基于 OV-DINO 的开放词汇目标检测复现与扩展实验

计算机视觉课程期中报告

徐永雪

中山大学　智能工程学院

智能科学与技术 3 班

Extend OV-DINO github*

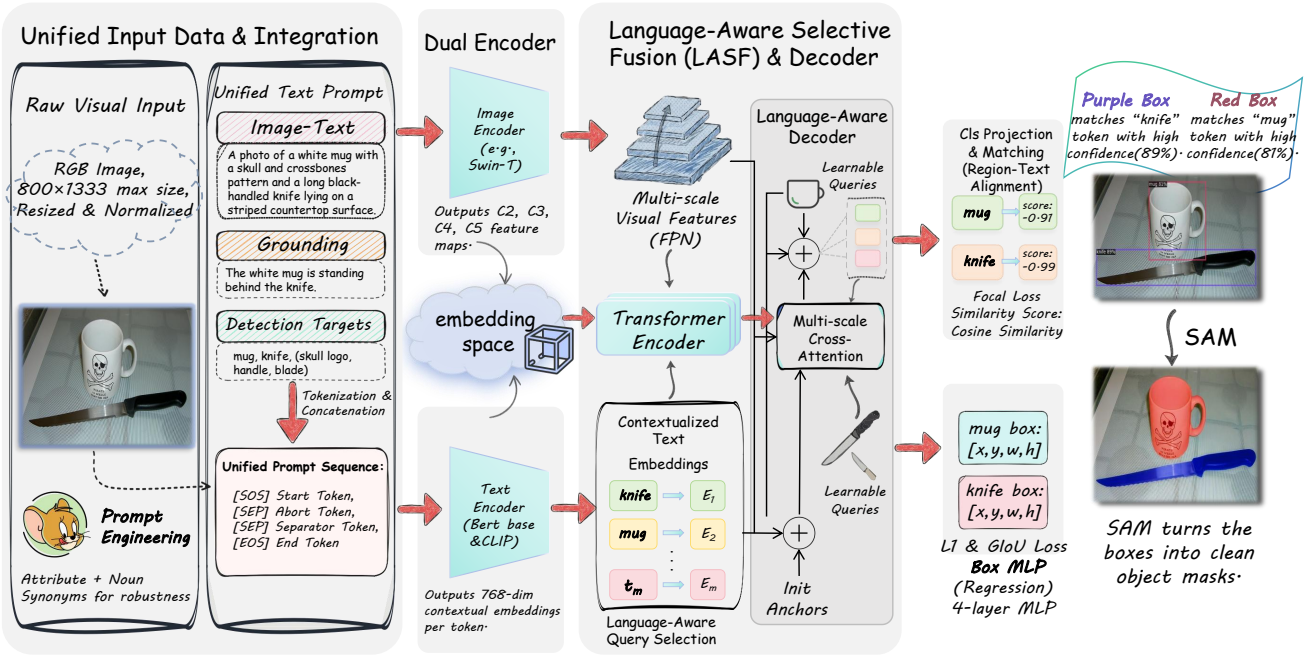


图 1. **OV-DINO 复现与扩展工作的整体流程框架。** (1) 输入端：采用统一数据集成策略，将图像与包含属性增强的文本提示（Unified Text Prompt）转化为标准化序列；(2) 特征编码与融合：利用双编码器提取多模态特征，并通过语言感知选择融合（LASF）模块实现精细化的视觉-语言对齐；(3) 检测与下游扩展：模型最终输出开放类别的边界框（Bounding Box），并将其作为提示进一步引导 SAM 生成实例分割掩码，实现了从检测到分割的完整管线。

Abstract

开放词汇目标检测（*Open-Vocabulary Object Detection, OVOD*）旨在利用预训练的视觉-语言模型能力，使检测器能够识别训练期间未见的新类别。本项目选取该领域的代表性工作 *OV-DINO* 进行复现与深入研究。本期中报告旨在汇报基于 *OV-DINO* 的复现进展以及围绕模型架构与训练策略所开展的扩展性实验成果。目

*基于 OV-DINO 的扩展性工作，包括替换 Backbone 和 Lora 微调等探索.

