arXiv:2303.02601v1 [cs.CL] 5 Mar 2023

基于知识的视觉问答反事实查询

狄奥多蒂斯托伊库[1], 玛丽亚林佩莱乌[1] 和乔治斯塔穆[1]

[1]雅典国立理工大学电气与计算机工程学院 *AILS* 实验室

摘要

视觉问答（VQA）是一项广受欢迎的任务，它结合了视觉与语言，相关实现已在文献中屡见不鲜。尽管已有一些研究尝试解决视觉问答模型的可解释性和鲁棒性问题，但其中极少有研究以模型无关的方式，利用反事实来探究这些挑战。在本研究中，我们提出一种系统化方法，通过反事实扰动解释视觉问答模型的行为并考察其鲁棒性。为此，我们利用结构化知识库，对语言模态进行确定性、最优且可控的词级替换，随后评估模型对这些反事实输入的响应。最后，我们基于反事实响应定性提取局部和全局解释，这些解释最终被证明有助于理解视觉问答模型的行为。通过针对输入问题中不同词性的多种扰动类型，我们比较模型在不同对抗情境下的响应，从而深入了解模型的推理过程。总体而言，我们揭示了模型决策过程中的潜在偏差，以及影响其性能的预期与非预期模式；我们的分析从定量和定性两方面印证了这一点。

关键词

视觉问答、知识图谱、可解释人工智能、反事实解释、鲁棒性

1. 引言

视觉语言（VL）学习 [1, 2, 3] 的普及无疑迅速推动了这一领域的发展，短时间内便为学界带来了各种令人瞩目的模型实现 [4, 5, 6, 7, 8, 9]。视觉问答（VQA）是一项视觉语言任务，在视觉对话 [10], 文本-图像检索 [11]和视觉常识推理 [12]等各类交互式视觉语言人工智能系统的发展中发挥了基础性作用。为此，众多现实世界应用都从视觉问答任务的新进展中显著受益，例如为视障人士提供辅助的系统 [13, 14] 以及自动驾驶汽车 [15]。

视觉问答涉及一个问题 q 和图像 I，二者的交互会生成文本答案 a。对视觉问答模型性能不断提升的追求不可避免地留下了一些未解决的问题，这些问题尤其归因于最先进的实现的黑箱特性 [16, 17, 18, 19, 20]。这种有限的访问权限使我们无法了解

载于 *A. Martin*、*K. Hinkelmann*、*H.-G.Fill*、*A.Gerber*、*D. Lenat*、*R. Stolle*、*F. van Harmelen*（编），《*AAAI 2023* 春季研讨会论文集：需要结合机器学习与知识工程的挑战》（*AAAI-MAKE 2023*），凯悦酒店，旧金山机场，加利福尼亚州，美国，*2023*年*3*月*27*日至*29*日。
dido.stoikou@gmail.com(T. Stoikou); marialymp@islab.ntua.gr (M. Lymperaiou); gstam@cs.ntua.gr (G. Stamou) 0000-0002-3653-0041 (T. Stoikou); 1-9442-4186 (M. Lymperaiou); 0000-0003-1210-9874 (G. Stamou)
©2023 年本文作者版权所有。允许依据知识共享署名 4.0 国际许可协议（CC BY 4.0）使用。CEUR 研讨会论文集（CEUR-WS.org）

此类模型在做出决策时所遵循的推理过程，凸显了它们可能产生任意行为的风险。这种危险主要在于偏差整合、缺乏适当关注的决策，以及结果缺乏可解释性和公平性的可能性。尤其是在基于此类系统做出关键决策时，其不透明性使它们对大多数应用而言不切实际，有时甚至具有危害性。这种不确定性表明，需要采用新的鲁棒性评估方法，并优先考虑视觉问答模型的透明度。

针对视觉问答模型去偏与可解释性的不同方法，着眼于问题的不同方面。例如，[21] 通过处理视觉模态上的变换来考察视觉问答的鲁棒性与可解释性，因为他们认为问题主要源于视觉偏差，即图像概念之间不必要的相关性。总体而言，现有研究主要关注视觉偏差 的影响，而非语言偏差，并将其视为视觉问答模型缺乏鲁棒性的原因。其他研究采用基于注意力的策略，这类策略需要深入了解模型架构，因此无法高效应对这些系统的黑盒 特性。为此，各种模型特定方法已证明在其严格限定的框架内行之有效，但无法推广至其他模型的评估，因此其适用范围仅限于单一特定情况。

我们认为，要解决视觉问答模型的可解释性难题，需要采用反事实方法，对问题q进行词级扰动。因此，我们不再一味探索视觉模态，而是考察语言在视觉问答模型中潜藏的偏差和虚假相关性所起的作用，同时追踪并解释其不透明的决策过程。我们提出的反事实扰动形式如下：“如果将问题 q 中的词 X 替换为词 Y，视觉问答模型会给出什么回答？”具体而言，我们将词视为概念，对其进行尽可能小且可行的变换，以促使模型的回答发生变化；随后，通过记录模型对各种此类变换的行为，进行有启发性的比较。我们所实施的反事实扰动，完全由层级知识 结构提供的确定性保障所引导。利用这些知识源，我们提出的变换不仅能针对每个具体的语言概念进行最优设计，而且在实现策略方面也完全具有可解释性。

从观察单个数据样本在不同语言扰动下的模型局部响应 入手，我们进一步识别出全局模式，即模型面对一组特定的受扰动概念时的整体行为。在此基础上，我们提出用于刻画模型响应的全局规则，这些规则能够指出模型的弱点，揭示哪些概念可能损害其鲁棒性和确定性。这一过程揭示了模型可能内化的偏差，并由此解释模型为何生成某个答案而非另一个答案；因此，无需访问模型的内部架构，我们也能深入了解模型的推理过程。我们的方法可推广到任何视觉问答模型及与之适配的数据集，因为它以完全模型无关的策略处理这一问题。总之，我们的贡献如下：

1. 我们按照层次化知识源的指示，对问题 $q \in Q$ 应用各种结构化的词级替换，设计反事实输入。我们的方法与模型无关，因为我们将任何视觉问答模型都视为黑箱。
2. 我们从模型对反事实

输入产生的意外响应中得出局部解释。3. 通过总结所有 $q \in Q$ 的局部模型行为，我们提取出一些全局解释，揭示模型对每种已设计反事实输入的整体响应。

2. 视觉问答（VQA）

视觉问答（VQA）最初于 [22], 被引入，属于多模态学习任务，因为它同时接收视觉和语言模态作为输入。具体而言，视觉问答模型 M 接收来自集合 i 的图像 I 以及相关问题 q，这些问题属于预定义的问题集合 Q，模型需要通过给出自然语言答案 q 准确回答这些问题 a。上述答案既可以是开放式的（由 M生成），也可以从一组预定义的候选答案中选出 A。一般而言，问题 q 形式多样，涵盖不同的计算机视觉子问题，例如目标识别与检测、属性和场景分类，以及计数 [23]。此外，更复杂的问题涉及更复杂的过程，例如物体之间的空间关系和常识推理。根据具体情况，视觉问题 q 会有选择地关注图像的不同区域 I，包括背景细节和潜在语境。因此，语言输入所关注的词语概念也会因图像—问题对而异。图 I 1 展示了一个视觉问答任务示例，其中包含一幅图像 q、相关问题 a 以及视觉问答模型 M 对这些问题给出的答案。1。



