专题综述

从因果视角量化和评估多模态大模型中的单模态偏见

北京大学 陈美琪 张岩　新加坡管理大学 曹艺馨　上海人工智能实验室 陆超超

本文是北京大学、新加坡管理大学、与上海人工智能实验室团队合作研究的成果。论文探讨了多模态大模型（MLLMs）过度依赖单一模态偏见（biases），如语言偏见和视觉偏见，而在复杂的多模态任务中给出错误的答案的问题。针对这一问题，论文提出了一个因果框架来解释视觉问答（VQA）问题中的偏见。通过该框架，论文设计了一个因果图来阐明论文探讨了多模态大模型在VQA问题上的预测，并通过深入的因果分析评估偏见的因果效应。基于因果图，论文介绍了一个新的数据集MORE，包含12000个VQA实例，这些实例设计用来挑战多模态大模型的能力，需要它们进行多跳推理和克服单一模态偏见。大量的定量和定性实验为未来的研究提供了有价值的见解。论文[1]的项目网页公开在 https://opencausalab.github.io/MORE.

一、研究背景

继大语言模型（LLMs）[2,3]的成功之后，多模态大模型（MLLMs）[4, 5] 已被提出用于各种视觉-语言任务[6, 7]。尽管取得了令人鼓舞的结果，但它们是否真正理解图像和文本在多模态推理上下文中的含义仍然不清楚。如图 1 所示的基于知识的视觉问题回答(VQA)问题中，当被问到“哪个国家将在这个场馆之后举办下一届世界杯？”时，多模态大模型，比如 GPT-4V 和 LLaVA，可能会捕捉到“下一届世界杯”的语言偏见，并认为下一届世界杯将是“在卡塔尔举行的 2022 年世界杯”（这也是过时的知识），同时忽略了图像中呈现的确切场馆。同样，当呈现出伦敦的“碎片大厦”图像时，受到视觉偏见的影响，多模态大模型直接识别出“代表性建筑是碎片大厦”，而忽略了问题中提到的特定限制“在柏林”。这些固有的问题对多模态大模型的推理能力提出了重大挑战，尤其是面对更复杂的问题时。

图 1　单一模态偏见过度依赖的例子。多模态大模型由于语言偏见（左侧图像下划线的文本所示）和视觉偏见（右侧图像）错误生成了答案。

为了调查多模态大模型对这种单一模态偏见的过度依赖问题，我们提出了一个因果框架来解释和量化语言和视觉偏见。具体来说，我们首先定义了多模态大模型在 VQA 问题上预测的因果图。因果图是基于预测过程中的各种因果因素构建的，如图像和问题文本。然后，我们在 VQA 问题的背景下识别了一系列干预，从而通过do-演算(do-calculus)[8]来确定单一模态偏见对多模态大模型预测能力的因果效应。通过量化这些因果效应，我们可以评估多模态大模型对单一模态偏见的敏感性和鲁棒性。

Datasets	Knowledge-based	Multi-hop Reasoning	Answer Type	Unimodal Biases Evaluation	Rationale	#Size
Visual7W[9]	✗	✗	Open-ended	✗	✗	327.9K
VQA(v2)[10]	✗	✗	Open-ended	✗	✗	1.1M
FVQA[11]	✓	✗	Open-ended	✗	✓	5.8K
OKVQA[12]	✓	✗	Open-ended	✗	✗	14K
S3VQA[13]	✓	✗	Open-ended	✗	✗	7.5K
A-OKVQA[14]	✓	✗	Open-ended	✗	✓	23.7K
INFOSEEK[15]	✓	✗	Open-ended	✗	✗	1.4M
MORE(Ours)	✓	✓	Open-ended	✓	✓	12K

表 1　MORE 和现有 VQA 数据集的对比

基于上述因果分析，我们创建了一个名为 MORE 的新数据集，包含 12,000 个 VQA 实例。该数据集通过引入专门的单一模态偏见评估，提升了现有 VQA 数据集。为了便于评估，我们采用多项选择题（MCQ）格式，每个实例由一张图像、一个问题和四个候选选项组成。图像来源于现有的 VQA 数据集。为了问题和选项的策划，我们纳入了一个知识图谱（KG），允许我们更好地模拟多模态大模型在因果图中导航对应的伪路径。具体而言，选项由一个正确答案和三个分别针对语言偏见、视觉偏见和多跳推理的干扰项组成。我们还为每个实例提供了 KG 中的推理路径，称为因果推理，为评估提供了可解释性。总得来说，与现有的 VQA 数据集相比，MORE 具有外部知识、多跳推理、单一模态偏见评估和推理路径，展现了更好的全面性。在六个领先的多模态大模型上的实验结果显示：1）大多数多模态大模型在 MORE 上的表现较差，明显倾向于依赖单一模态偏见。2）当处理多模态推理时，多模态大模型仍然难以实现精确的语义理解。

