arXiv:2611.30460v1 [cs.CL] 23 Nov 2026

重新思考对比解码在缓解多模态大语言模型幻觉方面的评估

Ting Chen[1,*] GengLi[2,*] Guohao Chen[2] HanjiaLyu[1]1新加坡管理大学 2
南洋理工大学
*同等贡献

摘要

对比解码（*CD*）通过对比不同条件下的词元分布来调整词元分数，从而缓解多模态大语言模型（*MLLMs*）中的物体幻觉。然而，现有的生成式与判别式基准测试主要衡量最终任务分数，而这些分数可能受到候选项筛选、生成轨迹变化以及响应偏差的混杂影响，因此难以分离底层对比调整所产生的效果。为应对这些局限并补充现有评估，我们提出***CONCOR***（对比校正评估），这是一项受控基准，利用对比接地边际（*CGM*）直接评估真实答案与不同类型幻觉候选项之间的竞争。*CGM*还支持分析对比校正的方向一致性、构成与可靠性。我们的评估表明，现有对比解码方法展现出不同的校正行为：一些方法持续偏向真实候选项，而另一些方法虽能产生较强校正，但也会引入更多错误。关键在于，这些差异有助于解释它们对对比强度增加的截然不同的响应：提高对比强度会改善某些方法，同时使另一些方法性能下降。这些发现强调了评估对比校正本身的重要性，而非仅依赖最终输出表现；***CONCOR*** 为理解并改进用于幻觉缓解的对比解码方法提供了一个受控且可解释的框架。

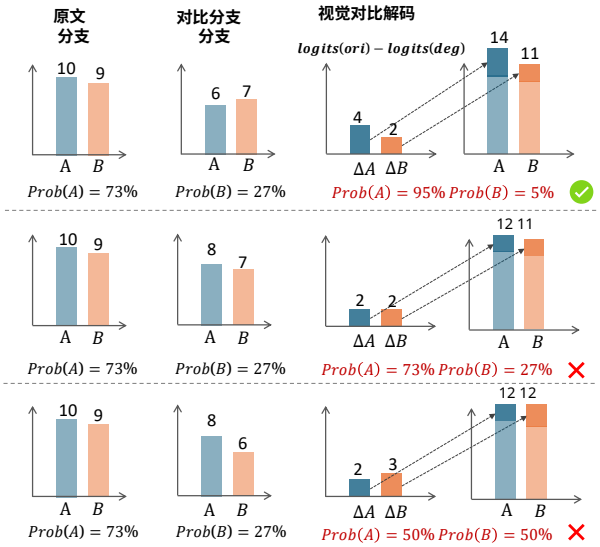


图1. 视觉对比解码的三种可能结果。只有当视觉依据词元的原始 logit 与对比 logit 之间的差值较大时，CD 才会提高其相对于幻觉竞争词元的相对偏好 A，即，$\Delta_A > \Delta_B$。当 $\Delta_A = \Delta_B$ 时，它们的相对偏好保持不变；当 $\Delta_A < \Delta_B$ 时，CD 反而会更偏好 B，而非 A。概率通过对图示的两个候选项应用 Softmax 计算，其中 $\alpha = 1$。

1. 引言

近年来，将视觉编码器与大型语言模型相结合的技术进展，使多模态大语言模型（MLLMs）具备了广泛的通用能力，可用于图像和视频描述 [23],视觉问题回答 [3], 多模态对话 [21], 视觉依据 [25],文档理解 [7], 以及视觉推理 [27, 42]。因此，MLLMs 正逐渐成为连接视觉感知与自然语言交互的重要基础 [27, 42]

。然而，当视觉证据薄弱或未得到充分利用时，这些模型往往过度依赖语言先验和场景共现模式，生成看似合理却缺乏视觉依据的物体、属性或关系 [6, 11, 15, 22, 36]。这一通常被称为视觉幻觉的现象 [12, 24], 削弱了模型输出的可信度，并限制了其在医疗辅助和自动驾驶等高可靠性设置中的部署 [1, 35, 46]。

研究人员探索了训练时对齐 [10,29、39–41],事后校正 [16, 38, 45],以及解码时干预，以缓解视觉幻觉 [8, 17,

31, 47]。在这些方法中，对比解码（CD）因无需额外训练且可直接在推理时运行而备受关注 [17, 30]。CD 对比不同条件下生成的词元分布，并利用它们之间的差异调整下一词元分数 [13]。从机制上看，只有当对比调整提高了视觉依据充分的候选项相对于缺乏视觉支持的候选项的相对偏好时，CD 才能有效纠正错误。如果这种相对偏好反而降低，对比干预可能会强化错误的候选项。

现有方法通常通过生成式和判别式任务的分数来评估 CD 缓解幻觉的能力 [4, 17, 33]。然而，这些分数同时受到解码选择和任务设计的影响。首先，CD 可以采用一种解码策略，过滤掉低概率的幻觉词元，从而走捷径提高基准测试分数 [37]。其次，生成式评估会在整个生成过程中应用 CD，此时对非目标内容的修改可能改变后续生成上下文，进而影响之后的物体判断 [32, 43, 44]。第三，判别式评估中的准确率容易受到是/否回答倾向的影响 [19]。因此，CD 可能通过提高肯定回答的比例来提升准确率，使回答倾向的变化被计作缓解幻觉的收益。综合来看，这些因素使得我们难以从报告的最终任务分数中，分离出对比干预对物体判断的直接影响。

为弥补这些局限并补充现有评估方法，我们提出 **CONCOR** （对比校正评估），这是一项受控基准测试，用于直接评估真实答案候选项与幻觉候选项之间的竞争；详情见附录A 。**CONCOR**构建候选集，其中包含真实答案物体，以及根据其与图像内容的共现关系选取的不同类型的幻觉物体。所有候选项均在相同的解码位置和上下文中进行评估，所用逻辑值取自候选项筛选之前的原始分支和对比分支。这一设计将对比调整与下游生成轨迹、候选项筛选及回答倾向隔离开来。在**CONCOR**中，我们提出对比接地边际（CGM），用于诊断 CD 是否以及如何增强真实答案候选项相对于幻觉候选项的优势。除了衡量对比校正的方向和幅度外，CGM 还提供三项互补的诊断属性：方向一致性、校正构成和校正可靠性。我们的贡献主要有三点：

- 我们指出，解码约束、生成轨迹和响应偏差会混淆现有的生成式与判别式评估，从而掩盖 CD 所产生的潜在对比校正。
- 我们提出 **CONCOR**，这是一项受控基准，用于直接评估真实答案与不同类型幻觉候选项之间的竞争；同时引入 CGM，用于测量和诊断对比校正。
- 我们揭示了现有对比解码方法在方向一致性、组成和可靠性方面存在显著差异，并说明了这些差异为何有助于解释对比强度增加为何会使某些方法受益、却使另一些方法表现下降；这些主要规律在不同模型和任务设置下仍然稳健。

2. 相关工作

对比解码。 现有对比解码方法的主要区别在于获取用于比较的分布的方式。VCD [17] 通过扰动输入图像，ICD [33]引入干扰指令，而 SID [9]和 AvisC [34]则根据词元相关性或注意力操纵视觉表征。其他方法则通过置信度感知的 logit 校准 [13]、无图像条件预测 [4]，或中间层表示 [2]构造对比信号。尽管存在这些差异，这些方法都遵循同一原则：对比不同的替代输出分布，以调整词元预测结果。

幻觉评估。 现有基准测试通过生成式、判别式和组合式设置评估多模态幻觉。CHAIR通过比较生成的对象提及与图像真实标注，衡量图像描述中的物体幻觉 [26]。它同时报告包含幻觉物体的描述所占比例，以及所有对象提及中幻觉物体所占比例。POPE将物体幻觉评估设定为二元存在性问题，并采用随机、热门和对抗采样策略构造负样本 [19]。AMBER提供了一个范围更广的基准测试，涵盖生成式和判别式幻觉，且无需依赖外部大语言模型评估器。其评估涉及物体、属性和关系方面的幻觉 [32]。这些基准测试共同提供了互为补充的物体级和响应级视觉忠实度测量，广泛用于评估多模态大语言模型中的幻觉缓解方法 [5]。

3. 预备知识

给定指令 x、图像 v 和生成前缀$y_{<t}$，令 $\ell_o(w \mid y_{<t},\ x,\ v)$ 表示原始分支为候选词元 w 分配的 logit，令 $\ell_d(w \mid y_{<t},\ x,\ v')$ 表示对比分支以输入 v' 产生的对应 logit。视觉对比解码计算

$$\ell_{\mathrm{CD}}(w) = (1+\alpha)\ell_o(w) - \alpha\ell_d(w) = \ell_o(w) + \alpha\Delta_w, \quad (1)$$

其中 $\Delta_w = \ell_o(w) - \ell_d(w)$，而 $\alpha > 0$ 控制对比强度。因此，$\alpha\Delta_w$ 是 CD 对

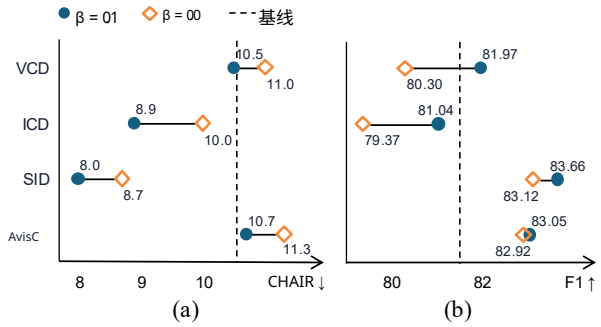


图2。自适应合理性约束（APC）对四种对比解码方法中（a）AMBER生成和（b）POPE判别的影响。蓝色圆圈表示已启用APC，橙色菱形表示已禁用APC。虚线表示未采用对比解码时的基线性能。

词元 w的原始 logit 所施加的加性调整。

为考察这一调整是否有利于视觉依据词元，考虑在相同前缀下同一解码步骤中相互竞争的两个候选项。令 A 表示一个由视觉内容支持的事实词元，令 B表示一个缺乏视觉支持、但由于模型的语言先验仍可能获得较高分数的幻觉竞争词元。分别将它们从原始到对比的logit差定义为 $\Delta_A = \ell_o(A) - \ell_d(A)$和 $\Delta_B = \ell_o(B) - \ell_d(B)$。由对比解码引起的二者成对逻辑值间隔变化为

$$[\ell_{\mathrm{CD}}(A) - \ell_{\mathrm{CD}}(B)] - [\ell_o(A) - \ell_o(B)] = \alpha(\Delta_A - \Delta_B). \tag{2}$$

由于Softmax下的成对概率比由相应的logit差决定，式（2）表明，如图1所示，仅当 A 相对于 B 时，对比解码才会提高对 $\Delta_A > \Delta_B$的相对偏好；否则，相对偏好保持不变或降低。

4. 现有对比解码评估的局限性

有效的对比解码要求对比分支提供纠正信号，使模型倾向于选择有视觉依据而非幻觉物体候选项。然而，现有评估主要通过下游生成与判别任务得分来评估对比解码，而这些得分未必能单独反映该信号的有效性。

