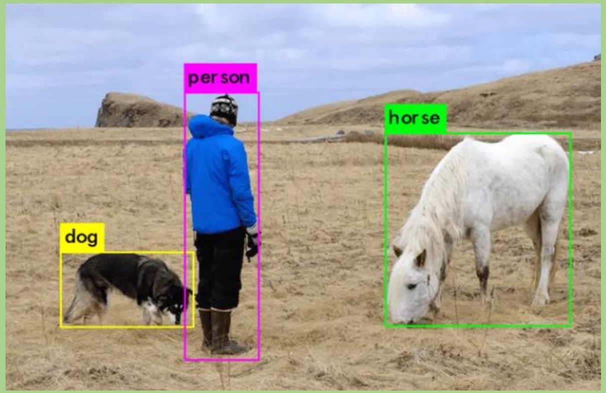
VLP for Object Detection

Pengchuan Zhang

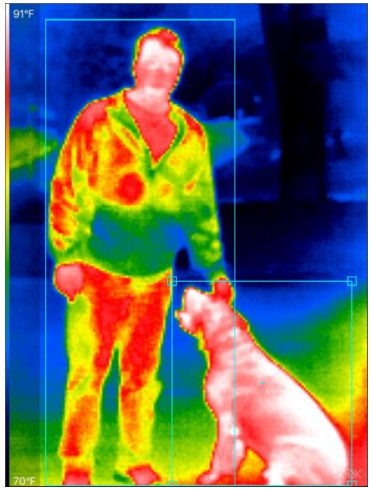
Recent Advances in Vision-and-Language Pre-training

Object Detection in the wild

(https://public.roboflow.com/object-detection)

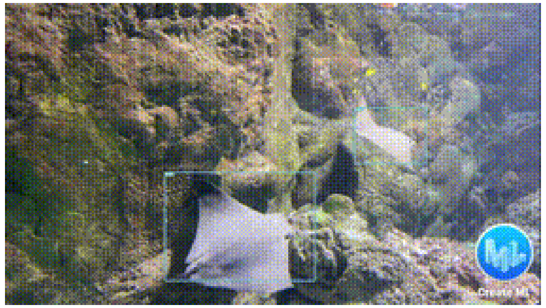


MS-COCO



Thermal Dogs and People

Wildfire Smoke



Aquarium (fish. jellyfish. penguin. puffin. shark. starfish. stingray)

Main challenges

1) **Open vocabulary:** unseen concepts
2) **zero/few-shot transfer**: zero or very few task-specific annotations
3) **Domain adaption**: data (images) in various domains/environments

Vision-Language Pre-training for Object Detection

1) Object detection as a vision-language grounding task
2) Pre-train the grounding model with both **region-level annotated data** (detection, grounding) and **weakly image-text paired data**

An overview of existing works

	VLP for region-level classification	VLP for end-to-end detection
Generic box proposals	ViLD (ICLR2022) RegionCLIP (CVPR2022)	X-Detr (Arxiv)
Text-guided box proposals		MDetr (ICCV2021) GLIP (CVPR2022) GLIPv2 (Arxiv) FIBER (Arxiv) FindIt (Arxiv)

Related topics

- **Zero-shot object detection**: Bansal et al (ECCV2018), Rahman et al (AAAI2020), …
- **Open-vocabulary object detection**: OV-Det (CVPR2021)
- **Phrase grounding, Referring Expression Comprehension**
- **General Purpose Vision System**: UniT, GPV, Florence, Gato, CoCa

An overview of existing works

	VLP for region-level classification	VLP for end-to-end detection
Generic box proposals	ViLD (ICLR2022) RegionCLIP (CVPR2022)	X-Detr (Arxiv)
Text-guided box proposals		**MDetr (ICCV2021)** **GLIP (CVPR2022)** GLIPv2 (Arxiv) FIBER (Arxiv) FindIt (Arxiv)

Related topics

- **Zero-shot object detection**: Bansal et al (ECCV2018), Rahman et al (AAAI2020), …
- **Open-vocabulary object detection**: OV-Det (CVPR2021)
- **Phrase grounding, Referring Expression Comprehension**
- **General Purpose Vision System**: UniT, GPV, Florence, Gato, CoCa

NYU Center for Data Science

FACEBOOK AI

MDETR - Modulated Detection for End-to-End Multi-Modal Understanding


Aishwarya Kamath
NYU


Mannat Singh
FAIR


Yann LeCun
NYU/FAIR


Gabriel Synnaeve
FAIR


Ishan Misra
FAIR


Nicolas Carion
NYU

2021 ICCV OCTOBER 11-17 VIRTUAL

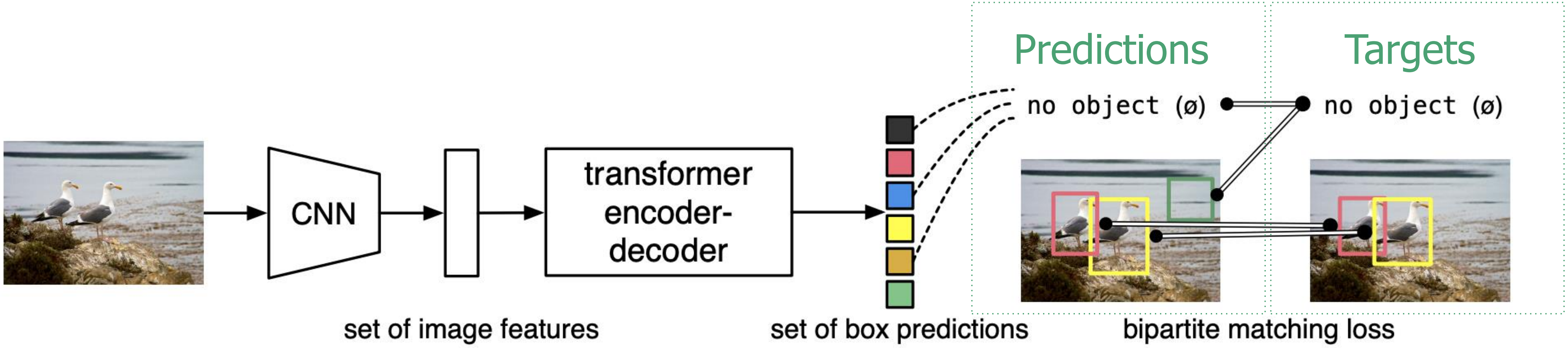
What is “modulated detection"?

- Free-form text conditioned detection
- End-to-end training
- Leverage compositionality of language

Output of MDETR for the query “A pink elephant”

