoder



W2-Net

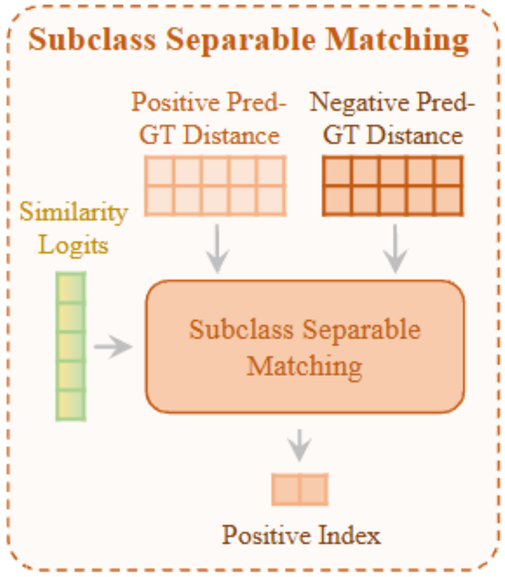
- Method
  - Query initialization
    - What-to-count(w2c)와 what-to-see(w2s) query를 구분하여 사용
      - w2s는 900개의 w2c가 복제되어 초기화
  - Parallel query refinement
    - 두 가지 queries를 동시에 처리하면서, 서로 다른 text feature를 적용
      - w2c: full text feature (red pen, walking person)
      - w2s: attributed text feature (red pen → red, walking person → walking)
    - w2s는 attribute-related word에 focus하도록 만들기 위함
  - Fusion and iterative update
    - Element-wise addition을 통해 w2s의 정보를 w2c에 injection
    - 각각 localization heads를 통해 reference point에 대한 positional offset 계산



$$Q_{w2c}^{txt} = \text{CrossAttn}(Q_{w2c}, F_t, F_t), \quad Q_{w2s}^{txt} = \text{CrossAttn}(Q_{w2s}, F_t \odot M_{w2s}, F_t \odot M_{w2s}),$$

W2-Net

- Method
  - Subclass Separable Matching
    - 잘못된 1대1 Hungarian matching이 발생하여 학습 신호가 불안정해지는 문제 완화
      - Subclass들이 시각적으로 매우 유사하기 때문에 feature 기반 matching 시 잘못된 매칭 발생
    - Repulsive term을 추가하여 prediction이 잘못된 subclass의 ground truth와 매칭되는 것에 penalty 부여



Classification cost

Repulsive term

$$\mathcal{C}_{match}(i,j) = \underline{\lambda_{cls}(1-\hat{s}_i)} + \underline{\lambda_{L1}\|\hat{p}_i - p_j\|_1} + \lambda_{rep}\mathcal{C}_{rep}(\hat{p}_i, p_j)$$

Localization cost

Prediction index

GT index

$$\mathcal{C}_{rep}(\hat{p}_i, p_j) = \exp(-\mathcal{R}(\hat{p}_i, p_j))$$

$$\mathcal{R}(\hat{p}_i, p_j) = \frac{\min_{p_k \in Y_{neg}} \|\hat{p}_i - p_k\|_2}{\|\hat{p}_i - p_j\|_2}$$

