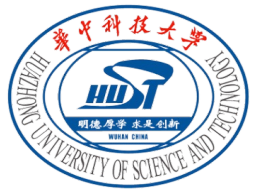
YOLO-World

Tianheng Cheng[2,3,*], **Lin Song**[1,2,*,✉], **Yixiao Ge**[1,2,🌟], **Wenyu Liu**[3], **Xinggang Wang**[3,✉], **Ying Shan**[1,2]

[1] Tencent AI Lab, [2] ARC Lab, Tencent PCG, [3] Huazhong University of Science and Technology

[*] Equal contribution 🌟 Project lead ✉ Corresponding author

thch@hust.edu.cn



程天恒

本科、博士就读于华中科技大学

- 主要研究方向为计算机视觉、通用物体检测与分割、自动驾驶、多模态视觉感知与理解

内容

- 开集物体检测
- **YOLO-World 模型与训练**
- **YOLO-World 微调与应用**
- 多模态大模型时代的物体检测

1. 开集物体检测

Traditional Object Detection

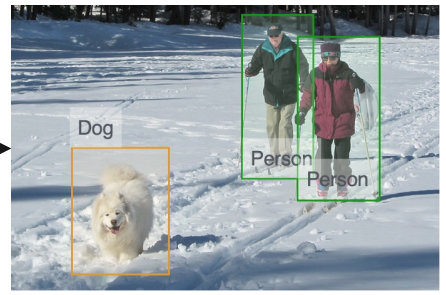
传统物体检测方法在闭集物体检测上已经取得了显著进展！

图像或视频

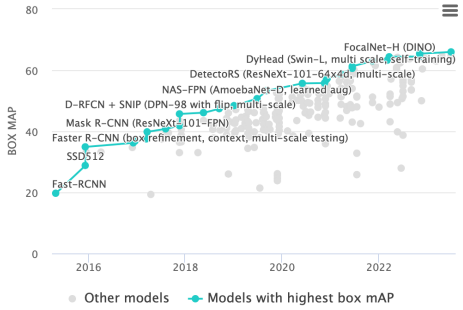
Object Detector

Faster R-CNN, YOLO, CenterNet,

FCOS, DETR, Deformable DETR



分类 & 定位



66.0 COCO AP

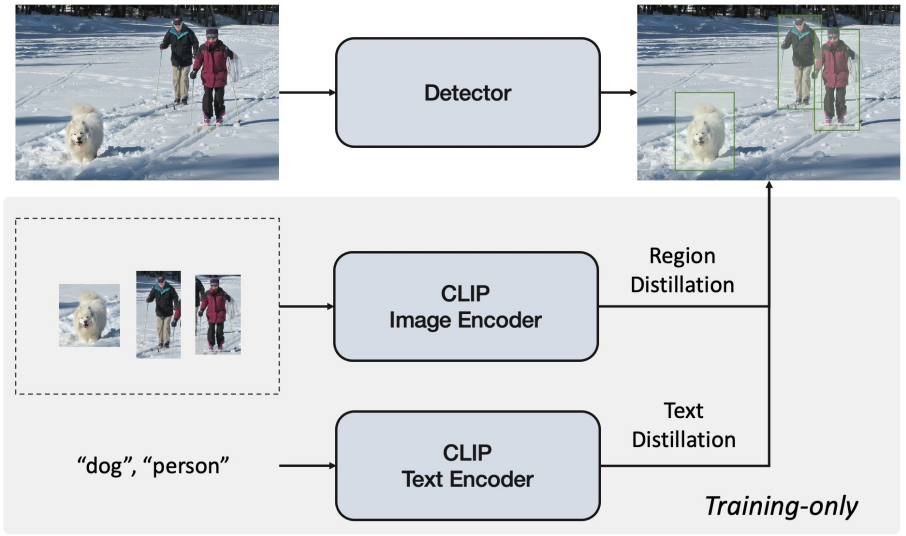
然而，传统物体检测方法**无法检测训练集中未出现的类别**

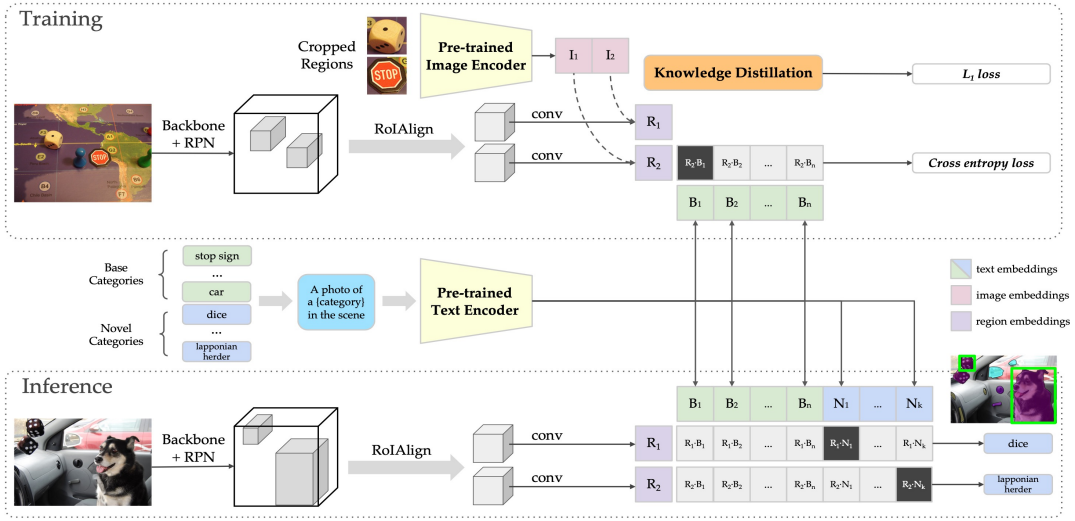
增加类别则需要重新训练，不同数据/任务需要需要训练不同的检测器，成本高！

Open-Vocabulary Object Detection

➢ 先前工作 - CLIP 蒸馏

蒸馏预训练的CLIP，在部分基类上训练，实现开集物体检测



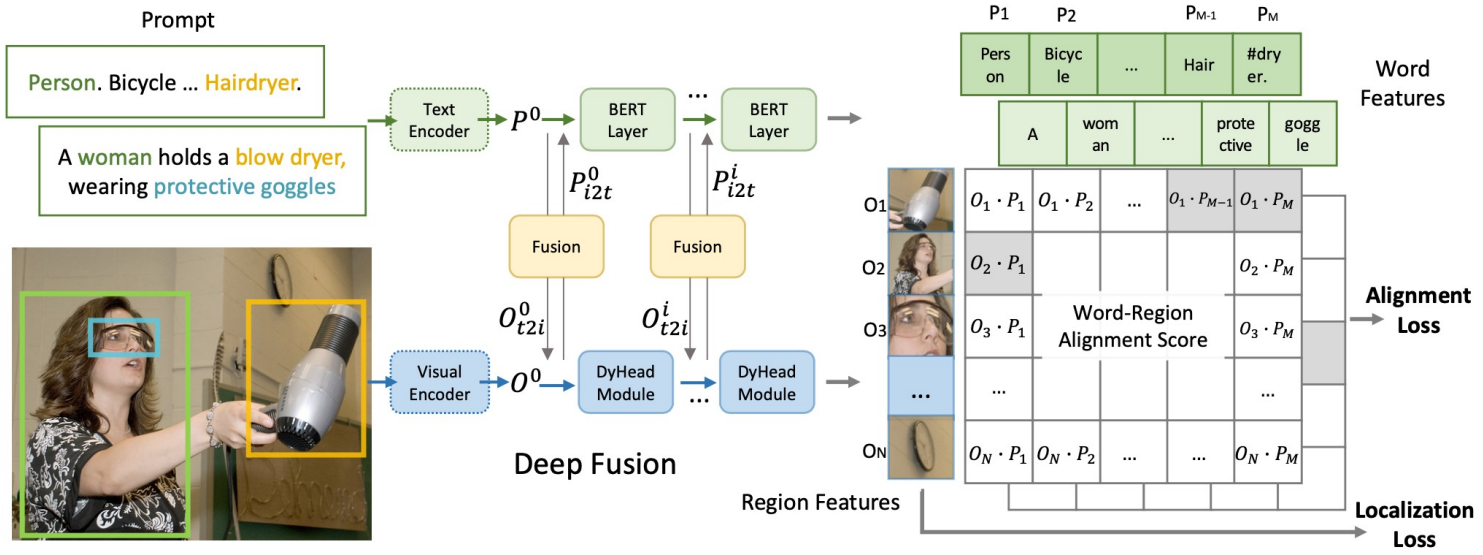


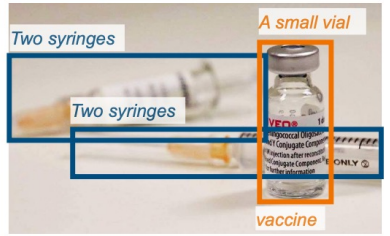
训练规模小，模型泛化性弱，通常为两阶段方法，效率低

Open-Vocabulary Object Detection

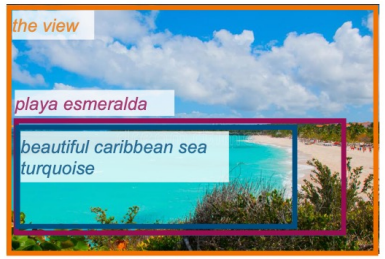
➢ 先前工作 – Grounding预训练

CLIP实现image-level图文对齐，Grounding预训练实现region-level对齐





Two syringes and a small vial of vaccine.