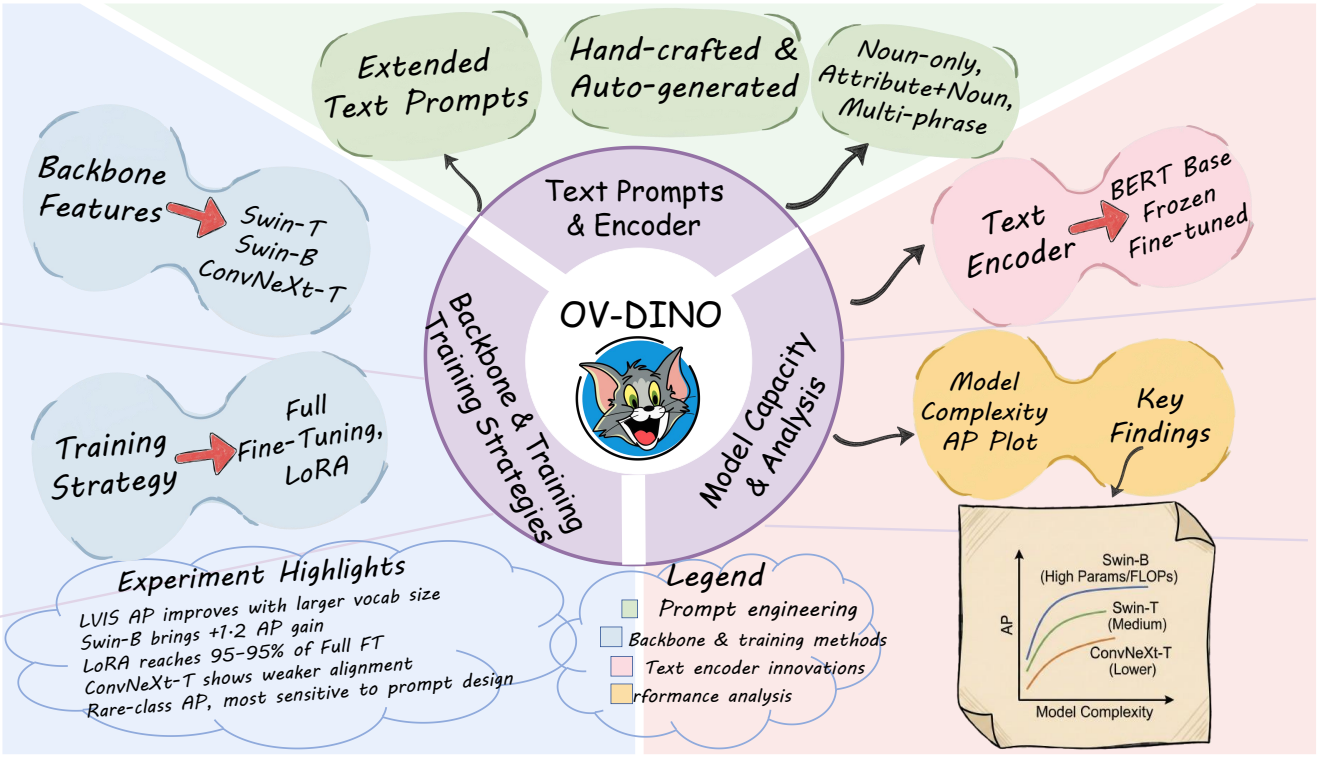


图 2. **OV-DINO 复现与扩展工作的实验概览**。我们将扩展性研究划分为三大核心维度：(1) **主干网络与训练策略**（左侧蓝区）：涵盖了 Swin-T/B 与 ConvNeXt 的架构对比，以及 LoRA 高效微调的有效性验证；(2) **文本编码与提示工程**（上方绿/粉区）：系统探究了属性增强、同义词扩展及 BERT 编码器微调对语义对齐的影响；(3) **模型容量与性能分析**（右侧黄区）：通过“复杂度-AP”曲线揭示了视觉-语言对齐的边际效应。图左下角列出了本项目截止目前的阶段性实验亮点（Experiment Highlights）。