DETR - Detection transformer

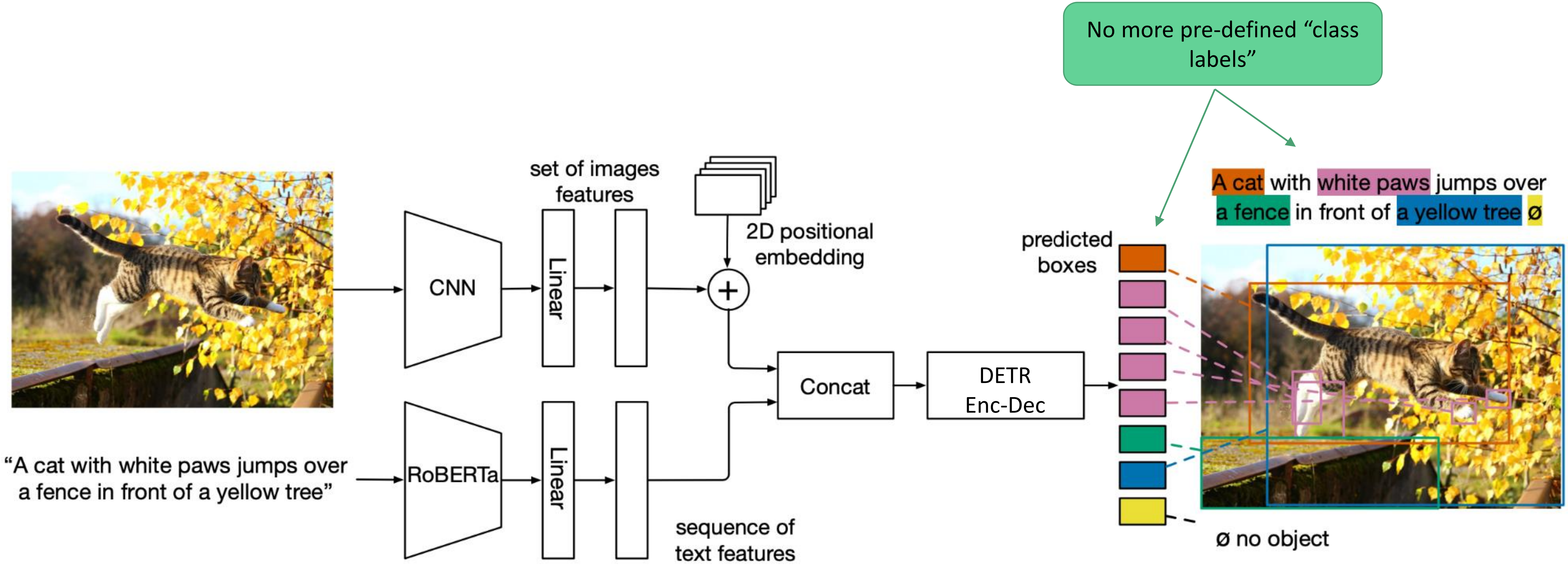
- End-to-end detection
- Encoder-decoder architecture



Loss = Box Regression + Label Prediction

Image credit: Carion, Nicolas, et al. "End-to-end object detection with transformers." *European Conference on Computer Vision*. Springer, 2020

MDETR: Architecture



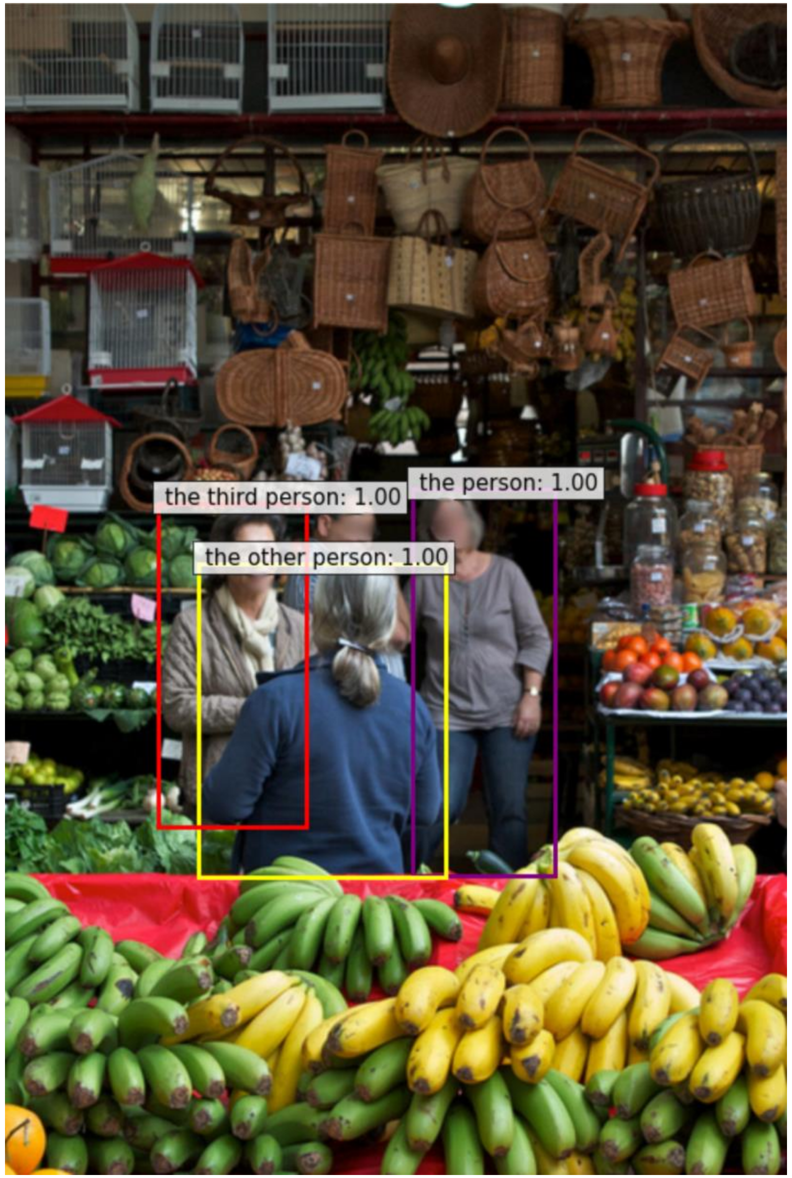
Loss = Box Regression + Soft Token Prediction

MDETR: Pre-training

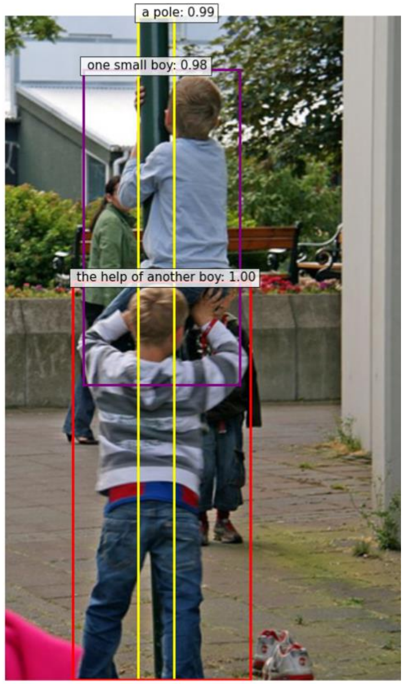
- Flicker30k-Entities, RefCOCO, RefCOCO+, RefCOCOg, Visual Genome Dense Captions, GQA with boxes
- Results in 1.3m aligned image-text pairs with box annotations (only 0.2m unique images)

Toy example:

"the person in the grey shirt with a watch on their wrist. the other person wearing a blue sweater. the third person in a gray coat and scarf."

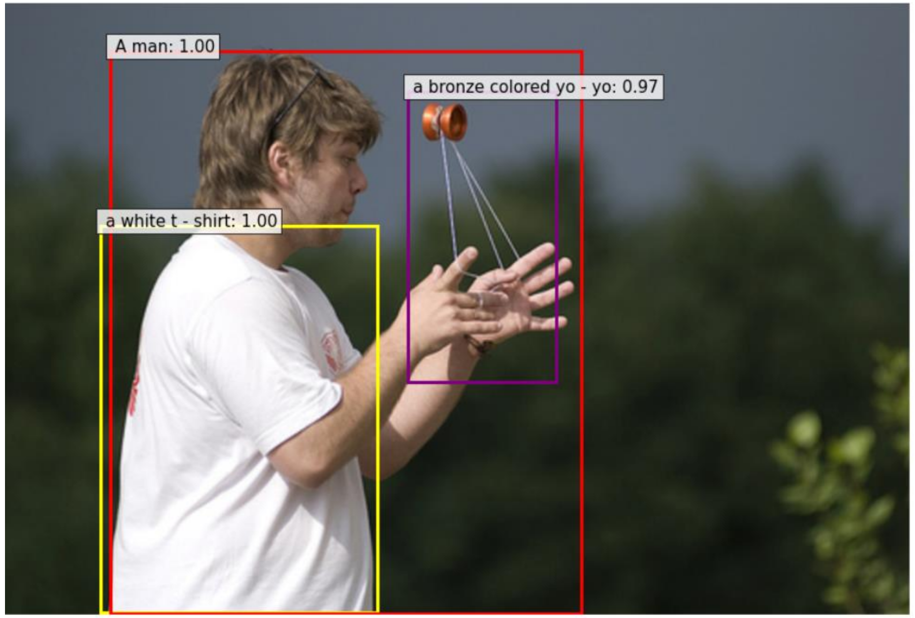


Phrase grounding on Flickr30k



"One small boy climbing a pole with the help of another boy on the ground"

