# MRFD：多区域融合解码与自洽以减轻 LVLM 中的幻觉

**Haonan Ge**[λ] **Yiwei Wang**[λ] **Ming-Hsuan Yang**[λ] **Yujun Cai**[†]
[λ]University of California, Merced [†]The University of Queensland
gehaonan82@gmail.com

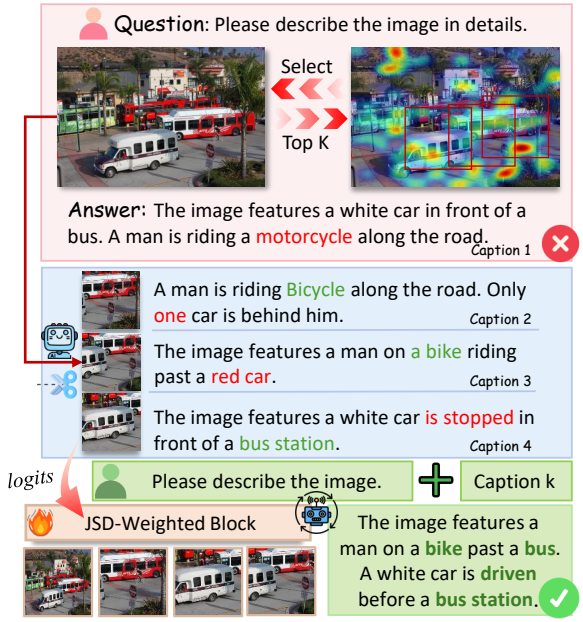


Figure 1: MRFD 过程：利用多个区域响应（图注 1-4），一个 JSD 加权块生成一致性权重来指导提示的融合解码，从而产生更可靠的输出。

## Abstract

大型视觉语言模型（LVLMs）在多模态任务中表现出色。然而，由于在图像的不同区域验证信息的能力有限，它们经常产生幻觉——与视觉输入不一致的文本。为了解决这个问题，我们提出了多区域融合解码（MRFD），一种无需训练的解码方法，通过建模区域间一致性来改进事实基础。MRFD 使用交叉注意力识别显著区域，为每个区域生成初始响应，并基于响应之间的 Jensen-Shannon 散度（JSD）计算可靠性权重。这些权重引导每个区域预测的一致性感知融合，使用受链式思维启发的区域感知提示。在多个 LVLMs 和基准测试的实验中表明，MRFD 在无需更新模型的情况下显著降低了幻觉并提高了响应的事实性。

## 1 介绍

大型视觉-语言模型（LVLMs）已成为多模态理解的强大工具，在图像描述、视觉问答和视觉推理等方面取得了显著进展 (????)。然而，这些模型经常生成幻觉内容，产生与视觉输入不一致的文本输出。这一现象在可靠性和实际部署方面提出了重大挑战 (??)。幻觉通常表现为识别错误的对象、捏造的属性或遗漏的视觉信息，从而降低图像与文本之间的事实对齐。

这一问题在需要对视觉复杂场景进行细粒度理解或解释的场景中尤为突出。目前的方法包括训练型解决方案，如数据增强和微调，这些方法改善了事实的基础但需要相当的资源，且通常缺乏泛化能力 (??)。无训练替代方案如链式思维提示 (?) 和对比或校正解码方法 (??????) 提供了灵活性，但通常依赖于精心设计的提示或未能充分利用内在的视觉信息。

这些方法的一个关键限制是它们倾向于整体处理图像或孤立分析区域，通常没有动态机制来评估不同视觉线索的可靠性或协调来自多个视角的潜在冲突解释。因此，在当前的解码过程中，基于显式一致性的推理在各种图像段中很少被整合 (??)。

我们的分析（第 3 节）强调了一些图像区域比其他区域提供了更可靠的证据。特别是，我们观察到由 Jensen-Shannon 散度（JSD）量化的区域级响应的一致性与其事实正确性相关（参见第 3.2 节）。这一强相关性表明，区域间的一致性是输出质量的关键决定因素：与多个区域的共识视图一致的响应不仅在事实准确性上更高，而且表现出更高的可靠性和更低的幻觉率。

为此，我们提出了一种名为多区域融合解码（MRFD）的解码策略，旨在将多视角推理纳入生成过程。如图 1 所示，MRFD 基于注意力机制识别图像中的几个显著区域，然后为每个区域生成初始响应。通过计算这些响应之间的 Jensen-Shannon Divergence（JSD），该方法得出一组基于一致性的可靠性权重。这些权重在解码过程中用于结合所有区域的预测。此外，受 CoT 提示的启发，我们通过将原始问题与每个区域的初始分析结合起来构建区域感知提示，在整个生成过程中保持局部化上下文。本文的

贡献包括：

- 我们提出了一种无训练解码方法，该方法整合了多个区域级别的视角，并通过区域间一致性加权，以减少 LVLM 中幻觉现象。
- 我们引入了一种基于 JSD 的评分方案，以量化区域性响应的一致性并指导基于可靠性的融合。
- 我们设计了一种区域感知提示策略，以增强生成过程中的上下文基础，而无需修改模型参数。

## 2 相关工作

LVLM 幻觉和落地不足。大型视觉-语言模型（LVLMs）经常会“幻想”——生成不受视觉输入支持的文本（**??**）。这通常源于视觉落地的缺陷，这是由于各种因素导致的，如数据偏见或视觉与语言组件之间的连接不良（**????**）。尽管标准的注意力机制（**?**）和基于区域的分析（**??**）旨在改善落地，LVLMs 仍然难以可靠地整合来自多个可能存在冲突的图像区域的信息，或评估它们之间的一致性。

高级解码和融合策略的局限性。旨在改善 LVLM 输出的高级策略同样存在局限性。Chain-of-Thought (CoT) 提示（**???**），虽然可以增强推理能力，但其步骤未必能与图像视觉上对齐，并且可能对设置敏感或资源密集。许多不依赖训练的矫正或对比解码方法（例如，（**???????**））难以捕捉细粒度的局部图像特征，可能漏掉有效信息，导致不良的基础对齐能力。此外，当尝试融合来自多个来源的信息时——这一概念在语言建模任务中已被证明具有一致性好处（**??**）——LVLM 在视觉区域面临特殊挑战。常见的简单聚合技术（如平均）通常不足以处理视觉数据，因为不同区域的可信度可能相差悬殊。因此，一个核心挑战仍然存在：开发合理的方法，以根据一致性恰当地衡量和融合来自不同、上下文理解的视觉区域的证据。

## 3 动机

### 3.1 全局解码遗漏局部证据

为了探究视觉定位能力是否与 LVLM 中的幻觉相关，我们分析了它们在问答过程中的注意模式。图 2 展示了一个场景，其中模型被问到：“图像中是否有笔记本电脑?”当使用完整图像作为输入时，尽管图中有笔记本电脑，模型输出为“没有”。注意力图显示，模型将焦点分布在不相关的区域。

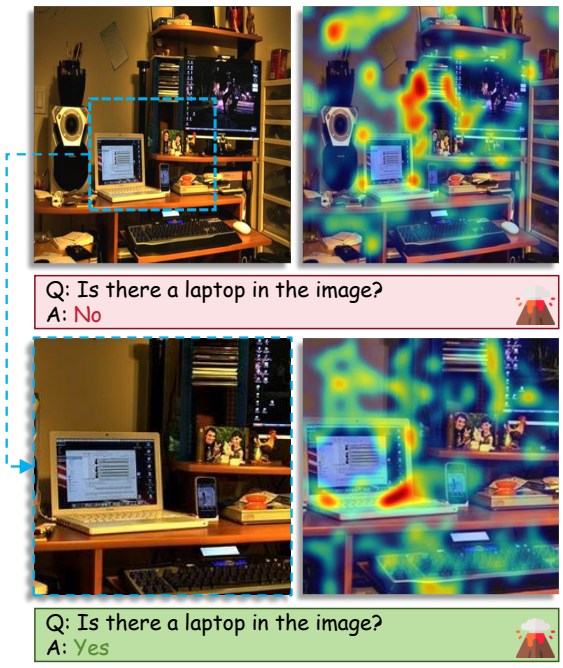


Figure 2: LVLM 交叉注意力模式用于“图像中有笔记本电脑吗?”。（上）整个图像输入导致注意力分散和潜在错误。（下）专注于笔记本电脑的裁剪图像输入使注意力集中并提高准确性。

