通过稳健的局部感知搜索来缓解对象幻觉

Zixian Gao[1,2], Chao Yang[1†], Zhanhui Zhou[1], Xing Xu[2], Chaochao Lu[1]
[1]Shanghai Artificial Intelligence Laboratory
[2]Center for Future Media & School of Computer Science and Engineering,
University of Electronic Science and Technology of China
{ zixian.gaoo, asap.zzhou } @gmail.com,
{ yangchao, luchaochao } @pjlab.org.cn, xing.xu@uestc.edu.cn

Abstract

最近在多模态大型语言模型（MLLMs）方面的进展使它们能够有效地整合视觉和语言，处理各种下游任务。然而，尽管取得了显著的成功，这些模型仍然存在幻觉现象，即输出看似合理但与图像内容不一致。为了缓解这一问题，我们引入了一种在推理过程中使用的解码方法——局部感知搜索（LPS），该方法既简单又无需训练，但能有效地抑制幻觉。这种方法利用局部视觉先验信息作为价值函数来纠正解码过程。此外，我们观察到局部视觉先验对模型性能的影响在高图像噪声情境中更加明显。值得注意的是，LPS 是一种即插即用的方法，兼容各种模型。在广泛使用的幻觉基准和噪声数据上的大量实验表明，与基线相比，LPS 显著降低了幻觉的发生率，在嘈杂环境中显示出卓越的性能。代码可在 https://github.com/ZixianGao/Local-Perception-Search 获取。

1 介绍

近年来，多模态大型语言模型（MLLMs）经历了快速发展 (Du et al., 2022; Ghosh et al., 2024; Gupta et al., 2023; Zhu et al., 2023)，深刻地改变了多模态学习领域。这些模型在广泛的任务中展示了卓越的能力，包括图像描述 (Zhang et al., 2024a; Kim et al., 2023)、视觉问答 (Antol et al., 2015; Kamalloo et al., 2023)、多模态推理 (Ma et al., 2023; Qi et al., 2024) 等。然而，尽管有这些进步，MLLMs 仍不是完全可靠的。在现实场景中，特别是当视觉模态受到对抗性扰动时，MLLMs 常常出现对象幻觉 (Zhang et al., 2024c; Gao et al., 2024) 的问题，即模型错误地生成了不存在的对象在其输出中，导致生成的文本与视觉内容之间的不匹配。

对象幻觉仍然是一个持续的挑战，限制了 MLLM 在安全关键应用中的部署。近年来，这一问题引起了越来越多的关注。扩大训练数据 (Zhao et al., 2024; Fu et al., 2024) 的规模和质量

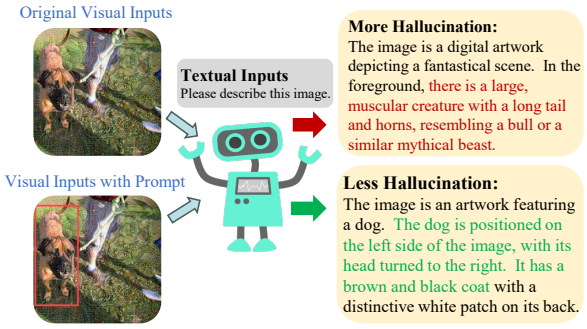

Figure 1: 我们观察到了在不同的 MLLMs 中存在一致的现象：当模型仅接受原始视觉输入时，会因图像中的对抗扰动而产生幻觉。相反，当给予视觉输入并附有提示（例如一个边界框）时，模型能够正确地识别图像中的物体为一只狗。

已被证明是增强模型性能和减少幻觉的有效方法。然而，这一策略也会带来显著的标注成本，并在训练过程中造成相当大的计算开销，从而阻碍扩展性。对此，最近的研究探索了推理时间搜索技术，作为一种有前途的补充策略，可以在不重新训练的情况下改善响应质量。

一些工作 (Zhou et al., 2024b; Liu et al., 2024b; Kim et al., 2024) 提出了推理时的搜索算法，旨在减轻幻觉。例如，(Wang et al., 2024) 将搜索过程表述为一个马尔可夫决策过程，并采用在精选数据上训练的高质量过程奖励模型（PRM）来指导解码。(Chen et al., 2024) 采用对比解码策略，利用预训练的基于目标检测的视觉识别对图像中的物体进行识别，并提供辅助视觉定位。虽然这些方法显示出潜力，但每种方法都有其局限性：前者需要训练专用的 PRM，这涉及到昂贵的注释和计算需求；后者则过分依赖外部的视觉定位，这降低了可扩展性并限制了其在实际中的应用。这些缺陷促使我们提出第一个研究问题：我们能否设计一种仅依赖 MLLM 内部功能的推理时搜索策略，而无需外部监督或辅助模型？

此外，最近的研究 (Pi et al., 2024; Guo et al., 2024; Ding et al., 2024; Zhang et al., 2024b) 强调了多模态大模型中视觉模态的脆弱性。模型的

† Corresponding author. Work done while ZG and ZZ were at Shanghai AI Lab.

