CRoPS：一种用于视觉语言模型的免训练幻觉缓解框架

Neeraj Anand[1]*，**Samyak Jha**[1]，**Udbhav Bamba**[2]，**Rahul Rahaman**[3]1印度理工学院（*ISM*），丹巴德，印度2*Transmute AI*3新加坡国立大学

在 OpenReview 上评审： https://openreview.net/forum?id=KQSoZDPVGX

arXiv:2601.00659v1 [cs.CV] 2 Jan 2026

摘要

尽管大型视觉语言模型（大型视觉语言模型）取得了迅速的成功，但其倾向于生成幻觉内容这一问题仍然存在，削弱了其在现实应用中的可靠性。现有的免训练方法能够缓解幻觉，但面临两项限制：（i）依赖于对幻觉来源的狭隘假设；（ii）随着生成接近末尾，其有效性会下降，而幻觉最有可能在这一阶段出现。一种常见策略是，通过完全或部分移除视觉词元来构建幻觉模型，并将其与原始模型进行对比。然而，仅凭这一点仍然不足，因为视觉信息仍会传播到生成文本中。基于这一洞见，我们提出了一种新颖的幻觉模型，通过有选择地移除关键文本词元来捕捉幻觉效应。我们进一步引入广义对比解码，通过整合多个幻觉模型来表示多样的幻觉来源。这些理念共同构成**CRoPS**，一种免训练幻觉缓解框架，使 CHAIR分数提高了20%，并在六个基准测试和三个大型视觉语言模型系列上持续取得提升，优于最先进的免训练方法。[1]

1 引言

近年来，大型视觉语言模型（LVLMs）取得了显著进展，展现出卓越的多模态能力（Liu 等人，2023c；Dai 等人，2023）。它们理解图像和语言的能力，促成了广泛应用。然而，与大语言模型（LLMs）类似，大型视觉语言模型容易产生“幻觉”，给出听起来令人信服但不够准确的回答，从而可能传播误导性信息（Rani 等人，2024年；Li 等人，2016年）。这一局限性对它们在现实世界应用中作为可信赖人工智能助手的可靠性构成了重大挑战（Wang 等人，2023b；Liu 等人，2023a）。

近期旨在缓解幻觉的工作可分为两类：基于训练的方法（Biten 等人，2022年；Kim 等人，2023；Rohrbach 等，2018），以及免训练方法（Wei 等人，2022年；Li等人，2023年；Leng 等人，2024年；Wang 等人，2024年b；Favero 等人，2024年）。在本研究中，我们关注一类非常具体的免训练方法，这类方法采用对比解码（CD）的通用框架（Li等人，2023年）。CD 是一种免训练方法，其提出通过对比（相减）具有严重幻觉的模型的输出，来消除 大型语言模型输出中的幻觉；本质上，这一问题被归结为识别并使用幻觉模型进行对比。这些方法通常存在若干缺陷。首先，从设计上看，用于对比的幻觉模型随着输出标记的生成而逐渐无法保持良好性能（图3），因此主要在生成早期阶段有效。此外，单个幻觉模型无法充分代表幻觉的所有来源，例如由信息量不足的视觉标记导致的幻觉，以及训练数据产生的偏差所造成的幻觉。

*通讯作者：neerajanandfirst@gmail.com[1]代码可在此处获取：https://github.com/ubamba98/CRoPS-Mitigate-Hallucinations-in-Vision-Language-Models

在本文中，我们首先分析现有方法（Favero 等人，2024年；Huo 等人，2024），并通过实证分析与理论计算展示现有方法如何在大型视觉语言模型生成更多词元时，难以消除幻觉。随后，我们提出一种新型幻觉模型，通过不再局限于移除全部或部分视觉词元，并将文本输入纳入移除范围，克服了这一挑战。我们提出的模型受到分析结果的启发：在生成的后期阶段，先前生成的输出标记与视觉词元同等重要，甚至更为重要。

接下来，我们通过实证展示现有对比解码方法在解决不同类型的幻觉来源时为何会失效。为缓解这一问题，我们将对比解码的概念加以泛化，并将其扩展为可与多个模型而非单个模型进行对比。我们提出的广义对比解码使我们能够将新提出的幻觉模型与 Huo 等人（2024年）的现有工作结合起来，从而设计出一种新颖的、无需训练的幻觉缓解方法 CRoPS（图 1）。我们在多种生成式和判别式基准测试上评估并比较了所提出的 CRoPS 与其他竞争方法。我们还展示了 CRoPS 在三种不同架构选择下均优于所有其他方法。

贡献： 下面概述我们的主要贡献。在本工作中，

- 我们详细分析了现有方法在大型视觉语言模型词元生成后期表现不佳的原因。这也使我们能够设计一种新颖的幻觉模型：不再局限于视觉词元，而是识别并移除提示和过去生成的词元中的文本输入。
- 我们通过实证表明，各种方法只能应对某一种特定的幻觉来源，却无法消除其他来源。为此，我们提出广义对比解码，将对比解码的适用范围扩展至允许对多个模型进行对比。
- 最后，我们在全新的广义对比解码框架下，将新提出的文本缺失幻觉模型与图像缺失幻觉模型结合起来，并提出 CRoPS。我们提出的方法显著优于基线，并在广泛的任务、数据集和大型视觉语言模型架构上，持续优于其他竞争方法。

2 相关工作

随着大语言模型的发展，近期研究通过将视觉编码器集成到预训练的大语言模型中，探索大型视觉语言模型。这些模型展现出了一些先进的多模态能力。然而，它们仍受到幻觉问题的困扰（Rohrbach 等，2018；Zhang 等人，2024年；Guan 等人，2024年；Wu 等人，2024b），限制了其在现实世界中的应用（Wang 等人，2023b；Liu 等人，2023a）。幻觉的成因有多种，包括缺乏对世界知识的理解、对特定训练数据模式的过拟合，以及常识推理能力不足。在大语言模型中，当生成的回答与现实世界知识或常识相矛盾时，通常会出现幻觉。相比之下，对于视觉语言模型，主要关注的是生成的回答是否与所提供的图像相冲突。

研究人员探索了多种缓解这一问题的方法，大致可分为以下两类：

基于训练的方法。 基于训练的方法包括使用精心整理的数据集重新训练视觉语言模型，或使用辅助模型来监督或修订生成结果。这些方法包括在针对幻觉设计的数据集上进行指令微调（Lee et al., 2022; Gunjal et al., 2024; Zhao 等人, 2024; Jiang et al., 2024; Yu et al., 2024; Yue et al., 2024），以及事后训练修订网络，由其分析模型输出，并借助辅助网络纠正幻觉（Manakul et al., 2023; Zhou et al., 2024; Yin 等人, 2024; Chen et al., 2024; Wu et al., 2024a; Feng et al., 2024）。虽然这些技术可能很有效，但需要大量计算资源，并且需要谨慎设计数据集。

免训练方法。相比之下，免训练方法会修改推理时的解码或注意力模式，而无需额外训练或监督。这些方法包括解码时修改

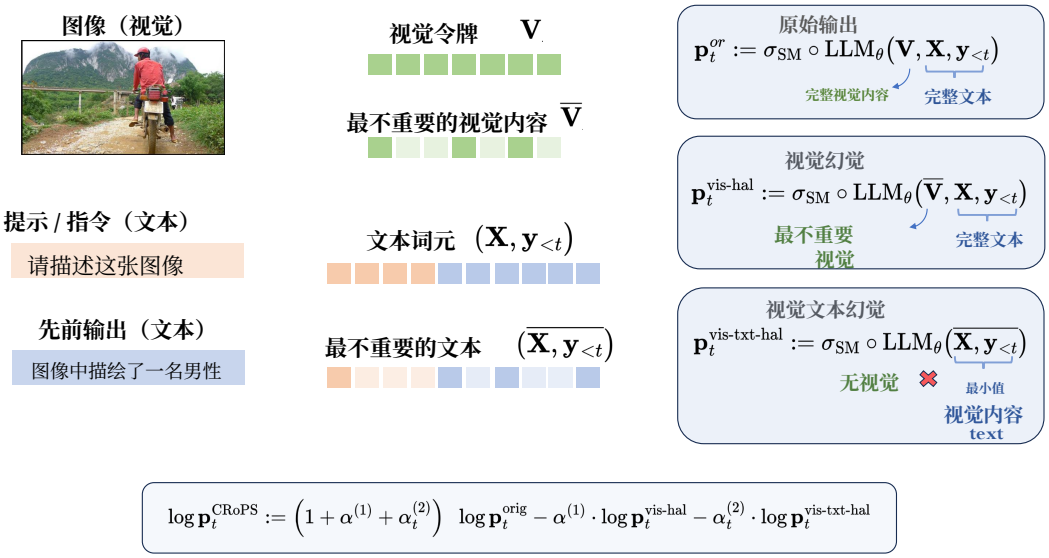


图1：**CRoPS 框架概览**。CRoPS 结合了两个幻觉模型：一个移除视觉词元以捕捉视觉相关幻觉，另一个移除关键文本词元以捕捉文本相关幻觉。随后，通过广义对比解码框架整合它们的输出，以减少大型视觉语言模型中的幻觉。

（Huo 等人，2024；Wang 等人，2023b；Favero 等人，2024；Li等人，2023年；Chuang 等人，2024；Liu 等人，2024b；Leng 等人，2024；Wang 等人，2024b；Kim 等人，2024b；Zhu 等人，2024；Huang 等人，2024；Kim 等人，2024a；Woo 等人，2024；Zheng 等人，2024；Li 等人，2025年）以及注意力调整机制（Tang 等人，2025年；Yin 等人，2025年）。这些方法是简单的即插即用模块，并且避免了基于训练的方法所带来的计算开销。

与我们的工作密切相关的是多模态互信息解码（M3ID）（Favero 等人，2024年）和自我内省解码（SID）（Huo 等人，2024年），它们也属于免训练方法。M3ID 提出利用并调整同一个大语言模型，以构建一个幻觉模型。他们认为，如果从大语言模型中移除图像输入，模型就会生成与视觉内容完全无关的任意文本，而仅依赖训练期间习得的偏差。另一方面，SID 并不从大型视觉语言模型输入中移除所有视觉词元，而是有选择地保留最不重要的视觉内容，即那些与查询词元连贯性较低的视觉词元。作者认为，这些低重要性词元对幻觉的贡献最大。这些方法构成了我们对比幻觉框架的基础，我们将在第4节进一步探讨该框架（形式化定义见附录A和B）。

与以往分别处理视觉令牌稀释和预训练期间引入的偏差的方法不同，*CRoPS* 能够有效缓解这些问题，且无需微调或辅助模型。

3 背景

本节将对大型视觉语言模型的推理与解码过程进行形式化描述，因为这些细节决定了视觉信息如何传播到生成的词元中（第3.1节）。随后，我们将介绍用于减少大型视觉语言模型中幻觉的对比解码（CD）框架（第3.2节），以及用于构建幻觉模型的基于注意力的词元剪枝方法（第3.3节）。

3.1 大型视觉语言模型中的推理与解码。

大型视觉语言模型接受视觉和文本输入，以生成文本输出。模型首先使用视觉编码器和跨模态投影模块，将输入图像编码为一系列视觉词元。我们将这些视觉词元表示为 $\boldsymbol{V} := (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m)$，其中 m 是视觉词元的总数。此外，模型还会接收文本提示 $\boldsymbol{X}$，它也可以表示为一系列 n 词元$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$。参数化为 θ的大语言模型随后会依次生成文本输出词元。设 $\boldsymbol{y}_{<t} := (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{t-1})$ 为截至推理时间步 t已生成词元的序列。推理过程可形式化表示如下：

$$\begin{aligned} \boldsymbol{p}_t &= \mathrm{softmax} \circ \mathrm{LLM}_\theta(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{y}_{<t}) \\ \boldsymbol{y}_t &= \mathrm{Decode}[\boldsymbol{p}_t] \\ \boldsymbol{Y}_{<t+1} &= [\boldsymbol{y}_{<t} : \boldsymbol{y}_t] \end{aligned} \tag{1}$$

