

1 [TODO: 1]
c

c

LMM-Det: 让大型多模态模型在目标检测中表现出色

Jincheng Li^{*} Chunyu Xie^{*} Ji Ao Dawei Leng[†] Yuhui Yin

360 AI Research

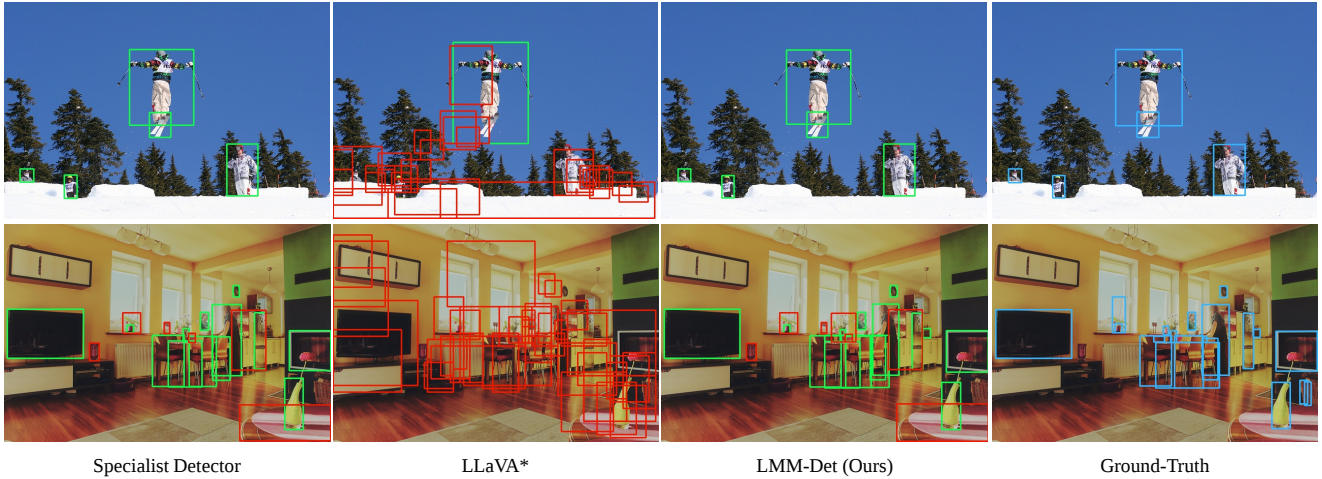


Figure 1. Visualizations of a specialist detector and large multimodal models for object detection on the validation set of COCO. For clarity, the green bounding boxes indicate the correctly predicted boxes, the red boxes represent the prediction errors, and the blue boxes denote the ground truth labels. Specifically, we employ Saliency-DETR [7] as the specialist detector. Following the prompt style of RefCOCO, we query the trained LLaVA-7B [15] with a sequence of questions, each targeting a single class category, which is called LLaVA *. In addition, a predicted box is deemed correct if it meets the following criteria: (1) The intersection over union (IoU) between the predicted box and the ground truth (GT) box is greater than 0.5; (2) The predicted class label matches the GT class label. As illustrated in the figure above, our approach (LMM-Det) enables a large multimodal model to perform object detection, achieving comparable results with state-of-the-art specialist detection methods. It is noteworthy that LMM-Det exhibits outstanding object detection capabilities without relying on additional specialist modules within its architecture. The detailed predicted labels of all images can be referred to the Figure A in Appendix.

Abstract

大型多模态模型 (LMMs) 由于其在多模态理解、推理和上下文学习等方面的卓越能力, 已经在人工智能研究和工业界中引起了广泛的关注和兴趣。虽然 LMMs 在处理图像字幕生成、视觉问答和视觉锚定等多模态任务上显示出令人鼓舞的结果, 但 LMMs 在目标检测方面的能力相比专业检测器存在显著差距。为了缩小这一差距, 我们摒弃了将重型检测器与 LMMs 结合的传统方法, 提出了 LMM-Det, 这是一种简单而有效的方法, 利用大型多模态模型进行普通目标检测, 而不依赖于专业的检测模块。具体来说, 我们进行了全面的探索性分析, 当大型多模态模型与目标检测结合时, 发现相较于专业检测模型, 召回率显著下降。为了缓解这一问题, 我们提出通过引入针对目标检测的数据分布调

整和推理优化来提高召回率。我们重新组织指令会话以增强大型多模态模型的目标检测能力。

1. 引言

大型多模态模型 (LMMs) [1, 5, 11, 15, 24, 26, 29] 最近因其在理解和处理多种形式的数 (如文本、图像、语音和视频) 方面的卓越能力而引起了广泛关注。LMMs 在弥合不同模态之间的差距方面显示出了显著的效果, blackblack 实现更高效的人机交互。特别是在各种多模态任务中表现出强大的性能, 如图像描述 [12]、视觉问答 (VQA) [31] 和视觉定位 [4]。在图像描述的情况下, 一个 GPT 助手为用户输入的图像生成详细的文本描述。此外, VQA 和视觉定位的目标是产生符合用户需求的详细 blackblack 和语境上适当的回应。这些

任务表明，大型多模态模型能够在细粒度水平上对齐图像和文本，这在物体检测和分割等细粒度图像-文本理解任务中具有 **blackblack** 显著潜力。

此外，获得物体的位置和类别对于有效的视觉理解是至关重要的。在 LMMs 中存在显著的差距。

更重要的是，他们没有充分发挥 LMMs 在执行目标检测任务时的潜力。例如，KOSMOS-2 将定位能力集成到下游应用中，并提供相应的边界框。Shikra 设计了一种简单的架构，无需使用前/后检测模块和外部插件模型，即可解锁 LMMs 的定位能力。

blackblack 观察到大型多模态模型表现出细粒度的图文对齐能力，并具备初步的检测能力 (*e.g.*, REC)，激发了我们进一步调查这一现象。我们的目标是探索大规模语言模型在不依赖额外专用检测模块的情况下执行目标检测任务的能力。**blackblack** 为此，我们强调大型多模态模型在遇到目标检测时面临的以下固有挑战。

大多数现有的大型多模态模型 [4, 17, 20] 在 Ref-COCO [9] 上评估其检测能力，而不是像 COCO [14] 这样专业的检测基准。blackblack 尽管使用了大型物体检测数据集（如 *e.g.*, Object365 [23]、LVIS [6]、Open-Images [10] 或 COCO）进行预训练。然而，在指令微调阶段经常忽略了用于 blackblackvanilla 物体检测的定制化指令数据，导致图像中物体的定位和识别不佳。这阻碍了机器人、自动驾驶和增强现实 [17] 等领域的实际应用。

如图 1 所示, 使用以下提示: **blackblack**“如果图像中存在该区域, 请提供该句子描述的区域边界框坐标: <类别>”。**blackblack** 对于每个查询, LLaVA 仅生成少量边界框, 其中大部分不准确, 导致目标检测任务的召回率较低。

在本文中，我们进行了综合分析，以研究专用目标检测模型和大规模多模态模型（LMMs）之间的性能差距。具体而言，我们在 COCO，一个专用目标检测基准上评估大规模多模态模型。通过样本可视化和分布比较，我们认为目标 **blackblack** 检测任务表现不佳的根本原因是低召回率。为了解决这个问题，我们提出了 LMM-Det，一种简单而有效的方法，使大规模多模态模型在目标检测中表现出色。**blackblack** 我们建议通过调整训练数据分布来提高召回率，因为我们观察到训练模型倾向于逼近这种分布。此外，我们引入了推断优化并呈现了一个指令微调数据集。总结来说，我们的主要贡献如下：

