基于智能体工作流的语言与视觉目标的重新对齐

Yuming Chen[1] **Jiangyan Feng**[2] **Haodong Zhang**[2] **Lijun Gong**[2] **Feng Zhu**[2]
Rui Zhao[2] **Qibin Hou**[1*] **Ming-Ming Cheng**[1] **Yibing Song**[*]

[1] 南开大学媒体计算实验室 [2]SenseTime Research
chenyuming@mail.nankai.edu.cn houqb@nankai.edu.cn yibingsong.cv@gmail.com

摘要

基于语言的目标检测（LOD）旨在将视觉目标与语言表达对齐。大量配对数据被用于提升 LOD 模型的泛化能力。在训练过程中，最新的研究利用视觉-语言模型（VLM）为视觉目标自动生成类人化表达，从而帮助扩展训练数据规模。在这一过程中，我们观察到 VLM 的幻觉会导致目标描述不准确（例如目标名称、颜色和形状），从而降低视觉-语言对齐的质量。为了减少 VLM 幻觉，我们提出了一种由大语言模型（LLM）控制的智能体工作流，通过自适应调整图像和文本提示来重新对齐语言与视觉目标。我们将该工作流命名为 Real-LOD，其中包含规划、工具使用和反思步骤。在给定一张包含已检测目标和 VLM 原始语言表达的图像，Real-LOD 能够自动推理其状态，并基于我们的神经符号设计安排动作（即规划）。该动作会自适应调整图像和文本提示，并将其输入到 VLM 中以重新描述目标（即工具使用）。随后，我们使用另一个 LLM 分析这些优化后的表达以获得反馈（即反思）。这些步骤以循环方式执行，以逐步改进语言描述，使其与视觉目标重新对齐。我们构建了一个包含 18 万张图像及其重新对齐的语言表达的数据集，并训练了一个流行的 LOD 模型，在标准基准测试上相较于现有 LOD 方法提升了约 50%。通过自动视觉语言优化，我们的 Real-LOD 工作流展示了在扩展数据规模的同时保持数据质量的潜力，并从数据对齐的角度进一步提高了 LOD 性能。

1 引言

将语言表达与视觉目标对齐的研究一直在不断发展。最初，单个名词被用作类别标签 (Redmon et al., 2016; Ren et al., 2016; Carion et al., 2020) 来连接视觉目标。随后，引入了短语 (Akbari et al., 2019; Li et al., 2022; Gao et al., 2023) 来描述目标。进一步地，指称表达 (Su et al., 2020; Zhang et al., 2022) 和完整描述 (Schulter et al., 2023; Yao et al., 2024) 被发展用于目标检测。尽管语言从粗略标签演变为细粒度表达，但目标检测的本质仍然是使语言数据与视觉目标对齐。由于语言表达为了表示各种人类意图而变得更加多样，这种对齐富有挑战性。对于相同的视觉目标，不同的人通常会以不同的方式描述它，因为他们关注目标属性的不同方面（例如颜色、形状、纹理以及与周围环境的关系）。这种多样性使得视觉-语言（VL）对齐变得复杂，需要收集全面的语言表达集合以用于模型训练。幸运的是，新兴的 VLM (Zhang et al., 2021; Liu et al., 2023a; Ye et al., 2023; Sun et al., 2024; Yuan et al., 2024; You et al., 2024; Zhang et al., 2024) 进来被用于生成类人化表达。自动生成视觉目标的语言表达缓解了训练数据对的收集难度。通过利用更多 VL 数据训练 LOD 模型，研究 (Pi et al., 2024; Dang et al., 2024; Kong et al., 2024) 相应地提升了检测性能，尤其是在语言查询多样化以描述目标对象的情况下。

尽管 VLM 生成的语言表达符合人类偏好，但由于模型幻觉，它们可能无法准确描述目标对象。图 1 展示了两个示例。如图 1(a) 所示，小目标会导致 VLM 生成错误的表达。此外，如图 1(c) 所示，未明确指定目标的通用文本提示会导致 VLM 错误地描述视觉内容。针对小目标的模型幻觉，我们分析认为，VLM 的训练依赖于大规模的图像-标题对数据 (Radford et al., 2021; Schuhmann et al., 2022)，其中标题主要描述全局图像内容而非局部目标。训练数据中对局部目标上下文的忽视，使得 VLM 对小目标产生了幻觉。另一方面，未明确指定目标目标的文本提示（例如“在红色框中”）会导致 VLM 生成错误的细节描述。提示词中

[*]Q. Hou 和 Y. Song 是通讯作者。J. Feng 和 H. Zhang 具有同等贡献。代码可从 https://github.com/FishAndWasabi/RealLOD 获取。

<Prompt>: What is the object in the red box?

<Response>: The object is a blue and red ***hat worn by a young girl***.

(a) VLM for a small object

<Prompt>: What is the object in the red box?

<Response>: The object is ***a flag*** with a design that resembles ***the Union Jack***.

(b) VLM with extended crop

<Prompt>: Describe the person in the red box.

<Response>: The person is a woman with long hair wearing a ***black top***.

(c) VLM with general prompt

<Prompt>: What is the color of person's clothes.

<Response>: The person is wearing ***a black and white striped shirt, dark pants, and black shoes.***

(d) VLM with specified prompt

图 1: 自适应图像和提示词修改能够优化语言表达的示例。对于 (a) 中的一个小目标，VLM 生成了错误的内容（红色标记）。在 (b) 中，我们裁剪了 (a) 的局部区域，并获得了优化后的内容（绿色标记）。另一个示例是，(c) 中的通用提示词导致了错误内容，而 (d) 中更具体的提示词则没有出现该问题。

缺乏目标标识，使得 VLM 对目标细节不够敏感，从而导致错误表达。当这些不准确的语言表达被加入时，目标与语言的对齐变得脆弱，阻碍了 LOD 性能的提升以及 VL 数据规模的扩展。

在这项工作中，我们提出了一种自动将语言表达与视觉对象重新对齐的方法，以从对齐的角度提高 VL 数据质量。我们的重新对齐通过由 LLM 驱动的智能体（即 Real-Agent）控制的工作流来实现。[1] 图 2 展示了构成循环的三个步骤，即规划 (planning)、工具使用 (tool use) 和反思 (reflection)。给定一张包含检测到的目标的输入图像，我们首先将该图像转换为标题，并与目标位置、类别以及由最初使用的 VLM 生成的原始语言表达一起输入到 Real-LOD 中。然后，我们的智能体会自动推理当前状态并安排后续动作。状态/动作体现了我们在工作流中的神经符号设计，其中我们预定义了五种状态，以指示语言如何与视觉对象对齐。每种状态之后均由一个预设的动作。在规划步骤之后，我们的智能体采取行动，为工具模型（即 VLM/LLM）构建自适应的 VL 提示。自定义的提示词使工具模型能够收集更多视觉信息或优化当前表达。在工具使用步骤之后，优化后的表达会被传递给基于 LLM 的反思模块以获取反馈。反馈随后被提供给智能体，以便在下一循环中进行规划。图 5 展示了一个示例，其中原始表达通过循环逐步优化，以对齐目标对象。

