Object Detection via Language Modality



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



Presented By
김태훈 석사과정

Outline

- Background
  - Open-Vocabulary Object Detection
- GroundingDINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection
- YOLO-World: Real-Time Open-Vocabulary Object Detection

Background

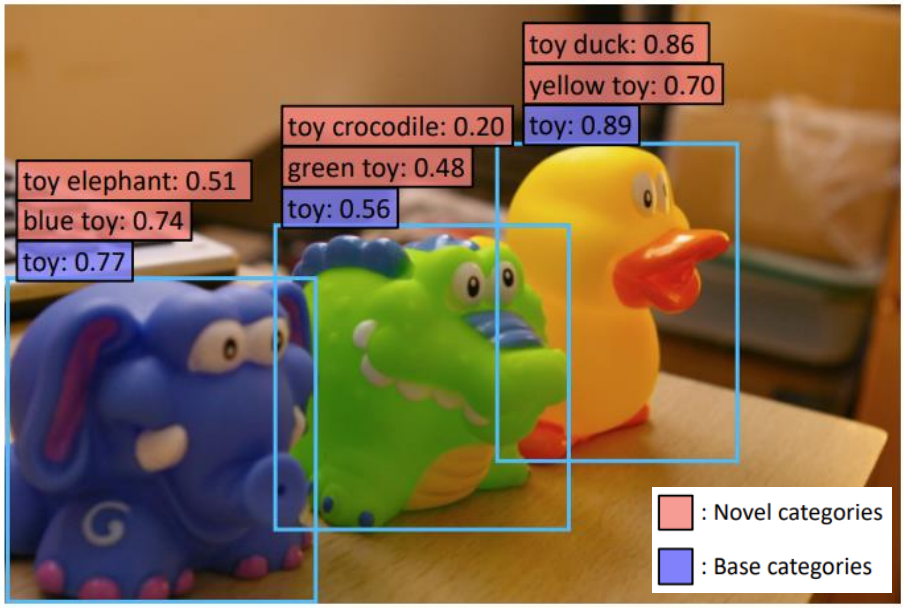
- Open-Vocabulary Object Detection
  - Traditional Object Detection
    - 이미지에서 물체의 위치와 종류를 동시에 예측하는 문제
      - Input: 이미지 → Output: 물체의 위치(Bounding Box), 종류(Class)
    - **Closed-Set Issue**
      - 학습된 객체 Class가 아닌 이상 탐지를 못함
      - 추가로 다른 객체 Class를 탐지하기 위해 Fine-tuning이 필요함



<Object Detection Example>

Background

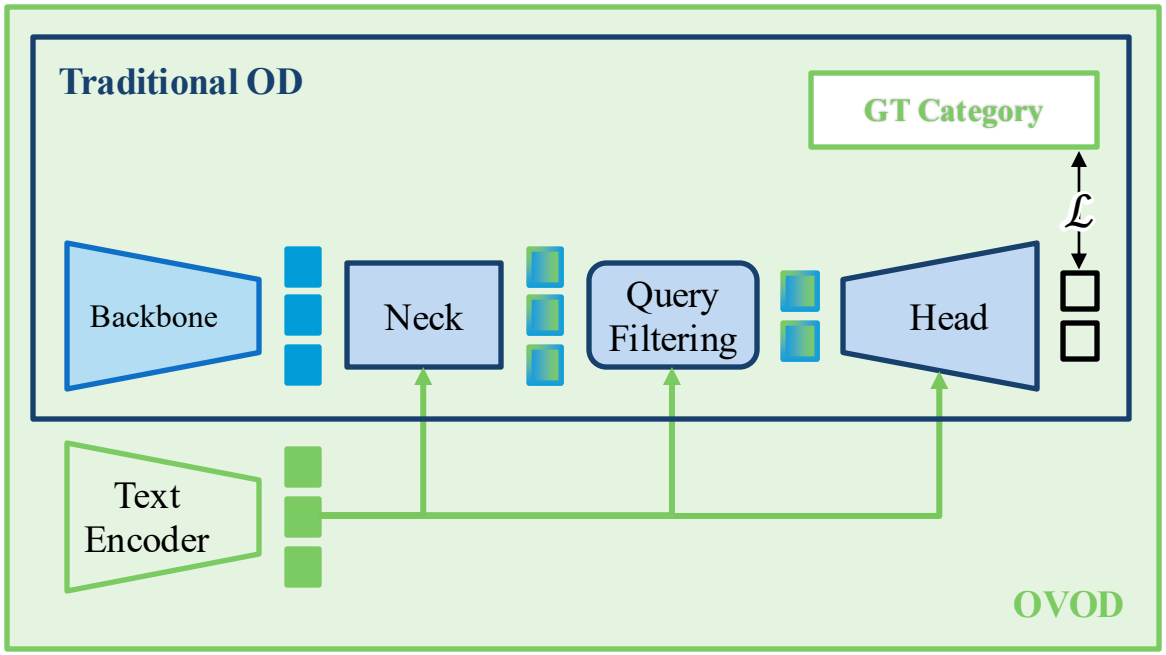
- Open-Vocabulary Object Detection
  - Open-Vocabulary Object Detection (OVOD)
    - 학습 시점에 객체와 언어 표현을 동시에 학습
      - Input: 이미지, **텍스트** → Output: 물체의 위치(Bounding Box), 종류(Category)
    - **Zero-shot Ability**
      - 학습 데이터에 없는 새로운 객체도 언어 표현(텍스트)을 통해 탐지 가능



<Open-Vocab Object Detection Example>

Background

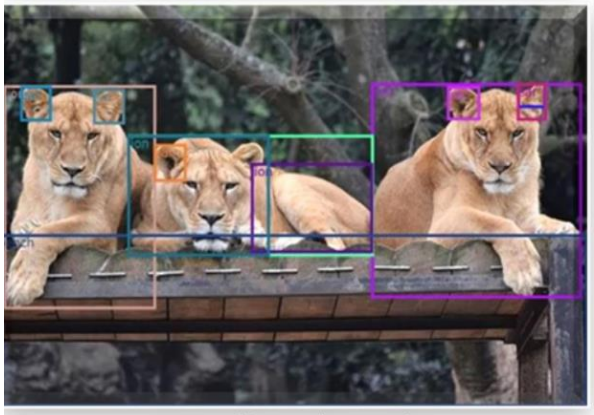
- Open-Vocabulary Object Detection
  - Traditional Object Detection vs. OVOD
    - Traditional Object Detection 학습시, Class에 대한 정보가 간접적으로 사용됨
      - **Loss Function에 의해 최종적으로 반영됨**
    - OVOD 학습시, Text를 Encode한 Text Embedding이 직접적으로 사용됨
      - Loss Function에 의해 반영되면서도 **추론 과정에서 직접 사용됨**



Background

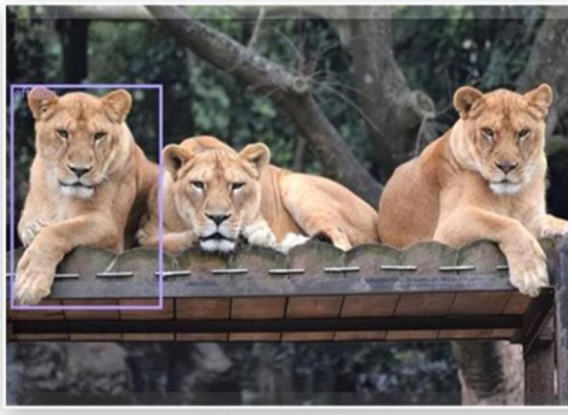
- Open-Vocabulary Object Detection
  - Traditional Object Detection vs. OVOD
    - Dataset
      - Traditional Object detection: COCO, OpenImage
      - Referring Object detection: RefCOCO/+/g, Visual Genome