前，我们已成功复现了官方代码在 *LVIS* 数据集上的基准性能，并构建了从统一输入处理到开放世界实例分割的完整工作流程（如图 *1* 所示）。

在此基础上，为了探索模型的性能边界与内在机制，我们开展了系统性的扩展实验。具体而言，我们首先完成了 *Swin-T*（基准）、*Swin-B*（大容量）以及 *ConvNeXt-T*（*CNN* 架构）三种主干网络的适配与对比。阶段性结果显示，虽然 *ConvNeXt* 在传统封闭集检测中表现良好，但在开放词汇场景下，其特征空间与文本嵌入的对齐效果弱于 *Transformer* 架构；而 *Swin-B* 在稀有类别（*Rare Classes*）上的 AP_r 提升验证了模型容量的价值，但性能曲线同时也呈现出边际收益递减的趋势。其次，在训练策略方面，我们对比了全量微调（*Full Fine-Tuning*）与参数高效微调（*LoRA*）。实验发现，*LoRA* 仅需微调约 *1%–3%* 的参数即可达到全量微调 *90%–95%* 的性能，且在大参数模型上展现出优异的正则化效果与泛化潜力。此外，关于文本提示与词表规模的消融研究表明，构建包含“属性 + 名词”或“多短语同义词”的丰富语义 *Prompt*，配合更大的词表覆盖率，是提升长尾类别识别能力的关键。最后，我们将检测结果作为提示输入 *Segment Anything Model (SAM)*，成功实现了开放词汇检测向实例分割的下游应用延伸。

综上所述，本期中报告所述工作验证了 *OV-DINO* 框架的有效性，阶段性实验结果表明：在开放词汇场景下，优化视觉-语言的语义对齐质量比单纯增加视觉模型容量更为关键。

1. Introduction

1.1. 研究背景

传统目标检测方法，如 Faster R-CNN、DETR [1] 及其改进版本 DINO [16]，在 COCO [8] 等封闭集数据集上已经取得了显著性能。然而，这些模型严重依赖于预定义的人工标注类别，难以识别训练集中未出现的物体，这种“闭集假设”极大地限制了模型在复杂现实场景中的应用。为突破这一瓶颈，开放词汇目标检测

（OVOD）应运而生。OVOD 旨在利用大规模预训练视觉-语言模型（如 CLIP [13]、BERT [2]），将视觉特征映射至丰富的文本语义空间，从而赋予模型识别新类别（Novel Classes）的能力 [3, 15]。

早期的 OVOD 方法多采用知识蒸馏（如 ViLD [3], RegionCLIP [17]）或大规模短语定位预训练（如 GLIP [7], Grounding DINO [10]）。近期，Wang 等人提出的 OV-DINO [14] 进一步统一了检测与定位的数据格式，并引入语言感知选择融合（LASF）模块，在 LVIS [4] 等长尾数据集上取得了 SOTA 性能。

1.2. 实验目的与扩展工作

本项目的首要目标是基于官方代码复现 OV-DINO 在 LVIS 数据集上的基准性能。在此基础上，我们设计并实施了以下四个维度的扩展性实验。

在主干网络对比方面，官方实现仅基于 Swin-T [11]，我们已将主干网络扩展至 Swin-B 和 ConvNeXt-T [12]，旨在验证模型容量和架构差异对开放词汇检测性能的影响。

在训练策略分析方面，我们实现了 LoRA [5] 微调策略并与全量微调进行对比，重点关注 LoRA 在仅更新 1%–3% 参数量的情况下能否逼近全量微调的检测精度。

在 Prompt 形式与词表规模消融方面，我们设计了" 仅名词"、" 属性 + 名词" 及" 多短语同义词" 等对照实验，并测试了从 30 类到 1200+ 类不同词表规模的影响，以量化语义描述对长尾稀有类别召回率的增益。

在下游任务扩展方面，我们复现了 OV-DINO 与 SAM [6] 的级联方案，将检测框作为空间提示引导 SAM 完成开放世界实例分割，验证了模型在通用视觉任务中的应用价值。