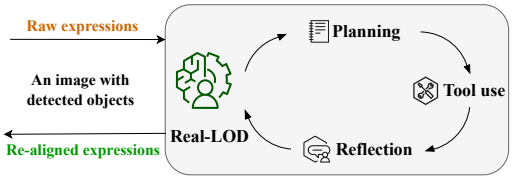

图 2: 我们 Real-LOD 的概览。它以带有检测到的物体和原始表达的图像标题作为输入，并逐步重新对齐表达以更好地匹配物体。通过使用对齐更好的训练数据对，我们提升了 LOD 的性能。

我们的 Real-LOD 通过重新对齐来优化语言-目标数据对以用于 LOD 模型训练。我们的 Real-Model 采用了一种流行的模型结构，使用 Swin-B 主干网络 (Liu et al., 2021)。训练该模型时，我们使用构建的数据集 Real-Data，其中包含 18 万张图像，涵盖 140 万条语言-目标配对数据。在标准基准测试 (Mao et al., 2016; Schulter et al., 2023; Xie et al., 2023) 中，我们比现有方法提升了约 50%。这表明数据的数量和质量对 LOD 训练至关重要。除了扩展图像数据的规模，我们的 Real-LOD 还能保持数据对的质量。这一潜力引领了一种新趋势，即从数据对齐的角度来看，扩展高质量配对数据可以进一步提升 LOD 性能。

2 相关工作

基于语言的目标检测。 LOD 要求模型根据不同表达方式定位相关实例。得益于视觉-语言检测器的发展，LOD 任务的准确率得到了快速提升。MDETR (Kamath et al., 2021) 首次

[1] 为了表达清晰，我们称 Real-LOD 为智能体工作流，称 Real-Agent 为 LLM 驱动的智能体，称 Real-Data 为我们构建的数据集，称 Real-Model 为我们训练的 LOD 模型。

提出了一种端到端调制检测器，通过给定的查询检测目标。GLIP (Li et al., 2022) 提出了一种语言-图像预训练模型，用于理解目标级别的、类别感知的视觉表示。GDINO (Liu et al., 2024b) 引入了一种开集目标检测器，采用高效融合模块，允许使用文本输入（如类别名称或指称表达）进行目标检测。FIBER (Dou et al., 2022) 设计了一种新的视觉-语言模型架构，能够处理多种任务，如视觉问答（VQA）、图像字幕、目标检测等。APE (Shen et al., 2024) 提出了一种通用视觉感知模型，可在大规模数据上一次性对齐视觉和语言表示，从而能够在无需特定任务微调的情况下执行不同的语言-视觉任务。OWL-V2 (Minderer et al., 2023) 提出了一种不含任何融合模块的架构，他们使用 10 亿语言-目标对数据直接对齐图像和文本特征。上述方法均利用语言-目标配对数据训练其检测器，包括 COCO (Lin et al., 2014)、Objects365 (Shao et al., 2019)、OpenImage (Kuznetsova et al., 2020)、SBU (Ordonez et al., 2011)、GoldG (Li et al., 2022)、CC (Sharma et al., 2018; Changpinyo et al., 2021; Xu et al., 2023)、LVIS (Gupta et al., 2019)、Flickr30K (Plummer et al., 2017)、GRIT (Gupta et al., 2022) 和 V3Det(Wang et al., 2023a)。

智能体工作流。由 LLM 驱动的智能体能够遵循用户指令解决各种复杂任务 (Askell et al., 2021; Liu et al., 2025; Significant Gravitas, 2023; Yohei Nakajima, 2023; Reworkd, 2023)。得益于 LLM 强大的理解和推理能力 (Wei et al., 2022; Wang et al., 2023b)，这些智能体能够制定计划以实现特定目标，掌握工具以执行任务 (Yao et al., 2023; Liu et al., 2023b; Tang et al., 2024; Yang et al., 2023; Guo et al., 2024; Shen et al., 2023; Cai et al., 2024)，生成反馈以优化输出 (Madaan et al., 2023; Shinn et al., 2023; Yu et al., 2024; An et al., 2023; Gou et al., 2024)，甚至能够与其他智能体协作 (Chen et al., 2025; Xu et al., 2024; Holt et al., 2024)。HuggingGPT (Shen et al., 2023) 是一个强大的智能体，它利用 LLM 连接各种 AI 模型，以解决不同的任务。该智能体旨在理解和拆解给定的 AI 任务，并自动规划和选择合适的 AI 模型来执行各个子任务。类似地，LLaVA-Plus (Liu et al., 2023b) 维护了一个技能库，其中包含大量视觉-语言工具，以满足现实世界中的多模态任务需求。其他示例包括 Gorilla (Patil et al., 2023)、GPT4tools (Yang et al., 2023) 和 ToolAlpaca (Guo et al., 2024)，这些都是经过微调的 LLM，具备调用可用 API 的能力。此外，最新研究还揭示了通过无训练方法提升智能体性能的方法。其中的核心思想之一是反思，即智能体对自身提供反馈，并利用这些反馈优化输出。Self-Refine (Madaan et al., 2023) 和 Reflexion (Shinn et al., 2023) 是通过语言反馈强化智能体的典型示例，而 CRITIC (Gou et al., 2024) 在反思过程中引入了外部工具，并使用相对固定的人为预设执行逻辑。与以往研究不同，Real-LOD 首次设计了一个涵盖上述三个步骤的完整智能体工作流，以提升 LOD 任务中 VL 数据的对齐质量。

3 重新对齐语言与视觉目标

在本节中，我们首先回顾了 LOD 框架，展示了如何利用 VL 配对输入预测目标对象，并在第 3.1 节中总结以往用于生成语言表达的方法。随后，我们在第 3.2 节中详细说明我们 Real-LOD 的关键步骤（即规划、工具使用、反思）。图 5 提供了一个示例，直观展示了如何通过我们的重新对齐机制优化语言表达。在第 3.3 节中，我们进一步分析了这些优化后的表达，它们构成了训练数据对，以提升 LOD 性能。