首先，语言模型输出逻辑值，记为 $\mathrm{LLM}_\theta(\cdot)$，然后通过应用 softmax 函数将其转换为概率向量 $\boldsymbol{p}_t$ 。生成的 $\boldsymbol{p}_t$ 是一个形状为 $|\mathcal{V}|$的概率向量，其中$\mathcal{V}$表示 LLM 的词汇表。接下来，应用解码函数将 $\boldsymbol{p}_t$ 转换为单个词元$\boldsymbol{y}_t$。为完成这一转换，可以使用多种可用的解码策略，例如贪心解码、束搜索和核采样。随后，新生成的输出词元 $\boldsymbol{y}_t$ 会被追加到先前词元中，作为输入提供给 LLM，以生成下一个词元。

3.2 幻觉与对比解码

当大型视觉语言模型生成的序列 $\boldsymbol{Y}$ 与视觉输入 $\boldsymbol{V}$完全或部分不一致时，我们称该模型产生了幻觉。如第 2 节所述，已有多种免训练方法被提出以缓解幻觉。在本文中，我们重点关注一类共享同一框架的方法，称为对比解码（*CD*）（Li等人，2023年）。CD 框架通过抑制分配给幻觉词元/对象的概率，试图消除大型视觉语言模型中的幻觉。具体而言，它首先创建或识别一个幻觉模型，然后从原始大型视觉语言模型的逻辑值中减去幻觉模型的逻辑值，从而降低分配给幻觉词元的概率。形式上，在 CD 框架下，最终的概率输出计算如下：

$$\log \boldsymbol{p}_t = (1+\alpha) \cdot \log \boldsymbol{p}_t^{orig} - \alpha \cdot \log \boldsymbol{p}_t^{hal}, \tag{2}$$

其中，$\alpha > 0$, $\boldsymbol{p}_t^{orig}$ 是根据公式 1，使用完整输入计算得出的原始大型视觉语言模型的概率输出，而 $\boldsymbol{p}_t^{hal}$是幻觉模型的概率输出。由于对比解码（CD）依赖于定义一个幻觉模型，因此模型的选择仍然至关重要。

3.3 基于注意力的词元剪枝

在推理时间 t，我们向大型视觉语言模型输入元组 $(\boldsymbol{V},\ \boldsymbol{X},\ \boldsymbol{y}_{<t})$，其中包括视觉词元 $\boldsymbol{V}$、文本词元 $\boldsymbol{X}$以及大语言模型迄今生成的输出 $\boldsymbol{y}_{<t}$。为了估计词元重要性，我们利用语言模型中选定的 Transformer 层 l 所产生的注意力权重。令$\mathbf{K} \in \{\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{y}_{<t}\}$ 表示键的集合，$\boldsymbol{y}_t$ 表示当前查询词元。令$\mathbf{K} \in \{\boldsymbol{V},\ \boldsymbol{X},\ \boldsymbol{y}_{<t}\}$ 表示键词元的集合，$\boldsymbol{y}_t$ 表示当前查询词元。对于具有 H 个注意力头的多头注意力机制，我们将每个键词元的重要性得分定义为所有注意力头上的平均注意力权重

$$\psi(\boldsymbol{y}_t) = \frac{1}{H} \sum_{h=1}^{H} \mathrm{Attention}^{(l,h)}(\mathbf{K}, \boldsymbol{y}_t), \tag{3}$$

其中，$\mathrm{Attention}^{(l,h)}(\mathbf{K}\ \boldsymbol{y}_t)$ 表示来自第 l 层第 h^{th} 个注意力头的注意力分布，用于衡量查询词元 $\boldsymbol{y}_t$ 对$\mathbf{K}$中各键的关注程度。

基于 $\psi(\boldsymbol{y}_t)$，我们找出重要性值最低、最少受到当前查询关注的词元。利用这一分数，我们可以选出底部 $\bar{u}(< u = |\mathbf{K}|)$ 个分数最低的词元，并

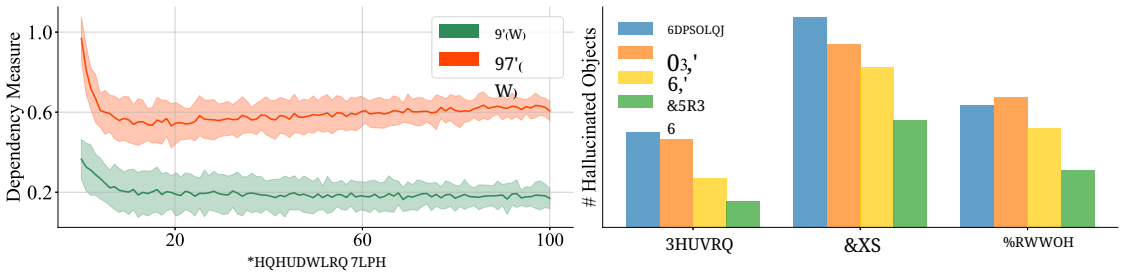


图2：**左：依赖性度量曲线图**（见第4.1节），该度量量化了视觉词元和视觉+文本词元对大型视觉语言模型生成的影响。我们观察到 VD(t) 随时间推移而降低，表明随着解码过程的推进，模型对视觉令牌的依赖减少。**右：幻觉对象的出现频率**这些对象经常与真实对象“餐桌”同时出现。我们观察到，SID 和$CRoPS$能有效缓解统计偏差，而 M3ID 的表现不尽如人意。

移除高重要性标记，从而创建一个更稀疏的词元集合。这一设计的依据是 SID（Huo 等人，2024年）的先前研究发现：低重要性标记更可能引发幻觉。

4 动机

本节重点介绍现有基于对比解码的幻觉缓解方法的局限性，这些局限性促使我们设计所提出的框架。

4.1 缺点 I：对视觉词元的依赖逐渐减弱

为了支持下文的讨论，我们首先定义一个度量，用于捕捉大语言模型生成行为对输入的依赖如何随生成时间变化 t。具体而言，我们扩展 Favero 等人 2024 提出的 提示依赖度量，定义两个新指标：视觉依赖性和视文依赖度，分别如方程4和方程8所示。视觉依赖性 定义如下：

$$\mathrm{VD}(t) := \mathrm{dist}\big(\mathrm{softmax} \circ \mathrm{LLM}_\theta\big(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{y}_{<t}\big),\ \mathrm{softmax} \circ \mathrm{LLM}_\theta\big(\boldsymbol{X}, \boldsymbol{y}_{<t}\big)\big) \tag{4}$$

其中，$\mathrm{dist}(P, Q) = \frac{1}{\sqrt{2}} \left\| \sqrt{P} - \sqrt{Q} \right\|_2$ 表示海林格距离。VD(t) 衡量省略视觉输入词元后生成的输出词元分布发生的变化。图2左侧子图绘制了 VD(t) 随生成步骤 t 变化的情况。显然，随着生成更多词元，VD(t) 会急剧下降；也就是说，在生成后期，忽略图像所导致的分布偏移较小。

这种依赖性减弱对对比解码方法具有重要影响，因为它直接影响生成后期所使用的对比信号的有效性。由于 Favero 等人 2024 将其幻觉模型定义为 $\boldsymbol{p}_t^{hal} :=$ 大语言模型$_\theta\big(\boldsymbol{X},\ \boldsymbol{y}_{<t}\big)$，因此，VD($t$) 的减小表明，随着时间推移，$\boldsymbol{p}_t^{orig}$ 与 $\boldsymbol{p}_t^{hal}$之间的相似性不断增加（推导见附录 E）。因此，在生成后期，对比信号的意义变得不那么显著，使幻觉模型的效果降低（如图 3 所示，后期生成结果表现出更严重的幻觉）。类似地，当 VD(t) 本身较低时，Huo 等人 2024 的方法中对图像词元进行掩码的效果也会减弱。

4.2 补救措施 I：一个用于对比的新模型

随着 VD(t) 逐渐减小，表明视觉词元的冗余度不断增加，后期生成主要依赖文本上下文。因此，部分或完全丢弃视觉词元的幻觉模型在后期会失去效用。为缓解这一问题，我们提出移除重要的文本词元，

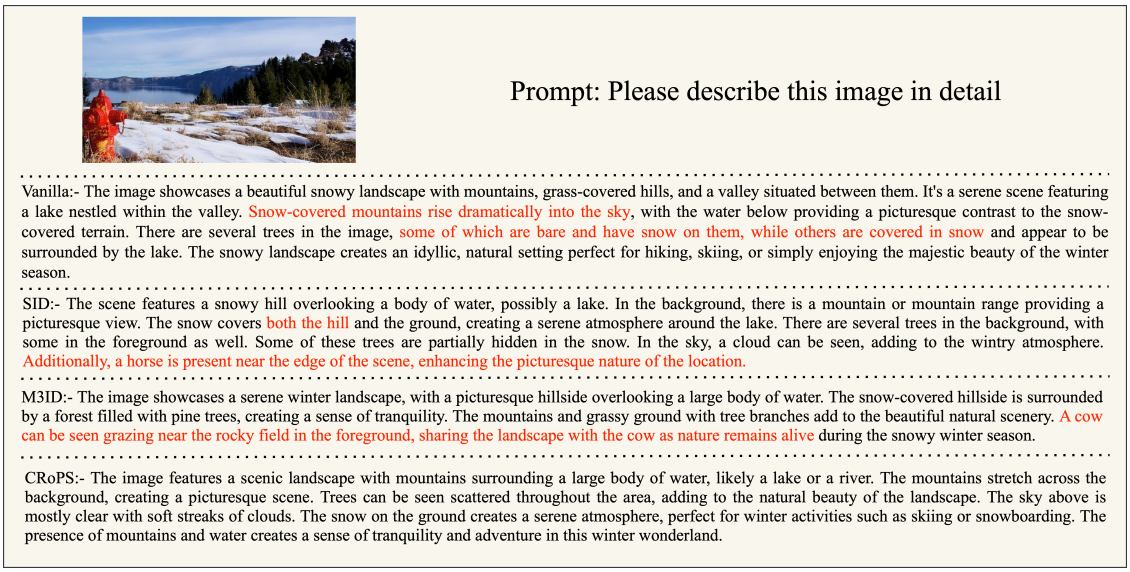


图 3：不同方法生成的图像描述对比。Vanilla、SID 和 M3ID 包含幻觉细节（以红色标出），例如动物和夸大的积雪覆盖。相比之下，CRoPS 生成的描述忠实于图像，不含这些幻觉。请注意，随着生成进入后期阶段，幻觉出现得愈加频繁。

其中，重要性分数根据模型早期层 l 的注意力权重计算得出，这与 SID（Huo 等人，2024年）针对视觉词元采用的策略类似。

$$\psi(\boldsymbol{y}_t) = \frac{1}{H}\sum_{h=1}^{H} \text{Attention}^{(l,h)}\big([\boldsymbol{X}, \boldsymbol{y}_{<t}], \boldsymbol{y}_t\big) \tag{5}$$

利用这些重要性分数，我们仅保留最不重要的文本词元（提示和先前词元合并后）的词元，并将元组（$\boldsymbol{X}$，$\boldsymbol{y}_{<t}$）转换为，

$$\overline{\boldsymbol{X}, \boldsymbol{y}_{<t}} := \text{LeastImp}\Big[\big(\boldsymbol{X}, \boldsymbol{y}_{<t}\big),\, \eta(\mu, t)\Big]. \tag{6}$$