Prediction이 negative GT에 얼마나 가까운지 보는 비율

Low cost

High cost

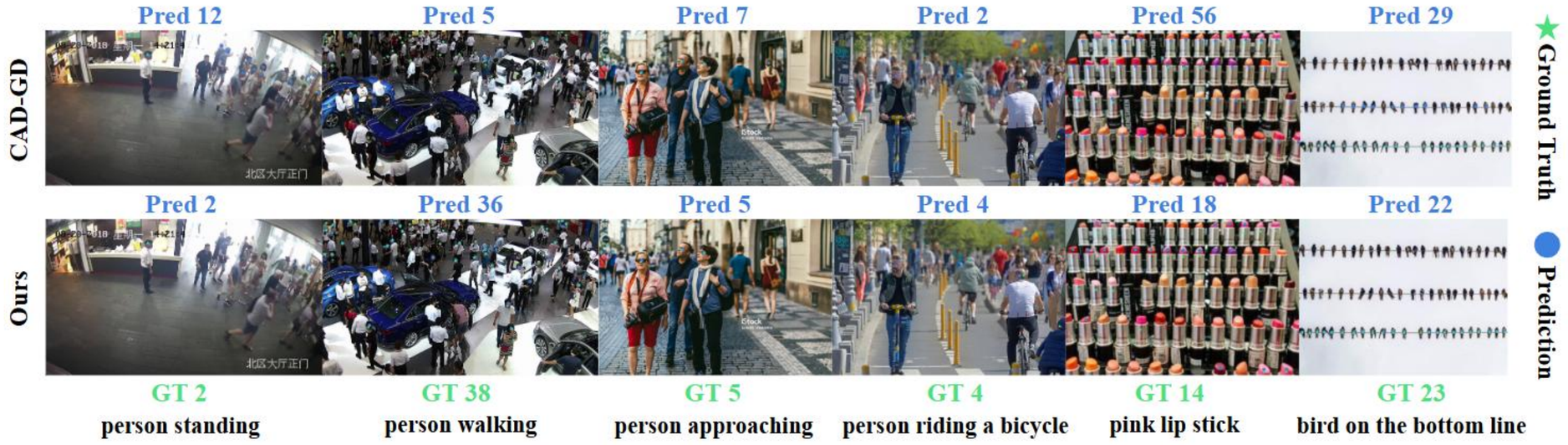
Expression: walking person

W2-Net

- Experiments
  - Referring expression counting 및 text기반 zero-shot counting SOTA

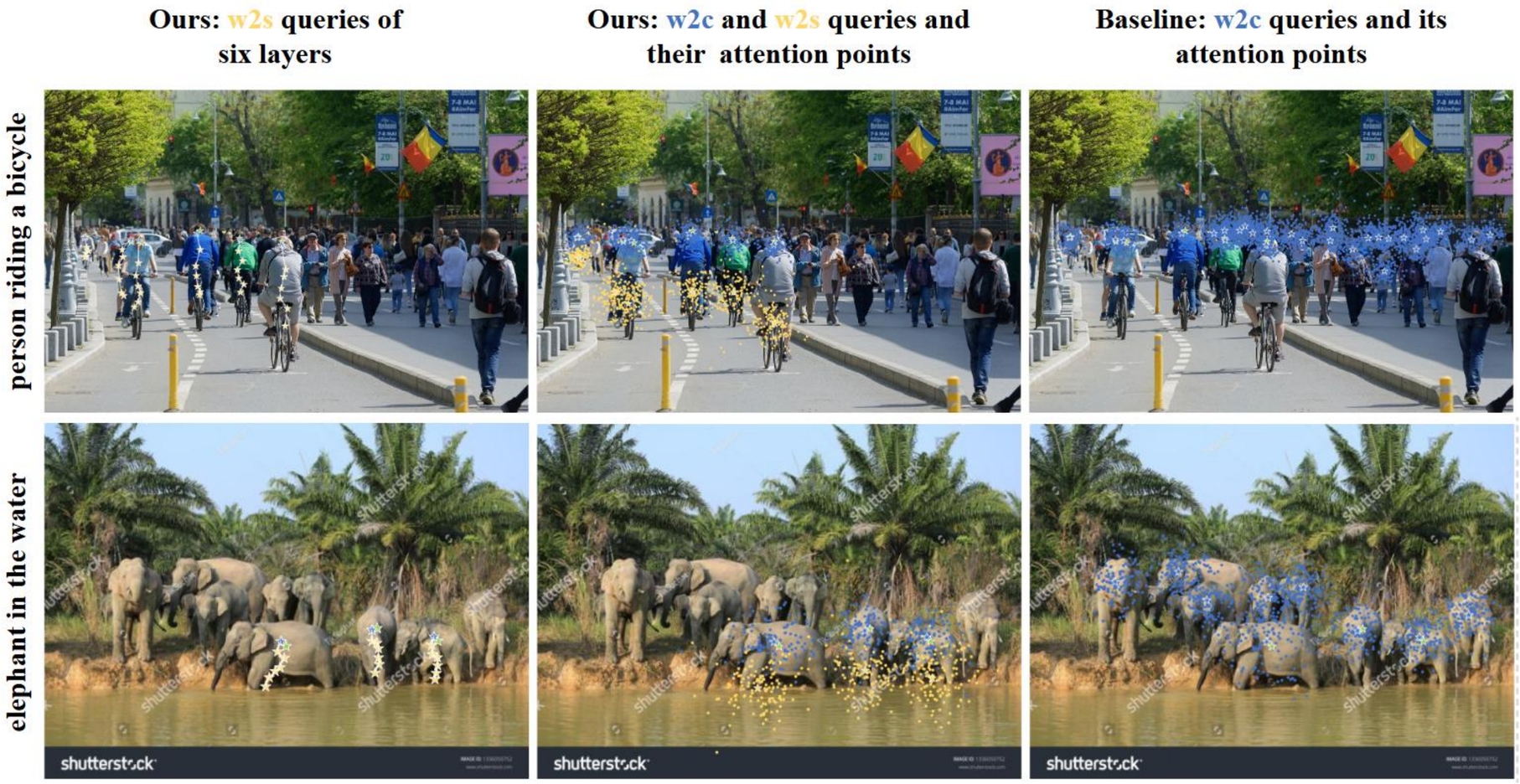
Method	Venue	Backbone	Test Set				
			MAE ↓	RMSE ↓	Prec ↑	Rec ↑	F1↑
ZSC (Xu et al. 2023)	CVPR23	Swin-T	13.00	29.07	-	-	-
CounTX (Amini-Naieni et al. 2023)	BMVC23	ViT-B	11.84	25.62	-	-	-
GroundingDINO (Liu et al. 2024)	ECCV24	Swin-T	8.88	21.95	0.59	0.76	0.66
GroundingREC (Dai, Liu, and Cheung 2024)	CVPR24	Swin-T	6.50	19.79	0.67	0.72	0.69
CAD-GD (Wang et al. 2025b)	CVPR25	Swin-T	5.29	17.08	0.71	0.73	0.72
CAD-GD† (Wang et al. 2025b)	CVPR25	Swin-T	4.59	14.68	0.72	0.70	0.71
W2-Net (Ours)	-	Swin-T	**3.59**	**9.59**	**0.77**	**0.80**	**0.79**
GroundingREC (Dai, Liu, and Cheung 2024)	CVPR24	Swin-B	5.42	18.47	0.71	0.69	0.70
CAD-GD (Wang et al. 2025b)	CVPR25	Swin-B	4.94	14.65	0.75	0.77	0.76
CAD-GD† (Wang et al. 2025b)	CVPR25	Swin-B	4.34	12.93	0.77	0.71	0.74
W2-Net (Ours)	-	Swin-B	**3.56**	**9.15**	**0.79**	**0.81**	**0.80**

