

LaMI-DETR: Open-Vocabulary Detection with Language Model Instruction(ECCV 2024)

Undergraduate Researcher at CVLab

Lee Dohyeong

2025.5.8

Contents

- Introduction
- Related Work
- Method
- Experiments
- Conclusion

Introduction

- VLM (Vision-Language Model)
 - Associate an image with a textual representation
- Open-Vocabulary
 - Recognize unseen classes using textual descriptions
- Zero-shot
 - Recognize classes without having seen labeled examples during training
- LVIS Dataset
 - Over 1,200 object categories for evaluating open-vocabulary and zero-shot detection

Introduction

- Previous Problem
 - Existing open-vocabulary detectors rely on category names to represent concepts.
 - Difficult to separate classes that are visually similar
 - Misclassify novel categories as background.
 - Novel Categories : 학습때 사용하지 않은 카테고리

Related Work

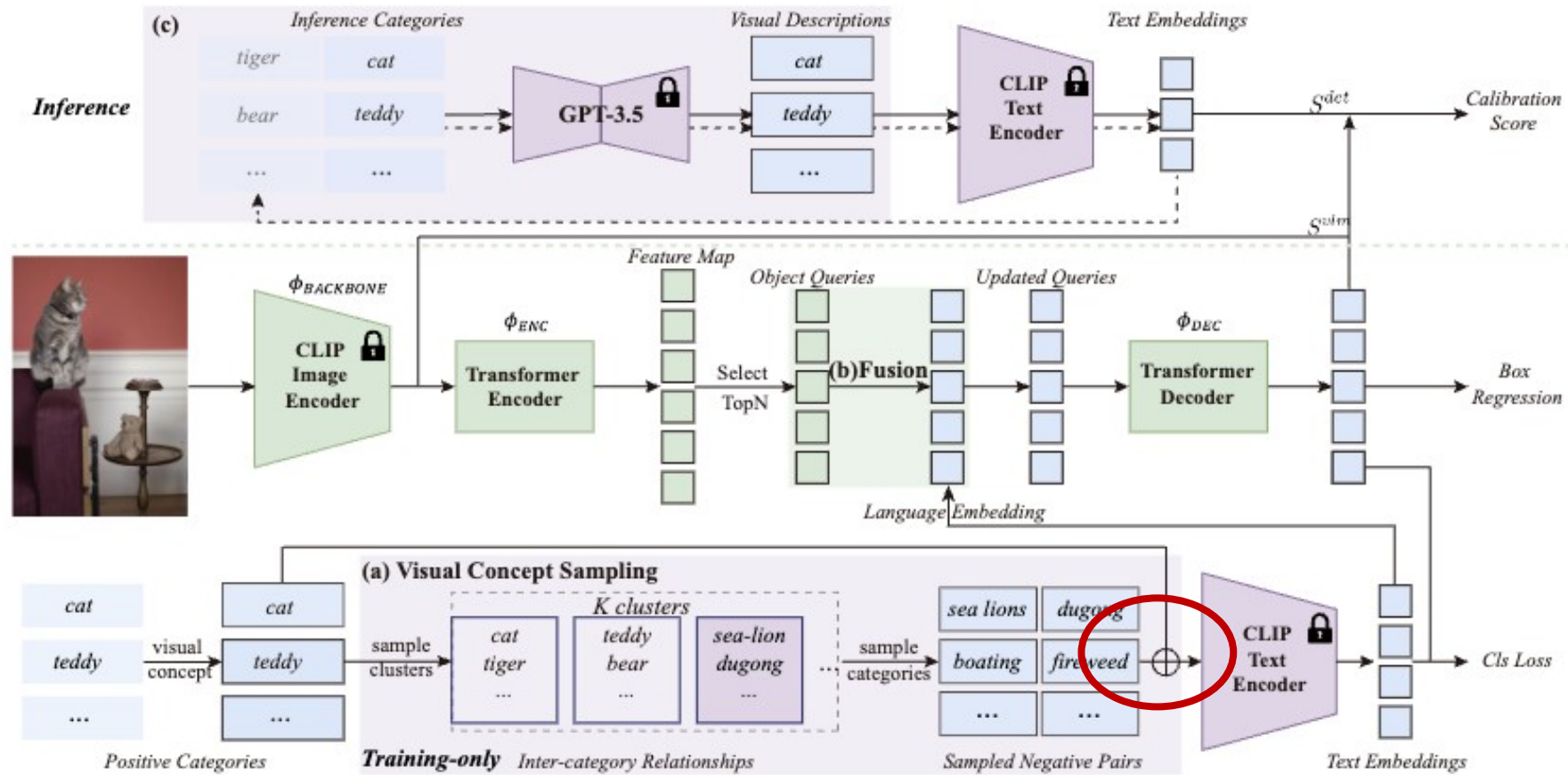
- T5(Text-To-Text Transfer Transformer)
 - semantic similarity (의미 기반 유사도)
 - Transformer 구조
 - 시각적 특성(visual characteristics) 이용
 - 예시 : “dugong”과 “sea-lion” 텍스트로는 유사하지 않음
 - Clip text encoder는 text만 같아도 유사 클러스터로 분류
 - T5는 같은 클러스터로 구분
 - 설명이 시각적으로 유사하면 같은 클러스터로 구분해준다.

