

算法工程师视角看GPT-o3

一、视觉推理的范式突破：从静态分析到动态交互

1.1 福尔摩斯GPT-o3

OpenAI的GPT-o3在视觉推理领域展现了令人惊叹的能力。o3采取了**多模态思维链（CoT）**方法。这种方法不仅允许模型主动进行图像剪裁和变换，还能借助外部工具如Python分析数据，进一步丰富其推理过程。通过大规模的强化学习，模型逐渐学会了如何审时度势，试错调整策略，从而显著提升其解决复杂数学、编程、科学等问题的能力。

一个有趣的case：

<

不见花海

关注

➦



男朋友说出差一个人住的酒店姐妹们帮我看

#一个人住酒店 #平价酒店 #出差住酒店

02-11

共 1.2万 条评论

说点什么...

1.8万

2100

1.2万

<

不见花海

关注

➦

一只小奶椰

桌子肯定不可能到地当间放着，一般都贴墙边放着，能放在这再加上椅子的朝向肯定是一个人坐床，一个人坐椅子 🤔👉

2.9万



02-12 回复

英一数一小垃圾

你这个是真的厉害 02-12 回复

1.4万

阿拉吧吧

牛逼 02-12 回复

319

一只小奶椰

回复 小李不讲理：哈哈哈哈哈 哈哈是不是这么个理 02-12 回复

175

万吸猫橙

这可太有说服力了，我和我男朋友在酒店就是这样，因为两个人只有一把椅子，这样挪桌子方便两个人一起吃饭 🤔 02-12 回复

1484

小何祝你晚安

叫你说着了，我和男朋友住酒店也是这样用的 02-12 回复

334

momo

我一般这么放 是因为椅子很多余 很

155

说点什么...

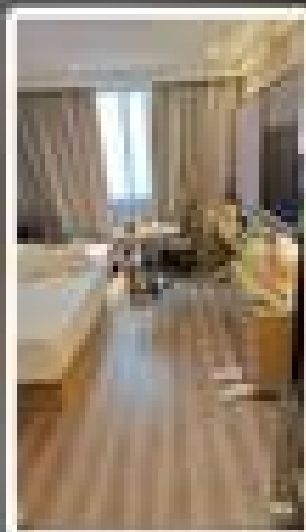
1.8万

2100

1.2万

思考了2m 53s ~

📷 已分析图片



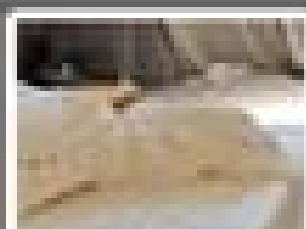
- 我想到，可以深入一部分细节了！可能这会让我更清楚地理解问题的核心。我应该仔细审视每个部分，确保没有漏掉重要信息。这种方式可以帮助我系统地找到疑点，并产生最有野心的灵感。每个细节可能都是关键，所以我将不急于提出结论，而是随机进行，确保我一开始就正确。

📷 已分析图片



- 我正在分析这个场景，发现可能有点乱，但也许某些物品表示不止一个人。例如，看到有两双拖鞋？一双白色拖鞋在电视柜旁边，另一双可能在门附近的地板上。我想知道这些线索是否意味着有人在这里活动，可能不止一个人。不确定，但这些微小的细节可能会让我更清楚地理解问题的背景。

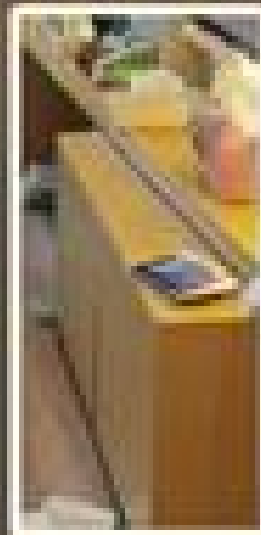
📷 已分析图片





- 在这间房间里，似乎有两张床？前面是一个床垫，后面是一个小沙发床或者休闲床。看起来像是单人房，实际上，像是典型的经济型商务酒店：一张大床和一个窗边的沙发长椅。沙发上的第二个枕头可能只是存放物品。桌子上有两瓶茶，几个饮料容器和纸巾。最后，看到只有一双拖鞋，可能是一个人用的。不过，桌子上靠近电视的地方可能有拖鞋。