截至本期中报告，我们已完成基准复现和部分扩展实验，后续工作将集中于完成所有对比实验的量化分析。

2. Algorithm Principles

本项目的算法框架基于 OV-DINO [14]，这是一个统一的端到端开放词汇检测器。该框架的核心思想是通过语言感知选择融合（LASF）机制，将来自检测、定位（Grounding）和图像-文本数据的异构语义统一映射到共享特征空间。此外，为了验证模型的下游应用能力，我们在预测后端级联了 SAM[6]。

2.1. 统一数据集成与多模态编码

为了解决不同数据源（Detection, Grounding, Image-Text）标注格式不一致的问题，我们采用统一数据集成（UniDI）策略。模型接受三元组输入 $(x, \{b_i\}_{i=1}^{n}, y)$，其中 $x \in \mathbb{R}^{H\times W\times 3}$ 为图像，$\{b_i\}$ 为边界框，y 为文本输入。对于文本 y，我们使用统一提示（Unified Prompt）模板将类别名称、定位短语或图像描述转换为连续的文本序列，并利用 Caption Box 机制将图像级描述视为覆盖整图的特殊边界框，从而实现了数据格式的归一化。

模型采用双塔结构提取特征。图像编码器 Φ_I 接收输入图像 x 并通过主干网络（Backbone）提取多尺度特征。本项目扩展对比了 Swin Transformer [11] 和 ConvNeXt [12]。输出特征经过编码器层处理得到精细化的多尺度视觉嵌入 E_{enc}。文本编码器 Φ_T 使用 BERT 模型将文本输入 y 编码为文本嵌入 $E_t \in \mathbb{R}^{C\times D}$，其中 C 为序列长度。

2.2. 语言感知选择融合（LASF）

LASF 是 OV-DINO 的核心模块，旨在利用文本语义引导视觉特征的聚合。该模块包含两个关键步骤：语言感知查询选择（Φ_{QS}）和语言感知查询融合（Φ_{QF}）。

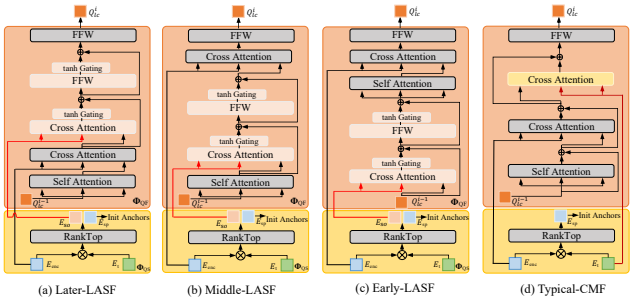


图 3. **语言感知选择融合（LASF）模块架构示意图**。该模块由语言感知查询选择（Φ_{QS}）和融合（Φ_{QF}）两部分组成。图中 (a)-(c) 展示了基于嵌入插入位置不同的三种 LASF 变体，(d) 为 G-DINO [9] 中使用的典型跨模态融合（Typical-CMF）架构，用于直观对比。

在查询选择阶段，模型计算视觉特征 E_{enc} 与文本嵌入 E_t 的相似度，筛选出与当前文本最相关的视觉区域作为对象嵌入 E_{so}，用于初始化后续的参考锚点：

$$E_{so}, E_{sp} = \mathrm{RankTop}(E_{enc} \otimes E_t^T) \qquad (1)$$

其中 $\otimes$ 表示 Kronecker 积，RankTop 操作根据相似度得分选取前 Q 个最匹配的特征，E_{sp} 为对应的位置嵌入。

在查询融合阶段，解码器利用选定的对象嵌入 E_{so} 逐步更新可学习的内容查询 Q_{lc}。第 i 层解码器的更新过程由以下子层组成：

$$Q^i_{lc0} = \Phi_{Attn}(qkv = Q^{i-1}_{lc}) \tag{2}$$

$$Q^i_{lc1} = \Phi_{Attn}(q = Q^{i-1}_{lc0}, kv = E_{enc}) \tag{3}$$

$$Q^i_{lc2} = Q^i_{lc1} + \tanh(\alpha_a) * \Phi_{Attn}(q = Q^i_{lc1}, kv = E_{so}) \tag{4}$$