大部分能力继承自大型语言模型的骨干，而不足的视觉-语言对齐使得视觉模态特别容易受到对抗性攻击的影响。虽然前述技术在一定程度上缓解了幻觉，但当视觉输入受损时，它们仍然无效。如图 1 所示，一次无目标的对抗性攻击导致严重的图像失真，然而模型继续产生不存在的对象幻觉（例如，红色文本中的幻觉动物），无法识别图像中的实际物种。最近的努力，例如 (Fang et al., 2024)，提出了不确定性感知解码策略，量化视觉不确定性并使用基于投票的集成来增强响应的可靠性。然而，这些方法在处理干净图像时会牺牲性能，并由于多次推理而计算成本高。这引导我们提出第二个研究问题：是否有可能在推理时间进行可信的干预，既能在友好输入上保持模型性能，又能在对抗性攻击下减轻幻觉?

为了解决这两个挑战，我们提出了一种新颖的局部感知先验引导搜索策略。我们的方法利用 MLLM 的内在能力，专注于局部视觉区域以生成指导解码的先验，从而产生稳健且高质量的响应。具体而言，如图 1 所示，我们观察到当边界框覆盖在对抗性扰动图像上时，模型成功识别出物体（例如，狗）。然而，要求手动或模型生成的边界框与我们的第一个研究目标相矛盾。受到这个观察结果的启发，我们进一步探索了模型的局部视觉注意力能力，发现 MLLM 能够通过关注局部化的区域来正确识别物体，即使没有显式的边界框监督。我们利用这一洞察，从模型自身对局部区域的注意力中构建先验，实现在推断时的指导，而无需依赖外部工具或标注。这种方法允许在不影响不同视觉条件下的性能的同时，有效地缓解幻觉现象。

我们的贡献可以总结如下：

- 我们识别并利用 MLLMs 的内在局部感知能力，展示了即使在视觉模式受到对抗扰动时，这一特性也可以被用来稳健地引导解码过程。
- 我们提出了一种称为 LPS（局部感知搜索）的可插拔推理时搜索算法，该算法利用模型的局部视觉感知而无需外部监督。LPS 显著减少了在各种场景中的物体幻觉。
- 我们对广泛采用的幻觉基准进行了全面评估，以验证我们方法的有效性。此外，我们收集并扩展了一个对抗性扰动的幻觉数据集，以验证在攻击条件下 LPS 的稳健性。

2 相关工作

多模态大型语言模型。在近年来，大型语言模型（LLMs）的快速发展也推动了多模态学习领域的重大创新。以 LLMs 为基础的多模态大型语言模型（MLLMs）已经成为解决视觉-语言任务的主导范式。早期的工作如 LLaVA 和 InstructBLIP 采用了视觉指令调优，实现了视觉和语言语义空间的有效对齐，从而能够在两种模态之间进行统一解码。随着训练数据的数量和质量的增加，以及算法的持续改进，MLLMs 展示了卓越的能力。然而，尽管这些模型表现出色，但在视觉模态受到对抗扰动时，它们并不总是可靠。在这样的情况下，MLLMs 往往会出现幻觉，生成与视觉内容不一致的文本输出，这显著限制了它们在安全关键应用中的部署。我们的工作旨在减轻当前 MLLMs 中普遍存在的幻觉问题，并促进它们在广泛应用领域中的可靠部署。

多模态大型语言模型中的幻觉。物体幻觉（OH）在多模态大型语言模型（MLLMs）中一直是一个持久的挑战，近年来备受关注。先前的研究 (Gunjal et al., 2024; Zhai et al., 2023) 将 OH 分为三种主要类型：物体存在幻觉、属性幻觉和关系幻觉。最近，许多研究 (Li et al., 2023; Rohrbach et al., 2019; Wang et al., 2023) 关注于评估和检测 OH。例如，POPE (Li et al., 2023) 将 OH 公式化为一个二元分类问题，要求模型判断图像中是否存在特定物体。与此同时，大量努力已经投入到开发旨在减少 MLLMs 中幻觉的方法，包括基于训练的 (Gunjal et al., 2024) 和推理时的 (Wang et al., 2024) 方法。然而，这些方法通常存在显著的资源开销，或在很大程度上依赖于外部模型提供的先验知识，限制了其在现实世界中可扩展性和适用性。这种持续改进清楚地表明，LPS 不仅在不同模型架构中有效，而且对模型容量的变化表现出显著的鲁棒性，进一步突出其可扩展性和广泛的现实世界适用性。

推理时搜索。推理时搜索 (Zhou et al., 2024b; Li et al., 2024; Tian et al., 2024) 已被广泛应用于大型语言模型（LLMs）领域，在推理和幻觉减少方面发挥着关键作用。鉴于在多模态大型语言模型（MLLMs）中广泛使用 LLM 主体，这些搜索策略在多模态环境中也获得了相当大的关注 (Kim et al., 2024; Chen et al., 2024)。这些策略的一个关键组成部分是在搜索过程中提供高质量的奖励信号。现有方法通常依赖于外部模型，例如 CLIP (Zhou et al., 2024a) 来提供奖励信号，或者通过使用经过筛选的数据集训练专门的过程奖励模型（PRM）(Wang et al., 2024) 来对候选输出进行评分。然而，这些方