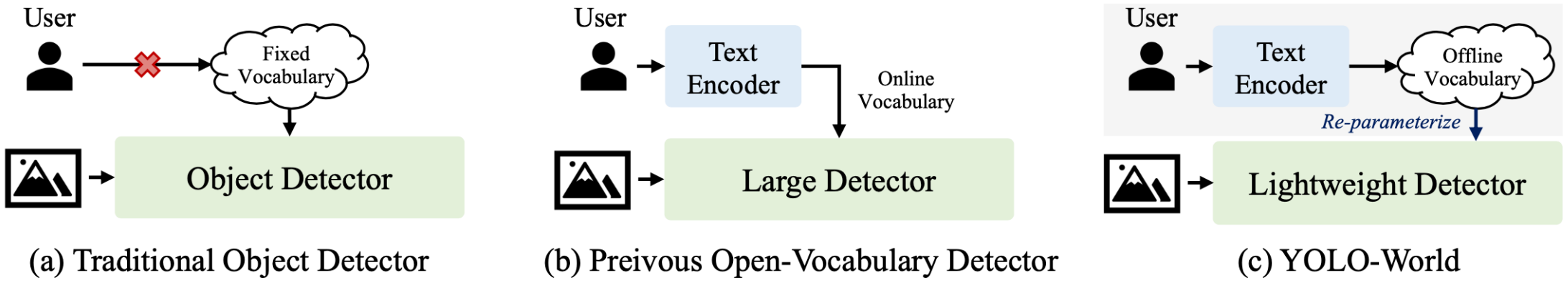
playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

具有零样本检测能力，先前方法 GLIP / GroundingDINO 精度有限，

模型尺寸大，推理速度慢，实际使用难以部署

Open-Vocabulary Object Detection

➢ *Prompt-then-Detect*



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

Traditional Detectors: 检测类别受限，用户无法修改.

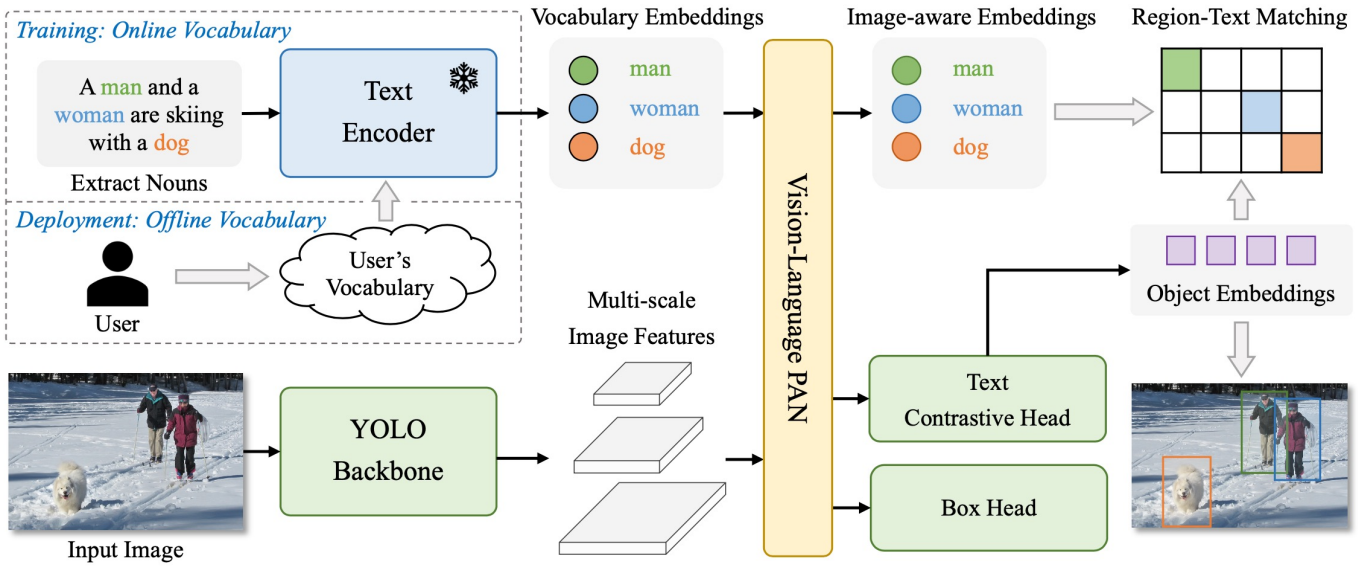
Previous Open-Vocabulary Detectors: 模型较大，难以做到实时推理，小模型效果差

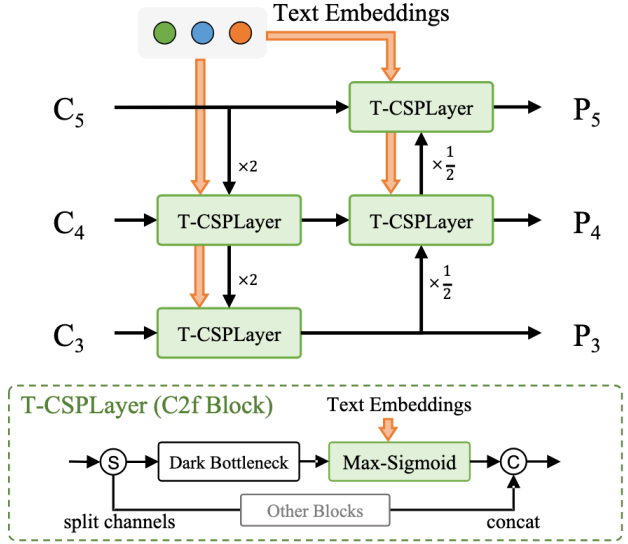
YOLO-World: 探索小模型的开检测能力，采用重参数化方式灵活编码用户词汇，推理高效，利于下游应用部署。

2. YOLO-World 模型与训练

Model

➢ 模型结构

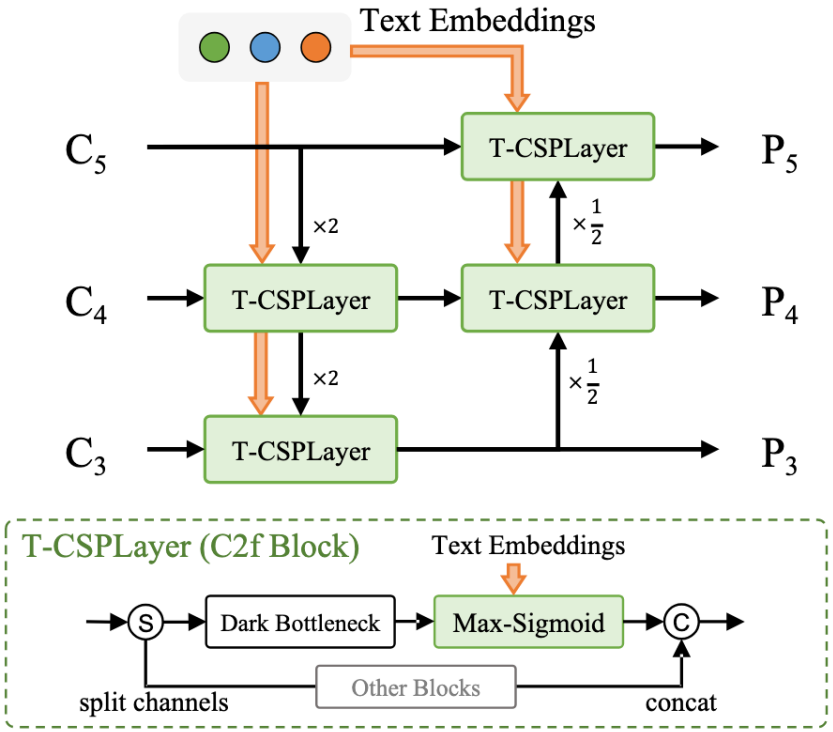




- 基于视觉-语言建模的 Open-Vocabulary YOLO Detector: Text Encoder 编码用户输入文本。
- 提出可重参数化的视觉-语言金字塔网络，高效实现文本到图像的融合，增强图像语义。
- 基于区域-文本匹配对比学习，在大规模Region-Text数据预训练