"A man talking on his cellphone next to a jewelry store"

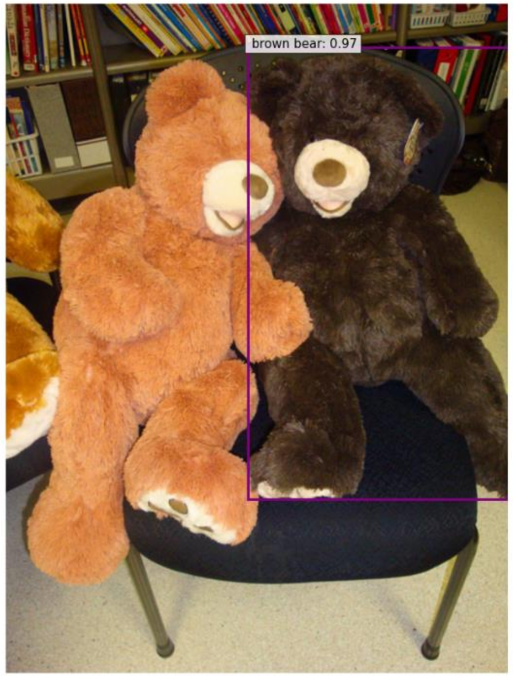


"A man in a white t-shirt does a trick with a bronze colored yo-yo"

Phrase grounding on Flickr30k - Quantitative results

Method	Val			Test		
	R@1	R@5	R@10	R@1	R@5	R@10
			ANY-BOX-PROTOCOL			
BAN [21]	-	-	-	69.7	84.2	86.4
VisualBert[25]	68.1	84.0	86.2	-	-	-
VisualBert†[25]	70.4	84.5	86.3	71.3	85.0	86.5
MDETR-R101	78.9	88.8	90.8	-	-	-
MDETR-R101†*	**82.5**	**92.9**	**94.9**	**83.4**	**93.5**	**95.3**
MDETR-ENB3†*	**82.9**	**93.2**	**95.2**	**84.0**	**93.8**	**95.6**
MDETR-ENB5†*	**83.6**	**93.4**	**95.1**	**84.3**	**93.9**	**95.8**
			MERGED-BOXES-PROTOCOL			
CITE [43]	-	-	-	61.9	-	-
FAOG [66]	-	-	-	68.7	-	-
SimNet-CCA [45]	-	-	-	71.9	-	-
MDETR-R101†*	**82.4**	**92.6**	**94.5**	**83.3**	**92.1**	**93.8**

Referring expressions



“brown bear”

RefCOCO

“zebra facing away”

RefCOCO+

“The man in the red shirt carrying baseball bats”

RefCOCOg

Results for referring expressions on RefCOCO

Method	Detection backbone	Pre-training image data	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val	test
MAttNet[69]	R101	None	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
ViLBERT[34]	R101	CC (3.3M)	-	-	-	72.34	78.52	62.61	-	-
VL-BERT_L [54]	R101	CC (3.3M)	-	-	-	72.59	78.57	62.30	-	-
UNITER_L[6]*	R101	CC, SBU, COCO, VG (4.6M)	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA_L[9]*	R101	CC, SBU, COCO, VG (4.6M)	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
ERNIE-ViL_L[68]	R101	CC, SBU (4.3M)	-	-	-	75.95	82.07	66.88	-	-
MDETR	R101	COCO, VG, Flickr30k (200k)	**86.75**	**89.64**	**81.47**	**79.52**	**84.72**	**69.76**	**81.64**	**80.98**
MDETR	ENB3	COCO, VG, Flickr30k (200k)	**87.51**	**90.38**	**82.90**	**81.13**	**85.52**	**72.96**	**83.35**	**83.45**

Few-shot detection on LVIS

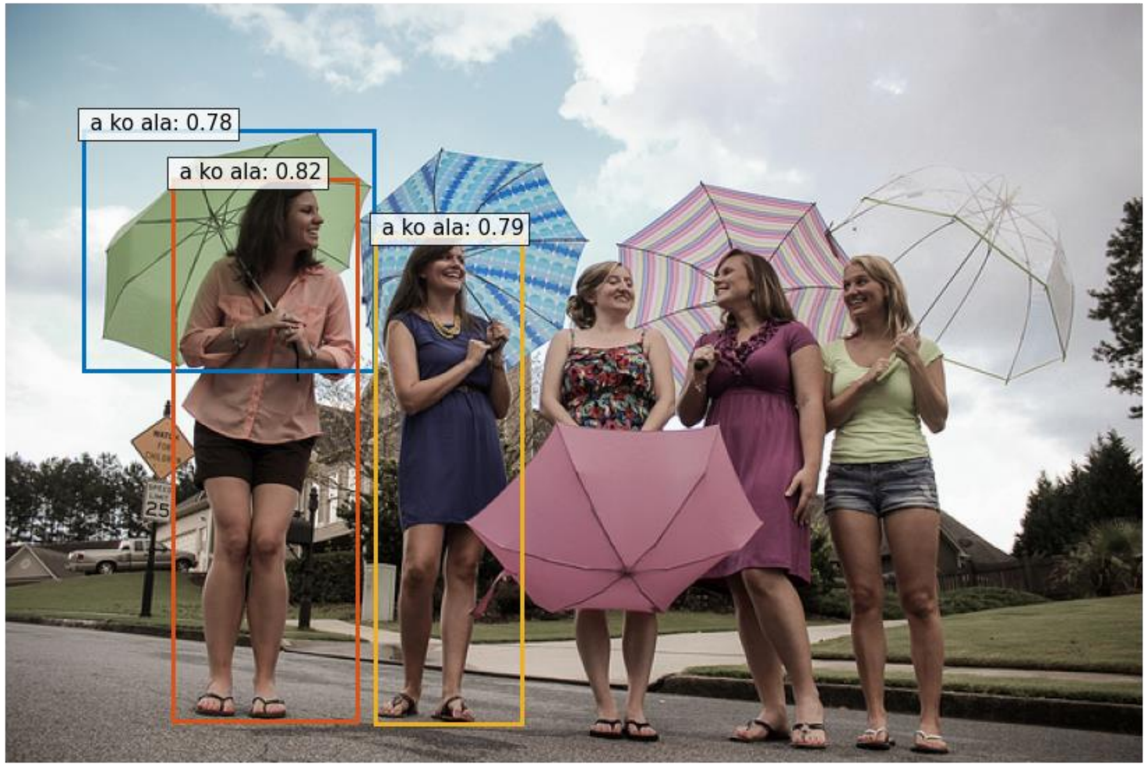
- Performs well with as low as 1 sample/class
- Due to overlaps between COCO/LVIS/… , we report results on the subset of 5k validation images (mini-val) that our model has never seen during training.