Related Work

- CLIP(Contrastive Language–Image Pretraining)
 - Use Large image-text dataset
 - Arrange images and text in the same semantic space
 - Output : similarity score to the entered text
 - Spelling similarity can lead to misleadingly close embeddings.

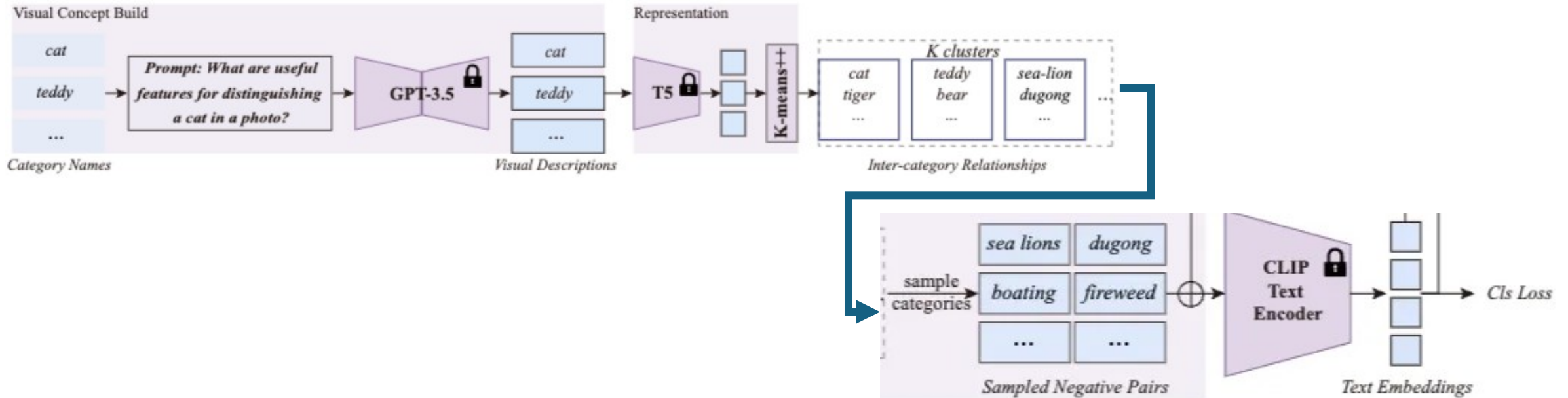
Method

- Model



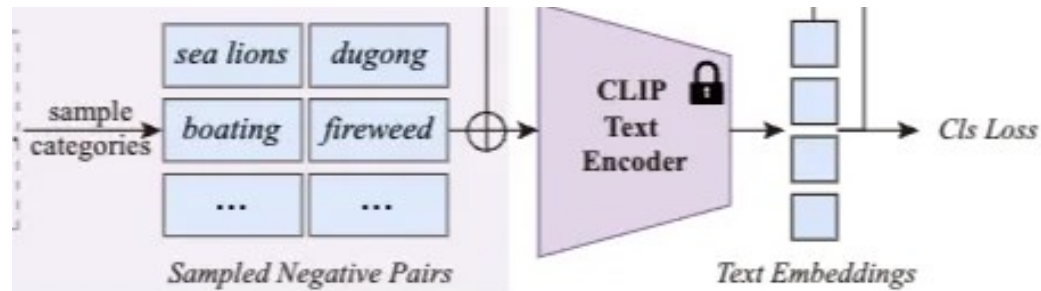
Method

- Train



Method

- Negative Sampling



$$p_{calc} = \begin{cases} 0 & \text{if } c \in C_g \\ p_c & \text{otherwise} \end{cases}$$