观察***1***：解码约束会干扰对比解码纠正能力的评估。自适应合理性约束 (APC) 常用于对比解码方法 [9, 17, 33], ，它利用原始分支的预测概率限制候选词元。APC 会在最终选择前滤除低概率候选项，因此即使没有增强对比分支的纠正信号，也可能影响幻觉相关性能。如图2所示，启用 APC 始终会降低

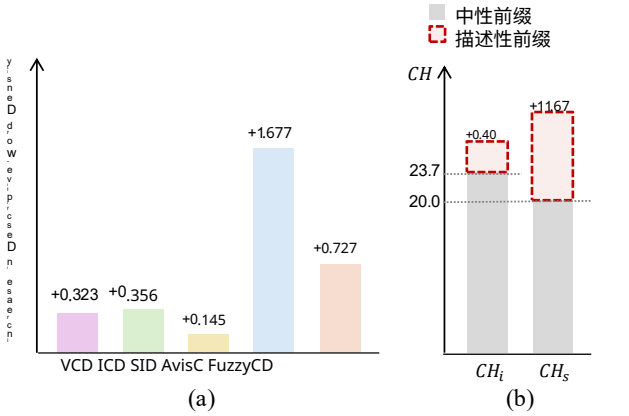


图 3. 非目标内容对生成式评估的影响。（**a**）在采样解码下，五种对比解码方法的描述性词语密度相对于基线的变化。数值表示在 500 张图像中，每生成 100 个词所增加的描述词均值。描述词包括形容词、副词、数词和空间关系表达式。（**b**）由中性前缀和描述性前缀生成的续写的幻觉率，且续写长度保持一致。

四种视觉对比解码方法的 CHAIR 分数，同时提高 F1 分数，而其对比分支的构造保持不变。实验细节见附录C.3。这些结果表明，下游性能的提升可能源于解码约束，而不只是对比重加权。因此，现有启用 APC 的评估无法单独衡量对比分支的纠正能力。

观察***2***：生成式评估会将对象层面的纠正与非目标内容的变化混为一谈。由于 CD 会在整个自回归生成过程中应用，因此它不仅会改变物体词元，也会改变非目标内容，例如属性和动作。这些变化会成为生成前缀的一部分，并可能随后影响对象预测结果。图3（a）显示，在采样解码下，与基线相比，所有五种对比解码方法都会提高描述性词语密度。为考察此类变化是否会影响后续幻觉，我们比较了从中性前缀和描述性前缀生成的续写；后者包含额外的描述性属性，但不包含额外的对象提及。实验详情见附录C.4。如图3（b）所示，在续写长度匹配的情况下，将中性前缀替换为描述性前缀，会使CHAIR_s 从20。00% 升至31。67% ；而CHAIR_i 仅从23。70% 略微变为24。10%。这项受控干预表明，非目标前缀内容能够改变后续的物体幻觉。因此，生成式评估可能会将物体词元位置上的直接对比效应

表1. 判别式评估将肯定回答捷径错误地归因于幻觉缓解效果的提升。AvisC 在降低假正例率的同时提高了真正例率，表明其确实抑制了幻觉。相比之下，SID 和 VCD 的真正例率与假正例率均有所提高，并通过肯定偏差获得更高的总体得分。这类变化提升了判别性能，却不一定意味着视觉依据更充分。实验详情请见附录 C.3 。

方法	TPR↑	FPR↓	总体肯定率	肯定回答准确率↑	否定回答准确率↑	总体准确率↑
基线	74.36%	8.11%	41.23%	74.36%	91.89%	83.12%
SID	78.31%	8.91%	43.61%	78.31%	91.09%	84.70%
VCD	76.64%	10.36%	43.50%	76.64%	89.64%	83.14%
ICD	72.91%	7.02%	39.97%	72.91%	92.98%	82.94%
AvisC	76.02%	7.04%	41.53%	76.02%	92.96%	84.49%

与生成轨迹变化所带来的间接效应混为一谈，从而难以单独评估对比分支在对象层面的纠正能力。

观察3：判别式评估可能会将响应偏差与幻觉缓解效果的提升混为一谈。有效的 CD 应当能够提升对图像中存在和不存在物体的区分能力：提高对存在物体的预测，同时减少对不存在物体的错误预测。然而，先前研究 [37] 表明，使用基于提示的调整（PBA）和输出层修改（OLM），也能通过肯定回答捷径提高判别得分。这两种方法都没有构建对比分支；它们通过诱导更强的“是”回答倾向，取得了与 CD 相当的 POPE 增益，而无需真正提升对存在和不存在物体的区分能力，汇总得分也可能因此提高。表1中的结果进一步展示了这两种模式的差异。与基线相比，AvisC 将真正例率（TPR）从 74.36% 提高到 76.02%，同时将假正例率（ FPR ）从 8.11% 降至 7.04%。因此，它提升了对正例（是）和负例（否）问题的准确率。相比之下，SID 和 VCD 的 TPR 和 FPR 均有所提高，其总体是率也分别从 41.23 % 升至 43.61% 和 43.50%。因此，它们在正例问题上的提升是以负例问题表现变差为代价的，这表明它们转向了更倾向于回答“是”，而非全面提升区分能力。即便如此，这种倾向仍可能带来更高的汇总得分：SID 的总体准确率达到 84.70%，超过 AvisC 的 84.49%。因此，汇总判别得分可能更偏向肯定回答偏差更强的方法，而不是更善于区分存在与不存在物体的方法，因而无法可靠地单独衡量 CD 缓解幻觉的能力。

5. CONCOR

前述观察表明，现有评估分数可能受到解码约束、生成上下文和响应偏差的影响，因此难以隔离对比分支提供的纠正信号。因此，评估这一信号需要在可比的对象判定条件下进行，以便直接衡量对比干预是否提升了真实答案相对于幻觉候选项的相对优势。为此，我们引入 **CONCOR** （对比校正评估），这是一种基准数据集，它将真实答案对象和幻觉对象置于同一个多选候选集中，固定生成上下文和解码位置，并在不进行候选项筛选的情况下提取两个分支的原始逻辑值。在**CONCOR**中，我们定义了**对比接地边际（CGM）**，用于衡量相对优势变化的方向和幅度（汇总见图4）。

5.1. 基准设计

我们从 COCO Val2014 [20],中抽取 2,000 张不同的图像，确保涵盖其 80 个物体类别且分布平衡，并为每张图像构建一个包含五个选项的物体识别问题。COCO 实例标注决定真实答案候选项，而未标注类别则构成幻觉候选项池。单选问题包含一个真实答案和四个幻觉候选项。多选问题包含两个至四个真实答案候选项，其余选项由幻觉候选项填充。

为评估真实物体与不同类型幻觉候选项之间的差异，我们构建候选集，将真实物体与依据不同共现标准选出的幻觉物体配对。令 P_I 表示图像 I中已标注类别的集合。对于已标注类别a 和未标注类别 o，我们将它们的条件共现定义为

$$C(a,o)=\frac{N(a,o)}{N(a)}, \tag{3}$$

其中， $N(a)$表示包含 a的 COCO 图像数量，而 $N(a,\ o)$表示同时包含 a 和 o的图像数量。对于每个未标注候选项 $o\notin P_I$，我们接着定义

$$S_{\text{scene}}(o,I)=\frac{1}{|P_I|}\sum_{a\in P_I}C(a,o), \tag{4}$$

$$S_{\max}(o,I)=\max_{a\in P_I}C(a,o). \tag{5}$$

这些分数分别衡量候选项与标注类别的平均共现程度和最高共现程度。我们利用这些分数构造四种类型的幻觉候选项：最大化$S_{\max}$的高共现候选项、最大化 S_{scene}的与场景相关候选项、最小化 S_{scene}的场景无关候选项，以及从其余未标注类别中均匀随机抽取的随机候选项。通过这种构造，每个真实标注对象都能面对具有不同共现特征的幻觉备选项。附录E中给出了候选项构造过程的示例。

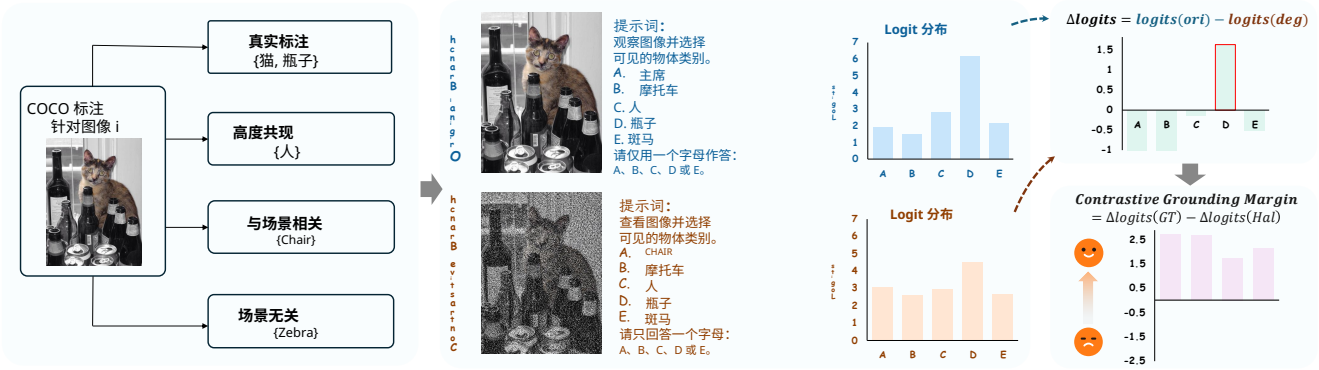


图4. CONCOR 和 CGM 概览。真实答案和幻觉候选项根据 COCO 标注和共现统计构建，并在相同提示词下进行评估。原始分支和对比分支的逻辑值用于计算对比信号 Δ_w，CGM 则据此衡量真实答案相对于幻觉候选项的 logit 优势变化。

5.2. 对比接地边际

5.2.1. 定义

在CONCOR的受控设置下，令 $\ell_o(w)$ 和 $\ell_d(w)$ 分别表示原始分支和对比分支分配给候选项 w 的逻辑值。对于每个候选项，我们将其对比信号定义为

$$\Delta_w = \ell_o(w) - \ell_d(w). \tag{6}$$

对于真实答案候选项 g 和幻觉候选项 h，我们将**对比接地边际（CGM）**定义为

$$\mathrm{CGM}(g, h) = \Delta_g - \Delta_h. \tag{7}$$

CGM 衡量对比信号是否更偏向真实候选项而非幻觉候选项，以及偏向程度。这一解读可直接从对比解码规则中得出，如式1所示，其中$\alpha > 0$ 控制对比强度。根据式2 和式 7，因此，逻辑值边际的变化为$\alpha\,\mathrm{CGM}(g, h)$。CGM 为正，表示对比干预提高了真实候选项的相对优势；CGM 为负，表示其提高了幻觉候选项的相对优势；CGM 为零，则表示两者的相对边际保持不变。CGM 的绝对值表示这一效应的强度，而逻辑值边际的实际变化则按 α缩放。

为兼顾单答案和多答案设置，令 G 和 H 分别表示某个问题中的真实答案候选项集合和幻觉候选项集合。我们将问题级 CGM 定义为

$$\mathrm{CGM}(G, H) = \frac{1}{|G|}\sum_{g\in G}\Delta_g - \frac{1}{|H|}\sum_{h\in H}\Delta_h. \tag{8}$$