二、因果框架介绍

在本节中，受 Stolfo 等人工作[16]的启发我们首先介绍多模态大模型在 VQA 问题上预测的因果图。然后，我们利用因果图阐明 VQA 中固有的偏见，特别是视觉和语言偏见。最后，我们通过执行受控干预[8]来评估这些偏见对多模态大模型预测的因果效应。

1. 问题设置

我们考虑一个以实体为中心的 VQA 问题，记为M，由一个问题Q和一个图像I组成。图像描绘了一个特定的实体，问题与该实体相关。问题Q由两个不同的元素组成：核心语义内容S，传达问题的真实意图；和文本表面形式T，与问题的核心含义无关。模型的最终答案/预测由A表示。在本文中，我们使用小写字母表示其对应的大写变量的一个实例。

2. 多模态大模型预测的因果图

受人类认知中观察到的直观推理机制的启发[16,17]，我们在 VQA 问题m中制定了人类问题解决的因果机制：

$$s = f_c(q) \ , g = f_s(i)$$

其中，认知过程f_c被用来解析问题q中的核心语义含义s。然后，函数f_s将s与图像i相关联，产生最终答案g。我们在图 2 的绿色子图G_h中展示了这些机制。

与此相反，模型解决同一 VQA 问题 m 的可能的因果机制如下：

$$a = f_b(i, q)$$

其中，f_b作为一个黑匣子，使得模型考虑q的哪些方面以及它如何与图像i交互是不确定的。

为了进一步分析，我们在图 2 中绘制了完整的因果图中可能发生的所有可能的因果机制。值得注意的因果机制包括：

- 视觉偏见：模型可能通过因果路径$I \rightarrow A$直接关注图像I，导致视觉偏见的出现。
- 语言偏见：模型可能直接以两种方式处理问题Q：通过因果路径$Q \rightarrow S \rightarrow A$关注核心语义$S$，或

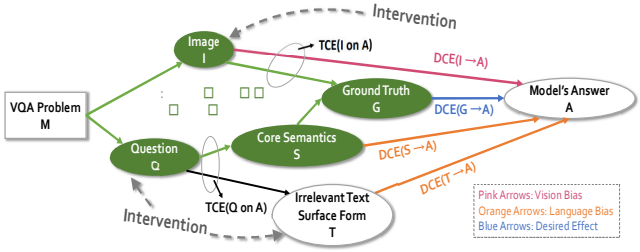

图 2　多模态大模型对 VQA 问题预测的因果图。

通过因果路径$Q \rightarrow T \rightarrow A$关注无关部分$T$。这两条路径都会导致语言偏见。

- 期望的因果机制
- 正确推理的本质在于模型对解决 VQA 问题所需的基本因果机制的把握。如图 2 的绿色子图G_h所示，它应该理解图像和问题如何共同贡献于正确答案G（通过$I \rightarrow G$和$S \rightarrow G$）。因此，模型的预测应该对正确答案的变化显示出敏感性和鲁棒性，即$G \rightarrow A$。没有任何假相关路径可以通过中介G而直接影响A。

基于上述分析，我们阐述模型在 VQA 问题上的敏感性和鲁棒性的概念：

- 敏感性：评估模型在正确答案变化时是否适当调整其预测，即A对G的变化做出反应。
- 鲁棒性：评估单模态偏见的直接因果效应，例如$I \rightarrow A$，$T \rightarrow A$，其中较低的效应表示对不改变正确答案的输入变化具有更好的鲁棒性。

3. VQA 偏见的因果分析

在定义了所期望的因果机制和单模态偏见的路径后，我们可以通过执行受控干预[8]来量化每个因子对另一个因子的因果效应。

因果干预 在 VQA 的上下文中，我们采用以下干预措施来量化图像和问题对模型预测的因果效应：

- 直接对图像I进行干预，将其替换为另一个图像I'。
- 对Q进行部分可控干预。问题Q可以通过两种方式修改：(i) 同时修改S和T，或 (ii) 修改T但保持S不变。

因果效应的计算 接下来，我们解释如何从干预中获得因果效应。考虑一个干预 $do(X{:}\, x \text{ -> } x')$，其中 $X \in \{I, Q, T\}$，并且 VQA 问题 $M = \{I, Q\}$。我们将干预前的分布 $\mathbb{P}(A \mid I, Q)$ 表示为 P，干预后的分布表示为 P'。

遵循 Pearl (1995) [8]提出的分布式因果效应定义，我们使用距离度量 δ 量化因子 X 在我们的因果图中的效应，即 $CE = \delta(P, P')$，其中 CE 表示因果效应，并且可以进一步细分为总因果效应（TCE，即通过所有从一个变量到另一个变量的定向因果路径的联合效应）或直接因果效应（DCE，即从一个变量到另一个变量的直接因果路径的效应，不经过任何中介变量）。