Method	Venue	Prompt	Val Set		Test Set	
			MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
Patch-Selection (Xu et al. 2023)	CVPR23	Text	26.93	88.63	22.09	115.17
CLIP-Count (Jiang, Liu, and Chen 2023)	ACMMM23	Text	18.79	61.18	17.78	106.62
VLCounter (Kang et al. 2024)	AAAI24	Text	18.06	65.13	17.05	106.16
CounTX (Amini-Naieni et al. 2023)	BMVC23	Text	17.10	65.61	15.88	106.29
DAVE (Pelhan et al. 2024)	CVPR24	Text	15.48	52.57	14.90	103.42
GroundingREC (Dai, Liu, and Cheung 2024)	CVPR24	Text	10.06	58.62	10.12	107.19
CountGD (Amini-Naieni et al. 2024)	NIPS24	Text	12.14	47.51	12.98	98.35
CAD-GD (Wang et al. 2025b)	CVPR25	Text	9.30	40.96	10.35	86.88
W2-Net (Ours)	-	Text	**8.73**	**37.32**	**9.53**	**83.46**



W2-Net

- Experiments
  - w2s에 대해 별다른 학습 신호가 없음에도 판단에 필요한 영역에 attention

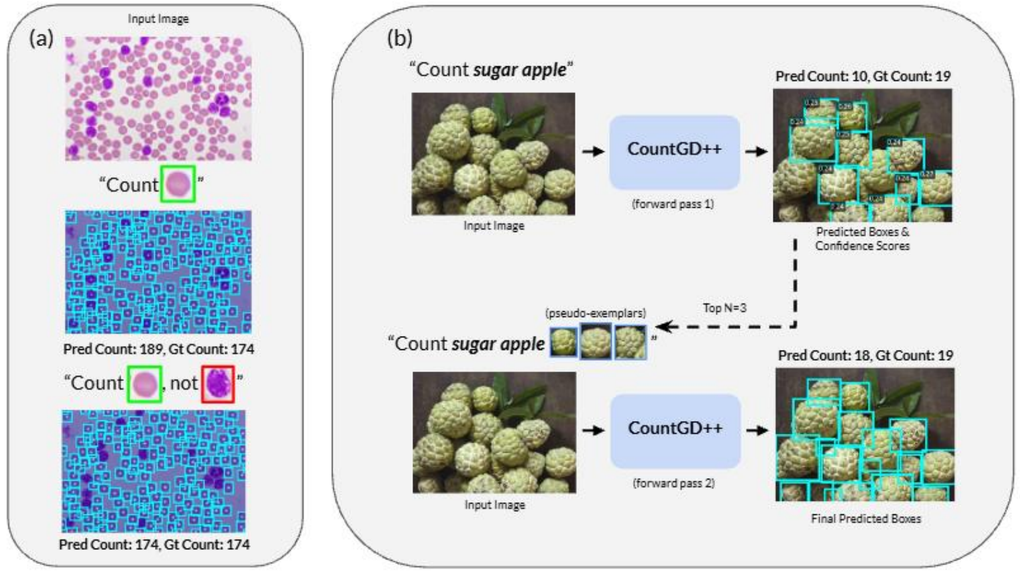


- Decoupling What to Count and Where to See for Referring Expression Counting [AAAI 2026]