法通常效果有限或带来显著的计算开销，其在 MLLMs 中的应用仍未被充分探索。在我们的工作中，我们利用模型自身的内在能力来生成先验信息并推导奖励信号，从而增强 MLLMs 在各种场景中的鲁棒性和泛化能力。

3 方法

3.1 预备知识

我们首先介绍物体幻觉现象。考虑一个具有参数 θ 的 MLLM M_θ 。该模型将提示图像对 (x, I) 作为输入，并自动回归解码为对应的文本 y 。形式上，我们可以将其数学表达为：

$$y_t \sim p_\theta(\cdot|v, x, y_{<t}) \propto exp f_\theta(\cdot|v, x, y_{<t}) \quad (1)$$

其中 y 表示第 $t-th$ 步的标记，$y_{<t}$ 表示在时间步 t 之前生成的标记序列。f 是由 M_θ 生成的对数分布。当生成的文本 y 的某些部分与输入图像 I 不一致时，就会发生物体幻觉现象，这种现象在 I 受到攻击时尤其明显。我们的最终目标是在有效防止图像攻击加剧幻觉现象的同时，始终保持高质量的文本生成，以缓解幻觉现象。

3.1.1 MLLM 推理的表述

我们输入一个提示图像对，并生成相应的 $y = [y_1, y_2, ..., y_m]$ ，其中 y 由 m 步骤级响应组成。每个步骤级响应 y 被视为从条件概率分布 $y_t = p_\theta(\cdot|v, x, y_{<t})$ 中抽取的样本。在本文中，我们使用句子级响应，其中每个步骤输出一个句子。因此，文本生成任务可以被表述为一个马尔可夫决策过程问题。在这个过程中，奖励函数 R 评估每个动作的奖励，这在大型语言模型（LLMs）中也被称为过程奖励模型（PRM）。一个更好的奖励函数往往能产生更好的响应 y 。许多方法已经在这个领域提出，例如用较小的模型指导大模型或者训练一个更好的价值模型。我们的方法旨在利用 MLLMs 的固有能力来指导搜索过程，从而改善生成质量并节省计算资源。

3.2 MLLMs 中的局部感知能力

正如在 1 中所展示的，我们发现视觉提示在帮助多模态大语言模型（MLLMs）识别图像中的物体，特别是在嘈杂条件下，起到了显著作用。然而，在实践中应用这样的视觉提示是具有挑战性的。一种方法涉及手动标注，而另一种依赖于专门的模型进行辅助标记。前者需要大量的人力劳动，使其在实际应用中不切实际。后者不仅会产生额外的计算成本，而且在嘈杂条件下也无法准确标记视觉提示。因此，我们正在探索更高效且精确的方法来整合视觉先验。

通过进一步探索，我们发现 MLLMs 本质上具备一种类似于可视化提示在感知图像中物体时所提供的帮助，这种能力我们称之为局部感知能力。如图 2 所示，我们对图像应用了一个无目标攻击 (Chen et al., 2023) ，导致图像模糊和失真。此时，MLLMs 难以辨别图像中物体的特征。当提示“请列出这张图中的物体”时，模型无法正确识别或描述其中的物体。然而，当提示“请列出这张图中左侧的物体”时，模型成功识别出图像左侧有一只狗。这个观察表明模型本质上具备关注局部物体的能力。此外，这种能力增强了模型的可靠性，帮助其减轻受到攻击的噪声影响，并实现准确的物体识别。这一发现为开发安全且值得信赖的推理模型奠定了基础。

从 MLLMs 的自回归推理过程可以看出，一个高效的奖励函数对于提供正确反馈是至关重要的，这有助于图像中的对象识别并促进精确解码。然而，现有方法要么需要从头训练一个奖励模型以提供反馈，要么依赖辅助模型提供比较解码的先验信息。两种方法都导致额外的资源消耗。因此，我们寻求一种高效且无需资源的方法来构建奖励函数。如前所述，MLLMs 本身就具备局部感知能力。受到这一点的启发，我们利用 MLLMs 的这种能力，通过模型自身生成先验信息。

如前所述，当多模态大语言模型（MLLMs）接受的视觉模态输入遭受对抗攻击时，模型通常无法识别图像中的物体，从而产生幻觉。然而，如图 X 所示，当模型的视觉注意力被限制在局部区域时，即使在攻击下，它仍然能够正确识别物体。利用这一特性，我们生成了先验信息以辅助模型。具体来说，我们使用以下提示：

> 请依次仔细观察图像的上、下、左、右部分，并列出每个部分存在的物体。

接下来，我们将目标图像和提示输入到 MLLM 中，使模型在进行主要任务之前，利用其全局-局部感知能力生成先验信息：

$$y = M_\theta(x, I), \quad (2)$$

，其中 y 表示模型生成的文本先验信息。

在 3.1.1 中，我们定义了多模态大型语言模型 (MLLMs) 的推理过程。我们将文本生成任务表述为马尔可夫决策过程 (MDP)，其特征为一个元组 (S, A, R, γ) 。其中，S 表示状态空间，在该状态空间中，每一个状态代表生成的句子与关联图像的组合。初始状态 s_0 对应于输入图像 I 和文本提示 x 。A 代表动作空间，其中每一个动作对应于在给定步骤生成的句子。奖