这等于所有真实答案–幻觉候选项配对的成对 CGM均值，反映了单位对比强度下真实答案相对于幻觉候选项的 logit 优势变化的平均值。

5.2.2. 诊断属性

除了量化对比校正的整体强度外，CGM 还直接刻画了对比干预如何改变真实答案与幻觉候选项之间的竞争。其公式具有三个诊断属性，分别涉及方向一致性、校正构成和校正可靠性。

属性 1：方向一致性。 对于每个问题，CGM(G, H) 的符号表示对比干预会增加还是减少平均相对优势。在包含 N 个问题的基准测试中，我们将正CGM率定义为

$$R_+ = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[\mathrm{CGM}(G_i, H_i) > 0\right]. \tag{9}$$

因此，正CGM率衡量了 CD方法在样本中朝期望方向进行调整的一致程度。较高的比率表明，对比干预通常会增强真实候选项的相对优势，而较低的比率则表明该干预经常偏离正确方向。

属性 2：校正构成。 虽然 CGM 衡量的是真实候选项相对于幻觉候选项的相对优势净变化，但这种变化可能源于不同的候选项层面调整。为刻画其构成，我们分别考察真实候选项、原始分支下得分最高的幻觉候选项，以及其余幻觉候选项的对比信号，分别记为 Δ_{GT}、Δ_{MaxHal} 和 $\Delta_{\mathrm{OtherHal}}$。它们的相对值揭示了校正主要是通过增强真实候选项、抑制最强的幻觉竞争项、更广泛地抑制幻觉候选项，还是通过这些效应的组合而产生。因此，CGM 与

其候选项层面的分解能够区分那些净校正效果相似、但底层调整模式不同的对比干预。

性质3：校正可靠性。 令 $M_o(g,\ h)\ = \ell_o(g) - \ell_o(h)$ 表示真实答案候选项 g 与幻觉候选项 h之间的原始边际。在对比解码下，相应的边际变为

$$M_{\mathrm{CD}}(g,h) = M_o(g,h) + \alpha \mathrm{CGM}(g,h). \qquad (10)$$

因此，CGM 衡量真实答案与幻觉之间的边际随单位对比强度发生的变化。当最终调整幅度足够大时，正值 CGM 可以扭转最初为负的边际，从而纠正错误的候选项排序。反之，负值 CGM 可以克服最初为正的边际，使正确的排序变为错误的排序。因此，我们考察错误转正确（W→C）和正确转错误（C→W）的预测转变，以刻画对比调整带来的有益和有害影响。

6. 实验

我们通过 **CONCOR** 评估现有的对比解码方法，并研究 CGM 的诊断属性在实践中如何体现。具体而言，我们围绕三个研究问题组织实验：

•RQ1. 现有对比解码方法的效果如何？其性能背后的对比校正模式是什么？

• 研究问题2. 对比强度如何影响这些校正及预测性能？

• 研究问题 3. 观察到的校正模式在不同模型和任务设置中是否一致？

6.1. 设置

我们将基线解码与 VCD [17], ICD [33],SID [9], 和 AvisC [34]在 **CONCOR**上进行比较，涵盖 LLaVA-1 的 7B 和 13B 版本。5 [21] 以及 InstructBLIP-Vicuna [3], , 同时包括单答案和多答案设置。附录B中简要介绍了这些对比解码方法。对于每种对比解码方法，我们在候选项筛选之前，从原始分支和对比分支提取答案首个位置的原始逻辑值。我们报告 CGM 均值和正CGM率，分别用于衡量对比校正的幅度和方向一致性，并报告图像级采样准确率，以衡量其对最终预测结果的影响。

6.2. RQ1：现有对比解码方法的效果如何?

我们首先在采用单选问题的 LLaVA-1.5-7B 主设置下评估现有对比解码方法。如表 2所示，基线解码的准确率为 71.1%。对比解码方法在预测准确率及其产生的对比校正方面均存在显著差异。VCD 的 CGM 均值为 -0.029，正 CGM率为 41.6%，准确率则降至 69.2%。相比之下，ICD、SID 和 AvisC 的 CGM 均值均为正，分别为 0。393、1。279 和 3。254，准确率则分别提升至 77.5%、81.5% 和 83.1%。这些结果表明，对比解码方法之间的性能差异，与其产生的对比校正存在显著差异有关。

方向一致性。不同对比解码方法的对比调整在方向一致性方面存在显著差异。表3 显示，VCD的正CGM率仅为41.6%，这表明在大多数样本中，其对比调整非但没有提高真实答案的优势，反而使其降低。相比之下，ICD和AvisC的正CGM率分别达到93.6% 和 98.3%，表明它们能够高度一致地朝真实答案的方向进行校正。SID的表现介于这些方法之间，正CGM率为74.7%：尽管其调整在大多数情况下有利于真实答案候选项，但一致性不如ICD和AvisC。这些差异有助于区分能够可靠地朝预期方向进行校正的方法，以及校正方向经常出错的方法。

校正构成。对比解码方法通过不同的候选项级调整模式实现正向校正。在成功校正的样本中，ICD对幻觉候选项进行幅度较小但相对均匀的调整，分别将最强幻觉项和其余幻觉项的逻辑值降低0.10和0.10。相比之下，SID将调整重点放在拉开真实答案与最强幻觉项之间的差距上：将前者提高2.05，将后者降低0.64，而其余幻觉项仅变化 $+0.03$。AvisC则呈现出不同的模式：它将三组候选项的逻辑值分别提高17.44、16.16和14.49。由于对最强幻觉项和其余幻觉项的增幅依次减小，AvisC使真实答案相对于这两组候选项的边际分别增加1.28和2.95。这些结果表明，正向对比校正可以通过不同机制实现，包括直接抑制幻觉候选项，或相对大幅提高真实答案的逻辑值，即便幻觉项的逻辑值本身有所上升。</output>

校正可靠性。对比解码方法在通过对比调整校正现有错误且不引入新错误方面存在显著差异。如表3所示，SID校正的样本数最多（288），但也引入了81 个错误；相比之下，ICD为8 个，AvisC为15 个。表4 进一步显示，在SID的W→C样本中，真实标注的logit上升，而最强幻觉项的logit下降，使二者的间隔增加了2.69。在C→W样本中，这一模式反转：真实标注的logit下降

表 2。CGM 和图像　级准确率在不同模型和模型规模下的单重和多重真实答案设置　.

设置	方法	LLaVA-1.5-7B			LLaVA-1.5-13B			InstructBLIP-Vicuna-7B			InstructBLIP-Vicuna-13B		
		CGM	正例	Acc.	CGM	正例	Acc.	CGM	正例	Acc.	CGM	正例	Acc.
Single GT	基线	—	—	71.1%	—	—	87.2%	—	—	47.2%	—	—	21.8%
	VCD	-0.029	41.6%	69.2%	0.178	60.8%	85.1%	-0.078	33.9%	41.2%	-0.025	40.9%	20.6%
	ICD	0.393	93.6%	77.5%	0.120	83.3%	87.7%	-0.059	42.4%	41.7%	-0.370	5.6%	12.8%
	SID	1.279	74.7%	81.5%	4.151	97.0%	93.0%	0.357	75.1%	50.1%	0.491	88.3%	31.6%
	AvisC	3.254	98.3%	83.1%	4.683	98.5%	91.0%	1.525	97.2%	71.3%	0.760	92.0%	39.8%
Multiple GT	基线	—	—	88.0%	—	—	82.0%	—	—	87.0%	—	—	52.5%
	VCD	-0.030	46.5%	87.5%	0.165	64.5%	79.0%	0.004	50.5%	85.5%	0.061	64.5%	52.5%
	ICD	0.115	79.5%	91.0%	0.075	84.0%	81.0%	0.043	67.0%	82.0%	0.074	86.0%	55.0%
	SID	0.545	73.5%	91.0%	1.836	97.0%	93.5%	0.170	72.5%	92.5%	0.836	97.5%	64.5%
	AvisC	2.131	99.5%	94.0%	2.267	99.0%	93.5%	0.859	94.5%	96.0%	0.996	98.0%	89.5%

表3. 候选级对比信号和预测转移。

方法	Δ_{GT}	Δ_{MaxHal}	Δ_{OtherHal}	C→W	W→C	Positive-CGM
基线	—	—	—	—	—	—
VCD	$+1.13$	-0.27	-0.17	100	62	41.6%
ICD	$+0.54$	-0.10	-0.10	**8**	135	93.6%
SID	$\mathbf{+2.05}$	$\mathbf{-0.64}$	$+0.03$	81	**288**	74.7%
AvisC	$+17.44$	$+16.16$	$+14.49$	15	255	98.3%

表4. 预测转变的候选级对比信号。

方法	转变	Δ_{GT}	Δ_{MaxHal}	Δ_{Margin}	Acc.
SID	W→C	$+2.05$	-0.64	$\mathbf{+2.69}$	81.5%
	C→W	-0.76	$+1.52$	$\mathbf{-2.28}$	
ICD	W→C	$+0.54$	-0.10	$\mathbf{+0.64}$	77.5%
	C→W	-0.04	$+0.31$	$\mathbf{-0.35}$	
AvisC	W→C	$+17.44$	$+16.16$	$\mathbf{+1.28}$	83.1%
	C→W	$+15.98$	$+16.79$	$\mathbf{-0.81}$	

表5。对比解码强度对 LLaVA-1.5-7B 准确率的影响。

方法	CGM	正例	准确率（$\alpha = 1$）	准确率（$\alpha = 3$）
基线	—	—	71.1%	71.1%
VCD	-0.029	41.6%	69.2%	63.3%
ICD	0.393	93.6%	77.5%	83.2%
SID	1.279	74.7%	81.5%	77.2%
AvisC	3.254	98.3%	83.1%	87.2%

了0.76，而成为最终预测结果的幻觉候选项则上升了1.52，使间隔变化为 -2.28。这一不利的间隔变化超过了ICD（-0.35）和AvisC（-0.81）的变化。当反向调整超过真实标注初始的logit领先幅度时，幻觉项便会反超真实标注，将正确预测转变为错误。

6.3. 研究问题2：对比强度如何影响对比解码方法?