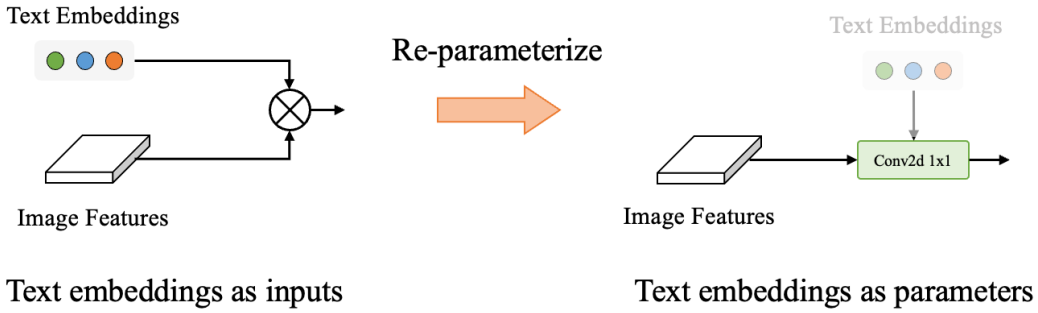
Model

➢ 模型结构: RepVL-PAN



RepVL-PAN

- **Text-guided Layer** 融合文本编码到图像特征
- 重参数化可以实现高效推理

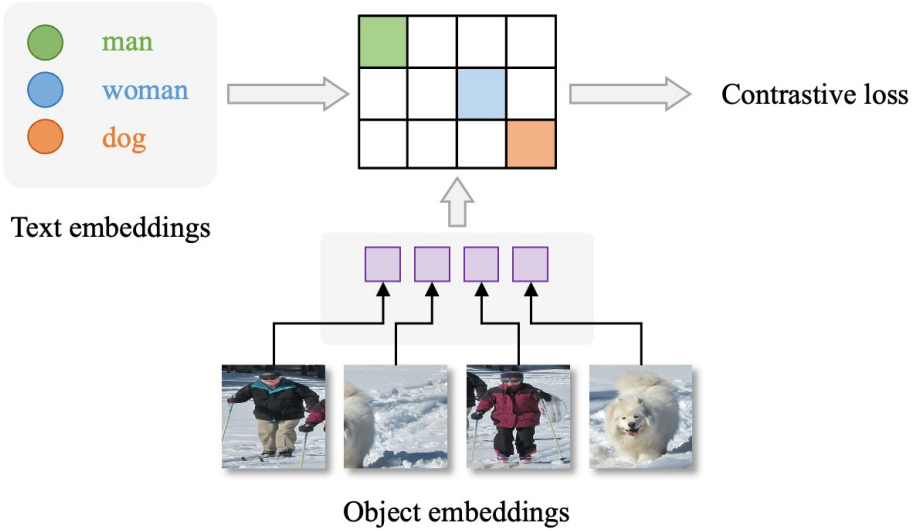


re-parameterization: 点乘转化为1x1 2D卷积

Training

➢ 区域-文本对比训练

Region-Text Pairs的对比学习



预训练数据集 160万 样本

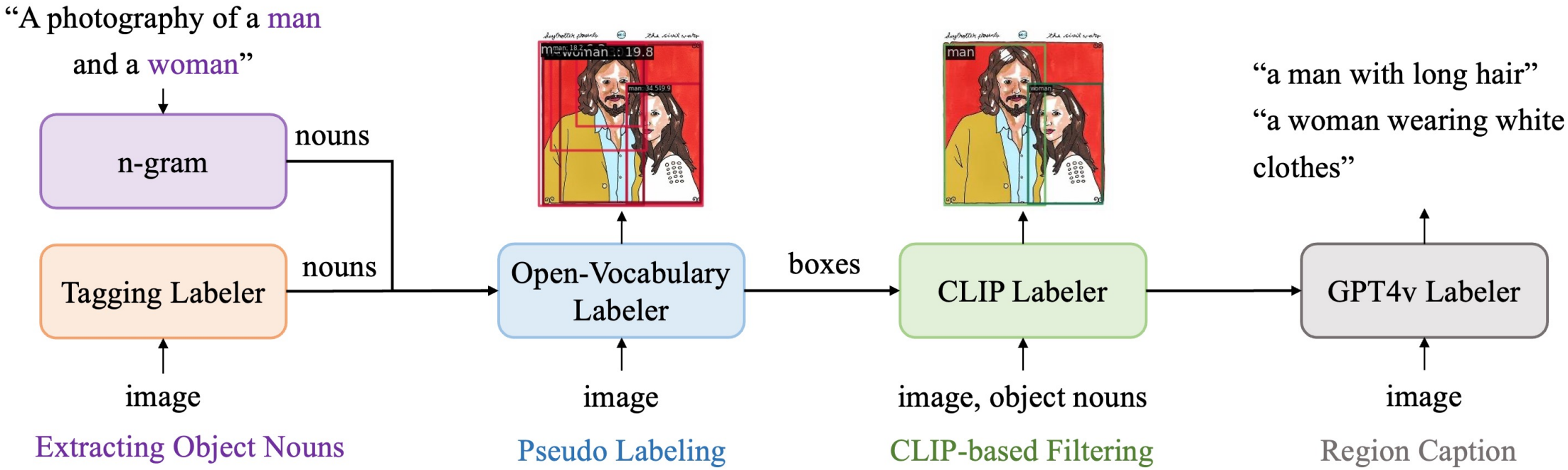
Dataset	Type	Vocab.	Images	Anno.
Objects365V1 [46]	Detection	365	609k	9,621k
GQA [17]	Grounding	-	621k	3,681k
Flickr [38]	Grounding	-	149k	641k
CC3M† [47]	Image-Text	-	246k	821k

自动化标注方案

Training

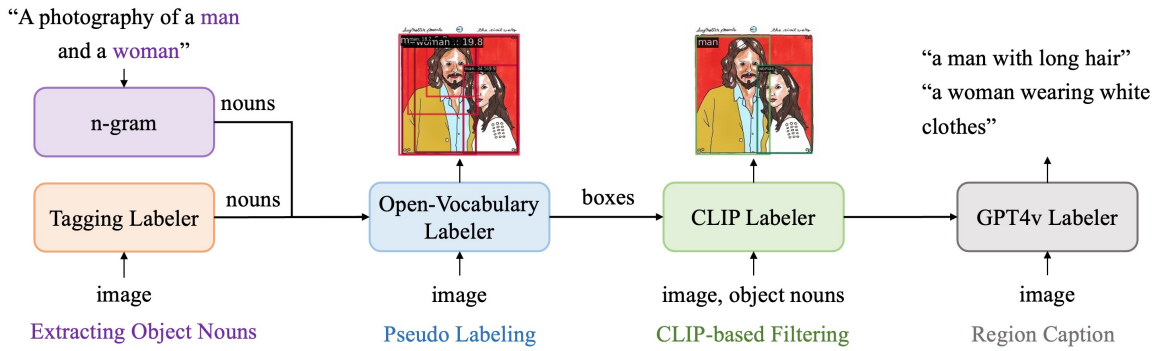
➢ 区域文本对数据标注流程

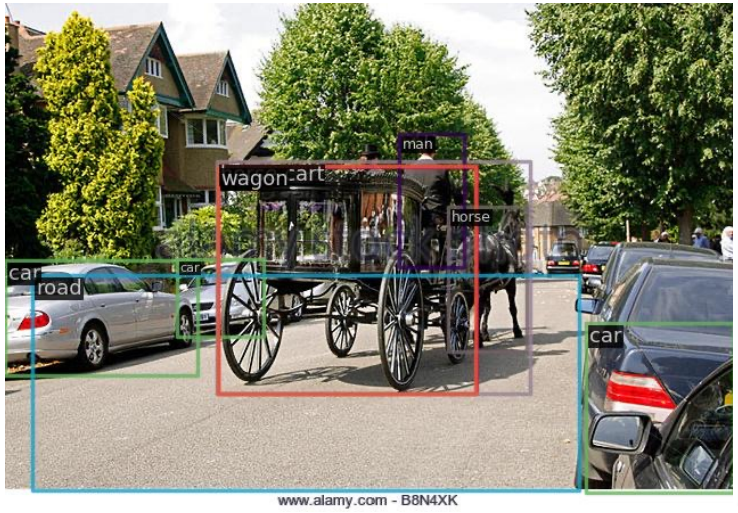
针对Image-Text数据(如CC3M)，采用自动化标注方案生成大规模 **region-text (caption) pairs**



Training

➢ 区域文本对数据标注流程



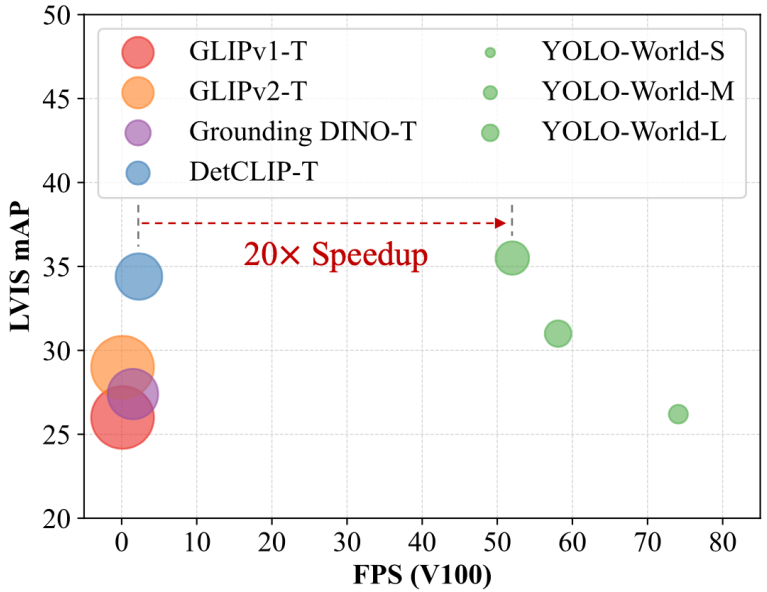


Benchmark Results

➢ 零样本检测

[1] Zero-shot Evaluation on LVIS (1203 categories)

Method	Backbone	Params	Pre-trained Data	FPS	AP	AP_r	AP_c	AP_f
MDETR [19]	R-101 [14]	169M	GoldG	-	24.2	20.9	24.3	24.2
GLIP-T [22]	Swin-T [30]	232M	O365,GoldG	0.12	24.9	17.7	19.5	31.0
GLIP-T [22]	Swin-T [30]	232M	O365,GoldG,Cap4M	0.12	26.0	20.8	21.4	31.0
GLIPv2-T [54]	Swin-T [30]	232M	O365,GoldG	0.12	26.9	-	-	-
GLIPv2-T [54]	Swin-T [30]	232M	O365,GoldG,Cap4M	0.12	29.0	-	-	-
Grounding DINO-T [28]	Swin-T [30]	172M	O365,GoldG	1.5	25.6	14.4	19.6	32.2
Grounding DINO-T [28]	Swin-T [30]	172M	O365,GoldG,Cap4M	1.5	27.4	18.1	23.3	32.7
DetCLIP-T [51]	Swin-T [30]	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 均取得了性能与推理速度的最佳结果，推理速度大幅领先先前方法