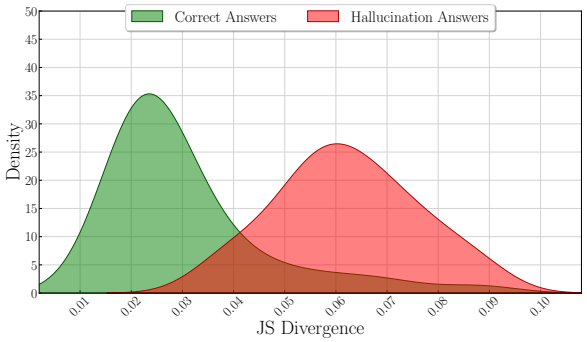


Figure 3: 正确与幻觉 LVLM 响应的 JS 散度的密度分布，表明较低的 JSD 与较高的事实准确性相关。

相比之下，当图像根据高关注度裁剪到一个显著区域（桌面区域）时，模型更专注于相关的证据，并正确地回答“是”。这种局部化的输入引导模型仅验证与查询相关的部分，避免分散其他图像部分的干扰。

这一比较揭示了全局解码的一个关键缺点：注意力容易在多个区域扩散，其中一些可能会误导模型。在杂乱或模糊的场景中，这通常会导致事实性错误。裁剪区域有助于缩小模型的视觉焦点，通过加强局部基础来提高答案的质量。

### 3.2 区域一致性体现可靠性

然而，并不是所有的区域——甚至是由注意力引导的区域——都是同样可靠的。有些可能在视觉上模棱两可或者在上下文中具有误导性。

为了避免对单个（可能具有误导性）区域过于自信，我们采用了一种自我一致性的方法：比较多个区域级别的响应。当不同的视角对一个答案达成一致时，该答案更有可能是值得信赖的。

为验证这一点，我们使用 LLaVA-1.5 在 3,000 个 MSCOCO 验证样本上进行实验，这些样本带有来自 POPE 基准的虚构标签。对于每个图像-问题对，我们从注意力引导的图像补丁中生成若干个响应，同时生成一个来自完整图像的响应。对于每个裁剪区域的输出，我们计算其响应分布与所有区域的平均响应分布之间的 Jensen-Shannon 散度（JSD）。

我们的结果（如图 3 所示）显示了明确的区别：对于幻想答案，JS 散度倾向于集中在 0.06-0.07 之间，而对于正确答案，它的中心大约在 0.02。这表明特定区域响应与整个区域平均值之间的 JSD 有效反映了该区域响应的幻想程度：较低的 JSD 与较少的幻想相关。这一发现与自我一致原则 (**?**) 相呼应，其中多个不同输出之间的一致性通常表明更高的质量和可靠性。受此启发，我们设计了一种解码策略：一种聚合多个区域证据并检查它们一致性的机制，可以帮助改善模型预测的稳健性和事实基础。

## 4 方法

### 4.1 概述

如第 3 节所述，当前 LVLM 解码方法通常因缺乏多视角一致性检查而导致幻觉。

为了解决这个问题，我们提出了多区域融合解码（MRFD），这是一种无需训练的解码策略，通过以下方式提高答案的可靠性：1）注意力引导的区域选择：根据注意力图选择显著区域；2）基于 JSD 加权的多区域分析：使用詹森-香农散度评估各区域之间的一致性；3）基于一致性的融合解码：融合根据其相互一致性加权的预测。完整框架如图 4 所示。

基于在第 3 节中提到的多视角验证方法，我们的 MRFD 框架中的第一步是识别并隔离输入图像中的多个信息丰富区域，以便进行集中和独立的分析。

利用交叉注意力进行区域识别。为了识别与输入查询最相关的图像区域，我们利用 LVLM 架构中已经存在的交叉注意力机制。给定原始图像 $I_0$ 和文本查询 $x$，我们从 LVLM 的交叉注意力层计算注意力矩阵 $A$。设 $H_T \in \mathbb{R}^{n\times d}$ 表示来自文本输入的隐藏状态序列，$H_V \in \mathbb{R}^{m\times d}$ 表示来自视觉输入的隐藏状态序列，其中 $n$ 和 $m$ 表示序列长度，$d$ 表示隐藏维度。

跨注意力权重矩阵 $A \in \mathbb{R}^{n\times m}$，通过使用缩放点积注意力 (**?**) 来计算，以捕捉文本和视觉模态之间的对齐：

$$\boldsymbol{A} = \text{softmax}\left(\frac{(H_T W^Q)(H_V W^K)^\top}{\sqrt{d_k}}\right) \quad (1)$$

，其中 $W^Q \in \mathbb{R}^{d\times d_k}$ 和 $W^K \in \mathbb{R}^{d\times d_k}$ 是查询和键的投影矩阵，$d_k$ 是它们的维度。

为了获得视觉焦点的整体图，我们在文本维度上汇总这些注意力权重：

$$a_j = \frac{1}{n}\sum_{i=1}^{n} A_{ij} \quad \text{for } j = 1, \ldots, m \quad (2)$$

这产生了一个聚合的注意力向量 $a \in \mathbb{R}^{1\times m}$。假设视觉隐藏状态对应于补丁的空间网格 ( $d' \times d' = m$ )，我们将这个向量重塑为二维空间注意力图：

$$\hat{A} = \text{reshape}(\boldsymbol{a}, (d', d')) \quad (3)$$

这个空间注意力图 $\hat{A}$ 直观地突出显示了模型认为与输入查询最相关的区域。利用模型自身的注意力模式，我们可以直接识别在回答查询时图像的哪些部分被主要考虑。

选择显著区域。使用空间注意力图 $\hat{A}$，我们在原始图像中识别出前 K 个最显著的区域 $R_1, R_2, ..., R_K$。为了确保多样性和全面覆盖，我们采用积分图像方法，来高效搜索 K 个不重叠或最小重叠且最大化总注意力分数的区域。对于实际的实现，我们将这些区域限制为矩形边界框，并设置任意两个区域之间的最大重叠（IoU）阈值，以确保所选视角的多样性。

对于每个选定的区域 $R_k$，我们裁剪原始图像以获得一个聚焦的子图像 $v_k$。这一过程产生了一组 K+1 个视觉输入：原始完整图像 $v_0 = I_0$ 和 K 个裁剪区域图像 $\{v_1, v_2, ..., v_K\}$。这些裁剪区域代表了相对于输入查询最有信息量的图像部分，这些部分是由模型自身的注意机制确定的。

### 4.2 基于 JSD 加权的多区域分析

选择了显著区域后，我们现在详细介绍其独立分析和基于 JSD 的加权机制。这种方法对于评估区域间一致性至关重要（章节 3.2 ）。对于每个选择的区域 $R_k$，我们通过处理相应的裁剪子图像 $v_k$ 通过 LVLM 生成初步分析：