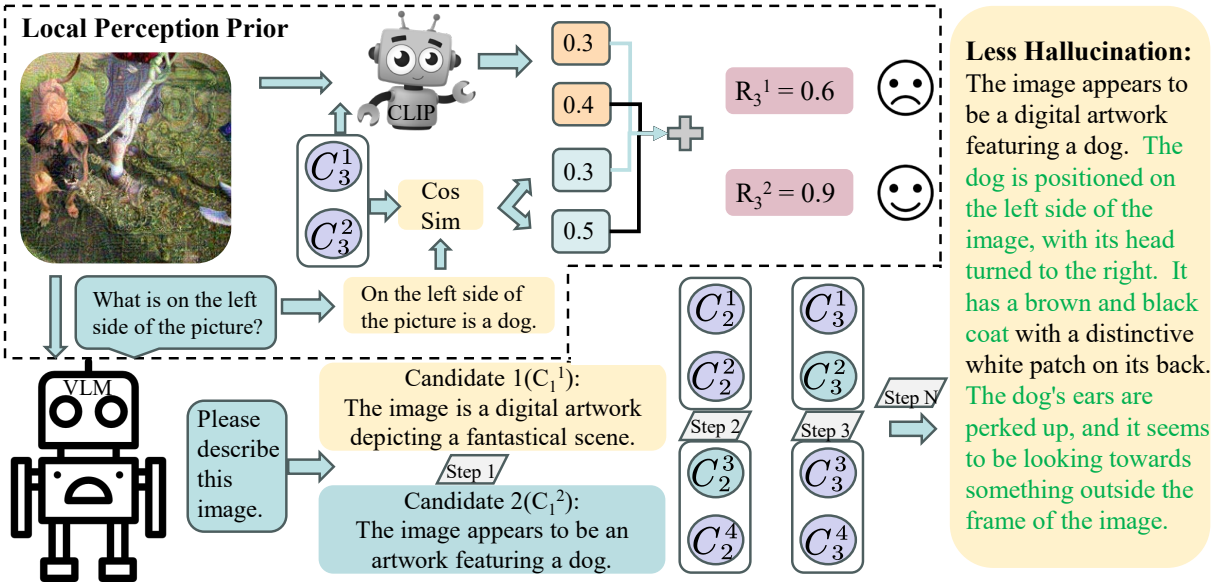

Figure 2: 我们方法的整体推理框架。虚线框强调了局部感知先验过程，该过程基于局部视觉线索优化候选选择。其余部分描述了单次推理过程中候选搜索过程。用绿色背景突出显示的候选者表示每个时间步长中选择的选项。

励函数 R 评估每个生成动作的质量，γ 表示折扣因子。

在这项工作中，我们旨在设计一个既节省资源又具有鲁棒性的奖励函数 R。对于条件分布 $p_\theta(\cdot|x, I, y_{<i}, T_n)$，我们在每个生成步骤 t 采样 k 个候选输出，记作 $\hat{y}_t = [c_t^1, c_t^2, ..., c_t^k]$。每个候选 c_t^i 都会使用先前阶段提取的对象先验信息 y 进行评估。具体来说，我们计算候选嵌入与先验嵌入之间的余弦相似度作为奖励：

$$R_{LPS}^t(i) = \text{CosSim}(y, c_t^i) = \frac{y \cdot c_t^i}{\|y\| \cdot \|c_t^i\|}, \quad i = 1, 2, \ldots, k. \tag{3}$$

我们定义在生成步骤 t 分配给 $i-th$ 候选的奖励为 $R_{LPS}^t(i)$，代表候选输出与对象级先验知识之间的语义一致性程度。该相似度评分作为每步的奖励，提供了一个语义评估信号，引导模型生成与提取的对象级先验更一致的输出。

此外，为了确保每个生成的候选句与输入图像在语义上相关，我们结合了基于视觉-语言相似性的第二个对齐信号。具体而言，我们使用 CLIP 嵌入来计算每个候选句与图像之间的余弦相似性。形式上，对于每个候选句 c_t^i，我们将基于 CLIP 的图像-文本相似性分数定义为：

$$R_{\text{CLIP}}^t(i) = \text{CLIP}(I, c_t^i). \tag{4}$$

这种双重对齐策略——利用对象级别的先验匹配和全局图文匹配——使我们能够生成的文本不仅在语义上与提取的先验对齐，而且与图像的视觉内容一致。为了整合这两个信号，我们将最终的奖励函数定义为基于先验和基于 CLIP 的相似性的加权组合：

$$R_t^i = \alpha R_{\text{LPS}}^t(i) + \beta R_{\text{CLIP}}^t(i) \tag{5}$$

，其中 α 和 β 控制局部先验匹配和全局视觉语言对齐的相对贡献。这样的设计通过鼓励局部语义一致性和全局视觉对齐，有助于缓解在多模态生成任务中常见的幻觉现象。

在每一代步骤 t 中，选择具有最高奖励 R_t^i 的候选 c_t^i，以确保最小的幻觉体验和与视觉输入的强语义对齐。这个逐步选择的过程在整个序列生成过程中重复进行，最终产生最终输出 $\hat{y}$。