研究问题1中发现的校正模式表明，对比解码方法对增强的对比强度可能会产生不同的响应。在对比解码公式下，增加 α 会在不改变方向的情况下调整对比调整的幅度。因此，当CGM为正时，更强的对比应会放大有益的调整；但当CGM为负时，也可能放大有害的调整。我们通过将 α 从1增加到3来考察这一关系。

表5显示，将 α 从1提升至3后，ICD的准确率从77.5%提高到83.2%，AvisC的准确率则从83.1%提高到87.2%。如RQ1所示，这些方法的正CGM率分别为93.6%和98.3%，表明它们的调整在大多数样本上都偏向真实候选项。因此，提高对比强度主要会放大有益的调整。对于ICD而言尤其如此：放大其初始幅度较小但方向可靠的调整，可带来5.7个百分点的提升。相比之下，SID的正CGM率较低，为74.7%。尽管SID能够产生强有力的有益修正，RQ1还显示，其不利调整也可能足够大，从而将最初正确的预测反转。提高 α 同样会放大这些调整，使其准确率从81.5%下降至77.2%。VCD进一步说明了放大不可靠对比方向的风险：其均值CGM为 -0.029，正CGM率仅为41.6%，准确率则从69.2%下降至63.3%。这些结果表明，CGM的作用不仅限于比较最终性能：它还揭示了不同CD方法如何抑制幻觉，以及在何种条件下可能放大错误。该分析能够为方法选择和解码配置提供指导，同时帮助发现改进对比分支及相关策略的机会，从而支持更有效的对比解码设计与应用。

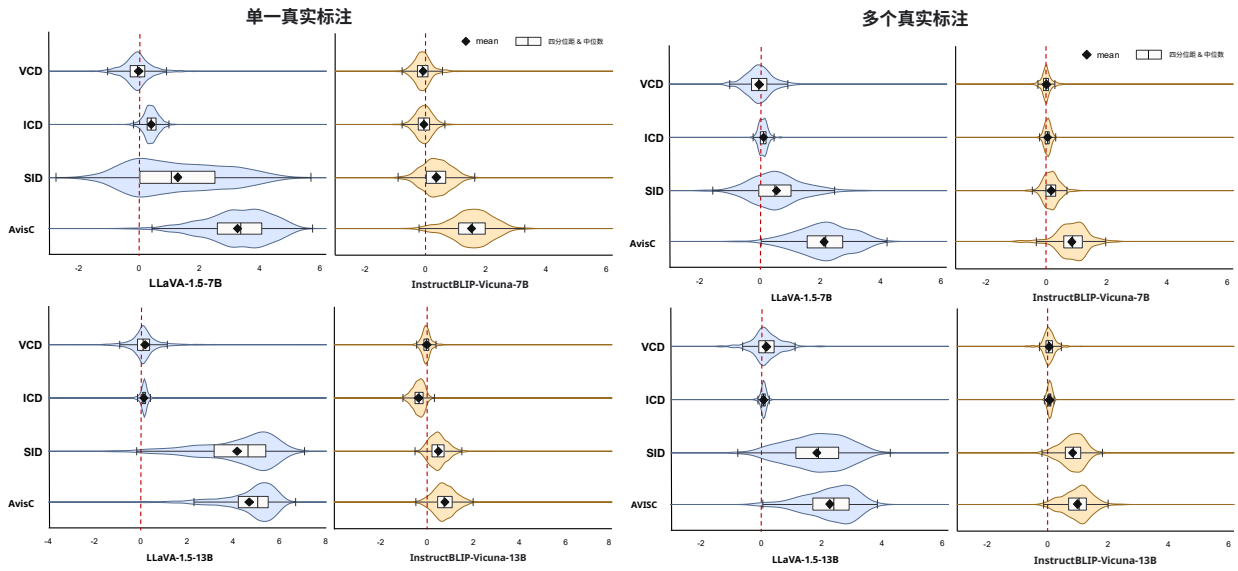


图 5. 不同模型和真实答案设置下 CGM 的分布。小提琴图展示了在单重和多重真实答案设置下，LLaVA-1.5 和 InstructBLIP-Vicuna 的 7B 与 13B 规模模型中 VCD、ICD、SID 和 AvisC 的 CGM 分布。菱形表示均值，箱体表示四分位距及中位数，须线延伸至四分位距的 1.5 倍。

6.4. 研究问题 3：观察到的效果在不同模型和任务设置下是否一致?

不同模型架构、参数规模和问题格式下，CD方法之间的主要差异仍然保持一致。表2 显示，当使用InstructBLIP-Vicuna-7B替代LLaVA-1.5-7B时，在单答案设置下，AvisC和SID的均值CGM分别为1。525 和0。357，仍然高于VCD的 −0.078 和ICD的−0.059。在更大的模型规模下也观察到相同的排序：在LLaVA-1.5-13B上，AvisC和SID的均值CGM分别为4。683 和4。151，相比之下，VCD为0。178，ICD为0。120 。这些结果与以下发现一致：特征级方法能够在不同模型中提供更强的纠正信号。

如图5所示，这种一致性也延伸到了多答案任务，在这类任务中，多个真实答案候选项会与幻觉候选项共同竞争。在 LLaVA-1.5-7B 上，AvisC 和 SID 的 CGM 均值分别为 2.131 和 0.545，再次超过 ICD 的 0.115 和 VCD 的 −0.030。在表2所列的全部八种设置中（涵盖四种模型变体和两种问题格式），AvisC 的 CGM 均值始终高于 SID，且二者均高于 ICD 和 VCD。进一步分析还证实，这些发现对于不同的选项选择设置（附录 C.6）、标注缺失程度（附录D）以及数据集构建方式（附录 D.1）均具有鲁棒性。这些结果表明，CGM 所识别出的对比校正主要差异，在所测试的模型架构、参数规模和问题格式。

7. 结论

我们重新审视了对比解码在缓解多模态大语言模型中的物体幻觉方面的评估。我们发现，现有的生成式和判别式评估可能会受到解码约束、生成轨迹变化和响应偏差的干扰。为弥补这些局限并补充现有评估，我们提出了**CONCOR**，这是一项直接评估真实答案与不同类型幻觉候选项之间竞争的基准测试；同时提出对比接地边际（CGM），用于衡量对比解码如何改变它们之间相对的 logit 边际。除了衡量对比校正的整体强度外，CGM 还能诊断其方向一致性、校正构成和可靠性。将这一框架应用于现有的 CD 方法后，我们发现它们产生校正的方式存在显著差异。展望未来，这些诊断结果为改进 CD 方法提供了若干方向。例如，后续研究可以开发能够更稳定地产生有益校正的对比分支，根据对比信号的方向和幅度调整对比强度，或明确控制调整如何分配至真实答案和幻觉候选项之间。总体而言，**CONCOR**和 CGM 提供了一个更可控且更易解释的框架，用于评估对比解码，并理解其底层调整如何助力缓解幻觉。

大语言模型使用声明

我们使用 ChatGPT 协助组织和撰写稿件，以提升其清晰度。它还通过识别可能相关的文献，协助进行初步的参考资料检索，以供进一步审阅。

参考文献

[1] Junzhou Chen 和 Sidi Lu。基于多模态大语言模型的自动驾驶车辆先进驾驶智能体。载于《2024年IEEE移动、运营、服务与技术国际会议（MOST）论文集》，第 1–11 页。IEEE，2024。1[2] Ting Chen、Geng Li、Guohao Chen、Yu Hu、Guan Huang、Mai Chen、Langsheng Lei 和 Yongyong Lin。Yard：用于缓解大型视觉语言模型幻觉的 Y 架构寄存器解码。载于《2026年自然语言处理经验方法会议论文集》，第 6864–6885 页，2026。2[3] Wenliang Dai、Junnan Li、Dongxu Li、Anthony Tiong、Junqi Zhao、Weisheng Wang、Boyang Li、Pascale N Fung 和 Steven Hoi。InstructBLIP：借助指令微调实现通用视觉语言模型。载于《神经信息处理系统进展》，36：49250–49267，2023。1，6[4] Alessandro Favero、Luca Zancato、Matthew Trager、Siddharth Choudhary、Pramuditha Perera、Alessandro Achille、Ashwin Swaminathan 和 Stefano Soatto。通过视觉信息定位控制多模态幻觉。载于《2024年IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，第 14303–14312 页。IEEE，2024。2[5] Mehrdad Fazli、Bowen Wei、Ahmet Sari 和 Ziwei Zhu。CAAC：通过置信度感知注意力校准减少大型视觉语言模型中的幻觉。载于《2026年IEEE/CVF计算机视觉应用冬季会议（WACV）论文集》，第 1242–1251 页。IEEE，2026。2[6] Tianrui Guan、Fuxiao Liu、Xiyang Wu、Ruiqi Xian、Zongxia Li、Xiaoyu Liu、Xijun Wang、Lichang Chen、Furong Huang、Yaser Yacoob 等。HallusionBench：用于诊断大型视觉语言模型中交织的语言幻觉与视觉错觉的先进诊断套件。载于《2024年IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，第 14375–14385 页。IEEE，2024。1[7] Anwen Hu、Haiyang Xu、Jiabo Ye、Ming Yan、Liang Zhang、Bo Zhang、Ji Zhang、Qin Jin、Fei Huang 和 Jingren Zhou。mPLUG-DocOwl 1.5：用于无OCR文档理解的统一结构学习。载于《计算语言学协会论文集：EMNLP 2024精选论文》，3120，2024 3096 1 pages –。[8] Qidong Huang、Xiaoyi Dong、Pan Zhang、Bin Wang、Conghui He、Jiaqi Wang、Dahua Lin、Weiming Zhang 和 Nenghai Yu。OPERA：通过过度信任惩罚与回溯分配缓解多模态大语言模型中的幻觉。载于《2024年IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，第 13418–13427 页。IEEE，2024。1[9] Fushuo Huo、Wenchao Xu、Zhong Zhang、Haozhao Wang、Zhicheng Chen 和 Peilin Zhao。自省解码：缓解大型视觉语言模型中的幻觉。在国际学习表征会议上，第 24272–24295 页，2025 年。2, 3, 6[10] 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。发表于 2024年IEEE/CVF计算机视觉与模式识别会议 (CVPR)，第 27026–27036 页。IEEE，2024 年。1[11] Zhangqi Jiang、Junkai Chen、Beier Zhu、Tingjin Luo、Yankun Shen 和 Xu Yang。大型视觉语言模型中间层中的恶魔：通过注意力透镜解释、检测和缓解物体幻觉。发表于 2025 IEEE/CVF计算机视觉与模式识别会议 (CVPR)，第 25004–25014 页。IEEE，2025 年。1[12] Liqiang Jing 、Ruosen Li、Yunmo Chen 和 Xinya Du。Faith-score：大型视觉语言模型幻觉的细粒度评估。发表于计算语言学协会：EMNLP 2024 论文集，5063，2024 5042 1 pages – 。[13] Jieun Kim、Jinmyeong Kim、Yoonji Kim 和 Sung-Bae Cho。通过模糊对比解码缓解大型视觉语言模型中的物体幻觉。发表于 2025 IEEE/CVF国际计算机视觉大会 (ICCV)，第 20572–20581 页。IEEE，2025 年。2[14] Alina Kuznetsova、Hassan Rom、Neil Alldrin、Jasper Uijlings、Ivan Krasin、Jordi Pont-Tuset、Shahab Kamali、Stefan Popov、Matteo Malloci、Alexander Kolesnikov 等。Open Images 数据集 v4：大规模统一图像分类、目标检测和视觉关系检测。《国际计算机视觉期刊》，128(7):1956–1981，2020 年。18[15] Kang-il Lee、Minbeom Kim、Seunghyun Yoon、Minsung Kim、Dongryeol Lee、Hyukhun Koh 和 Kyomin Jung。VLind-bench：衡量大型视觉语言模型中的语言先验。发表于计算语言学协会：NAACL 2025 论文集，第 4129–4144 页，阿尔伯克基，新墨西哥州，2025 年。计算语言学协会。1[16] Seongyun Lee、Sue Hyun Park、Yongrae Jo 和 Minjoon Seo。Volcano：通过自反馈引导修订缓解多模态幻觉。发表于计算语言学协会北美分会 2024 年会议论文集：人类语言技术（第 1 卷：长篇论文），第 391–404 页，2024 年。1[17] 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。发表于 2024年IEEE/CVF计算机视觉与模式识别会议 (CVPR)，第 13872–13882 页。IEEE，2024 年。1, 2, 3, 6[18] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。Blip-2：利用冻结图像编码器和大型语言模型引导语言-图像预训练。发表于国际机器学习会议，第 19730–19742 页。PMLR，2023 年。1[19] Yifan Li、Yifan Du、Kun Zhou、Jinpeng Wang、Xin Zhao 和 Ji-Rong Wen。评估大型视觉语言模型中的物体幻觉。发表于 2023 年自然语言处理经验方法会议论文集