$$r_k = \text{LVLM}(v_k, x) \quad (4)$$

，其中 $r_k$ 表示仅考虑来自区域 $k$ 的视觉信息时 LVLM 的响应。我们还使用原始完整图像

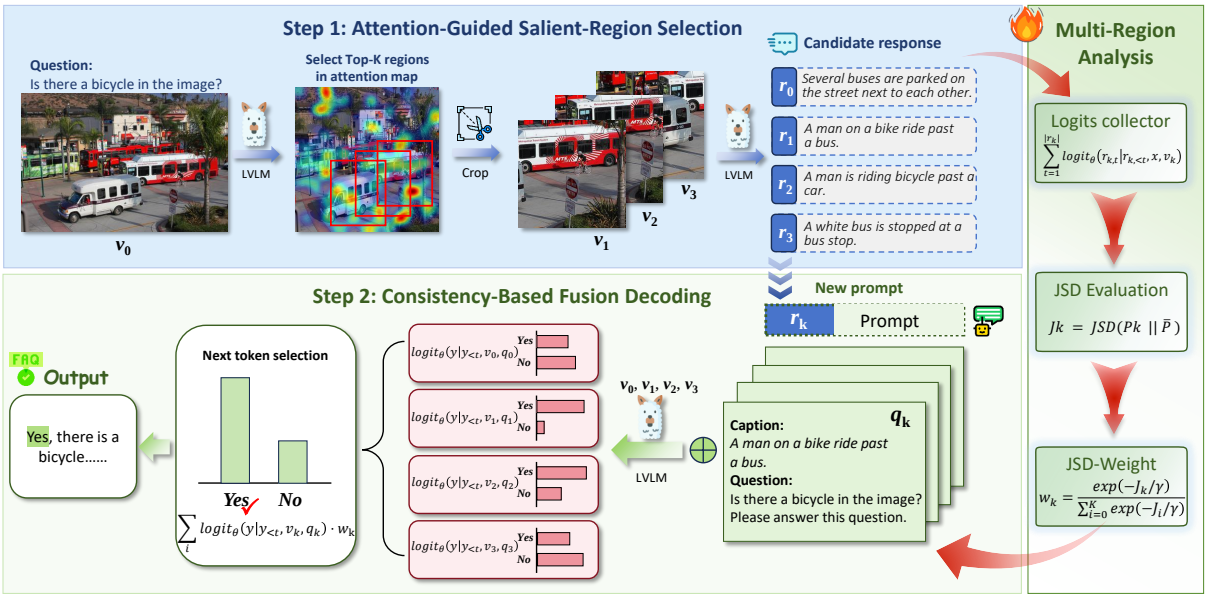


Figure 4: 多区域融合解码（MRFD）的总体框架：步骤 1 使用注意力选择和裁剪显著区域（$v_k$），为每个区域生成候选响应（$r_k$），并为每个响应计算基于 JSD 的一致性权重（$w_k$）。步骤 2 通过候选响应和原始提示形成每个区域的新输入。它们都被并行处理，在并行解码过程中使用权重 $w_k$ 融合各区域的逻辑以选择输出标记。

$v_0$ 生成响应 $r_0$。结果集合 $\{r_0, r_1, ..., r_K\}$ 提供了对查询的多重视角，每个视角都基于图像的不同显著区域。

为了评估来自每个区域的信息的可靠性，我们使用詹森-香农散度（JSD）来测量区域特定分析之间的一致性。JSD 提供了一种对称的相似性度量，用于衡量概率分布之间的相似性，定义为：

$$JSD(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{5}$$

，其中 $P$ 和 $Q$ 是两个概率分布，$M = \frac{1}{2}(P+Q)$，以及 $D_{KL}$ 是 Kullback-Leibler 散度。

我们为每个初始分析序列 $r_k$ 推导出一个代表性词汇分布 $P_k$，该分布是通过对生成过程中计算的下一个令牌概率分布进行平均得出的：

$$P_k = \text{softmax}\left(\frac{1}{|r_k|}\sum_{t=1}^{|r_k|}\text{logit}_\theta(y|r_{k,<t}, x, v_k)\right) \tag{6}$$

，其中 $|r_k|$ 是 $r_k$ 的长度，而 $\text{logit}_\theta$ 表示模型的输出 logits。这个分布 $P_k$ 捕捉了当模型考虑区域 $k$ 时的整体令牌预测模式。然后我们计算所有区域的平均分布：

$$\overline{P} = \frac{1}{K+1}\sum_{i=0}^{K} P_i \tag{7}$$

。对于每个区域 $k$，我们计算其与该平均分布的偏差：

$$J_k = JSD(P_k||\bar{P}) \tag{8}$$

。这些偏差分数衡量每个区域的预测与共识的偏离程度。较低的 $J_k$ 意味着区域 $k$ 提供的信息与其他区域更为一致，表明其可靠性更高。

最后，我们将这些散度得分转换为带有温度参数 $\gamma$ 的权重：

$$w_k = \frac{\exp(-J_k/\gamma)}{\sum_{i=0}^{K}\exp(-J_i/\gamma)} \tag{9}$$

温度参数 $\gamma$ 控制权重分布的锐度——较小的 $\gamma$ 会增加权重之间的对比，而较大的 $\gamma$ 则会导致权重更均匀。

通过为每个视觉输入 $v_k$（k = 0...K）计算一致性为基础的权重 $w_k$，我们现在通过以强调一致视觉证据的方式融合下一个标记预测来执行最终的解码步骤。

受链式思维（CoT）提示 (**?**) 在结构化推理中的有效性启发，我们采用其核心原则来增强我们多区域融合框架内的事实依据。对于每个区域 $k$，我们通过将原始问题 $x$ 和相应区域的初步分析 $r_k$ 连接，构建一个固定的区域感知提示 $q_k$。这一提示 $q_k$ 封装了从 $v_k$ 衍生出的全局查询和本地化的初步分析。它在解码过程中为区域 $k$ 提供静态的、丰富的上下文，提供特定于区域的基础信息，而无需动态更新提示。

在自回归解码过程中，在每个步骤 $t'$，LVLM 生成下一个标记的对数 $\ell_k^{(t')}$，其基于当前的部

分输出序列 $y_{<t'}$ ，视觉输入 $v_k$ ，以及固定的区域感知提示 $q_k$ ：

$$\ell_k^{(t')} = \text{logit}_\theta(y|y_{<t'}, v_k, q_k) \tag{10}$$

这些对数表示每个区域的视角下的预测，其基于预先计算好的分析。我们使用预先计算的一致性权重 $w_k$ 聚合它们：

$$\ell_{\text{fused}}^{(t')} = \sum_{k=0}^{K} w_k \cdot \ell_k^{(t')} \tag{11}$$

通过 softmax 获得下一个标记的最终概率分布：

$$P_{\text{fused}}^{(t')}(y) = \text{softmax}(\ell_{\text{fused}}^{(t')}) \tag{12}$$

然后从这个融合分布中选择下一个标记 $\hat{y}_{t'}$（例如，采样），并以自回归方式重复此过程直至完成。

# 5 实验

## 5.1 实验设置

评估的 LVLMs。我们在两个具有代表性的开源 LVLMs 上评估了多区域融合解码（MRFD）：即 LLaVA-1.5-7B (**?**) 和 InstructBLIP-7B (**?**) 。这些模型采用了不同的视觉-语言接口——LLaVA-1.5 中的直接投影和 InstructBLIP 中的基于查询的编码（Q-Former (**?**) ）——以便评估 MRFD 的通用性。MRFD 被应用为一种在冻结模型上进行的、无须训练的解码时间过程。架构细节在附录 **??** 中提供。

基线。作为一个基本基线，常规解码从模型的软最大化输出概率中采样标记。这些包括对比方法，如 VCD (**?**) 和 M3ID (**?**) ，增强稳健性的方法，例如 RITUAL (**?**) ，以及使用校正或反馈的方法，如 DeGF (**?**) 和 Woodpecker (**?**) 。我们还包括与其他相关方法的比较：HALC (**?**) 和 OPERA (**?**) 。这些基线的性能基于我们使用公开可用代码的重新实现。在附录 A 中提供了每个基线方法的详细描述。

实施细节。在所有实验中，我们的多区域融合解码（MRFD）方法选择 $K = 3$ 个显著区域，并对基于 JSD 的加权（方程 9 ）采用 $\gamma = 0.02$ 的温度。我们在解码的两个阶段都使用多项式采样。其他详细设置在附录 **??** 中提供。

## 5.2 数据集和基准测试

我们在多个基准测试上评估 MRFD，涵盖幻觉检测和一般视觉语言理解。下面我们简要描述关键数据集；完整细节和评估指标见附录 B 。

POPE (**?**)：一个针对物体存在幻觉的是/否问答基准，由 MSCOCO、A-OKVQA 和 GQA 构建，采用了各种负采样策略。
CHAIR (**?**)：通过在 MSCOCO 图像上使用 CHAIRi 和 CHAIRs 评分测量图像标题中的虚构物体提及。
MME (**?**)：我们专注于 MME-Hallucination 子集，该子集测试物体、数量、位置和属性幻觉。