图 1： 从 Visual Genome数据集 [24],中检索的图像和自由形式问题示例，旨在用于视觉问答任务。所显示的答案由 ViLT 模型 [5]作为响应给出。

我们关于语言替换的反事实方法围绕以下基本问题展开：“如果我们在问题 q 中用词语 Y 替换词语 X，M 其回答是什么？”。所实施的反事实 $X \rightarrow Y$ 替换应在语义上具有最小性，并且

在语言上可行。最小性是指保持与原词X含义相近的替换。例如，同义词能够保持这一最小性约束。为了确保替换的语义最小性，我们利用词汇知识源（如 WordNet [25]），通过选择与概念 X 最接近且满足特定约束的概念 Y，从而提供可能的最小 $X \rightarrow Y$ 转换。语言可行性要求有意义的替换始终涉及相同的词性（POS）；例如，名词只能被名词替换，而不能被动词替换。总体而言，这类$X \rightarrow Y$替换应用于整个 Q集合，每次针对一种词性。

此类反事实问题能够触发替代性的模型响应。因此，输入中的$X \rightarrow Y$概念替换可能会导致输出中出现替代性的 $X' \rightarrow Y'$ 响应，也可能不会。探测潜在的输出变化对于了解模型所遵循的推理过程 M具有很高的信息价值，可凸显对 M决策过程影响程度不同的概念或概念族。因此，在 q 上实施的反事实替换能够为模型观察到的行为提供有用的解释，并提升其可解释性，同时仍将其视为黑箱结构。

3. 相关工作

视觉问答模型 自视觉问答 [22], 的开创性工作以来，已有多项研究通过提出先进的模型架构或更具挑战性的数据集来扩展这一范式。当前处理视觉问答任务的最先进模型主要基于视觉语言变换器骨干网络；因此，ViLBERT [6],VisualBERT [4], FLAVA [8], ALBEF [9],ViLT [5]等模型主导了近期的视觉问答文献，并在相关基准数据集上展现出快速的性能提升。就数据集而言，对原始视觉问答（VQA-v2）的改进建议加入与同一问题对应的相似图像对 q，但这些图像对会导致不同的答案 [20]。Visual Genome（VG）是另一个大规模数据集，包含大量场景图像、对象、属性和关系标注，以及视觉问答对 [24]。我们的方法在VQA-v2和VG数据集上均进行了测试。其他流行的视觉问答数据集包括Flickr30k-Entities [26], COCO-QA [27], Visual7W [28] 等。关于视觉问答及相关主题的详细分析，我们建议读者参阅近期的专业综述论文 [1, 2, 3, 29, 30, 31]。

视觉问答中的可解释性 关于视觉问答中可解释性与鲁棒性的研究主题 [32, 33, 34], , 人们已经提出了多种方法，包括注意力图 [35, 36], 以及其他模型特定方法 [37]。通过反事实实现的策略相对较新，而现有的尝试则侧重于视觉扰动 [21], 掩蔽[18, 38], 在训练阶段引入反事实 [39, 38] ，以及探讨原始样本与反事实样本之间关系的方法 [40]。

语言扰动已有多项研究采用词级语言扰动，尽管这些研究针对的都是纯语言任务，主要是文本分类 [41, 42,43, 44, 45], , 也包括语义相似度 [46] 和机器翻译 [47]。我们采用的同义词替换和随机名词删除扰动，受到以下研究的启发： [42], 利用 WordNet [25]来指导替换。颜色扰动改编自 [46],

我们据此构建了一个基于颜色距离的适当层级结构。其余涉及名词和动词替换的替换方法均为全新构想。

4. 方法

我们的框架以一个数据集 D 为输入，该数据集包含对齐图像I、组成集合 Q 的文本问题以及候选文本答案 A。稍后我们将展示在 Visual Genome (VG) [24] 和 VQA-v2 [20], 上的结果，这些数据集符合上述要求。

我们选择 ViLT [5] 作为预训练视觉问答模型 M的概念验证。不过，我们提出的方法并不局限于 ViLT，因为它只考虑输入（问题）和输出（答案）。ViLT 接收来自 D 的问题 $q \in Q$ 和图像 $i \in I$，随后生成答案 a，而非从候选答案 $a \in A$中进行选择；由于这种行为是若干视觉问答模型的固有特点，因此允许对准确率指标作较宽泛的定义十分重要。更具体地说，ViLT 生成的输出 a 是自然语言文本，而同一个答案在英语中有许多不同的表达方式。在这种情况下，通过启发式方法将生成的答案与 A 中的真实答案进行比较，以判断 M 的预测是否准确。对所有 Q,I 对重复这一预测过程，并确定其答案是否正确后，我们最终提取准确率分数acc_Q，以反映正确答案占所有生成答案的比例。

我们的词语替换由外部知识源指导，每次针对不同的词性（名词、动词和形容词）。具体而言，WordNet 知识图谱 [25]提供了 VG 和 VQA-v2 词汇表中广泛存在的大量常见词之间的层级关系。因此，通过将特定词语与其 WordNet 匹配项相连接，并遵循第 4.1节所述的层级关系，构建替换词对。此外，我们扩展了 Matplotlib 颜色[2] 中提出的关系， [46], 形成了一个颜色相关性层级。该颜色层级基于每种 Matplotlib 颜色的 RGB 值所对应的颜色距离，更多细节见第 4.1节。7

随后，我们将设计的扰动应用于数据集问题 $q \in Q$，从而得到反事实问题 $q^* \in Q^*$。对于每次替换，我们将 M 对反事实问题 $q^* \in Q^*$的回答作为反事实问题准确率 acc^*_Q，并将其与真实准确率分数acc_Q进行比较。在整个过程中，我们通过测量acc_Q 与acc^*_Q之间的差异，评估 M 的回答是否以及如何发生变化，以此作为其对语义相关概念替换的鲁棒性指标。

尽管独立准确率分数有助于基准测试，但它们不足以解释为何原始 q 与反事实输入 q^*之间会出现如此差异。为此，我们单独考察在扰动作用下生成的 a发生变化的样本，并以如果 q 中的概念发生变化，那么 a 就会（错误地）发生变化的形式获得局部解释。对这些局部规则进行汇总后，便可得到全局解释，推导出适用于 D多个样本的如果—那么关系。

我们在图 2中展示了我们方法的可视化概览。来自 $q \in Q$ 以红色标出的词表示要替换的概念，而以蓝色标出的词则表示知识驱动的替换，这些替换会生成反事实问题 $q^* \in Q^*$。