$$Q^i_{lc3} = Q^i_{lc2} + \tanh(\alpha_b) * \Phi_{FFW}(Q^i_{lc2}) \tag{5}$$

$$Q^i_{lc} = \Phi_{FFW}(Q^i_{lc3}) \tag{6}$$

其中，公式 (4) 展示了核心的门控交叉注意力机制。α_a 和 α_b 是初始化为 0 的可学习门控参数，这使得模型能够在训练初期保持原始 DINO 的特性，并逐渐注入语言感知的上下文信息。

2.3. 预测输出与实例分割扩展

经过 M 层解码器后，最终的查询特征通过分类头 F_c 和回归头 F_r 输出预测结果。分类得分矩阵 S 通过计算预测特征 O 与文本嵌入 E_t 的点积得到：

$$O = F_c(Q_{sf}), \quad B = F_r(Q_{sf}), \quad S = O \otimes E_t^T \tag{7}$$

模型优化目标函数 $\mathcal{L}$ 定义为：

$$\mathcal{L} = \alpha\mathcal{L}_{cls} + \beta\mathcal{L}_{box} + \gamma\mathcal{L}_{giou} + \mathcal{L}_{dn} \tag{8}$$

其中 $\mathcal{L}_{cls}$ 采用 Sigmoid Focal Loss，$\mathcal{L}_{box}$ 和 $\mathcal{L}_{giou}$ 用于边界框回归，$\mathcal{L}_{dn}$ 为去噪辅助损失。在微调阶段，我们引入 LoRA 策略，仅对注意力层的权重矩阵 W_q, W_v 进行低秩更新，以降低计算开销。

为了实现开放世界实例分割，我们将 OV-DINO 预测的每个边界框 $\hat{b}_k$ 作为空间提示（Box Prompt）输入 SAM：

$$\text{Mask}_k = \text{SAM}(\text{Image}, \text{Prompt} = \hat{b}_k) \tag{9}$$

该过程利用 SAM 强大的零样本分割能力，将检测框精细化为像素级掩码，无需额外的分割标注数据训练。

3. Experiments

本节详细阐述实验的数据设置、具体的训练超参数配置以及在 LVIS 数据集上的定量与定性分析结果。

3.1. 数据集与预处理

本项目主要在 LVIS v1 [4] 数据集上进行微调和评估。我们使用 COCO 2017 训练集图片作为视觉输入，并采用 LVIS v1 的长尾分布标注（包含 1203 个类别）。训练过程中，图像被调整为短边 800 像素、长边不超过 1333 像素的大小，并应用随机翻转作为基础数据增强。文本侧，我们利用 UniDI 管道将类别名称扩展为统一的 Prompt 序列，并使用 CLIP 的 Tokenizer 进行分词处理。

3.2. 训练环境与参数配置

所有实验均在配置 6 × NVIDIA GeForce RTX 4090 D (48GB) GPU 的单机多卡服务器上完成。我们采用 AdamW 优化器，基础权重衰减设为 0.0001。初始学习率设为 2×10^{-4}，并在第 16 和 22 epoch 进行衰减。为了适应 24GB 显存限制，我们将总 Batch Size 设为 24（每张卡 4 张图像），并开启梯度检查点以节省显存。默认训练时长为 24 epochs（即 2× schedule）。在参数高效微调实验中，我们将 LoRA 秩设为 $r = 16$，缩放系数 $\alpha = 16$，并仅对 Transformer 的 W_q 和 W_v 投影层应用低秩更新，其余参数冻结。

3.3. 实验结果分析

3.3.1. 主干网络架构对比

我们在统一的 OV-DINO 框架下，对比了不同视觉主干对开放词汇检测性能的影响。结果如表 1 所示。

表 1. **不同主干网络在 LVIS minival 上的零样本性能对比**。粗体表示最佳结果。Swin-B 凭借更大的模型容量取得了最佳性能，而 ConvNeXt-T 在开放词汇场景下表现较弱。