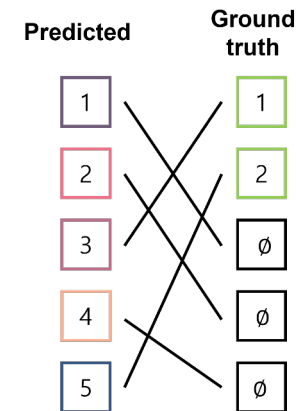
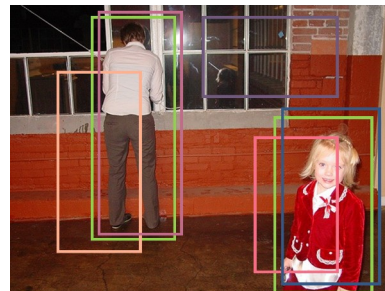
Method

- Cls Loss

Classification Score : Object query & text embedding $S_j^{\text{det}} = q_j \cdot T_{cls}$

Classification Loss : $\mathcal{L}_{cls} = -\log(\text{softmax}_{\text{class}})$

Hungarian matching:
Finds a 1:1 assignment



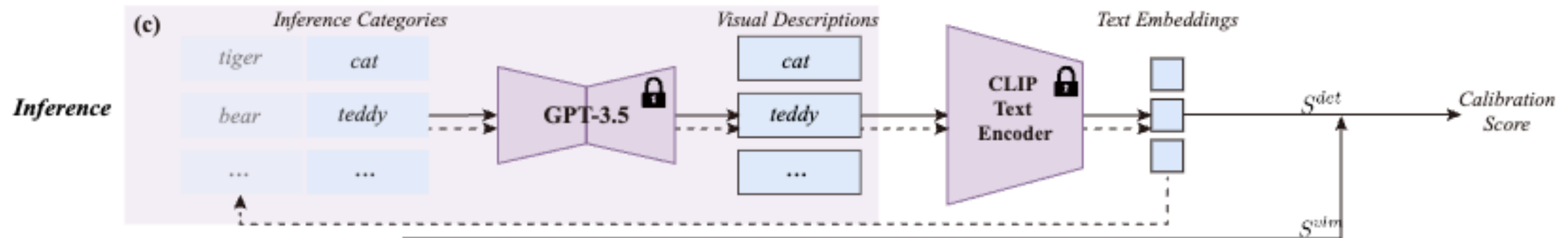
	1	2	\emptyset	\emptyset	\emptyset
1	12	11	1	1	1
2	4	2	8	5	9
3	1	3	5	7	8
4	2	5	6	7	4
5	2	1	9	10	6

Permutation = [3, 4, 1, 5, 2]