- **CountGD++: Generalized Prompting for Open-World Counting [CVPR 2026]**

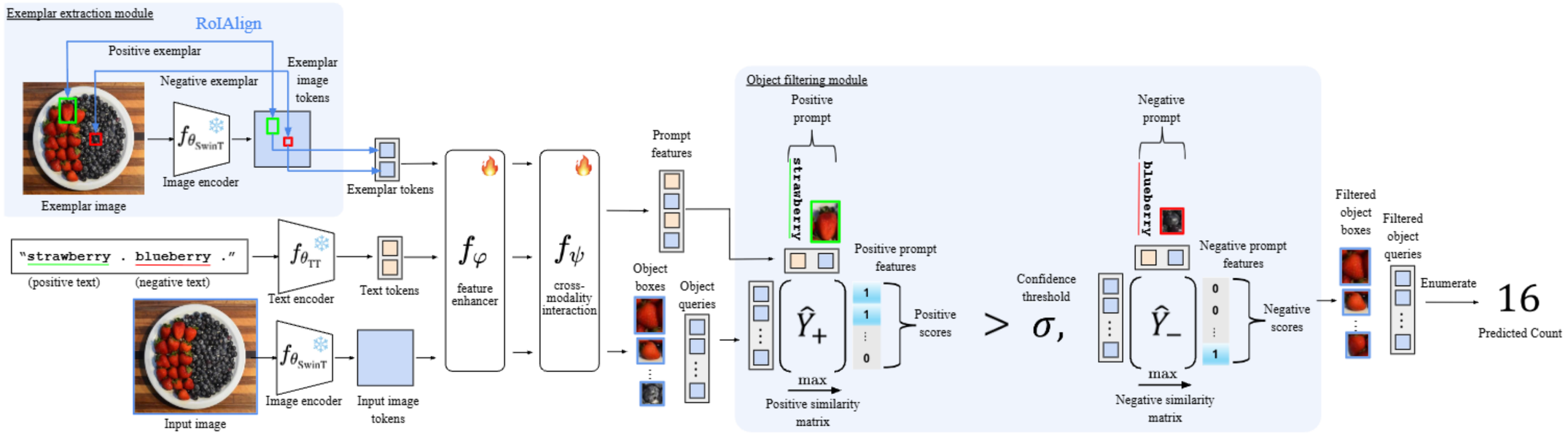
CountGD++

- Intro: Generalized Prompting for Open-World Counting
  - COUNTGD: Multi-Modal Open-World Counting의 후속 논문
  - Prompt에 대한 모호성, 실제 적용 시의 한계 존재
    - 어떤 text 혹은 visual prompt가 가장 적절한가?
    - Exemplar를 주기 위해 이미지마다 예시 객체에 bounding box를 annotation 하기 위한 overhead 존재
  - 1. Negative prompt
    - Count하면 안되는 객체에 대해서도 negative prompt 부여
  - 2. Pseudo exemplar scheme
    - Text기반으로 찾은 visual exemplar를 활용하는 2-stage 구조
      - “sugar apple”처럼 이름만으로는 visual concept이 명확하지 않은 경우
    - Synthetic exemplar image
      - 이미지 생성을 통해 만들어진 예시 이미지 활용



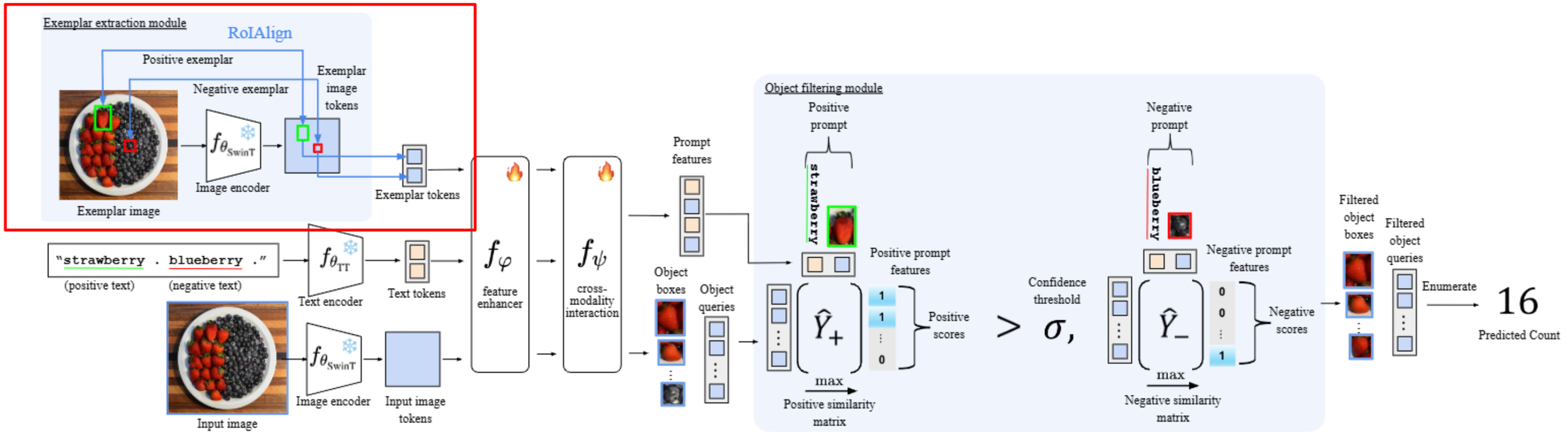
CountGD++

- Framework
  - GroundingDINO에서 feature enhancer와 decoder fine-tuning
  - Exemplar tokens 추가
  - Negative prompt를 고려한 object filtering module