通过保留 $\eta(\mu,\ t)$ 个文本词元，其中 $\eta(\mu,\ t) = \beta_0 + \beta_1\big(1 - e^{-\mu t}\big)$ 是关于 t 的非递减函数。由于文本词元的数量会随时间（先前生成的词元）增加，为了保持所保留的不连贯词元的稀疏性，$\eta(\mu,\ t)$ 需要是非递减函数。详见第 7 节。

因此，我们提出的幻觉模型形式如下，

$$\boldsymbol{p}_t^{\text{vis-txt-hal}} := \text{softmax} \circ \text{LLM}_\theta\big(\overline{\boldsymbol{X}, \boldsymbol{y}_{<t}}\big), \tag{7}$$

这意味着我们不仅完全移除了视觉词元，还移除了一些重要的文本词元。为了检查所提出的幻觉模型是否摆脱了依赖性减弱的问题，我们计算视文依赖度 VTD(t)，如下所示，

$$\text{VTD}(t) := \text{dist}\big(\text{softmax} \circ \text{LLM}_\theta\big(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{y}_{<t}\big),\, \text{softmax} \circ \text{LLM}_\theta\big(\overline{\boldsymbol{X}, \boldsymbol{y}_{<t}}\big)\big) \tag{8}$$

并与图2左侧子图中的 VD(t) 进行比较。显然，与 VD(t) 不同，VTD(t) 不会随着时间推移而减小。这意味着，所提出的幻觉模型与原始输出存在显著差异，因此，即使在生成后期，对比仍然有效。

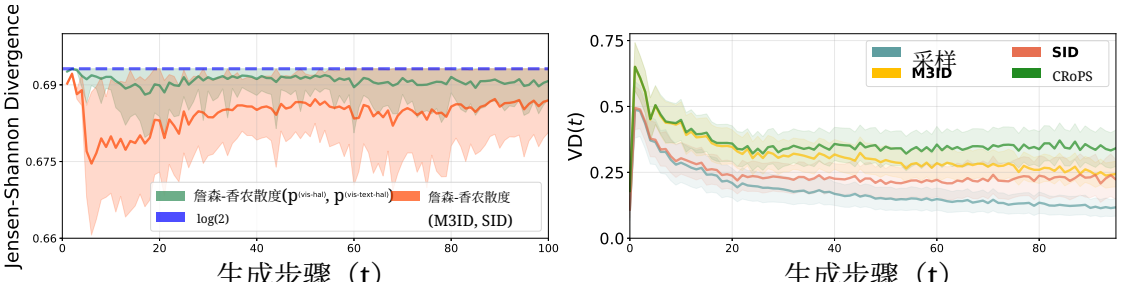


图4：**左：不同幻觉模型在生成时间内的 Jensen-Shannon（JS）散度曲线**。蓝色虚线表示 log(2)，即可能达到的最大散度值。**右：不同方法（采样、SID、M3ID 和 CRoPS）最终输出的视觉依赖性曲线**。

4.3 缺陷II：与单个幻觉模型进行对比是不够的

对比解码方法通常依赖幻觉模型，这些模型会忽略或扰动输入的部分内容，以暴露幻觉。这些模型旨在捕捉不同的失效模式，例如过度依赖先前生成的词元，或关注无关的视觉特征。尽管此类方法展现出良好的效果，但仅处理其中一种幻觉来源并不足够。

例如，M3ID 完全移除视觉输入，从而针对由过度依赖文本词元或语言先验所引发的幻觉。然而，它无法缓解由误导性或虚假视觉线索触发的幻觉。相比之下，SID 使用低重要性视觉词元构建幻觉模型，并将其与原始模型进行对比，以抑制由无关视觉线索导致的幻觉。然而，当文本依赖占主导地位、幻觉在生成后期出现时，它仍然无法有效应对。

正如第4.1节所述，并如图2（左）中不断下降的视觉依赖性 VD(t) 所示，视觉词元的影响会随时间减弱，从而降低了仅修改视觉输入的方法的作用。此外，图2（右）显示，M3ID 仍然会受到与真实物体“餐桌”同时出现的共现物体对（例如“人物”“杯子”“瓶子”）相关的幻觉影响，而这些幻觉源于训练时相关性。综合这些观察可以看出，无论是面向视觉还是面向文本，与单个幻觉模型进行对比都不足以全面解决生成各阶段的幻觉问题。

4.4 补救方法 II：对比解码的一般化形式

公式 2 中的标准对比解码形式，只允许将原始输出与一个幻觉模型进行对比。然而，正如第 4.3 节所述，单个幻觉模型可能不足以捕捉不同来源的幻觉，因此，对比后得到的模型仍会受到尚未解决的幻觉模式的影响。

我们提出了一种一般化的对比解码形式，可与多个幻觉模型进行对比。我们提出的一般化形式如下：

$$\log \boldsymbol{p}_t^{st} = (1 + \sum_{r=1}^{R} \alpha^{(r)}) \cdot \log \boldsymbol{p}_t^{orig} - \sum_{r=1}^{R} \alpha^{(r)} \cdot \log \boldsymbol{p}_t^{hal(r)}, \tag{9}$$

其中 $\alpha_r > 0 \ \forall r$。这一一般化形式使我们能够利用多个幻觉模型来代表不同来源的幻觉。

局限性与假设。参与的幻觉模型 $\boldsymbol{p}_t^{hal(r)}$ 应有足够差异，才能证明将它们纳入其中是合理的。在广义形式下加入相似的幻觉模型，结果将与传统的对比解码相同，但延迟可能更高。

5 CRoPS

我们提出的 CRoPS 在广义对比解码公式下整合了两个幻觉模型，这些模型

$$\begin{aligned} \boldsymbol{p}_t^{\text{vis-hal}} &:= \text{softmax} \circ \text{LLM}_\theta\big(\overline{\boldsymbol{V}}, \boldsymbol{X}, \boldsymbol{y}_{<t}\big), \\ \boldsymbol{p}_t^{\text{vis-txt-hal}} &:= \text{softmax} \circ \text{LLM}_\theta\big(\overline{\boldsymbol{X}}, \boldsymbol{y}_{<t}\big). \end{aligned} \tag{10}$$

第一个模型 $\boldsymbol{p}_t^{\text{vis-hal}}$受到 Huo 等人 2024年工作的启发，通过移除重要的视觉词元（第3.3节）来捕获共现相关幻觉（图2右侧子图），但 保留所有文本词元。第二个模型 $\boldsymbol{p}_t^{\text{vis-txt-hal}}$ （如第4.2节所述）则缺失所有视觉信息和重要的文本信息。在生成的早期阶段，生成高度依赖视觉词元，此时 $\boldsymbol{p}_t^{\text{vis-hal}}$代表幻觉的来源；而在视觉依赖性减弱的后期阶段， $\boldsymbol{p}_t^{\text{vis-txt-hal}}$ 通过剥夺模型的重要文本词元来产生幻觉输出。将二者结合后，所得的对比解码被我们称为 CRoPS，可表示为公式9，其组成部分定义如下:

$$\begin{aligned} \log \boldsymbol{p}_t^{\text{CRoPS}} := &\Big(1 + \alpha^{(1)} + \alpha_t^{(2)}\Big) \cdot \log \boldsymbol{p}_t^{\text{orig}} \\ &- \alpha^{(1)} \cdot \log \boldsymbol{p}_t^{\text{vis-hal}} - \alpha_t^{(2)} \cdot \log \boldsymbol{p}_t^{\text{vis-txt-hal}} \end{aligned} \tag{11}$$

其中， $\alpha^{(1)} \equiv \alpha$ 和 $\alpha_t^{(2)} := \left(1 - e^{-\gamma t}\right) / e^{-\gamma t}$。由于视觉驱动幻觉主要发生在生成的早期阶段，因此 $\alpha^{(1)}$ 保持常数。相比之下， $\alpha_t^{(2)}$ 被设计为随时间逐渐增加，以反映后期文本诱发幻觉日益增强的影响。与以往工作类似，我们加入 置信度 和合理性 约束，并通过公式11进行对比。整体算法概述见附录。

和 $\boldsymbol{p}_t^{\text{vis-hal}}$与 $\boldsymbol{p}_t^{\text{vis-txt-hal}}$ 的独特性。采用多个幻觉模型的广义对比解码是否合理，完全取决于参与模型的独特程度。图4左侧子图展示了我们提出的 $\boldsymbol{p}_t^{\text{vis-hal}}$和 $\boldsymbol{p}_t^{\text{vis-txt-hal}}$ 之间的詹森-香农散度（JSD）。我们将其与最大值（log2）以及先前工作所用幻觉模型之间对应的JSD进行了比较，结果表明，在广义对比解码中简单叠加先前工作中的幻觉模型并非最优方案。

最终输出的视觉依赖性。图4右侧子图展示了各竞争方法和 $\boldsymbol{p}_t^{\text{CRoPS}}$最终输出的 VD(t)。显然，与原始采样及其他方法相比，CRoPS 的视觉接地程度更高（即产生的幻觉更少）。

6 实验

6.1 实验设置

模型与基线方法。我们在三种广泛采用的大型视觉语言模型上进行评估：LLaVA-1.5（Liu 等人，2023b）、LLaVA-NeXT（Liu 等人，2024a）和 Qwen2-VL（Wang 等人，2024a）。在基线比较中，我们考察了近期提出的几种免训练幻觉缓解技术，包括 VCD（Leng 等人，2024）、ICD（Wang 等人，2024b）、OPERA（Huang 等人，2024）、ClearSight（Yin 等人，2025）、SID（Huo 等人，2024）和 M3ID（Favero 等人，2024）。这些方法代表了当前最先进的策略，旨在减少幻觉，同时无需对底层视觉语言模型进行额外微调或重新训练。

实现细节。 对于文本词元和视觉词元，我们均将剪枝层设置为 $l = 2$ 。我们遵循 Huo 等人，2024 的选择，因为他们证明，初始层中的注意力分数能够更好地反映词元重要性，并且相对不受注意力汇的影响。我们的实验采用了 LLaVA-1.5 的 7B 和 13B 骨干网络、Qwen2-VL 的 7B 骨干网络，以及 LLaVA-NeXT 的 8B 骨干网络。我们采用核采样，top-p 值设为 0.9，温度设为 1。所有实验均使用三个

表1：在 MS-COCO 验证集上对视觉-语言对齐的评估（Lin等人，2014年）。使用提示 “请详细描述这张图像”生成图注。CHAIR分数 C_S 和 C_I分别表示幻觉对象和图注的百分比，数值越低表示视觉定位能力越强。召回率表示生成的图注中被正确提及的标注对象所占的百分比。所有结果均取三个随机种子的平均值。

方法	LLaVA-1.5 (7B)			LLaVA-1.5（13B）			LLaVA-NeXT			Qwen2-VL		
	$C_S \downarrow$	$C_I \downarrow$	召回率 ↑	$C_S \downarrow$	$C_I \downarrow$	召回率 ↑	$C_S \downarrow$	$C_I \downarrow$	召回率 ↑	$C_S \downarrow$	$C_I \downarrow$	召回率 ↑
采样	57.0	17.0	75.0	50.2	13.7	76.4	37.4	8.9	66.3	33.2	8.0	68.1
ClearSight	54.1	16.2	74.3	49.4	13.9	74.8	35	8.5	63.6	13.5	8.7	38.2
VCD	53.3	15.3	77.9	49.5	13.7	77.7	36.4	8.8	68.6	29.0	7.8	67.3
ICD	52.5	14.6	77.7	49.2	13.9	78.1	36.6	9.4	67.1	28.0	7.9	66.1
OPERA	49.1	13.8	78.5	48.2	13.2	78.9	35.5	8.9	66.9	31.0	8.1	67.9
SID	48.9	13.0	77.9	47.0	12.3	77.9	37.0	10.7	70.2	30.6	8.1	66.1
M3ID	47.1	12.8	74.8	45.5	12.2	75.3	36.1	9.8	68.7	28.8	7.3	64.8
CRoPS	**39.5**	**10.2**	76.3	**38.5**	**9.1**	75.1	**33.2**	**8.1**	66.2	**26.9**	**6.9**	67.4