Matching score = 1 + 5 + 1 + 4 + 1 = 12

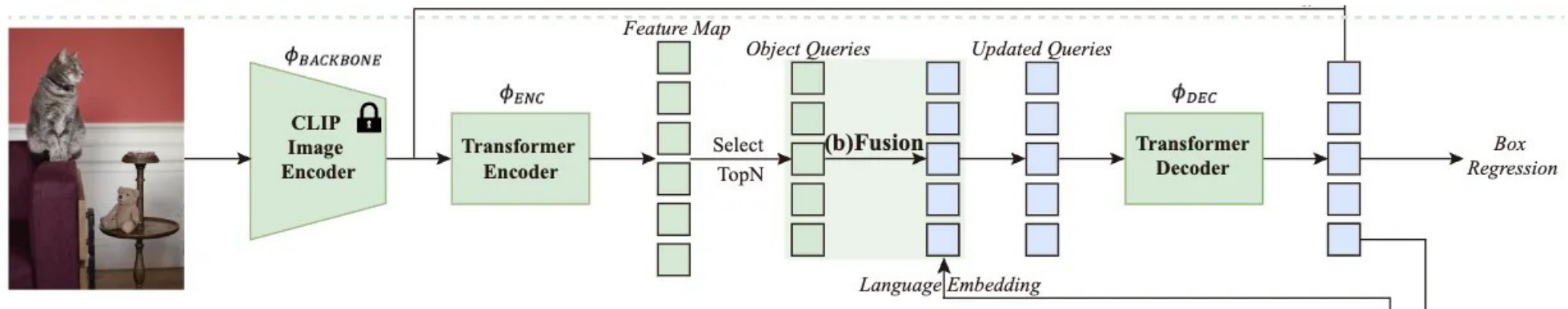
Method

- Inference



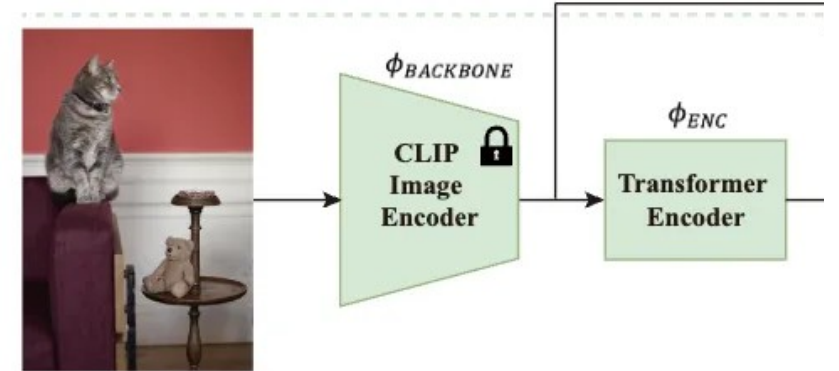
Method

- Main Model



Method

- CLIP Image Encoder(Frozen)
 - ConvNext based CNN(Backbone)
 - Preserve open-vocabulary recognition(Frozen)
 - Prevent overfitting to base categories. (Frozen)
 - Improves training speed and stability.
 - High-level visual features : (shape, texture, meaning)
- Transformer Encoder
 - Global context by relating CLIP features.



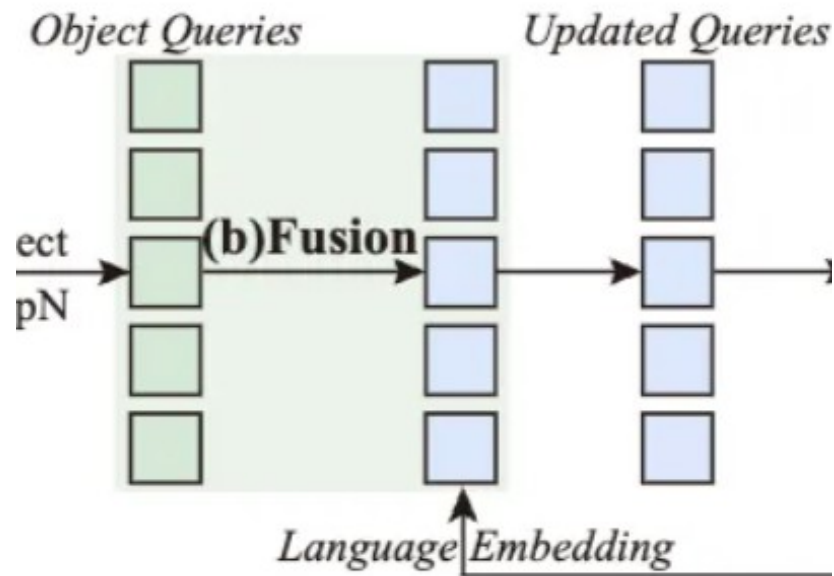
Method

- Select Top N
 - Higher dot product indicates higher similarity
 - Q_i : object Query
 - F_i : located i , feature vector
 - T_{cls} : Clip + Gpt 3.5 + T5 combination -> embedding vector