<Traditional OD>



ear, lion, bench

<Referring OD>



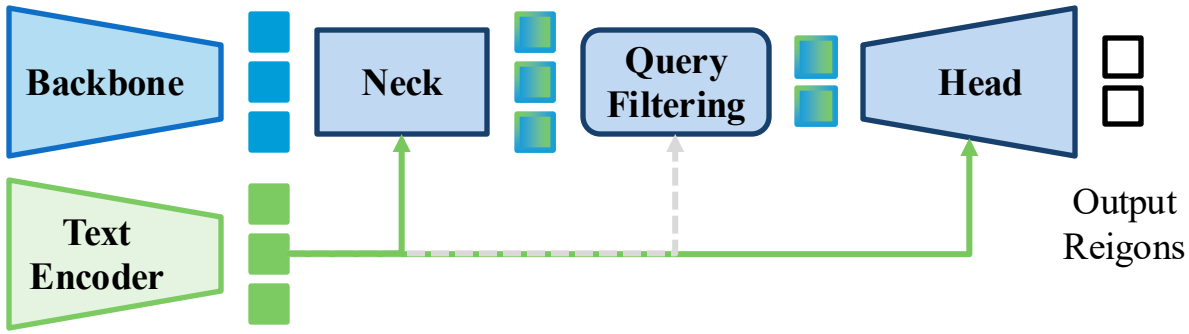
The left lion

Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection
[ECCV 24]

Grounding DINO

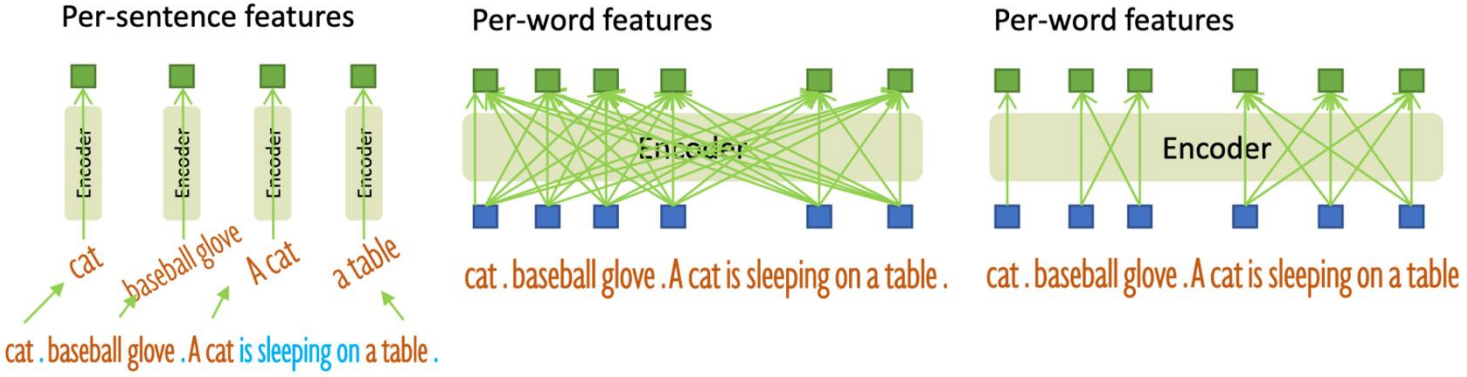
Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

- Problem of Previous Method
  - Vision-Language Fusion을 일부 모듈에만 적용돼 **언어능력이 sub-optimal**