遵循 Stolfo (2022) [16] 的方法，我们通过评估预测结果的变化来量化因子X对模型答案A的因果效应，即

$$\delta_{cp}(P, P') := \mathbb{I}(a \neq a')$$

其中$a = arg\ max_x P(x)$，$a' = arg\ max_x P'(x)$，$\mathbb{I}$ 表示“答案改变”事件的指示器。

图像的因果效应 当对图像 I 进行干预时，我们可以得到 I 对 A 的因果效应大小，即：

$TCE(I\ on\ A) := E_{i' \sim P(I)}[\delta(P, P')]$，其中 $P' = \mathbb{P}(A \mid Q, do(I = i'))$。

注意，这个 TCE 包含了两条不同的路径，说明了 I 如何影响 A，如图 2 中所示：

- 路径 $I \text{ -> } G \text{ -> } A$ 代表我们希望模型采用的理想决策路线，其中它响应于正确答案的变化。
- 路径 $I \text{ -> } A$ 描述了模型可能学习到的一种虚假关联，其中它依赖于某些可能与训练语料库中的普遍存在相关的视觉上下文。

我们可以量化 I 对 A 的 DCE，即路径 $I \text{ -> } A$的强度，通过在每次对 I 进行干预时保持G不变来实现，即：

$DCE(I \text{ -> } A) := E_{i' \sim P(I \mid G)}[\delta(P, P')]$，其中 $P' = \mathbb{P}(A \mid Q, do(I = i'))$。

问题的因果效应 对于问题，通过对 Q 进行干预，我们可以计算 Q 对 A 的总因果效应，即：

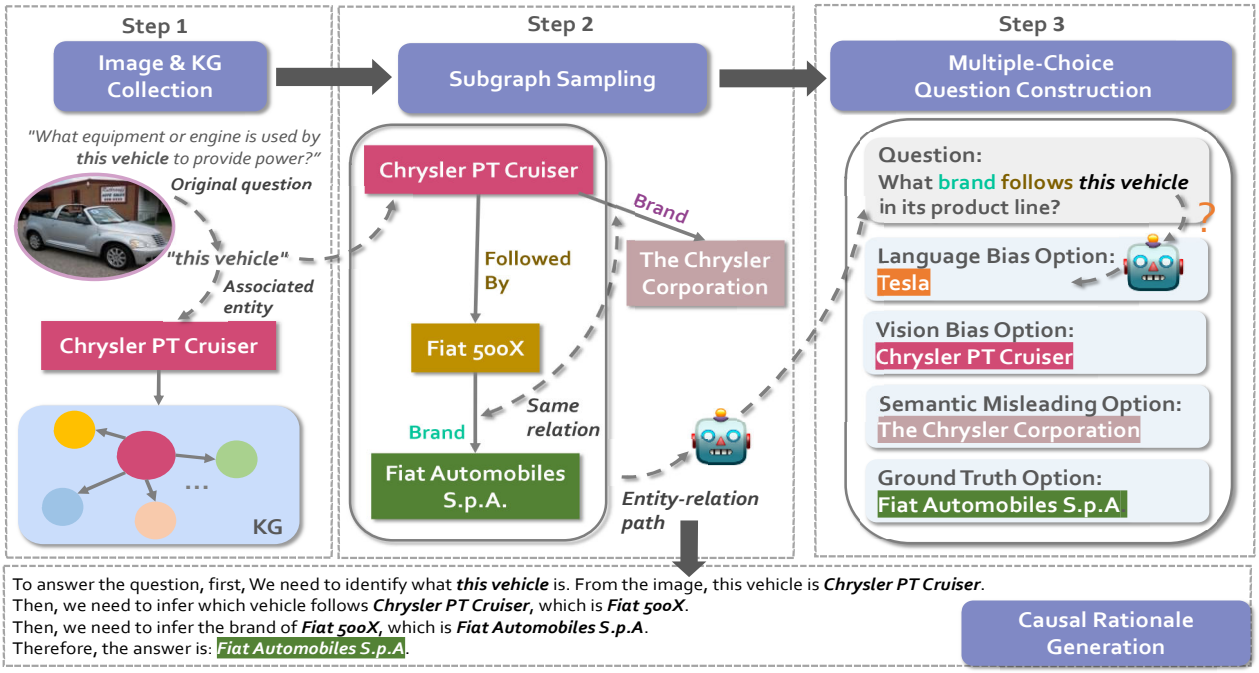

图 3　MORE 数据集的构建流程。

$TCE(Q\ on\ A) := E_{q' \sim P(Q)}[\delta(P, P')]$，其中 $P' = \mathbb{P}(A \mid I, do(Q = q'))$。