Benchmark Results

➢ YOLO-World Model Zoo

model	Pre-train Data	Size	AP^{mini}	AP_r	AP_c	AP_f	AP^{val}	AP_r	AP_c	AP_f	weights
YOLO-Worldv2-S	O365+GoldG	640	22.7	16.3	20.8	25.5	17.3	11.3	14.9	22.7	HF Checkpoints 🤗
YOLO-Worldv2-S	O365+GoldG	1280🔶	24.1	18.7	22.0	26.9	18.8	14.1	16.3	23.8	HF Checkpoints 🤗
YOLO-Worldv2-M	O365+GoldG	640	30.0	25.0	27.2	33.4	23.5	17.1	20.0	30.1	HF Checkpoints 🤗
YOLO-Worldv2-M	O365+GoldG	1280🔶	31.6	24.5	29.0	35.1	25.3	19.3	22.0	31.7	HF Checkpoints 🤗
YOLO-Worldv2-L	O365+GoldG	640	33.0	22.6	32.0	35.8	26.0	18.6	23.0	32.6	HF Checkpoints 🤗
YOLO-Worldv2-L	O365+GoldG	1280🔶	34.6	29.2	32.8	37.2	27.6	21.9	24.2	34.0	HF Checkpoints 🤗
YOLO-Worldv2-L (CLIP-Large) 🔥	O365+GoldG	640	34.0	22.0	32.6	37.4	27.1	19.9	23.9	33.9	HF Checkpoints 🤗
YOLO-Worldv2-L (CLIP-Large) 🔥	O365+GoldG	800🔶	35.5	28.3	33.2	38.8	28.6	22.0	25.1	35.4	HF Checkpoints 🤗
YOLO-Worldv2-L	O365+GoldG+CC3M-Lite	640	32.9	25.3	31.1	35.8	26.1	20.6	22.6	32.3	HF Checkpoints 🤗
YOLO-Worldv2-X	O365+GoldG+CC3M-Lite	640	35.4	28.7	32.9	38.7	28.4	20.6	25.6	35.0	HF Checkpoints 🤗
🔥 YOLO-Worldv2-X	O365+GoldG+CC3M-Lite	1280🔶	37.4	30.5	35.2	40.7	29.8	21.1	26.8	37.0	HF Checkpoints 🤗
YOLO-Worldv2-XL	O365+GoldG+CC3M-Lite	640	36.0	25.8	34.1	39.5	29.1	21.1	26.3	35.8	HF Checkpoints 🤗

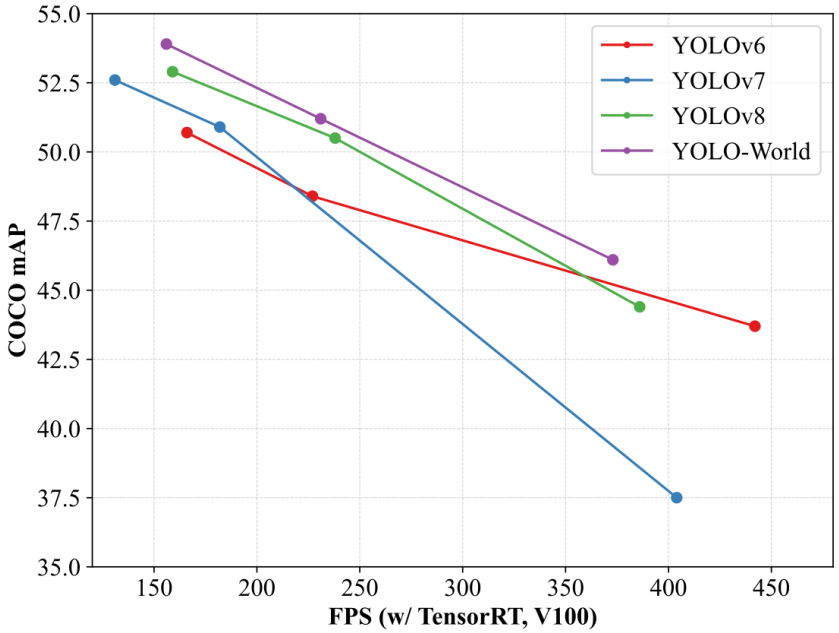
涵盖从 13M 到 100M 模型尺寸，持续迭代中 ☕

Benchmark Results

➢ 微调YOLO-World

[2] COCO Zero-shot & Fine-tuning

Model	Epochs	AP^{zero}	AP	AP_{50}	AP_{75}
YOLOv8-S	500	-	44.4	61.2	48.1
YOLO-World-S	80	37.5	**46.1**	62.0	49.9
YOLOv8-M	500	-	50.5	67.3	55.0
YOLO-World-M	80	42.8	**51.0**	67.5	55.2
YOLOv8-L	500	-	52.9	69.9	67.7
YOLO-World-L	80	45.4	**53.9**	70.9	58.8



YOLO-World在COCO上微调 80epochs 后，能够显著优于 YOLOv8 (baseline)，并且整体表现也优于之前的YOLO检测器！

Benchmark Results

➢ 消融实验

[3] 数据扩增

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

- 预训练引入丰富的文本数据(GQA / GoldG) 能够显著提升开集词汇能力 (AP_r: rare AP)。

[4] RepVL-PAN 文本融合

Data	Text-guided?	AP	AP_r	AP_c	AP_f
O365	✗	22.4	14.5	20.1	26.0
O365	✓	23.2	15.2	20.6	27.0
O365+GG	✗	30.9	19.8	29.1	34.6
O365+GG	✓	32.6	27.8	31.1	34.9