<Partial applied modality fusion>

  - Text Feature를 추출 시 **상관없는 다른 카테고리간 영향**을 미침



(a) Sentence level
Traditional OVOD

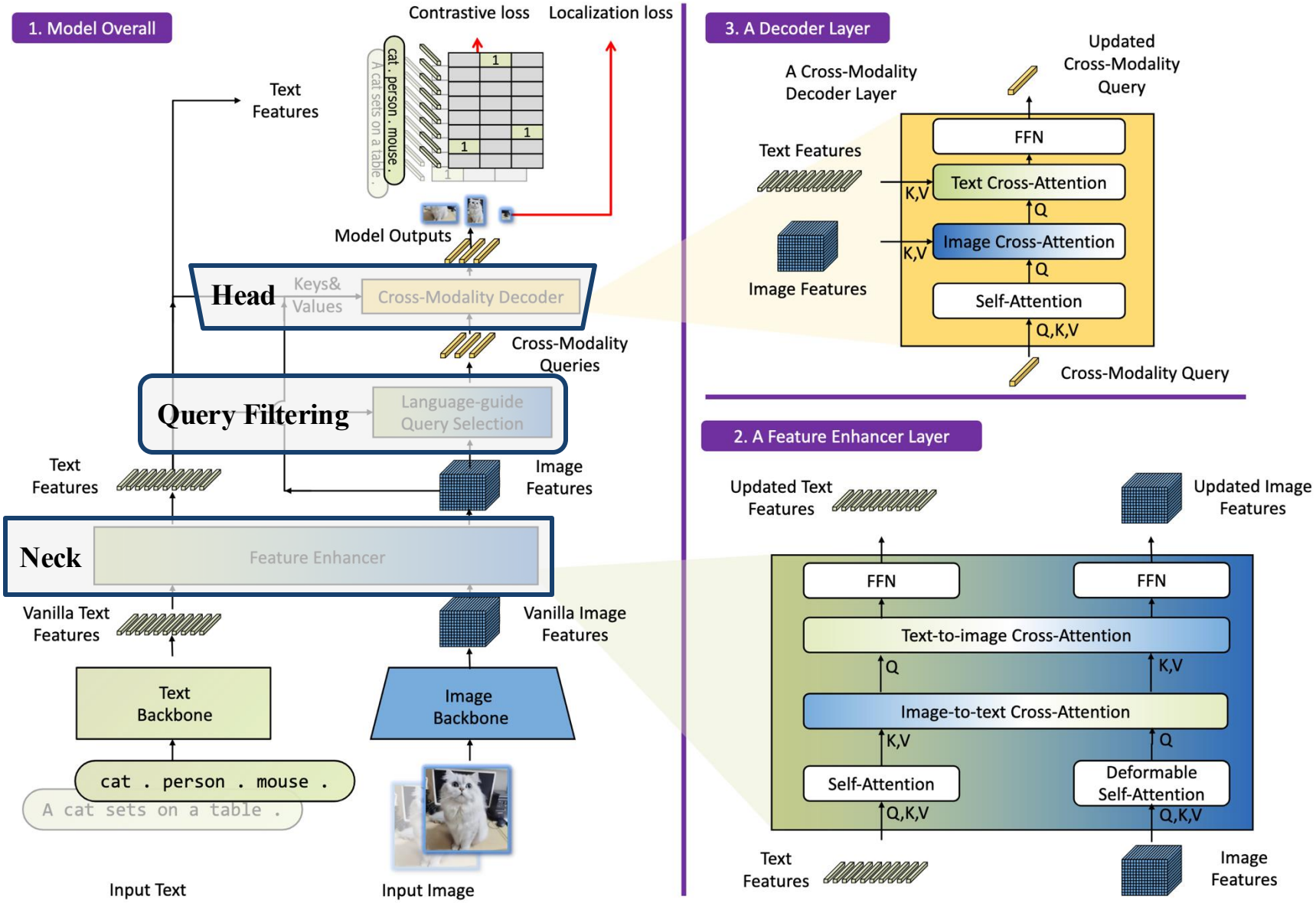
(b) Word level
e.g. **GLIP**

(c) Sub-sentence level
e.g. **GroundingDINO**

Grounding DINO

Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

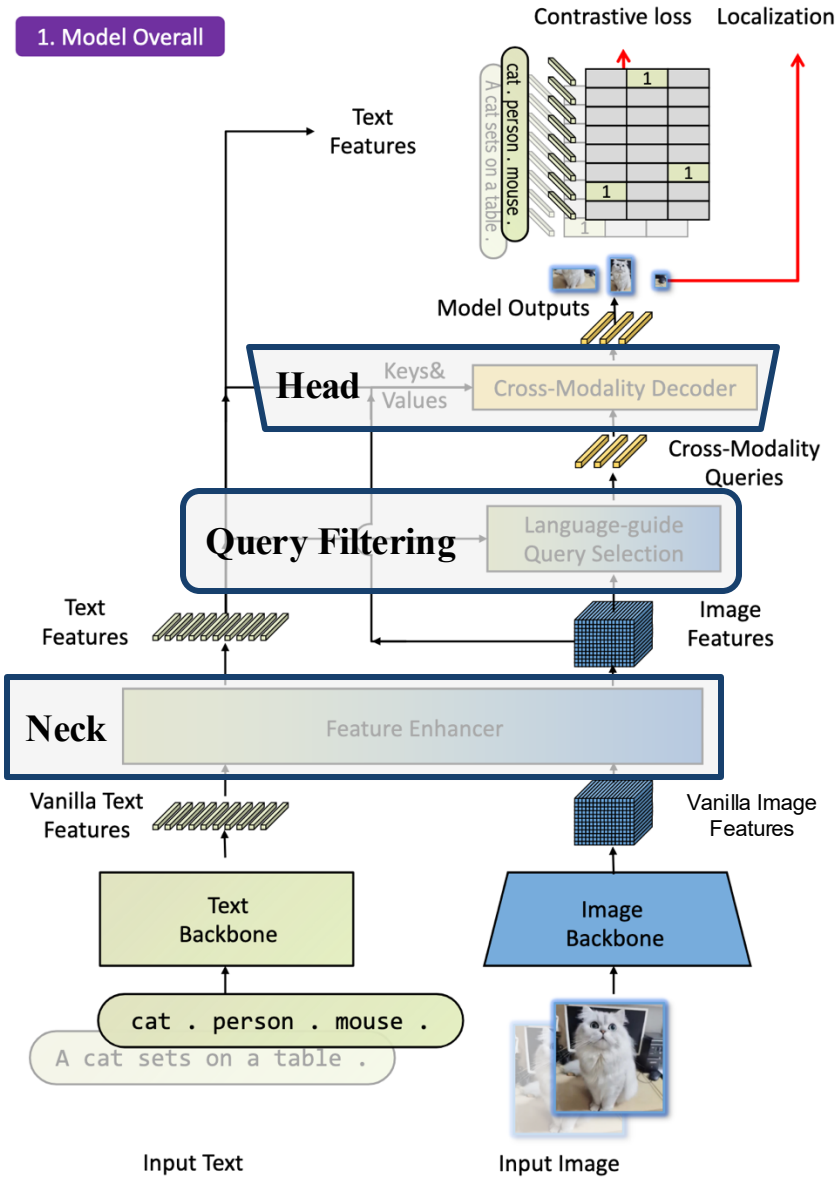
- Proposed Model Method – **Overview**



Grounding DINO

Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

- Proposed Model Method – **Overview**

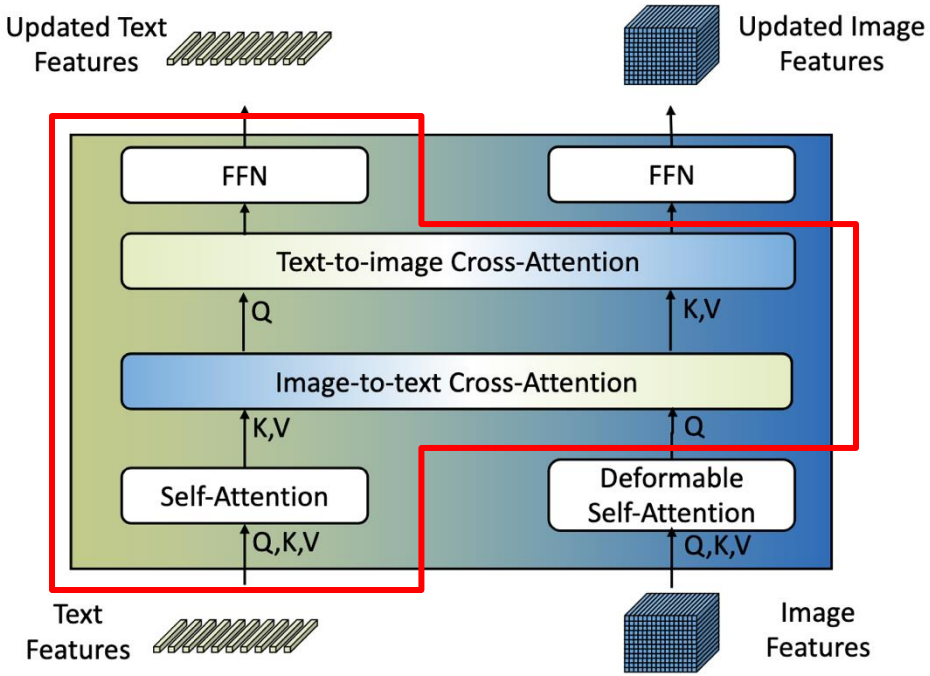


- Text Modality를 추가하기 편한 Transformer 기반 Detector인 DINO를 사용
- [1] Feature Enhancer (Neck)
  - Modality간 Cross-Attention을 추가해 Feature Fusion
- [2] Language-Guided Query Selection (Query Filtering)
  - Text Modality와의 유사도를 기반으로 Image Query를 선별함
- [3] Cross-Modality Decoder (Head)
  - DINO Decoder에 Image Modality Feature와의 Cross-Attention을 추가해 Feature를 Text에 Align

Grounding DINO

Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

- [1] Proposed Model Method – **Neck**
  - 먼저 Self-Attention으로 각 Modality의 Feature 추출
  - 기존의 DINO Encoder에 Image-to-Text, Text-to-Image Cross-Attention을 추가
    - 각 **Modality간 align을 맞춰줌**



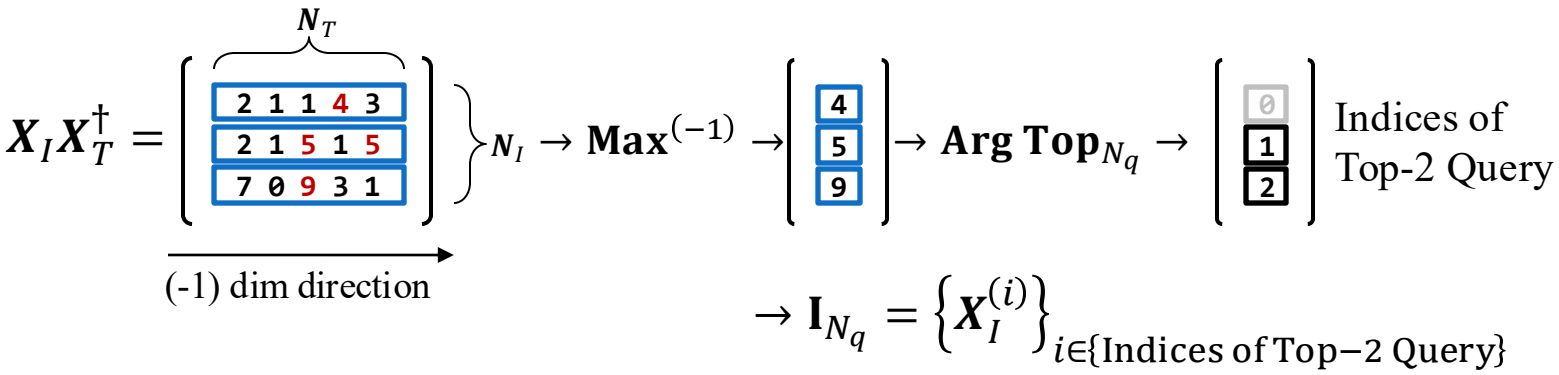
Grounding DINO

Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

- [2] Proposed Model Method – **Query Filtering**
  - 각 Query(Image Feature)에 대해 Text와 Align이 맞는 후보 Query를 Top-N_q 개 뽑음

$$\mathbf{I}_{N_q} = \mathbf{Top}_{N_q}\left(\mathbf{Max}^{(-1)}\left(\boldsymbol{X}_I \boldsymbol{X}_T^{\dagger}\right)\right)$$