，第 292–305 页，2023 年。2[20] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft coco：上下文中的常见物体。发表于欧洲计算机视觉会议，第 740–755 页。Springer，2014 年。4, 11[21] Haotian Liu、Chunyuan Li、Yuheng Li 和 Yong Jae Lee。通过视觉指令微调改进基线。发表于 IEEE/CVF计算机视觉与模式识别会议 (CVPR) 论文集，第 26296–26306 页，2024 年。1, 6[22] Hanjia Lyu、Jinfa Huang、Daoan Zhang、Yongsheng Yu、Xinyi Mou、Jinsheng Pan、Zhengyuan Yang、Zhongyu Wei 和 Jiebo Luo。GPT-4V(ision) 作为社交媒体分析引擎。《ACM智能系统与技术汇刊》，16(3)，2025 年。1[23] Muhammad Maaz、Hanoona Rasheed、Salman Khan 和 Fahad Khan。Video-chatgpt：借助大型视觉语言模型实现细致的视频理解。发表于计算语言学协会第62届年会论文集（第 1 卷：长篇论文），第 12585–12602 页，2024 年。1[24] Kyungmin Min、Minbeom Kim、Kang-il Lee、Dongryeol Lee 和 Kyomin Jung。通过摘要引导解码缓解大型视觉语言模型中的幻觉。发表于计算语言学协会：NAACL 2025 论文集，第 4183–4198 页，阿尔伯克基，新墨西哥州，2025 年。计算语言学协会。1[25] Zhiliang Peng、Wenhui Wang、Li Dong、Yaru Hao、Shaohan Huang、Shuming Ma、Qixiang Ye 和 Furu Wei。将多模态大语言模型 grounding 到现实世界。发表于国际学习表征会议，第 51575–51598 页，2024 年。1[26] Anna Rohrbach 、Lisa Anne Hendricks、Kaylee Burns、Trevor Darrell 和 Kate Saenko。图像描述中的物体幻觉。发表于 2018 年自然语言处理经验方法会议论文集，第 4035 4045 页，2018 2 – 。[27] Hao Shao、Shengju Qian、Han Xiao、Guanglu Song、Zhuofan Zong、Letian Wang、Yu Liu 和 Hongsheng Li。视觉思维链：通过全面的数据集和基准测试推动多模态语言模型的思维链推理。神经信息处理系统进展，37:8612–8642，2024 年。1[28] Shuai Shao、Zeming Li、Tianyuan Zhang、Chao Peng、Gang Yu、Xiangyu Zhang、Jing Li 和 Jian Sun。Objects365：用于目标检测的大规模高质量数据集。发表于 2019 IEEE/CVF国际计算机视觉大会 (ICCV)，第 8429–8438 页。IEEE，2019 年。18[29] Zhiqing Sun、Sheng Shen、Shengcao Cao、Haotian Liu、Chunyuan Li、Yikang Shen、Chuang Gan、Liangyan Gui、Yu-Xiong Wang、Yiming Yang 等。通过事实增强型RLHF 对齐大型多模态模型。发表于计算语言学协会：ACL 2024 论文集，13110，2024 13088 1 pages – 。[30] Wei Suo、Lijun Zhang、Mengyang Sun、Lin Yuanbo Wu、Peng Wang 和 Yanning Zhang。Octopus：通过动态对比解码缓解幻觉。发表于 2025 IEEE/CVF 计算机视觉与模式识别会议（CVPR），第 29904–29914 页。IEEE，2025。2[31] Zifu Wan、Ce Zhang、Silong Yong、Martin Q Ma、Simon Step- puttis、Louis-Philippe Morency、Deva Ramanan、Katia Sycara 和 Yaqi Xie。Only：单层干预足以缓解大型视觉语言模型中的幻觉。载于 2025年IEEE/CVF国际计算机视觉会议（ICCV），第 3225–3234 页。IEEE，2025。2[32] Junyang Wang、Yuhang Wang、Guohai Xu、Jing Zhang、Yukai Gu、Haitao Jia、Jiaqi Wang、Haiyang Xu、Ming Yan、Ji Zhang 等。Amber：用于评估多模态大语言模型幻觉的无大语言模型多维基准。arXiv 预印本 arXiv:2311.07397，2023。2[33] 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。载于计算语言学协会会议论文集：ACL 2024，第 15840–15853 页，2024。2, 3, 6[34] Sangmin Woo、Donguk Kim、Jaehyuk Jang、Yubin Choi 和 Changick Kim。不要只见树木，不见森林：针对大型视觉语言模型的注意力视觉校准。载于计算语言学协会会议论文集：ACL 2025，第 1927–1951 页，2025。2, 6[35] Shaoyuan Xie、Lingdong Kong、Yuhao Dong、Chonghao Sima、Wenwei Zhang、Qi Alfred Chen、Ziwei Liu 和 Liang Pan。视觉语言模型是否已准备好用于自动驾驶？从可靠性、数据和度量指标角度开展的实证研究。载于 2025年IEEE/CVF国际计算机视觉会议（ICCV），第 6285–6297 页。IEEE，2025。1[36] Tianyun Yang、Ziniu Li、Juan Cao 和 Chang Xu。通过模块化归因与干预理解并缓解大型视觉语言模型中的幻觉。载于国际学习表征会议，第 51546–51568 页，2025。1[37] Hao Yin、Guangzong Si 和 Zilei Wang。性能提升的假象：为什么对比解码无法缓解多模态大语言模型中的物体幻觉？神经信息处理系统进展，38:33056–33079，2026。2, 4[38] 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。1[39] Dokyoon Yoon、Youngsook Song 和 Woomyong Park。停止事事都学，以缓解视觉幻觉；聚焦幻觉目标。载于 2025 IEEE/CVF计算机视觉与模式识别会议（CVPR），第 4200–4208 页。IEEE，2025。1[40] 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工反馈实现行为对齐，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。[41] Tianyu Yu、Haoye Zhang、Qiming Li、Qixin Xu、Yuan Yao、Da Chen、Xiaoman Lu、Ganqu Cui、Yunkai Dang、Taiwen He 等。RLAIF-V：开源人工智能反馈带来更高的 GPT-4V 可信度

可信度发表于 *2025*年*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*，第 19985–19995 页。IEEE，2025 年。 1

[42] Daoan Zhang、Junming Yang、Hanjia Lyu、Zijian Jin、Yuan Yao、Mingkai Chen 和 Jiebo Luo。面向具有多个图像输入的大型多模态模型的基准测试与思维链提示策略。发表于 《模式识别》，第 226–241 页，沙姆，2025 年。施普林格·自然瑞士。 1

[43] Ge Zheng、Bin Yang、Jiajin Tang、Hong-Yu Zhou 和 Sibei Yang。为何大型视觉语言模型在视觉问答中更容易产生幻觉：上下文的作用。发表于 *2025*年*IEEE/CVF*国际计算机视觉会议 *(ICCV)*，第 4101–4113 页。IEEE，2025 年。 2

[44] Weihong Zhong、Xiaocheng Feng、Liang Zhao、Qiming Li、Lei Huang、Yuxuan Gu、Weitao Ma、Yuan Xu 和 Bing Qin。研究并缓解大型视觉语言模型中的多模态幻觉滚雪球效应。发表于 计算语言学协会第*62*届年会论文集（第 *1* 卷：长篇论文），第 11991–12011 页，2024 年。 2

[45] Yiyang Zhou、Chenhang Cui、Jaehong Yoon、LinjunZhang、Zhun Deng、Chelsea Finn、Mohit Bansal 和 Huaxiu Yao。大型视觉语言模型中的物体幻觉分析与缓解。发表于 国际学习表征会议，第 56969–56998 页，2024 年。 1

[46] Zhihong Zhu、Yunyan Zhang、Xianwei Zhuang、Fan Zhang、Zhongwei Wan、Yuyan Chen、Qingqing Long、Yefeng Zheng 和 Xian Wu。我们能信任人工智能医生吗？大型语言模型和大型视觉语言模型中的医学幻觉综述。发表于 计算语言学协会会议论文集：*ACL 2025*，第 6748–6769 页，2025 年。 1[47] Xin Zou、Yizhou Wang、Yibo Yan、Yuanhuiyi Lyu、Kening Zheng、Sirui Huang、Junkai Chen、Peijie Jiang、Jia Liu、Chang Tang 等。回答前请再看一遍：多模态大语言模型中的记忆空间视觉回溯与幻觉缓解。 *arXiv* 预印本 *arXiv:2410 03577*，2024 年2。。

A. CONCOR 详细信息

图 6 比较生成式、判别式和选择题设定下的幻觉评估。**CONCOR** 采用受控的选择题设置，在同一个问题和回答上下文中，让真实候选项与幻觉候选项直接竞争。

A.1. 数据集构建

CONCOR 可以在具有图像级物体类别标注的数据集上实例化。令 $\mathcal{C}$ 表示类别词汇表， P_I 表示在图像I中标注为存在的类别， $U_I = \mathcal{C} \setminus P_I$ 表示未标注为存在的类别。真实候选项从 P_I中选取，而幻觉候选项则由 U_I构造。

我们的主要实验基于 COCO2014 [20] 验证集中的图像、80 个物体类别及实例标注来实现 **CONCOR** 。

A.1.1. 单一真实标注设置

单一真实答案设置包含 2,000 个不同的图像问题，这些问题经过抽样，以覆盖 COCO 的 80 个物体类别，并兼顾类别分布。每个问题包含一个真实候选项和四个具有不同共现特征的幻觉候选项。模型收到以下提示词：

> 请研究图像，并选择其中可见的物体类别。A. {物体A}B. {物体B}C. {物体C}D. {物体D}E. {物体E}请仅使用一个字母作答：A、B、C、D 或 E。

四个幻觉候选项根据第5.1节所述的共现标准构建。它们按顺序选取，且不重复。若出现并列，则依次依据次级共现得分和全局类别频率进行判定。

为减少候选项位置偏差，我们为每个问题生成五种循环 A–E 排列，使每个语义候选项恰好出现在每个答案位置一次。这样共生成 $2{,}000 \times 5 = 10{,}000$ 个模型输入。

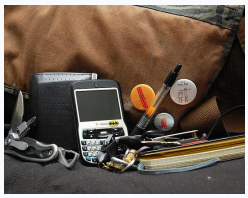
A.1.2. 多真实答案设置

多真实答案设置包含 200 个图像问题。每个问题有五个候选项，其中两个至四个是真实候选项，其余为幻觉候选项。模型收到以下提示词：

> 这张图像中有哪些物体？请选择所有可见物体类别对应的选项。A. {物体A}B. {物体 B}C. {物体C}D. {物体D}E. {物体E}请按字母顺序列出所有正确选项，并用逗号分隔（例如：A, C）。仅选择可见的选项，且勿选择其他选项。不要输出其他内容。

由于每个问题包含五个候选项，幻觉候选项的数量满足 $|H| = 5 - |G|$。数据集包含 67 个问题，其中 $(|G|,\ |H|) = (2,\ 3)$ 的有 67 个，$(3,\ 2)$ 的有 67 个，$(4,\ 1)$ 的有 66 个。

真实候选项从标注面积至少占图像面积 0.2% 的已标注对象中选取，并按标注面积排序。优先选择较大的对象。幻觉候选项从未标注类别中选取，并按 S_{scene} 降序排列。如遇并列，则依次依据$S_{\max}$ 和全局类别频率进行判定。



生成任务中的幻觉

描述这幅图像。

图像显示一个棕色布制**包**放在一个表面上，里面装着各种各样的物品。图像的焦点是放置在包中央的一**部手机**。包中的其他物品包括**钥匙**、**一个钱包**、一**支笔**以及各种办公工具，例如 **剪刀**、**铅笔**和一把 **夹子**。此外，包里还有一个 **电脑鼠标**，位于包中。这些物品和对象的组合表明这个包可能属于某个需要随时随地携带并使用这些物品的人。

判别任务中的幻觉

有一**副眼镜**吗?