$$\{q_j\}_{j=1}^N = \text{Top}_N(\{\mathcal{T}_{cls} \cdot f_i\}_{i=1}^M).$$

Method

- Fusion

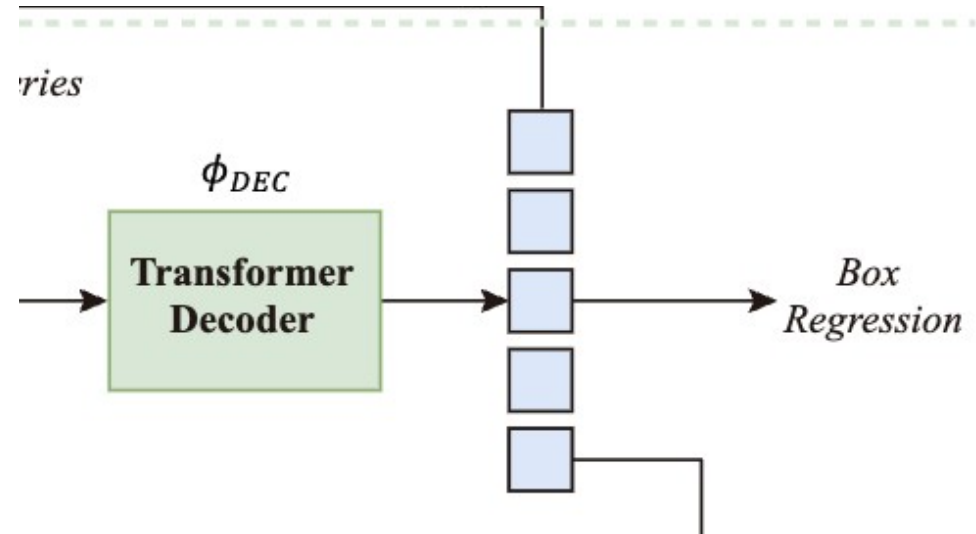


$$q_j = f_i \oplus t'_j$$

Method

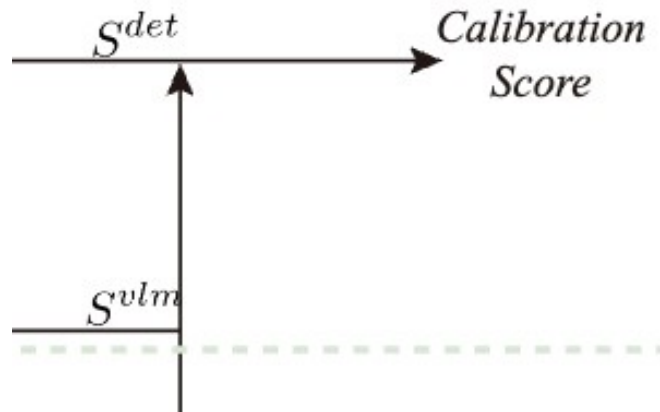
- Transformer Decoder
 - The refined features of object queries
- Box Regression
 - L1 loss + Giou loss

$$\Phi_{\text{bbox}}(f_j) = b_j = (x_j, y_j, w_j, h_j)$$



Method

- Calibration Score
 - Balancing detector and VLM contributions equally.

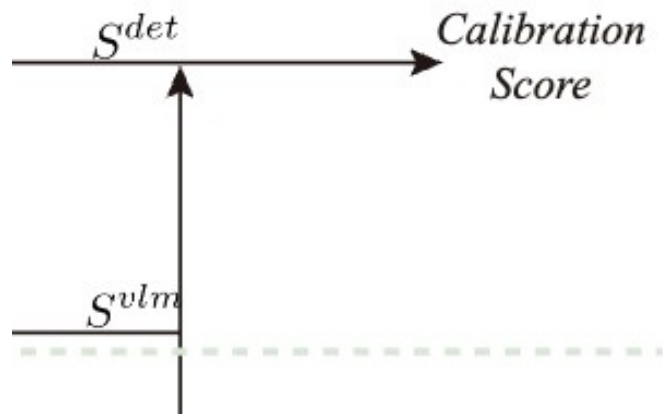


$$s_j^{vlm} = \mathcal{T}_{CLS} \cdot \Phi_{\text{pooling}}(b_j)$$

$$s_j^{det} = q_j \cdot T_{cls}$$

Method

- Calibration Score
 - Balancing detector and VLM contributions equally.