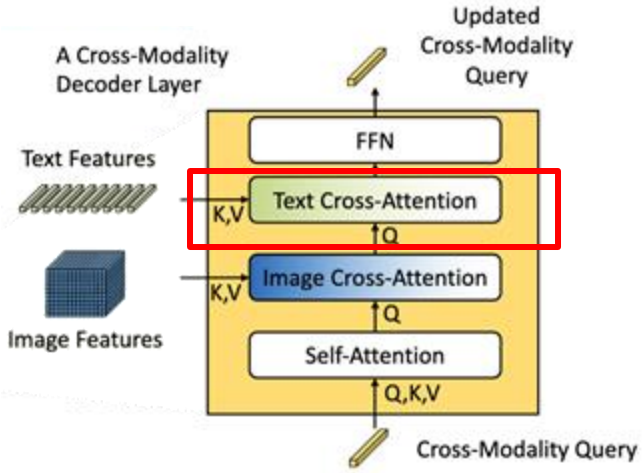
(예시) When N_q=2



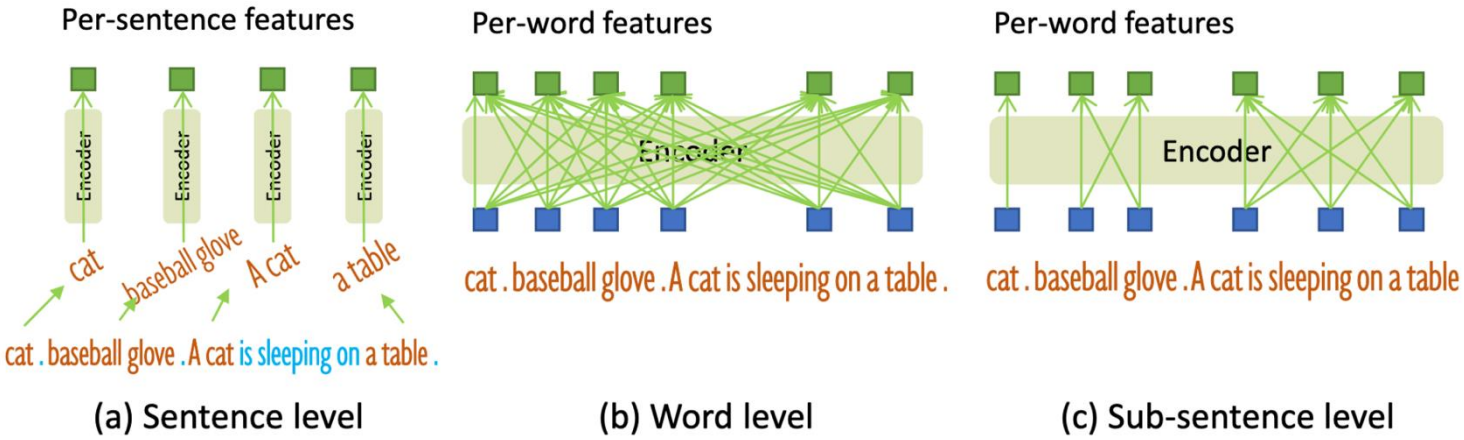
Grounding DINO

Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

- [3] Proposed Model Method – **Head**
  - 필터링한 쿼리에 대한 **Text Cross-Attention**을 기존 DINO Decoder 구조에 **추가**
    - 각 Modality간 align을 맞춰줌



- Sub-Sentence Level Text Feature
  - 동일 카테고리 사이에만 Self-Attention 허용
    - 서로 다른 Category에 대해서 Attention Mask를 적용



Grounding DINO

Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

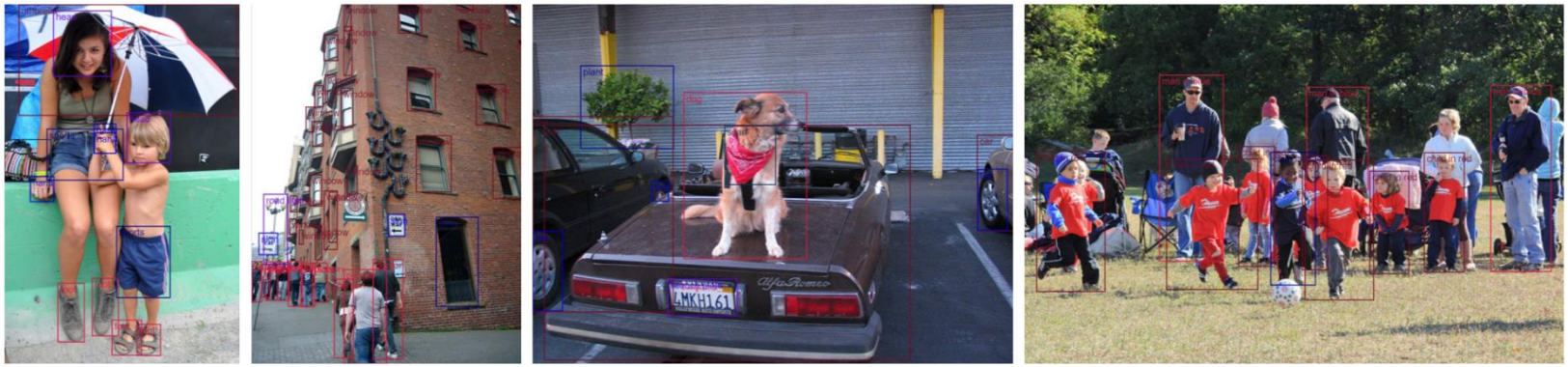
- Proposed Model Method – **Training Method**
  - Proposed Loss Terms
    - Classification Loss: $\mathcal{L}_{cls}$
      - Region-Text 불일치 정도를 나타내는 Cross-entrophy loss
  - DETR Loss Terms
    - L1 Loss, GIoU Loss : $\mathcal{L}_{L1}, \mathcal{L}_{GIoU}$
      - 예측한 Bounding Box의 불일치 정도
  - **Total Loss**

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{bbox} \cdot \mathcal{L}_{bbox} + \lambda_{GIoU} \cdot \mathcal{L}_{GIoU}$$

Grounding DINO

Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

- Proposed Model Method – **Training Data**
  - Region-Text Pair Dataset
    - Traditional Object detection: Object365(V1), OpenImage(V1)
    - Referring Object detection: GoldG(= GQA+Flickr), RefCOCO/+/g
    - Image Captioning: Cap4M for BERT Encoder



Grounding DINO

Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

- Experiments – Zero-shot Transfer
  - COCO Zero-shot Experiments
    - 동일 학습 데이터에 대해 GLIP보다 GroundingDINO가 좋은 **Zero-shot 성능**

	Model	Backbone	Pre-Training Data	Zero-Shot 2017val	Fine-Tuning 2017val/test-dev
(baseline)	DyHead-T† [5]	Swin-T	O365	43.6	53.3 / -
	GLIP-T (B) [25]	Swin-T	O365	44.9	53.8 / -
	GLIP-T (C) [25]	Swin-T	O365,GoldG	46.7	55.1 / -
	GLIP-L [25]	Swin-L	FourODs,GoldG,Cap24M	49.8	60.8 / 61.0
(baseline)	DINO(Swin-T)† [57]	Swin-T	O365	46.2	56.9 / -
	Grounding DINO T (Ours)	Swin-T	O365	46.7	56.9 / -
	Grounding DINO T (Ours)	Swin-T	O365,GoldG	48.1	57.1 / -
	Grounding DINO T (Ours)	Swin-T	O365,GoldG,Cap4M	48.4	57.2 / -
	Grounding DINO L (Ours)	Swin-L	O365,OI [19],GoldG	**52.5**	**62.6** / **62.7** (**63.0** / **63.0**)*
	Grounding DINO L (Ours)	Swin-L	O365,OI,GoldG,Cap4M,COCO,RefC	**60.7**‡	**62.6** / -

Grounding DINO

Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

- Experiments – Referring Object Detection
  - RefCOCO Zero-shot Experiments
    - GLIP이 학습 데이터를 더 사용 했으나 GroundingDINO가 비슷하거나 더 좋은 성능 달성
    - Fine-tuning시 기존의 방법론보다 더 좋은 성능 달성