是的，有一副眼镜。

有**电脑鼠标**吗?

是的，有一台电脑鼠标。

多项选择题中的幻觉

仔细观察图像并选择可见的物体类别。

A. 人
B. 斑马
C. 背包
D. 水槽
E. 桌子

请只用一个字母作答：A、B、C、D 或 E。

E

图 6. 幻觉评估的比较：生成式、判别式和多项选择题（MCQ）任务 .

与单一真实标注设置相同，每个问题都转换为五种循环排列，从而生成 $200 \times 5 = 1{,}000$ 个模型输入。

A.2. 度量指标计算

我们使用第5.2节定义的候选项级对比调整和对比定位边际（CGM）。对于每个语义候选项，首先对其在五种循环排列中的对比调整取平均，以减轻候选项位置的影响。然后根据公式8计算这些位置平均调整值，并据此计算问题级CGM。

单一真实答案准确率。当预测标签与唯一的真实答案标签一致时，该排列即为正确。若某个图像问题的五种排列中至少有三种正确，则计为正确。

多个真实答案准确率。我们采用无假正例标准。当预测集合非空且为真实答案集合的子集时，该排列即为正确。因此，允许遗漏部分真实候选项；但只要选中任何幻觉候选项，预测便判为错误。若某个图像问题的五种排列中至少有三种符合此标准，则计为正确。

B. 用于缓解幻觉的对比解码

本节介绍四种用于缓解多模态大语言模型幻觉的方法：视觉对比解码（VCD）、指令对比解码（ICD）、自省解码（SID）和注意力视觉校准（AvisC）。这些方法在推理时构建一个放大幻觉的参考分支，并将其输出分布与原始分支的输出分布进行对比，以调整下一词元分数。

B.1. 一般形式

给定图像 I、指令 x和先前生成的词元 $y_{<t}$，解码步骤 t 处的原始逻辑值为

$$\mathbf{z}_t = f_\theta(I, x, y_{<t}).$$

对比解码还会构造一个扰动条件 c^-，并计算相应的参考逻辑值：

$$\mathbf{z}_t^- = f_\theta(c^-, y_{<t}).$$

对比逻辑值通常可写为

$$\mathbf{z}_t^{\mathrm{CD}} = (1+\alpha)\mathbf{z}_t - \alpha\mathbf{z}_t^-,$$

其中，α 控制参考分支的贡献。一些方法还会应用合理性约束，将解码限制在原始分布下的高概率候选项之中。

B.2. 视觉对比解码

VCD 通过扰动视觉输入来构建其参考分支。逐渐向原始~图像 I中添加高斯噪声，生成失真图像 I。随后，模型计算

$$\mathbf{z}_t = f_\theta(I, x, y_{<t}), \qquad \widetilde{\mathbf{z}}_t = f_\theta(\widetilde{I}, x, y_{<t}).$$

VCD 将以原始图像为条件得到的逻辑值作为主分支，将以失真图像为条件得到的逻辑值作为参考分支。两组逻辑值通过加权减法进行组合。随后，可信度约束会移除在原始分布下概率较低的候选项，并从剩余的候选集中选出下一个词元。

B.3. 指令对比解码

ICD 通过扰动文本指令来构建其参考分支。在原始指令中添加预定义的扰动前缀或角色描述，以得到扰动指令 $\widetilde{x}$。模型计算

$$\mathbf{z}_t = f_\theta(I, x, y_{<t}), \qquad \widetilde{\mathbf{z}}_t = f_\theta(I, \widetilde{x}, y_{<t}).$$

在每个解码步骤中，ICD 将以扰动指令为条件得到的逻辑值，从以原始指令为条件得到的逻辑值中减去。对比系数控制减法的幅度。随后，根据原始分布构建一个高概率候选集，并在该集合内执行词元归一化和选择。

B.4. 自省解码

SID 使用模型内部的注意力信息构建其参考分支。其主要组成部分是上下文与文本感知词元选择（CT^2S）策略，该策略用于识别从当前生成位置接收到相对较低注意力的视觉词元。在较早的解码器层 i，CT^2S 会通过对所有注意力头上当前生成位置的注意力取平均，为每个视 觉词元 v_j 计算重要性分数：

$$s_i(v_j) = \frac{1}{H}\sum_{h=1}^{H} \mathbf{A}_i^{(h)}[t, j],$$

其中，H 表示注意力头的数量，而$\mathbf{A}_i^{(h)}[t, j]$ 表示当前词元位置 t 在头 h中分配给视觉词元 v_j 的注意力。重要性分数最低的视觉词元将被选为

$$\mathcal{V}_i^{\downarrow} = \mathrm{BottomK}\left(\{v_j\}_{j=1}^{N}; \{s_i(v_j)\}_{j=1}^{N}\right).$$

参考分支保留重要性分数较低的视觉词元，并将其传递给后续解码器层，以获得参考逻辑值。与此同时，原始分支使用完整的视觉词元集合生成原始逻辑值。随后，SID 对两个分支执行加权减法，并从校准分布中生成下一个词元。在自回归生成的不同步骤中，视觉词元选择会动态重复进行。

B.5. 注意力视觉校准

AvisC 首先从模型内部的视觉注意力分布中识别盲词元。它聚合选定解码器层和注意力头上的视觉注意力，为每个视觉词元 j获得一个注意力分数 A_j。令 μ_A 和 σ_A 分别表示这些分数的均值和标准差。满足以下条件的词元将被识别为盲词元：

$$\mathcal{B} = \{j \mid A_j > \mu_A + \lambda\sigma_A\},$$

其中，λ 控制盲词元选择的阈值。

随后，AvisC 通过保留索引为 $\mathcal{B}$ 的视觉词元，并将其余视觉词元置零，构建偏置视觉表示 $\mathcal{V}^*$。给定原始视觉表示 $\mathcal{V}$，模型计算

$$\boldsymbol{\ell}_t = \log p\left(y_t \mid \mathcal{V}, x, y_{<t}\right), \qquad \boldsymbol{\ell}_t^* = \log p\left(y_t \mid \mathcal{V}^*, x, y_{<t}\right),$$

其中，$\boldsymbol{\ell}_t$ 和 $\boldsymbol{\ell}_t^*$ 分别表示从原始视觉表示和偏置视觉表示中获得的下一词元逻辑值。AvisC 对这些逻辑值进行对比，以调整下一词元分数，并将偏置表示用作参考分支。

C. 补充实验细节

C.1. 计算资源

所有模型推理实验均在单张 NVIDIA A800-SXM4 80GB、NVIDIA RTX 6000D、NVIDIA GeForce RTX 5090 或 NVIDIA GeForce RTX 4090 图形处理器上运行，具体取决于实验组。同一实验组内，所有对比方法和设置均使用相同型号的图形处理器进行评估，以确保硬件条件一致。

C.2. 主要实验

C.2.1. 模型与一般设置

我们在 LLaVA-1.5-7B、LLaVA-1.5-13B、InstructBLIP-Vicuna-7B 和 InstructBLIP-Vicuna-13B 上评估 VCD、ICD、SID 和 AvisC，并以标准解码作为基线（记为 Baseline）。

所有模型均用于推理，不进行额外训练或微调。对于每个模型和样本，所有方法使用相同的图像、问题、候选集和候选项排序。对比解码方法使用相同的原始分支计算，仅在参考分支的构建方式上有所不同。Baseline 仅使用原始分支。

LLaVA 使用 llava v1 对话模板，而 InstructBLIP 使用官方的 InstructBlipProcessor 进行图像和文本预处理。

C.2.2. 对比解码配置

除非另有说明，所有对比解码方法均使用对比强度 $\alpha = 1$。所有模型和方法使用相同的生成参数：温度 $= 1$、top-$p = 1$、top-$k = 0$、重复惩罚 $= 1$，以及最大新生成词元数$= 16$。通过 do sample=true 启用采样。

VCD. VCD 通过向预处理后的图像张量添加扩散噪声来构建其参考分支。我们使用的扩散时间步为 $t = 500$。原始分支接收

干净图像，而参考分支接收空间维度和归一化方式相同的加噪图像。

ICD. ICD 使用以下扰动指令构建参考分支：

> 你是一个困惑的物体检测器，请对图像提供模糊的概览或印象。

对于 LLaVA，此指令会替换参考分支中的系统提示，而用户问题、图像和候选项保持不变。对于 InstructBLIP，此指令仅应用于参考分支的 Q-Former 指令；语言模型接收的提示词保持不变。

SID. 对于 LLaVA，SID 使用第二个语言模型层的末次查询注意力对 576 个视觉词元进行排序，并在后续层中保留注意力最低的 100 个词元。对于 InstructBLIP，我们在第五层计算视觉注意力，并保留注意力最低的 10% 的 Q-Former 词元，即 32 个视觉词元中的 3 个。从第六层开始应用所得掩码。

AvisC. AvisC 使用原始分支中的注意力识别视觉盲区词元。我们将层选择参数设为 $\gamma = 0.5$，将注意力阈值参数设为 $\lambda = 1.0$。参考分支会保留根据此标准选出的高注意力视觉词元，并将其余视觉表征置零。对于 LLaVA，此操作应用于视觉词元区域；对于 InstructBLIP，则应用于投影后的 Q-Former 视觉表征。

C.2.3. 对比强度实验

默认对比强度为 $\alpha = 1$。在对比强度实验中，我们仅针对单一真实标注设置下的 LLaVA-1.5-7B 将其提高至 $\alpha = 3$。所有其他设置，包括数据集、候选项排列、参考分支、随机种子和解码参数，均保持不变。CGM 由原始分支与参考分支之间的差异决定，因此不受 α 影响。提高 α 只会放大这一信号对最终候选项间隔和预测结果的影响。