表 2：在 AMBER基准（Wang 等人，2023a）上的性能，使用提示“请详细描述这张图像”进行评估。我们报告了幻觉的三个维度：**HAL**（总体幻觉率）、**Cog**（偏离正确属性/关系的认知偏差）和**CHAIR**（对象级幻觉），数值越低越好。结果取三个随机种子的平均值。

方法	LLaVA-1.5 (7B)			LLaVA-1.5（13B）			LLaVA-NeXT			Qwen2-VL		
	CHAIR ↓	HAL ↓	Cog ↓	CHAIR ↓	HAL ↓	Cog ↓	CHAIR ↓	HAL ↓	Cog ↓	CHAIR ↓	HAL ↓	Cog ↓
采样	10.6	44.3	4.0	9.3	41.3	4.2	10.5	57.1	4.1	6.4	38.9	2.8
ClearSight	10.5	44.1	4.2	9.2	40.5	4.0	9.1	54.4	4.5	5.9	34.0	2.2
VCD	9.0	42.9	4.6	8.4	38.3	3.9	9.5	55.7	4.2	5.7	33.9	2.4
ICD	10.0	44.8	4.3	8.5	39.5	4.1	10.1	53.0	4.5	6.0	34.1	2.5
OPERA	9.8	43.0	4.5	8.7	40.0	4.1	9.8	51.0	4.3	6.3	35.0	2.6
SID	9.3	43.7	3.7	6.9	35.0	3.5	9.1	54.2	3.9	5.4	30.6	1.8
M3ID	9.0	40.0	3.0	7.9	40.0	2.9	8.7	51.9	3.1	5.5	27.9	1.5
CRoPS	**6.3**	**29.3**	**2.8**	**5.7**	**27.8**	**2.5**	**7.2**	**44.6**	**2.6**	**5.1**	**24.2**	**1.1**

随机种子进行，并报告这些运行结果的平均性能。超参数配置及其消融分析见第 7 节。

6.2 实验结果

在本节中，我们将在一系列基准测试上评估**CRoPS**，这些基准测试涵盖了不同形式的幻觉和多模态推理。对于每项基准测试，我们会简要介绍其评测内容，然后讨论相应的结果。

CHAIR 基准测试。 CHAIR（图像相关性字幕幻觉评估）基准（Rohrbach 等，2018）通过比较生成的图注中的名词与真实对象标注来评估对象级幻觉，其中较低的 C_S 和 C_I 表示更强的视觉定位。如表 1 所示，**CRoPS** 在所有大型视觉语言模型上都取得了最低的 CHAIR分数，明显优于先前无需训练的方法。在 LLaVA-1.5 系列中，与 M3ID 相比，CRoPS 将幻觉率降低了约**15–25%**，同时保持了相近的召回率。在 LLaVA-NeXT 和 Qwen2-VL 上也呈现出相同趋势：CRoPS 进一步将 C_S 和 C_I 降低了**8–10%**，且描述质量没有任何下降。总体而言，这些在不同架构上始终如一的百分比提升表明，CRoPS 无需重新训练或辅助监督即可有效缓解幻觉。尽管 ClearSight 在 Qwen2-VL 上取得了更低的 C_S，但这是以牺牲召回率为代价的（38.2 对 67.4），这表明它过度抑制了视觉细节。相比之下，CRoPS 在定位准确率与语言丰富度之间保持了均衡，生成的图像描述既忠实于视觉内容，又细节丰富。

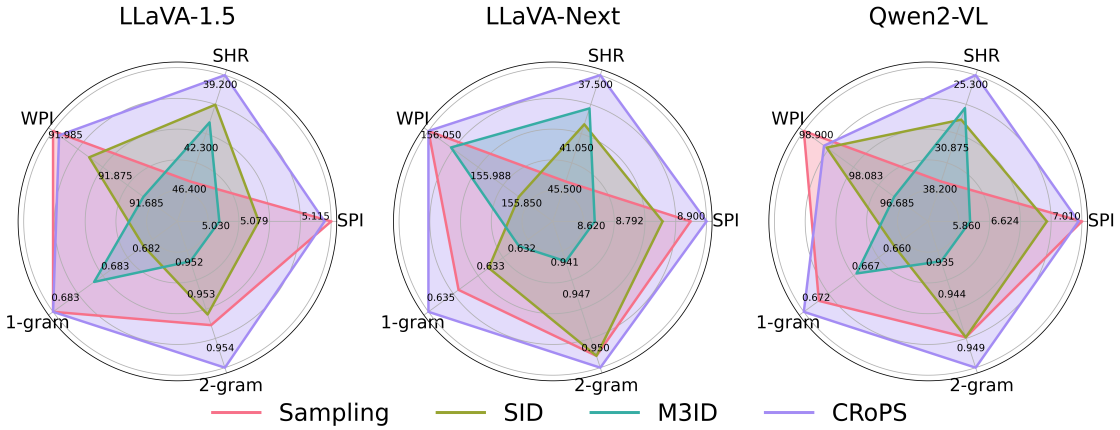


图 5：在 GPT-4o 辅助基准（Zhou 等，2024年）上的评估，对比幻觉（SHR）、流畅度（一元和二元语法精确率）以及描述细节（WPI 和 SPI）。封闭区域越大，整体表现越好。请放大查看，以便更清晰地观察。

表 3：POPE 视觉问答幻觉基准上的评估结果（Li等人，2023年）。POPE 包含三个子集：随机、热门和对抗。每个样本均采用以下查询模板：“图像中是否存在物体？”，其中物体 可以是随机选取的（随机），从出现频率最高的数据集对象中选取的（热门），也可以是从经常与目标实体共同出现的对象中选取的（对抗）。我们报告三个子集的平均表现；详细结果见附录 G。

方法	LLaVA-1.5		LLaVA-NeXT		Qwen2-VL	
	准确率 ↑	F1 ↑	准确率 ↑	F1 ↑	准确率 ↑	F1 ↑
采样	81.7	82.8	87.3	87.6	84.2	82.8
ClearSight	82.0	83.0	87.3	87.7	73.5	74.5
VCD	82.6	83.3	87.7	88.1	84.7	83.2
ICD	82.2	82.9	86.6	87.4	85.2	84.3
OPERA	82.3	82.9	87.4	<u>88.2</u>	85.5	84.5
SID	<u>82.7</u>	83.3	<u>88.6</u>	88.4	<u>85.6</u>	83.9
M3ID	82.4	<u>83.4</u>	88.0	88.1	84.9	<u>83.5</u>
CRoPS	**83.9**	**84.6**	**89.4**	**89.4**	**86.1**	**85.3**

AMBER基准。这是一个多维幻觉基准，不仅评估对象是否被错误纳入（通过 CHAIR），还评估模型如何错误表述属性或关系（HAL），以及生成的图注（Wang 等人，2023a）在认知一致性（Cog）方面的偏离程度。在此，**HAL** 衡量总体幻觉频率，而**Cog** 衡量所描述的属性或关系与真实标签之间的差异。如表 2 所示，**CRoPS**在各个模型上始终优于所有基线方法。在 LLaVA-1.5 系列上，CRoPS 使 CHAIR 和 HAL 均降低约**25–30%**，并使 Cog 降低**10–15%**，展现出更准确且语义更连贯的描述。对于 LLaVA-NeXT 和 Qwen2-VL，它也取得了类似的改进，在保留流畅度和描述丰富性的同时，将幻觉指标降低约**20–25%**。这些在不同架构上始终如一的降幅表明，与先前的免训练方法相比，CRoPS 更有效地缓解了感知幻觉和认知幻觉。

GPT-4o辅助基准。 GPT-4o辅助基准（Zhou 等人，2024年）使用 Visual Genome 标注作为真实标签，并提示 GPT-4o 以更细粒度评估幻觉。该基准衡量句子级幻觉率（SHR），即生成的句子中包含幻觉内容的频率，同时还评估流畅度（1 元和 2 元精确率）和描述性（每图词数，WPI）。如图 5 所示，**CRoPS** 在所有模型变体中的 SHR 均最低，展现了其

表 4：通用基准上的结果，展示了 LLaVA-1.5、LLaVA-NeXT 和 Qwen2-VL 在 **MME** 和 **MathVista** 上的得分。数值越高表示性能越好。粗体数字表示最佳结果，下划线数字表示各列中的次佳结果。***CRoPS*** 在大多数设置下均取得最高分。

方法	LLaVA-1.5		LLaVA-NeXT		Qwen2-VL	
	MME ↑	MathVista ↑	MME ↑	MathVista ↑	MME ↑	MathVista ↑
采样	1601	<u>27.4</u>	1669	34.8	2058	**56.9**
ClearSight	1569	27.1	1602	35.1	2001	54.7
VCD	1622	26.5	1630	36.2	2070	55.1
ICD	1605	25.9	1664	35.0	2060	54.2
OPERA	1615	27.0	1650	35.5	2080	54.8
SID	<u>1634</u>	26.3	1607	<u>36.8</u>	<u>2097</u>	54.3
M3ID	1607	26.9	<u>1683</u>	36.6	2090	52.0
CRoPS	**1662**	**28.9**	**1779**	**38.0**	**2184**	<u>55.6</u>

在抑制关系级和属性级幻觉方面的优势。此外，CRoPS 保持了出色的 1 元和 2 元流畅度以及较高的 WPI，说明其生成的描述流畅且详尽。与其他通过缩短输出以减少幻觉的基线方法不同，CRoPS 在提升事实依据的同时保留了文本的丰富性。**POPE**。 POPE 基准测试（Li 等人，2023年）采用二元视觉问答形式评估幻觉检测，其中模型必须确认或否认图像中是否存在某个物体。如表 3 所示，**CRoPS**在所有大型视觉语言模型中均取得最高的准确率和 F1，较最强基线方法平均提升约 **2%** 。这一持续的提升表明，即使在受限的是非回答格式下，CRoPS 仍然稳健，并能在不同模型中保持较强的视觉定位能力。

虽然 M3ID 的表现颇具竞争力，但在这种二元设置下，其表现略逊于 SID 和 CRoPS。这种差异源于 POPE 与图像描述基准测试的不同：POPE 只要求简洁的“是/否”回答，因此通常会限制开放式生成中出现的视觉令牌稀释程度。然而，正如 M3ID（Favero 等人，2024年）中所讨论的，由于视觉问答提示模板引入了图像词元与输出标记之间的结构性分离，视觉令牌稀释效应仍不可忽视。为弥补这一偏差，我们沿用相同的配置，并选择解码位置 $t = t_0$，其中 t_0对应图像与答案片段之间的词元数量。此调整确保 CRoPS 将视觉定位信号与输出区域对齐，从而在所有模型上持续实现更高的精确率和均衡的接地表现。详细结果见附录 G。

MME 和 MathVista 评估。MME 基准（Liang 等人，2024年）用于评估通用多模态感知与识别。表 4 显示，CRoPS 提升了不同架构上的 MME 得分，表明其在各类感知任务中具有更强的接地能力。MathVista（Lu 等人，2024年）考察视觉数学推理；CRoPS 在各模型上的表现均名列前茅或接近领先水平，这表明其解码策略有益于图像描述之外的结构化多模态推理。这些结果凸显了 CRoPS 的对比式设计能够提升各类大型视觉语言模型任务中的描述忠实度和逻辑推理能力。

