

治愈还是毒药？嵌入指令在视觉-语言模型中视觉上改变幻觉

Zhaochen Wang¹ Yiwei Wang² Yujun Cai¹

¹The University of Queensland ²University of California, Merced

zhaochen.wang@uqconnect.edu.au

Abstract

视觉语言模型 (VLMs) 常常会遭受幻觉问题，部分原因是由于多模态信息对齐的挑战。我们提出了一种名为 Prompt-in-Image 的简单方法，该方法将文本指令直接嵌入到图像中。这消除了对单独文本输入的需求，并迫使模型通过视觉通道处理所有内容。我们在三个流行的开源 VLM 上评估了这种方法：Qwen2.5-VL、LLaVA-1.5 和 InstructBLIP。结果显示有显著差异。Prompt-in-Image 提高了 Qwen2.5-VL 的性能，将 POPE 准确率提升了 4.1% (从 80.2% 提高到 84.3%)，并且在 MS-COCO 上也减少了幻觉率。相比之下，LLaVA-1.5 和 InstructBLIP 的表现严重下降，准确性从大约 84% 降至接近随机水平。通过详细分析，我们发现 LLaVA 和 InstructBLIP 中基于 CLIP 的编码器对嵌入文本区域表现出过度的关注偏差，破坏了视觉理解。相比之下，Qwen 的视觉编码器能够稳健地处理带有嵌入文本的图像。关键的是，Prompt-in-Image 减少了 Qwen 的模态差距，通过统一单个模态的信息处理来增强跨模态对齐。

1 介绍

大多数现代视觉-语言模型 (VLMs) 遵循一个标准架构：视觉编码器 (通常是 ViT)、投影器和语言模型 (LLM 解码器)。由于视觉和文本组件是分别预训练的 (Rabinovich et al., 2023)，这种方法引入了固有的跨模态对齐挑战，显著阻碍了模型的整体性能。这些对齐问题的一个突出表现是语言偏见现象，即 VLMs 过多依赖于文本信息而忽略了视觉信息 (Niu et al., 2021; Wang et al., 2024a, 2025)。为了应对这些跨模态

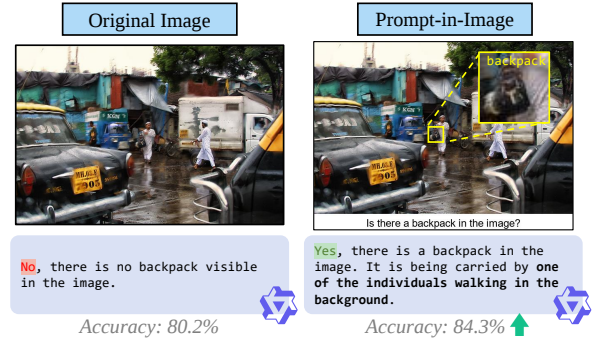


图 1: 示例的图像中包含提示文本：文本指令直接嵌入到图像中。使用图像中包含提示文本的方法，Qwen2.5-vl 的性能得到提升。

对齐挑战，之前的方法集中在改进跨模态融合上。然而，与其增强跨模态整合，我们问是否可以通过仅依靠单一模态的信息来完全避免跨模态对齐的挑战。

我们提出了 Prompt-in-Image (图 1)，该方法将文本指令直接嵌入图像中。通过迫使模型通过视觉通道处理所有信息，这种方法可能会增强融合并减少对齐问题。

我们使用幻觉作为主要评估任务有两个关键原因：(1) 幻觉代表了 VLM 开发中的一个重大挑战，模型会描述不存在的对象或忽略关键视觉细节，严重影响 VLM 的性能和可靠性；(2) 幻觉与模态对齐问题高度相关 (Liu et al., 2024)，使其成为评估我们方法的理想测试平台。

我们使用代表性的幻觉基准测试 POPE，广泛测试了三个流行的 VLMs：Qwen2.5-VL、InstructBLIP 和 LLaVA-1.5。令人惊讶的是，我们观察到了相反的效果。在 POPE 上，Qwen2.5-VL 的准确率提高了 4 - 5%，而 LLaVA-1.5 的表现

从 84% 急剧下降到 55% (近乎完全崩溃)。同样地, InstructBLIP 表现出与 LLaVA-1.5 一致的行为, 从 74.4% 下降到了 54%。我们的分析揭示了两个关键发现。首先, LLaVA 和 InstructBLIP 的崩溃源于其基于 CLIP 的编码器对文本的强大偏见, 这给予了嵌入文本区域过多的关注。这导致了严重的幻觉。相比之下, Qwen 的视觉编码器处理嵌入文本的图像要好得多。其次, 图片中的 Prompt 有效地减少了 Qwen 的模态差距, 提高了跨模态对齐并提升了性能。

总之, 我们的贡献有三个方面:

1. 我们提出了图像中的提示, 这是一种新颖的输入策略, 将文本嵌入图像以提高模态融合。
2. 我们对 Qwen-VL、InstructBLIP 和 LLaVA-1.5 进行了系统评估, 揭示了图像中的提示对幻觉性能产生不同影响。
3. 我们对性能差距进行了深入分析, 并解释了 Prompt-in-Image 如何通过减少模态差距和增强对齐来提升性能。

本文其余部分组织如下。第 2 节回顾了与 VLM 架构、幻觉问题及现有缓解方法相关的研究工作。第 3 节介绍了我们的 Prompt-in-Image 方法, 包括设计细节、评估基准 (POPE 和 MSCOCO)、测试模型以及实验配置。第 4 节报告了我们的实验结果, 展示了 Prompt-in-Image 对不同模型的对比效果。第 5 节提供了深入分析以解释这些不同的结果。最后, 第 6 节总结全文并讨论未来的研究方向。

2 相关工作

2.1 视觉-语言模型和幻觉

视觉-语言模型 (VLMs) 近年来迅速发展, 许多强大的模型如 GPT-4o (OpenAI, 2024)、LLaVA (Liu et al., 2023b) 和 Qwen-VL (Bai et al., 2025) 都取得了令人印象深刻的表现。这些模型通常具有相似的架构: 一个视觉编码器处理图像, 一个投影层, 以及一个语言模型生成文

本。尽管它们在各种任务中取得成功, 但幻觉仍然是 VLMs 面临的重大挑战。幻觉可以分为两类。判断性幻觉发生在模型对用户的查询响应与实际视觉数据不一致时。描述性幻觉是未能忠实描绘视觉信息 (Liu et al., 2024)。为了全面评估和量化这个问题, 研究人员提出了各种视觉语言模型幻觉评估方法和基准, 包括 POPE (Li et al., 2023), NOPE (Lovenia et al., 2023), CHAIR (Rohrbach et al., 2018), MMHal-Bench (Sun et al., 2023) 和 AMBER (Wang et al., 2023)。

2.2 幻觉原因及缓解方法

视觉语言模型中幻觉产生的原因复杂且多方面。几个重要因素导致了这个问题, 例如数据偏差 (Liu et al., 2023a), 视觉编码器的限制 (Li et al., 2024; Cho et al., 2022; Gong et al., 2024), 模态对齐不佳 (Sun et al., 2023), 以及大语言模型固有的幻觉。为了解决这些问题, 提出了各种方法。其中, 无训练对比解码 (CD) 策略在减少幻觉方面表现出有效性。对比解码通过比较原始输入和扰动输入的模型输出——如视觉噪声图像或修改后的指令——来抑制对语言先验的过度依赖, 从而减少幻觉。代表性的例子包括视觉对比解码 (VCD) (Leng et al., 2024) 和指令对比解码 (ICD) (Wang et al., 2024b)。然而, 这些方法也存在一些局限性, 包括推理速度变慢和性能提升有限。此外, 一些最近的研究 (Yin et al., 2025) 认为这样的解码策略可能与减少幻觉的原始目标完全无关。

3 方法

3.1 图像中的提示设计

在传统的视觉问答任务中, 用户会提供一张图片和文本指令 (提示), VLMs 处理图像和文本输入以完成任务。为了使模型依赖单一模态的信息, 我们采取了一种直接的方法: 将指令直接嵌入到图像中, 类似于电影字幕。然后我们只向 VLM 提供图像, 消除了单独的文本输入。我们将这种方法称为“提示在图像中”。