- 我们专注于探索大型多模态模型在实际应用中中对 **blackblack** 释放他们的全部潜力 的检测能力。为此我们提供了全面的分析以促进大型多模态模型的检测性能提升。
- 我们提出了一种简单的 **blackblack** 然而有效 方法，命名为 **LMM-Det**，不需要任何额外的检测模块。**blackblack** 我们引入数据分布调整和推理优化，以提高大型多模态模型在与目标检测结合时的固有召回率。
- **blackblack** 大量实验表明，**LMM-Det** 不仅表现出检测

- 我们提出了一种简单的 **blackblack** 然而有效 方法, 命名为 **LMM-Det**, 不需要任何额外的检测模块。**blackblack** 我们引入数据分布调整和推理优化, 以提高大型多模态模型在与目标检测结合时的固有召回率。

- blackblack 大量实验表明, LMM-Det 不仅表现出检测

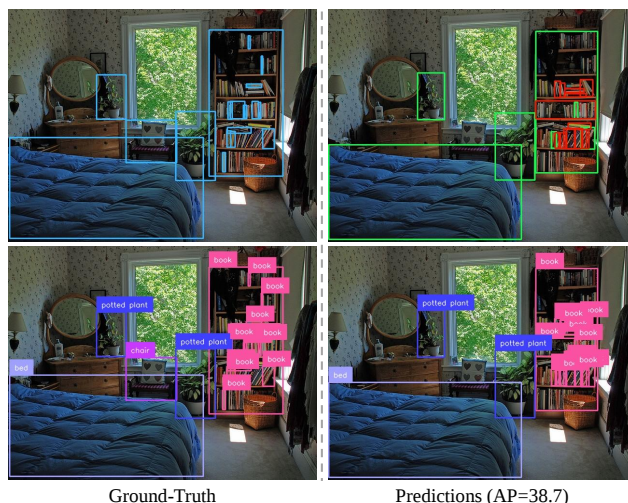


Figure 2. 由表格 1 第五行中的模型生成的真实值和预测边框的可视化。

能力,还保留了内在的多模态能力,如标题生成和视觉问答。

为了评估和分析大型多模态模型 (LMMs) 的检测性能, 我们首先从数据规模和图像分辨率的角度设计了一系列实验。然后, 我们进行深入分析, 以促进 LMMs 更好地适应目标检测。在本节中, 我们使用标准的大型多模态模型 LLaVA-7B 进行所有实验。我们选择 RT-DETR 作为比较的代表性专业检测器。

本节进行探索性实验以 `blackblack` 评估 LMMs 的检测能力，具体的实现细节在附录的 [第 A 节](#) 中提供。

在 COCO 数据集上的零样本实验。检验 LMMs 检测性能的一种直接方法是进行零样本评估。因此，我们首先在 COCO 数据集上使用 LLaVA 进行零样本实验。实验结果显示，LLaVA 在训练期间，在 COCO 数据集的零样本设置下表现出较差的检测性能。

blackblack 使用 **COCO** 进行监督微调。为了验证上述说法，我们利用 **COCO** 进行监督微调。如表格 1 的第二行和第三行所示，检测数据 (*i.e.*, **COCO**) 的加入使得 **AP** 显著提升，尽管仍有进一步改善的空间。受传统目标检测方法中扩大数据规模是一种有效提高检测准确性的技术这一事实的启发，我们尝试通过额外的检测数据进行增强，以使大型多模态模型在目标检测方面表现出色。

扩展检测数据。如前所述，我们通过训练期间加入 Object365 来扩展检测数据。从表 1 的第四行可以看出，增加更多的数据未能显著提升性能。可能的原因有：(1) 输入图像分辨率较低，不足以执行目标检测；(2) COCO 数据的指令组织需要改进。

提高图像分辨率。为了与专业检测器进行公平比较，我们采用插值法将图像分辨率从 336 提升到 644，从而使其与检测器的原生分辨率保持一致。

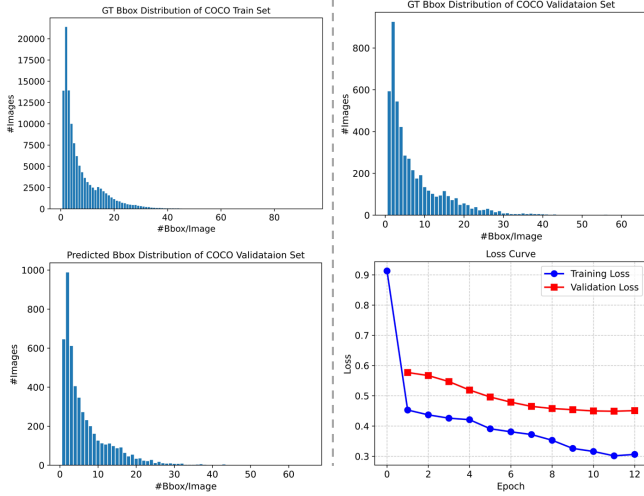


Figure 3. 真实值和预测的边界框 (Bbox) 分布。我们使用表 1 中第 5 行的模型 (AP=38.7) 来生成框。此外，我们提供损失曲线以确认模型没有过拟合。

Method	SFT data	COCO	Object365	Res.	AP	AP ₅₀	AP ₇₅	AR (100)
LLaVA	black(r,g,b), 0, 0x	black(r,g,b), 0, 0x	black(r,g,b), 0, 0x	336	0.2	0.6	0.2	11.2
	black(r,g,b), 0, 0x	black(r,g,b), 0, 0x	black(r,g,b), 0, 0x	336	14.5	29.5	12.7	24.3
	black(r,g,b), 0, 0x	black(r,g,b), 0, 0x	black(r,g,b), 0, 0x	336	16.0	31.2	14.4	26.7
	black(r,g,b), 0, 0x	black(r,g,b), 0, 0x	black(r,g,b), 0, 0x	336	15.6	28.5	15.1	21.5
	black(r,g,b), 0, 0x	black(r,g,b), 0, 0x	black(r,g,b), 0, 0x	blackblack 644	blackblack 17.7	blackblack 30.3	blackblack 17.5	blackblack 26.9
blackgray 专家检测器	black(r,g,b), 0, 0x	black(r,g,b), 0, 0x	black(r,g,b), 0, 0x	644	38.7	55.8	41.3	50.5
				blackgray 640	blackgray 49.2	blackgray 66.6	blackgray 53.5	blackgray 70.9

Table 1. 对 COCO 验证集的 LLaVA-7B 进行检查。“SFT 数据”表示在 LLaVA 中监督微调 (SFT) 期间的 665K 指令数据。“Res.”是图像分辨率。

1.1. 现象分析

blackblack 为了缩小 LMMs 和专业检测模型之间的性能差距，我们调查了其潜在原因。

预测和 GT 的可视化。blackblack 我们在图 2 的底部行中将生成的边界框与相应的类别以及真实值 (GT) 进行可视化展示。尽管在 AP 上比专家模型低 10.5%，但可视化结果出乎意料的令人满意。blackblack 根据图 1 中指定的标准，我们绘制了不同颜色的框，并在图 2 的顶行中提供了结果。我们观察到，红色的边界框被认为是假阳性，但实际上具有正确的预测标签。同时，真值 (GT) 的标签并未完全标注，e.g., “书”。此外，我们的统计结果（如附录中的表格 A 所示）表明，blackblackCOCO 的训练集和验证集大约为 7 中每张图像的平均边界框数量，也与训练模型产生的盒子平均数量相符。一个可能的原因是训练的模型捕捉到了 COCO 数据集的数据分布。blackblack 适应于

