关于先画辅助线再输出 action 的思考

Yang

2025 年 7 月 9 日

1 Related Work

https://arxiv.org/pdf/2506.09965这篇论文中提及了当下 VLM 完成涉及图像的推理任务时先将图像转化为文字，再使用文字的推理功能完成任务。这样的转化势必会造成信息的损失，因此推理任务的执行效果并不算良好。论文提出了一种直接基于图像的处理方式：先培养模型画辅助线的能力 (利用别的模型生成轨迹，再根据函数确定良好的轨迹，再培养模型生成轨迹的能力，轨迹中包含推理过程和操作，操作就是想要的处理图像的能力，每执行一个操作就会对图像进行裁剪或者缩放)

逻辑链是用别的模型按照传统的老方法生成推理结果和中间操作，利用函数过滤出认为比较好的数据用于培养训练操作。数据包含操作的坐标，模型可以调用内部函数进行坐标处理得到新的图像。因此整体来说是从生成 action 到画辅助线

目的：让模型先画辅助线再有 action

意义：从模仿人类行为的逻辑上来讲，如果一个模型真正具备视觉推理能力，那么他应该仅基于图像和任务信息完成推理任务。而论文中提出的方法更像是“伪视觉推理”，因为模型训练的数据来自于传统的方法加上过滤，本质上在训练模型的决策能力，后续的 RL 和 reasoning rejection 也只是让这个能力具备了更多的反思和优化。图像作为输入让模型解决问题，模型在预训练之后看到图像会生成更高质量的决策步骤，似乎还是没有绕开转化为文本的隐形过程

2 Thought