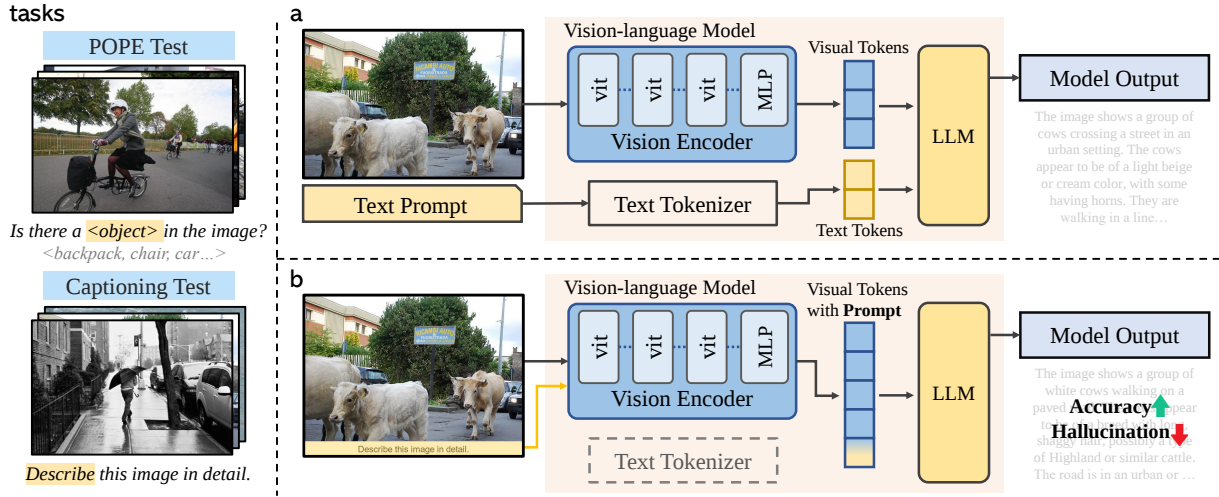


图 2: 交互范式的比较。(a) 传统的 VLM 交互需要同时输入图像和单独的文本提示。(b) 图像内嵌入指令直接将指令嵌入到图像中, 允许用户仅提供图像而无需额外的文本输入。

为了创建图像中的提示, 我们在每张图片的底部渲染问题文本。为了避免遮挡图片内容, 我们在每张图片下方添加一个独立的白色矩形区域。问题以黑色 Arial 字体 (26pt) 呈现, 确保机器可读性。这段文字区域仅覆盖总图高度的大约 5%, 最小化对原始视觉内容的干扰。为了控制与文本无关的视觉变化, 我们包括了一个空白白盒条件: 在该对照组中添加相同的白色矩形区域但不包含任何文本。该组中的图片大小保持与图像中的提示条件相同。

3.2 基准和评估指标

我们在两个互补的基准上评估 Prompt-in-Image:

基于采样的对象探测评估 (POPE) (Li et al., 2023): 该数据集通过二元问题 (如 “图像中有 <object> 吗?”) 来评估对象存在检测, 其中 <对象> 从三个拆分中选择: 随机 (随机选取的对象)、流行 (经常出现的对象) 和对抗 (与图像中的对象密切相关的对象)。我们专注于最具挑战性的对抗子集, 它最能反映现实世界的幻觉场景。完整的对抗数据集包含 3,000 个问题 (500 张图片 6 个问题)。为了平衡全面评估与计算效率之间的关系, 我们随机选择了 1,000 个问题用于测试。模型性能使用准确率和 F1 分数进行衡量。

MS-COCO 标题: 我们从 MS-COCO 2017 (Lin et al., 2014) 验证集中随机选择了 500 张图片, 并用 “详细描述这张图片。” 提示 VLMs。幻觉是通过椅子度量进行评估的 (Rohrbach et al., 2018), 该度量将生成的标题与真实物体标签进行了比较。我们报告了两个分数: 椅子 i , 所有提到的物体中幻觉物体的比例; 以及椅子们, 包含至少一个幻觉物体的标题的比例。

这两个数据集是成熟的基准, 并提供了一个全面的评估框架, 涵盖了二元问题回答和开放生成任务。

4 实验结果

4.1 模型

我们在两个广泛使用的开源 VLM 上评估了我们的方法: Qwen2.5-VL-7B (Bai et al., 2025), InstructBLIP-vicuna-7B (Dai et al., 2023) 和 LLaVA-v1.5-7B (Liu et al., 2023b). 这三个模型遵循相似的基于变压器的架构, 并在多模态任务中表现出强大的性能。重要的是, 这三种模型都能够准确地识别嵌入图像中的问题, 使它们适合评估 Prompt-in-Image 的有效性。

我们评估了四种输入配置以测试 Prompt-in-Image 的效果:

设置	Qwen2.5-VL-7B			InstructBLIP-7B			LLaVA-v1.5-7B		
	Acc.	F1	Yes Ratio	Acc.	F1	Yes Ratio	Acc.	F1	Yes Ratio
Baseline	80.2	0.76	0.32	74.4	0.72	0.53	84.0	0.86	0.62
Control	80.5(+0.3)	0.77	0.33	75.0(+0.6)	0.77	0.53	84.0(±0)	0.86	0.62
Hybrid	83.0(+2.8)	0.81	0.38	63.2(-11.2)	0.69	0.68	64.0(-20.0)	0.74	0.82
Prompt-in-Image	84.3(+4.1)	0.82	0.38	—	—	—	—	—	—
Prompt-in-Image [†]	82.1(+1.9)	0.81	0.43	54.0(-20.4)	0.70	0.99	55.0(-29.0)	0.70	0.99

表 1: 不同输入设置在三个视觉语言模型上的表现。图像中的提示（无指令）仅适用于 Qwen2.5-VL。[†] 带有明确指令“回答图像中的问题”。

设置	椅子们 (%)	椅子 i (%)
Baseline	32.3	8.8
Hybrid	34.2(+1.9)	9.9(+1.1)
Prompt-in-Image	24.7(-7.6)	6.7(-2.1)