分布 blackblack 比较。我们提供了预测和实际值的箱型分布，以进一步分析上述观察结果。从图 3 可以看出，blackblack 预测的框分布近似于 COCO 的训练集。然而，不完整的真实标注导致预测过早截断，导致生成的边界框数量较少。此外，目前简单的自动回归训练方法导致预测的边界框少于专业检测模型。

实际上，传统的目标检测方法在候选区域上保持了一个适当的召回率，例如 Faster RCNN [22] 中的 300 个

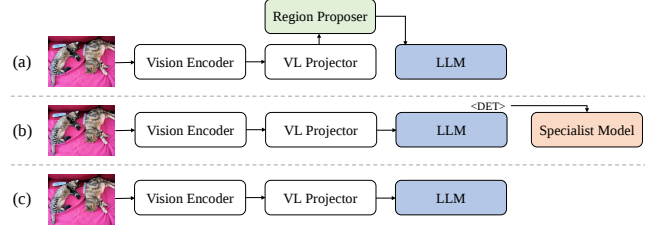


Figure 4. 提出的 LMM-Det 概述。相比于其他具有额外模块如区域建议生成器 (a) 或专门检测模型 (b) 的 LMMs，LMM-Det (c) 使得 LMMs 能够以直接的方式释放检测能力。

候选区域或 H-Deformable-DETR [8] 中的 900 个候选区域，以平衡检测性能和计算成本。在这种情况下，不足的召回率会显著降低检测性能。然而，由于大多模态模型 (LMMs) 在不完整真实标签下的下一个标记预测损失的固有限制，让大型多模态模型作为区域提议网络 (RPN) [22] 并生成大量高质量的候选区域是具有挑战性的。

2. LMM-Det

blackblack 本节介绍了 LMM-Det，它通过提高召回率增强了 LMMs 的检测能力。我们首先在第 2.1 节介绍模型架构。LMM-Det 的进展包括数据分布调整（第 2.2 节）和推理优化（第 2.3 节）。

2.1. 初步：模型架构

blackblack 如图 4 (c) 所示，LMM-Det 由一个视觉编码器、一个投影器和一个大型语言模型组成。

视觉编码器。LMM-Det 使用 OWLv2-ViT 模型 [18] 作为视觉编码器，该模型支持高分辨率图像输入 (1008 $\times$ 1008)，并捕获用于目标检测的详细信息。我们没有压缩视觉标记的序列并将它们 blackblack 直接输入到 LMMs 中。

投影算符。blackblackLLaVA [15] 证明了线性投影器足以实现视觉和语言的对齐。此外，Honeybee [3] 显示线性投影器可以通过一对一投影无损地保留视觉特征的所有局部上下文。因此，我们使用线性投影器将视觉特征映射到文本标记的嵌入空间中。

大型语言模型。我们将 blackblack 利用 Vicuna-1.5-7B 最大序列长度设为 16,000 作为大型语言模型。LMM-Det 被训练以使用语言建模损失的范式进行标记预测。形式上，给定一幅图像和组织良好的指令文本标记，我们最大化以下优化问题：

$$\max_{\theta} \sum_{i=1}^L \log p_{\theta}(\tilde{y}_i | x_v, x_t, y_{1:i-1}), \quad (1)$$

其中 θ 是可训练参数， x_v 来自视觉编码器和投影仪的视觉标记， x_t 是来自大型语言模型标记器的指令文本标记， $y_{1:i-1}$ 是当前预测标记 $\tilde{y}_i$ 之前指令数据中的答案标记， L 是输出序列的长度。

2.2. blackblack 数据分布调整

blackblack 如在第 ?? 节中所述, 训练集中的不完整真实标注可能导致大规模多模态模型的预测过早终止。为了缓解这一问题, 我们调整训练集的数据分布, 因为训练后的大规模多模态模型倾向于逼近这一分布。此外, 数据分布调整的过程应达到提高召回率的目标, 以增强大规模多模态模型的检测性能。blackblack 为此, 一种可能的方法是使用自训练策略, *i.e.*, 通过大型多模态模型生成伪标签并迭代地用真实标签进行训练。然而, 在不具备强大目标检测能力的大型多模态模型中, 生成高质量的伪标签可能面临挑战。在这项工作中, 我们使用半监督学习策略来调整训练数据分布, 增加召回率, 从而提高 LMMs 的检测精度。

blackblack 具体来说, LMM-Det 的数据分布调整包括三个步骤: 1) 伪标签生成。此步骤旨在为训练集中不完整的真实标签准备更多高质量的数据。2) 数据合并与设计。将制作的伪标签与原始的真实标签合并, 然后设计以便将目标检测应用于 LMMs。3) 标记表示验证。我们研究了不同的标记表示, 以代表坐标和置信分数。

blackblack 伪标签生成。我们通过高质量的伪标签增强真实标签, 从而提升标注的多样性。具体来说, 我们使用预训练的专业检测器 (*i.e.*, Saliency-DETR [7]) 制作若干伪标签。通过这样的方式, 我们开发了大量提案以改进大型多模态模型的整体检测性能。值得注意的是, 我们的修改仅涉及对数据的扩展, 而我们的 LLM-Det 架构在整个预训练和监督微调阶段不依赖于额外的检测模块。

blackblack 数据合并与设计。为了去除冗余框并进一步提高数据质量, 我们通过执行非极大抑制 (NMS) 合并生成的伪标签和原始的真实标签。此外, 由于缺乏每个预测框的置信度分数, 现有的数据组织在计算 COCO 平均精度时并不容易。在实践中, 我们实验证实通过计算每个坐标令牌的平均概率 (经过 softmax 处理后) 的置信度分数不可区分。为了解决这个问题, 我们让一个大型多模态模型输出坐标和相应的置信度分数, 而不仅仅是坐标。我们将真实标签的置信度分数设置为 1, 而伪标签的置信度分数由预训练的专业检测器决定。

blackblack 令牌表示验证。我们研究如何表示坐标及其对应的置信分数。我们探索了两种方案: 直接输出令牌预测和用额外的令牌扩展词汇表。第一种方法增加了模型输出令牌的长度, 而不需要额外训练词汇嵌入, 而第二种方法则相反, 虽然需要额外训练词汇嵌入, 但输出令牌变得更短。第 3.3 节中的实验表明, 第一种方法实现了更好的检测准确性。因此, 所有实验均采用该方法。

blackblack 通过调整训练数据分布, LMM-Det 能够生成更准确和数量更多的边界框, 这可以提高召回率, 从而改善整体检测性能。此外, 我们通过实验表明, LMM-Det 不仅可以近似检测数据分布, 还可以保持原有能力去近似其他数据分布, 例如图像字幕和视觉问答, 如第 3.2 节所示。

2.3. 为检测量身定制的推理优化

blackblack 在推理阶段, 对于大型多模态模型中的输入图像检测所有对象, 一个简单的方法是同时预测所有边界框, 例如 InternVL 2.5 [5]。然而, 我们从经验上观察到, 这种方法在使用固定采样策略的 LMMs 时很难生成足够的建议框。一个可能的原因是当前的 LMMs 难以在单一步预测中处理所有细粒度的图像细节。我们尝试优化这一解决方案, 使 LMM-Det 输出更多的建议框, 以提高推理时的召回率。

blackblack 具体来说, 我们在计算成本上做出妥协, 以获得 LMM-Det 更好的检测准确性。我们没有在一步中输出所有边界框, 而是让 LMM-Det 在一步中仅输出属于一个类别的所有物体, 并多次重复此步骤。也就是说, LMM-Det 独立地预测每个类别的边界框。为了在推理和训练之间保持一致性, 我们重新组织指令对话以采用特定类别的预测策略。