4 实验

4.1 实验设置

实验设置。为了验证我们提出的方法在不同多模态大语言模型（MLLMs）中的有效性，我们在以下模型中实现了我们的方法：Llava 1.5（7B，13B）(Liu et al., 2024a)，Qwen 2.5 VL（7B）(Yang et al., 2024)，LLaMA 3.2-V（11B）(Grattafiori et al., 2024)，以及 Phi 3.5-V（4B）(Abdin et al., 2024)。我们针对使用 CLIP 作为过程奖励模型的传统基线进行了比较评估。此外，我们在选定的数据集上使用视觉对比解码（VCD）(Leng et al., 2023) 策略进行了进一步的比较实验。
基准测试和评估指标。

根据标准协议 (Leng et al., 2023; Chen et al., 2024)，我们使用了两个基准数据集——POPE (Li et al., 2023) 和 CHAIR (Rohrbach et al., 2019)

Method	POPE								CHAIR			
	adversarial		popular		random		overall					
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	C_s ↓	C_i ↓	B_1 ↑	B_4 ↑
Qwen 2.5 VL+ CLIP PRM	82.2	78.8	82.9	79.8	83.3	80.1	82.8	79.6	24.0	8.6	4.4	0.1
Qwen 2.5 VL+ LPS	83.7	81.2	84.4	81.9	85.4	83.0	84.5	82.0	20.4	7.9	8.1	1.7
Llama 3.2 Vision + CLIP PRM	82.4	83.8	85.8	86.5	89.8	89.8	86.0	86.7	24.0	10.3	3.0	0.7
Llama 3.2 Vision + LPS	83.0	84.3	86.0	86.7	90.7	90.7	86.6	87.2	22.2	8.9	3.1	0.7
Phi 3.5 Vision + CLIP PRM	81.4	81.8	81.7	81.5	84.8	84.8	82.7	82.7	22.2	10.1	14.7	2.9
Phi 3.5 Vision + LPS	82.7	82.1	83.8	82.8	85.8	84.6	84.1	83.2	21.2	8.5	14.7	3.1

Table 1: 关于 POPE 和 CHAIR 的结果。CLIP-PRM 表示使用 CLIP 作为奖励模型的基准方法，而 LPS 是我们提出的方法。在每个设置中的最佳表现以加粗显示。

——来评估模型在幻觉现象下的鲁棒性。POPE 针对 VQA 中的对象级幻觉，包含诸如“图中有狗吗？”的二元问题，这些问题是通过图像-对象标注三元组形成的。CHAIR 通过测量生成的描述中不存在对象的存在来关注描述级幻觉。它包括 CHAIR-S（句子级幻觉率）和 CHAIR-I（实例级幻觉率）。对于 CHAIR，我们从 MSCOCO 2017 验证集中抽取了 500 张图像 (Lin et al., 2014) 。

此外，为了评估模型在噪声条件下的鲁棒性，我们遵循了 Multitrust (Zhang et al., 2024c) 中的设置并扩展了他们的无目标攻击数据集。具体而言，我们在对抗性扰动的视觉模态下测试了模型和方法。在此设置中使用的攻击本质上是黑箱的，并遵循文献 [24] 中提出的 SSA-CWA (Chen et al., 2023) 策略，该策略以其对未见多模态语言模型的高转移性而闻名。令 $\{f_i^v\}_{i=1}^N$ 表示一组代理模型的视觉编码器。对抗性攻击的目标被表述为：

$$\max_{x_{\text{adv}}} \quad \sum_{i=1}^{N} \|f_i^v(x_{\text{adv}}) - f_i^v(x)\|_2^2$$
$$\text{s.t.} \quad \|x_{\text{adv}} - x\|_\infty \leq \epsilon \tag{6}$$

，其中 ϵ 表示允许的最大扰动幅度。此目标旨在生成对抗性样本 x_{adv} ，这些样本在多个代理编码器的特征空间中显著偏离，同时在 L_∞ 范数约束下视觉上与原始输入 x 保持相似。

实现细节。我们使用 LPS 策略实现了句子级别的推理时间搜索，其中句号符号（“.”）被用作分割定界符。在每个解码步骤中，生成了四个候选延续，最大解码步骤数设置为 10。所有实验都在一台 NVIDIA A100 GPU 上进行。

4.2 评估结果

一般幻觉基准测试结果。我们在三个最先进的多模态大语言模型（MLLMs）上进行了实验，分别是 Qwen 2.5 VL、Llama 3.2 Vision 和 Phi 3.5 Vision。为了评估我们提出的方法在解决幻觉问题上的有效性，我们在两个被广泛采用的基准测试上进行了实验：POPE 和 CHAIR。结果总结在表 1 中，最佳分数以粗体突出显示。实验结果表明，我们提出的 LPS 方法在这两个数据集的所有指标上始终优于基线，带来了显著的性能提升。与使用 CLIP 模型作为过程奖励模型的基线相比，我们的方法在 Qwen 2.5 VL 上实现了超过 1 % 的绝对增益。此外，它在另外两个模型上也显示出显著的优势。然而，我们也观察到对于 LLaMA 3.2 Vision 模型，我们的方法在 POPE 数据集的热门和随机类别上的性能提升相对较小。我们推测，这可能是因为所提出的局部感知敏感性在模型受到对抗攻击时更加有效。在挑战性较低的场景中，视觉模态特征受到的干扰较少，这种能力的优势可能不太明显。