Method	Backbone	Pre-Training Data	Fine-tuning	RefCOCO			RefCOCO+			RefCOCOg	
				val	testA	testB	val	testA	testB	val	test
MAttNet [53]	R101	None	✓	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
VGTR [10]	R101	None	✓	79.20	82.32	73.78	63.91	70.09	56.51	65.73	67.23
TransVG [7]	R101	None	✓	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73
VILLA_L* [11]	R101	CC, SBU, COCO, VG	✓	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
RefTR [26]	R101	VG	✓	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01
MDETR [18]	R101	GoldG,RefC	✓	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
DQ-DETR [45]	R101	GoldG,RefC	✓	88.63	91.04	83.51	81.66	86.15	73.21	82.76	83.44
GLIP-T(B)	Swin-T	O365,GoldG		49.96	54.69	43.06	49.01	53.44	43.42	65.58	66.08
GLIP-T	Swin-T	O365,GoldG,Cap4M		50.42	54.30	43.83	49.50	52.78	44.59	66.09	66.89
Grounding DINO T (Ours)	Swin-T	O365,GoldG		50.41	57.24	43.21	51.40	57.59	45.81	67.46	67.13
Grounding DINO T (Ours)	Swin-T	O365,GoldG,RefC		73.98	74.88	59.29	66.81	69.91	56.09	71.06	72.07
Grounding DINO T (Ours)	Swin-T	O365,GoldG,RefC	✓	89.19	91.86	85.99	81.09	87.40	74.71	84.15	84.94
Grounding DINO L (Ours)*	Swin-L	O365,OI,GoldG,Cap4M,COCO,RefC	✓	**90.56**	**93.19**	**88.24**	**82.75**	**88.95**	**75.92**	**86.13**	**87.02**

Grounding DINO

Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

- Experiments – Zero-shot Transfer
  - LVIS Zero-shot and Fine-Tuning Experiments
    - 동일 학습 데이터에 대해 GLIP보다 GroundingDINO가 좋은 **Zero-shot 성능**
    - 동일 학습 데이터에 대해 DetCLIPv2보다 GroundingDINO 가 좋은 Fine-Tuning 성능

Model	Backbone	Pre-Training Data	MiniVal [18] AP	MiniVal [18] APr/APc/APf
Zero-Shot Setting				
GLIP-T (C)	Swin-T	O365,GoldG	24.9	17.7/19.5/31.0
GLIP-T	Swin-T	O365,GoldG,Cap4M	26.0	20.8/21.4/31.0
DetCLIPv2	Swin-T	O365,GoldG,CC15M	40.4	36.0/41.7/40.0
Grounding DINO T	Swin-T	O365,GoldG	25.6	14.4/19.6/32.2
Grounding DINO T	Swin-T	O365,GoldG,Cap4M	27.4	18.1/23.3/32.7
Grounding DINO L	Swin-L	O365,OI,GoldG,Cap4M, COCO,RefC	33.9	22.2/30.7/38.8
Fine-Tune Setting				
MDETR	RN101	GoldG,RefC	24.2	20.9/24.9/24.3
Mask R-CNN	RN101	-	33.3	26.3/34.0/33.9
DetCLIPv2 [51]	Swin-T	O365,GoldG,CC15M	50.7	44.3/52.4/50.3
Grounding DINO T	Swin-T	O365,GoldG	**52.1**	35.4/51.3/55.7

Grounding DINO

Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

- Experiments – Vision-Language Fusion과 sub-sentence의 실효성
  - 실험[1-3]
    - Full Model에 대해 각 Language Modality를 지우기 → 모두 성능이 떨어짐
      - 기존 DINO Model에서 가장 많이 Complexity 커지는 encoder fusion의 성능 하락 폭이 가장 큼
  - 실험[4]
    - Full Model에 대해 Sub-sentence 대신 word-level text prompt 사용 → 성능이 떨어짐

#ID	Model	COCO minival		LVIS minival
		Zero-Shot	Fine-Tune	Zero-Shot
0	Grounding DINO (Full Model)	46.7	56.9	16.1
1	w/o encoder fusion	45.8	56.1	13.1
2	static query selection	46.3	56.6	13.6
3	w/o text cross-attention	46.1	56.3	14.3
4	word-level text prompt	46.4	56.6	15.6

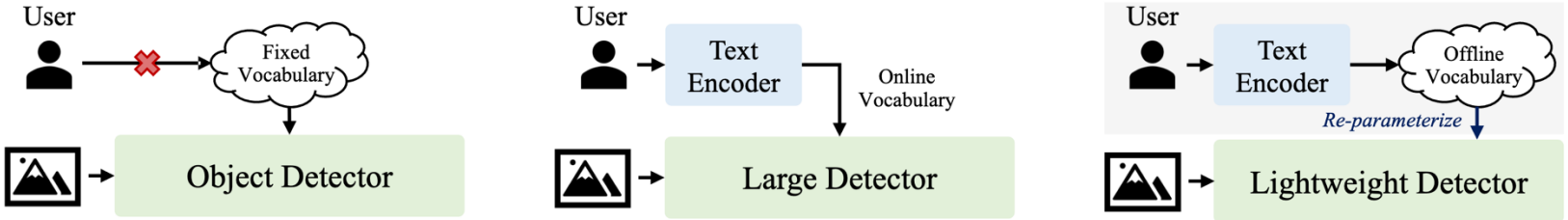
YOLO-World: Real-Time Open-Vocabulary Object Detection [CVPR 24]

YOLO-World

Cheng, Tianheng, et al. "Yolo-world: Real-time open-vocabulary object detection."
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

- Problem of Previous Method
  - Edge computing incompatible Issue
    - 큰 컴퓨팅 자원을 요구함
    - Edge Device에 올리기에 변환 과정이 복잡함

- Proposed Solution
  - Prompt-then-detect paradigm 제안 → **'Personalized Object Detector'**
    - **Traditional Object detection의 속도의 장점과 OVOD의 Zero-shot ability 활용**
    - 탐지 대상 객체를 User Prompt로 받아 Offline Vocabulary (Closed-Set)을 구성
      - OVOD의 Zero-shot ability로 인해 추가적인 데이터 셋 학습 없이 사용자가 원하는 객체 탐지 가능
    - Detector re-parameterization을 통해 Traditional Object detection과 유사한 속도를 가짐

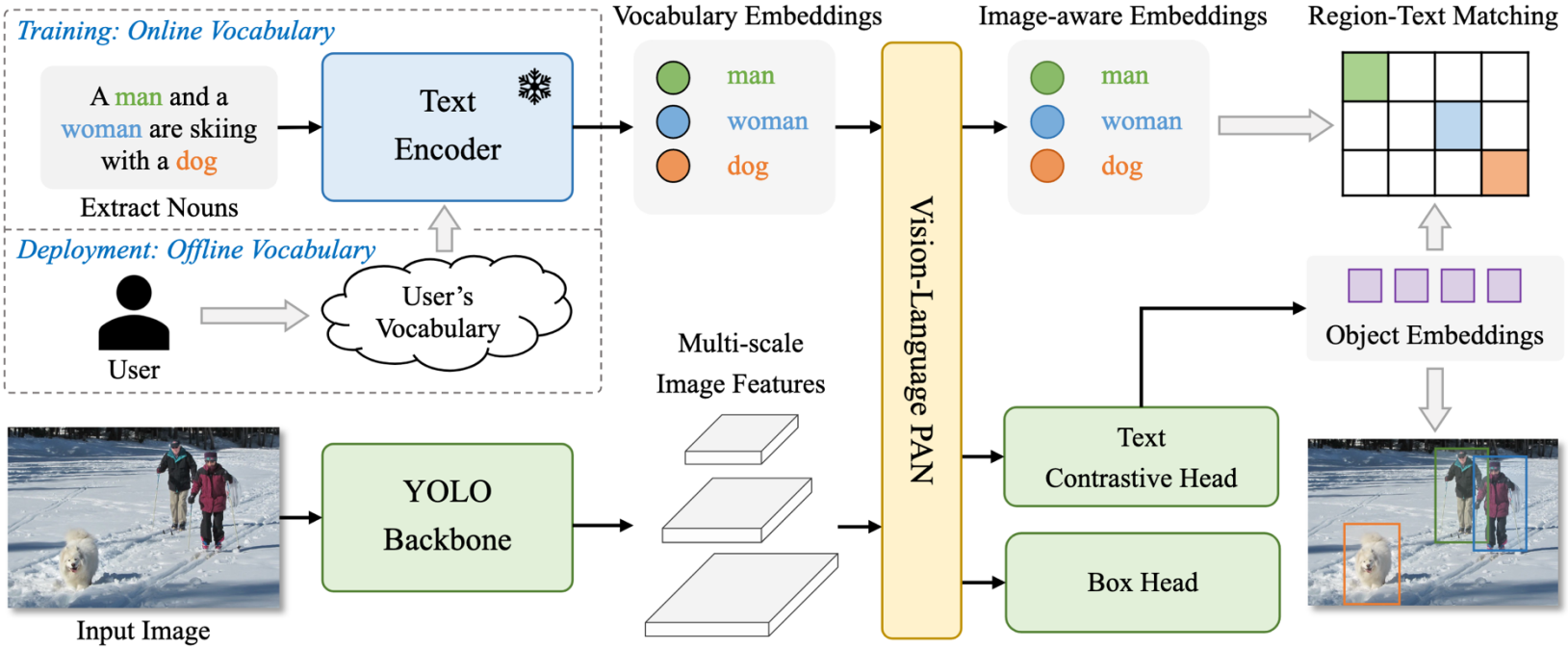


(a) Traditional Object Detector (b) Preivous Open-Vocabulary Detector (c) YOLO-World