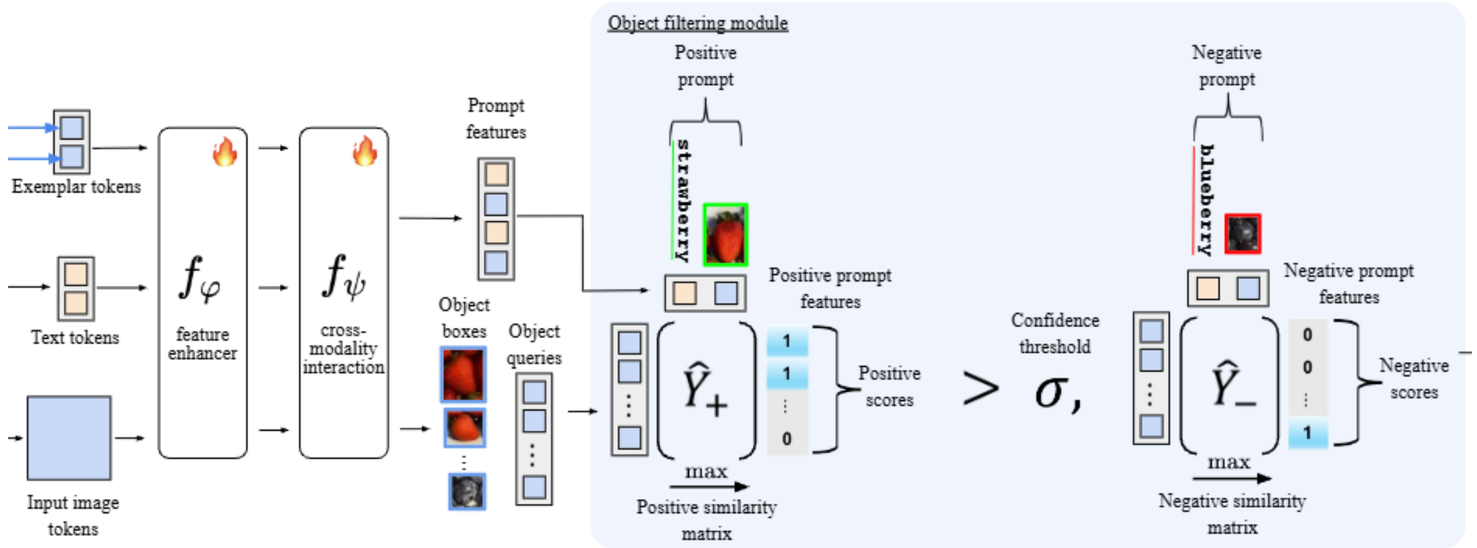
CountGD++

- Method
  - Exemplar extraction module
    - Exemplar image: Target 객체가 bbox 형태로 표시된 이미지
    - Swin-T를 통해 추출된 feature map에서 bbox에 해당하는 영역 feature를 pooling하여 단일 exemplar token(d=256) 생성
    - 이전 방식들은 이 exemplar image를 만들기 위한 cost가 존재함