Step Count	Method	Model		
		Qwen 2.5 VL	Llama 3.2 Vision	Phi 3.5 Vision
100	CLIP PRM	47.7	78.0	61.4
	LPS	49.5	80.6	65.4
500	CLIP PRM	48.8	78.3	61.5
	LPS	49.3	78.4	64.2

Table 2: 多重信任在不同的视觉-语言模型和解码策略中的结果。

攻击幻觉基准上的结果。为了进一步评估我们方法在视觉模态攻击下的鲁棒性，我们遵循 Multitrust 设置并将其无目标攻击数据集从原始的 100 个样本扩展到 1000 个样本。此外，为调查不同攻击强度对模型性能的影响，我们进行了使用 100 和 500 攻击步的实验。结果在表 2 中展示。从结果中我们观察到我们的方法在所有攻击配置下都始终优于基线。特别是，对于 Phi 3.5 Vision，我们的方法在两个攻击步设置下均实现了将近 3 % 的性能提升。值得注意的是，我们还发现基于 CLIP 的 PRM 基线的性能在 100 和 500 攻击步之间基本保持不变，表明这种方法对无目标攻击极为脆弱——即便是轻微的扰动也能导致不可逆的性能退化。相比之下，我们提出的 LPS 方法在轻微和严重攻击设置下均表现出性能提升，清楚地强调了其在对抗性场景中的鲁棒性。

4.3 进一步分析

与其他解码方法的比较。为证明我们提出的 LPS 方法的有效性，我们将其与最先进的方法 VCD 进行比较。实验在 POPE 数据集和我们构建的 Multitrust 数据集上进行。如图 **??** 所示，结果表明，我们的方法在不同模型上均优于 VCD，包括 Qwen 2.5 VL 和 Phi 3.5 Vision。值得注意的是，在 Multitrust 数据集上，我们的方法取得了显著的性能领先。这些结果强烈表明，与 VCD 相比，我们的方法在对抗攻击方面表现出更强的鲁棒性。

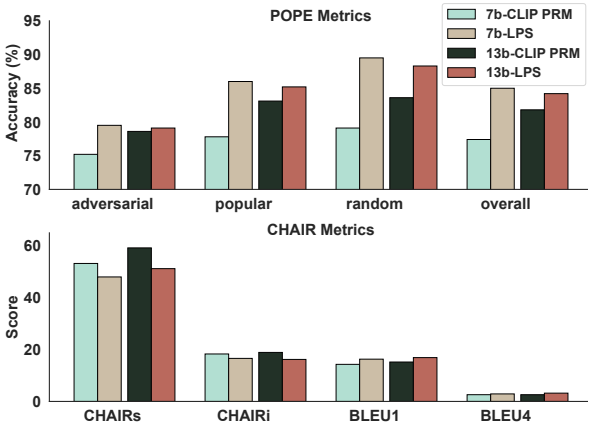

Figure 3: 比较 CLIP-PRM 和 LPS 在不同参数规模（7B vs. 13B）的 LLaVA 1.5 模型上在 POPE 和 CHAIR 数据集上的表现。

不同参数规模模型的对比。为了进一步验证我们提出的 LPS 方法的通用性，我们在同一架构家族中对具有不同参数规模的模型进行了比较研究。具体来说，我们在 POPE 和 CHAIR 数据集上评估了 LLaVA 1.5-7B 和 LLaVA 1.5-13B，并应用了基线 CLIP-PRM 和我们的方法 LPS。如图 3 所示，LPS 在这两个数据集上的表现一致优于 CLIP-PRM，并且适用于两种模型尺寸。值得注意的是，在 POPE 数据集上，LPS 在所有评估指标上均比 CLIP-PRM 提高了超过 2 %，无论基础模型是 7B 还是 13B。这种持续的改进清楚地表明，LPS 不仅在不同的模型架构上有效，而且在模型容量变化时仍然表现出显著的鲁棒性，进一步突显了其可扩展性和广泛的通用适用性。

候选数量的消融研究。为了评估我们提出的算法的有效性，我们对每一步考虑的候选数量进行了消融研究。具体来说，我们测试了候选数量从 1 到 8 时模型的性能。如表 **??** 所示，结果表明对于几乎所有模型，随着每个时间步候选数量的增加，性能持续提升。这个趋势支持了所提出的局部感知先验在选择合理候选时的有效性。此外，随着候选数量继续增加，性能提升逐渐减少。我们假设这是因为模型在任务上的性能有一个固有的上限，并且随着模型接近这个极限，进一步的改进变得越来越不明显。