7 分析

延迟比较。 表 5 展示了 LLaVA-1.5 7B 在 CHAIR 基准测试上的推理时间和 GPU峰值内存。尽管额外增加了一次前向传递，CRoPS 凭借轻量级输入设计带来的开销仍然极小。性能提升源于有针对性的对比式设计，而非反复进行模型调用。与 5 束搜索相比，CRoPS 仅略微增加时间和内存，就实现了更低的幻觉率。本分析采用 Vanilla 实现；优化后的版本还能进一步降低开销。

表 5：NVIDIA A100 上的效率比较。时间以秒为单位，内存使用量以 MB 为单位。

方法	时间 ↓	内存 ↓	C_S ↓
采样	215	15699	57
束搜索（5 个束）	531	16737	50.7
VCD	550	17864	53.3
SID	510	16574	48.93
OPERA	1947	21943	49.1
CRoPS	652	16934	**39.46**

表 6：**左侧：**图像和文本移除对幻觉模型的影响。**右侧：**不同 token 保留策略在 CHAIR 指标和召回率上的比较。

图像移除	文本移除	$C_S \downarrow$	$C_I \downarrow$	召回率 ↑
Full	No	45.4	12.4	73.3
No	部分（最不重要）	47.8	13.2	73.8
Full	部分（最不重要）	**43.2**	**10.8**	73.3
Full	部分（随机）	50.7	13.8	76.4

策略	$C_S \downarrow$	$C_I \downarrow$	召回率 ↑
常数 仅保留一个策略	43.6 46.8	10.9 13.1	75.0 73.6
线性	41.2	10.7	74.6
Ours	**39.5**	**10.2**	**76.3**

图像和文本移除的影响。 表 6 左侧子表分析了图像和文本移除对幻觉行为的影响。第 *1* 行 移除所有图像词元，同时保留文本词元。第 *2* 行 保留所有图像词元，但移除最不重要的文本词元。第 *3* 行 为我们提出的设置，即移除所有图像词元，同时部分保留文本词元。第 *4* 行 移除数量相同的文本词元，但移除对象是随机选择的。我们的方法优于其他方法，证实有选择地移除最不重要的文本词元更有效（见第4.2节）。

幻觉模型的选择。 为测试我们的对比式幻觉缓解策略，我们将第二个产生幻觉的模型替换为 M3ID，同时保持 CRoPS 不变。表 7 显示，这样做的效果较差，凸显了我们提出的第二模型设计的有效性。

表 7：在 MS-COCO 数据集上，针对幻觉模型选择进行 CHAIR 指标消融。

方法	$C_S \downarrow$	$C_I \downarrow$	召回率 ↑
M3ID + SID	45.1	11.8	78.3
CRoPS	**39.5**	**10.2**	76.3

令牌保留策略。 我们通过策略 $t(\eta)$ t 控制每个生成步骤中保留的词元数量 t。基线策略包括：**常数** ($\eta(t) = \beta_0$)，保留固定数量 β_0 的词元；**仅保留一个策略** ($\eta(t) = t - 2$)，丢弃除最新词元以外的所有词元；以及**线性** ($\eta(t) = \beta_0 + \beta_1 t$)，确保随着 t 增长，保留的词元数量稳步线性增加。为实现平滑且逐渐饱和的增长，我们引入一种**指数策略**，定义为 $\eta(\mu, t) = \beta_0 + \beta_1(1 - e^{-\mu t})$，其中 $\mu > 0$ 控制增长速率。该策略从 β_0 开始，并渐近趋近于 $\beta_0 + \beta_1$，从而避免上下文丢失突增，并在实证中带来更稳定的生成。如表 6 右侧子表所示，我们的指数策略取得了最低的幻觉评分，优于所有基线策略。

超参数消融。 我们分析了 CRoPS 对公式 11 中关键超参数的敏感性。系数 $\alpha^{(1)}$ 和 $\alpha_t^{(2)}$ 分别控制对比惩罚的总体强度和随时间变化的增长。参数 β_0、 β_1和 μ 控制动态加权函数 η，该函数决定每一步掩码多少个文本词元。此处，β_0 和 β_1 定义保留词元范围的下限和上限，而 μ 控制 η随时间增长的平滑程度。如表8所示，适中的对比强度和渐进式掩码效果最佳，

表8：改变超参数对 CRoPS 性能的影响。

α	γ	b_0	b_1	μ	C_S
1.0	0.02	10	30	1×10^{-3}	39.5
0.5	0.05	5	20	1×10^{-3}	42.3
1.0	0.02	5	40	1×10^{-3}	41.2
1.5	0.01	10	30	1×10^{-2}	43.5
1.0	0.02	10	40	1×10^{-3}	40.7

而极端 的 α 或 μ 取值会因削弱或过度抑制 e
对比信号而降低性能。

所有实验均采用一套共享的超参数配置： $\alpha = 1.0$、 $\gamma = 0.02$、 $\beta_0 = 10$、$\beta_1 = 30$和 $\mu = 1\text{e}{-3}$。所有骨干网络和基准测试均一致使用此配置，不进行针对特定模型或数据集的调优。

8 结论与局限性

在本研究中，我们首先指出，现有缓解技术分别针对视觉令牌稀释和对无关视觉特征的注意力，却无法全面解决这两个问题。为弥补这一

不足，我们提出了 CRoPS，这是一种新颖的解码策略，能够有效缓解这两种幻觉来源，同时只增加极少的计算开销，并在多个基准测试中取得显著提升。

局限性。 尽管我们的研究对对比解码的弱点提供了宝贵见解，但并未面面俱到。我们重点关注了某些类型的幻觉，将其他方面的探索留待未来工作。此外，我们的分析仅限于最新的 SOTA 方法，且只考虑免训练对比解码方法。探索其他框架，包括涉及特定任务训练的框架，有助于更全面地理解对比解码中的权衡。

我们的方法也会增加一些推理延迟。尽管 CRoPS 比原始采样慢约 $3\times$ （652 秒对比 215 秒），其运行时间仍与现有免训练方法相当或更短，例如 VCD（550 秒）、SID（510 秒）和 OPERA（1947 秒）。额外开销来自一次额外的前向传播，但由于文本缺失模型仅处理一小部分词元，因此这次传播的计算负担相对较轻。延迟仍是一项实际限制。未来工作可以降低这部分开销，例如并行执行生成幻觉模型所需的前向传播。

参考文献

Ali Furkan Biten、Lluís Gómez 和 Dimosthenis Karatzas。在海滩上放一个时钟：减少图像描述生成中的物体幻觉。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，第 1381–1390 页，2022年。

Zhaorun Chen、Zhuokai Zhao、Hongyin Luo、Huaxiu Yao、Bo Li 和 Jiawei Zhou。Halc：通过自适应焦点对比解码减少物体幻觉。*arXiv* 预印本 *arXiv:2403.00425*，2024年。

Yung-Sung Chuang、Yujia Xie、Hongyin Luo、Yoon Kim、James R. Glass 和 Pengcheng He。Dola：通过对比层进行解码，提升大语言模型的事实性。载于第十二届国际学习表征会议，2024年。URL https://openreview.net/forum?id=Th6NyL07na。

Wenliang Dai、Junnan Li、DONGXU LI、Anthony Tiong、Junqi Zhao、Weisheng Wang、Boyang Li、Pascale N Fung 和 Steven Hoi。Instructblip：通过指令微调构建通用视觉语言模型。载于 A. Oh、T. Naumann、A. Globerson、K. Saenko、M. Hardt 和 S. Levine（编），*神经信息处理系统进展*，第 36 卷，第 49250–49267 页。Curran Associates, Inc.，2023年。URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf。

Alessandro Favero、Luca Zancato、Matthew Trager、Siddharth Choudhary、Pramuditha Perera、Alessandro Achille、Ashwin Swaminathan 和 Stefano Soatto。通过视觉信息接地控制多模态幻觉。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 14303–14312 页，2024年。

Shangbin Feng、Weijia Shi、Yike Wang、Wenxuan Ding、Vidhisha Balachandran 和 Yulia Tsvetkov。不要产生幻觉，而应弃答：通过多大语言模型协作识别大语言模型的知识盲区，2024年。URL https://arxiv.org/abs/2402.00367。

Tianrui Guan、Fuxiao Liu、Xiyang Wu、Ruiqi Xian、Zongxia Li、Xiaoyu Liu、Xijun Wang、Lichang Chen、Furong Huang、Yaser Yacoob、Dinesh Manocha 和 Tianyi Zhou。Hallusionbench：大型视觉语言模型中语言幻觉与视觉错觉交织现象的高级诊断套件。载于*2024*年 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第 14375–14385 页，2024年。doi：10.1109/CVPR52733.2024.01363。

Anisha Gunjal、Jihan Yin 和 Erhan Bas。在大型视觉语言模型中检测和防止幻觉。载于*AAAI*人工智能会议论文集，第 38 卷，第 18135–18143 页，2024年。

Qidong Huang、Xiaoyi Dong、Pan Zhang、Bin Wang、Conghui He、Jiaqi Wang、Dahua Lin、Weiming Zhang 和 Nenghai Yu。《Opera：通过过度信任缓解多模态大语言模型中的幻觉》

惩罚与回顾分配。在 *IEEE/CVF*计算机视觉与模式识别会议，第 13418–13427 页，2024年。

Fushuo Huo、Wenchao Xu、Zhong Zhang、Haozhao Wang、Zhicheng Chen 和 Peilin Zhao。自省式解码：缓解大型视觉语言模型的幻觉。*arXiv* 预印本 *arXiv:2408.02032*，2024年。

Chaoya Jiang、Haiyang Xu、Mengfan Dong、Jiaxing Chen、Wei Ye、Ming Yan、Qinghao Ye、Ji Zhang、Fei Huang 和 Shikun Zhang。面向多模态大语言模型的幻觉增强对比学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 27036–27046 页，2024年。

Jae Myung Kim、A Koepke、Cordelia Schmid 和 Zeynep Akata。揭示并缓解跨模态检索中的虚假相关性。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 2585–2595 页，2023。

Junho Kim、Hyunjun Kim、KIM YEONJU 和 Yong Man Ro。CODE：通过对比自生成描述来对抗大型多模态模型中的幻觉。载于第三十八届神经信息处理系统年会，2024a。URL https://openreview.net/forum?id=g6nn2AijDp。

Taehyeon Kim、Joonkee Kim、Gihun Lee 和 Se-Young Yun。指令式解码：指令微调大语言模型是含噪指令的自我精炼器，2024b。URL https://arxiv.org/abs/2311.00233。

Ranjay Krishna、Yuke Zhu、Oliver Groth、Justin Johnson、Kenji Hata、Joshua Kravitz、Stephanie Chen、Yannis Kalantidis、Li-Jia Li、David A. Shamma、Michael S. Bernstein 和 Li Fei-Fei。Visual Genome：利用众包密集图像标注连接语言与视觉。国际计算机视觉期刊，123：32–73，2017。doi：10.1007/s11263-016-0981-7。URL https://doi.org/10.1007/s11263-016-0981-7。

Nayeon Lee、Wei Ping、Peng Xu、Mostofa Patwary、Pascale N Fung、Mohammad Shoeybi 和 Bryan Catan-zaro。用于开放式文本生成的事实性增强语言模型。神经信息处理系统进展，35:34586–34599，2022年。

Sicong Leng、Hang Zhang、Guanzheng Chen、Xin Li、Shijian Lu、Chunyan Miao 和 Lidong Bing。通过视觉对比解码缓解大型视觉语言模型中的物体幻觉。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第13872–13882页，2024年。