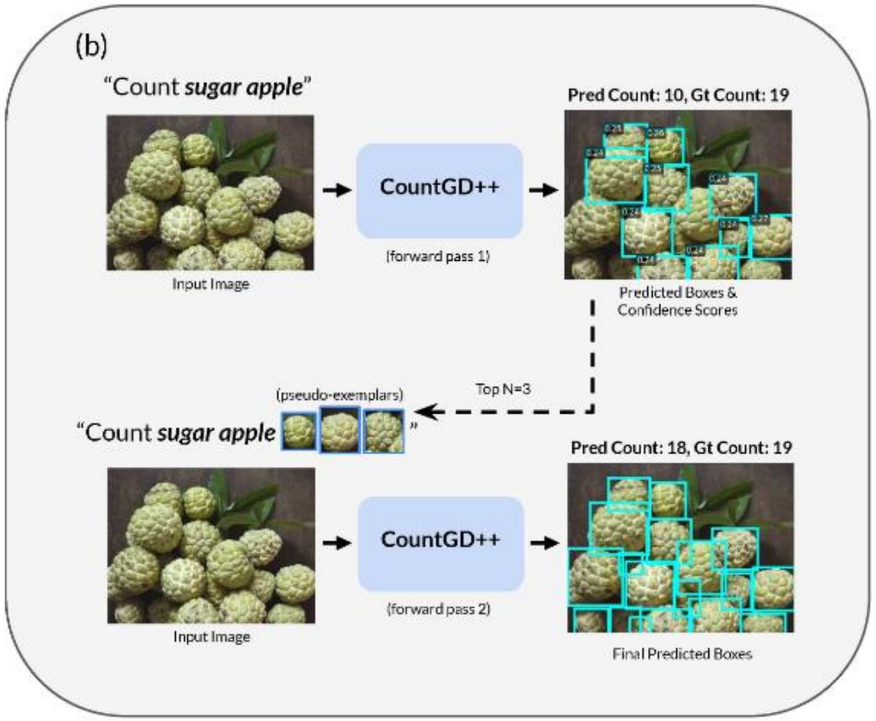
CountGD++

- Method
  - Negative prompt
    - Query 900개를 추론하면서 prompt feature를 동시에 정제
      - 기존 GroundingDINO 구조에서 text feature에 해당하는 부분을 확장
    - Positive prompt와의 similarity가 threshold를 넘는지 판단 & Negative prompt와의 similarity보다 높은 지 판단
    - Exemplar와 text token은 서로 같은 class 인 경우만 self-attention을 통한 정보 교환 수행



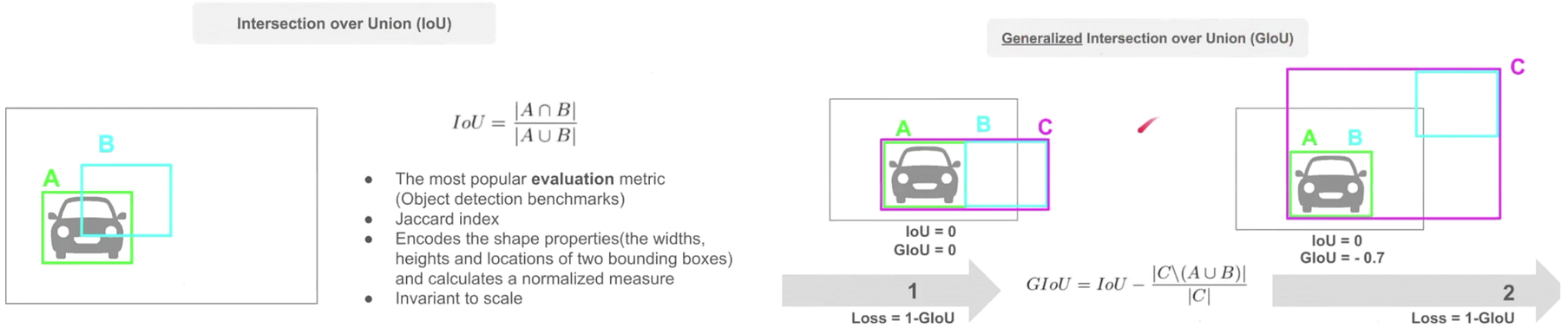
CountGD++

- Method
  - Pseudo Exemplar (automatically detecting visual examples)
    - Text prompt만으로 1차 추론을 수행한 뒤, confidence가 높은 Top-N prediction box를 자동 선택
    - 선택된 box를 visual exemplar처럼 다시 입력하여 2차 추론 수행
    - 별도의 수동 annotation 없이 visual cue를 추가해 text-only counting 성능 개선