[2]Matplotlib颜色

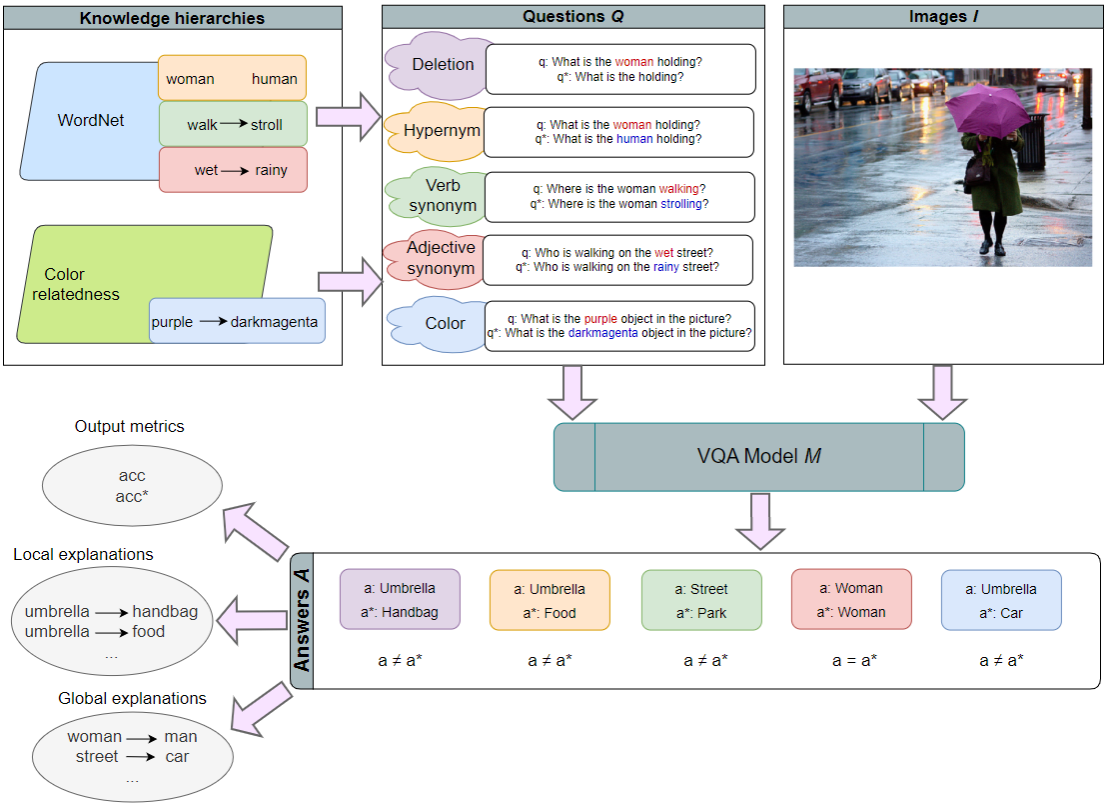


图 2： 我们提出的基于知识的反事实视觉问答框架概览。

4.1. 扰动

在我们的工作中，我们会对Q的语言表示进行多种替换或删除操作。这种反事实策略利用了Q的多种不同形态学属性，并试图揭示对模型响应影响最大的语义 a。我们的目标是通过这些反事实扰动，促使 M 产生不同的响应，从而识别 M 面对不同概念时行为的变化。由此，我们可以推断 M中存在的潜在偏差或鲁棒性不足之处。上述替换可根据所用的知识来源和目标词性分为以下几类。表 1总结并列出了具有代表性的替换示例。

A. Wordnet层级结构： 知识驱动的词语替换涉及将问题 $q \in Q$ 中的一个名词替换为具有层级关系的词语（下位词、上位词、同级词），或将动词和形容词替换为其同义词。通过使用 WordNet层级结构的确定性结构，可以保证替换的质量和相关性，从而实现可控且最优的词级替换。

- **同义词：** 我们对原始问题 $q \in Q$ 中的形容词和动词采用同义词变换。例如，根据 WordNet，“talk”和“speak”是同义动词，而“small”和“minuscule”是形容词同义词。
- **上位词 - 下位词：** WordNet 提供了更一般以及更具体 的名词概念，分别以上位词 和下位词 的形式呈现。例如，

对于给定的名词（例如“狗”），我们可以提取其直接名词上位词（例如“犬科动物”），或直接下位词（例如“拉布拉多犬”）。

- **同义词兄弟：** 我们通过沿着 WordNet 知识树先向上遍历一步，再向下遍历一步，构造名词同级词替换。同义词兄弟指共享同一个直接上位概念的名词实体。例如，根据 WordNet，“胡萝卜”和“萝卜”是同义词兄弟，因为它们的上位概念都是“植物根”。

B. 颜色相关性层级： 语义上相似、因而 RGB 值彼此接近的颜色，在颜色相关性层级中也会彼此接近。例如，Matplotlib颜色“紫罗兰色”和“兰花色”在颜色层级中相距较近（两者之间的颜色距离为 6.16），而“紫罗兰色”和“深天蓝色”则相距较远（两者的颜色距离为 207.88）。可以从该颜色相关性层级中替换为距离较远或较近的颜色，从而得到以下**颜色最大替换**和**颜色最小化**颜色替换。最大化/最小化替换既可以涉及常见颜色（已存在于数据集中），也可以涉及不常见颜色（属于 Matplotlib颜色列表，但不在 VG/VQA-v2词汇表中）：

- **颜色最大替换：** 对于提及某种特定颜色的问题，我们根据原始问题的输入a 与 M 以及扰动问题的输出 a^* 来质疑输出 $q \in Q$。在 q^* 中，将原始颜色替换为与其颜色距离很大的颜色，例如“紫罗兰色” $\rightarrow$ “深天蓝色”。在这一类别中，我们还使用较少见的颜色实例（即“天蓝色”“绿松石色”“鲑鱼色”）来检验模型。这种替换偏离了最初要求进行最小 更改的反事实问题；不过，与相关最小 更改进行比较，将凸显不同颜色距离对最终 acc_Q^*所造成的差异。

- **颜色最小化：** 与上述内容一致，我们使用颜色距离最小的颜色进行颜色替换，例如“紫罗兰色” $\rightarrow$ “兰花色”。同样，我们还使用较少见的颜色替换来检验模型。

C. 删除操作： 我们随机选择每个问题中的一个名词 $q \in Q$并将其删除。

表 1用于反事实查询的问题扰动示例。

扰动	问题
原始	你看到那只白色的小狗了吗?
颜色最大替换	你看到那只**黑色**小狗了吗?
颜色最小化	你看到那只**米色**小狗了吗?
同义形容词	你看到那只白色**微小的**狗了吗?
同义动词	你**观看**那只白色的小狗吗?
上位词名词	你看到那只白色的小**犬科动物**了吗?
下位词名词	你看到那只白色的小**拉布拉多犬**了吗?
同级词名词	你看到白色的小**狼** 了吗?
删除名词	你看到白色的小 _ 了吗?