C.3. 自适应合理性约束（APC）对 AMBER 和 POPE 的影响

我们研究了自适应合理性约束（APC）对 AMBER 生成式基准测试和 POPE 判别式基准测试的影响。我们以 LLaVA-1.5-7B 作为基础视觉语言模型，在受控解码条件下评估了四种对比解码方法：VCD、ICD、SID 和 AvisC，并以标准的 LLaVA-1.5-7B 解码作为基线。

表 6. AMBER 和 POPE 所用的采样与解码配置。

参数	AMBER	POPE
温度	1.0	1.0
Top-p	0.9	1.0
Top-k	50	已禁用
重复惩罚	1.0	–
对比强度 α	1.0	1.0
APC参数 β	0.1 或 0	0.1 或 0
批次大小	–	1
最大生成长度	256 个词元	8 个词元
VCD 扩散噪声步	500	500
基础随机种子	42	42

APC 将对比解码限制为原始分支下概率足够高的词元。具体而言，若候选词元 w 满足以下条件，则予以保留：

$$\ell_o(w) > \max_v \ell_o(v) + \log \beta, \tag{11}$$

其中，β 控制约束的强度。β越大，候选项筛选就越严格。我们将$\beta = 0.1$ 与 $\beta = 0$ 进行比较，后者表示 APC 已禁用的无约束设置。

实现细节。对于 AMBER，所有方法均使用相同的模型权重、数据、提示模板、随机种子和采样参数。我们将最大上下文长度设为 2,048 个词元。

对于 POPE，我们在基于 COCO 的随机、热门和对抗子集上评估模型。每个子集包含 3,000 个问题，总计 9,000 个评估样本，其中真实答案为是的有 4,500 个，为否的有 4,500 个。所有方法使用完全相同的问题、图像和分词后的提示。

解码配置。 两个基准测试均采用多项式采样，并使用不同的解码配置以适应各自的任务：AMBER 评估开放式生成，而 POPE 要求简短回答。完整配置汇总于表 6。

对于 AMBER，在对比解码和 APC 筛选之后应用温度、top-p 和 top-k 处理。随后通过多项式采样选出下一个词元。

对于 POPE，当生成结束序列词元时，生成会提前终止。生成并未明确限制为二元是/否 词汇表。

POPE 评估指标。所有方法和 APC 设置均采用相同的归一化和解析流程，将生成的回答转换为是/否预测结果。我们汇总随机、热门和对抗子集中的预测结果，并报告准确率、是准确率（TPR，真阳性率）、否准确率、FPR（假阳性率）、是率和 F1 分数。

C.4. 非目标内容对生成式评估的影响

C.4.1. 描述性词语密度

我们使用提示词“请详细描述这张图像。”，分析 LLaVA-1.5-7B 针对 500 张 MS-COCO 验证图像生成的自由形式回答。我们将这种标准生成方式称为基线，并将其与 VCD、ICD、SID、AvisC 和 FuzzyCD 进行比较。所有方法均采用温度为 1.0 的多项式采样、top-$p = 1.0$、top-$k = 0$，且最大生成长度为 256 个词元。对于每张图像，所有方法都使用相同的确定性采样种子。

使用 NLTK 的 Treebank 分词器对回答进行词元化，并使用其英语平均感知机词性标注器进行标注。统计词数时不计标点。我们将形容词（JJ、JJR 和 JJS）、副词（RB、RBR 和 RBS）、基数词（CD）以及空间关系表达式定义为描述词。空间词汇表包括多词表达式，如 在……前面、紧挨着 和 在背景中，以及单词关系词，如 上方、下方、后面 和 之间。优先匹配多词表达式，以避免重复计数。

对于回答 y，描述性词语密度定义为

$$q(y) = 100\,\frac{N_{\mathrm{desc}}(y)}{N_{\mathrm{word}}(y)}.$$

对于方法 m，我们报告其相对于基线的配对差值均值：

$$\Delta_m = \frac{1}{500}\sum_{i=1}^{500}\left[q(y_i^m) - q\big(y_i^{\mathrm{base}}\big)\right].$$

正值表示相对于基线，每生成 100 个词中的描述词数量有所增加。

C.4.2. 受控前缀干预

我们进一步考察生成前缀中的非对象描述性内容是否会影响后续生成中的幻觉。我们从同一批 500 张图像的基线运行结果中，保留首个生成句子中不含 CHAIR 对象幻觉的示例，并选取了 120 张图像。

对于每张图像，我们都构造一个中性前缀和一个描述性前缀。两者都以相同的基线首句开头，后接衔接短语“此外，”。中性前缀包含以下文本：

> 总体而言，图像以一种与前一句中已经描述的信息保持一致的方式呈现场景。内容以平常的方式延续，使用熟悉的措辞，没有添加任何不寻常的主张、具体细节或新的视觉信息。

表 7. LLaVA-1.5-7B 在 AMBER 生成任务上的完整响应幻觉评估结果。

方法	CH↓	封面↑	Hal↓	Cog↓
基线	10.6	50.7	46.4	4.7
VCD	10.5	51.5	46.4	4.5
ICD	8.9	51.5	40.2	4.0
SID	**8.0**	52.2	**39.0**	**3.3**
AvisC	10.7	**52.3**	48.0	4.8

相反，描述性前缀包含了额外的形容词、颜色术语及其他描述性表达，但没有引入新的对象提及：

> 总体而言，场景看起来明亮、色彩丰富、清晰、生动且细节丰富，红、蓝、绿、黄等色调随处可见。措辞简单直接，没有在此添加任何其他信息。

每个构造的前缀都会以教师强制的方式输入基线解码器，并在生成续写前从头重新计算模型状态。两个条件使用相同的图像和续写采样种子，温度为 1.0，top-$p = 1.0$、top-$k = 0$，总生成长度上限为 256 个词元。仅评估生成的续写中是否存在幻觉。

为控制续写长度差异，每张图像的续写都会截断至较短的词元长度。使用标准 CHAIR物体词汇表和 MS-COCO 真实标注，语料级 CHAIR_i is 计算为 $\mathrm{CHAIR}_i = \sum_i H_i / \sum_i O_i$，其中 H_i 和 O_i 分别表示幻觉对象提及数和对象提及总数。CHAIR_s is 计算为续写中至少包含一个幻觉对象的图像所占比例。

结果。 中性前缀条件在 135 次对象提及中产生了 32 次幻觉，其中 120 个续写中的 24 个出现了幻觉，对应的 $\mathrm{CHAIR}_i = 23.70\%$ 和 $\mathrm{CHAIR}_s = 20.00\%$。相比之下，描述性前缀条件在 195 次对象提及中产生了 47 次幻觉，并且在 120 个续写中的 38 个出现了幻觉，对应的 CHAIR_i =24。10% 和 $\mathrm{CHAIR}_s = 31.67\%$.

C.5. 方法生成的前缀对后续生成中幻觉的影响

桌子 7显示了完整响应评估中各去幻觉方法之间的明显差异。SID 的整体表现最佳，CH、Hal 和 Cog 分数分别为 8.0、39.0 和 3.3。ICD 的表现也优于基线，相应分数为 8.9、40.2 和 4.0，而 VCD

的表现与基线相当。相比之下，尽管 AvisC 取得了最高的 Cover 分数 52.3，但其 CH、Hal 和 Cog 分数分别为 10.7、48.0 和 4.8，均略差于基线的 10.6、46.4 和 4.7。为解释 AvisC 在完整响应中的相对较弱表现，图 7进一步考察了方法生成的前缀如何影响后续生成的幻觉率。

在固定基线续写的情况下，将基础前缀替换为相应的方法前缀，会使所有四种方法的幻觉率上升。具体而言，VCD 的幻觉率从 21.26% 上升至 22.51%，增加了 1.25 个百分点；ICD 从 21.26% 上升至 22.45%，增加了 1.19 个百分点；SID 从 21.26% 上升至 22.50%，增加了 1.24 个百分点。对于 AvisC，幻觉率从 21.26% 上升至 23.28%，增加了 2.02 个百分点。这些结果表明，方法生成的前缀通常会对后续解码造成干扰，从而提高续写的幻觉率。值得注意的是，AvisC 前缀造成的增幅远大于其他方法。

当固定每种方法自身的续写策略时，同样的趋势依然存在。相较于基础前缀，VCD、ICD、SID 和 AvisC 的方法生成的前缀分别使幻觉率增加了 0.85、0.61、0.19 和 1.44 个百分点。SID 前缀的影响最小，而 AvisC 前缀再次带来了最大的幻觉率增幅。两组比较均得出一致结果，表明前缀替换本身是导致后续生成中幻觉增加的重要因素，尽管这一影响的幅度因方法而异。

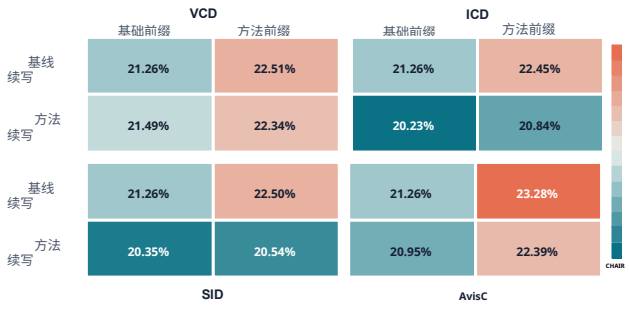


图 7。前缀-续写析因实验。各项数据为报告的幻觉率。“Base”和“Method”表示在固定前缀之后解码的续写类型。

C.6. 不同候选选择设置下的鲁棒性

候选项的构建方式会改变真实答案与幻觉候选项之间竞争的难度。因此，我们考察不同候选选择策略下观察到的方法之间的差异是否具有稳健性。表 8 比较了四种设置：与场景相关、共现、随机和平衡。在这些设置下，基线准确率从 58.2% 变化至 70.6%。具体而言，共现设置产生的基线准确率最低，而随机设置产生的基线准确率最高，这表明不同的候选构建策略会形成难度各异的竞争。

表8。不同候选选择设置下的鲁棒性。在四种候选选择设置下报告均值 CGM、正CGM率和准确率。

选项设置	方法	CGM	Positive-CGM	Acc.
与场景相关	基线	—	—	59.1%
	VCD	−0.106	39.8%	58.0%
	ICD	0.420	93.3%	64.4%
	SID	1.015	70.6%	68.7%
	AvisC	3.005	97.6%	73.6%
共现	基线	—	—	58.2%
	VCD	−0.068	40.8%	56.5%
	ICD	0.421	92.9%	63.0%
	SID	1.088	73.0%	69.6%
	AvisC	3.040	97.6%	73.0%
随机	基线	—	—	70.6%
	VCD	−0.034	36.8%	69.9%
	ICD	0.368	95.0%	75.1%
	SID	1.078	66.8%	78.7%
	AvisC	3.844	99.4%	86.1%
平衡	基线	—	—	65.6%
	VCD	−0.084	37.1%	63.6%
	ICD	0.382	91.1%	69.9%
	SID	1.102	71.4%	73.0%
	AvisC	3.354	98.1%	79.2%