## 5.3 结果与讨论

POPE 的结果。在表 1 中，我们将我们的 MRFD 方法与其他基准方法在 POPE 基准测试 (**?**) 下的三个不同负采样设置（随机、流行、对抗）中进行比较，涵盖 LLaVA-1.5 和 InstructBLIP。结果显示，MRFD 在这两种 LVLM 上的表现始终优于其他解码方法，在所有六种配置中均取得领先的 F1 分数，与第二好的方法相比，准确度提升高达 2.44 %，精确度提升 6.52 %，F1 分数提升 3.05 %。这表明 MRFD 的核心策略——通过交叉注意力识别多个显著区域，使用 JSD 加权评估其响应一致性，并融合其上下文丰富的预测——使得 LVLM 能够更好地在相关视觉证据中定位响应，从而有效解决对象幻觉问题。此外，尽管大多数解码方法在不存在情况下往往过于自信，但 MRFD 固有的一致性验证似乎促使其响应更加谨慎和准确，这可以从其在挑战性的对抗环境下（例如 LLaVA-1.5 上为 76.16 和 InstructBLIP 上为 74.26）的强大精确度中得到证明，突出其在过滤假阳性和抑制虚假信息方面的增强能力。POPE 的详细结果附在附录 **??** 中。

CHAIR 上的结果。我们使用 CHAIR 基准来评估 MRFD 在开放式图像标注中减少对象幻觉的效果 (**?**) ，报告了 LLaVA-1.5 和 InstructBLIP 在表 2 中的 CHAIRs (Cs) 和 CHAIRi (Ci) 分数（越低越好）。MRFD 在两个 LVLMs 上始终实现了最先进的性能，显著超越了强大的 DeGF 基准，在 LLaVA-1.5 上相对 CHAIR 分数下降最高可达 16.4 %，在 InstructBLIP 上下降 20.8 % 。鉴于图像标注的 CHAIR 任务需要全面关注多样的局部细节，MRFD 的卓越性能凸显了其在稳健的多区域分析方面的高级能力，以及有效整合来自不同视觉片段的关键信息，从而生成更具事实依据的图像标注。CHAIR 的详细结果附录于附录 C.1 。一些定性示例展示在附录 E 。

在 MME-Hallucination 上的结果。我们在 MME 幻觉子集中评估 MRFD (**?**) ，该子集评估 LVLMs 中包括对象级别（存在）、属性级别（计数、颜色）和关系级别（位置）理解的多种幻觉类型。如图 5 所示（展示了 LLaVA-1.5

| Setting | Method | LLaVA-1.5 (?) | | | InstructBLIP (?) | | |
|---|---|---|---|---|---|---|---|
| | | Acc. ↑ | Prec. ↑ | F1 ↑ | Acc. ↑ | Prec. ↑ | F1 ↑ |
| Random | Regular | 82.42 | 78.30 | 83.67 | 79.85 | 80.33 | 83.45 |
| | VCD (?) | 84.69 | 80.30 | 85.85 | 84.47 | 83.07 | 84.81 |
| | M3ID (?) | 85.46 | 81.54 | 86.42 | 85.32 | 83.51 | 85.58 |
| | RITUAL (?) | 86.71 | 82.84 | 87.51 | 87.12 | 85.64 | 87.23 |
| | DeGF (?) | 87.79 | 86.33 | 88.08 | 87.21 | 89.01 | 86.70 |
| | MRFD | 88.15 | 88.91 | 88.23 | 88.03 | 88.74 | 88.01 |
| Popular | Regular | 76.57 | 71.23 | 79.56 | 75.17 | 70.90 | 77.54 |
| | VCD (?) | 77.30 | 71.61 | 80.57 | 78.12 | 73.77 | 80.10 |
| | M3ID (?) | 78.66 | 73.09 | 81.45 | 78.32 | 73.75 | 80.30 |
| | RITUAL (?) | 79.75 | 74.55 | 82.31 | 78.40 | 73.63 | 80.55 |
| | DeGF (?) | 81.94 | 78.33 | 83.31 | 80.37 | 78.84 | 81.24 |
| | MRFD | 81.99 | 78.81 | 83.29 | 82.17 | 79.52 | 83.15 |
| Adversarial | Regular | 71.09 | 65.77 | 75.93 | 71.02 | 66.58 | 74.70 |
| | VCD (?) | 71.13 | 65.28 | 76.37 | 73.07 | 68.50 | 76.36 |
| | M3ID (?) | 72.10 | 66.27 | 76.87 | 72.97 | 67.92 | 76.55 |
| | RITUAL (?) | 71.87 | 66.17 | 76.88 | 73.06 | 67.83 | 76.91 |
| | DeGF (?) | 76.13 | 71.50 | 79.01 | 75.96 | 73.45 | 77.36 |
| | MRFD | 77.99 | 76.16 | 79.22 | 77.72 | 74.26 | 79.72 |

Table 1: POPE 基准测试的结果。更高的（↑）准确率、精确率和 F1 值表示更好的性能。最佳结果以粗体显示，第二好的结果以下划线标示。

| Method | LLaVA-1.5 | | InstructBLIP | |
|---|---|---|---|---|
| | Cs ↓ | Ci ↓ | Cs ↓ | Ci ↓ |
| Regular | 26.2 | 9.4 | 31.2 | 11.1 |
| VCD | 24.4 | 7.9 | 30.0 | 10.1 |
| M3ID | 21.4 | 6.3 | 30.8 | 10.4 |
| RITUAL | 22.4 | 6.9 | 26.6 | 8.9 |
| Woodpecker | 24.9 | 7.5 | 31.2 | 10.8 |
| HALC | 21.7 | 7.1 | 24.5 | 8.0 |
| DeGF | 18.4 | 6.1 | 24.0 | 7.7 |
| Ours (MRFD) | 14.1 | 5.0 | 21.3 | 6.1 |

Table 2: 针对描述生成在 CHAIR 基准测试上的结果。我们将新标记的最大数量限制在 64 个。较低的↓、CHAIRs (Cs) 和 CHAIRi (Ci) 表示幻觉更少。最佳结果以粗体显示，第二佳结果以下划线标注。

和 InstructBLIP 的结果），我们的 MRFD 方法始终表现出强劲的性能，在与常规解码和其他先进基线的比较中，在大多数这些幻觉类别中在被评估的 LVLMs 上实现了领先或最先进的结果。跨越一系列具有挑战性的幻觉类型的这种稳健表现，强调了 MRFD 的核心机制——多区域分析结合一致性驱动融合——有效增强了细粒度的视觉理解和整体的事实一致性。MME-Hallucination 的详细结果附录于 **??**

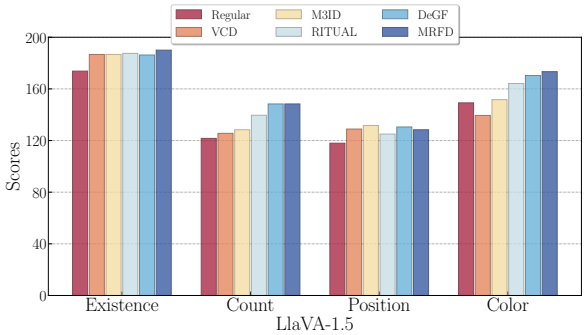


Figure 5: MME 在一个具有不同解码策略的幻觉子集上的实验结果。

效率讨论。尽管采用了多步推理，MRFD 仍保持着竞争力的效率，尤其是在与其他采用更复杂迭代或反馈机制的先进幻觉缓解技术相比时。详细内容见附录 **??** 。

我们进行消融研究，以评估关键组件和超参数在 MRFD 中的影响，并在 POPE-MSCOCO 数据集上使用 LLaVA-1.5。

首先，我们评估 MRFD 核心设计的贡献：基于 JSD 的一致性加权（CW）、融合提示（FP）和注意引导的区域选择（RS）。我们将完整的 MRFD 与三个主要变体进行对比测试，这些变体中这些组件被单独改变：（1）没有 CW 的

| Model Variants | Acc. ↑ | Prec. ↑ | F1 Score ↑ |
|---|---|---|---|
| MRFD (Full) | 86.50 | 88.11 | 86.21 |
| w/o CW | 83.76 | 84.27 | 83.74 |
| w/o FP | 82.87 | 83.95 | 82.70 |
| w/o RS | 82.77 | 83.94 | 82.58 |
| Regular | 80.57 | 78.84 | 81.59 |