替换和删除操作是量化视觉问答模型 M是否理解特定问题-图像对，还是其输出在很大程度上取决于有偏

估计的一种绝佳方式。通过这种方式，我们可以揭示被错误地整合进 M 的虚假相关性。进行颜色替换的动机，源于输入数据集（VG 和 VQA-v2）中与颜色相关的问题数量众多。我们通过对 M 施加与原始颜色差异极大的颜色扰动（**颜色最大替换**实验），旨在检测 M 是否能正确、合理地感知这一语义上的重大变化。在大多数 **颜色最大替换**实验中，我们预期 M 会改变其回答 a^*；反之则表明模型存在忽视颜色属性的潜在模式。同样，我们进行 **颜色最小化**替换实验，以考察模型面对颜色概念的轻微变化时的行为。我们预期这一实验中的替换对模型的回答 a^* 几乎不会产生影响，甚至毫无影响。相反的行为则会揭示模型对特定颜色存在偏差，由此可得出结论： M 无法恰当地、稳健地适应轻微的颜色变化并进行相应泛化。当然，在最小化和最大化两种情况下，不常见的颜色替换都会带来更大的挑战，因为 M 需要适应性地应对数据集外颜色概念。

关于**同义词** 替换，我们旨在研究模型能否有效处理轻微的形态语言变更，同时保持原意。在这种情况下，如果模型未能妥善作答（给出另一个 $a^* = a$），就会暴露其对特定语义的过拟合，使模型在词汇上不灵活，因此无法抵御语义上无关紧要的扰动。**上位词-下位词** 扰动旨在刻画模型相应地进行泛化和具体化的能力，同时保持可靠的鲁棒性水平。上位词和下位词关系是现实世界中人们深刻理解的概念，因此也被纳入大规模数据集，而这些数据集广泛用于视觉问答模型的预训练。因此， M 也应能够正确理解并推理这些关系。理想情况下，我们期望 M 在上位词替换时保持相同的回答，而在下位词替换时则能合理地给出更具体的回答，以体现其对语义具体程度的把握。通过**同级词** 替换实验，我们旨在衡量相应的具体化能力。根据具体情况，我们期望 M 适当地修改或保持回答，以验证其理解并区分不同但仍相关的含义的能力。

最后，我们实施了**删除**实验，理想情况下预期 M的性能会下降。 acc_Q^* 下降的幅度取决于被删除名词对于问题含义的重要性。因此，无偏的 M 应当能够确定这种重要性并采取相应行动，而不会得出通过无法辩护的回答表达出的无根据结论。

5. 实验

我们使用准确率指标呈现结果，该指标展示了模型预测答案与真实答案之间的相似程度，既适用于每个数据集的原始 Q ，也适用于反事实问题集 Q^*。在我们的分析中，准确率不足以针对模型面对特定概念时的行为提供具体的情境解释和洞见。然而，准确率仍然展现了模型在

所实施的反事实扰动下效率变化的高层次表现。由于 ViLT 模型是在 VQA-v2 数据集上训练和优化的，因此可以预期，与 VG 相比，它在该数据集上的表现会更好（两个数据集包含相似的词汇表）。我们在表2 & 3中呈现的结果确实验证了这一观察结果，这些结果表明，在所有实施的实验中，前者的 acc_Q 始终高于后者。我们指出，acc_Q 每项实验的得分仅包含相应的问题，例如，颜色实验仅包含提及颜色的问题。这导致了每项实验中原始acc_Q 得分之间的差异。然而，在两个数据集中，当acc_Q 面对 acc^*_Q 反事实问题时，我们都注意到每项实验中的 M 与 $q^* \in Q^*$之间存在类似差异。这通常可能表明潜在偏差的存在： M 表现出对原始 $q \in Q$的某种过拟合，这使得它在处理仅经过轻微扰动的反事实问题时效率较低。在所有实验中，准确率从原始acc_Q 到acc^*_Q 的下降幅度约为 15–20% 或更高。

两个数据集检索到的准确率的详细结果见表格2 （基于颜色的替换）和表格3 （基于WordNet的替换和名词删除操作）。具体而言，在 VQA-v2 中，**颜色最大替换**使常见颜色的准确率下降 34.2%，罕见颜色则下降 37.4%。即使是语义上最小的替换（**颜色最小化**实验），常见和罕见颜色的准确率也都下降了 31%。至于 VG，进行常见颜色的准确率下降了 38.2%，罕见颜色的准确率下降了 52.2%，此时使用的是**颜色最大替换**。相应地，使用常见和罕见颜色进行**颜色最小化**替换时，准确率均下降了 35%。

表 2VQA-v2 和 VisualGenome（VG）上颜色扰动的准确率。常见扰动指替换为数据集内的颜色，而罕见扰动指替换为任意 Matplotlib 颜色。

	acc_Q%		acc^*_Q%（常见）		acc^*_Q%（罕见）	
扰动	VQA-v2	VG	VQA-v2	VG	VQA-v2	VG
颜色最大替换	69.6	46.9	45.8	29.0	43.6	22.4
颜色最小化	70.0	47.5	48.3	30.9	48.3	30.9

表3基于 WordNet 的扰动在 VQA-v2 和 Visual Genome (VG) 上的准确率。

	acc_Q%		acc^*_Q%		acc_Q 降低 %	
扰动	VQA-v2	VG	VQA-v2	VG	VQA-v2	VG
同义形容词	75.1	47.0	56.9	37.4	20.6	20.4
同义动词	76.8	52.3	64.1	44.4	16.5	15.1
上位词名词	75.2	54.6	60.8	41.5	19.1	24.0
下位词名词	75.1	53.5	56.3	36.1	25.0	32.5
同级词名词	76.9	54.0	54.0	33.4	29.8	38.1
删除名词	76.9	53.9	59.1	36.5	23.1	32.3

发现全局模式能够更深入、更有针对性地了解模型 M 的鲁棒性。为此，我们指出，在所研究的所有情况下，我们关注的并非问题 q 的真实答案，而是反事实问题的答案 a^* 与原始答案 a 之间的差异；原始答案由 M 给出，无论 a 是否正确。我们选择这种方法，是因为我们希望了解模型在存在反事实输入时决策方式的变化；与衡量 a^* 在语义上偏离真实响应的程度相比，这种方法能提供更多信息。

基于对实验结果的深入考察，以及对以下局部解释的汇总（如后续图中所示），我们推导出了一些有意义的全局规则，这些规则既体现了 M 对我们反事实问题的鲁棒性 $q^* \in Q^*$，也提供了可靠的解释，揭示了模型决策背后的推理过程。此外，我们还分析了 M 中存在的潜在偏差，这些偏差可由这些全局规则逻辑推导得出。在以下图中，我们将原始 q用 a 标出，其中红色表示原始内容；反事实 q^*用 a^* 标出，其中蓝色表示反事实内容。

5.1. 颜色最大化解释

在**颜色最大替换**中，我们注意到 M 会错误地保持相同答案$a^* = a$，即使我们将颜色灰色和银色替换为任何其他语义最大化颜色，无论常见与否（带下划线标示）。因此，我们发现模型存在与这两种颜色相关的偏差：将它们替换为其他颜色后，模型不会做出合乎逻辑的决策，也无法对这种替换进行恰当推理。图3a展示了一个相关示例。

然而，与上述情况相反，M 当我们将颜色绿色和红色替换为任意差异较大的颜色（无论常见还是罕见，带下划线标出），如图3b所示时，模型会在逻辑上修正其答案。因此，模型能够识别并定性理解这些替换，并未对这两种颜色形成任何有问题的关联。



(a)**问题**：什么东西围绕着银色/黑色/海军蓝消防栓?
答案：柱子/柱子/柱子。



(b) **问题**：How many 玻璃杯 have 红色/金色/深绿松石色 葡萄酒?
a: 6/0/0.