YOLO-World

Cheng, Tianheng, et al. "Yolo-world: Real-time open-vocabulary object detection."
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

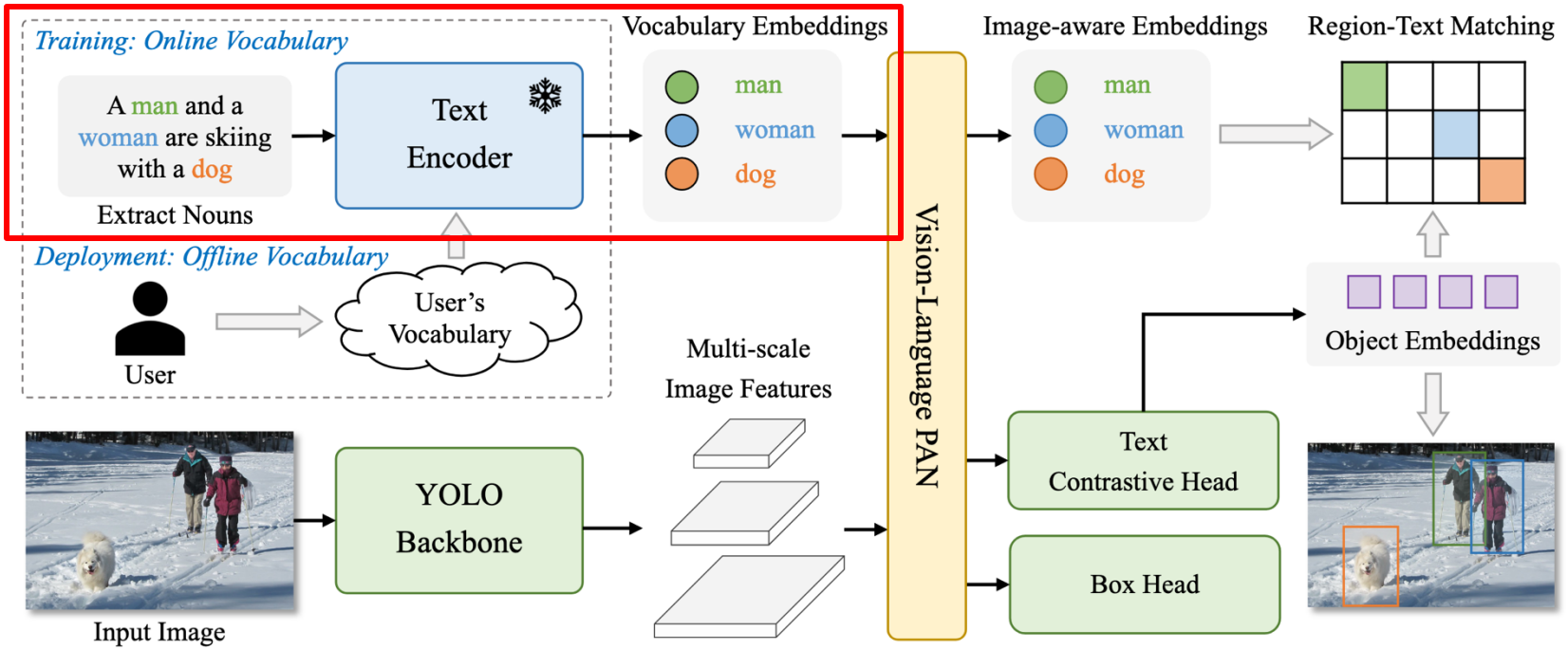
- Proposed Model Method – **Overview**
  - 속도의 이점을 얻기 위해 **YOLO Backbone**을 사용
  - Vision-Language Align된 **CLIP Text Encoder** 사용
  - Multi-scale Image Feature와 Text Embedding을 Fusion하기 위해 **Re-parameterizable Vision-Language PAN 제안**
    - Re-parameterize를 통해 fast offline vocabulary inference 가능



YOLO-World

Cheng, Tianheng, et al. "Yolo-world: Real-time open-vocabulary object detection."
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

- Proposed Model Method – **Text encoding**
  - 학습 시 Text input으로 Class, Phrase, Sentence(or caption)
    - Sentence는 N-gram algorithm을 통해 noun phrase를 추출함
  - CLIP Text Encoder 사용
    - Visual object에 대해 text를 연결하는 **visual-sementic ability가 Text-only encoder보다 좋음**



YOLO-World

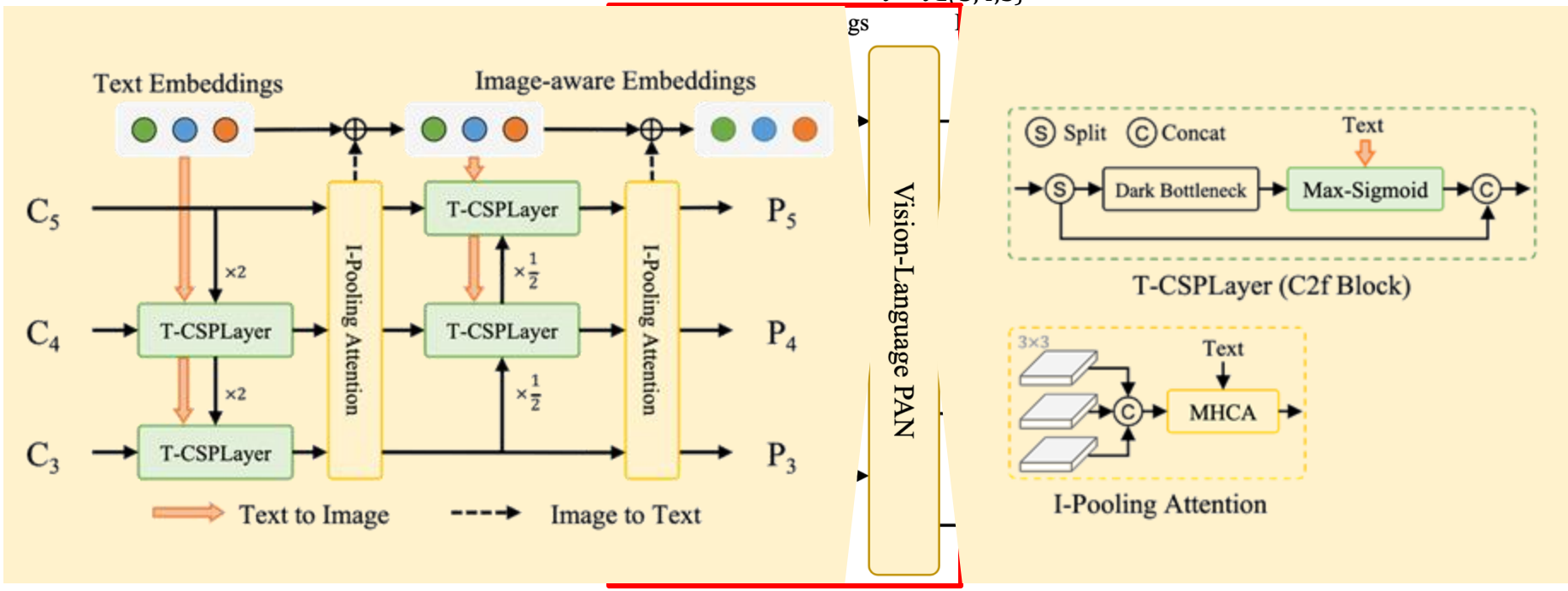
Cheng, Tianheng, et al. "Yolo-world: Real-time open-vocabulary object detection."
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