Method	Data	AP	AP50	AP_r	AP_c	AP_f
Mask R-CNN	100%	33.3	51.1	26.3	34.0	33.9
DETR	1%	4.2	7.0	1.9	1.1	7.3
DETR	10%	13.7	21.7	4.1	13.2	15.9
DETR	100%	17.8	27.5	3.2	12.9	24.8
MDETR	1%	16.7	25.8	11.2	14.6	19.5
MDETR	10%	24.2	38.0	20.9	24.9	24.3
MDETR	100%	22.5	35.2	7.4	22.7	25.0

Limits of MDETR

- Not for zero-shot detection

Training data has no “negative examples” - i.e. when the text does not correspond to any object in the image. Model will always try to find something (usually salient objects in the image)



- Pre-training data does not scale up

All pre-training data are aligned image-text pairs with box annotations

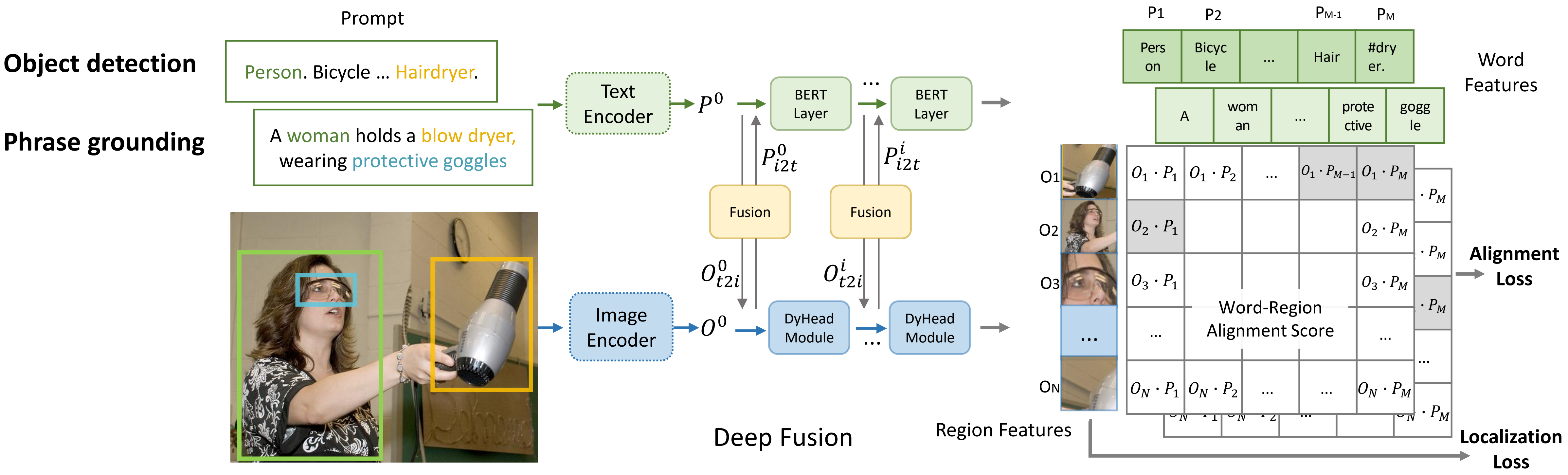
GLIP: Grounded Language-Image Pre-Training

Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, Jianfeng Gao

UCLA Microsoft UNIVERSITY of WASHINGTON WISCONSIN UNIVERSITY OF WISCONSIN-MADISON idea INTERNATIONAL DIGITAL ECONOMY ACADEMY

Work done at Microsoft

Unify Object Detection and Phrase Grounding



Phrase grounding data: 0.08M images, 0.8M image-text-boxes triplets

Object detection data: Objects365 + OpenImages + VisualGenome, 2.5M image-text-boxes triplets

Self-training on massive image-text paired data

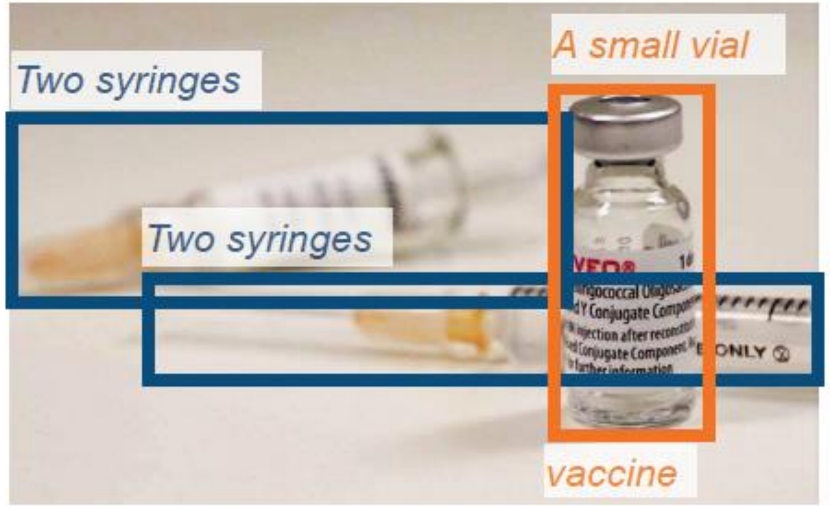
person battles with person in the production sedans

Two syringes and a small vial of vaccine.

playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

Self-training on massive image-text paired data

person battles with person in the production sedans



Two syringes and a small vial of vaccine.

playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

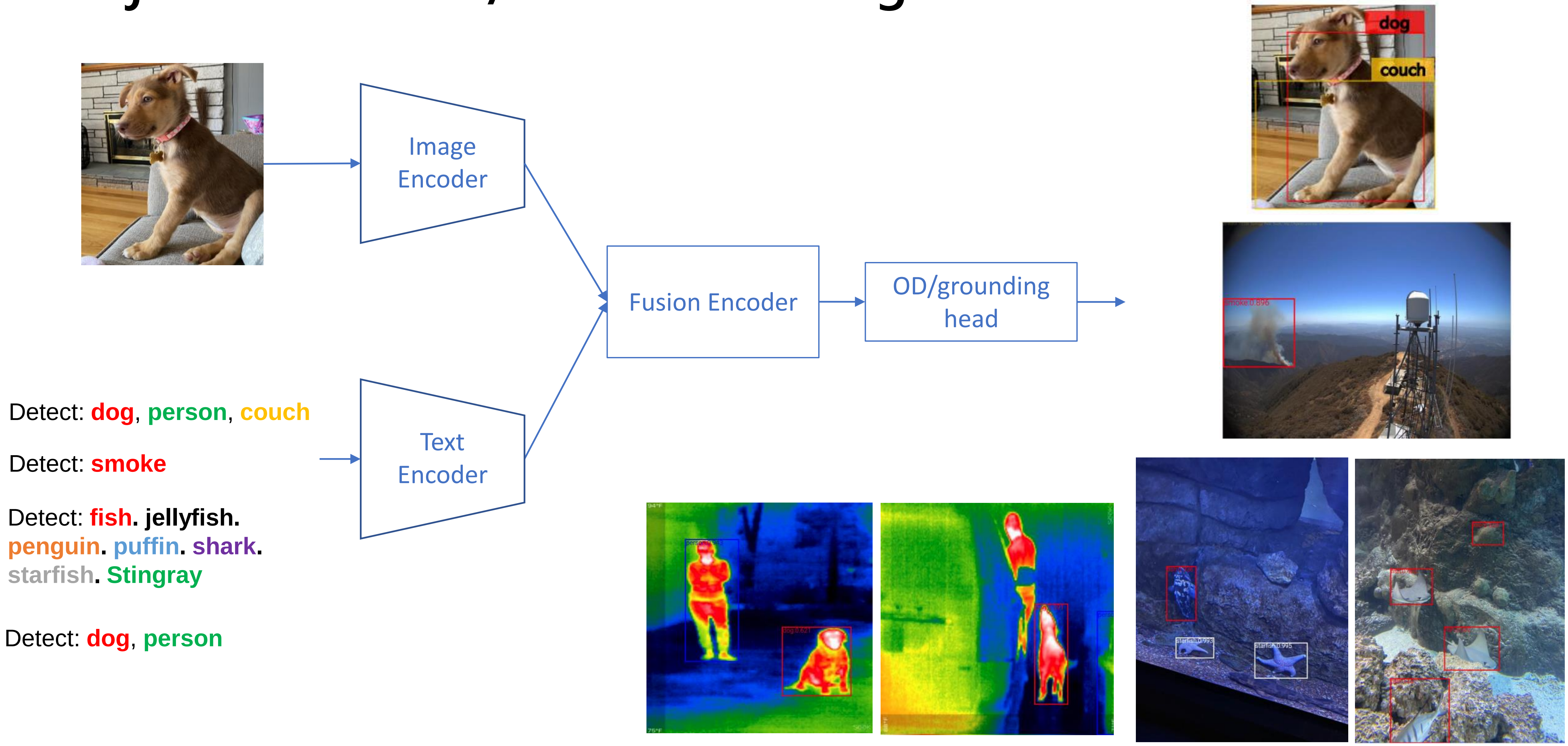
From 24M image-text paired data:

- 78.1M high-confidence (> 0.5) phrase-box pseudo annotations
- 58.4M unique noun phrases

Compared with traditional object detection self-training:

- Visual concepts are significantly scaled up, from ~2k to ~60m; massive visual attributes and relationships
- More accurate bounding boxes thanks to the text clues

Object Detection / Text Grounding in the Wild



Results on Benchmarks

	Backbone	COCO 2017 val Zero-Shot / Fine-Tune	LVIS Minival APr
MDETR	R101	-	20.9
Mask RCNN	R101	-	26.3
Faster RCNN	R101	- / 42.0	-
DyHead-T	Swin-T	- / 49.7	-
GLIP-T	Swin-T	46.3 / 54.9	20.8
GLIP-L	Swin-L	49.8 **/** 61.5*	**28.2**

Zero-shot

Fine-tuned/supervised

Zero-shot GLIP rivales with **supervised** models (No COCO images seen during pre-training)

- COCO: GLIP-T (46.3 AP, zero-shot) v.s. Faster RCNN (42.0 AP, supervised)
- LVIS: GLIP-T (20.8 APr, zero-shot) v.s. MDETR (20.9 APr, supervised)

Strong **fine-tuning** performance

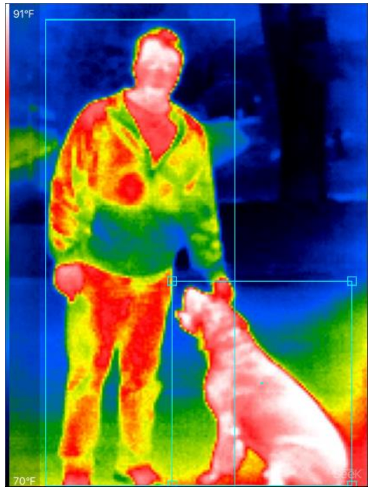
- GLIP-T outperforms DyHead-T (same backbone) by 5 AP on COCO
- GLIP-L achieves 61.5 AP on COCO (SOTA when released)

* 61.5 is tested on COCO test-dev with multi-scale testing; this version of GLIP-L is pre-trained with COCO

Object Detection in the Wild (13 real world detection tasks)

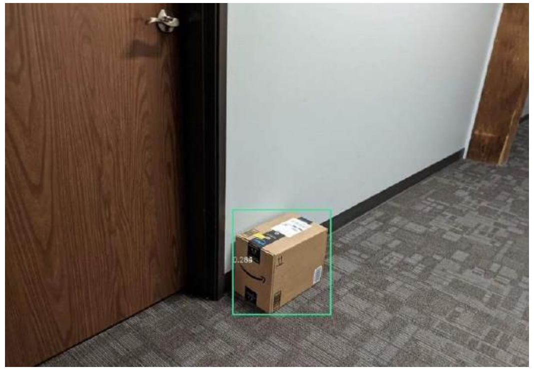
Wildfire Smoke Dataset

Aquarium Dataset (fish. jellyfish. penguin. puffin. shark. starfish. stingray**)**