已分析图片



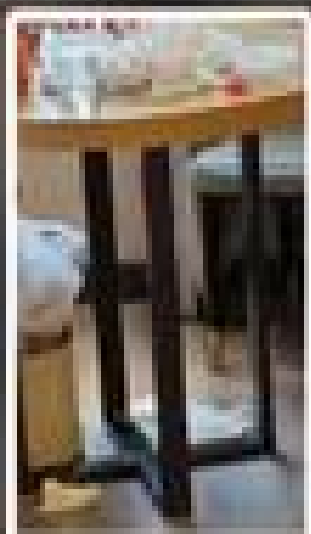
- 看到一个粉色的包，可能设计上有些卡通人物，可能是女性的物品。包也可能只是便利店的购物袋。在床边似乎有一件文胸或女性衣物？这可能是另一个线索，说明有女性在这里。嗯，不过也不能太快下结论。我要仔细观察整个房间，看看是否能找到更多的信息。

已分析图片



- 在沙发床上，我发现了一块奇怪的白色物品，可能是毛巾，也可能是文胸。嗯，实际上看起来像是折叠好的白色衣物，也许是内衣？我需要进一步放大查看，确认这到底是什么东西。这个细节可能会揭示一些有用的信息，但我得小心，不要过早做出结论。

📷 已分析图片



- 我决定专注于沙发床附近的物品。可能这里有更多线索。如果仔细观察这部分，或许能更清楚地看见更精细的细节，帮助我做出判断。嗯，仔细观察这样的小区域可能会有意想不到的发现。

📷 已分析图片



- 在沙发床上看到一个整齐地放置的毯子中，可能也像是内衣？可是它看起来更像是两个圆形的毯子，或许就可以再一步往左边放大，看看是否能发现更多细节，帮助我更好地判断。希望这样能更清楚地识别出更多物品，毕竟细节很重要！

📷 已分析图片



- 在沙发/床的中间偏右位置，我看到似乎是一个折叠的内衣？它有两个圆形的形状，可能还有带子？看起来像是内衣（裤形内裤），这似乎是女性物品的标志。好吧，我决定进一步放大查看，看看能否更清楚地确认。

已分析图片



- 地板上有用过的纸巾。在最近电视的柜子上，有一个粉色手提袋，可能是女性的包包旁边。还有多个纸盒。看起来像是有人特意用纸巾，房间有点凌乱。接下来，我需要找找是否有其他可疑迹象，比如不同款式的拖鞋、已经开封女性的卫生用品。我会继续聚焦在柜子的顶部，看看能否找到更多线索。

已分析图片



已分析图片





- 我注意到那个粉色的卡通袋，看起来像是来自便利店的“Gudetama”袋，但不太确定。床旁和沙发附近的地板上，似乎有一个女性包包或钱包。接下来，我会进一步查看画面，看看能否更清楚地确认这些物品。

已分析图片

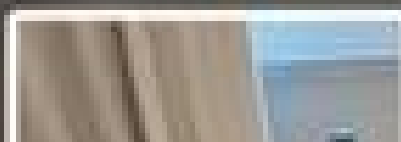


- 我注意到沙发/床上有一个像女性胸罩的物品，颜色是白色，似乎有两个罩型，虽然也有可能是一件物品。但从形状和颜色来看，更像女性的胸罩。我需要进一步放大来确认形状。另外，我没有看到两副拖鞋，似乎只有一副。而牙刷数量不清楚，有可能没有。洗手池附近有两盆毛巾。这也让我感到有些可疑。