3.1 LOD 框架和语言表达生成

语言驱动的目标检测（LOD）框架通常由两个编码器、若干交互模块和一个解码器组成。图 3 展示了其概览。LOD 的输入包括一张图像和由单词、短语或句子组成的语言表达。LOD 使用图像和文本编码器分别获取它们的嵌入。随后，表达与视觉对象进行交互，以构建一个联合的跨模态特征空间。这些交互通常通过交叉注意力作进行。之后，LOD 引入解码器模块，根据每个表达来定位相应的目标。训练损失（例如 L1 损失、GIoU 损失 (Rezatofighi et al., 2019)、对比损失 (Li et al., 2022)）通常源自基于 DETR 的方法 (Carion et al., 2020; Kamath et al., 2021; Zhang et al., 2023)。

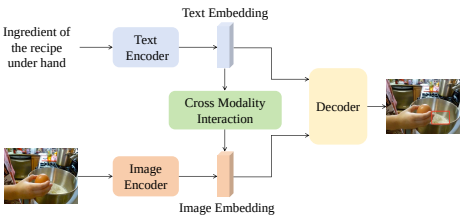

图 3: 一个通用 LOD 框架的概览。配对的 VL 数据被独立编码，然后交互以解码结果。

LOD 框架建立了语言与目标之间的关联。训练数据包含图像、目标边界框（bbxs）和语言表达。以往的数据集 (Mao et al., 2016; Plummer

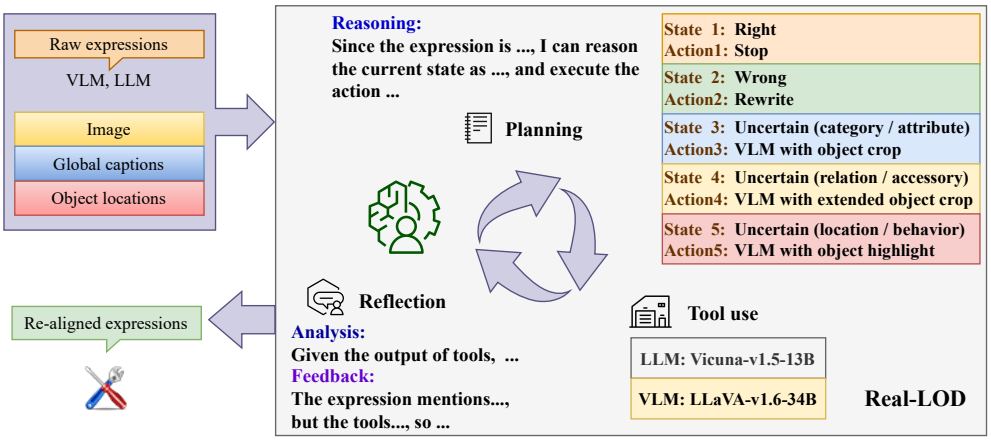

图 4: 我们的智能体工作流概览。输入包括带有标题的图像、检测到的物体和原始表达。我们的 Real-Agent 推理状态并安排动作（即规划）。在执行过程中，Real-Agent 使用 VLM 和 LLM 重新感知视觉内容并优化表达（即工具使用）。然后，LLM 对输出结果进行分析（即反思）。反馈将提供给 Real-Agent 以进行下一轮规划。

et al., 2017; Krishna et al., 2017) 倾向于收集人类参与者的表达，从而构建了有限数量的配对数据，成为检测性能的瓶颈。最近的研究 (Dang et al., 2024; Pi et al., 2024) 利用 VLM 为视觉对象生成类人表达。训练数据量大幅扩展，所学习的 LOD 模型能够捕获多样化的目标描述。遵循这一思路，我们使用 VLM 模型 LLaVA-v1.6-34B (Liu et al., 2024a) 为 18.8 万张图像、33.65 万个目标生成 67.3 万条语言表达。同时，我们利用 LLM Vicuna-v1.5-13B (Zheng et al., 2023)，通过生成同义表达将语言表达数量从 67.3 万扩展至 134.61 万。关于原始表达生成的具体细节详见附录第 ?? 节。获得语言-目标配对数据后，我们使用 SigLIP (Zhai et al., 2023) 计算 VL 匹配得分。对于得分低于 0.5 的配对数据，我们采用 Real-LOD 重新对齐原始语言表达，如第 3.2 节所示。这是因为我们利用 SigLIP 从工作流中过滤掉了约 75% 的训练数据，仅保留近 25% 进行后续处理。

3.2 用于语言表达重新对齐的智能体工作流

生成的语言表达可能与视觉对象不匹配。如第 1 节所示，对局部上下文的忽略或不明确的文本提示词都可能导致模型幻觉。为了解决这一问题，我们设计了一个循环工作流，使 VLM 能够自适应地聚焦于局部区域，并根据目标对象具体化文本提示词。基于 LLM 在纯语言模式下比在 VL 模式下推理更准确的发现，我们选择了经过微调的 ChatGLM-6B (Zeng et al., 2023)，只输入文本作为我们的 Real-Agent 来控制该工作流。图 4 展示了整体概览，包括规划、工具使用和反思步骤，以逐步优化原始语言表达。为了实现 VL 重新对齐，我们在规划和工具使用步骤中进行了神经符号设计，预定义了 5 种状态和动作，如下所示。

规划 (Planning)。在此步骤中，我们预定义了五种状态，指示表达如何从 VLM 的角度与目标对象对齐。每种状态对应一个需要执行的动作。制定这些状态和动作的动机源于我们在第 3.3 节的数据分析，其中我们观察了 VL 失配在实践中是如何发生的。给定包含表达、图像描述、目标类别以及上一个周期反思模块输出的纯文本输入（如果是第一个周期，则为空），我们的 LLM 驱动的 Real-Agent 将推理当前状态并相应地安排动作。五种预定义的状态和动作如下：

状态 1：正确。动作 1：停止。Real-Agent 确信当前的语言表达与目标对象匹配。Real-Agent 将终止工作流并输出当前的表达。

状态 2：错误。动作 2：重写。Real-Agent 确信当前的表达与目标对象不匹配。因此，Real-Agent 将使用 LLM 重新生成表达。用于重写的上下文提示词将按照附录表 ?? 中的模板生成。

状态 3：不确定（类别/属性）。动作 3：使用 VLM 裁剪目标。Real-Agent 无法确定当前表达是否与目标对象匹配，这种不确定性涉及对象类别或属性。因此，Real-Agent 规划裁剪出目标区域并使用 VLM 进行进一步重新感知。VLM 生成的描述将保留在下一步的文本提示词中。