3. 实验

3.1. 实现细节

训练方案。我们在三个连续阶段训练 LMM-Det。附录中的表格 B 总结了所有阶段的超参数。blackblack 我们总共使用了 595K 个图像文本对和 1.86M 张图像来训练 LMM-Det。在一个由 6 个节点组成的集群上训练 LMM-Det 需要 176 小时, 每个节点配备 8 个 Nvidia H800 GPU。下面是每个阶段的详细信息。

阶段 I。具体来说, 我们利用了 blackblack 我们通过冻结视觉编码器和大型语言模型, 同时预训练投影器来对齐视觉和语言模块。在 LLaVA [15] 中使用的 59.5 万图文对。

第二阶段。我们利用一个大规模的目标检测数据集 (blackblack *i.e.*, Object365 [23]) 对提出的 LMM-Det 进行预训练。blackblack 在这个阶段, 我们在冻结视觉编码器的同时训练投影器和大型语言模型。

阶段 III。我们重新组织了一套对象检测指令数据集 (如附录中的图 B 所示), 其基于广泛使用的对象检测任务基准 COCO 数据集构建。然后, 我们使用这些指令数据对投影器和大型语言模型进行微调, 以进一步提高 LMM-Det 的检测能力。

blackblack (可选) 阶段 IV。在这个阶段, 投影器和大型语言模型在视觉编码器冻结的情况下进行训练, 使用 665k LLaVA [15] 数据集以及我们提议的重新组织的指令数据, 并采用与阶段 III 相同的超参数。我们将其称为 LMM-Det[†]。

推理与评价。我们在 COCO 的验证集上评估 LMM-Det 的检测能力。对于验证集的每个类别, 我们构建相应的提示进行提问。我们收集所有预测输出并将其视为最终的提案。为简单起见, 我们将 AP 表示为 mAP。我们采用 AP, AP₅₀, AP₇₅, AP_S, AP_M, AP_L 和 AR@100 作为指标。

在本节中, 我们将 LMM-Det 与最新的 LMMs 在零样本设定中进行比较。我们采用他们的 blackblack 官方培训模型并以零样本方式测试它们。实现细节在

Method	blackblack 视觉骨干网络	LLM	w. Specialist	AP	AP ₅₀	AP ₇₅	AP _s	AP _M	AP _L	AR @100
LLaVA [15]	blackblackCLIP 大模型	Vicuna-7B	<i>black[rgb]0, 0, 0</i> ×	0.2	0.6	0.2	0.0	0.1	0.7	11.2
Shikra [4]	blackblackCLIP-L	Vicuna-7B	<i>black[rgb]0, 0, 0</i> ×	0.4	0.8	0.4	0.0	0.4	1.0	18.7
KOSMOS-2 [20]	blackblackCLIP-L	MAGNETO-24-layers	<i>black[rgb]0, 0, 0</i> ×	7.6	13.7	7.3	0.8	6.7	15.8	18.2
blackblackInternVL-2.5 [5]	InternViT-300M	Internlm2.5-7B	<i>black[rgb]0, 0, 0</i> ×	11.8	18.4	12.0	3.6	13.0	23.4	27.5
Groma [17]	DINOv2	Vicuna-7B	<i>black[rgb]0, 0, 0</i> ✓	12.8	17.0	13.8	3.3	12.8	24.2	22.5
LMM-Det (Ours)	blackblackOWLv2-L	Vicuna-7B	<i>black[rgb]0, 0, 0</i> ×	24.5	34.7	26.3	15.4	27.4	37.3	46.6

Table 2. COCO 数据集上的零样本结果与最新的多模态大模型相比。我们没有报告 VisionLLM v2 [25] 的零样本结果，因为他们没有发布预训练的检查点。“w. Specialist” 代表 “带有专业检测模型”。

Method	blackblack 视觉主干	LLM	w. Specialist	AP	AP ₅₀	AP ₇₅	AP _s	AP _M	AP _L	AR @100
Faster R-CNN [22]	RN50-FPN	-	<i>black[rgb]0, 0, 0</i> ✓	40.2	61.0	43.8	24.2	43.5	52.0	54.0
Cascade R-CNN [2]	RN101-FPN	-	<i>black[rgb]0, 0, 0</i> ✓	42.7	61.6	46.6	23.8	46.2	57.4	-
Deformable-DETR [32]	RN50	-	<i>black[rgb]0, 0, 0</i> ✓	43.7	63.0	47.6	26.7	47.0	58.0	63.2
RT-DETR [30]	RN50	-	<i>black[rgb]0, 0, 0</i> ✓	55.3	73.4	60.0	38.0	59.9	71.6	74.4
H-Deformable-DETR [8]	Swin-L	-	<i>black[rgb]0, 0, 0</i> ✓	56.1	75.1	61.3	39.2	60.4	72.4	73.1
Salience-DETR [7]	FocalNet-L	-	<i>black[rgb]0, 0, 0</i> ✓	57.3	75.5	62.4	40.9	61.8	74.5	75.4
Groma [17]	blackblackCLIP-L	Vicuna-7B	<i>black[rgb]0, 0, 0</i> ✓	32.4	-	-	-	-	-	-
Groma [17]	DINOv2	Vicuna-7B	<i>black[rgb]0, 0, 0</i> ✓	43.6	-	-	-	-	-	-
VisionLLM v2 [25]	Swin-T	Vicuna-7B	<i>black[rgb]0, 0, 0</i> ✓	56.3 ↓	74.3	61.6	-	-	-	-
blackgray 基础 DINO [16]	blackgraySwin-T	blackgray -	blackgray <i>black[rgb]0, 0, 0</i> ✓	blackgray 57.2	blackgray -	blackgray -	blackgray -	blackgray -	blackgray -	blackgray -
blackblack 狮鹫-13B [28]	blackblackCLIP-L	LLaMA2-13B	<i>black[rgb]0, 0, 0</i> ×	24.8	40.6	25.1	5.9	25.5	48.7	-
blackblack 格里芬 v2 [27]	blackblackEVA2-CLIP-L	LLaMA2-13B	<i>black[rgb]0, 0, 0</i> ×	38.5	54.3	41.2	19.4	43.2	57.6	-
LLaVA* [15]	blackblackCLIP-L	Vicuna-7B	<i>black[rgb]0, 0, 0</i> ×	38.7	55.8	41.3	20.1	43.6	57.3	50.5
LMM-Det (Ours)	blackblackOWLv2-L	Vicuna-7B	<i>black[rgb]0, 0, 0</i> ×	47.5	66.5	51.1	34.7	51.8	60.3	63.6
LMM-Det [†] (Ours)	blackblackOWLv2-L	Vicuna-7B	<i>black[rgb]0, 0, 0</i> ×	47.1	66.2	50.5	35.0	51.6	60.1	63.1

Table 3. blackblack 在 COCO 上的微调结果与传统的最新检测模型和依赖额外专用检测模型的大型多模态模型相比（使用专家）。VisionLLM v2 使用 Grounding DINO 作为额外的专用模块进行物体检测，但这种集成会影响 Grounding DINO 本身的检测性能。LLaVA* 表示我们用 Object365 和 COCO 重新训练 LLaVA。LMM-Det[†] 表示我们应用了可选的第四阶段来获得一个通用的 LMM-Det。

blackblack 附录 中给出。LMM-Det 仅采用如第 3.1 节中描述的两个阶段，并舍弃了第三阶段。

blackblack 从表格 2 可以看出，LMM-Det 在 COCO 上以零样本的方式取得了最佳结果，展示了其在检测数据训练后进行检测的潜力。LLaVA 未在检测数据上训练，仅在 COCO 上达到了 0.2 AP。像 blackblackInternVL-2.5 这样的模型利用大量检测数据，表现优于那些使用有限数据的模型。Groma 结合了一个专业的检测模型，表现也相当出色。在没有大量检测数据和专业模型的情况下，LMM-Det 显著优于其他 LMMs，验证了我们的现象分析和我们提出方法的有效性。