Jiwei Li、Michel Galley、Chris Brockett、Jianfeng Gao 和 Bill Dolan。面向神经对话模型的多样性促进目标函数。载于 Kevin Knight、Ani Nenkova 和 Owen Rambow（编），*2016*年计算语言学协会北美分会会议论文集：人类语言技术，第110–119页，圣迭戈（加利福尼亚州），2016年6月。计算语言学协会。doi：10.18653/v1/N16-1014。URL https://aclanthology.org/N16-1014/。

Xiang Lisa Li、Ari Holtzman、Daniel Fried、Percy Liang、Jason Eisner、Tatsunori Hashimoto、Luke Zettle-moyer 和 Mike Lewis。对比解码：将开放式文本生成视为优化。在 Anna Rogers、Jordan Boyd-Graber 和 Naoaki Okazaki（编）编的《计算语言学协会第*61*届年会论文集（第*1*卷：长篇论文）》中，第12286–12312页，多伦多，加拿大，2023年7月。计算语言学协会。doi：10.18653/v1/2023.acl-long.687。URLhttps://aclanthology.org/2023.acl-long.687/。

Zhuowei Li、Haizhou Shi、Yunhe Gao、Di Liu、Zhenting Wang、Yuxiao Chen、Ting Liu、Long Zhao、Hao Wang 和 Dimitris N Metaxas。词元的隐秘生命：通过视觉信息引导减少大型视觉语言模型的幻觉。*arXiv* 预印本 *arXiv:2502.03628*，2025年。

Zijing Liang、Yanjie Xu、Yifan Hong、Penghui Shang、Qi Wang、Qiang Fu 和 Ke Liu。多模态大语言模型综述。载于第三届计算机、人工智能与控制工程国际会议论文集，第 405–409 页，2024年。

Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。Microsoft COCO：情境中的常见对象。载于计算机视觉 —*ECCV 2014*：第 *13* 届欧洲会议，苏黎世，瑞士，*2014*年 *9* 月 *6–12* 日，论文集，第 *V* 卷第 *13* 部分，第 740–755 页。施普林格，2014年。

Haokun Liu、Yaonan Zhu、Kenji Kato、Izumi Kondo、Tadayoshi Aoyama 和YasuhisaHasegawa。基于大语言模型的人机协作框架，用于操作任务。*arXiv* 预印本 *arXiv:2308.14972*，2023a。

Haotian Liu、Chunyuan Li、Yuheng Li 和 Yong Jae Lee。通过视觉指令微调改进基线方法，2023b。

Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。视觉指令微调。载于 *NeurIPS*，2023c。

Haotian Liu、Chunyuan Li、Yuheng Li、Bo Li、Yuanhan Zhang、Sheng Shen 和 Yong Jae Lee。Llava-next：改进推理、光学字符识别和世界知识，2024年1月。URL https://llava-vl.github.io/blog/2024-01-30-llava-next/。

Shi Liu、Kecheng Zheng 和 Wei Chen。更加关注图像：一种缓解大型视觉语言模型幻觉的免训练方法。载于欧洲计算机视觉会议，第 125–140 页。施普林格，2024b。

Pan Lu、Hritik Bansal、Tony Xia、Jiacheng Liu、Chunyuan Li、Hannaneh Hajishirzi、Hao Cheng、Kai-Wei Chang、Michel Galley 和 Jianfeng Gao。MathVista：评估基础模型在视觉语境中的数学推理能力。载于国际学习表征会议（*ICLR*），2024年。

Potsawee Manakul、Adian Liusie 和 Mark Gales。SelfCheckGPT：面向生成式大语言模型的零资源黑盒幻觉检测。载于*2023*年自然语言处理实证方法会议，2023。网址 https://openreview.net/forum?id=RwzFNbJ3Ez。

Anku Rani、Vipula Rawte、Harshad Sharma、Neeraj Anand、Krishnav Rajbangshi、Amit Sheth 和 Amitava Das。视觉幻觉：定义、量化与规范性补救措施，2024年。网址 https://arxiv.org/abs/2403.17306。

Anna Rohrbach、Lisa Anne Hendricks、Kaylee Burns、Trevor Darrell 和 Kate Saenko。图像描述生成中的对象幻觉。载于 Ellen Riloff、David Chiang、Julia Hockenmaier 和 Jun'ichi Tsujii（编），*2018*年自然语言处理实证方法会议论文集，第4035–4045页，比利时布鲁塞尔，2018年10月至11月。计算语言学协会。doi：10.18653/v1/D18-1437。URL https://aclanthology.org/D18-1437/。

Feilong Tang、Chengzhi Liu、Zhongxing Xu、Ming Hu、Zile Huang、Haochen Xue、Ziyang Chen、Zelin Peng、Zhiwei Yang、Sijin Zhou 等。看得更远、更清晰：利用注意力因果解码缓解多模态大语言模型中的幻觉。载于计算机视觉与模式识别会议论文集，第26147–26159页，2025年。

Junyang Wang、Yuhang Wang、Guohai Xu、Jing Zhang、Yukai Gu、Haitao Jia、Jiaqi Wang、Haiyang Xu、Ming Yan、Ji Zhang 等。AMBER：用于多模态大语言模型幻觉评估的无大语言模型多维基准。*arXiv* 预印本 *arXiv:2311.07397*，2023a。

Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhihao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Yang Fan、Kai Dang、Mengfei Du、Xuancheng Ren、Rui Men、Dayiheng Liu、Chang Zhou、Jingren Zhou 和 Junyang Lin。Qwen2-VL：在任意分辨率下增强视觉语言模型对世界的感知能力，2024年a。网址：https://arxiv.org/abs/2409.12191。

Sheng Wang、Zihao Zhao、Xi Ouyang、Qian Wang 和 Dinggang Shen。ChatCAD：使用大语言模型对医学图像进行交互式计算机辅助诊断。*arXiv* 预印本 *arXiv:2302.07257*，2023b。

Xintong Wang、Jingheng Pan、Liang Ding 和 Chris Biemann。通过指令对比解码缓解大型视觉语言模型中的幻觉，2024b。URL https://arxiv.org/abs/2403.18715。

Jason Wei、XuezhiWang、DaleSchuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。思维链提示能够激发大语言模型的推理能力。神经信息处理系统进展，35：24824–24837，2022年。

Sangmin Woo、Jaehyuk Jang、Donguk Kim、Yubin Choi 和 Changick Kim。Ritual：随机图像变换作为大型视觉语言模型的通用抗幻觉手段。*arXiv* 预印本 *arXiv:2405.17821*，2024年。

Junfei Wu、Qiang Liu、Ding Wang、Jinghao Zhang、Shu Wu、Liang Wang 和 Tieniu Tan。逻辑闭环：揭示大型视觉语言模型中的物体幻觉。载于 Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编），计算语言学协会会议论文集：*ACL 2024*，第 6944–6962 页，泰国曼谷，2024年8月。计算语言学协会。doi：10.18653/ v1/2024.findings-acl.414。URL https://aclanthology.org/2024.findings-acl.414/。

Mingrui Wu、Jiayi Ji、Oucheng Huang、Jiale Li、Yuhang Wu、Xiaoshuai Sun 和 Rongrong Ji。评估并分析大型视觉语言模型中的关系幻觉。*CoRR*，abs/2406.16449，2024b。网址 https://doi.org/10.48550/arXiv.2406.16449。

Hao Yin、Guangzong Si 和 Zilei Wang。ClearSight：通过视觉信号增强缓解多模态大语言模型中的物体幻觉。载于计算机视觉与模式识别会议论文集，第 14625–14634 页，2025年。

Shukang Yin、Chaoyou Fu、Sirui Zhao、Tong Xu、Hao Wang、Dianbo Sui、Yunhang Shen、Ke Li、Xing Sun 和 Enhong Chen。Woodpecker：多模态大语言模型的幻觉纠正。中国科学：信息科学，67(12):220105，2024年。

Tianyu Yu、Yuan Yao、Haoye Zhang、Taiwen He、Yifeng Han、Ganqu Cui、Jinyi Hu、Zhiyuan Liu、Hai-Tao Zheng、Maosong Sun 等。《RLHF-V：通过细粒度纠正性人类反馈实现行为对齐，迈向可信多模态大语言模型》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 13807–13816 页，2024年。

Zihao Yue、Liang Zhang 和 Qin Jin。《少即是多：从 EOS决策视角缓解多模态幻觉》。载于 Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编），计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第 11766–11781 页，泰国曼谷，2024年8月。计算语言学协会。doi：10.18653/v1/2024.acl-long.633。URL https://aclanthology.org/2024.acl-long.633/。

Muru Zhang、Ofir Press、William Merrill、Alisa Liu 和 Noah A. Smith。语言模型中的幻觉如何滚雪球般加剧。载于第四十一届国际机器学习大会，2024年。URLhttps://openreview.net/forum?id=FPlaQyAGHu。

Zhiyuan Zhao、Bin Wang、Linke Ouyang、Xiaoyi Dong、Jiaqi Wang 和 Conghui He。超越幻觉：通过幻觉感知直接偏好优化增强大型视觉语言模型。*arXiv* 预印本 *arXiv:2311.16839*，2023。

ZhiyuanZhao、BinWang、LinkeOuyang、XiaoyiDong、JiaqiWang 和ConghuiHe。超越幻觉：通过幻觉感知直接偏好优化增强大型视觉语言模型，2024年。URL https://arxiv.org/abs/2311.16839。

Kening Zheng、Junkai Chen、Yibo Yan、Xin Zou 和 Xuming Hu。Reefknot：一个用于多模态大语言模型关系幻觉评估、分析和缓解的综合基准。*arXiv* 预印本 *arXiv:2408.09429*，2024年。

Yiyang Zhou、Chenhang Cui、Jaehong Yoon、Linjun Zhang、Zhun Deng、Chelsea Finn、Mohit Bansal 和 Huaxiu Yao。分析并缓解大型视觉语言模型中的物体幻觉，2024年。URLhttps://arxiv.org/abs/2310.00754。

Lanyun Zhu、Deyi Ji、Tianrun Chen、Peng Xu、Jieping Ye 和 Jun Liu。IBD：通过图像偏置解码缓解大型视觉语言模型中的幻觉。*arXiv* 预印本 *arXiv:2402.18476*，2024年。

A 多模态互信息解码（M3ID）

M3ID 提出使用同一个大语言模型，通过从输入中完全移除视觉词元 $\boldsymbol{V}$ 来生成幻觉概率。形式上，M3ID 定义为:

$$\begin{aligned} \boldsymbol{p}_t^{orig} &:= \text{softmax} \circ \text{LLM}_\theta\big(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{y}_{<t}\big), \\ \boldsymbol{p}_t^{hal} &:= \text{softmax} \circ \text{LLM}_\theta\big(\boldsymbol{X}, \boldsymbol{y}_{<t}\big), \\ \alpha_t &:= (1 - e^{-\gamma t})/e^{-\gamma t}, \\ \log \boldsymbol{p}_t^{\text{M3ID}} &:= \log \boldsymbol{p}_t^{or} + \alpha_t \cdot (\log \boldsymbol{p}_t^{orig} - \log \boldsymbol{p}_t^{hal}). \end{aligned} \tag{12}$$

请注意，混合系数 α 取决于时间，是推理时间步 t的单调递增函数。M3ID 通过指出，随着 t 增加，幻觉会愈发严重，或者说， $\boldsymbol{p}_t^{hal}$的影响会愈发显著，来对此作出解释。