Model	Candidate Number				
	1	2	4	6	8
Qwen 2.5 VL	46.8	48.1	47.7	49.5	49.3
LLaMA 3.2 Vision	76.1	77.3	78.0	80.6	80.3
Phi 3.5 Vision	62.2	63.7	62.9	65.4	63.8

Table 3: 关于 Qwen 2.5 VL、LLaMA 3.2 视觉模块和 Phi 3.5 视觉模块在 Multitrust 数据集上候选数量的消融研究。

定性分析。为了更直观地理解我们方法的有效性，我们进行了定性分析，如图 4 所示。在图中，我们展示了不同方法在对抗攻击下图像描述任务生成的完整响应。正确的描述以绿色突出显示，而虚构或不正确的描述以红色标记。从结果中可以明显看出，我们提出的方法 LPS 能够准确识别和描述物体，例如蝴蝶，即使在严重的视觉失真下也是如此。相比之下，基线方法 CLIP-PRM 在解读对抗扰动的图像时表现挣扎，常常依赖于诸如颜色或纹理等肤浅的线索，并且无法做出正确的内容层级判断。

我们提出了一种新颖的解码方法，称为局部感知搜索（LPS），该方法利用局部感知先验来指导多模态大型语言模型（MLLMs）的解码过程，旨在减轻在各种条件下的幻觉现象，特别是在视觉模态受到对抗攻击时。我们首先提供实证证据，证明 MLLMs 中存在局部感知能力。在此观察的基础上，我们增强并利用这种固有能力，让模型生成自己的先验来指导解码，从而避免依赖外部模型或额外组件。我们在标准设置下对广泛采用的幻觉基准 POPE 和 CHAIR 进行评估，以衡量我们方法的有效性。此外，我们采用并扩展了 Multitrust 中的非目标攻击数据集，以评估在视觉干扰输入下的性能。大量实验和与基线方法（如 CLIP-PRM 和其他解码策略）的比较表明，我们的方法具有优越的性能。

5

限制条件 本研究的局限性主要体现在两个方面。首先，尽管我们的方法消除了对诸如基线检测器等额外组件的需求，但仍然依赖于 CLIP 模型来提供视觉-语言相似性先验。其次，获取局部感知先验需要额外的推理步骤，这会导致额外的计算和时间开销。这些局限性也突出了未来研究的潜在方向。

LPS: The image provided does not contain a chart. It appears to be a colorful illustration with various elements, including a butterfly, a character that resembles a mouse or a small creature, and a figure that could be a human or a fantastical being. The image has a vibrant, artistic style with a mix of warm and cool colors. There is no data or information that can be extracted from a chart as there is no chart present in the image.
CLIP PRM: The image appears to be a painting or illustration featuring a large, yellow, furry creature with a long snout and large eyes, resembling a troll or a similar fantastical being. It is standing on a rocky outcrop with a fiery background, suggesting a volcanic or fiery landscape. The creature seems to be observing the butterfly, and there is a small, green, humanoid figure with a round head and a pointed nose, possibly a troll or a similar creature, sitting on the ground to the right of the butterfly.

LPS: The image appears to be a painting depicting a scene with a group of people gathered around a large, ornate vase or urn. The vase is placed on a pedestal and is the central focus of the scene. The people are dressed in period clothing, suggesting the painting is set in a historical context. The background features a cityscape with buildings that have a classical architectural style, and there are trees and foliage in the foreground. The overall impression is that of a historical or classical scene, possibly from a past era.
CLIP PRM: The image depicts a lively scene of a public gathering in a park-like setting. In the center, there is a large, ornate fountain with a tall, slender spire rising above it. The fountain is surrounded by a crowd of people, some of whom are standing on the steps leading up to it, while others are seated on the surrounding grassy area. The crowd appears to be engaged in various activities, with some individuals standing and conversing, while others are seated, possibly enjoying the view or the music. The overall atmosphere is one of leisure and social interaction.

Figure 4: 处理对抗性扰动图像时，LPS 和 CLIP-PRM 方法的定性比较。结果显示了两种方法的完整响应，突出了 LPS 对视觉扰动的稳健性。

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. 2023. Rethinking model ensemble in transfer-based adversarial attacks. *arXiv preprint arXiv:2303.09105*.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*.

Yi Ding, Bolian Li, and Ruqi Zhang. 2024. Eta: Evaluating then aligning safety of vision language models at inference time. *arXiv preprint arXiv:2410.06625*.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*.

Yixiong Fang, Ziran Yang, Zhaorun Chen, Zhuokai Zhao, and Jiawei Zhou. 2024. From uncertainty to trust: Enhancing reliability in vision-language models with uncertainty-guided dropout decoding. *arXiv preprint arXiv:2412.06474*.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, and 1 others. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. 2024. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26876–26885.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Yangyang Guo, Fangkai Jiao, Liqiang Nie, and Mohan Kankanhalli. 2024. The vllm safety paradox: Dual ease in jailbreak attack and defense. *arXiv preprint arXiv:2411.08410*.

Devaansh Gupta, Siddhant Kharbanda, Jiawei Zhou, Wanhua Li, Hanspeter Pfister, and Donglai Wei. 2023. Cliptrans: transferring visual knowledge with pre-trained models for multimodal machine translation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2875–2886.