Table 3: 在三种设置的平均值下，对 POPE-COCO 的不同模型变体进行消融研究。

MRFD，采用统一的融合权重；(2) 没有 FP 的 MRFD，仅使用原始问题进行区域解码；(3) 没有 RS 的 MRFD，仅通过后续管道阶段处理全局图像。如表 3 所示，完整的 MRFD 实现了 86.21 的 F1 得分。去掉 CW 使 F1 表现下降了 2.94 %，强调了动态的一致性加权的重要性。省略 FP 使 F1 下降了 3.93 %，突出了丰富区域上下文的价值。跳过 RS 导致 F1 最大幅度下降 4.07 %，强调了稳健的多区域分析是基础。尽管有这些影响，所有三种去除组件的 MRFD 变体仍然优于常规解码（81.59 F1），而完整的 MRFD 配置展示了最强的协同效益。

其次，我们分析了对 JSD 加权温度 $\gamma$（方程 9 ）的敏感性。如图 **??** 所示（$K = 2, 3, 4$ 的 POPE COCO 精度），性能在 $\gamma = 0.02$ 时达到峰值（$\approx$ 时的精度为 87.8 %）。较高的 $\gamma$ 值（$\geq 0.08$）导致权重更加均匀，相应地降低了精度，接近于未加权融合。相反，非常低的 $\gamma$ 值（$\approx 0.01$）由于极端权重化导致过度依赖于少数区域，这可能对鲁棒性产生负面影响，尤其是对较大的 $K$ （例如，$K = 4$ 的性能下降）。因此，采用 $\gamma = 0.02$ 来提供最佳平衡。更多细节见附录 D 。

## 6 结论

我们提出了多区域融合解码（MRFD），这是一种无需训练的方法，通过模拟自我一致性多视图验证来提高 LVLM 的可靠性。MRFD 通过注意力机制识别显著区域，使用詹森-香农散度（JSD）估计其可靠性，并通过一致性加权，以链式思维启发的提示融合区域级预测。实验表明，MRFD 能够有效减少幻觉现象并提高各种 LVLM 及基准的真实性，提供了一种简单而可靠的解码策略。

## 7

局限性 MRFD 依赖于底层 LVLM 生成的注意力图的质量和可解释性，这可能会因架构和输入条件的不同而有所变化。区域间一致性表明事实正确性的假设在经验上成立，但可能无法推广到所有推理任务，特别是那些需要抽象或常识性推理的任务。此外，尽管 MRFD 在基于图像的任务中提升了事实性，但尚未在更广泛的多模态背景中（如视频或基于对话的定位）进行评估，这些背景可能涉及更复杂的时间或对话依赖关系。

我们的多区域融合解码（MRFD）工作旨在提高 LVLMs 的可靠性和事实基础，从而通过减少视觉幻觉来助力于更可信赖的人工智能系统。我们相信这对需要准确事实的应用具有积极影响。

本文件由张悦、Ryan Cotterell 和 Lea Frermann 改编自先前 ACL 和 NAACL 会议的格式文件，包括由 Steven Bethard、Ryan Cotterell 和 Rui Yan 用于 ACL 2020 的文件，Douwe Kiela 和 Ivan Vulić 用于 ACL 2019 的文件，Stephanie Lukin 和 Alla Roskovskaya 用于 NAACL 2019 的文件，Shay Cohen、Kevin Gimpel 和魏露用于 ACL 2018 的文件，Margaret Mitchell 和 Stephanie Lukin 用于 NAACL 2018 的文件，Jason Eisner 为（NA）ACL 2017/2018 提供的 Bib TEX 建议，Dan Gildea 和 Min-Yen Kan 用于 ACL 2017 的文件，Margaret Mitchell 用于 NAACL 2017 的文件，Maggie Li 和 Michael White 用于 ACL 2012 的文件，Jing-Shin Chang 和 Philipp Koehn 用于 ACL 2010 的文件，Johanna D. Moore、Simone Teufel、James Allan 和 Sadaoki Furui 用于 ACL 2008 的文件，Hwee Tou Ng 和 Kemal Oflazer 用于 ACL 2005 的文件，Eugene Charniak 和 Dekang Lin 用于 ACL 2002 的文件，以及包括 John Chen、Henry S. Thompson 和 Donald Walker 在内的多人撰写的早期 ACL 和 EACL 格式文件。额外元素来自国际人工智能联合会议和计算机视觉与模式识别会议的格式说明。

我们在三个具有代表性的开源大型视觉语言模型（VLMs）上评估了我们提出的多区域融合解码（MRFD）框架：LLaVA-1.5-7B、Qwen-VL-7B 和 InstructBLIP-7B。对于视觉编码器，LLaVA-1.5 使用了 ViT-L-336px，从 CLIP-L/14-336px 预训练得到的。InstructBLIP 使用了从 EVA-CLIP 预训练得到的 ViT-g/14。Qwen-VL 使用了 Openclip ViT-bigG。关于语言模型模块，LLaVA-1.5 和 InstructBLIP 都使用了 Vicuna-7B。

视觉-语言接口因评估的模型而异。LLaVA-1.5 直接使用 MLP 层投影视觉特征。相反，InstructBLIP 采用 Q-Former (**?**)，通过 32 个可学习查询处理视觉特征，以生成 LLM 的标准化输入。我们对 MRFD 的评估涵盖了这些不同的策略——直接的 MLP 投影和基于查询的瓶颈（Q-Former）——以展示其在无论采用何种连接模块的情况下，广泛减轻幻觉的有效性。

## A 基线

我们将我们的 MRFD 方法的性能与几种最近设计用于减轻幻觉的无需训练的解码方法进行比较。这些关键的基线方法及其核心方法如下所述：

我们报告了这些基线的性能，这是基于我们使用其发布的代码库进行重新实现的结果。

在所有使用 MRFD 的实验中，我们设置区域数量 $K = 3$ 以获得裁剪后的图像，并设置基于 JSD 加权的温度参数 $\gamma = 0.02$ （方程 9）。在解码过程中，我们在两个阶段中都采用多项式采样。具体来说，在生成初始分析 $r_k$ 的第一步中，我们使用 0.7 的采样温度。在生成最终输出序列 $y$ 的第二步中，我们在融合 logits 后使用 0.1 的较低采样温度。

为了获得聚合的空间注意力图 $\hat{A}$ （方程 3），我们首先在相关层内对所有注意力头的注意力权重进行平均。对于 LLaVA-1.5，我们使用最后的交叉注意力层的注意力图。对于使用 Q-Former 的 InstructBLIP，我们识别具有最大聚合注意力评分的查询标记，并使用其对应的注意力图。地图 $\hat{A}$ 的空间维度 $d'$ 对应于视觉补丁的网格大小（$m = d' \times d'$），其中 LLaVA-1.5（ViT-L-336px）为 $24 \times 24$，而 InstructBLIP 使用的 ViT-g/14 为 $16 \times 16$。

为了在 $\hat{A}$ 的基础上有效识别出前 K 个显著区域 $\{R_1, ..., R_K\}$，我们采用了一种积分图像方法 (**?**)，以快速计算任意矩形边界框内注意力分数的总和。我们搜索使这些注意力分数总和最大的 K 个不重叠或最小化重叠的矩形区域。为确保选择区域的多样性，我们对任意一对所选边界框 $R_i$ 和 $R_j$ （$i \neq j$）实施最大 40

% 的交并比 (IoU) 限制。

## B 数据集和基准测试

我们在一组多样化的基准测试中评估我们的 MRFD 框架，目标是幻觉检测和通用视觉-语言能力。

- POPE (**?**)：POPE（基于轮询的对象探测评估）是用于评估 LVLMs 中对象存在幻觉的广泛使用的基准。它通过关于特定对象是否存在的是/否问题（例如，“图像中是否有 {对象}?”）向模型提问。基准数据被分为三个主要子集，分别来自 MSCOCO (**?**)、A-OKVQA (**?**) 和 GQA (**?**)。每个子集进一步基于三个针对不存在对象的负采样策略进行划分：随机、热门和对抗，这三者在难度上有所不同。对于评估，我们报告包括准确性、精确率、召回率和 F1 得分在内的标准指标。