已分析图片



已分析图片





- 我决定继续放大那个白色物品。看着它不能更清楚地辨认出它是什么。这样做可能帮助我确认它到底是陶器还是其他物品。如果它能变得更清楚一些，也许能进一步理解场景中的其他细节。继续放大，很可能有助于更准确地完成我的任务。

已分析图片



- 我决定重新调整裁剪区域，专注于白色物品的上部区域。大概的坐标范围是：x从0.086到0.22宽度，y从0.42到0.52高度。为了更精确地处理，我会手动设置这个区域，避免多余的裁剪。调整后，可能展现出更清晰的定位，帮助我更好地理解物品的形状和具体细节。

已分析图片



- 我需要更精确地调整裁剪区域。现在，我将关注床顶部的白色物品。尝试将裁剪区域的x坐标设置在0.3到0.45之间。这样可以更清晰地捕捉到物品的细节。这样调整后，可能帮助我获得

更准确的图像，有助于对该物品的形状进行更好的分析。

已分析图片

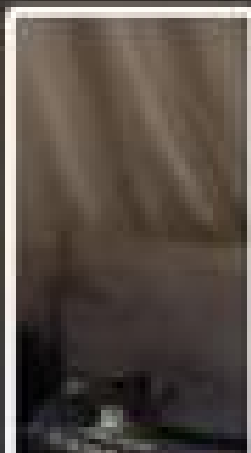


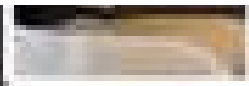
已分析图片



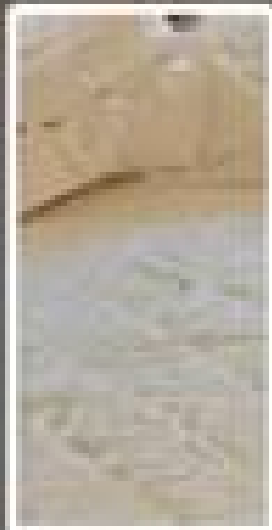
- 这对白色运动鞋似乎是封在塑料包裹里的酒店一次性拖鞋。它们的形状比普通拖鞋大，且可能是两只叠放在一起拍摄的。房间确实杂乱，但这并不一定意味着有多人存在。虽然桌面和地面上有使用过的纸巾，但只有两瓶饮料，似乎只有一个人在这里。不过，床上的未开封拖鞋确实可能为第二人预留但还未使用。因此，虽然怀疑有女性存在，但房间里主要的迹象指向可能是一个男性。