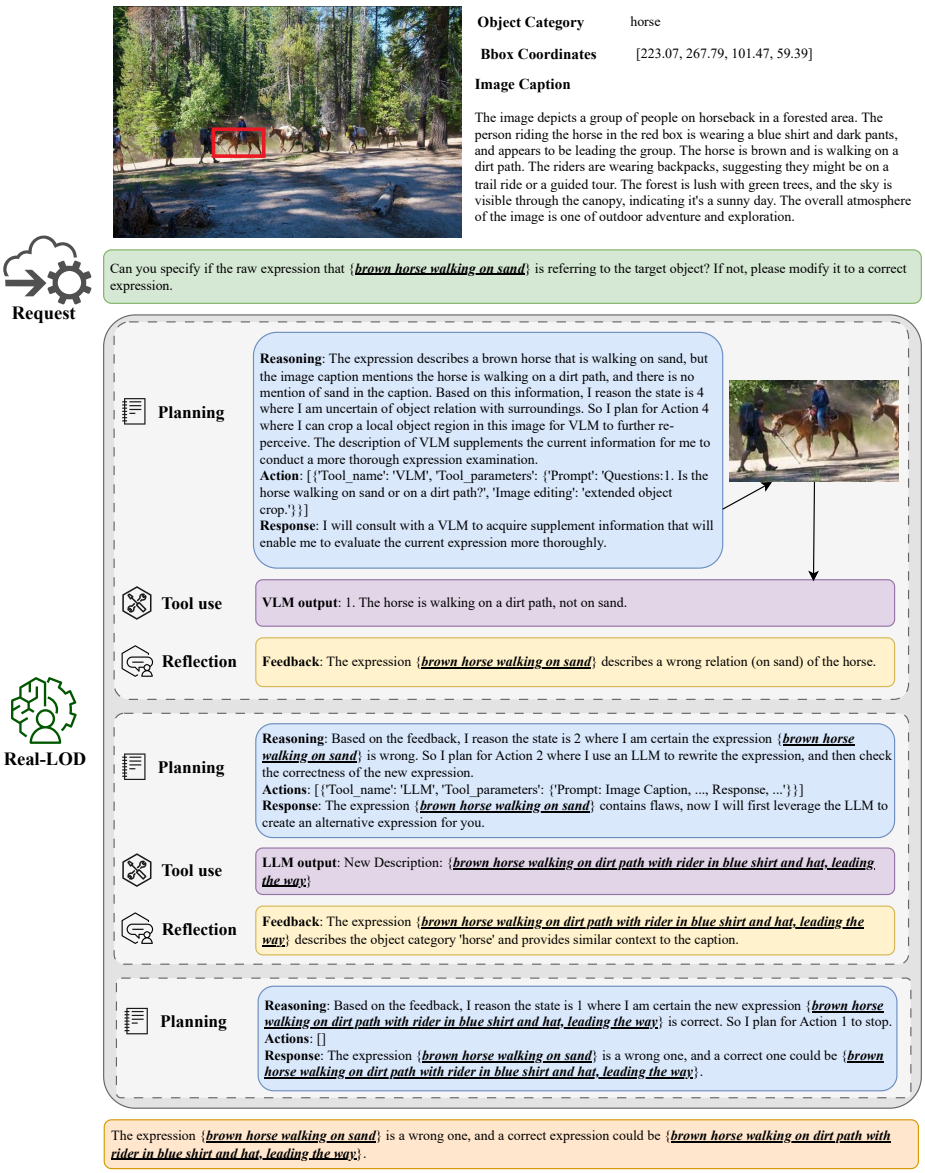

图 5: 一个 Real-LOD 如何将原始表达与给定图像重新对齐的示例。基于输入图像、标题和检测到的对象，Real-LOD 在循环工作流中执行规划、工具使用和反思，以进行状态推理、动作执行和结果反馈。图像和提示词会自适应调整，以辅助工具模型补充定制的对象描述，从而促进表达的重新对齐。

状态 4：不确定（关系/附属物）。动作 4：使用 VLM 裁剪扩展目标。类似于状态 3，Real-Agent 不能确定目标的关系（与周围环境）或附属物。它规划裁剪一个覆盖目标对象的较大区域，并使用 VLM 进行重新感知。VLM 生成的描述将保留在下一步的文本提示词中。

状态 5：不确定（位置/行为）。动作 5：使用 VLM 突出显示对象。类似于状态 3，Real-Agent 不能确定目标在图像中的位置或其行为。它规划使用红色矩形 (Shtedritski et al., 2023) 突出显示对象区域，并使用 VLM 进行重新感知。VLM 生成的描述将保留在文本提示词中。

在执行动作时，我们仅在动作 2 中优化语言表达，而在动作 3、4、5 中，我们使用 VLM 补充描述作为上下文提示词。这些提示词将在下一个周期中用于状态推理和动作执行。

工具使用（Tool use）。在规划阶段，Real-Agent 已经规划了在执行动作时使用多个工具（即 VLM 和 LLM）。我们预先准备了一个工具集，包括一个 LLaVA-v1.6-34B 模型（作

为 VLM），以及一个 Vicuna-v1.5-13B 模型（作为 LLM）。基于对当前表达状态的推理，Real-Agent 通过为规划的工具设置“提示词和“图像编辑”参数，自适应地调整视觉内容和文本提示。这样，该工具可以有效地用于从 VLM 中获得期望的响应，来优化表达方式。例如，在执行 VLM 进行视觉内容重新感知时，Real-Agent 会根据目标对象的边界框（bbxs）裁剪或突出显示图像。此外，Real-Agent 设计的定制文本提示会更具体地与目标对象相关。通过这种方式，Real-LOD 能有效减少模型幻觉，通过重新对齐表达方式来提升语言与对象的连接性。VLM 所使用的视觉和语言提示在附录表 **??** 中展示。

反思（Reflection）。在使用工具后，Real-Agent 已完成动作执行。我们使用 LLM（即 Vicuna-v1.5-13B）作为反思模块来结合图像标题分析结果。反思模块会验证当前表达方式是否与目标对象匹配。对于状态 3-5，Real-Agent 处于不确定状态，反思模块将提高其判断表达是否正确的置信度。对于状态 2（Real-Agent 计划重写表达方式），反思模块将检查新表达方式的正确性。反思模块的分析结果将被整理成反馈信息，来帮助 Real-Agent 在下一个周期进行规划。