表 2: 使用 CHAIR 指标在 MS-COCO 上的幻觉率。Prompt-in-Image 相较于其他输入模式减少了 CHAIRs 和 CHAIRi。

4.2 实验配置

- **基线:** 原始图像 + 文本提示。
- **图像中的提示:** 嵌入问题的图像 + 无文本提示（系统消息如“您是一个乐于助人的助手。”可能仍然存在）。
- **混合型:** 嵌入问题的图像 + 文本提示。
- **控制:** 图像中有一个空白的白框（无文字）+ 文本提示。这控制了由提示区域引入的视觉变化。

所有实验均在相同的推理参数和固定的随机种子下进行，以确保可重复性。解码温度设置为 0.7（默认值）。这些配置使我们能够隔离视觉输入、文本输入及其相互作用的影响。

VLMs 在处理我们的图像内提示样本时表现得非常不同。Qwen2.5-VL 可以自然地处理嵌入的问题——它可以直接读取并回答放置在图像中的问题，而无需我们提供任何额外的指令。然而，LLaVA-1.5 和 InstructBLIP 的表现则

有所不同。当我们没有给它们具体的文本指令时，它们往往会将嵌入的文本视为图像中需要描述的一部分。它们通常会给出对所看到内容的一般性描述，而不是回答嵌入的问题。

为了确保公平比较，我们为 InstructBLIP 和 LLaVA-1.5 在所有 Prompt-in-Image 测试中给出明确指令：“回答图像中的问题。”这防止了 LLaVA 仅仅描述图像，并确保两个模型都在尝试回答嵌入的问题。

此外，为了实验的严谨性，我们也使用 Qwen 模型进行了相同的实验（见表 1）。我们将在下一节详细分析这些结果。

4.3 POPE 评估

表 1 总结了三个 VLM 在 POPE 上的性能比较。结果揭示了模型之间的显著差异：

Qwen2.5-VL 证明了在图像中使用提示词的一致改进。准确率提高了 4.1%（从 80.2% 到 84.3%）。混合配置也显示出收益（+2.8%），尽管略低于图像中的提示词效果。我们通过具体示例来更好地了解图像中的提示词如何改变 Qwen 的行为（图 3，顶部）。结果显示了基线和图像中使用提示词响应之间的明显差异。令人惊讶的是，我们发现图像中的提示词有助于模型检测它之前遗漏的小的、不寻常的或部分隐藏的对象。更重要的是，Qwen 不仅能够识别这些具有挑战性的对象——还能告诉我们它们在图像中的确切位置（例如，“位于图像右侧”）。这表明图像中的提示词实际上提高了 Qwen 对视觉的理解能力以及将物体定位在场景中的能

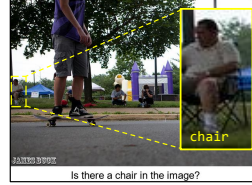
POPE Example 1:



Baseline: **No**, there is no backpack visible in the image.

Prompt-in-Image: **Yes**, there is a backpack in the image. It is being carried by one of the individuals walking in the background.

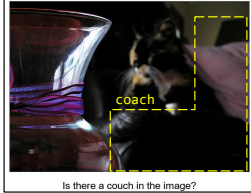
POPE Example 2:



Baseline: **No**, there is no chair in the image.

Prompt-in-Image: **Yes**, there is a chair in the image. It can be seen on the left side of the image, partially obscured by the person standing on the skateboard.

POPE Example 3:



Baseline: **No**, there is no couch in the image. The background appears to be a dark, blurred setting, possibly a room, but there is no visible couch.

Prompt-in-Image: **Yes**, there appears to be a couch in the background of the image. It is partially visible and out of focus, but the shape and cushioning suggest it is a couch.

POPE Example 4:



Baseline: **No**, there is no laptop in the image.

Prompt-in-Image: **Yes**, there is a laptop in the image. It is partially visible on the right side of the image, with the screen and part of the keyboard showing.

Caption Example:



Baseline: The image is a vintage advertisement featuring athletes on a tennis court showcasing tennis gear. The advertisement highlights "adidas" alongside the slogan, "get into top gear with adidas." Two tennis players are featured; one in the foreground appears poised to serve, carrying a tennis racket with the ball visible, and the other plays peek-a-boo scrape on the bench of the bag Court.