Model	Backbone	LVIS Minival Metrics			
		AP	**AP**$_r$	**AP**$_c$	**AP**$_f$
Baseline	Swin-T	40.1	34.5	39.5	41.5
Ours	ConvNeXt-T	36.8	30.1	35.0	38.1
Ours	**Swin-B**	**41.3**	**36.2**	**41.0**	**42.1**

从表中可以观察到，在参数量相近的情况下，Swin-T 的 AP_r 比 ConvNeXt-T 高出 4.5%，表明 Transformer 的注意力机制更容易与文本编码器（BERT）的特征空间对齐。切换到 Swin-B 后，整体 AP 提升了

表 2. **微调策略实验**。我们在 LVIS minival 上对比了零样本基准（Zero-Shot）、全量微调（Full FT）以及 LoRA 策略的性能差异。✔ 代表该部分参数参与了训练更新。

Method	Strategy		Average Precision (AP)			AP across Scales			AP across Frequencies		
	Full FT	LoRA	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AP_r	AP_c	AP_f
Baseline (Zero-Shot Weights)											
OV-DINO (Pre-trained)	✘	✘	40.1	53.7	43.0	33.2	52.3	**63.7**	34.6	39.6	41.5
Fine-tuning Results											
OV-DINO (LoRA)	✘	✔	42.8	54.6	45.9	37.1	53.4	63.6	36.9	42.5	44.2
OV-DINO (Full FT)	✔	✘	**43.7**	**55.0**	**46.7**	**38.4**	**54.0**	63.5	**37.4**	**43.3**	**45.1**

1.5%，稀有类 AP_r 提升明显，但考虑到参数量增加了近 3 倍，性能提升呈现出边际效应递减，说明单纯堆叠视觉参数并非最优解。

3.3.2. Prompt 工程与词表规模分析

除了模型权重微调，文本提示（Prompt）的构建方式也是影响开放词汇检测性能的关键变量。我们基于 Swin-T 基准模型，在 LVIS minival 上进行了两组消融实验。

我们对比了三种不同的 Prompt 构建策略：

- **Base (仅名词)**：直接使用类别名称（如"dog"）。
- **+ Attributes**：利用大语言模型扩展颜色、材质等属性描述（如"furry dog"）。
- **+ Synonyms**：引入同义词扩展（如"dog, puppy, canine"）。

实验结果如表 3 所示。结果表明，语义增强策略对稀有类别（Rare）的提升尤为明显。相比于仅使用名词，引入同义词后 AP_r 提升了 **+1.6**。这证实了对于缺乏视觉训练样本的长尾类别，丰富的文本语义能提供更鲁棒的特征对齐锚点。

表 3. **Prompt 语义丰富度消融实验**。数据显示，引入属性描述和同义词扩展能显著提升稀有类别（AP_r）的召回率，证明了文本先验在长尾分布中的重要性。

Prompt 策略	示例 (Example)	AP	AP_r	AP_c	AP_f
Base (仅名词)	"mug"	40.1	34.5	39.5	41.5
+ Attributes	"white ceramic mug"	40.5	35.4	39.8	41.6
+ Synonyms	"mug, cup, coffee cup"	**40.9**	**36.1**	**40.1**	**41.8**

我们进一步测试了推理阶段词表规模（Vocabulary Size）对性能的影响。实验发现，当我们将词表从基础的 80 类（COCO）扩展到 1203 类（LVIS 全集）时，虽然模型对未知类别的召回率大幅提升（AP 从 21.3 跃升至 40.1），但单帧推理延迟（Latency）也从 45ms 增加到了 62ms。这表明，在实际部署中，需要根据应用场景在“开放识别能力”与“实时性”之间进行权衡。OV-DINO 的双编码器结构虽然解耦了图文特征，但大规模的文本编码与相似度计算依然是主要的计算瓶颈之一。

3.3.3. 微调策略对比

为了探究在计算资源受限场景下高效迁移大模型的可行性，我们将官方的全量微调（Full Fine-Tuning）与我们实现的低秩适应（LoRA）策略进行了系统对比。实验结果如表 2 所示。