3.3 语言和视觉目标的数据分析

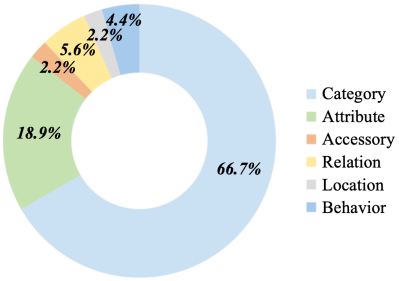

图 6: 不匹配表达的 6 种情形的百分比，其中类别和属性占多数。

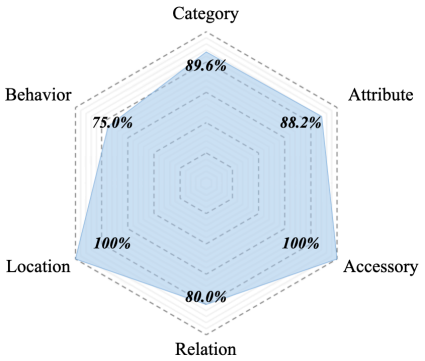

图 7: 通过 Real-LOD 在 6 种情形重新对齐表达的成功率。

Real-Agent 的训练数据。我们准备了文本形式的训练数据，以用于从 ChatGLM-6B 微调得到 Real-Agent。首先，我们从 Objects365 数据集 (Shao et al., 2019) 中随机采集包含已检测目标的图像。然后，与 Sec. 3.1 类似，我们使用 VLM 生成原始表达，并收集那些被 SigLIP 过滤掉的数据对，即匹配分数低于 0.5 的数据对。我们总共准备了 1.5 万条输入数据用于训练 Real-Agent。每条输入数据包含目标类别、原始表达以及来自基于 LLM 的反思模块（定义在 Sec. 3.2）的推理，以检查原始表达是否与目标对象匹配。然后，我们手动为每条输入数据设置状态，并通过 LLM（即 Vicuna-v1.5-13B）使用文本提示收集响应（包括“推理”和“动作”），其中提示包含若干手工设计的上下文示例（见表 **??**），这借鉴了 LLaVA (Liu et al., 2023a) 的思想。最后，我们进行人工检查，以确保微调数据中没有错误。训练过程以参数高效的形式即 LoRA (Hu et al., 2021) 进行，不会影响 ChatGLM-6B 的推理能力。

语言与视觉对象的分析。我们的 Real-LOD 修正了被 SigLIP 滤除的原始表达。由于我们预先设计了用于表达修正的动作，因此我们能够分析这些表达如何与目标对象错误匹配。我们随机选取了三百条被滤除的表达，并对每条进行人工检查以详细观察它们。总体而言，基于观察到的表达，我们将错误匹配的原因总结为六种情形：1) **类别（Category）**：表达描述的是另一个对象，而非目标对象；2) **属性（Attribute）**：表达提供了错误的属性，如颜色、形状和纹理等；3) **附属物（Accessory）**：目标对象的附属物描述错误；4) **位置（Location）**：目标对象在图像中的相对位置错误；5) **关系（Relation）**：目标对象与周围环境的关系错误；6) **行为（Behavior）**：目标对象或人的行为错误。图 8 展示了每种情形的典型表达错误的修正方式。这六种情形激发了我们在第 3.2 节中的神经符号动作设计，其中利用了图像编辑操作（即目标裁剪‘`object crop`’、扩展目标裁剪‘`extended object crop`’和目标突出显示‘`object highlight`’）来增强 VLM 对目标相关内容的感知能力。图 6 展示了这六类问题在我们观察到的表达中的占比，其中大多数错误匹配集中在类别和属性方面。经过修正后，我们计算了每类问题的表达的修正成功率，如图 7 所示，其中类别和属性方面的大多数错误表达都可以被有效修正。

Real-LOD 的分析实验。除了总结原始表达错误匹配的 6 种情形外，我们还分析了 Real-LOD 在重新对齐中的有效性。尽管我们为 VLM 设计了相应的动作以实现重新感知，但准确的状态推理与动作规划将决定修正的质量。对于图 4 中列出的输入，Real-Agent 需要准确推理其所属状态，并据此执行相应的动作。为了分析 Real-Agent 的推理能力，我们抽取了 1.1 万个样本，并引入了一种比较方案，即将规划步骤替换为随机选择一个状态/动作以进一步修正表达。在这两个工作流中，我们都使用反思模块来判断最终表达的修正是否成功，并将最大轮次数设置为 3。在使用 Real-LOD 和随机选择方案进行表达修正后，我们得到成功率[2]分别为 74.7% 和 35.6%。这一对比表明，Real-Agent 的准确推理显著提高了表达的正确性。此外，我们使用 SigLIP 比较了图像与修正后表达之间的匹配分数。我们的 Real-Agent 将平均匹配分数提升了 66.27%（即从 0.0673 提升至 0.1119），而随机选择方案提升了 32.69%（即从 0.0673 提升至 0.0893）。这表明 Real-Agent 相比随机选择方案提高了更多的 SigLIP 匹配分数（即 66.27% 和 32.69%）。从重新对齐的成功率和 SigLIP 分数提升的对比来看，我们的 Real-Agent 在推理输入状态、正确规划动作以及成功修正原始表达方面表现出了显著的效果。我们还在第 **??** 节中提供了更多分析实验。

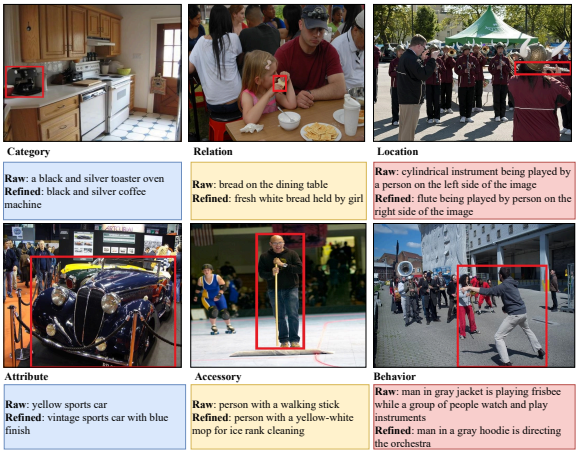

图 8: 示例从 6 种情形总结了未对齐的原始表达，以及我们重新对齐后的表达。

4 基于语言的目标检测的实验

Real-Model 是一种流行的语言引导目标检测（LOD）模型结构，如第 3.1 节所示。我们使用重新对齐后的数据 Real-Data 来训练该模型。训练细节见附录第 **??** 节。在本节中，我们将重点评估我们模型在 LOD 场景下的表现。我们将展示所采用的基准数据集、消融实验、现有方法的评估，以及计算开销分析。

标准基准。我们用于评估的基准是 OmniLabel (Schulter et al., 2023)、DOD (Xie et al., 2023)、RefCOCO/g/+（即 RefCOCO、RefCOCOg、RefCOCO+） (Yu et al., 2016; Mao et al., 2016) 和 OVDEval (Yao et al., 2024)。OmniLabel 收集自三个目标检测数据集，即 Objects365 (Shao et al., 2019)、OpenImage (Kuznetsova et al., 2020) 和 COCO (Lin et al., 2014)，并划分为这三个子集进行评估。共包含 1.22 万张图像、2.04 万个目标框（bbxs）、1.58 万条表达。评估指标包括 AP、AP-des-pos 和 AP-des-S/M/L，分别衡量整体、仅正例和不同长度语言表达的平均精度。DOD 包含 1 千张图像、1.8 万个目标框和 422 条语言描述，使用 “Presence”（存在）和 “Absence”（不存在）分别评估模型在正负查询下的检测表现。RefCOCO/g/+ 来自 COCO 数据集，共有 9.9 千张图像、2.29 万个目标框和 4.65 万条描述。OVDEval 的具体细节见第 **??** 节。所有基准实验均遵循标准协议，以保证比较的公平性。