Prompt-in-Image: This image is an advertisement for Adidas, a well-known sportswear brand. The ad features two individuals engaged in a tennis match, set against a backdrop that resembles a tennis court. The woman in the foreground is dressed in a white tennis outfit with a red skirt and a white headband, holding a tennis racket mid-swing, with a tennis ball visible in the air. The man in the background is also dressed in a tennis outfit, wearing a blue and white striped shirt and shorts, and is holding a tennis racket as well. Both individuals appear to be in motion, suggesting an active and dynamic scene. The advertisement is framed within a large, dark blue box that resembles a product box, with the Adidas logo prominently displayed at the bottom. The text "adidas The science of sport" is written in white, emphasizing the brand's focus on sports performance. Above the image, the tagline "get into top gear with adidas" is written in white, encouraging viewers to associate the brand with high performance and readiness. At the bottom of the image, there is additional text that reads, "Official Supplier To Canada's Olympic Teams" which highlights Adidas's association with elite sports and its support for national teams. The overall design of the advertisement is clean and professional, with a focus on the brand's connection to sports and athleticism.

图 3: 案例研究关于 Qwen2.5-VL。上: POPE 示例比较基线和图像内提示对目标存在检测的响应。下: 一个 MS-COCO 标题生成示例。图像内提示有助于模型生成更详细的答案并减少幻觉。

力，而不仅仅是更频繁地说“是”。

LLaVA-v1.5 和指令 BLIP 表现出灾难性的性能下降。LLaVA-v1.5 的准确率从 84.0% 基线水平急剧下降到 55.0%，而 InstructBLIP 则表现出非常相似的行为，从 74.4% 下降到 54.0%。更令人担忧的是，两个模型的“是”比率跃升至 0.99，表明它们几乎对所有查询都默认回答“是”。这代表了两个模型辨别能力的完全丧失。

在所有模型中，对照组的表现几乎与基线相同，证实了性能变化是由图像中的文本存在引起的，而不是布局或格式的变化。

4.4 MS-COCO 验证集

鉴于 Qwen2.5-VL 在 POPE 上取得了有希望的结果，我们使用 MS-COCO Caption 进行了进一步的评估以衡量其在开放生成任务上的表现。我们测试了 500 张图像，并使用 CHAIR 指标对模型输出进行了评估。Prompt-in-Image 在幻觉指标上均表现出一致的改进。如表 2 所示，CHAIRs 下降了 7.6% (32.3% 24.7%)，而

CHAIRi 下降至 2.1% (8.8% 6.7%)，这表明在句子和实例级别上的幻觉都有所减少。图 3 (底部) 显示了这种改进的一个很好的例子。使用 Prompt-in-Image，模型生成的标题包含更多详细信息，包括图像中的衣物颜色和小字体等具体细节。这一点很有趣，因为该模型实际上提供了更多信息并给出了更多细节，但出错更少。Prompt-in-Image 帮助模型看到并描述实际存在的内容，而不是仅仅编造信息。

5 图像中提示的发散效应

我们的实验揭示了 Prompt-in-Image 在不同模型中产生的矛盾效果。本节将解答两个关键问题：

- 为什么 InstructBLIP 和 LLaVA 的性能在图像中的指令下会下降？
- 如何通过图像中的提示提升 Qwen 的性能？

我们在本节中详细考察这两个问题。

5.1 为何图像中的提示会伤害 LLaVA

我们分析了 CLIP ViT-L/14 (Radford et al., 2021), 它作为 LLaVA-v1.5 的视觉编码器, 并且与 InstructBLIP 中的视觉编码器密切相关。我们在不同层上可视化补丁级别的注意力权重 (图 4), 使用所有头部的平均注意力。具体来说, 我们检查了三个具有代表性的层: 浅层 (第 4 层)、中层 (第 12 层) 和深层 (第 24 层)。目标是比较 CLIP 处理两种输入类型的方式: 嵌入文本的图像 (图中有提示) 及其无文本对应物 (对照组)。

我们观察到, 在浅层 (第 4 层) 和中层 (第 12 层), 两种图像类型的注意力分布是相似的, 这表明在早期阶段对嵌入文本的敏感度有限。然而, 在深层 (第 24 层), 差异变得更加明显: CLIP 显示出明显的倾向, 即高度集中在文本区域上。

为了进一步确认 CLIP 的文字偏差, 我们通过分析来自 POPE 数据集的 35 张随机选择的 Prompt-in-Image 图像来量化这一影响。我们检查了 CLIP 编码器最终层 (第 24 层) 中的自我注意力模式, 重点是注意力矩阵的对角线, 这显示了每个补丁给予自身的注意程度。图 5 显示, 最后 32 个补丁, 对应于文字区域, 始终具有非常高的自我注意力分数。这证实了 CLIP 对嵌入的文字过于敏感, 在更深层次中给予了文本区域过多的注意。许多先前的研究 (Darcet et al., 2023; Gong et al., 2024; Zhang et al., 2024) 表明, 对某些视觉标记给予过度关注会导致 VLM 输出中的幻觉现象。当注意力权重不均衡地集中在特定的视觉标记 (如包含嵌入文字的那些) 上时, 模型失去了平衡局部和全局视觉信息的能力。这些主导标记可以覆盖视觉上下文和语言先验, 导致模型忽略实际图像内容并默认给出肯定响应, 而不考虑其他视觉证据。