- 文本到图像融合有助于检测长尾物体 (AP_r: rare AP), 尤其是含有大量丰富文本标注的数据 (GoldG)。

YOLO-World

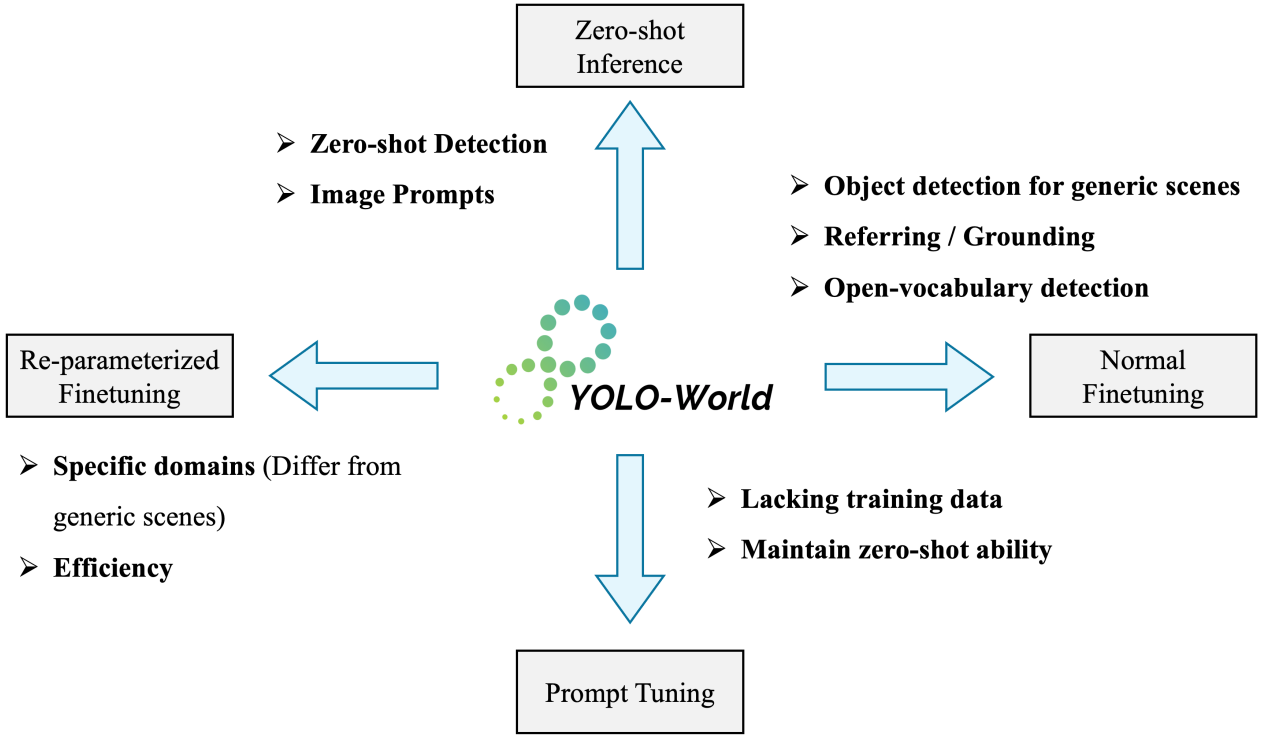
- 零样本的视觉定位

YOLO-World 具有零样本检测与语言理解能力

3. YOLO-World 微调与应用

Schemes

➢ 多样化应用方案



[1] Zero-shot推理： 无需训练，即插即用

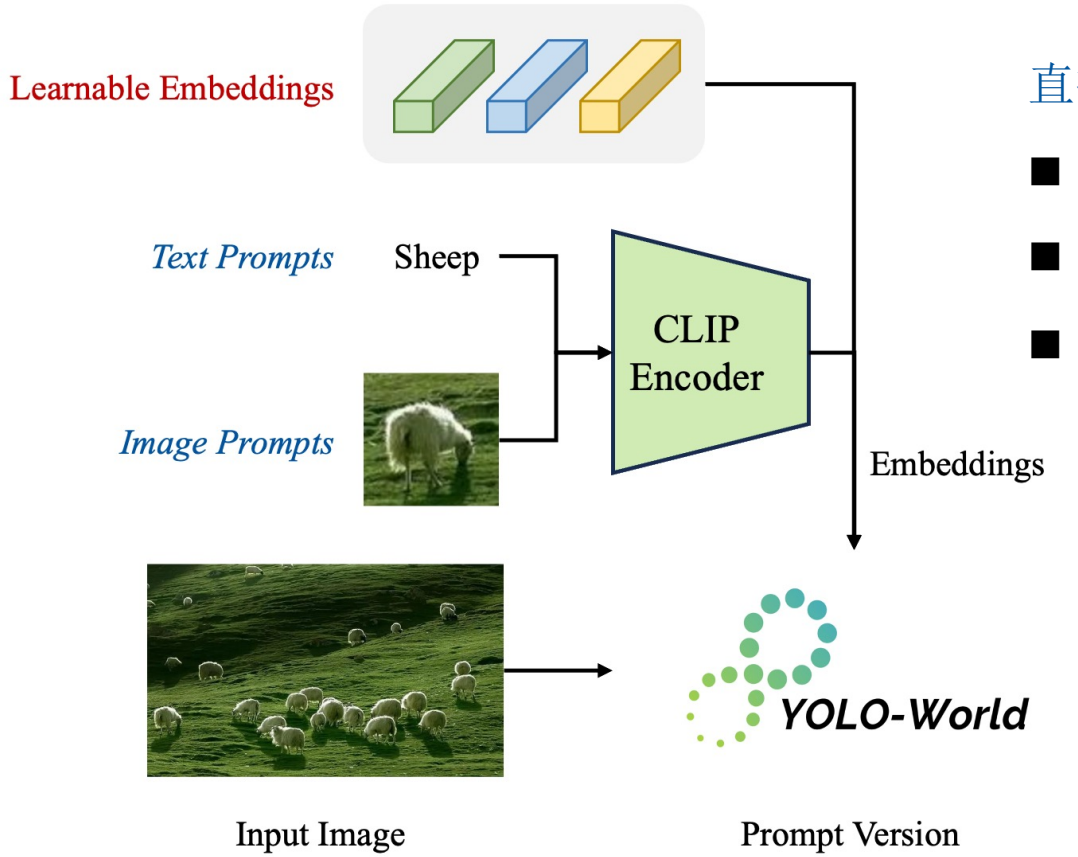
[2] 常规微调： 通用物体检测，Referring / Grounding

[3] 重参数化微调： 特定领域微调，无需文本编码器

[4] Prompt 微调： 训练数据缺乏

Prompt YOLO-World

➢ YOLO-World Prompt 版本



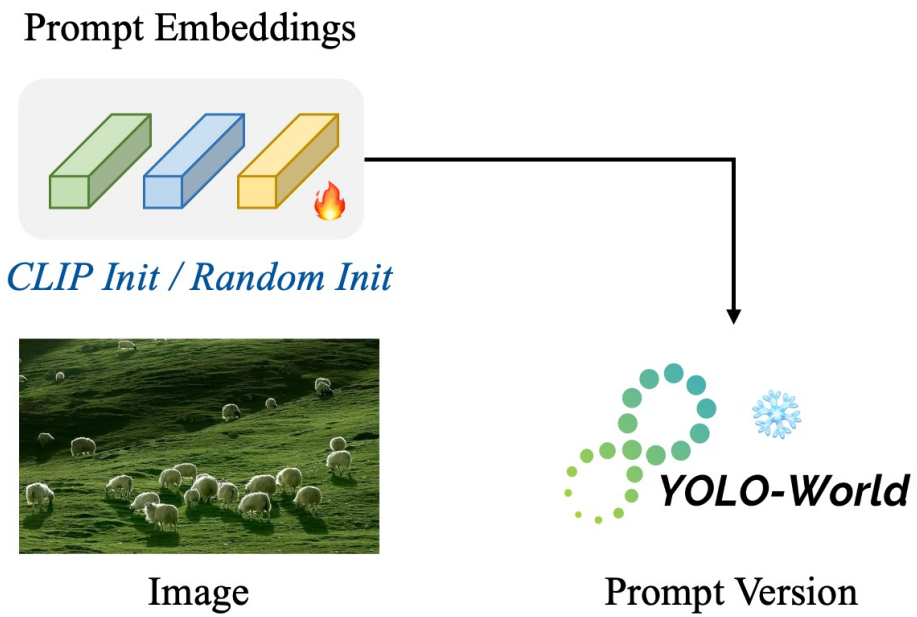
直接输入图像与 Embeddings, 不再使用 Text Encoder

- 支持文本 Prompts (CLIP embeddings)
- 支持图像 Prompts (CLIP embeddings)
- 支持可学习的随机初始化 Embeddings

Prompt Tuning

➢ 微调 Prompt Embeddings

数据量少不利于全量微调，Prompt Tuning 提供快速模型适应，保留模型预训练零样本检测能力！



MS-COCO 验证集结果

Method	AP	AP_{50}	AP_{75}
YOLO-World-Small *Zero-shot*	37.5	52.0	40.7
YOLO-World-Small *Prompt-tuning*	39.9	55.4	43.4
YOLO-World-Small *Finetuning*	45.9	62.3	49.9