我们进一步在 COCO 上微调 LMM-Det。我们将 LMM-Det 与传统检测模型以及依赖于额外检测专家进行对象检测任务的 LMM 进行比较（如 Groma [17] 和 VisionLLMv2 [25]）。

blackblack 如表 3 所示，即使将 Groma 与专业检测模型结合，传统的最新检测模型仍然表现优于 Groma。同样，VisionLLMv2 尽管结合了 Grounding DINO，其表现也出现下降。我们还使用 COCO 和 Object365 数据集重新训练了 LLaVA。实验结果显示，传统的最新检测模型与大型多模态模型之间存在显著的性能差距。然而，LMM-Det 缩小了这一差距，支持了大型多模态模型可以在不需要额外专业检测模型的情况下本身执行物体检测的观点。

3.2. 多功能 LMM-Det

blackblack 我们在图 5 中提供了在 COCO 验证集上的 LMM-Det 可视化，展示了其在没有额外专业检测器

Model	COCO AP	Image Captioning				VQAv2 Accuracy
		BLEU@4	METEOR	CIDEr	SPICE	
LLaVA	0.2	29.4	29.3	108.9	23.6	78.5
blackblackLMM-Det [†]	47.1	26.7	28.1	99.0	22.4	74.1

Table 4. 关于多功能 blackblackLMM-Det[†] 的定量结果。

的情况下的检测性能。此外，定性和定量结果（分别见图 6 和表 4）证明了 LMM-Det 的多功能性，支持 blackblackLMM-Det[†] 不仅能释放检测能力，还能在图像描述和视觉问答中保持高性能。blackblack 我们在附录的表格 D 中提供了更多结果。

3.3. 消融研究

视觉编码器。在消融研究中，我们首先用 OWLv2-ViT [18] 替换 CLIP-ViT [21]，以进一步提高输入分辨率。如表 5 的第一行所示，该策略在 AP 中实现了 3.4 % 的提升。blackblack 我们在附录的表格 ?? 中进一步研究了 DINOv2 [19] 的有效性。

blackblack 数据分布调整。这种方法旨在解决召回率不足的挑战。blackblack 在表格 5 中，它有效地提高了召回率并增强了整体检测性能。blackblack 我们还在附录的图 C 中绘制了调整后的边界框分布。

blackblack 推断优化。我们进一步研究了推断优化的有效性。从表格 5 中可以看到，该策略导致了 AP 从 44.2 % 增加到 47.5 %，以及 AR @100 从 56.0 % 增加到 63.6 % 的显著提升。为了更好地可视化召回率的增强，我们扩展了图 2，并在附录的图 D 中展示了定性结果。



Figure 5. COCO 数据集上 LMM-Det 的可视化结果。

blackblackOWLv2-ViT	DDA	INO	AP	AP ₅₀	AP ₇₅	AR @100
<code>black[rgb]0,0,0×</code>	<code>black[rgb]0,0,0×</code>	<code>black[rgb]0,0,0×</code>	38.7	55.8	41.3	50.5
<code>black[rgb]0,0,0✓</code>	<code>black[rgb]0,0,0×</code>	<code>black[rgb]0,0,0×</code>	42.1	57.8	45.8	51.3
<code>black[rgb]0,0,0✓</code>	<code>black[rgb]0,0,0✓</code>	<code>black[rgb]0,0,0×</code>	44.2	61.3	47.5	56.0
<code>black[rgb]0,0,0✓</code>	<code>black[rgb]0,0,0✓</code>	<code>black[rgb]0,0,0✓</code>	47.5	66.5	51.1	63.6

Table 5. 消融研究。第一行的基线实验是 LLaVA* [15]，它使用 CLIP-ViT 作为视觉编码器。DDA: 数据分布调整；INO: 推理优化。

额外词汇的有效性。blackblack 我们研究了不同的标记表示法来表示坐标和置信度分数。blackblack 从附录中的表格 F 可以看出，直接输出标记比额外的词汇表能取得更好的性能，这一结果与 Shikra 的发现 [4] 一致。

采样策略的有效性和推理成本。blackblack 我们在推理过程中使用贪婪解码、束搜索和 top-p 采样进行消融研究。如附录中的表格 G 所示，使用 beam=2 的束搜索获得了最佳结果，但需要更多的推理时间。特别是，LMM-Det 处理单个图像大约需要 4.0 秒的计算时间。

4. 相关工作

4.1. 大型多模态模型

关于大型多模态模型 (LMMs) 的开创性工作主要集中在视觉-语言对齐，并支持基础的多模态任务，如图像描述和视觉问答。在这方面，LMMs 展现了其在执行多模态任务上的巨大潜力。为了进一步展示视觉-语言任务的全面能力，后续的大型多模态模型支持分辨率敏感任务，如视觉定位和光学字符识别 (OCR)。这些模型利用精心策划的预训练和指导数据集，展示了一种有效的战略，突显了它们强大的能力。这类 LMMs

包括了一系列广泛的模型，例如 Shikra, KOSMOS-2, Ferret v2, MiniCPM-V 2.6, DeepSeek-VL2, InternVL-2.5 和 Qwen2.5-VL。与粗粒度的视觉问答相比，这些 LMMs 的目标定位能力是在实际应用中，如代理、机器人技术、自动驾驶和安全监控中有效部署 LMMs 的关键先决条件。在本文中，我们研究了 LMMs 的检测潜力。

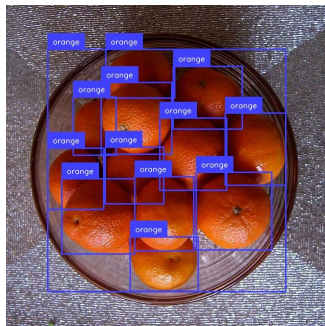
4.2. LMMs 的检测能力探索

虽然 LMMs 在大多数视觉语言基准测试中表现出色，但它们在物体检测任务上难以取得相当的性能。blackblack 现有的方法 [5, 13, 17, 25, 27, 28] 尝试在大型多模态模型上实现目标检测。例如，VisionLLM-v2 [25] 将 Grounding-DINO [16] 引入 LMMs，并使用一个特殊的标记 <DET> 来执行物体检测。Groma [17] 使用区域建议器来发现感兴趣的区域，从而激发 LMMs 的检测能力。而本文旨在研究在没有专门的检测模型或建议网络帮助下的大型多模态模型的物体检测能力。

5. 结论和局限性

LMM-Det 已经解决了 LMM 与专业检测器之间的检测性能差距，而不依赖于额外的模块。blackblack 具体而言，我们提供了全面的探索分析，并总结出关键挑战是召回率不足。我们使 LMM-Det 在保持生成标题和 VQA 等固有能力的同时，在目标检测中表现出色。此外，LMM-Det 存在 blackblack 不可忽略的推理时间延迟，这使它在与传统实时检测相比时竞争力较低。在未来，我们将尝试减少 LMM-Det 的推理时间。

Visual Input Example



User: Detect all the objects in the image that belong to the category set {orange}.