CountGD++

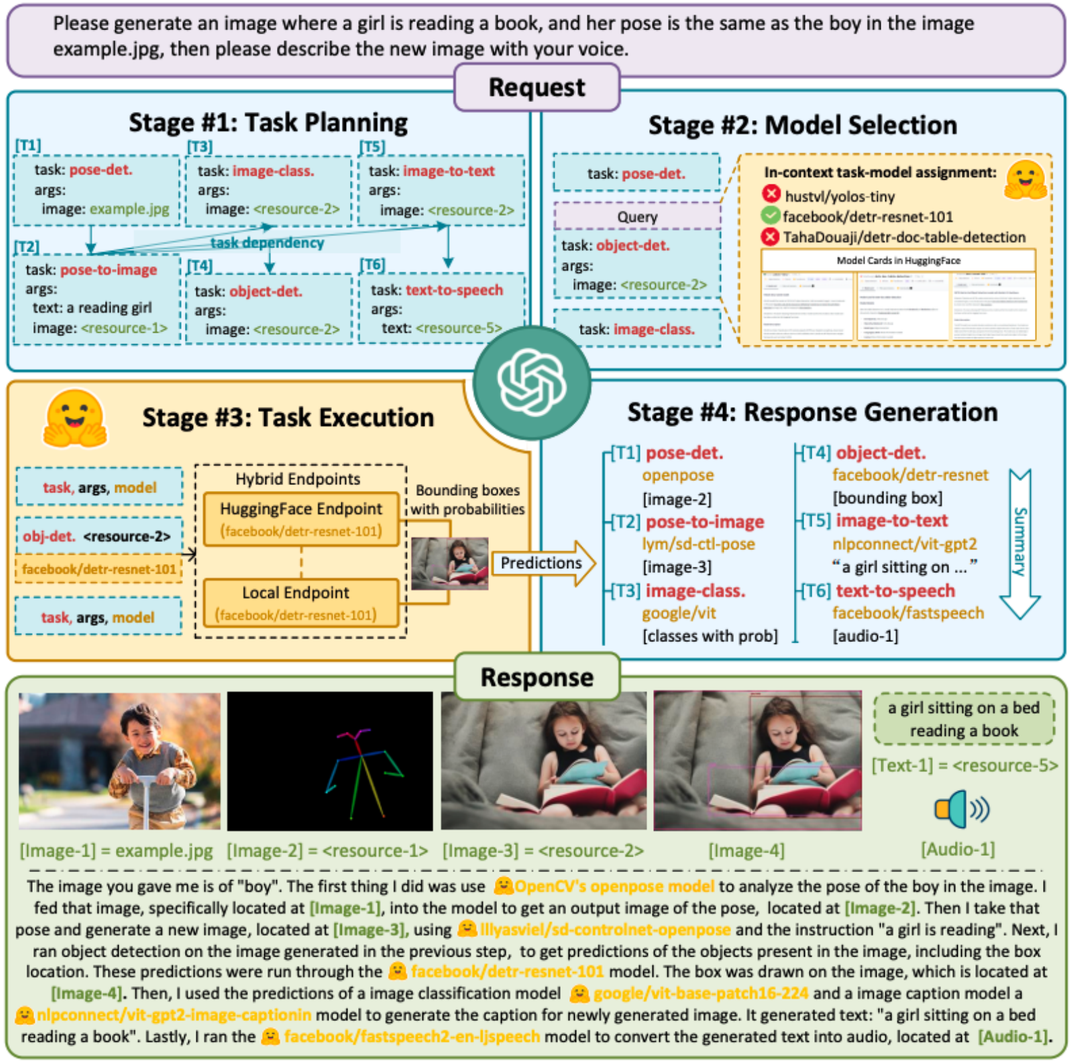
- Method
  - Training objective
    - Classification loss
      - Hungarian matched pair에 sigmoid 적용 후 Focal Cross Entropy loss
    - Localization loss
      - Bbox의 center가 거리와 bbox의 height, width의 MAE
    - GIoU loss



$$\mathcal{L} = \lambda_{loc}(\mathcal{L}^{e}_{h,w} + \mathcal{L}_{center}) + \lambda_{GIoU}\mathcal{L}^{e}_{GIoU} + \lambda_{cls}\mathcal{L}_{cls}$$

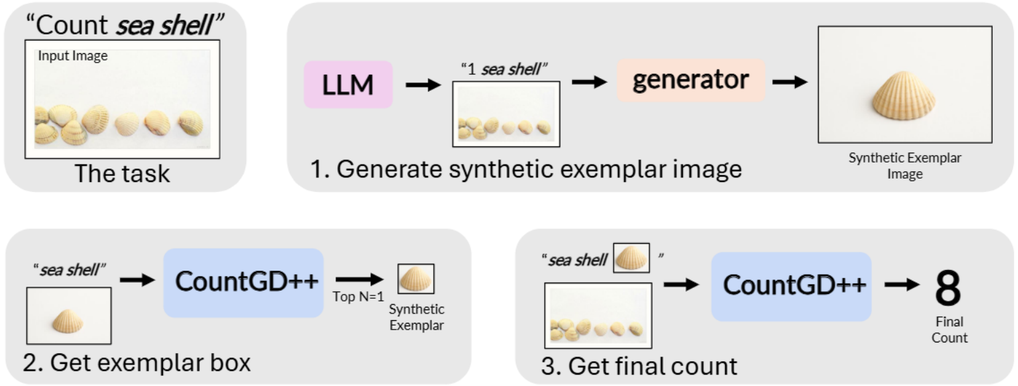
CountGD++

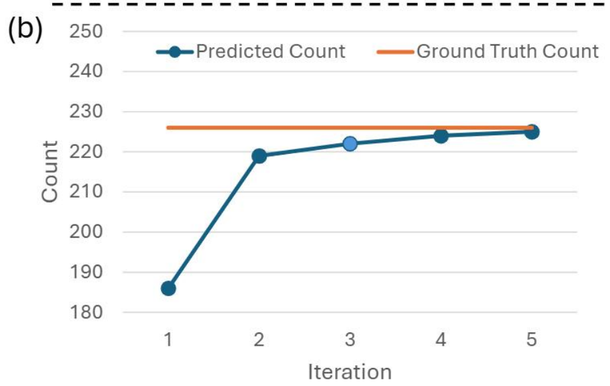
- CountGD++ as an Expert Agent
  - LLM as high-level planner
    - ViperGPT(CVPR 2023), HuggingGPT(NeurIPS 2023), VideoAgent(ECCV 2024)
    - User request를 sub-task로 분해
    - 각 sub-task에 적절한 vision model 선택
    - Vision model의 출력(box, mask, count 등)을 다시 활용하여 최종 응답 생성
  - Agent가 사용 하기 유용한 expert로서의 의의
    - Text prompt를 입력으로 받으며 visual exemplar와 negative prompt를 선택적으로 사용할 수 있음
    - Negative prompt와 exemplar 방법 자체도 이미지 내에 해당 객체들이 존재함을 알 때 유용함