- MME (**?**)：MME 作为评估整体 LVLM 感知和认知的综合基准。我们的评估特别利用了 MME-Hallucination 子集，该子集旨在评估一系列常见的幻觉类型。这些被分类为对象级评估，如存在性和计数，以及属性级评估，如对象位置和颜色。该子集中的问题通常是是/否查询。我们依据官方基准协议报告分数，这通常涉及结合准确性度量，以反映问题级和图像级的正确性。

- CHAIR (**?**)：CHAIR（Caption Hallucination Assessment with Image Relevance）基准测试在开放式图像描述的背景下量化对象幻觉。我们提示 LVLMs 为图像生成描述性字幕，按照先前的工作 (**?**) 使用 MSCOCO (**?**) 验证集中的 500 个随机选择的图像。然后，将生成的字幕与图像内的真实对象进行比较，以计算 CHAIRi（实例级）和 CHAIRs（类别级）得分，其中较低的得分表示较少的幻觉。

- MMBench (**?**)：MMBench 通过精心策划的多项选择题评估广泛的多模态能力，这些问题涵盖了各种认知维度和技能。在此基准测试中，我们遵循官方评估协议并报告整体准确率得分。

## C 实验的详细结果

### C.1 CHAIR 的详细结果

CHAIR 的详细结果显示在表 4 和表 5 中，报告了 LLaVA-1.5 和 InstructBLIP 的 CHAIR (Cs) 和 CHAIR (Ci) 分数（数值越低越好）。MRFD 始终在这两个 LVLMs 上实现最新的性能。

| Method | LLaVA-1.5 | | InstructBLIP | |
|---|---|---|---|---|
| | Cs ↓ | Ci ↓ | Cs ↓ | Ci ↓ |
| Regular | 58.0 | 17.7 | 61.0 | 18.2 |
| VCD | 58.2 | 16.7 | 63.0 | 18.6 |
| M3ID | 56.8 | 16.1 | 65.8 | 19.9 |
| RITUAL | 51.0 | 15.1 | 50.4 | 15.3 |
| DeGF | <u>49.8</u> | <u>14.7</u> | <u>49.8</u> | <u>15.1</u> |
| Ours (MRFD) | 39.0 | 11.0 | 38.6 | 11.3 |

Table 5: 在 CHAIR 基准测试中生成字幕的结果。我们将新符号的最大数量限制为 256。较低的 ↓ CHAIRs（Cs）和 CHAIRi（Ci）表示幻觉较少。最好的结果是加粗显示，次佳结果是下划线显示。

| Method | LLaVA-1.5 | | InstructBLIP | |
|---|---|---|---|---|
| | Cs ↓ | Ci ↓ | Cs ↓ | Ci ↓ |
| Regular | 55.0 | 16.3 | 57.0 | 17.6 |
| VCD | 54.4 | 16.6 | 60.4 | 17.8 |
| M3ID | 56.6 | 15.7 | 62.2 | 18.1 |
| RITUAL | 49.6 | 14.8 | <u>48.4</u> | 14.5 |
| Woodpecker | 57.6 | 16.7 | 60.8 | 17.6 |
| HALC | 51.0 | 14.8 | 53.8 | 15.7 |
| DeGF | <u>48.8</u> | <u>14.6</u> | 49.2 | <u>14.4</u> |
| Ours (MRFD) | 37.1 | 9.2 | 38.2 | 10.9 |

Table 4: CHAIR 基准测试的字幕生成结果。我们将新标记的最大数量限制为 128。较低的 (↓)CHAIRs (Cs) 和 CHAIRi (Ci) 表示幻觉较少。最佳结果用粗体表示，次佳结果用下划线标注。

在表格 7 和图 6 中，我们提供了关于 MME-Hallucination 基准 (**?**) 的详细结果，涵盖了 LLaVA-1.5 和 InstructBLIP。该表包括对象级别（存在、计数）和属性级别（位置、颜色）任务的分数，并在三个随机种子中平均得分。最佳结果以粗体显示，次优结果以下划线标出。

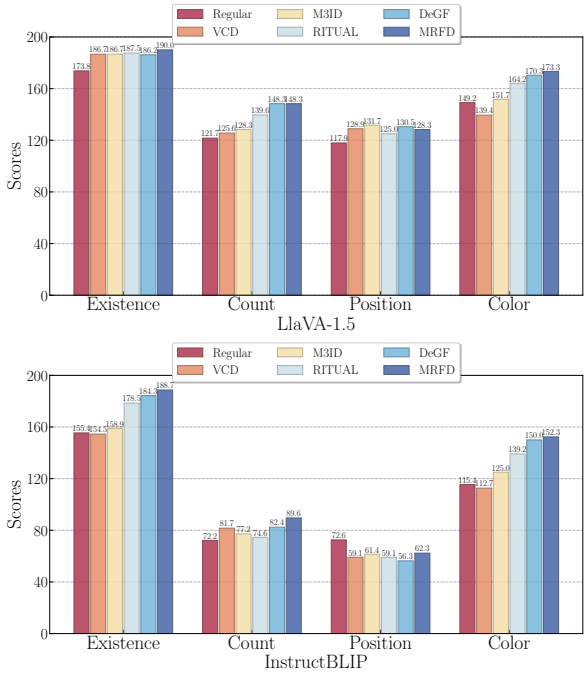


Figure 6: MME 在不同解码策略下的实验结果。

在表格 6 中，我们提供了关于 POPE 基准 (**?**) 的详细结果，适用于 LLaVA-1.5 和 InstructBLIP 的三种不同设置（随机、流行、对抗）。该表格包括每种方法的准确率（Acc.）、精确率（Prec.）、召回率和 F1 分数。最佳结果以粗体显示，次佳结果以下划线标记。

表 **??** 展示了 MRFD 在 CHAIR 基准上的效率（InstructBLIP，128 个最大 tokens，RTX 3090 GPU）。其多区域（$K = 3$）分析和融合过程导致了大约 2.96 × 的延迟增加（达到 10.21s）和大约 1.07 × 的峰值 GPU 内存增加，与标准多项式解码相比。因此，尽管 MRFD 比简单的方法如 VCD 或事后 Woodpecker 更具计算密集型，但它仍然比复杂的迭代或基于反馈的技术如 DeGF、OPERA 和 HALC 更加高效。鉴于 MRFD 在幻觉减少方面的最新技术（在比较方法中获得了最佳 Cs 得分 34.6），它在计算成本和效果之间提供了一个引人注目的平衡。

## D 详细消融研究

### D.1 组件消融

为了评估 MRFD 核心组件的贡献——注意力引导的区域选择（RS）、基于 JSD 的一致性加权（CW）和融合提示（FP）——我们进行了消融研究。POPE-COCO 上的结果（在所有设置下取平均，LLaVA-1.5）见表 8 。评估的模型变体为：

(a) MRFD（完整）：我们的完整方法，整合了所有三个组件。

(b) 无 CW：采用 RS 和 FP，但对来源于多个区域的预测使用统一权重进行融合，绕过基于 JSD 的一致性加权。

(c) 无 FP：使用 RS 和 CW，但忽略丰富的融合提示，只在每次区域解码中使用原始问题。

(d) 无 RS（全局影像 +FP）：绕过了注意力引导的区域选择，仅操作于全局影像。融合提示根据全局影像的初始响应构建。在这种单视图情况下，定义的多区域一致性加权（CW）机制不可适用或变得无意义（对单视图效果上为 1 的权重）。

(e) 仅 RS：利用注意力引导的区域，但有统一的融合权重，并且只有原始问题作为提示，单独隔离多区域视角带来的益处。

(f) 仅 FP（全局影像）：直接对全局影像解码应用融合提示（从全局影像的初始响应中获得），无需多区域选择或任何形式的一致性加权。

(g) 全局影像（标记层次投票）：使用原始提示操作于全局影像。CW 组件被调整为通过汇总单一全球视图隐性多样化的预测来执行标记层次投票，以测试标记层次的一致性益处。