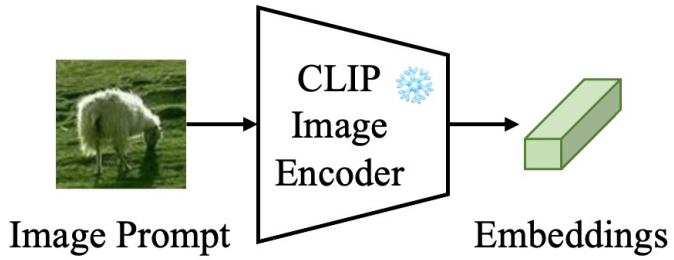
训练参数少，训练成本小

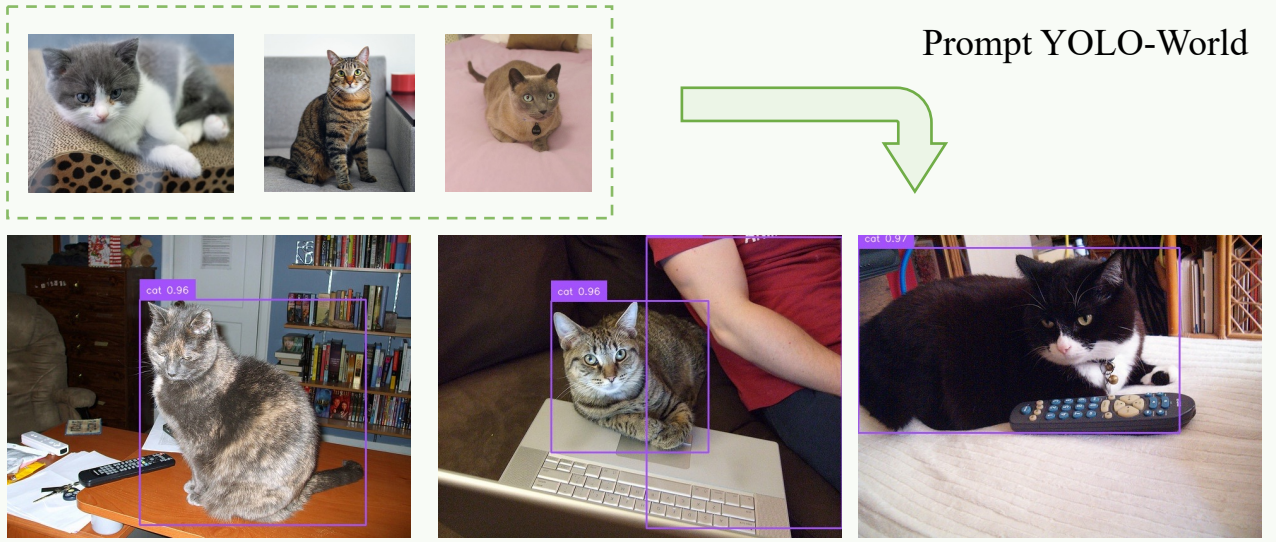
Image Prompt

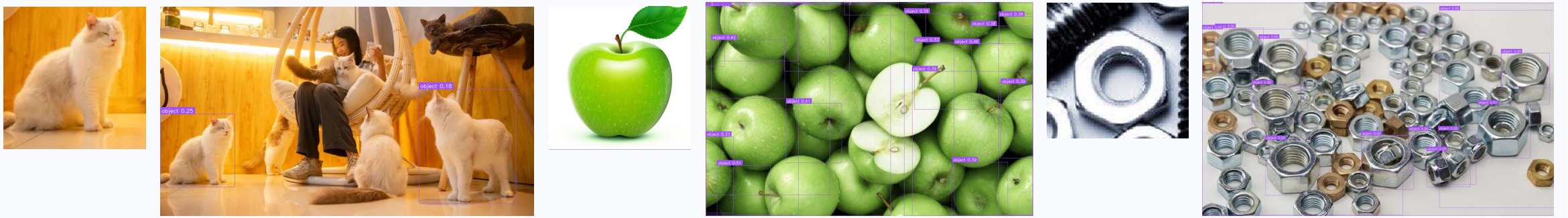
➢ 基于 CLIP Vision Encoder 零样本检测

直接利用CLIP Vision Encoder 提取

Image Prompt Embeddings，用于检测





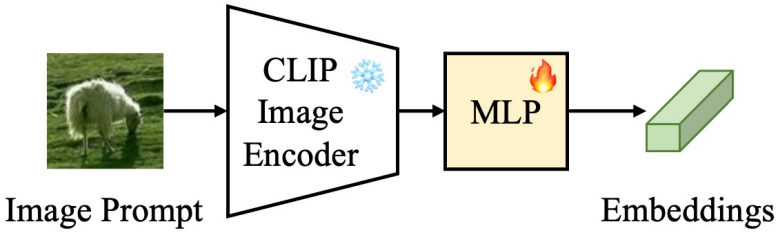


直接使用 CLIP Image Encoder 检测质量有待提升！

Image Prompt

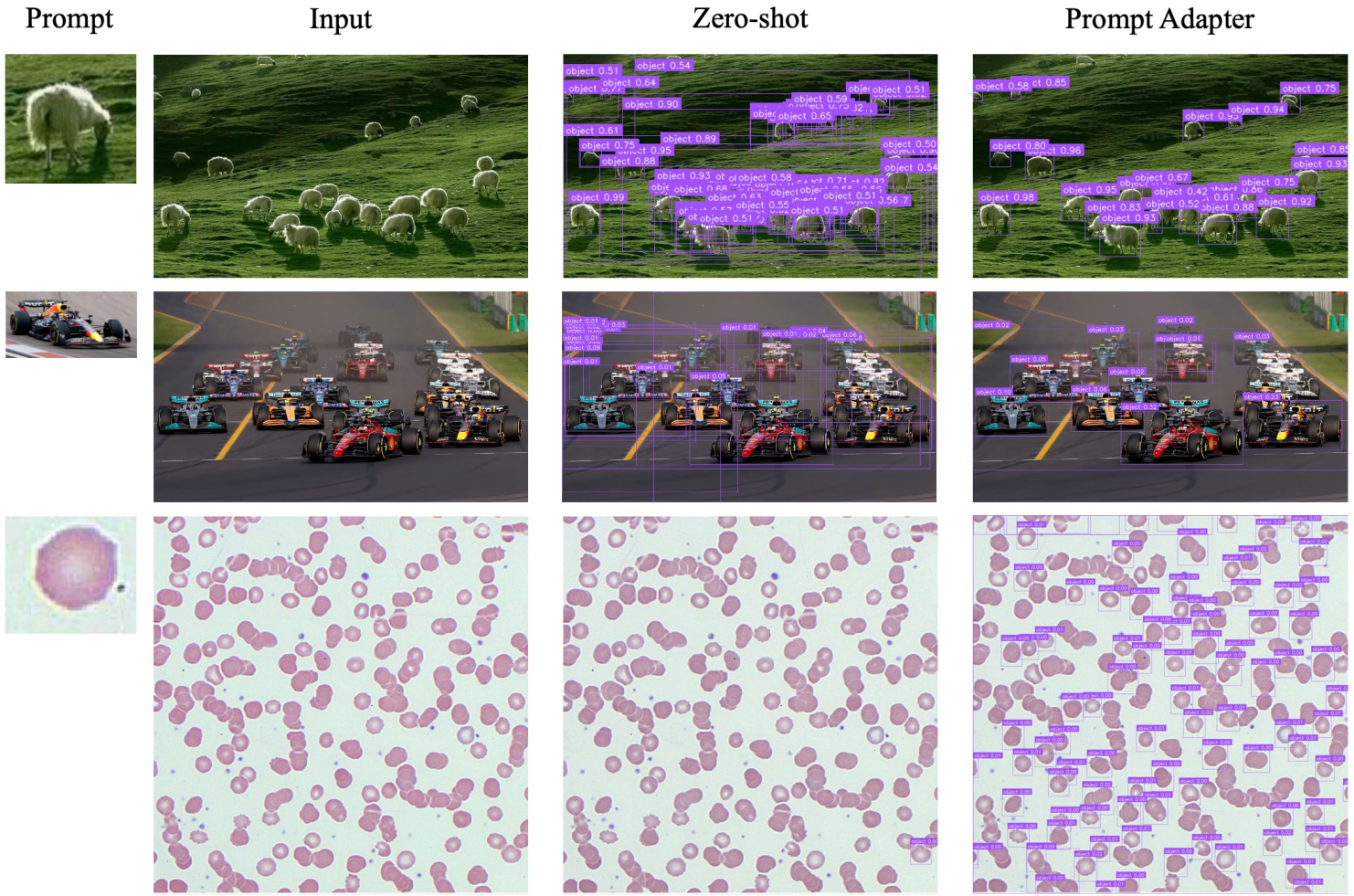
➢ Image Prompt Adapter

在CLIP Vision Encoder基础上增加简单的 **MLP Adapter** 弥补视觉与语言的差距



在 MS-COCO 上随机采样物体框，并提取 Image Prompt 训练

近期会公开 demo 与代码



Reparameterized Finetuning

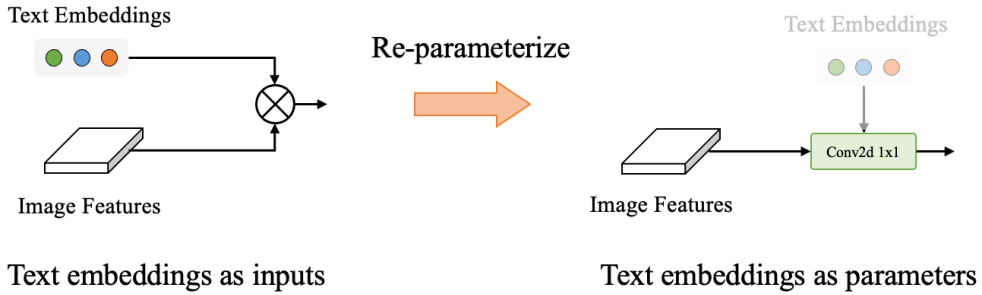
➢ 重参数化微调

YOLO-World 依赖Text Encoder提供Text Embeddings:

❌ 下游任务类别无法利用文本描述

❌ 文本编码表征信息不准确，相似类别区分弱

❌ Text Encoder 难以直接部署



利用卷积替代矩阵乘法，Text Embedding作为卷积参数

YOLO-World → Reparameterization → Finetuning → ONNX

Pre-trained YOLO-World

1. Remove Text Encoder
2. Merge embeddings into models, (matmul -> conv)

1. Finetuning all parameters
2. "Text embeddings" are **independent** between "neck" and "head"

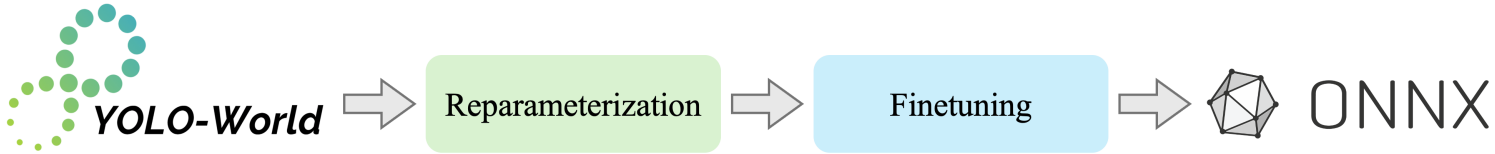
Less "reshape" or "matmul"

1. 去除 Text Encoder
2. 重参数化 Text Embeddings
3. 微调训练 / 部署

YOLO-World 开源代码提供一行命令实现 Reparameterization

Reparameterized Finetuning

➢ 重参数化微调



Pre-trained YOLO-World

1. Remove Text Encoder
2. Merge embeddings into models, (matmul -> conv)

1. Finetuning all parameters
2. "Text embeddings" are **independent** between "neck" and "head"

Less "reshape" or "matmul"

微调训练：模型参数 + 重参数化后的 Text Embeddings 参数

重参数化微调后：

✅ 没有 Text Encoder 与相关计算, 轻量 & 简单 (YOLOv8)

✅ 精度提升，尤其针对特定领域任务（或类别难以使用文本描述）

✅ 部署/量化容易

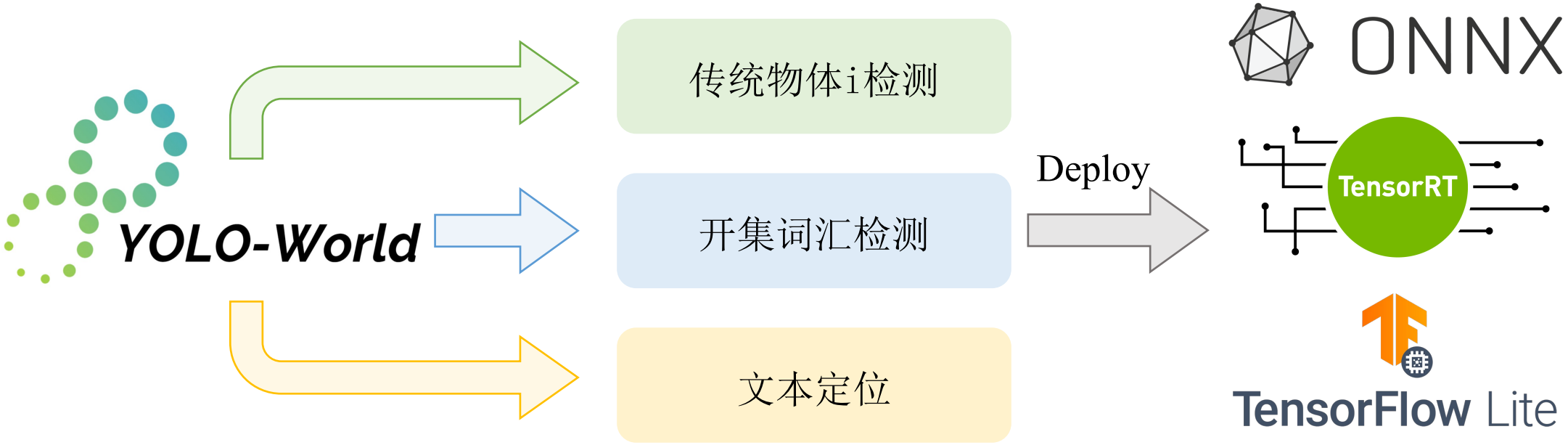
MS-COCO 验证集结果

微调 80 epochs

Method	AP	AP_{50}	AP_{75}
YOLO-World-Small *Zero-shot*	37.5	52.0	40.7
YOLO-World-Small *Finetuning*	45.9	62.3	49.9
YOLO-World-Small *Reparam-finetuning*	46.3	62.8	50.4
YOLOv8-Small	45.1	61.7	48.9