此外, 我们在处理图像中的提示和控制图像对时, 从所有变压器层中提取特征, 并在每

个层的补丁之间计算余弦相似度。图 6 显示了最终 12 层的相似性分布, 其中高级语义表示出现。

结果显示了一个显著的差异: Qwen-ViT 在其深层中始终保持较高的相似性 (>0.95) 在 Prompt-in-Image 和 Control 图像之间, 而 CLIP-ViT 在相同层中的相似性则下降。这表明 Qwen 的视觉编码器即使在嵌入文本的情况下也保留了图像语义, 而 CLIP 随着层的加深对文本元素变得越来越敏感。

这种鲁棒性很可能源于 Qwen 多样化的预训练方案, 该方案不仅包括标准的图像-标题配对, 还包括交错的图文文档和 OCR 数据 (Bai et al., 2025)。通过在预训练过程中学习处理自然嵌入文本的图像, Qwen-ViT 发展出将文本视为普通视觉元素而非干扰信号的表示方法——这解释了它为何能成功解析视觉提示而不会发生灾难性的注意力转移。

5.2 为什么图像中的提示有助于 Qwen

与 InstructBLIP 和 LLaVA-1.5 相比, Qwen2.5-VL 在 Prompt-in-Image 条件下表现出一致的性能提升。在 POPE 数据集上, 准确率提高了 4 - 5%, 而在 MS-COCO 图像的开放性描述生成任务中, 幻觉率也有所下降。

我们提出了一种可能的解释: Prompt-in-Image 通过视觉通道统一输入模式, 从而增强模式融合并减轻对齐问题。这一假设得到了两个观察结果的支持。

文本输入干扰图像中提示的表现性能: 如第 4 节所示, 混合设置 (带有嵌入问题的图像和独立文本提示) 的表现不如图像中含问题设置 (仅带有嵌入问题的图像)。这表明引入额外模式 (即文本) 并不能帮助模型, 反而会降低性能。相比之下, 图像中含问题将所有相关信息整合到单一模式中, 有助于模型集中注意力并提升整体性能。

图像中提示减少模态差距: 以前的研究 (Liang et al., 2022) 已经识别出了视觉语言模型中的“模态差距”现象。虽然这些模型设

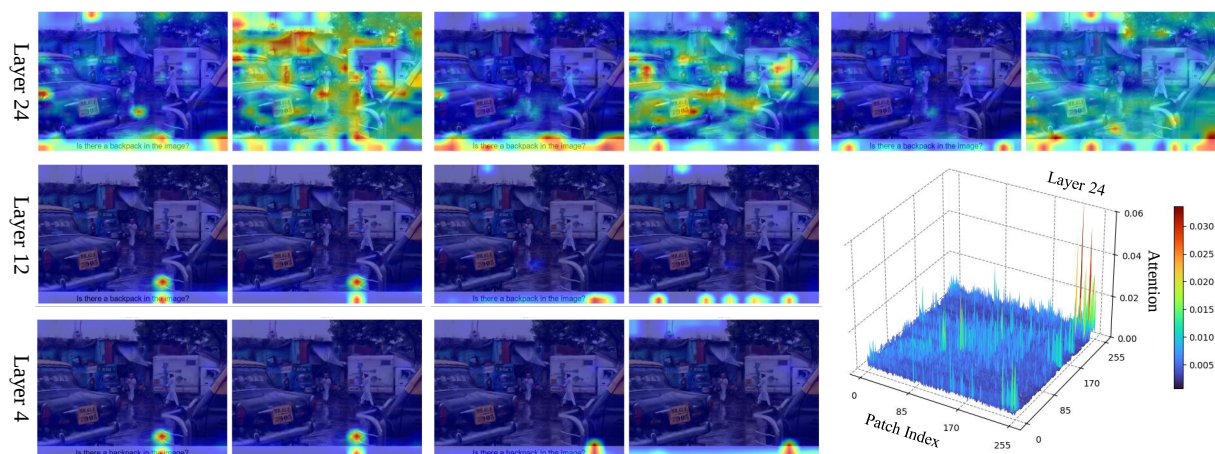


图 4: CLIP 注意力可视化在第 4 层、第 12 层和第 24 层上的示例图像。每行显示了图像内提示（嵌入文本）与对照组（无文本图像）的补丁级注意力权重。深层（第 24 层）表现出对文本区域的强大注意力偏置。