- Proposed Model Method – **Vision-Language PAN**
  - Text-guided CSPLayer

$$X_l' = X_l \cdot \textbf{Sigmoid}\left(\max_{j\in\{1..C\}} X_l W_j^{\dagger}\right)^{\dagger}, \qquad X_l \in \mathbb{R}^{H\times W\times D}(l \in \{3,4,5\}), W_j \in \mathbb{R}^{1\times D}(j \in \{1..C\})$$

  - Image-Pooling Attention

$$W' = W + \textbf{MultiHeadAttention}(W, \tilde{X}, \tilde{X}),$$
$$W = TextEncoder(T) \in \mathbb{R}^{C\times D}, \tilde{X} = [MaxPool(X_l)]_{l\in\{3,4,5\}}$$



YOLO-World

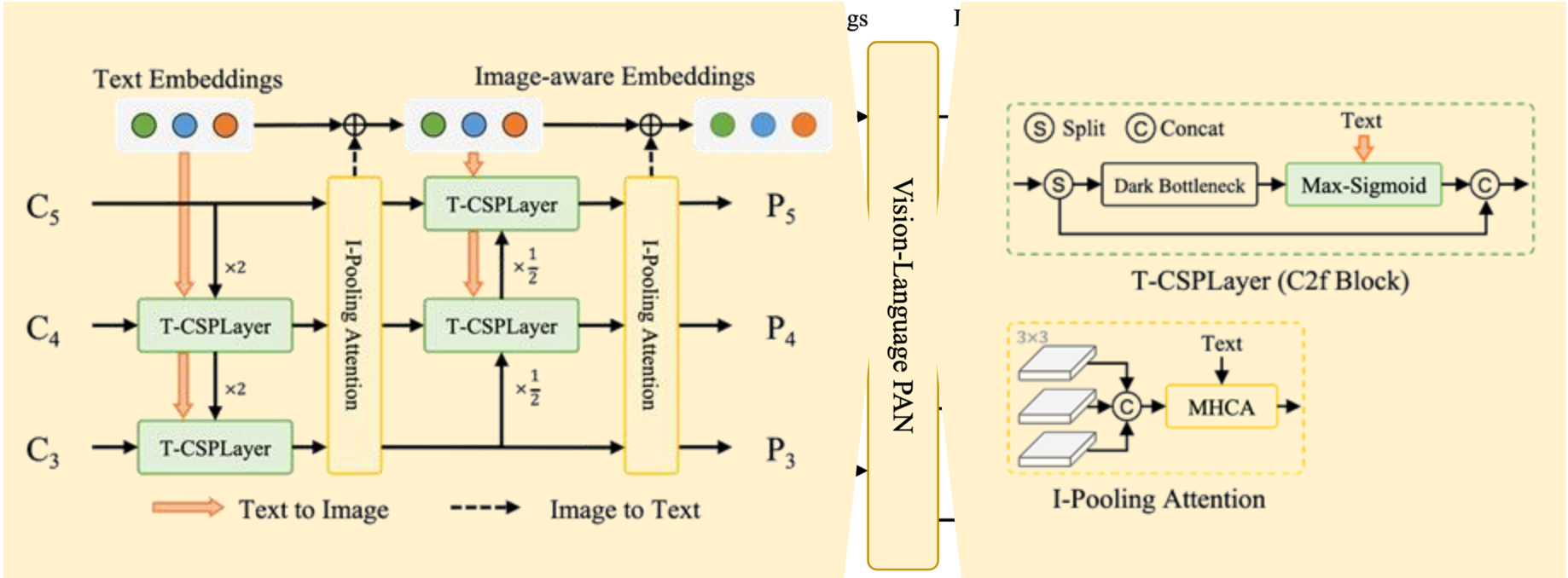
Cheng, Tianheng, et al. "Yolo-world: Real-time open-vocabulary object detection."
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

- Proposed Model Method – **Re-parameterize Vision-Language PAN**
  - Re-parameterize Text-guided CSPLayer

    $X_l' = X_l \odot \mathbf{Sigmoid}(\max(\mathbf{Conv}(X, W), \dim = 1))$ → 행렬 연산을 1x1Conv 로 대체
  - Re-parameterize Image-Pooling Attention

    $\tilde{X} = [MaxPool(X_l)]_{l \in \{3,4,5\}}$,

    $W' = W + \mathbf{Softmax}(W \odot \tilde{X}, \dim = -1) \odot W$ → Simplified Attention

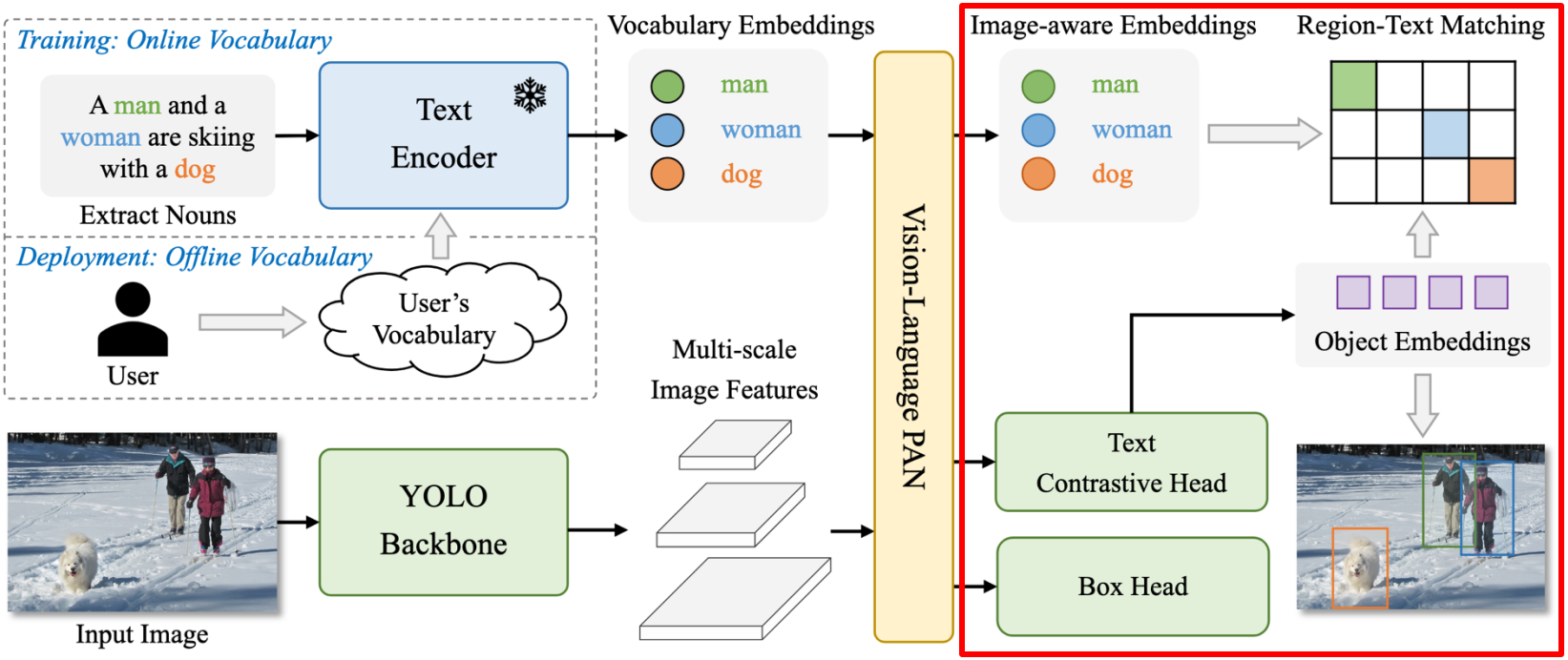


YOLO-World

Cheng, Tianheng, et al. "Yolo-world: Real-time open-vocabulary object detection."
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

- Proposed Model Method – **Head**
  - VL-PAN을 통과한 Multi-Scale Feature에 대해 3x3 Conv. 학습
    - Bounding box regression과 Object Embedding을 출력한 두 Head를 학습
    - Object Embedding과 Image-aware (text) Embedding간의 similarity를 계산

$$s_{k,j} = \alpha \cdot L2Norm(e_k) \cdot L2Norm(w_j)^{\dagger} + \beta, \qquad (\alpha, \beta \; is \; \text{learnable})$$