Ehsan Kamalloo, Nouha Dziri, Charles LA Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. *arXiv preprint arXiv:2305.06984*.

Junho Kim, Hyunjun Kim, Kim Yeonju, and Yong Man Ro. 2024. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems*, 37:133571–133599.

Yeonju Kim, Junho Kim, Byung-Kwan Lee, Sebin Shin, and Yong Man Ro. 2023. Mitigating dataset

bias in image captioning through clip confounder-free captioning network. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1720–1724. IEEE.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *Preprint*, arXiv:2311.16922.

Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. 2024. Common 7b language models already possess strong math capabilities. *arXiv preprint arXiv:2403.04706*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

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.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Zhixuan Liu, Zhanhui Zhou, Yuanfu Wang, Chao Yang, and Yu Qiao. 2024b. Inference-time language model alignment via integrated value guidance. *arXiv preprint arXiv:2409.17819*.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Renjie Pi, Tianyang Han, Jianshu Zhang, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. 2024. Mllm-protector: Ensuring mllm's safety without hurting performance. *arXiv preprint arXiv:2401.02906*.

Zhenting Qi, Hongyin Luo, Xuliang Huang, Zhuokai Zhao, Yibo Jiang, Xiangjun Fan, Himabindu Lakkaraju, and James Glass. 2024. Quantifying generalization complexity for large language models. *arXiv preprint arXiv:2410.01769*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.

Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. 2024. Toward self-improvement of llms via imagination, searching, and criticizing. *Advances in Neural Information Processing Systems*, 37:52723–52748.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Xiyao Wang, Zhengyuan Yang, Linjie Li, Hongjin Lu, Yuancheng Xu, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2024. Scaling inference-time search with vision value model for improved visual comprehension. *arXiv preprint arXiv:2412.03704*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. 2023. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints, pp. arXiv–2310*.

Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, and 1 others. 2024a. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2024b. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *arXiv preprint arXiv:2411.09968*.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, and 1 others. 2024c. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. *Advances in Neural Information Processing Systems*, 37:49279–49383.

Henry Hengyuan Zhao, Pan Zhou, Difei Gao, Zechen Bai, and Mike Zheng Shou. 2024. Lova3: Learning to visual question answering, asking and assessment. *Advances in Neural Information Processing Systems*, 37:115146–115175.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang,

Yun Li, Linjun Zhang, and Huaxiu Yao. 2024a. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*.

Zhanhui Zhou, Zhixuan Liu, Jie Liu, Zhichen Dong, Chao Yang, and Yu Qiao. 2024b. Weak-to-strong search: Align large language models via searching over small language models. *arXiv preprint arXiv:2405.19262*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A 实验设置的详细信息

A.1 幻觉率指标

（1）椅子。 带图像相关性的字幕幻觉评估（CHAIR） (Rohrbach et al., 2019) 提出了一个被广泛采用的幻觉评估指标。这个指标通过计算模型输出中提到的物体总数中实际不存在于图像中的参考物体的比例来评估幻觉。它包含两个变体：CHAIR- S ，在句子层面评估幻觉；CHAIR- I ，在物体实例层面操作。这两种表述提供了捕捉物体幻觉现象的互补视角：

$$\text{CHAIR}_{\text{I}} = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects}\}|}, \quad (7)$$

$$\text{CHAIR}_{\text{S}} = \frac{|\{\text{hallucinated responses}\}|}{|\{\text{all responses}\}|}, \quad (8)$$

，其中幻觉响应是指包含至少一个幻觉物体的响应。

为了评估模型生成的描述与人为撰写的描述之间的贴合度，我们采用了 BLEU 分数 (Papineni et al., 2002) ，这是一种用于评估文本相似性的标准指标。BLEU 量化了模型输出与参考描述之间的重叠程度，从而反映出模型捕捉到人类语言真实性以及符合预期描述标准的程度。

（2）POPE。 与之前的研究一致，我们的评估结合了基于投票的对象探查评估（POPE）方法 (Li et al., 2023) 。POPE 采用自动分割系统来识别并勾勒图像中的对象。然后，它向模型询问这些检测到的对象的存在，同时引入随机选择的虚构对象以评估错误阳性。由此产生的 F1 分数提供了模型准确感知和解释视觉内容能力的全面度量。

我们将 Multitrust 数据集中无目标攻击子集

（3）多重信任。 从 100 个扩展到 1,000 个图像，以便在视觉模态损坏的情况下进行更全面的评估。对于每个图像描述实例，我们检查模型的响应，以确定它是否提及图像中存在的任何真实对象。明确包含至少一个真实对象的响应被认为是准确的，而完全没有这种引用的情况会被视为幻觉。

在方程 5 中，我们使用两个超参数 a 和 b 来控制两个奖励组件 $R^t_{\text{LPS}}(i)$ 和 $R^t_{\text{CLIP}}(i)$ 之间的平衡。在我们的实验中，两个超参数都设置为 $a = 1$ 和 $b = 1$ 。此外，我们在每个解码步骤生成 6 个候选标记，并将最大搜索步骤设置为 10。