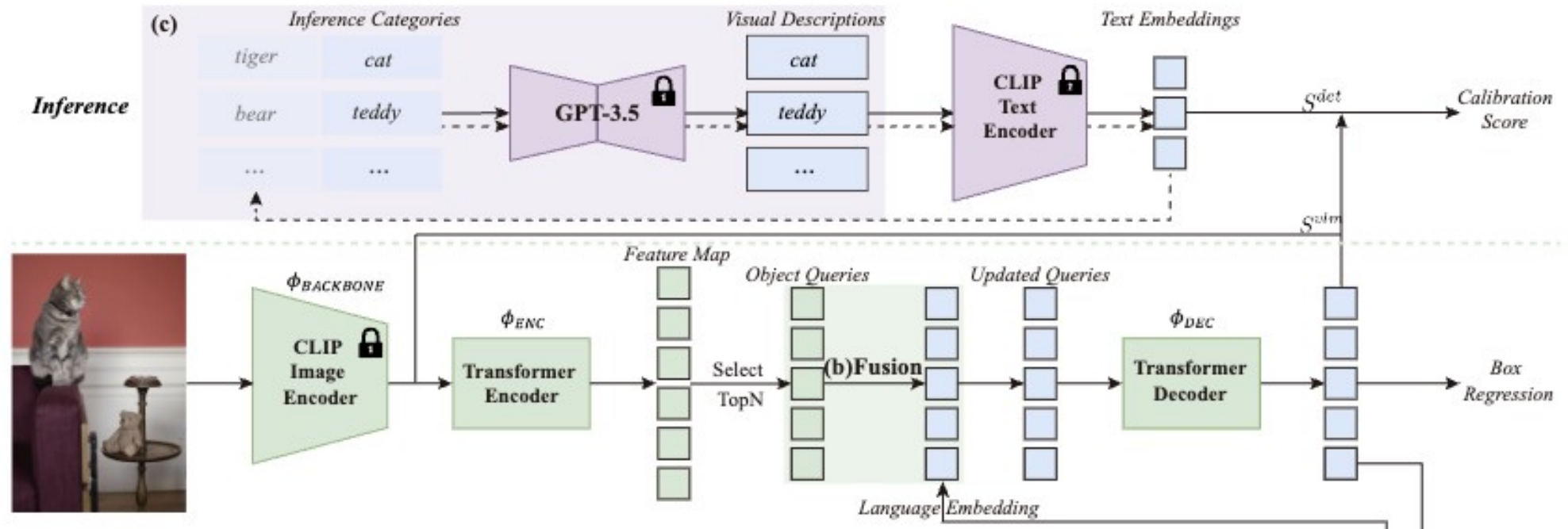


$$S_c^{cal} = \begin{cases} S_c^{vlm^\alpha} \cdot S_c^{det^{(1-\alpha)}} & \text{if } c \in \mathcal{C}_B \\ S_c^{vlm^\beta} \cdot S_c^{det^{(1-\beta)}} & \text{if } c \in \mathcal{C}_N \end{cases}$$

set $\alpha, \beta = (0.0, 0.25)$

Method

- Final result
: Calibration Score + Box Regression



Experiments

- LVIS open-vocabulary detection

- Image-level Dataset : Only labels
- AP_r : LVIS Dataset(Rare cls)
- AP : Average Precision