图3： 颜色最大替换**颜色最大替换** 反事实扰动的局部解释。

这一观察表明，M 对鲜明且视觉上易于区分的颜色更敏感，而往往会忽略涉及较中性色彩的变化，转而关注物体身份（例如

图中的“消防栓”和“柱子” 3a）。如果平等地考虑所有问题语义，那么更不确定的 a^*（例如，模型的回答 a^* 可能是“什么也没有”）会更合适。

5.2. 颜色最小化解释

根据我们对语义最小化颜色替换的实验，我们得出以下全局规则：当我们将颜色灰色和紫色替换为任何其他密切相关的颜色时，M 往往会给出相同的答案 $a^* = a$。因此， M 不会对这种颜色变化给予应有的重视，这一事实凸显了模型在前述两种颜色相关问题上的稳健表现。无论我们用常见颜色还是不常见颜色进行替换，模型都能保持这种不变的行为，如图4a所示。

相比之下， M 在我们将绿色替换为任何其他语义相似的颜色时，都会重新给出答案，无论这种颜色是常见还是不常见。因此，这种细微变化会使其产生混淆，无法给出有意义的答案，如图4b所示。对于反事实问题，即便给出更不确定的答案（例如“什么都没有”），也比模型给出的语义不相符的答案（“公共汽车”和“包”，而非“灯”）更合适。同样， M 在将粉色替换为常见的最小颜色，以及将银色替换为不常见的颜色时，也表现出类似的变化。



（a）**问题**：横幅上印有哪个组织的标志紫-色/蓝色/梅紫色横幅?
答案：奥运会/奥运会/奥运会。



（b）**问题**：正在举行什么绿色/森林绿/橄榄绿?
答案：灯/公交车/包。

图 4： **颜色最小化**反事实扰动的**局部解释**。

5.3. 同义形容词解释

一般而言，问题中出现的形容词-名词搭配包含某种共同的特殊概念含义，这不同于形容词在句子中独立、分开存在时所具有的独立含义。在这种情况下，我们注意到模型 M 会改变其答案 a^*，当我们对问题中的形容词实施同义词替换时。这一发现表明， M 模型能够从定性上感知此类形容词-名词搭配的含义，并相应地区分其响应，如图5a所示。

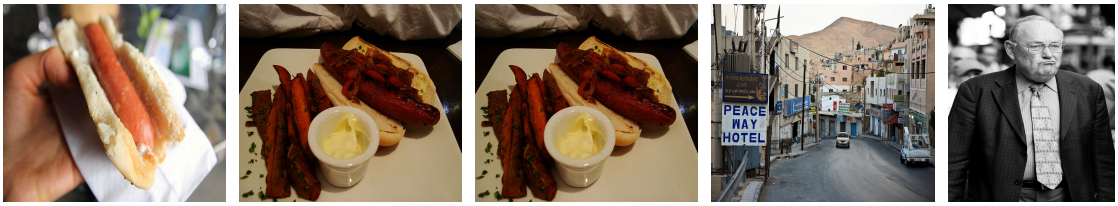
另一项发现与 M 在面对形容词的词汇上正确的同义词时能否正确调整回答有关；不过，该同义词并不完全适用于

问题给定的语言环境。因此，我们得出结论： M 能够理解形容词在句子语境中的含义（“典型食物”有意义，但“特色食物”没有意义），如图 5b所示。

此外，我们还推导出一条全局规则，涉及 M 的行为：当我们将一个具有多重含义的形容词替换为其某个同义词时，尽管该同义词对前述含义而言是最佳选择，却不适合被替换形容词所处的语义语境（例如“美味的”与“令人愉快的”）。我们注意到，在这种情况下， M面对这类替换时表现稳定，这证明它能够以语境化方式推理形容词，不会被不适用于问题确切语境的同义词误导 q。图5c展示了这一观察的一个示例。

图5d展示了尺寸相关的形容词替换。我们观察到， M 对此类替换尤其稳健，因此能够正确捕捉深层含义，不会偏向特定词语。这条全局规则体现了 M恰当处理语义和语境等价的形容词替换的灵活性。

最后， M 在处理罕见或晦涩的形容词同义词时表现不稳定，如图5e所示。因此，它在处理这类罕见形容词时存在词汇弱点，并且可能偏向它更熟悉的特定词语。



(a) **问题**: Is this a 热的/狂暴的狗?
答案: 是/否。

(b) **问题**: 属于哪种餐 is this kind of food 典型-的/独特的?
a: 午餐/热狗。

(c) **q**: How 美味的- 美味的/令人愉快的 看起来怎么样?
a: 非常/不太。

(d) **问题**: Is this a 小/小小的 镇?
回答: 是/是。

(e)**问题**: 是否 man look 高兴/愉悦?
回答: 否/是。

图5：针对**同义形容词**反事实扰动的局部解释。

5.4. 同义动词解释

关于涉及动词同义词的替换， M 在处理具有多重含义的动词替换时并不特别稳定，如图6a所示。这表明模型难以从多个含义中区分出正确且符合预期的含义。

我们提取出的一个更具体的规则是， M 当我们将动词“see”替换为其定性同义词时，会错误地改变其原始答案 a 。图6b给出了一个相关示例。这一发现表明，模型对单词“see”存在不恰当的依附，这被视为一种偏差。这是一个更普遍情形的例子：在特定语境中，最优同义词未必是最佳选择。在这类情况下，模型往往会改变其回答，正如这个例子所示。

当需要处理简单或常见的英语词汇时，该模型 M 表现出令人满意的鲁棒性

动词，这意味着它在简单的词汇挑战中已经具备了一定程度的多功能性，如图 6c所示。



(a) 问题: What 说/状态 STAPLES?
回答: 什么都没有/纽约。



(b) 问题: Do you 看/理解 any 摩托车头盔?
一: 否/是。



(c)问: 墙壁是完成/制作在一个夏日色调?
是: 是/是。



(d) q: 什么种类的鸟类被拍摄到/被可视化?
a: par- 鹦/鹦鹉。

图6： 针对**同义动词**反事实扰动的局部解释。

最后， M 在替换的动词对应于名词对应形式（例如，picture 作为动词和名词）甚至形容词对应形式（例如，pictured）时仍相当稳定，例如图6d所示。因此， M 能够捕捉此类动词在问题语境中的一般含义；用对应的名词（picture→visualization）或形容词（pictured→visualized）进行等价替换，大多会产生相同的反事实响应 a^*。