Real-Data。我们的 Real-LOD 通过语言表达重新对齐自然构建出一个数据集。我们从 Objects365、OpenImage 和 LVIS 数据集中随机选取覆盖所有类别的图像，总共包含 18.8 万张图像和 134.61 万对目标-查询对。其中，47.38 万对样本被 SigLIP 筛选出，这些样本中 30.71 万对经由 Real-LOD 重新对齐。最终用于训练 Real-Model 的样本为 117.94 万对。我们将该数据集命名为 Real-Data。

4.1 和最先进 LOD 方法的对比

我们在标准基准上评估了我们 Real-Model 与现有 LOD 方法的表现，标准基准包括 OmniLabel、DOD 和 RefCOCO/g/+，如表 1-3 所示。在每个表中，我们列出了 LOD 方法所使用的视觉主干网络、训练图像的来源（即 “数据源”）以及用于训练的图像数量（即 “图像数”）。我们使用 VG、OI、O365、RefC/g/+ 和 CC 分别表示 Visual Genome (Krishna et al., 2017)、

[2]设 N 和 N_s 分别表示总表达数和正确修正的表达数，则成功率可表示为 $\frac{N_s}{N}$。

表 1: 在 OmniLabel 基准上最先进方法的比较。

子集	LOD 方法	主干网络	数据源	图像数	AP-des	AP-des-pos	AP-des-S	AP-des-M	AP-des-L
COCO	MDETR (Kamath et al., 2021)	ENB3	COCO, VG, Flickr30K	0.3M	13.2	31.6	15.4	13.5	12.4
	GLIP (Li et al., 2022)	Swin-L	O365, OI, RefC/g/+, etc	17.5M	13.9	36.8	28.9	12.9	11.5
	mm-GDINO (Zhao et al., 2024)	Swin-B	GoldG, O365, COCO, etc	12M	15.2	47.0	29.3	14.9	15.1
	FIBER (Dou et al., 2022)	Swin-B	COCO, CC3M, SBU, etc	4M	14.3	38.8	31.3	12.7	16.1
	Real-Model	Swin-B	Real-Data	0.18M	**26.2**	**59.7**	**39.4**	**25.4**	**24.3**
O365	MDETR (Kamath et al., 2021)	ENB3	COCO, VG, Flickr30K	0.3M	3.2	5.9	3.0	3.2	2.7
	GLIP (Li et al., 2022)	Swin-L	O365, OI, RefC/g/+, etc	17.5M	24.0	35.2	44.5	20.5	11.8
	mm-GDINO (Zhao et al., 2024)	Swin-B	GoldG, O365, COCO, etc	12M	19.6	31.0	32.3	17.8	12.4
	FIBER (Dou et al., 2022)	Swin-B	COCO, CC3M, SBU, etc	4M	25.9	38.2	44.7	22.5	14.1
	Real-Model	Swin-B	Real-Data	0.18M	**36.0**	**52.1**	**55.7**	**32.3**	**23.7**
OI	MDETR (Kamath et al., 2021)	ENB3	COCO, VG, Flickr30K	0.3M	6.1	10.6	9.6	5.7	4.1
	GLIP (Li et al., 2022)	Swin-L	O365, OI, RefC/g/+, etc	17.5M	20.1	31.2	33.3	18.7	10.3
	mm-GDINO (Zhao et al., 2024)	Swin-B	GoldG, O365, COCO, etc	12M	23.2	34.5	32.3	23.8	16.9
	FIBER (Dou et al., 2022)	Swin-B	COCO, CC3M, SBU, etc	4M	20.1	30.9	34.1	18.5	10.5
	Real-Model	Swin-B	Real-Data	0.18M	**40.5**	**51.4**	**54.9**	**37.8**	**30.6**
ALL	MDETR (Kamath et al., 2021)	ENB3	COCO, VG, Flickr30K	0.3M	4.7	9.1	6.4	4.6	4.0
	GLIP (Li et al., 2022)	Swin-L	O365, OI, RefC/g/+, etc	17.5M	21.2	33.2	37.7	18.9	10.8
	mm-GDINO (Zhao et al., 2024)	Swin-B	GoldG, O365, COCO, etc	12M	20.8	33.1	31.9	19.8	14.1
	FIBER (Dou et al., 2022)	Swin-B	COCO, CC3M, SBU, etc	4M	22.3	34.8	38.6	19.5	12.4
	Real-Model	Swin-B	Real-Data	0.18M	**36.5**	**52.1**	**54.4**	**33.2**	**25.5**

表 2: DOD 基准上的评估结果。

LOD 方法	主干网络	数据源	图像数	Full	Presence	Absence
OWL-V2 (Minderer et al., 2023)	ViT-L	WebLI	10B	9.6	10.7	6.4
UNINEXT (Yan et al., 2023)	ViT-H	O365, RefC/g/+	0.7M	20.0	20.6	18.1
GDINO (Liu et al., 2024b)	Swin-B	CC4M, O365, RefC/g/+, etc	5.8M	20.1	20.7	22.5
mm-GDINO (Zhao et al., 2024)	Swin-B	GoldG, O365, COCO, etc	12M	24.2	23.9	25.9
OFA-DOD (Xie et al., 2023)	RN101	CC12M, SBU, VG, etc	16M	21.6	23.7	15.4
APE-B (Shen et al., 2024)	ViT-L	LVIS, O365, RefC/g/+, etc	2.6M	30.0	29.9	30.3
Real-Model	Swin-B	Real-Data	0.18M	**34.1**	**34.4**	**33.2**

OpenImage (Kuznetsova et al., 2020)、Objects365 (Shao et al., 2019)、RefCOCO/g/+ (Yu et al., 2016) 和 Conceptual Captions (Sharma et al., 2018; Changpinyo et al., 2021; Xu et al., 2023)。此外，各方法训练图像的详细来源可见表 **??**。

在 OmniLabel 基准上，我们的 Real-Model 在所有测试集上均显著优于现有 LOD 方法。特别是在 OI 数据集上，在 AP-des 指标下，Real-Model 显著超过了第二名的 mm-GDINO（指标分别为 40.5% 和 23.2%）；同时，在 AP-des-pos 指标下，Real-Model 也显著超过了同样第二名的 mm-GDINO（指标分别为 51.4% 和 34.5%）。Real-Model 的优异性能得益于 Real-Data 所提供的高质量语言-目标配对数据。另一方面，我们观察到 GLIP 使用的训练数据规模大于 Real-Model，但准确率仅为我们方法的约 50%。这表明，要获得卓越的结果，数据的质量与数量同样重要。