Thermal Dogs and People Dataset

Mask Wearing

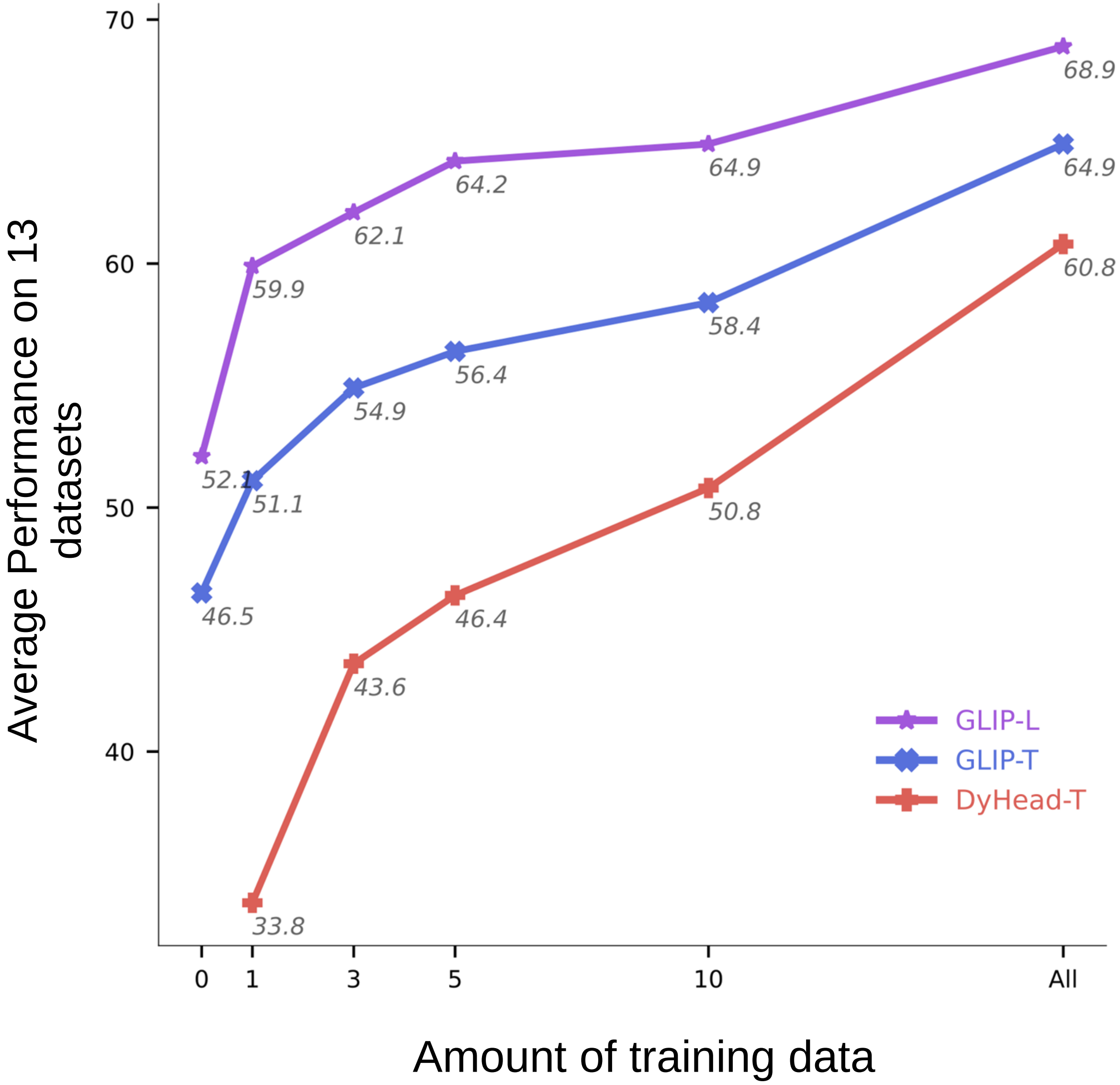


Packages

Pistols

Potholes

Object Detection in the Wild : Data Efficiency

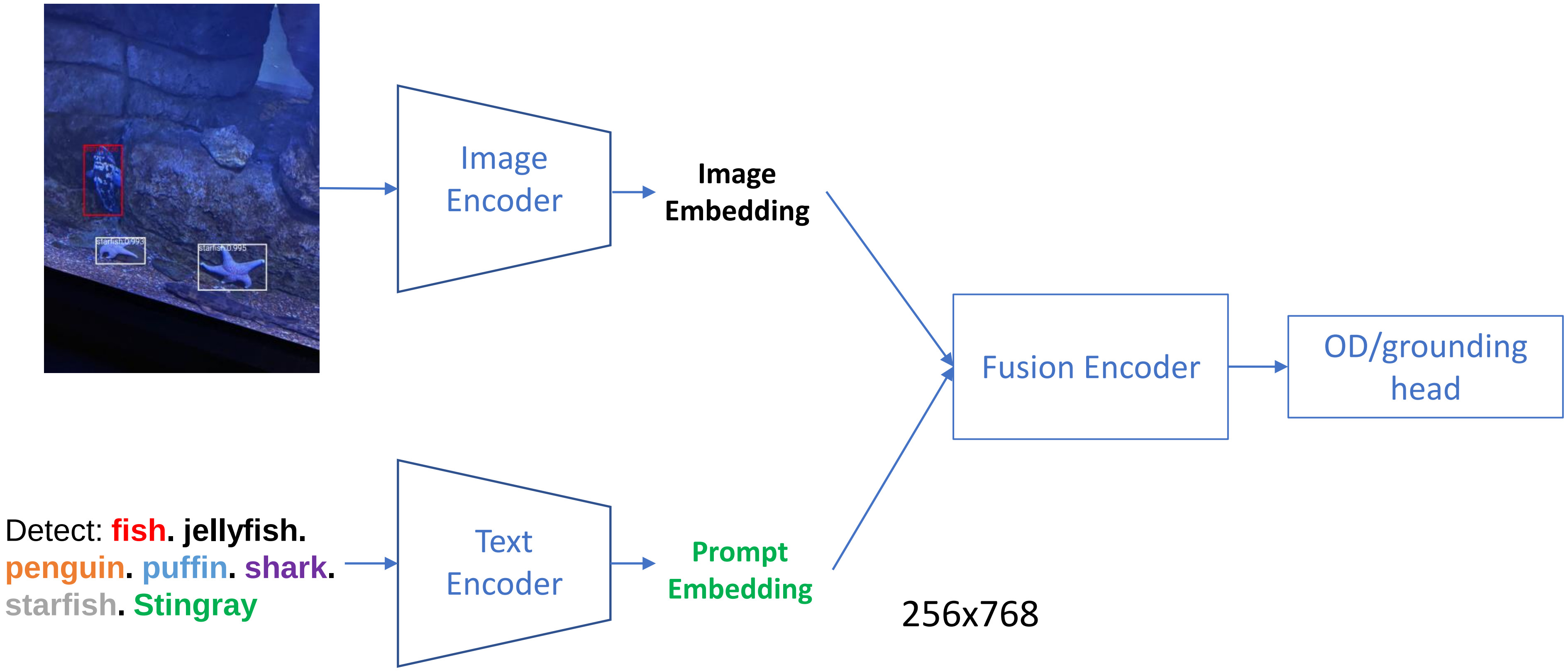


0-shot GLIP-T ~= 5-shot DyHead-T

1-shot GLIP-T / 0-shot GLIP-L ~= 10-shot DyHead-T

1-shot GLIP-L ~= Fully-supervised DyHead-T

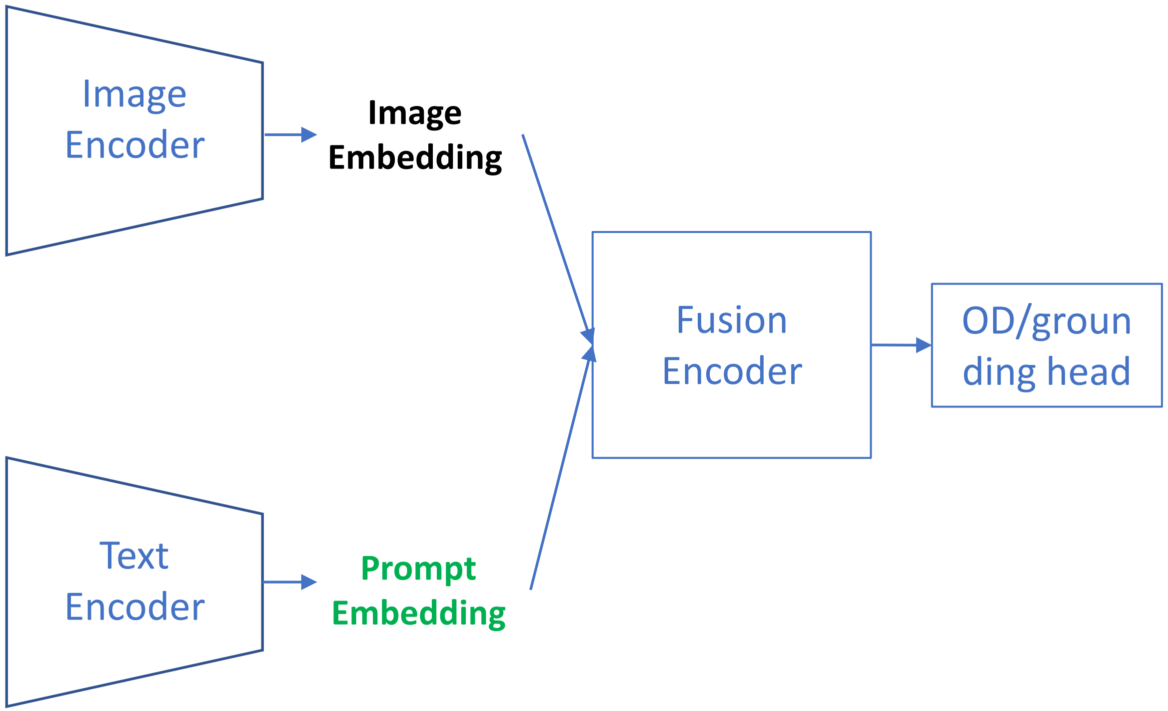
One Model for All Tasks: Prompt Tuning



COCO	PascalVOC	AerialDrone	Aquarium	Rabbits	EgoHands	Mushrooms	Packages	Raccoon	Shellfish	Vehicles	Pistols	Pothole	Thermal
58.8	72.9/86.7	23.0	51.8	72.0	75.8	88.1	75.2	69.5	73.6	72.1	73.7	53.5	81.4

Table 1. AP (evaluated with COCO-API) of one GLIP-L model on 14 tasks with prompt tuning – tuning only the embedding of each task's prompt. Thus, one set of GLIP model weights can simultaneously serve many tasks. For PascalVOC (2012 Val), we report AP/AP50.