LMM-Det: [0.388, 0.722, 0.601, 0.891], 0.773; [0.298, 0.465, 0.322, 0.539], 0.183; [0.130, 0.446, 0.306, 0.637], 0.758; [0.310, 0.443, 0.493, 0.620], 0.751; [0.130, 0.446, 0.402, 0.779], 0.147; [0.311, 0.138, 0.527, 0.290], 0.771; [0.130, 0.446, 0.602, 0.893], 0.136; [0.130, 0.138, 0.528, 0.638], 0.179; [0.130, 0.138, 0.874, 0.893], 1.000; [0.298, 0.242, 0.518, 0.465], 0.791; [0.594, 0.521, 0.831, 0.763], 0.848; [0.480, 0.352, 0.690, 0.578], 0.781; [0.210, 0.286, 0.344, 0.468], 0.729; [0.402, 0.535, 0.587, 0.767], 0.791; [0.522, 0.192, 0.739, 0.386], 0.780; [0.298, 0.242, 0.690, 0.580], 0.137; [0.685, 0.337, 0.874, 0.562], 0.838; [0.174, 0.542, 0.402, 0.775], 0.790; [0.298, 0.138, 0.739, 0.466], 0.167; [0.300, 0.242, 0.874, 0.766], 0.156; [0.497, 0.281, 0.554, 0.373], 0.324

User: Detect all the objects in the image that belong to the category set {apple}.

LMM-Det: There are no objects in the image that belong to the required category set.

User: Describe the image in detail.

LMM-Det: The image features a clear glass bowl filled with a variety of oranges. There are at least 13 oranges in the bowl, with some of them being closer to the edge of the bowl and others being more centrally located. The oranges are arranged in a way that allows for a clear view of their vibrant colors and textures. The bowl is placed on a table, which is covered with a grey tablecloth. The oranges are arranged in a way that showcases their unique shapes and sizes, creating an appealing and visually pleasing display.

User: What color are oranges in the image?

LMM-Det: The oranges in the image are orange.

Figure 6. 聊天实例展示了 LMM-Det 不仅可以实现检测, 还保留了其固有的多模态功能。左上图代表输入图像, 左下图则显示输出坐标的可视化。为了更好地可视化, 我们对输出坐标应用后处理, 使用阈值为 0.5 的 NMS。

6. blackblack 致谢

blackblack 我们衷心感谢来自南洋理工大学的牛帅城和 360 人工智能研究院的朱晓乐, 他们为我们提供了宝贵的讨论和反馈。

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, pages 23716–23736, 2022.
- [2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018.
- [3] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *CVPR*, pages 13817–13827, 2024.
- [4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.
- [5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [6] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019.
- [7] Xiuquan Hou, Meiqin Liu, Senlin Zhang, Ping Wei, and Badong Chen. Saliencetr: Enhancing detection transformer with hierarchical salience filtering refinement. In *CVPR*, pages 17574–17583, 2024.
- [8] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. *CVPR*, 2023.
- [9] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014.
- [10] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from <https://github.com/openimages>*, 2017.
- [11] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.
- [12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023.

- [13] Chuang Lin, Yi Jiang, Lizhen Qu, Zehuan Yuan, and Jianfei Cai. Generative region-language pretraining for open-ended object detection. In *CVPR*, pages 13958–13968, 2024.
- [14] 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.
- [15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.
- [16] 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.
- [17] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. *arXiv preprint arXiv:2404.13013*, 2024.
- [18] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *NeurIPS*, 36, 2024.
- [19] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *ICLR*, 2025.
- [20] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.
- [21] 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, 2021.
- [22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, 2016.
- [23] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019.
- [24] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [25] Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Wenhui Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *NeurIPS*, 2024.
- [26] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- [27] Yufei Zhan, Yousong Zhu, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon v2: Advancing multimodal perception with high-resolution scaling and visual-language co-referring, 2024.
- [28] Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *ECCV*, pages 405–422. Springer, 2025.
- [29] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.
- [30] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Dets beat yolos on real-time object detection, 2024. *CVPR*.
- [31] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [32] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 2021.

A. 更多实现细节。

COCO 的统计结果列在表 A 中。

	train	validation
# images	118,287	5,000
# bboxes	860,001	36,781
# bboxes per image	7.3	7.4

Table A. COCO 的统计结果。

LMM-Det 在三个阶段的训练超参数列在表格 B 中。

Configuration	Stage I	Stage II	Stage III
Training epochs	1	5	12
Global batch size	192	480	288
Learning rate	1e-3	2e-5	2e-5
Learning rate schedule	Cosine decay		
Warmup ratio	0.03	0.05	0.05
Weight decay	0	0.05	0.05
Optimizer	AdamW		
Optimizer hyperparameters	$\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$		
Deepspeed ZeRO stage	ZeRO-2	ZeRO-3	ZeRO-3
Text max sequence length	2k	4k	10k
Training precision	bf16		

Table B. 三个阶段训练超参数。

探索性实验的更多细节。blackblack 在第 ?? 节，我们使用检测数据 (*i.e.*, Object365 和 COCO) 重新训练 LLaVA，采用表格 B 中阶段 III 的相同超参数，除文本最大序列长度 blackblack 在探索性实验中设为 2k。在训练和推理阶段，我们同时输出所有预测边界框。此外，在进行 COCO 上的零样本检测时，我们还为每个类别预测边界框。不幸的是，这种方法未能提高检测性能，反而增加了不正确的预测边界框数量。外。

更多 blackblack 训练和推理中的实现细节。在训练过程中，我们为每个输入图像构建多轮对话。为了减轻潜在的过拟合，我们在每个训练时期，对话的每一轮以及目标序列中的每个边界框进行随机化。blackblack 对于来自 COCO 或 Object365 的每一张图像，我们根据图像中现有类别的数量按 1:1 的比例构建正面和负面的对话。具体来说：1) 对于含有 n 个可见类别的图像 (例如，猫，狗)，我们生成 n 个正面指令，要求模型输出边界框。2) 然后，我们通过从剩余标签集中随机选择不存在的类别 (COCO 中的标签集为 *i.e.*, $80 - n$ ，Object365 中的为 $365 - n$) 来抽取 n 个负面指令。3) 每张图像的指令轮数上限分别为 COCO 的 80 次和 Object365 的 365 次。重要的是，我们不对 Object365 数据集进行筛选，而是保留其所有类别和实例，以保留模型的广泛检测能力。

blackblack 在推理过程中，我们预测带有置信度分数的边界框。特别是，由于伪标签的集成和重新组织的指令数据，平均每张图像的边界框数量从 7 增加到 31。然而，计算 AP 时生成的提案数量仍低于专门化模型

(例如，Saliency-DETR 中的 900 个提案)。因此，我们没有应用非极大值抑制 (NMS)，而是设置了一个阈值来计算 AP 和 AR。为了更好的可视化，我们使用了 0.5 的分数阈值和 0.5 的 NMS 阈值。blackblack 零样本实验的更多细节。在第 ?? 节中，我们以零样本的方式将 LMM-Det 与多种大型多模态模型进行比较。表格 C 中提供了所有模型的详细提示。blackblack 为简单起见，对于 KOSMOS-2 和 Groma，我们省略了 <image> 标记。我们在 QwenVL-2.5 的提示中省略了特殊标记，比如 <lim_start> 和 <lim_end>。特别是，我们随机选择 Shikra 中指代表达理解 (REC) 任务的模板，在 COCO 上进行零次试验。我们在表格 C 中举了一个 Shikra 的例子。

关于多功能 LMM-Det 的更多细节。blackblack 在第 3.2 节，实验表明，LMM-Det 不仅解锁了目标检测功能，还保留了图像标注和视觉问答等固有的多模态能力。具体来说，在按照第 3.1 节中的训练方案进行的三个阶段之后，我们添加了第四阶段。在这个阶段，我们冻结视觉编码器，训练投影器和大型语言模型，使用来自 LLaVA [15] 的 665K 数据和重新组织的指令数据。我们使用与 LLaVA 微调阶段相同的超参数。