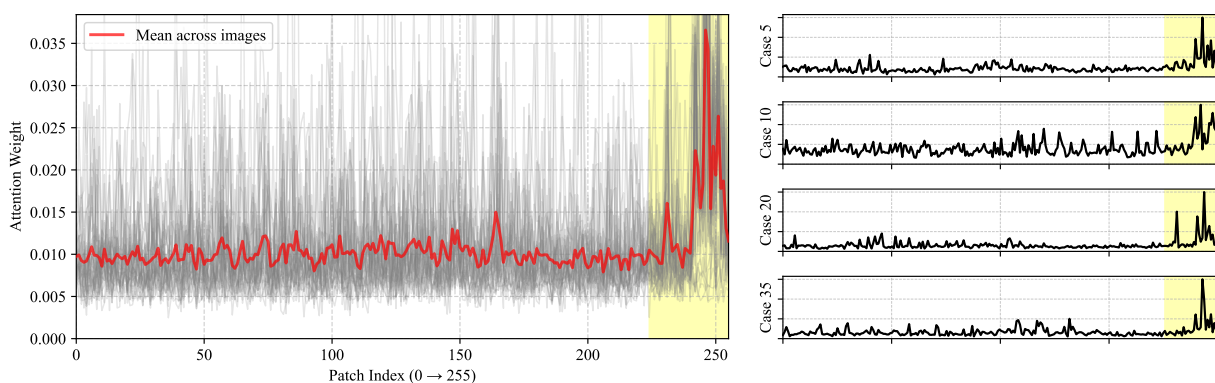


图 5: CLIP 最终层的自我注意力分析跨越 35 张 Prompt-in-Image 图像。红线 显示了所有图像的平均注意力权重，而 灰线 代表个别样本。黄色区域 对应于与文本相关的补丁，这些补丁始终获得更高的自我注意力分数。

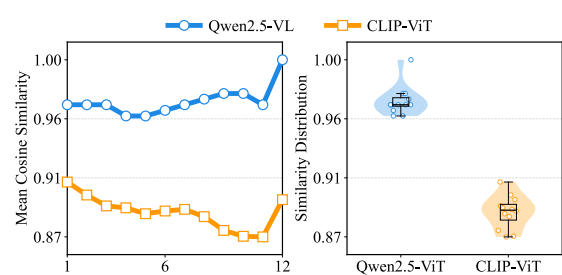


图 6: Qwen2.5-ViT 和 CLIP 在一个示例图像对上的逐层相似性比较。

计用于将图像和文本映射到共享的表示空间，但实际上不同模态在这个空间中明显分离。这种分离对多个下游任务的表现产生了负面影响，并且减少这个差距已经被证明可以提高模型性能 (Role et al., 2025)。

理想情况下，图像嵌入及其相应的标题嵌入应在表示空间中紧密重叠，这表明良好的跨模态对齐。为了测试 Prompt-in-Image 是否能改善 Qwen 中的这种对齐，我们检查了图像和标题嵌入之间的模态差距。

我们从 MS-COCO 测试集 4 中随机选择了 35 个样本，并形成了两组：

- 一幅带有空白白框（无文字）的图像及其通过明确的文字指令“详细描述此图像”生成的图注。
- 图像中嵌入了一个问题（图像中的提示）及其生成的标题，没有任何文本指令。