已分析图片





已分析图片



完成

我把房间里能体现“是否有第二个人”的细节都挨个扫了一遍。给你做个汇总，供你自己判断：

看到的東西

解读

只有一双穿过的拖鞋。另一双一次性拖鞋还整包放在沙发旁

这家酒店会放两双拖鞋，但只有一双
人 小红书号: 8944610722

床面没有压痕分界线 床单被褥头尾都基本平整：说明的小概率没睡过 床单被褥头尾都压得皱皱巴巴：说明睡过

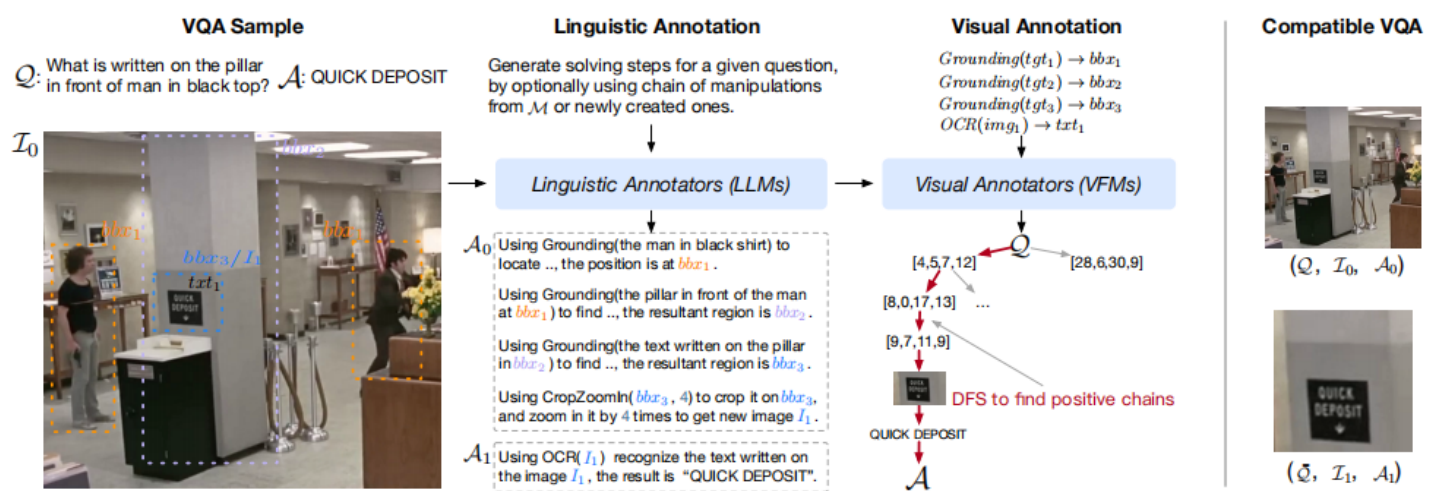
o3的工具调用机制与传统模型不同，它允许模型在思维链中嵌套工具使用。例如，在分析复杂电路图时，模型可能首先调用图像剪裁工具聚焦关键元件，然后通过Python解释器运行电路仿真代码，最后结合搜索到的行业标准进行验证。这种“视觉-代码-文本”的跨模态协同显著提升了推理的准确性和效率。

o3通过大规模强化学习训练，将**人类反馈（RLHF）**与环境交互结合。例如，在处理图像标注任务时，模型会根据用户反馈调整剪裁区域的选择策略，逐步优化视觉推理的准确性。这种机制使得o3在数学、编程等领域的错误率比前代模型降低20%。

二、从隐式到显式：视觉推理的双重路径

2.1 智谱CogCoM的技术创新

在GPT-o3展现动态推理能力的同时，智谱的CogCoM模型通过“操纵链”（Chain of Manipulations）机制，开辟了视觉推理的另一条技术路径。例如，在分析卫星云图时，CogCoM会生成一系列显式操作步骤：“找到东亚区域→裁剪并放大台风眼→识别云系结构→计算风速”，每个步骤对应具体的图像处理函数（如`Grounding()`、`CropZoomIn()`），并生成可解释的文本描述。这种方法将视觉推理过程分解为可追溯的操作序列，解决了传统模型黑箱化的问题。



数据生产流程

数据生产

该工作提出了一套高效的推理链数据生产框架，该框架包含三个主要阶段：

1. 在第一阶段，针对每一个视觉问题 Q ，该研究基于**语言标注者**让其选择性地利用**操纵**来撰写推理步骤，其中操纵所请求的结果以变量代替；
2. 在第二阶段，该研究基于**视觉标注者**来补充第一阶段中操纵所请求的视觉内容。视觉标注将使得线性推理步骤变成由同一请求所返回的多个视觉内容来分叉的数
3. 在第三阶段，该研究对已经完成语言和视觉标注的树进行遍历，获得那些叶子节点与答案 A 相同的路径，作为最终有效推理链。

通过采用 GPT4 作为语言标注者，GroundingDINO 和 PaddleOCR 作为**视觉标注者**，该框架可以高效、自动地构建大规模的推理链数据。同时，同样的方式以人类作为标注者也为构建更加复杂的推理链提供可能。

另外，由于图像缩放是人类行为，该研究在最终操纵链中满足一定条件（如 box 小于原图的 1/6）的操纵之前随机插入**缩放操纵**（CropZoomIn）以补充。基于这种方式，该工作构建了 70k 条推理链数据，并对数据进行了全面的分析。