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Method	Pretrained Model	Detector Backbone	Backbone Size	Image-level Dataset	AP_r	AP
VL-PLM [44]	ViT-B/32	R-50	26M	IN-L	17.2 [†]	27.0 [†]
OV-DETR [40]	ViT-B/32	R-50	26M	✗	17.4 [†]	26.6 [†]
DetPro-Cascade [6]	ViT-B/32	R-50	26M	✗	21.7	30.5
Rasheed [26]	ViT-B/32	R-50	26M	IN-L	21.1 [†]	25.9 [†]
PromptDet [7]	ViT-B/32	R-50	26M	LAION-novel	21.4 [†]	25.3 [†]
OADP [33]	ViT-B/32	R-50	26M	✗	21.9	28.7
RegionCLIP [45]	R-50x4	R-50x4	87M	CC3M	22.0 [†]	32.3 [†]
CORA [36]	R-50x4	R-50x4	87M	✗	22.2	-
BARON [35]	ViT-B/32	R-50	26M	CC3M	23.2	29.5
CondHead [34]	R-50x4	R-50x4	87M	CC3M	25.1	33.7
Detic-CN2 [46]	ViT-B/32	R-50	26M	IN-L	24.6 [†]	32.4 [†]
ViLD-Ens [9]	ViT-B/32	R-50	26M	✗	16.7	27.8
F-VLM [15]	R-50x64	R-50x64	420M	✗	32.8 [†]	34.9 [†]
OWL-ViT [22]	ViT-L/14	ViT-L/14	306M	✗	25.6	34.7
RO-ViT [14]	ViT-B/16	ViT-B/16	86M	ALIGN*	28.4	31.9
RO-ViT [14]	ViT-L/16	ViT-L/16	303M	ALIGN*	33.6	36.2
CFM-ViT [13]	ViT-B/16	ViT-B/16	86M	ALIGN*	29.6	33.8
CFM-ViT [13]	ViT-L/16	ViT-L/16	303M	ALIGN*	35.6	38.5
ours	ConVNext-L	ConVNext-L	196M	✗	43.4	41.3

Experiments

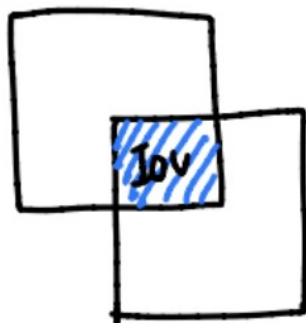
- LVIS zero-shot detection
 - Without relying on additional large-scale image-text datasets.

Method	Detector Backbone	Datasets	AP _r	AP
GLIP-L [16]	Swin-L	O365,GoldG,Cap4M	17.1	26.9
GroundingDINO [17]	Swin-L	O365,GoldG,OI,Cap4M,COCO,RefC	22.0	32.3
DetCLIP [§] [39]	Swin-L	O365,GoldG,YFCC1M	27.6	31.2
DetCLIPv2 [§] [38]	Swin-L	O365,GoldG,CC15M	33.3	36.6
OWL-ViT [22]	ViT-L/14	O365,VG-dedup	31.2	34.6
OWL-ST [20]	ViT-L/14	O365,VG-dedup	34.9	33.5
ours	ConVNext-L	O365,VG-dedup	37.8	35.4

Experiments

- Cross-dataset Transfer
 - Train : LVIS
 - Val : COCO / Objects 365
 - AP_{50,75} : IoU 50%,70% -> considered correct

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



Method	Backbone	Parameters	COCO			Objects365		
			AP	AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅
ViLD [9]	RN50	26M	36.6	55.6	39.8	11.8	18.2	12.6
DetPro [6]	RN50	26M	34.9	53.8	37.4	12.1	18.8	12.9
F-VLM [15]	RN50	38M	32.5	53.1	34.6	11.9	19.2	12.6
BARON [35]	RN50	26M	36.2	55.7	39.1	13.6	21.0	14.5
CoDet [19]	EVA02-L	304M	39.1	57.0	42.3	14.2	20.5	15.3
CFM [13]	ViT-L/16	303M	-	-	-	18.7	28.9	20.3
ours	ConvNext-L	196M	42.8	57.6	46.9	21.9	30.0	23.5

Conclusion

- Language Model Instruction (LaMI) 전략
 - GPT와 T5를 활용하여 시각 개념 간의 유사성을 추출
 - 음성(negative) 범주를 샘플링하여 과적합을 방지하고, 보다 일반화된 객체 특징을 학습
- OV-LVIS 기준 rare class box AP에서 기존 최고 모델 대비 +7.8 AP 향상
- end-to-end 구조 유지
- ConvNeXt-L 기반에 한정 \rightarrow ViT 계열 VLM 적용은 향후 연구로 제시