我们将图像及其对应的标题输入到 Qwen2.5-VL 模型中，并提取最终层的嵌入。然后，我

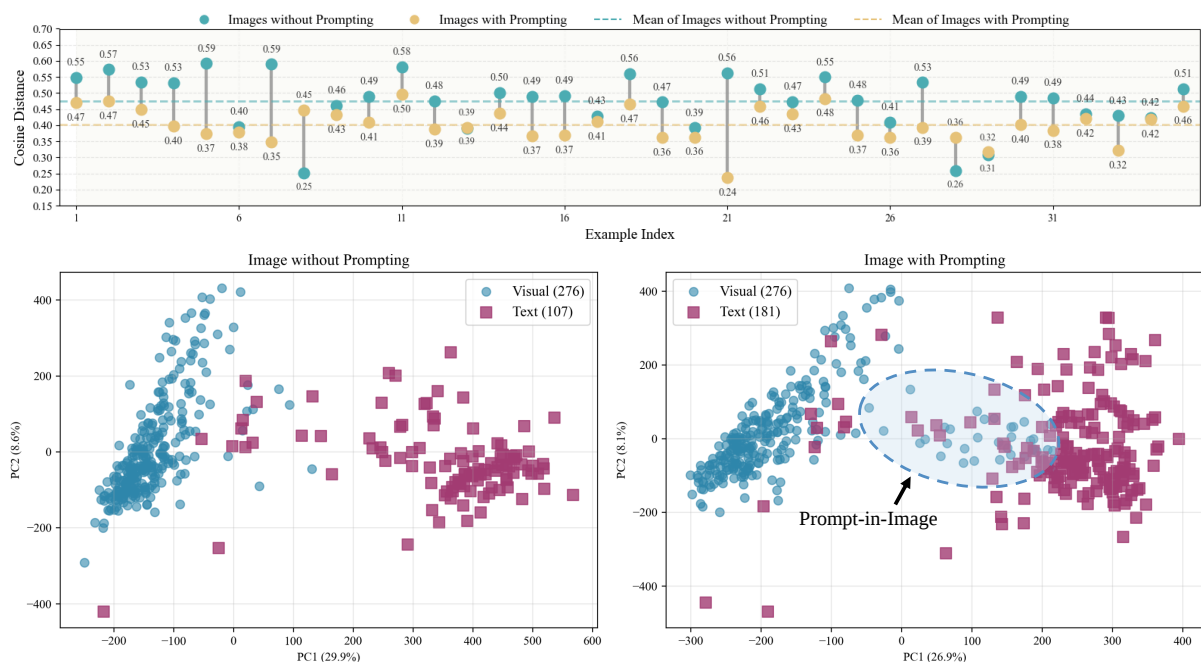


图 7: 模态差距分析对于 Qwen2.5-VL。顶部: 35 个样本中图像和标题嵌入之间的平均余弦距离比较。大多数含提示的图像样本 (黄色) 显示出比基线 (蓝色) 更低的模态差距。底部: 一个示例的图像 (蓝色) 和文本 (紫色) 嵌入的 PCA 可视化。含提示的图像 (右侧) 将两种模态拉得更近, 相比之下基线条件 (左侧)。

我们在共享语义空间中计算了图像标记和文本标记之间的平均余弦距离。结果显示, **Prompt-in-Image** 组的一致表现出较小的余弦距离, 平均减少了 12%。在图 7 中, 我们也展示了两组测试样本的 PCA 可视化。右侧图表 (**Prompt-in-Image**) 显示了一个更小的模态差距, 视觉和文本标记更加紧密对齐。这表明 **Prompt-in-Image** 作为两个模态之间的桥梁, 有效地缩小了差距并增强了多模态对齐。

6 结论

在这项工作中, 我们提出了 **Prompt-in-Image**, 这是一种简单而有效的策略, 将文本指令直接嵌入图像中以统一输入模式。通过对三个代表性的 VLMs (Qwen2.5-VL、InstructBLIP 和 LLaVA-1.5) 进行系统的评估, 我们观察到不同的效果: 虽然 **Prompt-in-Image** 持续提高了 Qwen 的性能并减少了幻觉现象, 但它显著降低了 InstructBLIP 和 LLaVA 的输出质量。

我们进一步分析了这一现象, 并识别出两个模型之间的关键差异。一方面, InstructBLIP

和 LLaVA (基于 CLIP) 表现出对嵌入文本区域的过度关注, 导致过分依赖局部模式并产生幻觉。相比之下, Qwen 展现出更强的鲁棒性。另一方面, 图像中的提示语有助于 Qwen 提升模态融合并缓解对齐问题。实证结果证实了图像中的提示语能够缩小模态差距并提升跨模态一致性。

这项工作表明, 在多模态数据上训练模型的方式真的很重要。它还建议, 可能值得进一步探索更简单、统一的视觉语言模型架构方法。

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023.

- Instructblip: Towards general-purpose vision-language models with instruction tuning.
- Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2023. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*.
- Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. 2024. Damro: Dive into the attention mechanism of l1vm to reduce object hallucination. *arXiv preprint arXiv:2410.04514*.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26763–26773.
- Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*.
- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.
- Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.
- OpenAI. 2024. Gpt-4o: Openai’s new flagship model. <https://openai.com/index/gpt-4o>. Accessed: 2025-05-23.
- Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby-Tavor. 2023. Predicting question-answering performance of large language models through semantic consistency. *arXiv preprint arXiv:2311.01152*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.
- François Role, Sébastien Meyer, and Victor Amblard. 2025. Fill the gap: Quantifying and reducing the modality gap in image-text representation learning. *arXiv preprint arXiv:2505.03703*.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.
- Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.
- Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*.

- Zhaochen Wang, Bryan Hooi, Yiwei Wang, Ming-Hsuan Yang, Zi Huang, and Yujun Cai. 2025. Text speaks louder than vision: Ascii art reveals textual biases in vision-language models. *arXiv preprint arXiv:2504.01589*.
- Hao Yin, Gunagzong Si, and Zilei Wang. 2025. The mirage of performance gains: Why contrastive decoding fails to address multimodal hallucination. *arXiv preprint arXiv:2504.10020*.
- Yudong Zhang, Ruobing Xie, Jiansheng Chen, Xingwu Sun, Yu Wang, et al. 2024. Dhcp: Detecting hallucinations by cross-modal attention pattern in large vision-language models. *arXiv preprint arXiv:2411.18659*.