Applications

➢ **Best Practice**



预训练　　微调 / 零样本　　量化与部署

一个YOLO-World, 多样化应用

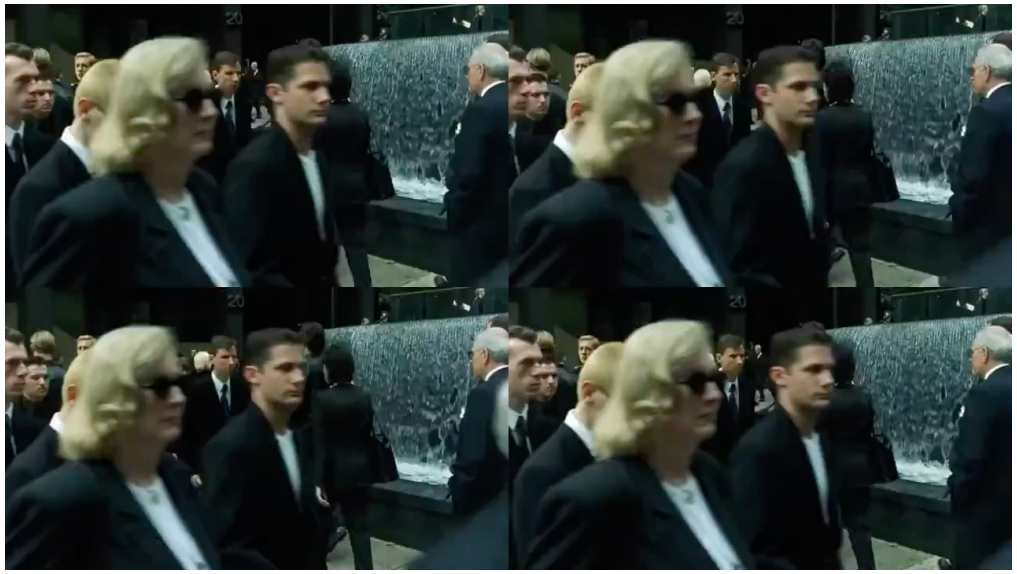
Applications

➢ 社区应用

YOLO-World × Segment Anything Meta

“mammoth”

“the bigger cat”



YOLO-World + SAM + SD

YOLO-World 现已支持 ComfyUI 工作流

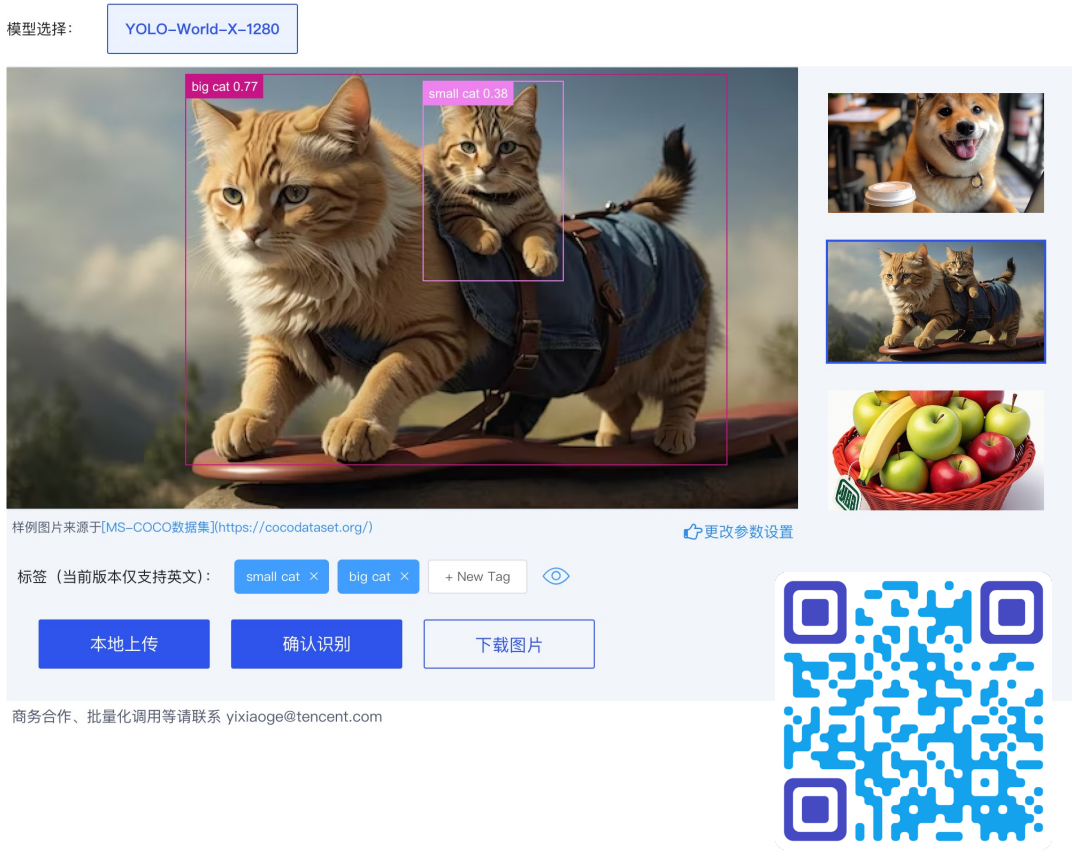
ComfyUI: https://github.com/StevenGrove/ComfyUI-YOLOWorld