5.5. 上位名词解释

通过上位名词替换，我们得出结论， M 对于涉及生物的替换尤其稳健，例如动物或人类。这表明它能够恰当地推理此类概念所遵循的层级关系，如图7a所示。

相反， M 无法清晰区分与服装类型相关的概念，也就是说，将其替换为更宽泛的概念时，它往往会错误地改变答案。因此， M 无法很好地泛化到此类实体，并且表现出偏向更具体、更明确的服装类型的倾向。图7b给出了一个相关示例。

作为上述内容的扩展， M 在面对过于宽泛、包容性强且多义的上位词替换时表现出不稳定性。因此，当我们用一个概念普遍性高得多的最优上位词替换一个名词时， M 无法定性地感知支配它们的层级关系，因此给出了错误答案，如图7c所示。

5.6. 下位名词解释

与上位词替换实验类似， M 能够在下位词替换指向生物实体的情况下做出恰当回答。因此，它能够正确感知更具体的生物实体概念中的特殊性，如图8a所示。

相应地， M 在代表服装的下位词替换中也表现出稳定性。因此，它能够熟练地处理更具体的服装相关实体



(a) **问**：这只猫/猫科动物在哪里?
答： 床/床。



(b)问：所有球员都穿着黑色——色 衬衫/服装?
答： 否/是。

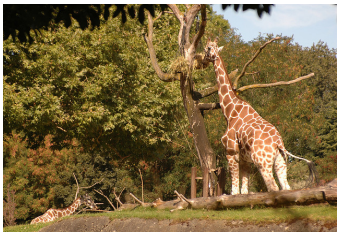


(c) 问：是否有多个植——盘子上的盘/底座?
a： 是/否。

图7： 上位词名词反事实扰动的**局部解释**。

并根据具体情况，通过适应变化恰当地调整回答。图中给出了一个相关示例8b。

相反，对于指代交通工具的下位词替换，模型并未表现出鲁棒性。因此，模型偏向于这类更宽泛的概念，无法充分理解其专业化含义，如图8c所示。



（a）**q**：有动物/齿冠在吃东西吗?
a： 是/是。



(b)问题：所有球员都穿着——着黑色衬衫/衬衫?
a： 否/是。



(c)问题：摩托车后面的物体是什么那辆摩托车/小型摩托车?
a： 标志/标志。

图8： 下位词名词的局部解释**下位词名词** 反事实扰动。

5.7. 同级名词解释

关于同级名词替换，我们注意到一个普遍规律： M 在同级名词指建筑物房间或房屋时，区分能力不足。例如，在图9a中， M 无法恰当地分辨所描述的内部空间，很容易被涉及不同功能地点的替换所迷惑。 M 在图9d中也会受到干扰，此时替换的是同级交通工具。具体而言， M 即使图像中出现了一个不存在的概念（自行车→卡车），仍坚持自己的答案。这表明， M 它反而越过了语言模态的语境，基于视觉模态给出一个“轻松”的答案，因为图像中唯一出现的鸟是鹦鹉。在这种情况下，鸟的相关位置——在男人的自行车/卡车上——被忽略了。

测试涉及动物的同级概念时，我们观察到一种有趣的行为，如图9e所示。在这种情况下，M 似乎绕过了对图像的推理，而是根据它很可能在预训练阶段获得的知识作答（斑马是黑白相间的）。然而，M 并未被马→斑马替换所迷惑；若受此迷惑，它就会得出“斑马是棕色的”这一错误的事实陈述。图 M 展示了另一个未受迷惑的例子，见图9c。在这种情况下，M 对于表示人体部位的同级实体表现得非常一致，这意味着它能正确识别它们之间的差异，而不会随意将它们归为一类。此外，M 展现了正确的推理过程：当同级概念之间的含义差异很大时，它会给出不同的答案，例如图→9b中的空气水。

总体而言，同级名词替换带来了丰富的见解，在许多情况下，它不均衡地依赖于q或 I 来得出答案，而不是给出不确定结果（例如，在图9d、9e的示例中回答“什么都没有”，这与图9c中正确的推理类似）。总体而言，这表明 M 面对不同的同级词对时表现不稳定，在相同概念距离的不同替换下会产生不可预测的结果。



(a)**问**：这个浴-房间/工作间整齐吗?
答：是/是。

(b) **问**：那些是风筝 in the 空中/水中?
a：是/否。

(c) **q**：她在做什么戴在她身上的头部/喉咙?
a：hel-遇见/什么也没有。

(d) **问题**：什么鸟在男人的自行车/卡车?
a：鹦鹉/鹦鹉。

(e) **q**：它是什么颜色这只马/斑马?
a：棕色/黑色和白色。

图9： 针对**同级词名词** 反事实扰动的局部解释。

5.8. 删除名词解释

关于涉及删除名词的反事实问题，我们首先观察到以下模式：当被删除的名词对问题中已出现的另一个名词起限定作用时，M 即使删除该名词，仍会给出原来的答案。因此，我们发现其倾向于依附于被限定名词，同时没有充分关注限定词名词。因此，M 会随意回答此类问题，尽管其答案不能被认为是错误的；人类也可能给出相同的答案，尤其是在是非问句中。图10a中给出了一个相关示例。

我们发现的另一种模式涉及这样的问题：问题询问某个名词的颜色，而该名词已被删除。在这种情况下，M 往往会回答图像中最显眼的颜色，而不考虑本应由该颜色描述的名词已不存在，如图10b所示。当然，我们认为这种行为是合理的，因为人类很可能也会以同样的方式回答这类问题。类似地，对于询问某个

已被删除的名词的位置的问题， M 会回答给定图像中最显眼的实体。图10c展示了一个相关示例。

最后，我们列出一些名词。我们注意到，在被要求作答时， M 并未给予它们应有的重视，如图10d所示。具体而言，即使删除这些词， M 仍经常返回最初的答案。这些词包括：图像、照片、人类、男人、动物、房间。一般来说，这些词在问题中十分常见，通常会与其他更具体的实体搭配出现。再次说明，这种行为是合理的。