YOLO-World

Cheng, Tianheng, et al. "Yolo-world: Real-time open-vocabulary object detection."
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

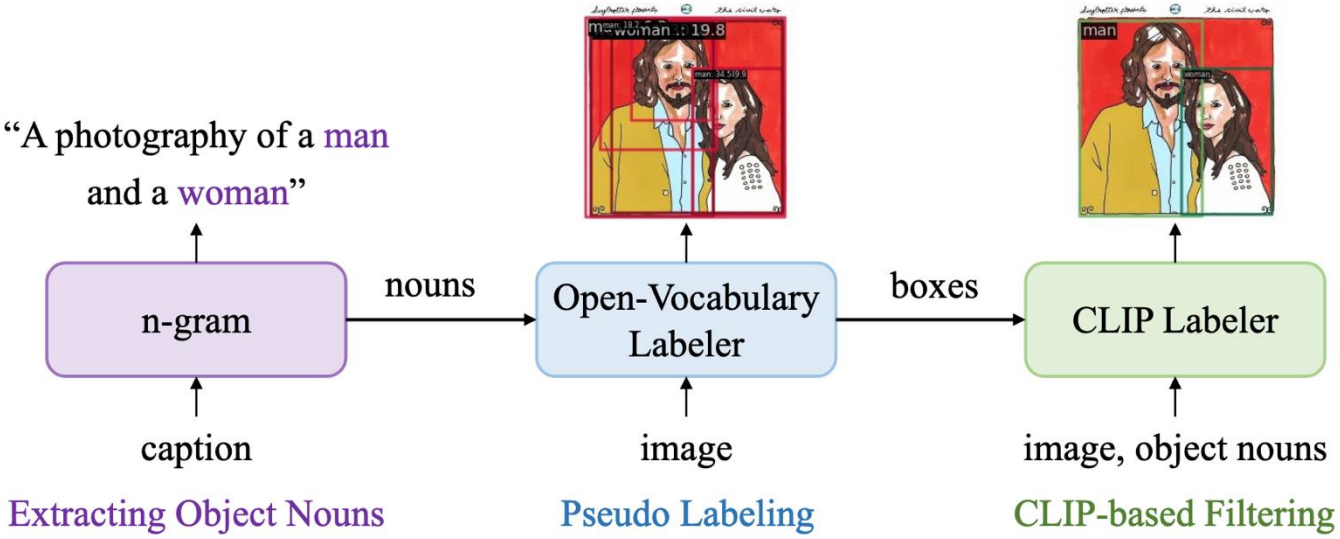
- Proposed Model Method – **Training Method**
  - Proposed Loss Terms
    - Region-Text Contrastive Loss: $\mathcal{L}_{con}$
      - Region-Text 불일치 정도를 나타내는 Cross-entrophy loss
  - YOLOv8 Loss Terms
    - IoU Loss: $\mathcal{L}_{IoU}$
      - 예측한 Region(Area)의 불일치 정도
    - Distributed Focal Loss: $\mathcal{L}_{DFL}$
      - 예측한 Region(Point)의 불일치 정도
  - **Total Loss**

$$\mathcal{L} = \mathcal{L}_{con} + \lambda_I \cdot (\mathcal{L}_{IoU} + \mathcal{L}_{DFL})$$

YOLO-World

Cheng, Tianheng, et al. "Yolo-world: Real-time open-vocabulary object detection."
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

- Proposed Model Method – **Training Data**
  - Region-Text Pair Dataset
    - Traditional Object detection: Object365(V1)
    - Referring Object detection: GoldG(= GQA+Flickr)
    - Image Captioning → Region-Text Pair Dataset: CC3M
  - Image Captioning Dataset을 Region-Text Dataset으로 만들어 추가 학습
    1. Caption을 n-gram algorithm을 통해 noun phrase 추출
    2. GLIP을 통해 Pseudo region annotate 및 NMS을 통한 Low Confidence Filtering
    3. CLIP을 통해 낮은 region-text score를 Filtering



YOLO-World

Cheng, Tianheng, et al. "Yolo-world: Real-time open-vocabulary object detection."
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

- Experiments – Zero-shot Transfer
  - LVIS Zero-shot Experiments
    - 압도적인 FPS 차이

Method	Backbone	Params	Pre-trained Data	FPS	AP	AP_r	AP_c	AP_f
MDETR [21]	R-101 [15]	169M	GoldG	-	24.2	20.9	24.3	24.2
GLIP-T [24]	Swin-T [32]	232M	O365,GoldG	0.12	24.9	17.7	19.5	31.0
GLIP-T [24]	Swin-T [32]	232M	O365,GoldG,Cap4M	0.12	26.0	20.8	21.4	31.0
GLIPv2-T [59]	Swin-T [32]	232M	O365,GoldG	0.12	26.9	-	-	-
GLIPv2-T [59]	Swin-T [32]	232M	O365,GoldG,Cap4M	0.12	29.0	-	-	-
Grounding DINO-T [30]	Swin-T [32]	172M	O365,GoldG	1.5	25.6	14.4	19.6	32.2
Grounding DINO-T [30]	Swin-T [32]	172M	O365,GoldG,Cap4M	1.5	27.4	18.1	23.3	32.7
DetCLIP-T [56]	Swin-T [32]	155M	O365,GoldG	2.3	34.4	26.9	33.9	36.3
YOLO-World-S	YOLOv8-S	13M (77M)	O365,GoldG	74.1 (19.9)	26.2	19.1	23.6	29.8
YOLO-World-M	YOLOv8-M	29M (92M)	O365,GoldG	58.1 (18.5)	31.0	23.8	29.2	33.9
YOLO-World-L	YOLOv8-L	48M (110M)	O365,GoldG	52.0 (17.6)	35.0	27.1	32.8	38.3
YOLO-World-L	YOLOv8-L	48M (110M)	O365,GoldG,CC3M†	52.0 (17.6)	35.4	27.6	34.1	38.0

YOLO-World

Cheng, Tianheng, et al. "Yolo-world: Real-time open-vocabulary object detection."
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

- Experiments – Text-only Encoder vs. CLIP Encoder
  - Object365 자체가 Text 정보가 부족해 Fine-tuning시 CLIP의 일반화 능력이 오히려 떨어짐

Text Encoder	Frozen?	AP	AP_r	AP_c	AP_f
BERT-base	Frozen	14.6	3.4	10.7	20.0
BERT-base	Fine-tune	18.3	6.6	14.6	23.6
CLIP-base	Frozen	**22.4**	**14.5**	**20.1**	**26.0**
CLIP-base	Fine-tune	19.3	8.6	15.7	24.8

- Experiments – Vision-Lanuage Fusion의 실효성
  - T-CSPLayer (T→I), I-Pooling Attention (I→T) 중에서 적어도 하나라도 빠지면 성능 감소

GQA	T→I	I→T	AP	AP_r	AP_c	AP_f
✗	✗	✗	22.4	14.5	20.1	26.0
✗	✓	✗	23.2	15.2	20.6	27.0
✗	✓	✓	23.5	16.2	21.1	27.0
✓	✗	✗	29.7	21.0	27.1	33.6
✓	✓	✓	**31.9**	**22.5**	**29.9**	**35.4**

YOLO-World

Cheng, Tianheng, et al. "Yolo-world: Real-time open-vocabulary object detection."
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

- Experiments –Pseudo Labeling의 실효성
  - 생성한 데이터셋 CC3M에 대해 추가 학습 시 LVIS Zero-shot 성능 증가

Pre-trained Data	AP	AP_r	AP_c	AP_f
O365	23.5	16.2	21.1	27.0
O365,GQA	31.9	22.5	29.9	35.4
O365,GoldG	32.5	22.3	30.6	**36.0**
O365,GoldG,CC3M†	**33.0**	**23.6**	**32.0**	35.5

Summary

- Grounding DINO vs. YOLO-World
  - Both are
    - Open-Vocabulary Object Detection Task
    - Sub-sentence Level Text Embedding
  - Grounding DINO
    - Language Fusion with DINO Detection Transformer
      - Fusing language modality for each detection modules
  - YOLO-World
    - Language Fusion with YOLO-based Detector
    - **Edge-device compatible methodology**
      - **Re-parameterizable Vision-Language Feature Fuser**
    - Prompt-then-detect paradigm