Prompt Tuning is Comparable with Full-model Finetuning



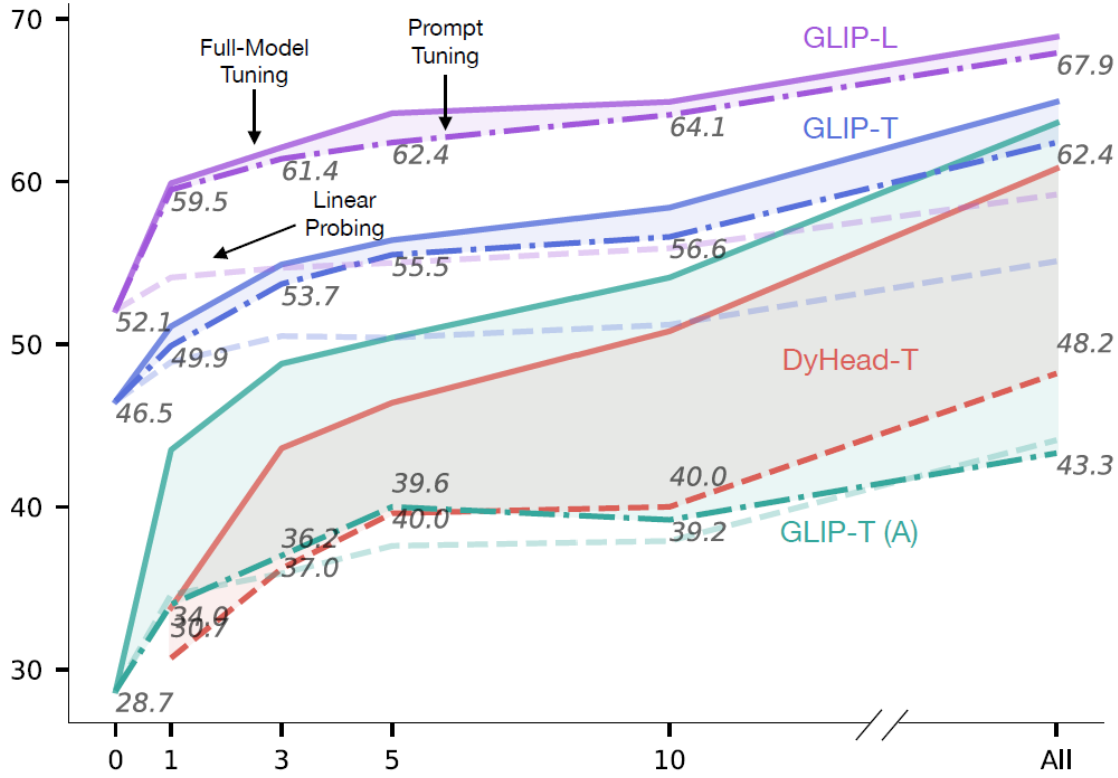


Figure 5. Effectiveness of prompt tuning. Solid lines are full-model tuning performance; dashed lines are prompt/linear probing performance. By only tuning the prompt embeddings, GLIP-T and GLIP-L can achieve performance close to full-model tuning, allowing for efficient deployment.

An overview of existing works

	VLP for region-level classification	VLP for end-to-end detection
Generic box proposals	ViLD (ICLR2022) RegionCLIP (CVPR2022)	X-Detr (Arxiv)
Text-guided box proposals		MDetr (ICCV2021) GLIP (CVPR2022) GLIPv2 (Arxiv) FIBER (Arxiv) FindIt (Arxiv)

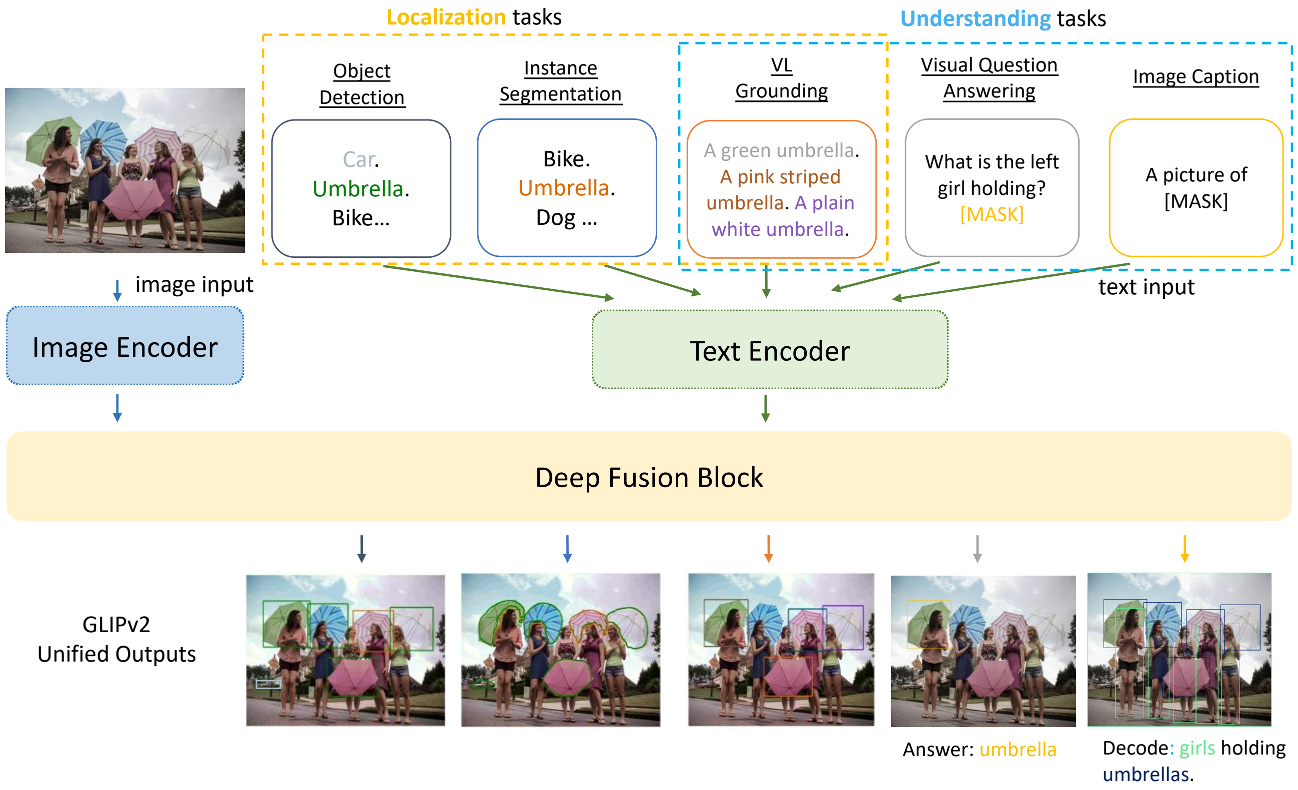
Related topics

- **Zero-shot object detection**: Bansal et al (ECCV2018), Rahman et al (AAAI2020), …
- **Open-vocabulary object detection**: OV-Det (CVPR2021)
- **Phrase grounding, Referring Expression Comprehension**
- **General Purpose Vision System**: UniT, GPV, Florence, Gato, CoCa

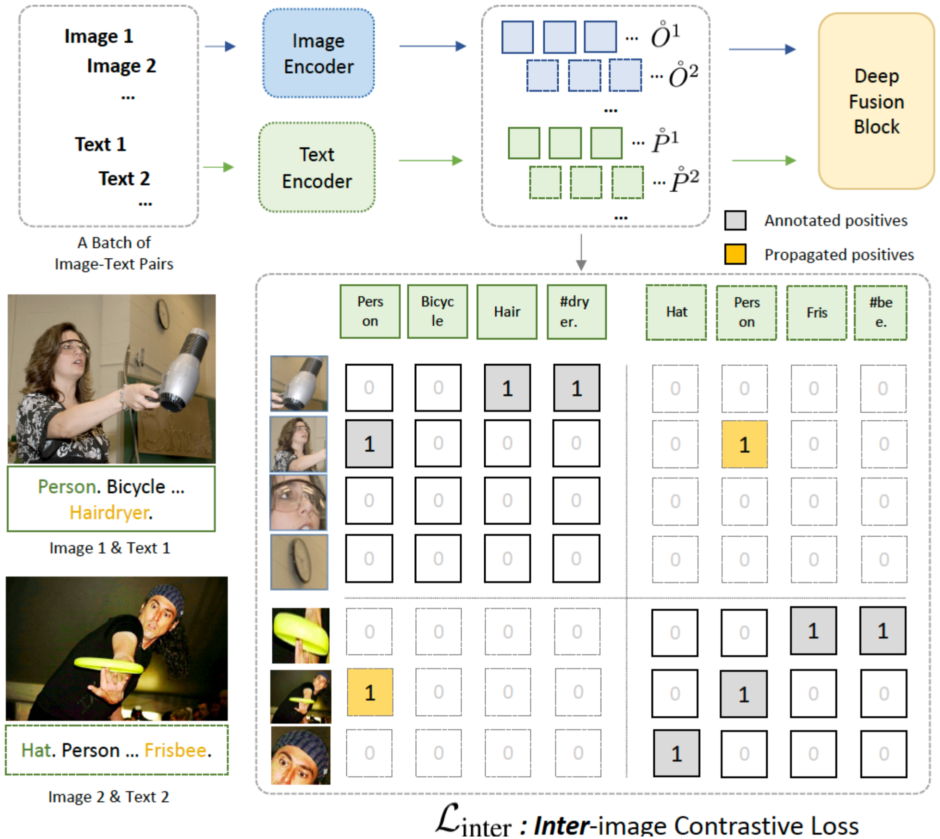
GLIPv2: Unifying Localization and Vision-Language Understanding