B. 更多量化结果。

blackblack 本文揭示了在目标检测 (ODet) 领域中 LMMs 表现欠佳的根本原因。LMM-Det 可以处理多模态任务 (*e.g.*, ODet + 指称表达理解 (REC) + 图像描述 + 视觉问答)，而 Griffon [28] 和 Griffon v2 [27] 则专注于 ODet+REC。表格 D 进一步表明 LMM-Det[†] 也可以将目标检测与 REC 任务统一，并展示了互惠的性能提升 (例如， $81.4 \Rightarrow 85.7$)。

Model	COCO	RefCOCO val	MMStar
LLaVA-7B	0.2	81.4	30.3
Griffon-13B	24.8	88.0	-
Griffon v2-13B	38.5	89.6	-
LMM-Det [†] -7B	47.1	85.7	32.1

Table D. 更多关于多功能 LMM-Det[†] 的定量结果。

blackblack 表格 ?? 显示 OwlV2-ViT 比其他视觉编码器表现更好。特别是，这些实验 (表格 ??、F、G) 为了简化仅在 COCO 上训练。

额外词汇的效果显示在表格 F 中。

采样策略的有效性列在表 G 中。blackblack 我们在手稿的所有实验中使用贪婪解码作为推理采样策略。

C. 更多的定性结果。

在手稿的图 1 中所有图像的预测边界框显示在图 A 中。

blackblack 在推理优化中，我们重新组织了目标检测指令数据，以保持推理和训练之间的一致性。我们在图 B 中展示了一个例子。

Model	Multi-step	CLIP emb	Prompt
LLaVA [15]	<code>black[rgb]0, 0, 0✓</code>	<code>black[rgb]0, 0, 0×</code>	<code><image>\nProvide the bounding box coordinate of the region this sentence describes if region exists in the image: <category></code>
Shikra [4]	<code>black[rgb]0, 0, 0✓</code>	<code>black[rgb]0, 0, 0×</code>	<code>May I have the coordinates of <category> in <image>?</code>
KOSMOS-2 [20]	<code>black[rgb]0, 0, 0✓</code>	<code>black[rgb]0, 0, 0×</code>	<code><grounding> Where is the <category>?</code>
InternVL-2.5 [5]	<code>black[rgb]0, 0, 0×</code>	<code>black[rgb]0, 0, 0✓</code>	<code><image>\nPlease detect and label all objects in the following image and mark their positions.</code>
Groma [17]	<code>black[rgb]0, 0, 0×</code>	<code>black[rgb]0, 0, 0✓</code>	<code>[grounding] Please summarize the content of this image in detail.</code>
LMM-Det (Ours)	<code>black[rgb]0, 0, 0✓</code>	<code>black[rgb]0, 0, 0×</code>	<code><image>\nDetect all the objects in the image that belong to the category set <category>.</code>

Table C. blackblack 详细提示，执行 COCO 上的零样本目标检测任务。“多步”表示是否使用多步骤推理来预测图像。对于每张图像，我们构建了 80 个步骤让 LLaVA 在 COCO 上预测边界框。“CLIP emb”表示是否使用 CLIP 嵌入。通过这种方式，我们将未知类别映射到预定义类别（e.g., 80 个 COCO 上的类别）。

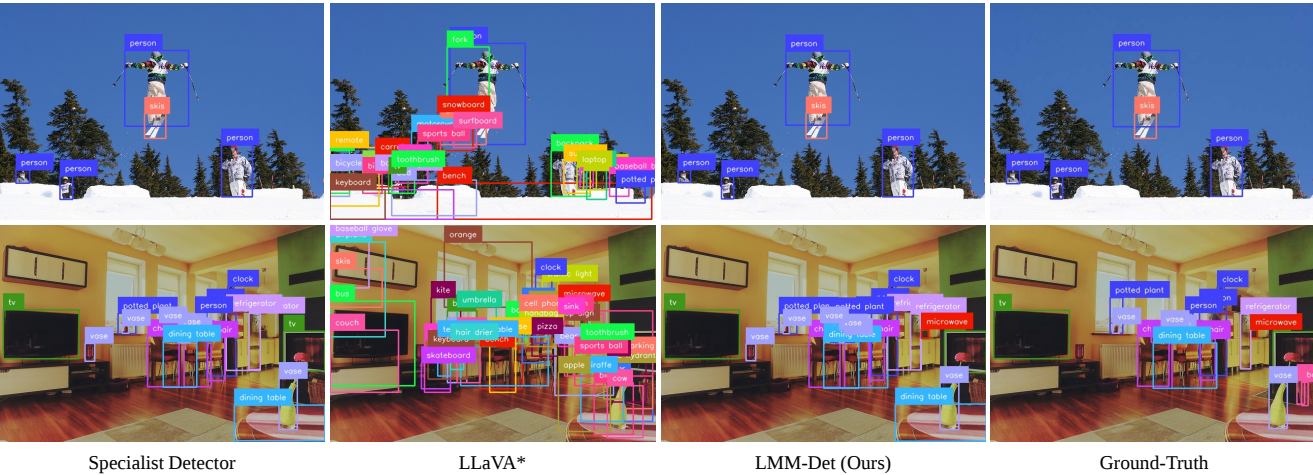


Figure A. 图 1 的详细预测标签在原稿中。

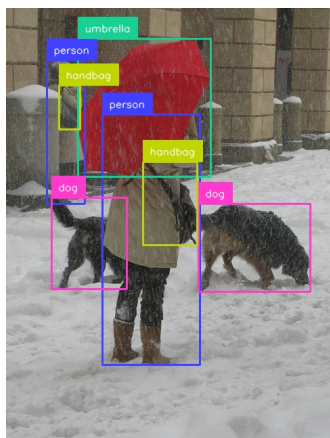
Techniques	AP	AP ₅₀	AP ₇₅	AR @100
LMM-Det	32.6	50.5	34.4	43.1
+ extra vocabularies	29.2	47.3	29.9	40.8

Table F. 额外词汇的有效性。

Model	Techniques	AP	AP ₅₀	AP ₇₅	AR @100	Cost/Img
LMM-Det	greedy decoding	32.6	50.5	34.4	43.1	4.0 s
	beam search (beam=2)	33.0	51.3	34.9	43.7	8.2 s
	beam search (beam=3)	32.8	51.1	34.5	43.6	10.9 s
	beam search (beam=4)	24.6	38.2	25.8	32.5	13.8 s
	top-p sampling	27.6	45.6	28.0	41.7	4.3 s

Table G. 推理采样策略。Cost/Img 表示单张图片的推理计算成本，其是在不使用 TensorRT 的情况下在一台 Nvidia H800 GPU 上测量的。我们对 COCO 验证集上所有图片的推理成本取平均值来获得 Cost/Img。

blackblack 我们在图 C 中进一步提供了调整后的边界框分布。从图 C 可以看出，数据分布调整有效地提高了召回率并改善了整体检测性能（模型的 AP=47.5）。如论文中的消融研究所示，LMM-Det 的 AP 在引入所提出的策略后从 38.7 提高到 47.5。我们在图 D 中提供了这些显著结果的定性结果，以展示召回率的提升。



User: <image>\n Detect all the objects in the image that belong to the category set {person}.

LMM-Det: [0.095, 0.121, 0.182, 0.452], 1.000; [0.223, 0.246, 0.448, 0.824], 1.000

User: Detect all the objects in the image that belong to the category set {dog}.

LMM-Det: [0.106, 0.439, 0.279, 0.649], 1.000; [0.445, 0.452, 0.702, 0.656], 1.000

User: Detect all the objects in the image that belong to the category set {handbag}.