B 自我内省解码（SID）。

SID 会有选择地保留连贯性较低的视觉词元，而不是从大语言模型的输入中移除所有视觉词元。视觉词元的连贯性由分配给每个视觉输入词元的汇总注意力分数确定。随后，每个视觉词元 $\boldsymbol{v}_i$ 的分数 $\psi(\boldsymbol{y}_t)$ 按如下方式计算:

$$\psi(\boldsymbol{y}_t) = \frac{1}{H} \sum_{h=1}^{H} \text{Attention}^{(h)}\big(\boldsymbol{v}_i, \boldsymbol{y}_t\big), \tag{13}$$

利用该分数，SID 选取分数最低的底部 $\bar{m}$，$(< m = |\boldsymbol{V}|)$ 个视觉词元，并移除高重要性标记，以创建一个由视觉词元组成的更稀疏集合 $\boldsymbol{V}$：$=$ 稀疏$(\boldsymbol{V})$与$|\boldsymbol{V}| = \bar{m}$。
考虑到 LLaVA-NeXT 和 Qwen2-VL 等现代架构的图像词元数量并非固定，我们动态地将 $\bar{m}$设为图像词元总数的 25%。最后，从弱模型获得的概率输出如下:

$$\begin{aligned} \boldsymbol{p}_t^{orig} &:= \text{softmax} \circ \text{LLM}\theta\big(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{y}_{<t}\big), \\ \boldsymbol{p}_t^{hal} &:= \text{softmax} \circ \text{LLM}\theta\big(\overline{\boldsymbol{V}}, \boldsymbol{X}, \boldsymbol{y}_{<t}\big), \\ \log \boldsymbol{p}_t^{\text{SID}} &:= (1 + \alpha) \cdot \log \boldsymbol{p}_t^{orig} - \alpha \cdot \log \boldsymbol{p}_t^{hal}. \end{aligned} \tag{14}$$

公式 2 中的混合系数 α 是 SID 中的一个超参数。与 M3ID 中依赖时间变化的混合不同，SID 在整个推理时间步中使用一个常数 α 。

C M3ID 与我们的新型幻觉模型的详细比较

方法	LLaVA-1.5			LLaVA-NeXT			Qwen2-VL		
	$C_S \downarrow$	$C_I \downarrow$	召回率 ↑	$C_S \downarrow$	$C_I \downarrow$	召回率 ↑	$C_S \downarrow$	$C_I \downarrow$	召回率 ↑
M3ID	45.4	12.4	73.3	36.1	9.8	68.7	28.8	7.3	64.8
新模型 $\boldsymbol{p}^{\text{vis-txt-hal}}$	43.2	10.8	73.3	35.2	9.4	66.2	28.0	7.4	66.6

表9：使用 CHAIR 指标对比不同模型中 M3ID 和新型幻觉模型的消融结果。 $\boldsymbol{p}^{\text{vis-txt-hal}}$

表9展示了第5节提出的幻觉模型 $\boldsymbol{p}^{\text{vis-txt-hal}}$ 的性能提升。该模型仿照 M3ID，不使用全部文本词元，而是选择最不重要的文本词元。按照 SID（视觉词元选择）采用的相同方法，仅选择最不重要的词元后，我们观察到所有模型在 CHAIR 基准测试上的性能均有所提升。

算法1 *CRoPS*

需要： 大语言模型 LLM_θ，文本 $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$，图像 $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$
需要： 超参数：$\eta(\cdot, \cdot)$，$\bar{m}$，γ，α
确保： 输出序列 $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$

$\boldsymbol{y}_0 \leftarrow$ 序列起始标记，$t \leftarrow 1$, $\boldsymbol{Y}_{<1} \leftarrow (\boldsymbol{y}_0)$
当 $\boldsymbol{y}_{t-1} \neq$ 序列结束标记 **时执行**
  $\overline{\boldsymbol{V}} \leftarrow$ 最不重要项$[\boldsymbol{V}, \bar{m}]$
  $\overline{\boldsymbol{X}, \boldsymbol{y}_{<t}} \leftarrow$ 最不重要项$[(\boldsymbol{X}, \boldsymbol{y}_{<t})\ n + t - \eta(n, t)]$
  $\boldsymbol{p}_t^{orig} \leftarrow$ softmax函数 $\circ$ 大语言模型$_\theta(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{y}_{<t})$
  $\boldsymbol{p}_t^{\text{vis-hal}} \leftarrow$ softmax函数 $\circ$ 大语言模型$_\theta(\overline{\boldsymbol{V}}, \boldsymbol{X}, \boldsymbol{y}_{<t})$
  $\boldsymbol{p}_t^{\text{vis-txt-hal}} \leftarrow$ softmax函数 $\circ$ 大语言模型$_\theta(\overline{\boldsymbol{X}, \boldsymbol{y}_{<t}})$
  $\alpha_t^{(1)} \leftarrow \frac{1-e^{-\gamma t}}{e^{-\gamma t}}$, $\alpha_t^{(2)} \leftarrow \alpha$
  日志 $\boldsymbol{p}_t^{\text{CRoPS}} \leftarrow \left(1 + \alpha_t^{(1)} + \alpha_t^{(2)}\right) \cdot$ 日志 $\boldsymbol{p}_t^{orig}$
  $-\alpha_t^{(1)} \cdot$ 日志 $\boldsymbol{p}_t^{\text{vis-hal}} - \alpha_t^{(2)} \cdot$ 日志 $\boldsymbol{p}_t^{\text{视觉文本幻觉}}$
  $\boldsymbol{y}_t \leftarrow$解码$(\boldsymbol{p}_t^{\text{CRoPS}})$
  $\boldsymbol{Y}_{<t+1} \leftarrow (\boldsymbol{y}_{<t} : \boldsymbol{y}_t)$
  $t \leftarrow t + 1$
结束 while 循环

D 基准测试与评估指标

大型视觉语言模型的幻觉基准大致可分为生成式和判别式方法。**生成式基准测试**用于评估自由形式文本生成中的幻觉。在对 CRoPS 的评估中，我们使用：1）CHAIR（Rohrbach 等，2018）：该指标通过比较生成的图注中提及的对象与对应图像中存在的标注对象来衡量幻觉。2）AMBER（Wang 等人，2023a）：AMBER 量化幻觉响应（Hal）的比例，并评估其与人类认知（Cog）的一致性。3）GPT-4 辅助基准（Zhao 等人，2023）：这些基准测试利用 Visual Genome 数据集（Krishna 等人，2017）中的细粒度对象级描述，并使用 GPT-4 计算句子级幻觉率（SHR）。此外，我们还计算 n 元语法流畅度（$n = 1, 2$）以评估文本流畅度，并通过测量每图词数（WPI）和每图句数（SPI）来分析文本的冗长程度和细节丰富度。**判别式基准测试**在视觉问答场景中评估幻觉，其中回答通常为二元形式（例如“是”或“否”），因此评估类似于分类任务。我们使用 POPE（Li 等人，2023年），该方法将物体幻觉定义为一个二元分类问题，问题形式为“图像中是否存在〈物体〉？”

除针对幻觉的专门评估外，我们还使用以下基准评估大型视觉语言模型的通用能力：1) MME（Liang 等人，2024年）：一个综合性基准，包含十个用于评估感知能力的子任务，以及四个通过“是/否”问题评估识别能力的子任务；以及 2) MathVista（Lu 等人，2024年）：一个旨在分析视觉复杂场景中数学推理能力的基准。

E 过去生成的标记的信息量

根据方程12,

$$\boldsymbol{p}_t^{orig} := \text{softmax} \circ \text{LLM}_\theta\big(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{y}_{<t}\big),$$
$$\boldsymbol{p}_t^{hal} := \text{softmax} \circ \text{LLM}_\theta\big(\boldsymbol{X}, \boldsymbol{y}_{<t}\big),$$
$$\log \boldsymbol{p}_t^{\text{M3ID}} := \log \boldsymbol{p}_t^{orig} + \alpha_t \cdot (\log \boldsymbol{p}_t^{orig} - \log \boldsymbol{p}_t^{hal})$$

其中，$\alpha_t := (1 - e^{-\gamma t})/e^{-\gamma t}$

此外，根据方程4，视觉依赖（VD）定义为:

$$\text{VD}(t) := \text{dist}\Big(\text{softmax}\circ\text{LLM}_\theta(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{y}_{<t}),\ \text{softmax}\circ\text{LLM}_\theta(\boldsymbol{X}, \boldsymbol{y}_{<t})\Big)$$

现在，由于 VD(t) $\to 0$,

$$\text{softmax}\circ\text{LLM}_\theta\left(\boldsymbol{X},\boldsymbol{y}_{<t}\right) \to \text{softmax}\circ\text{LLM}_\theta\left(\boldsymbol{V},\boldsymbol{X},\boldsymbol{y}_{<t}\right)$$

$$\implies \boldsymbol{p}_t^{hal} \to \boldsymbol{p}_t^{orig}$$

因此，

$$\lim_{\text{VD}(t)\to 0} \mathbf{p}_t^{\text{M3ID}} = \mathbf{p}_t^{orig}.$$

F 对比排序策略的消融

为评估我们提出的对比式幻觉缓解策略的有效性，我们在 **LLaVA-1.5-7B**设置下，对弱模型采用不同对比应用顺序进行了消融研究。具体而言，我们比较了三种变体：

- **SID → M3ID**：使用 SID 的幻觉模型进行对比解码，随后使用 M3ID 的幻觉模型进行对比。
- **M3ID → SID**：使用 M3ID 的幻觉模型进行对比解码，随后使用 SID 的幻觉模型进行对比。
- **新型弱模型（我们的方法）**：对 M3ID 的幻觉模型进行修改后应用对比解码，随后与 SID 的幻觉模型进行对比

结果总结如下：

表 10：针对对比策略顺序，在 MSCOCO数据集上进行的 CHAIR 指标消融。较低的 CHAIR指标表示幻觉较少，而较高的召回率表示答案忠实度较高。

方法	CHAIRs ↓	CHAIRi ↓	召回率 ↑
SID → M3ID	45.5	11.9	78.9
M3ID → SID	45.1	11.8	78.3
新型弱模型（我们的方法）	**39.5**	**10.2**	76.3

直接量化幻觉的**CHAIRs** 和**CHAIRi** 指标显著提升，表明我们的新型幻觉模型更善于拒绝幻觉内容。

这项消融揭示了一个关键见解：简单地组合现有的对比解码是不够的。仅以不同顺序应用 M3ID 和 SID 虽然能提升幻觉评分，但通过我们的新型幻觉模型还能进一步改善。

G POPE 基准测试的详细结果

表 11 报告了 POPE 基准测试三个子集上的详细结果：随机、热门和对抗。在所有模型变体和子集上，无论是准确率还是 F1，CRoPS 的表现始终优于所有基线方法。在对抗子集上的提升尤为显著，因为该子集以高度共现的物体对对模型提出挑战，这凸显了 CRoPS 即使面对令人困惑的视觉语境，也能保持视觉接地能力。在所有子集上，相较于 M3ID 和 SID，CRoPS 在两项指标上的平均提升约为 1–2%，表明它能够很好地适应不同难度级别。与 ClearSight 等注意力调整方法在 Qwen2-VL 上性能急剧下降不同，CRoPS 在不同架构上始终保持均衡而稳健的表现。这些结果进一步证明，CRoPS 能有效缓解二元接地任务中的幻觉，同时保持响应准确率。