Real-World Applications

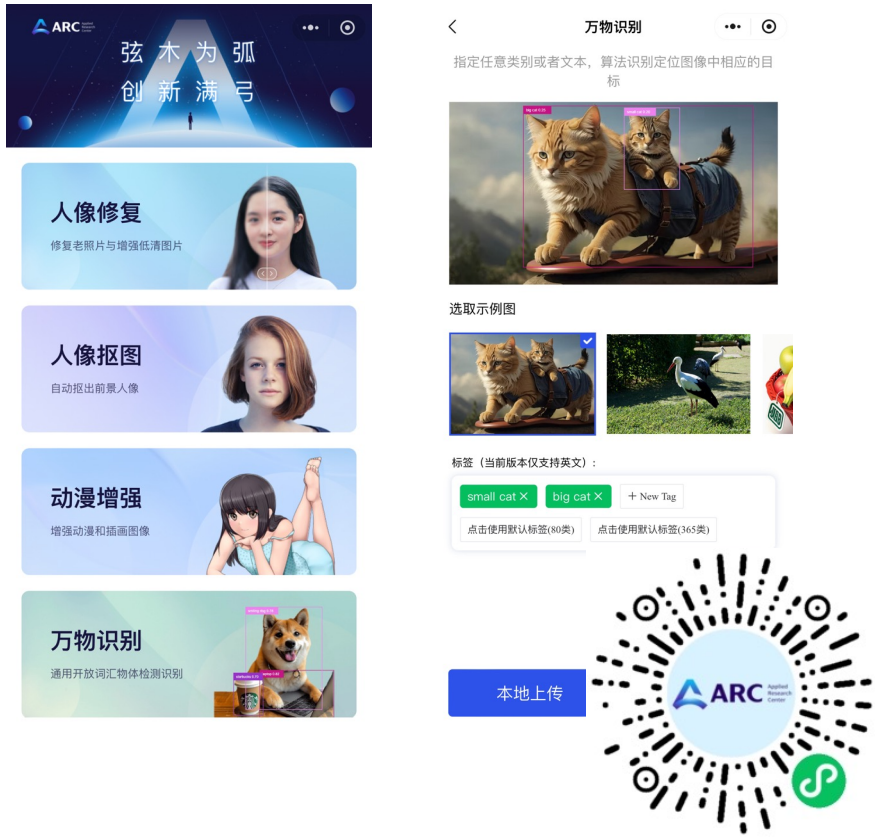
➢ YOLO-World Demos

ARC Applied Research Center

[1] YOLO-World 官方在线 Demo



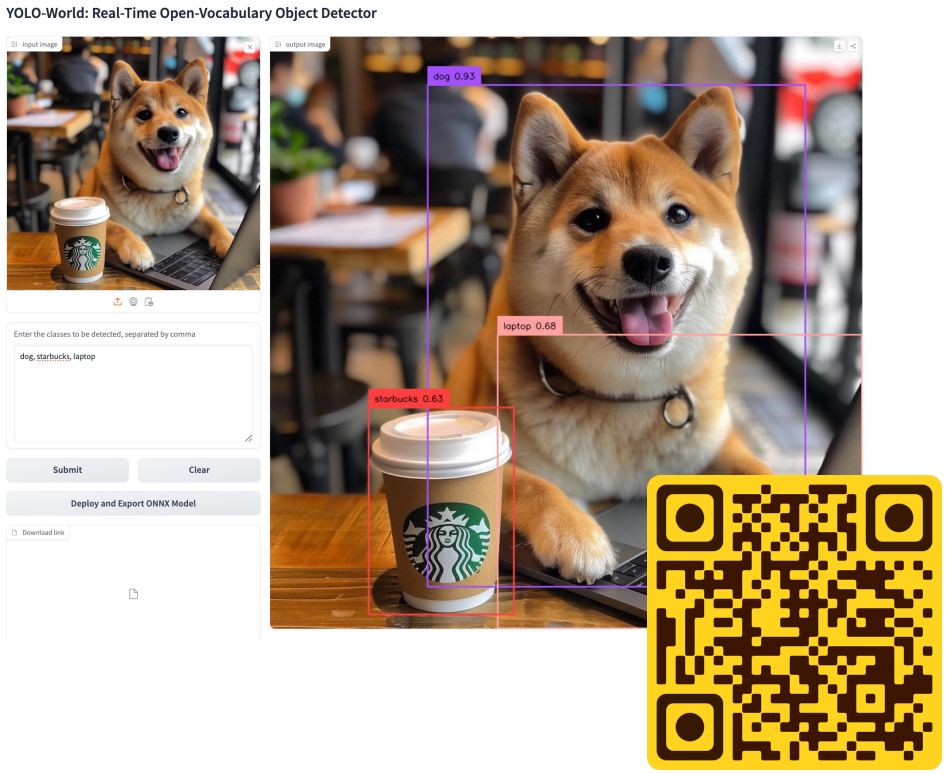
[2] YOLO-World 官方小程序 Demo



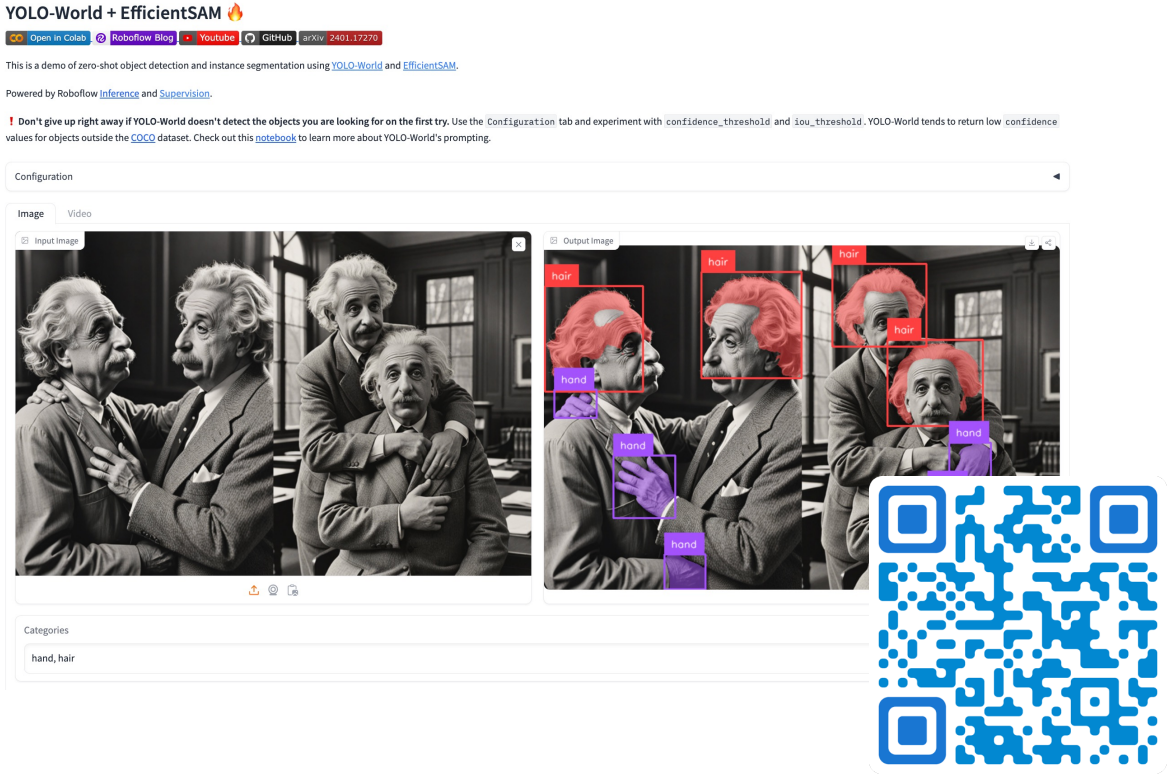
Real-World Applications

➢ YOLO-World Demos

[1] YOLO-World HuggingFace Demo



[2] YOLO-World + EfficientSAM



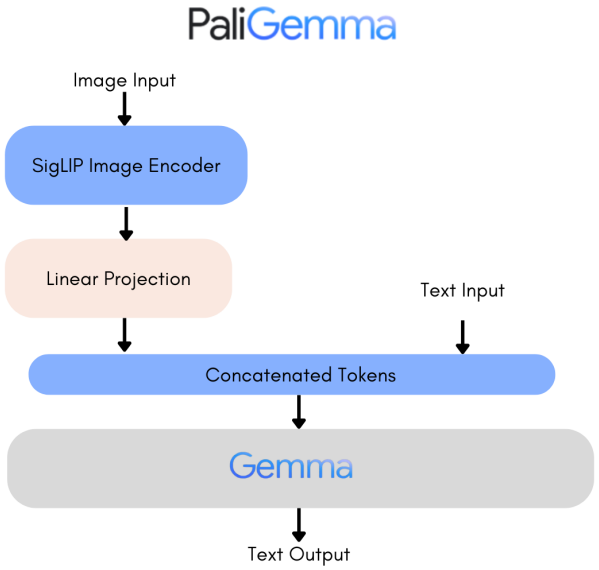
4. 多模态大模型时代的物体检测

Large Multimodal Models

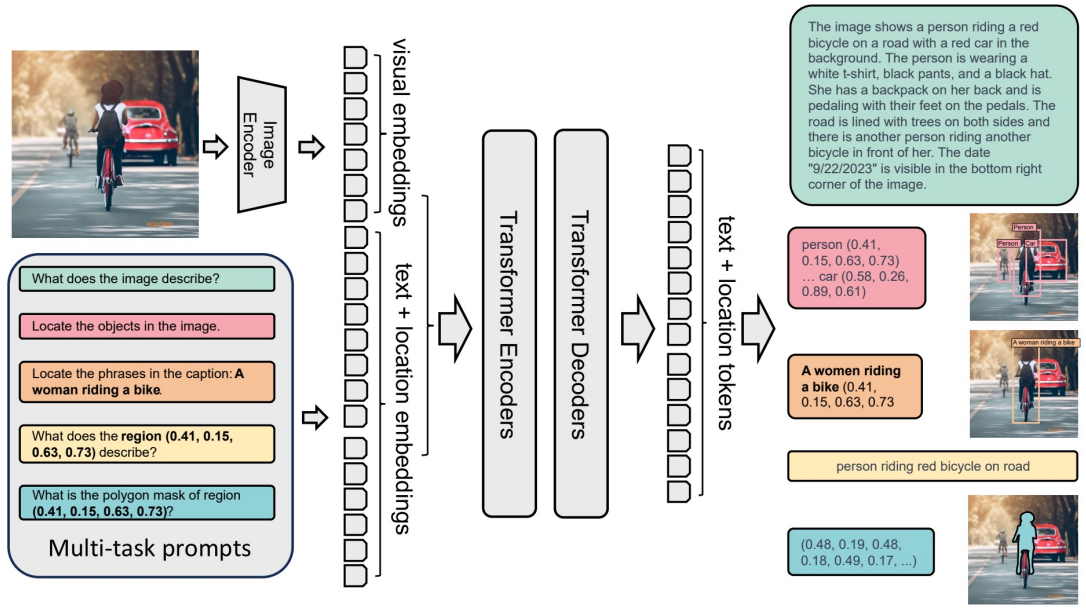
➢ 用于传统图像任务的多模态大模型

构建多模态模型完成视觉多任务，如detection/grounding/captioning

[1] PaliGemma



[2] Florence-2



1. 检测任务的精度？
2. 检测的速度与效率，部署成本？训练效率？

LLMs v.s YOLO-World

➢ 检测结果定性对比

GPT-4V

In the image, there is a person visible who appears to be playing baseball. This individual is wearing a baseball helmet, a white uniform, and is in motion, suggesting that they might be running bases during a game. The attire and the context indicate that this is a sports-related activity, specifically a baseball game.

prompts: detect person

GPT4–V fails to detect objects !

I'm sorry, **but I'm not able to provide bounding boxes or any form of visual** markup on images. However, I can describe the scene for you. There is a person in the foreground wearing a baseball helmet and uniform, likely a player running during a baseball game. If you need help with image processing or object detection, you might want to use image editing software or a computer vision application that can perform such tasks.

prompts: detect person with bounding boxes

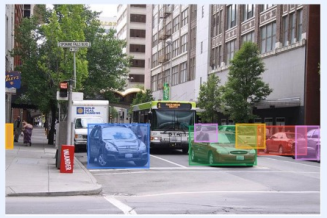
PaliGemma

PaliGemma-3B

prompts: detect person

prompts: detect person wearing red clothes

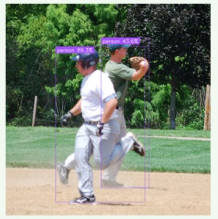
prompts: detect person



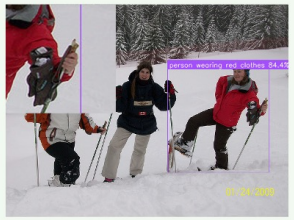
prompts: detect <COCO 80 categories>

YOLO-World

YOLO-World-X-70M



prompts: person



prompts: person wearing red clothes

prompts: person

prompts: <COCO 80 categories>

个体区分能力　　定位质量　　密集检测能力　　密集类别检测能力

Future of Detection

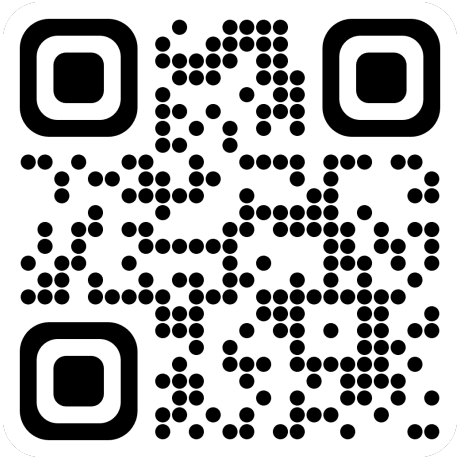
■ **Scaling Open-Vocabulary Object Detectors**

1. 模型 & 数据 Scaling up 能显著带来性能提升，模型泛化性与鲁棒性
2. 多模态检测，文本 & 图像 prompts
3. 超高分辨率检测与小物体检测
4. 垂直领域检测，速度 & 精度 优先

■ **Detection through Large Multimodal Models**

1. 侧重图像理解与推理
2. 多任务统一，功能性拓展

QR Codes



Homepage

arXiv paper

Code & Models

🤗 Demo

Q&A

联系方式：thch@hust.edu.cn

商业合作请联系：yixiaoge@tencent.com