(h) 常规：仅使用全局影像和原始问题的标准贪婪解码。

| | Components | | | Acc. ↑ | Prec. ↑ | F1 ↑ |
|---|---|---|---|---|---|---|
| | CW | FP | RS | | | |
| (a) | ✓ | ✓ | ✓ | 86.50 | 88.11 | 86.21 |
| (b) | | ✓ | ✓ | 83.76 | 84.27 | 83.74 |
| (c) | ✓ | | ✓ | 82.87 | 83.95 | 82.70 |
| (d) | ✓ | ✓ | | 82.77 | 83.94 | 82.58 |
| (e) | | | ✓ | 82.58 | 82.01 | 82.26 |
| (f) | | ✓ | | 81.96 | 81.36 | 82.14 |
| (g) | ✓ | | | 81.52 | 81.05 | 81.68 |
| (h) | | | | 80.57 | 78.84 | 81.59 |

Table 8: 在 POPE-COCO 上的不同模型变体的消融研究，基于三种设置的平均值。CW: 基于 JSD 的一致性加权，FP: 融合提示，RS: 区域选择。

为了评估不同解码策略对 MRFD 的影响，我们在与 LLaVA-1.5 的对抗设置下，使用 POPE-COCO 基准进行消融研究。结果总结在表格 9 中。我们比较了三种解码策略：低温采样、高

温采样和带 top-p 过滤的高温采样。最佳结果以粗体显示。

低温策略是我们实验中的默认设置，其中 $t$ 为 0.1。高温策略设置为 $t$ =0.5，增加了采样过程的随机性。高温 +top-p 策略结合了高温采样和 top-p 过滤，其中我们设置 $p$ =0.7 以保留概率质量的前 70 %。

| Decoding | Acc. ↑ | Prec. ↑ | F1 ↑ |
|---|---|---|---|
| Low temp. | 82.75 | 85.22 | 82.12 |
| High Temp. | 83.04 | 82.29 | 83.23 |
| High Temp.+Top P | 82.50 | 81.25 | 82.67 |

Table 9: 不同解码策略的消融研究。

## E 定性示例

为了提供对 MRFD 在缓解幻想和提高字幕准确性方面影响的更直观理解，本节展示了我们使用 LlaVA-1.5 在 CHAIR 基准上的实验中的定性示例。我们将由我们的多区域融合解码（MRFD）方法生成的字幕与标准规则基线生成的字幕进行比较。这些示例被选中用来突出方法在感知和描述视觉场景中的物体、属性及其关系时的差异。图 **??** 和图 7 展示了这样一种比较。

| Dataset | Setting | Method | LLaVA-1.5 | | | | InstructBLIP | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Acc. ↑ | Prec. ↑ | Recall ↑ | F1 ↑ | Acc. ↑ | Prec. ↑ | Recall ↑ | F1 ↑ |
| MS-COCO | Random | Regular | 83.13 | 81.94 | 85.00 | 83.44 | 83.07 | 83.02 | 83.26 | 83.08 |
| | | VCD | 87.00 | 86.13 | 88.18 | 87.15 | 86.23 | 88.14 | 83.73 | 85.88 |
| | | M3ID | 87.50 | 87.38 | 87.67 | 87.52 | 86.67 | 88.09 | 84.79 | 86.41 |
| | | RITUAL | 88.87 | 89.23 | 88.39 | 88.81 | 88.83 | 90.48 | 86.80 | 88.60 |
| | | DeGF | 89.03 | 91.20 | 86.41 | 88.74 | 88.83 | 93.73 | 82.42 | 87.71 |
| | | Ours | 89.50 | 92.55 | 85.94 | 89.12 | 88.52 | 93.83 | 82.61 | 87.86 |
| | Popular | Regular | 81.17 | 78.28 | 86.26 | 82.08 | 77.00 | 73.82 | 83.68 | 78.44 |
| | | VCD | 83.10 | 79.96 | 88.34 | 83.94 | 80.07 | 77.67 | 84.39 | 80.89 |
| | | M3ID | 84.30 | 81.58 | 88.63 | 84.95 | 80.97 | 77.93 | 86.19 | 81.85 |
| | | RITUAL | 85.83 | 84.17 | 88.26 | 86.17 | 81.97 | 78.90 | 87.26 | 82.87 |
| | | DeGF | 86.63 | 87.75 | 84.86 | 86.28 | 82.73 | 84.02 | 80.27 | 82.10 |
| | | Ours | 87.24 | 86.56 | 88.22 | 87.38 | 83.69 | 85.22 | 81.58 | 83.36 |
| | Adversarial | Regular | 77.43 | 73.31 | 86.26 | 79.26 | 74.60 | 71.26 | 82.46 | 76.45 |
| | | VCD | 77.17 | 72.18 | 88.40 | 79.47 | 77.20 | 74.29 | 83.19 | 78.49 |
| | | M3ID | 78.23 | 73.51 | 88.28 | 80.22 | 77.47 | 73.68 | 85.48 | 79.14 |
| | | RITUAL | 78.80 | 74.43 | 87.74 | 80.54 | 78.73 | 74.57 | 87.21 | 80.39 |
| | | DeGF | 81.63 | 80.59 | 83.33 | 81.94 | 80.30 | 80.90 | 79.33 | 80.11 |
| | | Ours | 82.75 | 85.22 | 79.25 | 82.12 | 82.49 | 83.14 | 81.51 | 82.32 |
| A-OKVQA | Random | Regular | 81.90 | 76.63 | 91.78 | 83.53 | 80.63 | 76.82 | 87.75 | 81.92 |
| | | VCD | 83.83 | 78.05 | 94.13 | 85.34 | 84.20 | 80.90 | 89.54 | 85.00 |
| | | M3ID | 84.67 | 79.25 | 93.94 | 85.97 | 85.43 | 81.77 | 91.20 | 86.23 |
| | | RITUAL | 85.17 | 79.79 | 94.21 | 86.40 | 87.13 | 83.92 | 91.87 | 87.71 |
| | | DeGF | 86.93 | 84.28 | 90.80 | 87.42 | 87.40 | 87.67 | 86.86 | 87.26 |
| | | Ours | 87.13 | 87.56 | 86.76 | 87.16 | 88.33 | 86.81 | 90.40 | 88.57 |
| | Popular | Regular | 75.07 | 68.58 | 92.53 | 78.77 | 75.17 | 70.15 | 87.60 | 77.91 |
| | | VCD | 76.63 | 69.59 | 94.59 | 80.19 | 78.63 | 73.53 | 89.46 | 80.72 |
| | | M3ID | 77.80 | 70.98 | 94.07 | 80.91 | 78.80 | 73.38 | 90.39 | 81.00 |
| | | RITUAL | 78.83 | 71.99 | 94.37 | 81.68 | 78.73 | 72.83 | 91.68 | 81.17 |
| | | DeGF | 80.90 | 75.68 | 91.05 | 82.66 | 81.47 | 78.61 | 86.47 | 82.35 |
| | | Ours | 80.99 | 76.68 | 89.28 | 82.51 | 83.19 | 78.74 | 91.00 | 84.43 |
| | Adversarial | Regular | 67.23 | 61.56 | 91.81 | 73.70 | 69.87 | 64.54 | 88.20 | 74.54 |
| | | VCD | 67.40 | 61.39 | 93.79 | 74.21 | 71.00 | 65.41 | 89.13 | 75.45 |
| | | M3ID | 68.60 | 62.22 | 94.74 | 75.11 | 70.10 | 64.28 | 90.47 | 75.16 |
| | | RITUAL | 68.57 | 62.26 | 94.27 | 74.99 | 70.27 | 64.15 | 91.89 | 75.55 |
| | | DeGF | 72.70 | 66.70 | 90.68 | 76.86 | 73.93 | 69.36 | 85.70 | 76.67 |
| | | Ours | 75.23 | 71.03 | 85.39 | 77.65 | 75.62 | 69.40 | 91.78 | 79.04 |
| GQA | Random | Regular | 82.23 | 76.32 | 93.47 | 84.03 | 79.67 | 76.05 | 86.62 | 80.99 |
| | | VCD | 83.23 | 76.73 | 95.38 | 85.05 | 82.83 | 80.16 | 87.26 | 83.56 |
| | | M3ID | 84.20 | 78.00 | 95.26 | 85.77 | 83.07 | 80.06 | 88.06 | 83.87 |
| | | RITUAL | 86.10 | 80.30 | 95.66 | 87.31 | 84.87 | 82.52 | 88.47 | 85.39 |
| | | DeGF | 87.40 | 83.51 | 93.20 | 88.09 | 85.40 | 85.64 | 84.61 | 85.12 |
| | | Ours | 87.81 | 86.62 | 90.32 | 88.41 | 87.24 | 85.57 | 89.69 | 87.58 |
| | Popular | Regular | 73.47 | 66.83 | 93.20 | 77.84 | 73.33 | 68.72 | 85.67 | 76.26 |
| | | VCD | 72.37 | 65.27 | 95.60 | 77.58 | 76.13 | 71.10 | 88.07 | 78.68 |
| | | M3ID | 73.87 | 66.70 | 95.35 | 78.49 | 75.17 | 69.94 | 88.26 | 78.04 |
| | | RITUAL | 74.80 | 67.50 | 95.66 | 79.15 | 74.50 | 69.17 | 88.39 | 77.61 |
| | | DeGF | 78.10 | 71.56 | 93.25 | 80.98 | 76.90 | 73.89 | 83.20 | 78.27 |
| | | Ours | 77.72 | 73.17 | 88.24 | 80.00 | 79.62 | 74.61 | 90.15 | 81.65 |
| | Adversarial | Regular | 68.60 | 62.43 | 93.41 | 74.84 | 68.60 | 63.94 | 85.31 | 73.10 |
| | | VCD | 68.83 | 62.26 | 95.67 | 75.43 | 71.00 | 65.75 | 87.66 | 75.14 |
| | | M3ID | 68.67 | 62.16 | 95.42 | 75.28 | 71.17 | 65.79 | 88.19 | 75.36 |
| | | RITUAL | 68.23 | 61.75 | 95.81 | 75.10 | 70.17 | 64.76 | 88.48 | 74.78 |
| | | DeGF | 74.07 | 67.42 | 93.14 | 78.22 | 73.63 | 70.08 | 80.92 | 75.11 |
| | | Ours | 76.00 | 72.22 | 84.50 | 77.88 | 75.06 | 70.25 | 87.18 | 77.80 |