表 11：POPE 基准测试的 MS-COCO 划分结果

数据集类型	方法	LLaVA-1.5（7B）		LLaVA-NeXT		Qwen2-VL	
		准确率 ↑	F1 ↑	准确率 ↑	F1 ↑	准确率 ↑	F1 ↑
随机	采样	85.7	86.0	91.0	90.8	86.1	84.5
	SID	87.1	86.8	91.4	91.5	87.7	84.9
	M3ID	86.8	87.0	91.0	91.4	86.4	85.7
	ClearSight	86.8	86.5	90.7	90.9	71.5	70.8
	VCD	86.5	86.2	91.1	91.2	86.2	85.0
	ICD	86.1	86.3	89.7	89.5	86.7	86.0
	OPERA	86.3	86.6	89.9	90.5	86.8	86.2
	CRoPS	**87.8**	**87.7**	**92.0**	**92.3**	**87.9**	**86.8**
热门	采样	82.8	83.6	88.1	88.2	84.3	82.7
	SID	84.3	84.2	89.6	89.4	86.0	85.3
	M3ID	83.0	83.8	89.0	89.1	84.7	84.0
	ClearSight	82.3	83.2	87.9	88.0	72.2	73.0
	VCD	83.8	83.6	88.0	88.7	85.6	83.3
	ICD	83.3	83.2	87.8	88.4	85.8	84.7
	OPERA	83.5	82.9	88.1	89.3	86.9	84.6
	CRoPS	**84.4**	**84.9**	**90.7**	**89.7**	**86.5**	**85.7**
对抗性	采样	76.4	78.5	82.9	83.9	82.3	81.2
	SID	77.9	79.3	84.4	84.3	83.1	83.2
	M3ID	77.5	79.6	83.1	84.2	81.6	83.5
	ClearSight	77.0	79.2	83.3	84.3	76.9	79.7
	VCD	77.7	79.9	83.9	84.4	82.2	81.2
	ICD	77.1	79.2	82.3	84.3	83.0	82.1
	OPERA	77.0	79.1	84.2	84.8	82.9	82.7
	CRoPS	**79.6**	**81.1**	**85.1**	**86.2**	**83.8**	**83.3**

H 受剪枝影响的文本词元定性分析

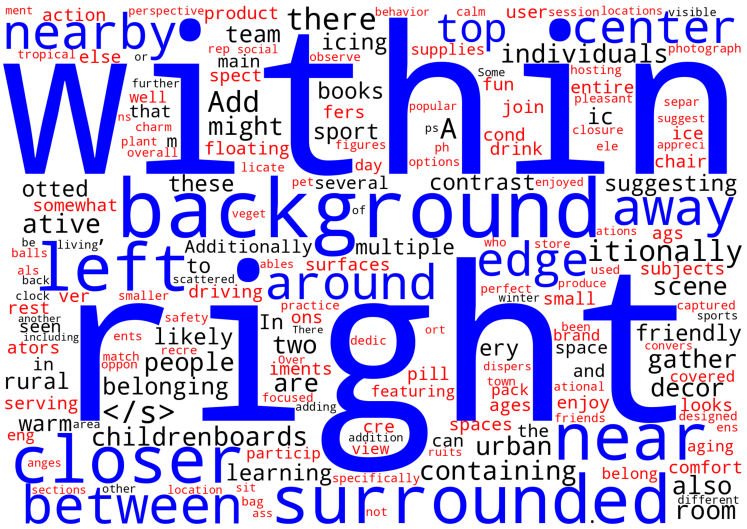
我们检查了根据公式 5 中重要性分数排名的文本词元。图 6 和图 7 展示了在视觉-文本缺陷模型中始终获得较高重要性、因而被移除的词元。左、右、中心、周围、内部、附近和上方等词语经常出现在被剪除的词元中，表明原始模型依赖提示中的显式位置线索。移除这些线索会削弱模型的接地能力，并增加出现位置相关不一致性的可能性。

第二组包含罕见或细粒度的内容词元，包括 *containing*、*depth*、*focus*、*unusual*、*expl*、*transport* 以及若干子词片段。虽然其中一些在我们的示例中只出现一次，但它们受到高度关注，也经常被剪枝。移除这些词元会使模型倾向于采用更通用的语言模式，而非提供具体细节。

剪枝后保留下来的其余低重要性词元大多是作用较弱的修饰语或片段。总的来说，分析表明，视觉-文本缺失扰动移除的是连贯且信息丰富的词元组，而非随机文本，这有助于解释它为何会可靠地诱发易产生幻觉的行为。

I 定性示例

图8、9和10展示了不同方法之间的定性比较 s.


within
background
right
left
around
edge
closer
between
surrounded
nearby
center
top
near
away
left
Additionally
multiple
itionally
scene
friendly
gather
decor
rural
belonging
are
people
likely
two
In
Add
might
floating
these
contrast
suggesting
individuals
team
there
product
action
perspective
behavior
calm
user
session
locations
visible
photograph
supplies
observe
main
spect
books
fers
sport
join
fun
entire
pleasant
ic
closure
ice
chair
drink
cond
popular
ps
A
figures
ph
options
day
licate
pet
several
somewhat
ative
otted
veget
of
ags
ations
who
store
used
subjects
perfect
produce
winter
small
captured
sports
driving
surfaces
ables
practice
ons
safety
ents
recre
match
oppon
iments
over
pill
featuring
focused
adding
cre
addition
view
ruits
spaces
the
can
urban
aging
participl
learning
containing
belong
comfort
also
room
different
specifically
sections
other
location
sit
bag
ass
hot
anges
eng
warm
area
childrenboards
serving
looks
designed
ens
friends
enjoy
ages
pack
dispers
town
ational
and
ery
space
brand
been
convers
covered
ators
including
seen
another
rest
clock
ver
smaller
scattered
back
balls
be
living
als
ment
tropical
else
or
rep
social
further
well
that
ns
charm
plant
overall
m
hosting
Some
entire
separ
suggest
appreci
ele
closure
ice
chair
enjoyed
contrast
suggesting
ort
dedic
There
ons
In
two
people
belonging
are
rural
serving
warm
childrenboards
learning
containing
urban
near
closer
between
surrounded
also
room
different
friendly
gather
decor
looks
designed
aging
comfort
belong
spaces
enjoy
ages
pack
ational
and
dispers
town
ery
space
brand
been
convers
covered
ens
friends
sports
captured
used
subjects
perfect
produce
winter
small
ags
ations
who
store
practice
ons
safety
ents
recre
match
oppon
iments
over
pill
featuring
focused
adding
cre
addition
view
ruits
spaces
the
can
urban
aging
participl
learning
containing
belong
comfort
also
room
different
specifically
sections
other
location
sit
bag
ass
hot
anges
eng
warm
area
childrenboards
serving
looks
designed
ens
friends
enjoy
ages
pack
dispers
town
ational
and
ery
space
brand
been
convers
covered
ators
including
seen
another
rest
clock
ver
smaller
scattered
back
balls
be
living
als
ment
tropical
else
or
rep
social
further
well
that
ns
charm
plant
overall
m
hosting
Some
entire
separ
suggest
appreci
ele
closure
ice
chair
enjoyed
contrast
suggesting
ort
dedic
There
ons
In
two
people
belonging
are
rural
serving
warm
childrenboards
learning
containing
urban
near
closer
between
surrounded
also
room
different
friendly
gather
decor
looks
designed
aging
comfort
belong
spaces
enjoy
ages
pack
dispers
town
ational
and



Prompt: Please describe this image in detail

Vanilla :- The image depicts a woman holding a white package while crossing the street at an intersection. She is wearing a purple jacket and carrying a handbag. As she is in the process of crossing, it can be inferred that she is looking for the appropriate sidewalk to step onto. The street is equipped with traffic lights and a stop sign, ensuring pedestrian safety. There are multiple traffic lights visible in the scene, with one near the left side, one at the top right corner, and another towards the top left corner of the image. The setting appears to be a residential or suburban area, with a building visible in the background, providing a sense of scale to the scene.

SID :- A woman stands on a crosswalk, holding a delivery of papers or white sheet across her body. She is likely waiting for a traffic light to change so she can continue crossing the street. The woman appears to be walking across the intersection as she makes her way. The street is equipped with multiple traffic lights. One traffic light is located in the upper-left corner, another is situated in the lower-left, and two more traffic lights can be seen in the lower-middle section. There is also a handbag on the ground, probably belonging to the woman. The scene appears to be a black and white photo.

M3ID :- The image features a woman walking across a street in a city, carrying a large white paper parachute or kite. She is standing in the middle of a crosswalk at an intersection with street lights, waiting to cross. \n\nThere are multiple traffic lights in the scene, with three close to the right side of the intersection and two more in the background. Window coverings can be seen on the buildings on both the left and right sides of the intersection. Her handbag is also visible, placed close to her body while she stands in the crosswalk.

CRoPS :- In the image, a woman is standing at an intersection holding a large sheet of paper or a newspaper in her hand. She appears to be crossing the street while navigating through traffic lights. There are multiple traffic lights visible in the scene, some closer to the woman while others are further away. The woman seems to be focused on her task while navigating the urban environment.

图8：不同方法生成的图注的定性比较。幻觉词以红色标出。



Prompt: Please describe this image in detail

Vanilla :- The image features two children standing in a snow ski store, surrounded by various skis and snowboards on display. A girl is looking at the skis, while a boy, standing next to her, is curiously examining the snowboards. There are numerous skis in different colors and positions, some closer to the children while others are further away on the right side of the store. A snowboard is also displayed in the background, close to the girl. The children appear to be fully engaged in the winter sports equipment, possibly contemplating which one to choose or simply enjoying the visual variety.

SID :- In the store, two little girls are looking at the display of skis and snowboards. The skis are arranged on shelves above the girls, and the snowboards are situated on separate shelves, some placed higher and others lower. The girls are standing close to the display, with one closer to the skis and the other closer to the snowboards. Both girls are wearing black coats, and pink hats are visible on their heads. It appears the girls are browsing the ski shop's offerings, possibly considering which one to purchase for their skiing or snowboarding activities.

M3ID :- The image depicts a young girl standing next to a counter filled with various skiing equipment. She is pointing towards a display of skis on the wall, engaging in conversation with another child. Together, they seem to be admiring the store's collection of skis and poles. There are multiple skis throughout the scene, with a particularly large snowboard display among the collection. A smaller skateboard can also be seen. The women at the shop might be mother and daughter, exploring the store and sharing their observation of the skiing gear.

CRoPS :- The image features two little girls standing next to each other in a ski shop, admiring the various ski equipment displayed around them. There are multiple pairs of skis visible in the scene, some placed higher up on shelves while others are positioned closer to the ground. The girls seem interested in trying out the skis or learning more about them. One girl appears closer to the left side of the image, while the other girl is positioned more towards the right side of the scene. Both girls seem engaged in the ski shop environment and eager to explore the ski equipment available for purchase or rent.

图9：不同方法生成的图注的定性比较。幻觉词以红色标出。



Prompt: Please describe this image in detail

Vanilla :- The image features a woman standing in a grassy field, holding the reins of a horse. The horse is tied in the field, allowing the woman to keep a close eye on it. A fence can be seen in the background, adding to the rural setting of the scene.

SID :- The woman, who is feeding hay to the white horse in a grassy field, stands in a black and white picture. She has short hair and appears happy while tending to the horse. The horse stands close to her, displaying a friendly demeanor as she feeds it. In the background, there are potted plants, and a fence, possibly in the distance, providing context to the surroundings. Overall, the scene is a calm and peaceful depiction of the woman's connection with her horse.

M3ID :- The image portrays a woman standing on a grassy field next to a white horse. The woman is petting the horse, taking the time to engage with it. There is a bridle on the horse's head, indicating that it is a riding horse. The scene includes a potted plant nearby, adding a touch of greenery to the lawn area.

CRoPS :- The image captures a woman standing in a grassy field next to a white horse. She appears to be petting the horse while holding its bridle in her hand. The scene takes place near some trees, adding a natural element to the environment. The woman appears to be enjoying her time with the horse in this serene setting.

图10：不同方法生成的图注的定性比较。幻觉词以红色突出显示。