总而言之，我们观察到，随机删除一个名词会导致模型表现出相当符合预期的行为，这种行为受给定图像中占主导地位的视觉概念驱动。



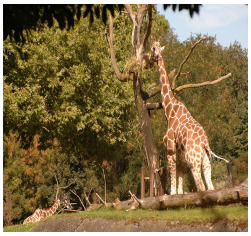
(a)**问**：女人的头发是向后扎起来的吗?
答： 不/不。



(b) **问**：是什么颜色的浴室?——
答： 黄色/白色。



(c) **问**：在哪里 蛋糕?
答： 桌子/桌子。



(d)**q**：有多少只动物在这张照片中?
a： 2/2。

图10： 删除名词 反事实扰动的局部解释。

6. 结论与未来工作

对视觉问答模型进行反事实扰动，能够为模型鲁棒性和结果的可解释性提供新颖且有用的见解。在我们的研究中，我们提出了一种基于知识的反事实框架，专门针对问题中的替换操作。具体而言，我们的框架提出了多种词级语言变换，以黑盒方式探测选定的视觉问答模型，并考察反事实问题的出现是否会导致意料之外的模型响应。通过这一过程，我们揭示了潜在的语言偏差；同时，通过定性地提取全局规则，我们也为模型行为提供了有信息价值的解释，最终呈现出这些已有偏差。我们在 Visual Genome 和 VQA-v2 数据集上使用 ViLT 模型进行概念验证，结果展现了该方法的优势，并以不依赖特定模型的方式突显了引发模型偏差的概念。作为对该方法的直接扩展，我们计划对数据集中的答案应用相同的语言扰动，以处理那些基于多项选择答案进行推理的视觉问答模型。未来，我们计划将该方法扩展到其他相关的视觉语言任务，例如文本-图像检索、视觉蕴含和视觉常识推理；另一个研究方向则是设计针对视觉模态的反事实扰动。

致谢

这项研究工作得到了希腊研究与创新基金会（HFRI）在HFRI第三轮博士奖学金申请项目下的支持（奖学金编号：5537）。

参考文献

[1] A. Mogadala、M. Kalimuthu、D. Klakow，《视觉与语言研究融合趋势：任务、数据集与方法综述》，《人工智能研究杂志》71 (2021) 1183–1317。URL：http://dx.doi.org/10.1613/jair.1.11688。doi:10.1613/jair.1.11688.[2] Y. Du、Z. Liu、J. Li、W. Zhao，《视觉语言预训练模型综述》(2022)。[3] M. Lymperaiou、G. Stamou，《知识增强多模态学习综述》，ArXiv abs/2211.12328 (2022)。[4] L. H. Li、M. Yatskar、D. Yin、C.-J. Hsieh、K.-W. Chang，《VisualBERT：一种简单且性能优异的视觉与语言基线模型》，2019。URL：https://arxiv.org/abs/1908.03557。doi:10.48550/ARXIV.1908.03557.[5] W. Kim、B. Son、I. Kim，《ViLT：无需卷积或区域监督的视觉与语言变换器》，2021。URL：https://arxiv.org/abs/2102.03334。doi:10.48550/ARXIV.2102.03334.[6] J. Lu、D. Batra、D. Parikh、S. Lee，《ViLBERT：面向视觉与语言任务的任务无关视觉语言表征预训练》，2019。URL：https://arxiv.org/abs/1908.02265。doi:10.48550/ARXIV.1908.02265.[7] V. Kazemi、A. Elqursh，《展示、提问、关注并回答：一种强大的视觉问答基线模型》，2017。URL：https://arxiv.org/abs/1704.03162。doi:10.48550/ARXIV.1704.03162.[8] A. Singh、R. Hu、V. Goswami、G. Couairon、W. Galuba、M. Rohrbach、D. Kiela，《FLAVA：一种基础语言与视觉对齐模型》，2021。URL：https://arxiv.org/abs/2112. 04482。doi:10.48550/ARXIV.2112.04482.[9] J. Li、R. R. Selvaraju、A. D. Gotmare、S. Joty、C. Xiong、S. Hoi，《先对齐后融合：利用动量蒸馏进行视觉与语言表征学习》，2021。URL：https://arxiv. org/abs/2107.07651。doi:10.48550/ARXIV.2107.07651.[10] A. El-Nouby、S. Sharma、H. Schulz、D. Hjelm、L. E. Asri、S. E. Kahou、Y. Bengio、G. W. Taylor，《讲述、绘制并重复：基于持续语言指令生成和修改图像》，2019。 arXiv:1811.09845.[11] S. R. Dubey，《基于深度学习的基于内容的图像检索十年综述》，《IEEE视频技术电路与系统汇刊》(2021) 1–1。URL：http: //dx.doi.org/10.1109/TCSVT.2021.3080920。doi:10.1109/tcsvt.2021.3080920.[12] R. Zellers、Y. Bisk、A. Farhadi、Y. Choi，《从识别到认知：视觉常识推理》，2019。 arXiv:1811.10830.[13] K. Baker、A. Parekh、A. Fabre、A. Addlesee、R. Kruiper、O. Lemon，《勺子在水槽里：在厨房中协助视障人士》，载于：推理

与交互会议 (ReInAct 2021) 论文集，计算语言学协会，瑞典哥德堡，2021，第 32–39 页。URL：https://aclanthology.org/2021.reinact-1.5。[14] C. Chen、S. Anjum、D. Gurari，《为视障人士提出的视觉问题定位答案》，2022。URL：https://arxiv.org/abs/2202.01993。doi:10.48550/ARXIV.2202.01993.[15] H. Ben-Younes、Éloi Zablocki、P. Pérez、M. Cord，《利用多级融合解释驾驶行为》，《模式识别》123 (2022) 108421。URL：https://www.sciencedirect.com/science/article/pii/S0031320321005975。doi:https://doi.org/10.1016/j.patcog.2021.108421.[16] R. Cadene、C. Dancette、H. Ben-younes、M. Cord、D. Parikh，《Rubi：减少视觉问答中的单模态偏差》(2019)。URL：https://arxiv.org/abs/1906.10169。doi:10.48550/ARXIV.1906.10169.[17] V. Agarwal、R. Shetty、M. Fritz，《迈向因果视觉问答：通过不变与协变语义编辑揭示并减少虚假相关性》，2019。URL：https://arxiv.org/abs/ 1912.07538。doi:10.48550/ARXIV.1912.07538.[18] L. Chen、X. Yan、J. Xiao、H. Zhang、S. Pu、Y. Zhuang，《为稳健视觉问答合成反事实样本》，2020。URL：https://arxiv.org/abs/2003.06576。doi:10.48550/ARXIV.2003.06576.[19] P. Anderson、X. He、C. Buehler、D. Teney、M. Johnson、S. Gould、L. Zhang，《用于图像描述和视觉问答的自底向上和自顶向下注意力》，2017。URL：https://arxiv.org/abs/1707.07998。doi:10.48550/ARXIV.1707.07998.[20] Y. Goyal、T. Khot、D. Summers-Stay、D. Batra、D. Parikh，《让视觉问答中的“视觉”发挥作用：提升图像理解在视觉问答中的作用》，2016。URL：https: //arxiv.org/abs/1612.00837。doi:10.48550/ARXIV.1612.00837.[21] Z. Boukhers、T. Hartmann、J. Jürjens，《Coin：用于解释视觉问答的反事实图像生成》，2022。URL：https://arxiv.org/abs/2201.03342。doi:10.48550/ARXIV.2201.03342.[22] A. Agrawal、J. Lu、S. Antol、M. Mitchell、C. L. Zitnick、D. Batra、D. Parikh，《视觉问答》，2015。URL：https://arxiv.org/abs/1505.00468。doi:10.48550/ARXIV.1505.00468.[23] K. Kafle、C. Kanan，《视觉问答：数据集、算法与未来挑战》，《计算机视觉与图像理解》163 (2017) 3–20。URL：https://doi.org/10.1016%2Fj.cviu.2017.06.005。doi:10.1016/j.cviu.2017.06.005.[24] R. Krishna、Y. Zhu、O. Groth、J. Johnson、K. Hata、J. Kravitz、S. Chen、Y. Kalantidis、L.-J. Li、D. A. Shamma、M. S. Bernstein、F.-F. Li，《Visual Genome：利用众包密集图像标注连接语言与视觉》，2016。URL：https://arxiv.org/abs/1602.07332。doi:10.48550/ARXIV.1602.07332.[25] C. Fellbaum，《WordNet：电子词汇数据库》(1998)。[26] B. A. Plummer、L. Wang、C. M. Cervantes、J. C. Caicedo、J. Hockenmaier、S. Lazebnik，《Flickr30k Entities：收集区域到短语对应关系，以构建更丰富的图像到句子模型》，载于：2015年IEEE国际计算机视觉会议 (ICCV)，2015，第 2641–2649 页。doi:10.1109/ICCV.2015.303.[27] M. Ren、R. Kiros、R. Zemel，《探索用于图像问答的模型与数据》，2015。arXiv:1505.02074。