LMM-Det: [0.123, 0.179, 0.171, 0.280], 1.000; [0.317, 0.356, 0.442, 0.548], 0.510

User: Detect all the objects in the image that belong to the category set {umbrella}.

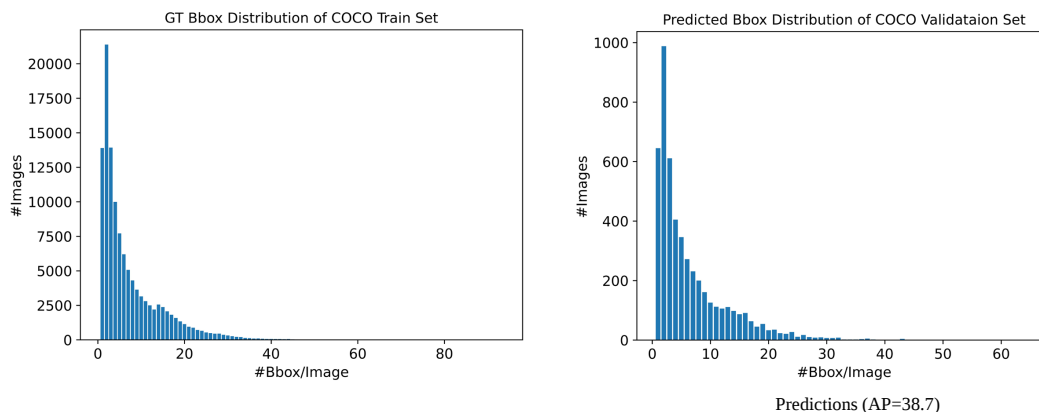
LMM-Det: [0.167, 0.071, 0.472, 0.390], 1.000

User: Detect all the objects in the image that belong to the category set {motorcycle}.

LMM-Det: There are no objects in the image that belong to the required category set.

Figure B. blackblack 重组数据的示意图是针对大型多模态模型中的目标检测设计的。关于边界框预测置信度和 blackblackLMM-Det 的答案 的后处理的更多细节，可以参考附录的第 A 节。

Before data distribution adjustment



After data distribution adjustment

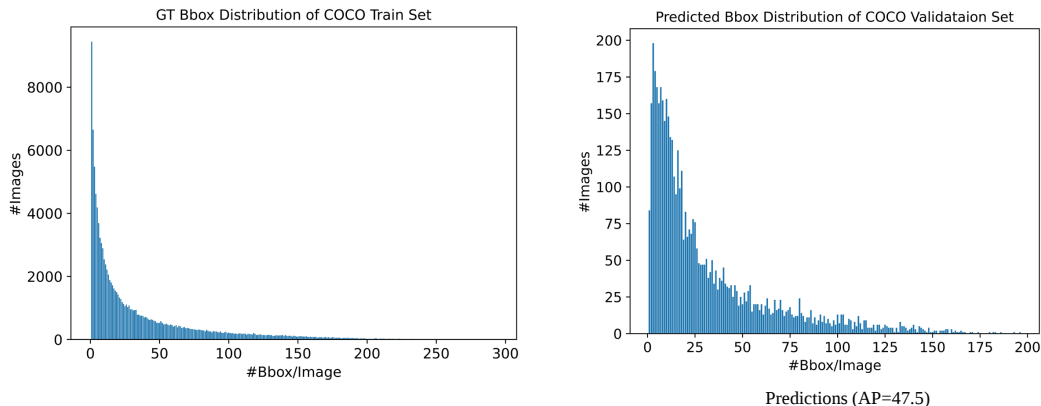
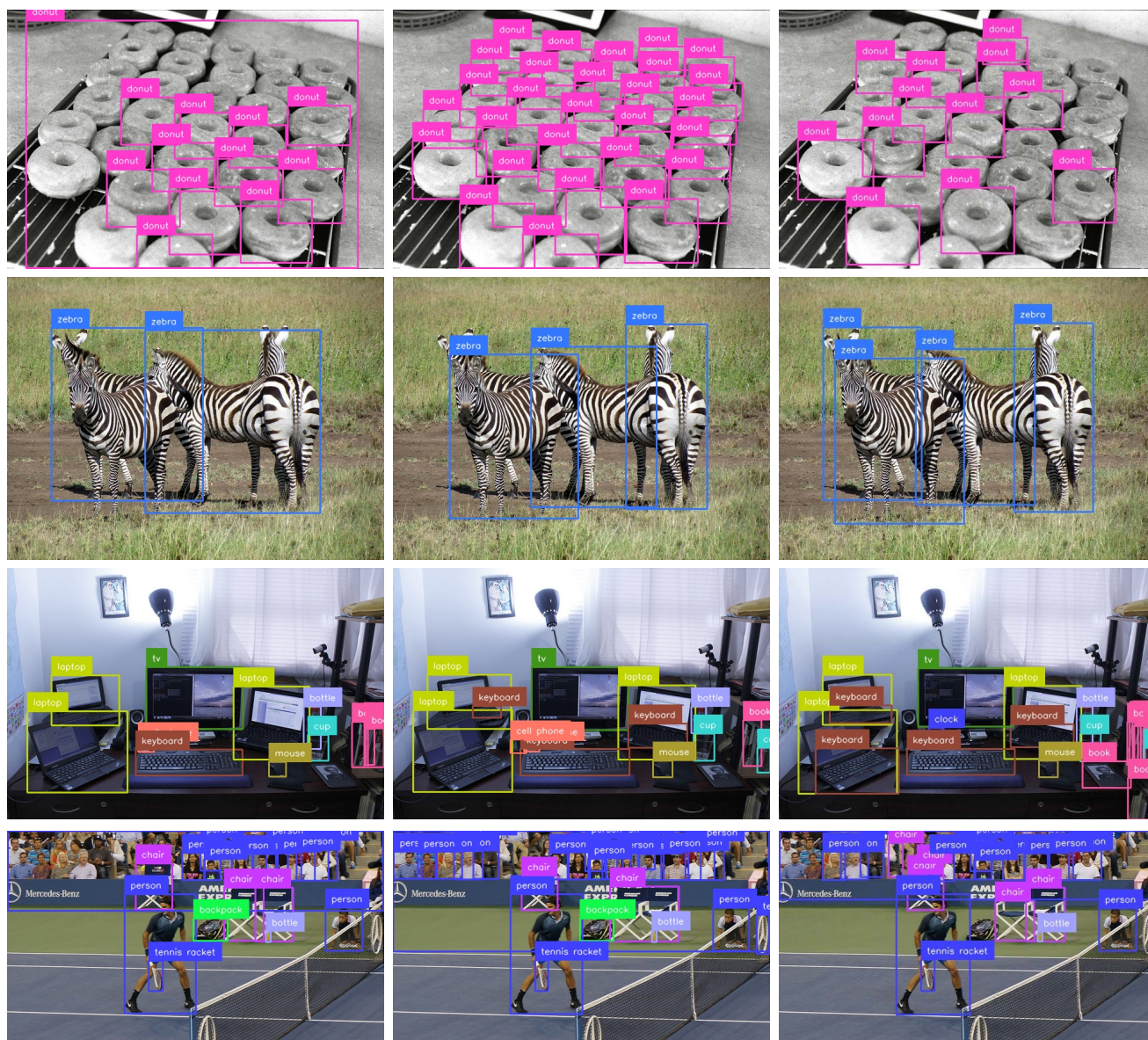


Figure C. blackblack 在数据分布调整前后的真实值和预测值的边界框分布。平均精度（AP）为 38.7 和 47.5 的模型可以参考表格 3。



LLaVA* (AP=38.7)

LMM-Det (AP=47.5)

Ground-Truth

Figure D. blackblack 召回增强的定性结果。