模型架构

CogVLM 是一个通用的大规模视觉语言模型框架，其通过引入**视觉专家**来对不同的模态进行深度融合。本工作的模型框架基于 CogVLM，具体来讲，模型包含四部分：1) Vicuna-7B-1.5 的语言基座；2) EVA2-CLIP-E 的视觉编码器；3) 基于 SwiGLU 的映射层；4) 视觉专家模块。模型总体包含 17B 参数，其中视觉专家模块占比 6.5B。

在此基础之上，该研究设计了基于 **KV-memory** 的多图多轮视觉语言模型架构。即，在训练阶段，在包含多次图片输入的一个推理链轮次，其保存 Transformer 每一层的 KV-memory 的累积，并在每一步的 attention 计算时考虑当前累积的 KV-memory。

当有新图片输入时，模型无需输入历史文本，而基于固定的 prompt 来要求模型基于历史 memory 进行推理。这种方式使得模型能够在兼容已有 VLMs 架构的同时接受多次图片输入，并有效缓解了上下文序列过长带来的问题。

CogCoM 基于两个主要阶段的训练，以构建具备推理能力的同时能够解决多个多模态任务的通用视觉语言模型。

- 在第一阶段，其首先基于由 LAION-2B 和 COYO-700M 获得的 1.5B 图文对模型进行预训练，其次基于由 LAION-115M 获得的 40M 清洗数据对模型进行 grounding 训练。
- 在第二阶段，该研究通过混合涵盖四种能力（instruction-following、OCR、detailed-captioning、CoM）的 570K 数据对模型进行对齐训练。

2.2 技术原理对比

维度	GPT-o3	CogCoM
推理机制	隐式思维链（CoT）+ 动态工具调用	显式操纵链（CoM）+ 结构化操作步骤
工具调用	自主选择工具（如搜索、代码执行）	预定义操作集合（如剪裁、计数）
可解释性	黑箱式推理路径	透明的操作序列与文本描述

训练数据	大规模多模态数据 + 强化学习	合成操纵链数据 + GPT-4标注
应用场景	复杂问题求解（如科研、编程）	需要可解释性的场景（如教育、医疗）
技术互补性	动态灵活性强，适合开放性任务	可解释性高，适合需透明推理的场景

虽然o3和CogCoM采用不同的技术路径，但它们共同解决了视觉推理的核心挑战：**如何将图像信息转化为可执行的推理步骤**。o3通过强化学习实现工具调用的自主性，而CogCoM通过显式操作链提升可解释性。这种差异并非对立，而是互补——例如，在医疗影像分析中，CogCoM的结构化操作链可用于生成诊断报告，而o3的动态工具调用可用于跨模态数据整合（如结合患者病历和基因数据）。

智谱的技术生态进一步强化了这种互补性。例如，GLM-4-Air的轻量化架构支持消费级显卡运行，而CogCoM的操纵链数据生成框架已被用于医疗影像分析，研究者通过微调模型实现了对病理切片的自动化标注。这种开源生态不仅加速了技术迭代，还促进了多模态推理的跨领域应用。

三、未来趋势：从单模态到全模态智能

4.1 技术挑战

- 模态对齐：如何更有效地融合视觉、语言、音频等多模态信息，避免“模态鸿沟”问题。
- 实时推理：在保证准确性的前提下，提升模型处理高分辨率图像和长视频的速度。
- 伦理安全：防止模型生成误导性或有害的视觉推理结果，特别是在医疗、金融等关键领域。

4.2 智谱的技术布局

智谱的技术路线图显示，其下一代模型将聚焦于**全模态智能**：

- **多模态强化学习**：通过模拟人类“观察-思考-行动”的闭环，训练模型在真实环境中执行复杂任务（如自动驾驶、机器人操作）。
- **神经符号整合**：将符号逻辑（如数学定理、物理定律）嵌入神经网络，提升推理的可解释性和可靠性。
- **边缘端部署**：优化模型架构，使其能够在手机、嵌入式设备等资源受限环境中运行。