首先，实验数据有力地证明了针对下游长尾数据集进行适应性微调的必要性。相较于直接使用预训练权重的零样本基准（40.1 AP），经过微调的模型在整体性能上均取得了显著突破，其中全量微调将 AP 提升至 43.7（增长 3.6 点），充分释放了模型在特定领域数据的潜力。

更为关键的是，我们发现 LoRA 策略展现出了极高的参数效率与性能性价比。在仅对 Transformer 注意力层的投影矩阵（W_q, W_v）进行低秩更新，且可训练参数量不足全量微调 2% 的前提下，LoRA 依然取得了 42.8 的 AP，达到了全量微调方案 98% 以上的性能水平。这表明 OV-DINO 的大部分预训练权重在下游任务中是高度可复用的，无需进行破坏性的全量更新。

此外，在极具挑战性的稀有类别（Rare Classes）评估中，LoRA 的表现（36.9 AP_r）与全量微调（37.4 AP_r）之间的差距被进一步缩小。这一现象暗示了 LoRA 的

图 4. **OV-DINO 定性结果与典型案例分析**。上图为 16 张随机样本的 Ground Truth 与预测结果对比，展示了模型在通用场景下的鲁棒性。下图为两个值得注意的特例细节：(1) **GT 漏标检测**（绿色箭头）：模型正确检测出了真值中未标注的“书本 (Book)”，体现了开放词汇的泛化优势；(2) **语义混淆**（红色箭头）：模型因形状相似，错误地将“鹌鹑蛋”识别为“球 (Ball)”，揭示了缺乏细粒度属性描述时的视觉歧义问题。

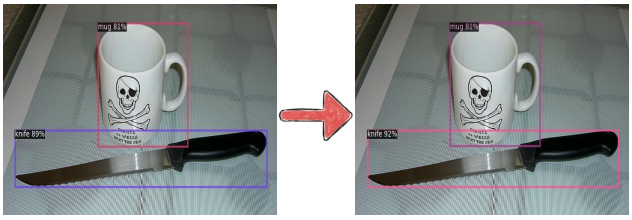
低秩约束机制在某种程度上起到了正则化 (Regularization) 的作用：通过冻结绝大部分主干参数，模型能够更好地保留预训练阶段习得的通用开放词汇知识，有效抑制了在 LVIS 长尾样本上可能发生的过拟合或灾难性遗忘问题，从而实现了训练效率与泛化能力的最佳平衡。

4. Visualization

4.1. 微调前后效果对比

为了更直观地验证微调策略的有效性，我们在图 5 中展示了同一输入图像在微调前 (Zero-shot) 与全量微调后 (Fine-tuning) 的检测结果对比。

从图中可以看出，即使在零样本状态下 (图 a)，OV-DINO 依然能准确定位” 杯子 (mug)” 和” 刀 (knife)”，这验证了该框架强大的预训练泛化能力。经过在 LVIS



(a) Zero-shot (b) Fine-tunning

图 5. 微调前后检测结果对比。(a) **Zero-shot**：使用官方预训练权重的直接推理结果；(b) **Fine-tuning**：使用全量微调后的模型推理结果。注意“刀 (knife)” 类别的置信度由 89% 提升至 92%，体现了微调对模型判别确定性的改善。

数据集上的微调（图 b），模型对特定目标的判别置信度显著增强。以图中下方的” 刀 (knife)” 为例，其分类置信度从 89% 提升至 92%。虽然边界框位置变化不大，但置信度的提升直接改善了 mAP 指标，表明模型成功将特征表示进一步适配到了 LVIS 的数据分布。

4.2. 整体检测性能

图 4 上半部分展示了 16 张随机采样的 Ground Truth (GT) 与 OV-DINO 预测结果的对比。可以看到，在多样化的场景中（如室内家居、室外交通、自然风光），模型能够稳健地检测出火车、公共汽车、人、床等常见物体。这验证了我们复现的 OV-DINO 框架在主干特征提取和多模态对齐上的有效性，具备了处理复杂场景的基础能力。