表 2-3 分别展示了在 DOD 和 RefCOCO/g/+ 基准上的评估结果。结果与 OmniLabel 基准中的表现类似。只使用少量训练数据，我们的 Real-Model 即在多项指标下均得了优异的结果，超过了现有的 LOD 方法。这一性能提升得益于我们 Real-Data 中数据对的丰富多样的语言表达，它提升了语言与目标对齐的泛化能力。因此，我们的 Real-Data 数据集在图像与目标不变的前提下，通过多样化的语言描述，推动了 Real-Model 的性能达到最先进水平。此外，关于在 OVDEval 上的评估结果以及我们方法在其他 LOD 模型上的应用，详见第 **??** 节。

4.2 消融实验

我们使用三种训练数据对配置（即 A、B 和 C 形式）来训练我们的 Real-Model，并在 OmniLabel 基准上评估相应的 LOD 性能。我们从 O365 和 OI 数据集中随机选取了 9.4 万张涵盖所有类别的图像，这是 Real-Data 的一个子集。这些图像与其对应的目标对象和原始表达构成了我们的原始训练数据对，总数为 93.3 万（即 A 形式）。随后，我们利用 SigLIP 滤除表达与目标对象不匹配的数据对，剩余数据对有 69.5 万（即 B 形式）。进一步地，我们使用 Real-LOD 重新对齐这些错误配对数据对，并将其补充回 B 形式中，从而将数据对数量增加至 86.3 万（即 C 形式）。我们分别使用 A、B 和 C 三种形式的数据对来训练 Real-Model 并评估相应的性能，从而分析 Real-LOD 在数据对齐角度上对 LOD 性能提升的贡献。

表 4 展示了在三种数据配置下（即 A、B 和 C 形式）的 LOD 结果。结果表明，在 COCO 测试集上，使用全部训练数据对（A 形式）时，Real-Model 达到了 21.2% 的 AP；在剔除表

表 3: RefCOCO/g/+ 基准上的评估结果。其中，“*”表示模型在训练时使用了 RefCOCO/g/+ 数据集。

LOD 方法	主干网络	数据源	图像数	RefCOCO			RefCOCO+			RefCOCOg	
				val	testA	testB	val	testA	testB	val-u	test-u
MDETR (Kamath et al., 2021)	ENB3	COCO, VG, Flickr30K	0.3M	73.4	-	-	58.8	-	-	57.1	-
APE-A (Shen et al., 2024)	ViT-L	COCO, LVIS, O365, etc	2.0M	34.2	34.8	36.1	33.5	32.3	36.0	38.9	40.5
Real-Model	Swin-B	Real-Data	0.18M	74.0	79.6	66.0	76.4	83.1	68.5	80.8	81.2
GLIP* (Li et al., 2022)	Swin-L	O365, OI, RefC/g/+, etc	17.5M	53.1	59.4	46.8	54.0	59.4	47.0	60.7	60.4
GDINO* (Liu et al., 2024b)	Swin-B	CC4M, O365, RefC/g/+, etc	5.8M	-	-	-	73.6	82.1	64.1	78.3	78.1
APE-B* (Shen et al., 2024)	ViT-L	LVIS, O365, RefC/g/+, etc	2.6M	84.6	89.2	80.9	76.4	82.4	66.5	80.0	80.1
Real-Model*	Swin-B	RefC/g/+, Real-Data	0.24M	**91.3**	**93.1**	**88.0**	**85.4**	**90.3**	**78.6**	**88.4**	**89.0**

表 4: OmniLabel 基准的消融实验。我们的训练数据对由图像、目标对象和表达组成。我们通过不同地处理（即 SigLIP 过滤器和 Real-LOD）原始表达来调整训练数据对，并评估相应的性能。请注意，我们使用了 Real-Data 的子集。

测试子集	训练数据类型	图像数	AP-des	AP-des-pos	AP-des-S	AP-des-M	AP-des-L
COCO	原始表达（A）	933k	21.2	59.4	31.3	21.1	18.6
	原始表达（过滤后）(B)	695k	22.2	59.4	32.4	21.9	19.4
	原始表达（过滤后）+ Real-LOD (C)	863k	**24.2**	**59.6**	**35.2**	**24.2**	**21.1**
O365	原始表达（A）	933k	27.6	43.1	39.8	25.5	17.9
	原始表达（过滤后）(B)	695k	28.5	43.7	40.9	26.2	18.5
	原始表达（过滤后）+ Real-LOD (C)	863k	**32.4**	**48.5**	**47.5**	**30.0**	**21.3**
OI	原始表达（A）	933k	30.5	43.0	37.2	30.3	23.2
	原始表达（过滤后）(B)	695k	31.4	43.5	38.1	31.2	24.0
	原始表达（过滤后）+ Real-LOD (C)	863k	**33.5**	**44.9**	**42.2**	**32.9**	**24.8**

达与目标不匹配的数据对后（B 形式），Real-Model 提升至 22.2%。这一提升表明，数据质量从根本上有利于 LOD 性能。接着，我们使用 Real-LOD 对被剔除的数据对进行优化，并补充至训练集（C 形式），这使得 Real-Model 进一步提升至 24.2%。这表明 Real-LOD 在高质量的同时扩大数据规模，从而进一步提升 LOD 性能。在其他两个测试集（即 O365 和 OI）上的结果也呈现出相似的现象。在训练 Real-Model 的过程中，数据质量也会对 LOD 性能产生影响，特别是在数据规模扩大的情况下。我们的 Real-LOD 通过重新对齐不匹配的语言-目标对，在提升数据数量的同时保留了数据质量。因此，在使用 C 形式（即重对齐数据）进行训练时，Real-Model 在 OmniLabel 基准上表现最佳。

4.3 计算开销分析

在我们的 Real-LOD 中，我们还采用了两种策略以进一步降低工作流的时间开销：1）我们利用 SigLIP 滤除工作流中 75% 的训练数据，仅保留 25% 进行处理；2）我们将工作流的最大循环次数设置为 4，以权衡时间成本和性能。我们在表 **??** 中详细说明了我们的计算开销。对于每条表达的优化，我们报告了每个步骤的平均调用次数以及每次调用的时间开销。时间开销基于在 48 张 V100 32G GPU 上执行我们的工作流测得。优化一条表达总耗时为 1.579 秒，平均需要 3.08 次循环。我们还在图 **??** 中提供了迭代次数的分布。需要注意的是，为便于分析，此处设置的最大迭代次数为 10。此外，我们的工作流完全离线，不会为 LOD 模型的推理带来额外的计算负担。