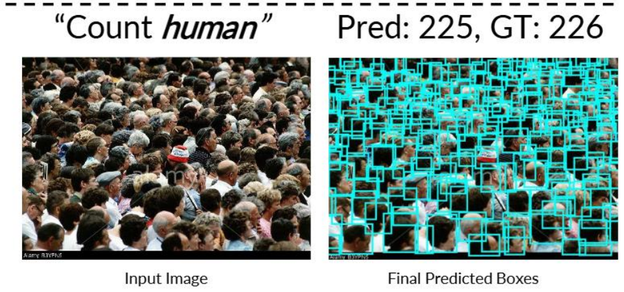


CountGD++

- CountGD++ as an Expert Agent
  - Example 1. Synthetic exemplar
    - Image generator를 통해 타겟에 대한 synthetic example image를 생성
      - 부정 프롬프트가 있다면 부정 example도 생성
    - CountGD++를 통해 exemplar box 추론 후 visual exemplar까지 활용하여 최종 출력 생성
  - Example 2. Iterative agent for images
    - Top-N box를 CountGD++에 pseudo exemplar로 다시 입력하여 count가 수렴할 때까지 반복

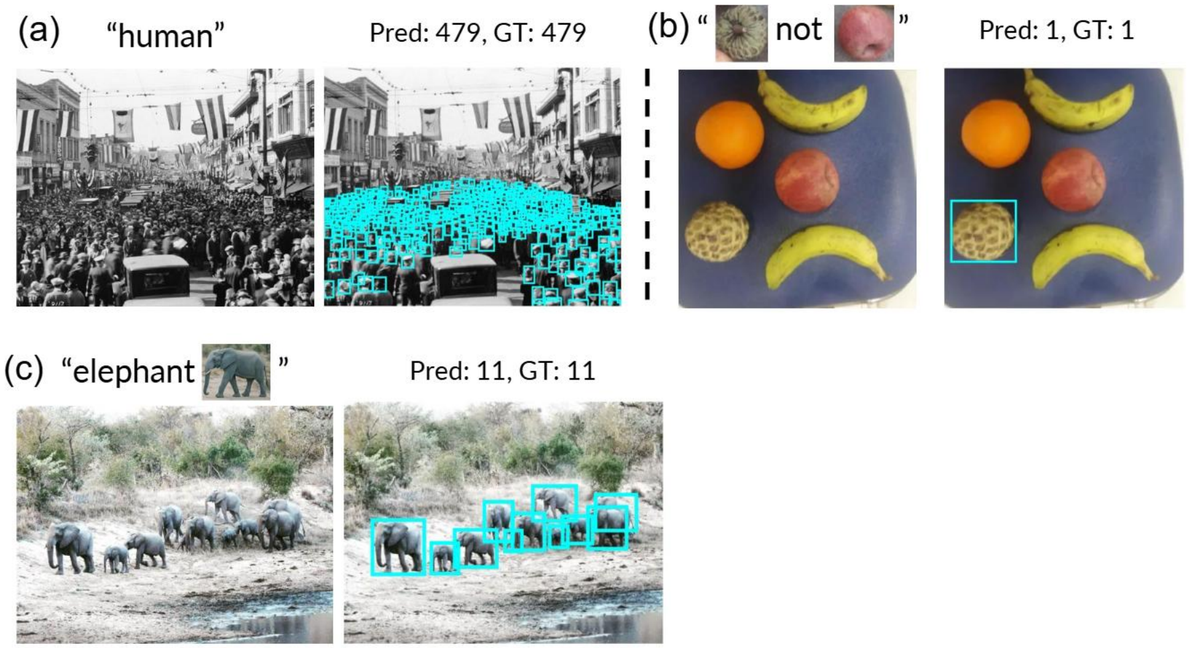






CountGD++

- Experiment



Method	FSCD-147 Test Text Only			
	Counting		Detection	
	MAE ↓	RMSE ↓	AP ↑	AP50 ↑
DAVE$_{prm}$ [42]	14.90	103.42	✗	✗
CountGD [7]	12.98	98.35	✗	✗
GrREC [16]	10.12	107.19	✗	✗
CAD-GD [55]	10.35	86.88	✗	✗
CountSE [32]	**7.84**	82.99	✗	✗
PseCo [61]	16.58	129.77	37.91*	62.45*
DAVE$_{prm}$ [42]	15.52	114.10	18.50	50.24
CGD-B [5]	15.01	118.16	30.44	61.56
Ours$_t$	16.55	129.76	33.01	61.75
Ours$_{t+p}$	10.29	33.52	37.78	68.90
Ours$_{t+p+s}$	**8.39**	**27.03**	**38.93**	**71.35**

Thank you