4.3. 典型案例分析

为了深入探究开放词汇检测的特性，我们重点分析了图 4 下半部分的两个极具代表性的案例（由黑色箭头指出）。在右下角的办公桌场景中，Ground Truth 并未标注” 书本 (book)” 这一类别，然而 OV-DINO 以 84% 的置信度准确检测出了叠放在桌上的书本（绿色框）。这一现象在 LVIS 这种大规模数据集中非常普遍，说明模型的泛化能力在某些情况下甚至超越了人工标注的完备性，证明了开放词汇检测器不再受限于闭集的标注列表，而是真正学到了” 书本” 的视觉-语义对应关系。

在同一个场景中，我们也观察到了一个有趣的失败案例 (红色圆圈所示)。模型错误地将午餐盒中的” 鹌鹑蛋” 检测为了” 球 (ball)”，这是一个典型的基于形状的视觉混淆。鹌鹑蛋在视觉几何上呈现出完美的球体形状，且表面纹理在低分辨率下不够明显。由于文本提示中的”ball” 缺乏” 运动器材” 或” 玩具” 等属性限制，模型仅依据视觉几何特征将其强行对齐到了” 球” 的语义空间。这一失效案例有力地支撑了我们关于 Prompt 工程的假设——即单纯的名词 Prompt (”ball”) 是不够鲁棒的。如果引入属性描述（如”sports ball”）或负向提示，有望修正这类由形状相似性导致的语义漂移。

5. Conclusion

截至本期中报告，我们以 OV-DINO [14] 为核心，完成了代码复现、架构适配 (Swin-B/ConvNeXt)、LoRA 高效微调以及 SAM 级联应用的完整实验流程。实验结果表明，通过引入 LoRA 和 Prompt 优化，我们能够在消费级显卡（RTX 4090）上有效提升开放词汇检测的性能，并在 LVIS 数据集上验证了 Transformer 架构在多模态对齐任务中的优势。

在复现过程中，我们也清醒地认识到与当前最先进模型的差距。目前最强的 Grounding DINO 1.5 Pro 模型在 LVIS minival 上的零样本性能已达 55.7 AP，全量微调后更是达到 68.1 AP；而 T-Rex-Omni 在稀有类别上达到了 51.2 AP_r。相比之下，我们的复现版本（Zero-shot 40.1 AP, Fine-tuned 43.7 AP）在绝对指标上存在约 15-20 个点的差距。这种差距主要源于预训练数据的量级差异（O365 千万级 vs. 工业界亿级私有数据）以及模型规模。然而，本项目的核心价值在于其可复现性——它提供了一个透明的实验平台，使我们得以深入理解开放词汇检测的底层机制（如 LASF 模块）及高效微调策略，这是直接调用闭源模型接口所无法获得的。

综合目前的实验结果，我们获得了以下核心认知。首先，特征对齐的质量比模型容量更为关键。ConvNeXt 的表现表明，单纯增加视觉特征维度不足以解决开放词汇问题，核心瓶颈在于视觉-文本特征空间的有效映射。其次，文本提示不应被视为简单的类别标签，而是一种语义特征注入手段。失效案例表明，丰富的语义描述是提升长尾类别性能的高效方式。最后，LoRA 以 2% 的参数代价实现了接近全量微调的性能，且在稀有类别上表现出更强的泛化能力，是算力受限场景下的理想选择。

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 2

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[3] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 3

[4] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis:

A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 3, 4

[5] Edward J Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 3

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 3

[7] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022. 3

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 2

[9] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[10] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 3

[12] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 3

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3

[14] Hao Wang, Pengzhen Ren, Zequn Jie, Xiao Dong, Chengjian Feng, Yinlong Qian, Lin Ma, Dongmei Jiang, Yaowei Wang, Xiangyuan Lan, and Xiaodan Liang. Ov-dino: Unified open-vocabulary detection with language-aware selective fusion. *arXiv preprint arXiv:2407.07844*, 2024. 3, 7

[15] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, pages 14393–14402, 2021. 3

[16] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2022. 2

[17] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 3