5 结论

语言与视觉目标的重新对齐已经从人工描述发展到 VLM 自动生成。数据对的规模正不断扩大以增强 LOD 的连接性能。由于模型幻觉，生成的描述可能与目标不匹配。因此，我们提出了 Real-LOD，通过智能体工作流逐步优化语言表达，在提升数据量的同时保障数据质量。我们利用这些数据训练了一个流行的 LOD 模型，大幅超过现有的 LOD 方法。我们的自动工作流具有重新对齐任意目标的语言描述的扩展潜力。结合开放词表检测器用于定位带有短类别标签的目标，以及 VLM 用于扩展表达，我们的 Real-LOD 将持续生成高质量训练对，以提升 LOD 性能。

致谢

本研究得到了国家自然科学基金（编号：62225604, 62176130）和天津市科技支撑计划项目（编号：23JCZDJC01050）的资助。南开大学超级计算中心在部分计算资源上给予了支持。

伦理声明

我们声明本研究不存在任何潜在的伦理问题。研究过程中未涉及人类受试者、敏感数据，亦未采用可能导致有害后果或偏见的方法。本研究中使用的所有数据均为公开数据，未涉及隐私或安全相关问题。

可复现性声明

透明性和可靠性是我们研究的核心。在本声明中，我们总结了为促进研究可复现性所采取的措施，并在正文及附录中提供了相应内容的引用。

源代码。我们计划在论文接收后公开源代码、模型权重以及数据集。这将使后续研究者能够获取并使用我们的代码以复现实验与结果。详细的安装与运行指南将包含在“README.md”文件中。

实验设置。我们在第 3.1 节和第 3.2 节中提供了 Real-LOD 的基本实现信息。此外，我们在第 4 节和第 **??** 节提供了实验设置与评估配置。Real-Data 的详细信息则见第 3.3 节和第 4 节。Real-Model 的训练与架构细节可在附录的第 **??** 节和第 **??** 节中找到。

我们提供上述资源和引用，旨在确保本研究工作的可复现性，便于其他研究人员验证我们的方法。我们也欢迎任何对进一步的咨询或对方法细节的询问。

参考文献

Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Learning from mistakes makes llm better reasoner, 2023. arXiv preprint arXiv:2310.20689.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment, 2021. arXiv preprint arXiv:2112.00861.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. In *International Conference on Learning Representations*, 2024.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. In *International Conference on Learning Representations*, 2025.

Ronghao Dang, Jiangyan Feng, Haodong Zhang, Chongjian Ge, Lin Song, Lijun Gong, Chengju Liu, Qijun Chen, Feng Zhu, Rui Zhao, et al. Instructdet: Diversifying referring object detection with generalized instructions. In *International Conference on Learning Representations*, 2024.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *Advances in Neural Information Processing Systems*, 2022.

Shangqian Gao, Burak Uzkent, Yilin Shen, Heng Huang, and Hongxia Jin. Learning to jointly share and prune weights for grounding based vision and language models. In *International Conference on Learning Representations*, 2023.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024.

Zhicheng Guo, Sijie Cheng, Hao Wang, Shihao Liang, Yujia Qin, Peng Li, Zhiyuan Liu, Maosong Sun, and Yang Liu. Stabletoolbench: Towards stable large-scale benchmarking on tool learning of large language models. In *International Conference on Learning Representations*, 2024.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. Grit: General robust image task benchmark, 2022. arXiv preprint arXiv:2204.136533.

Samuel Holt, Max Ruiz Luyten, and Mihaela van der Schaar. L2MAC: Large language model automatic computer for extensive code generation. In *International Conference on Learning Representations*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetrmodulated detection for end-to-end multi-modal understanding. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Fanjie Kong, Yanbei Chen, Jiarui Cai, and Davide Modolo. Hyperbolic learning with synthetic captions for open-world detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 2020.

Liunian Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, and Jenq-Neng Hwang. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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 *European Conference on Computer Vision*, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. Llava-plus: Learning to use tools for creating multimodal agents, 2023b. arXiv preprint arXiv:2311.05437.

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. In *European Conference on Computer Vision*, 2024b.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, 2025.

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 *IEEE/CVF International Conference on Computer Vision*, 2021.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, 2023.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, 2011.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis, 2023. arXiv preprint arXiv: 2305.15334.

Renjie Pi, Lewei Yao, Jianhua Han, Xiaodan Liang, Wei Zhang, and Hang Xu. Ins-detclip: Aligning detection model to follow human-language instruction. In *International Conference on Learning Representations*, 2024.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2016.

Reworkd. AgentGPT, 2023. URL https://github.com/reworkd/AgentGPT.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, 2022.

Samuel Schulter, Yumin Suh, Konstantinos M Dafnis, Zhixing Zhang, Shiyu Zhao, Dimitris Metaxas, et al. Omnilabel: A challenging benchmark for language-based object detection. In *IEEE/CVF International Conference on Computer Vision*, 2023.

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 *IEEE/CVF International Conference on Computer Vision*, 2019.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2018.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. In *Advances in Neural Information Processing Systems*, 2023.

Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Significant Gravitas. AutoGPT, 2023. URL https://github.com/Significant-Gravitas/AutoGPT.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases, 2024. arXiv preprint arXiv:2306.05301.

Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *IEEE/CVF International Conference on Computer Vision*, 2023a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating object detection with flexible expressions. In *Advances in Neural Information Processing Systems*, 2023.

Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Peng Xu, Haoran Wang, Chuang Wang, and Xu Liu. Caca agent: Capability collaboration based ai agent, 2024. arXiv preprint arXiv:2403.15137.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching llm to use tools via self-instruction. In *Advances in Neural Information Processing Systems*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Yiyang Yao, Peng Liu, Tiancheng Zhao, Qianqian Zhang, Jiajia Liao, Chunxin Fang, Kyusong Lee, and Qing Wang. How to evaluate the generalization of detection? a benchmark for comprehensive open-vocabulary detection. In *AAAI Conference on Artificial Intelligence*, 2024.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl: Modularization empowers large language models with multimodality, 2023. arXiv preprint arXiv:2304.14178.

Yohei Nakajima. BabyAGI, 2023. URL https://github.com/yoheinakajima/babyagi.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, 2024.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, 2016.

Zishun Yu, Yunzhe Tao, Liyu Chen, Tao Sun, and Hongxia Yang. $\mathcal{B}$-coder: Value-based deep reinforcement learning for program synthesis. In *International Conference on Learning Representations*, 2024.

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130b: An open bilingual pre-trained model. In *International Conference on Learning Representations*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, 2023.

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 *International Conference on Learning Representations*, 2023.

Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: unifying localization and vl understanding. In *Advances in Neural Information Processing Systems*, 2022.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Change Loy Chen, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. In *Advances in Neural Information Processing Systems*, 2024.

Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection, 2024. arXiv preprint arXiv:2401.02361.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.