尽管任务难度存在这些差异，CGM 的方向和排序在全部四种设置中仍**保持一致**。在场景相关、共现、随机和平衡设置下，VCD 的 CGM 值始终为负，分别为 -0.106、-0.068、-0.034 和 -0.084，正值率仅在 36.8% 至 40.8% 之间。相比之下，ICD、SID 和 AvisC 在每种设置下的 CGM 值均为正。ICD 的 CGM 值介于 0.368 至 0.421 之间，正值率为 91.1%–95.0%；SID 的 CGM 值为 1.015–1.102，正值率为 66.8%–73.0%；AvisC 的 CGM 值最高，介于 3.005 至 3.844 之间，且正值率始终较高，为 97.6%–99.4%。因此，在全部四种设置下，CGM 排序始终为 AvisC $>$ SID $>$ ICD $>$ VCD，这表明该排序并不依赖于任何特定的候选选择策略。

D. 标注缺失的鲁棒性

COCO 对象标注并不完整。因此，图像中存在但未被标注的对象可能会被错误地视为不存在。这一问题不仅会影响对象共现统计的估算，也会影响由此产生的幻觉候选项。因此，我们考察此类标注缺失是否会改变我们对不同方法所产生的对比校正的结论。

实验设置。 我们在新抽取的 200 张图像上进行实验，并通过人为移除对象实例标注来模拟标注不完整的情况。我们考虑两种缺失机制。在均匀缺失机制下，每个符合条件的对象实例被移除的概率相同。在优先缺失小型对象机制下，较小的实例更有可能被移除，以近似图像标注中小型对象更容易被漏标的情况。我们考虑 $q = 10\%$ 和 $q = 20\%$ 的缺失率，其中 q 表示被移除的符合条件的对象实例标注所占比例。

由于即使缺失率相同，不同的标注移除集合也可能对基准构建产生不同影响，因此我们进一步考察 $q = 10\%$ 缺失率下的这种变化。对于每种缺失机制，我们生成 50 组独立的缺失实现，每组对应于以相同的 $q = 10\%$ 缺失率移除的一组不同标注。我们采用两个标准来描述每组实现的影响。首先，受污染问题是指这样的问题：某个已从标注中移除、但实际存在于图像中的对象随后被选为幻觉候选项。其次，变化问题是指至少有一个结构化幻觉候选项不同于$q = 0$ 设置的问题。我们先按受污染问题的数量对这些实现排序，再按变化问题的数量排序，并选取 P10、P50 和 P90 对应的实现，分别作为轻度、中位数和重度条件。这些条件的缺失率相同，但标注缺失的具体模式对基准构建的影响程度不同。

对于每种条件，我们都会重新推导图像级对象集、重新计算类别共现概率，并重新构造三个结构化幻觉候选项。随机幻觉候选项固定不变，作为对照。每个问题均通过五种循环答案位置排列进行评估。对于所有 CD 方法，我们使用 LLaVA-1.5-7B 和 $\alpha = 1$ 及 $\beta = 0$ 评估基线、VCD、ICD、SID 和 AvisC。

结果。 表 9 报告了均匀缺失和小目标优先缺失条件下的结果。在这两种缺失机制下，主要的 CGM 模式均保持高度稳定。

表9. 均匀缺失和优先缺失小对象情况下的鲁棒性。

		均匀分布			优先缺失小对象		
条件	方法	CGM	正例-CGM	准确率	CGM	正例-CGM	准确率
$q = 0$	基线	–	–	70.0	–	–	70.0
	VCD	0.031	44.1	68.5	0.031	44.1	68.5
	ICD	0.377	93.9	72.0	0.377	93.9	72.0
	SID	1.250	72.5	76.5	1.250	72.5	76.5
	AvisC	3.251	98.4	80.5	3.251	98.4	80.5
$q = 10\%$ mild	基线	–	–	71.5	–	–	68.5
	VCD	0.030	43.8	70.5	0.028	44.3	69.0
	ICD	0.384	94.5	77.5	0.378	94.5	72.5
	SID	1.282	72.6	82.0	1.267	72.1	78.5
	AvisC	3.251	98.9	82.0	3.246	98.5	81.5
$q = 10\%$ 中位数	基线	–	–	68.0	–	–	71.0
	VCD	0.031	43.8	67.0	0.023	44.1	72.0
	ICD	0.382	93.5	72.0	0.374	94.4	74.0
	SID	1.270	73.9	79.5	1.286	72.5	80.0
	AvisC	3.258	98.6	78.5	3.243	98.5	81.0
$q = 10\%$ 严重	基线	–	–	65.0	–	–	64.0
	VCD	0.040	44.4	64.5	0.032	44.4	66.0
	ICD	0.381	93.4	74.0	0.374	93.6	73.0
	SID	1.275	73.0	78.0	1.280	72.8	76.5
	AvisC	3.243	98.5	80.0	3.244	98.3	80.0
$q = 20\%$ 中位数	基线	–	–	71.0	–	–	69.5
	VCD	0.044	45.0	67.0	0.035	46.3	69.0
	ICD	0.372	93.9	74.5	0.380	93.9	75.0
	SID	1.214	72.3	77.5	1.315	73.8	78.5
	AvisC	3.229	98.3	80.0	3.239	98.4	80.5

AvisC 始终取得最大的 CGM，均值范围为 3.229 至 3.258，正例-CGM 范围为 98.3% 至 98.9%。SID 的 CGM 保持在 1.214 至 1.315 之间，明显高于 ICD；其正例-CGM 范围为 72.1% 至 73.9%。尽管 ICD 的 CGM 较小，但其正例-CGM 仍保持在 93.4% 至 94.5% 之间，表明其参考分支提供了高度一致的校正方向。相比之下，VCD 的 CGM 始终接近于零，其正例-CGM 一直低于 50%，这表明其参考分支无法持续提供有益的对比方向。

更重要的是，按 CGM 均值排序后，所有九种条件的顺序均保持不变：

$$\text{AvisC} > \text{SID} > \text{ICD} > \text{VCD}.$$

因此，无论是均匀模拟缺失标注，还是优先针对小型物体进行模拟以更好地近似现实中的标注偏差，重新计算共现统计并重建干扰项都不会 改变基于 CGM 的对比解码方法排序。在所有条件下，正例 CGM 的排序也保持稳定，遵循

$$\text{AvisC} > \text{ICD} > \text{SID} > \text{VCD}.$$

表10. 三个数据集上的鲁棒性。

数据集	方法	CGM	正例-CGM	Acc.
COCO	基线	—	—	65.6%
	VCD	-0.084	37.1%	63.6%
	ICD	0.382	91.1%	69.9%
	SID	1.102	71.4%	73.0%
	AvisC	3.354	98.1%	79.2%
Open Images	基线	—	—	68.4%
	VCD	-0.055	42.6%	67.3%
	ICD	0.424	94.7%	73.8%
	SID	1.431	76.6%	78.1%
	AvisC	3.439	98.8%	83.5%
Objects365	基线	—	—	54.6%
	VCD	-0.114	40.0%	53.7%
	ICD	0.345	90.1%	59.7%
	SID	1.096	69.8%	68.4%
	AvisC	2.753	94.1%	69.2%

D.1. 跨数据集的鲁棒性

由于 **CONCOR**是在 COCO 上实例化的，我们考察观察到的对比解码方法之间的差异是否与其图像和物体分布有关。我们在 Open Images 上重复评估 [14] 和 Objects 365 [28], , 这两个数据集提供了不同的图像来源和物体分布，以评估主要的 CGM 模式能否推广到其他数据集。

表 10 总结了 LLaVA-1.5-7B 在 COCO、Open Images 和 Objects365 各自的 200 张图像上的结果。这三项实验均采用相同的五选一、单一真实答案 MCQ 设置，并将基线与四种对比解码方法进行比较：VCD、ICD、SID 和 AvisC。

三种数据集上各方法的相对表现模式一致。就 CGM 而言，四种对比解码方法的排序相同：AvisC $>$ SID $>$ ICD $>$ VCD。AvisC 在每个数据集上都取得最高的 CGM 均值，其次是 SID 和 ICD，而 VCD 在三个数据集上的 CGM 均值均为负值。

按正CGM率排序的结果也一致：AvisC $>$ ICD $>$ SID $>$ VCD。AvisC 在所有三个数据集上的正CGM率均超过 94%，而 ICD 始终高于 90%。尽管其 CGM 均值低于 SID，但较高的正CGM率表明其在各个样本上的提升更加稳定。SID 带来的平均增益更大，但样本级差异也更明显。

在所有三个数据集上，五候选采样准确率的排序一致：AvisC $>$ SID $>$ ICD $>$基线 $>$ VCD。AvisC 在每个数据集上的准确率都最高，而 SID 和 ICD 的表现始终优于基线。相比之下，VCD 的表现略低于基线



图8。COCO Val2014 中的候选项构造示例。指定的真实候选项为背包，而人、餐桌、斑马和水槽 分别被选为高度共现、与场景相关、场景无关和随机幻觉候选项。

在所有三个数据集中。总体而言，各数据集上方法排序一致，表明 CGM 识别出的主要差异对于测试的图像来源具有稳健性，并非 COCO 所特有。

E. 幻觉候选项构造示例

图8 展示了 **CONCOR** 中的一个示例（COCO 图像 ID 37900）。指定的真实目标对象是背包。该图像中经 COCO 标注的物体类别集合为

$$P_I = \{backpack, book, cell\ phone, knife\}. \quad (12)$$

对于每个未标注候选类别 o，我们计算其与每个已标注类别 $a \in P_I$ 的条件共现，计算方式为 $C(a,o) = N(a,o)/N(a)$，其中 $N(a)$ 是包含类别 a 的 COCO 图像数量，而 $N(a,o)$ 是同时包含 a 和 o 的图像数量。表11报告了四个选定幻觉候选项的这些数值。

候选项人的 $S_{\max}$ 最高，为 0.9039，因此被选为高度共现候选项。排除人后，餐桌 的 S_{scene} 最高，为 0.2560，因此被选为场景相关候选项。对于每个已标注类别，斑马的共现值都接近于零，因此它是场景无关候选项。最后，从剩余的未标注类别中均匀采样水槽 作为随机候选项。

表11。图中用于构造幻觉候选项的条件共现分数 8。每个条目报告 $C(a\ o\ =\ N\ a\ o\ /N\ a\ a\ o\ S_{\text{scene}}\ S_{\max}\ ,\ )\ (,\ )()$ 已标注类别与未标注候选项之间的条件共现。最后几行报告由此得到的共现分数和分配给每个选定幻觉候选项的候选类型。

已标注类别 a	$N(a)$	人	餐桌	斑马	*sink*
背包	1,832	0.9039	0.0535	0.0011	0.0076
book	1,828	0.4152	0.2030	0.0000	0.0481
手机	1,695	0.7900	0.1504	0.0006	0.0177
刀	1,410	0.5312	0.6170	0.0007	0.1121
$S_{\text{场景}}$	–	**0.6601**	**0.2560**	**0.0006**	**0.0464**
$S_{\max}$	–	**0.9039**	**0.6170**	**0.0011**	**0.1121**
候选类型	–	高度共现	与场景相关	场景无关	随机