Table 6: POPE（Li 等，2023d）基准测试的详细结果。较高的（↑）准确率、精确度、召回率和 F1 值表明更好的性能。最佳结果以粗体显示，次佳结果以下划线标注。

| Model | Method | Object-level | | Attribute-level | | Total Score ↑ |
|---|---|---|---|---|---|---|
| | | Existence ↑ | Count ↑ | Position ↑ | Color ↑ | |
| LLaVA-1.5 | Regular | 173.75 | 121.67 | 117.92 | 149.17 | 562.50 |
| | VCD | 186.67 | 125.56 | 128.89 | 139.45 | 580.56 |
| | M3ID | 186.67 | 128.33 | 131.67 | 151.67 | 598.11 |
| | RITUAL | 187.50 | 139.58 | 125.00 | 164.17 | 616.25 |
| | DeGF | 186.22 | 148.33 | 130.50 | 170.33 | 635.38 |
| | Ours | 190.00 | 148.33 | 128.33 | 173.33 | 640.00 |
| InstructBLIP | Regular | 155.42 | 72.17 | 72.58 | 115.43 | 415.60 |
| | VCD | 154.49 | 81.67 | 59.11 | 112.67 | 407.94 |
| | M3ID | 158.89 | 77.22 | 61.44 | 125.00 | 422.55 |
| | RITUAL | 178.50 | 74.58 | 59.08 | 139.17 | 451.33 |
| | DeGF | 184.32 | 82.44 | 56.33 | 150.00 | 473.09 |
| | Ours | 188.67 | 89.58 | 62.33 | 152.33 | 492.91 |

Table 7: 在 MME-幻想基准上的详细表现。评分以平均值报告。得分越高 ↑ 表示性能越好。对于每个模型组 LLaVA-1.5、InstructBLIP，"Ours" 的结果如果在该列中是最好的则用黑体显示。其他列中的最佳结果也用黑体显示。带下划线的值 score 通常代表第二好的方法或是一个显著的基准。"Ours" 行是被高亮的。

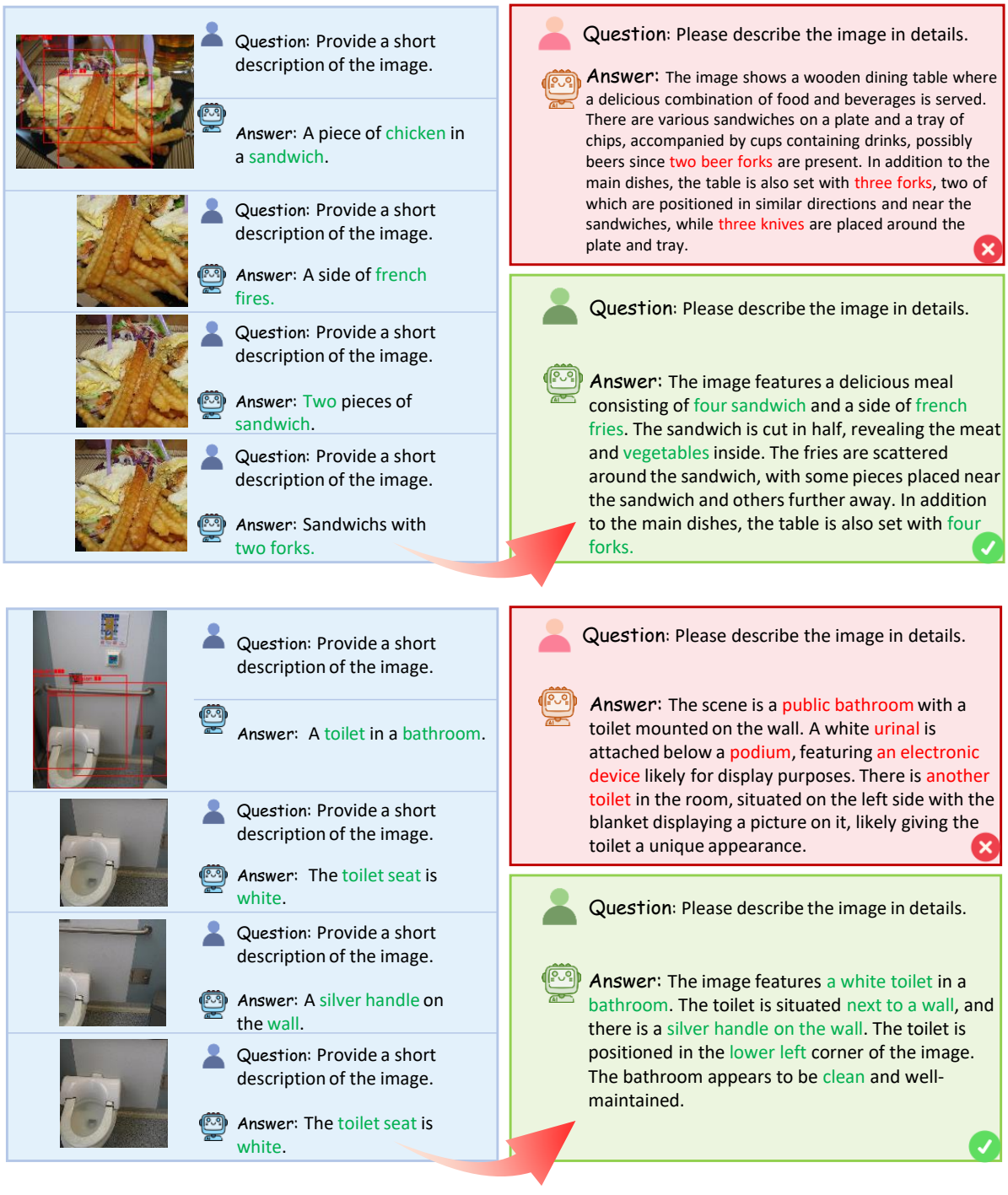


Figure 7: 对 CHAIR 评估集中的一张图片，进行由常规解码（红色）和我们的 MRFD（绿色）生成的标题的定性比较。