[28] Y. Zhu, O. Groth, M. Bernstein, L. Fei-Fei，Visual7w：图像中的有依据问答，2016年，第4995–5004页。doi:10.1109/CVPR.2016.540.[29] Y. Zou, Q. Xie，视觉问答综述：数据集与方法，载于：2020年第2届国际信息技术与计算机应用会议（ITCA），IEEE，2020年。URL：https://doi.org/10.1109%2Fitca52113.2020.00069。doi:10.1109/itca52113.2020.00069.[30] M. Banchhor, P. Singh，视觉问答综述，载于：2021年第2届全球技术进步会议（GCAT），2021年，第1–5页。doi:10.1109/GCAT52182.2021.9587797.[31] H. Sharma, A. S. Jalal，视觉问答的方法、数据集和评估指标综述，《图像与视觉计算》116（2021）104327。URL：https://www.sciencedirect.com/science/article/pii/S0262885621002328。doi:https://doi.org/10.1016/j.imavis.2021.104327.[32] A. Panesar, F. I. Doğan, I. Leite，通过利用深度信息并改进可解释性来提升视觉问答，载于：2022年第31届IEEE机器人与人类交互通信国际会议（RO-MAN），2022年，第252–259页。doi:10.1109/RO-MAN53752.2022.9900586.[33] K. Alipour, J. P. Schulze, Y. Yao, A. Ziskind, G. Burachas，视觉问答多模态交互式解释研究（2020）。URL：https://arxiv.org/abs/ 2003.00431。doi:10.48550/ARXIV.2003.00431.[34] J.-H. Huang, M. Alfadly, B. Ghanem, M. Worring，评估视觉问答模型的鲁棒性，2019年。URL：https://arxiv.org/abs/1912.01452。doi:10.48550/ARXIV.1912.01452.[35] J. Lu, J. Yang, D. Batra, D. Parikh，用于视觉问答的层次化问题-图像协同注意力，2016年。URL：https://arxiv.org/abs/1606.00061。doi:10.48550/ARXIV.1606.00061.[36] M. Jiang, S. Chen, J. Yang, Q. Zhao，精彩答案及其所在之处：沉浸式问题引导的视觉注意力，载于：2020年IEEE/CVF计算机视觉与模式识别会议（CVPR），2020年，第2977–2986页。doi:10.1109/CVPR42600.2020.00305.[37] F. Sammani, T. Mukherjee, N. Deligiannis，NLX-GPT：一种用于视觉及视觉语言任务自然语言解释的模型，2022年。URL：https://arxiv.org/abs/2203.05081。doi:10.48550/ARXIV.2203.05081.[38] L. Chen, Y. Zheng, Y. Niu, H. Zhang, J. Xiao，合成反事实样本并进行训练，以提升视觉问答的鲁棒性，2021年。URL：https://arxiv.org/abs/2110.01013。doi:10.48550/ARXIV.2110.01013.[39] E. Abbasnejad, D. Teney, A. Parvaneh, J. Shi, A. van den Hengel，反事实视觉与语言学习，载于：2020年IEEE/CVF计算机视觉与模式识别会议（CVPR），2020年，第10041–10051页。doi:10.1109/CVPR42600.2020.01006.[40] Z. Liang, W. Jiang, H. Hu, J. Zhu，学习对比反事实样本，以提升视觉问答的鲁棒性，载于：2020年自然语言处理实证方法会议（EMNLP）论文集，计算语言学协会，线上，2020年，第3285–3292页。URL：https://aclanthology.org/2020.emnlp-main.265。doi:10.18653/v1/2020.emnlp-main.265。

[41] S. Ren, Y. Deng, K. He, W. Che，通过概率加权词显著性生成自然语言对抗样本，载于：第57届计算语言学协会年会论文集，计算语言学协会，意大利佛罗伦萨，2019年，第1085–1097页。URL：https://aclanthology.org/P19-1103。doi:10.18653/v1/P19-1103.[42] J. Wei, K. Zou，Eda：提升文本分类任务性能的简易数据增强技术，2019年。URL：https://arxiv.org/abs/1901.11196。doi:10.48550/ARXIV.1901.11196.[43] S. Garg, G. Ramakrishnan，BAE：基于BERT的文本分类对抗样本，载于：2020年自然语言处理实证方法会议（EMNLP）论文集，计算语言学协会，线上，2020年，第6174–6181页。URL：https://aclanthology.org/2020.emnlp-main.498。doi:10.18653/v1/2020.emnlp-main.498.[44] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, Y. Qi，Textattack：用于自然语言处理对抗攻击、数据增强和对抗训练的框架，2020年。URL：https: //arxiv.org/abs/2005.05909。doi:10.48550/ARXIV.2005.05909.[45] A. Karimi, L. Rossi, A. Prati，AEDA：一种更简便的文本分类数据增强技术，载于：计算语言学协会会议论文集：2021年EMNLP，计算语言学协会，多米尼加共和国蓬塔卡纳，2021年，第2748–2754页。URL：https://aclanthology.org/2021.findings-emnlp.234。doi:10.18653/v1/2021.findings-emnlp.234.[46] M. Lymperaiou, G. Manoliadis, O. M. Mastromichalakis, E. G. Dervakos, G. Stamou，面向语言模型对视觉概念语义相似度的可解释评估，2022年。URL：https://arxiv.org/abs/2209.03723。doi:10.48550/ARXIV.2209.03723.[47] J. Wan, J. Yang, S. Ma, D. Zhang, W. Zhang, Y. Yu, Z. Li，Paeg：面向神经机器翻译的短语级对抗样本生成，2022年。URL：https://arxiv.org/abs/2201. 02009。doi:10.48550/ARXIV.2201.02009。