Haotian Zhang*, Pengchuan Zhang*, et al, Arxiv 2022

Localization + VL understanding = grounded VL understanding



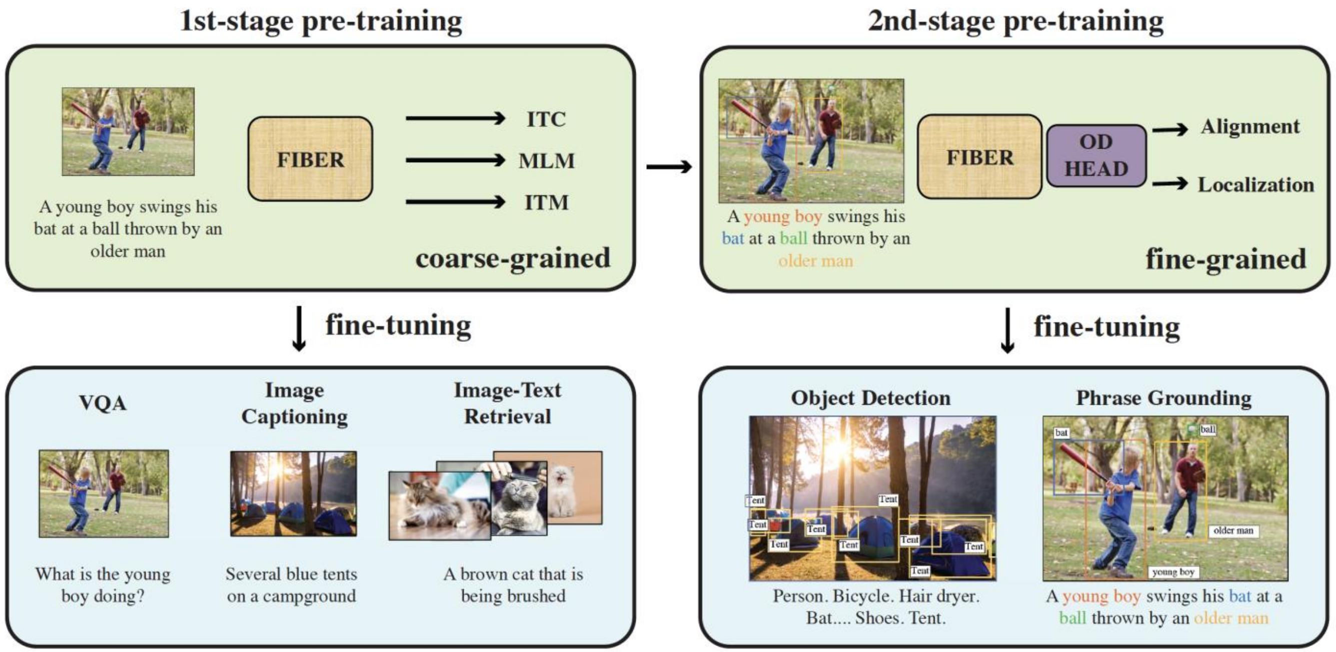
Inter-image region-word level contrastive loss



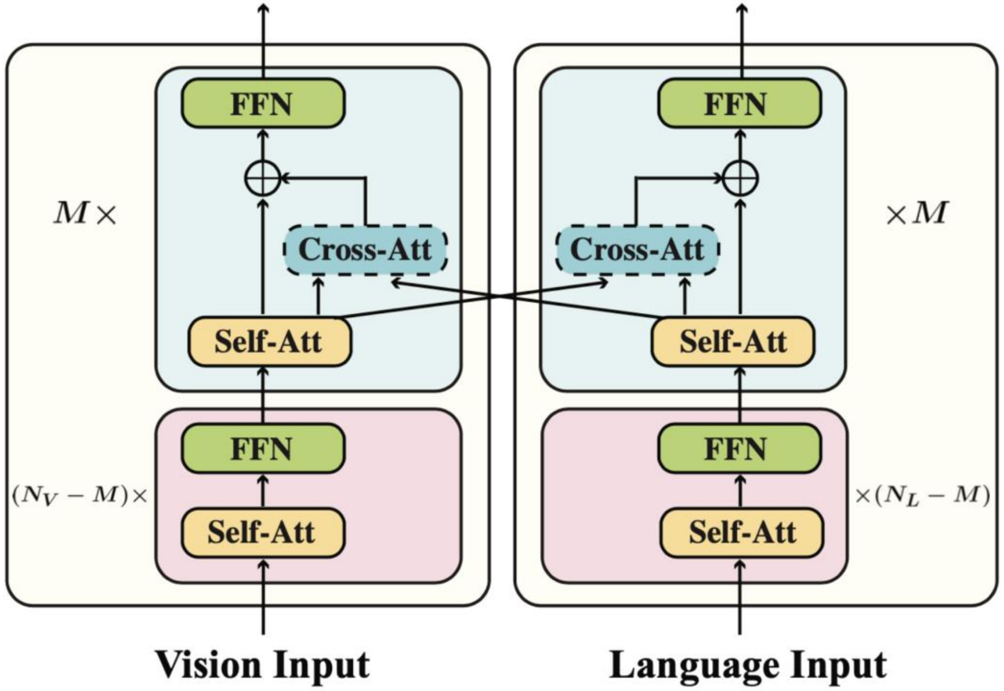
FIBER: Coarse-to-Fine Vision-Language Pre-training with Fusion in the Backbone

Zi-Yi Dou*, Aishwarya Kamath*, Zhe Gan*, et al, Arxiv 2022

Two-stage coarse-to-fine pre-training framework



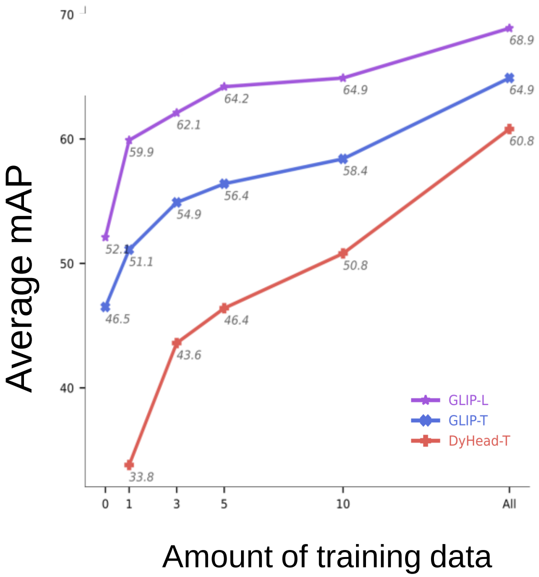
Fusion **I**n the **B**ackbone Transform**ER** (FIBER)



Several Future Directions

1) Large scale region-aware pre-training for object detection
   - How to better use weakly supervised data, e.g., image-text pairs
   - Scalable object detection model architecture
2) Zero-shot and few-shot object detection
   - More data-efficient
   - More training efficient, e.g., full-finetune -> prompt tuning
   - More efficient/compact model on device
3) Computer vision in the wild
   - More tasks: segmentation, action recognition, human-object interaction, …
   - More modalities: video, audio, IMU, …
   - A true multimodal foundation model

Wildfire Smoke



Thanks!