控制核心语义意义 S 将允许我们获得文本表面形式 T 对 A 的 DCE，即：

$DCE(T \rightarrow A) := E_{q' \sim P(Q \mid S)}[\delta(P, P')]$，其中 $P' = \mathbb{P}(A \mid I, do(Q = q'))$。

请注意，由于T和A之间没有中介，所以$DCE(T -> A)$也是T对A的TCE。因为枚举T的所有可能扰动通常不可行，我们可以通过干预 Q 而不影响 S，在 T 的某个子集上获得实际结果[16]。此外，在 VQA 问题的上下文中，我们不能只干预 S 而不影响文本表面T。然而，通过比较我们已经知道的两个量，即$TCE(Q\ on\ A)$和$DCE(T -> A)$，可以帮助我们理解S对A的因果影响。

总的来说，计算TCE帮助我们评估模型的敏感性（对正确答案变化的反应），而 DCE 评估其鲁棒性（对固定正确答案时假相关性预测的稳定性）。

三、构建MORE数据集

本章构建了一个新颖 MORE 数据集，要求多模态大模型超越单模态偏见，并从文本和图像中彻底整合信息以选择正确答案。数据生成过程如图 3 所示。

1. 图像和知识图谱收集

我们从一个现有的视觉问答（VQA）数据集 INFOSEEK[15]开始，该数据集将图像中描绘的实体与从 Wikipedia 来源的信息链接起来，要求 VQA 模型回答有关关联实体的问题。基于图像和对应的实体信息，我们在知识图谱（KG）- Wikidata5M[3]中识别所有与关联实体相关的n阶邻居（$n \in \{1,2\}$）。

2. 子图采样

受到因果分析的启发，我们旨在构造需要克服单模态偏见才能正确回答的多跳查询。为此，我们首先识别实体及其在 KG 中的n阶邻居的子图。然后，过滤满足两个标准的路径：1）路径的唯一性：从关联实体到选定邻居的路径是唯一的；2）共享类型关系：它们共享指向唯一实体的相同类型关系，这两个指向的实体不相同。

3. 多项选择题构造

在此小节，我们详细说明构造四个候选选项的多项选择题的过程。

问题生成 获取满足标准的子图后，我们使用子图中的实体-关系路径生成问题。为了获得流畅且连贯的问题，我们将路径输入到一个大语言模型中产生目标问题文本。我们采用了上下文学习（ICL）[3]技术，并为大语言

Model	MORE (Two-hop, acc (%))		MORE (Three-hop, acc (%))		MORE (Overall, acc (%))	
	Open-ended	Multi-choice	Open-ended	Multi-choice	Open-ended	Multi-choice
Random	/	25.0	/	25.0	/	25.0
BLIP2	4.0	16.4	1.4	15.4	2.7	15.9
InstructBlip	3.0	17.0	1.6	16.2	2.3	16.6
mPLUG-Owl	4.0	12.4	8.2	11.4	6.1	11.9
LLaVA	8.0	20.8	6.8	13.6	7.4	17.5
GPT-4V	15.8	25.6	15.3	23.2	15.6	24.4
Gemini Pro	14.2	33.5	10.1	24.4	12.2	28.9

表 2　多模态大模型在 MORE 上的结果对比。

模型提供了几个例子。在比较了不同的大语言模型并调整指令后，我们发现 ChatGPT 生成的多跳问题质量最高，因此选择其结果进行后续的评估。最后，为了防止信息泄露，问题中的实体名称被替换为 “this <OBJECT_NAME>”。

语言偏见选项 如前所述，语言偏见指的是模型过度关注问题文本中的信息。为了模拟这种情况，我们在纯文本设置下使用生成的问题测试多模态大模型。为确保所有多模态大模型的最终选项相同，我们统一使用 GPT-4V 生成得到的答案。

视觉偏见选项 为了探索沿着$I \rightarrow A$路径的视觉偏见，我们将与视觉相关的实体名称（例如，“Chrysler PT Cruiser”）作为一个选项。这允许我们观察模型在遇到与视觉信息对齐的选项时是否直接选择它。

语义误导选项 此外，我们引入了一个语义误导选项，例如“The Chrysler Corporation”，挑战多模态大模型在 KG 中的多跳推理。这个选项指的是被两个关联实体和它们的采样邻居共同拥有的关系所指向的实体。

正确答案选项 与通过$I \rightarrow G$和$S \rightarrow G$的因果路径相对应，这个选项是实体-关系路径的最终实体（例如，“Fiat Automobiles S.p.A.”）。最后，我们检查并确保每个选项与其余三个选项不同，以消除重叠样本。

因果推理路径生成 此外，实体-关系路径可以帮助生成针对当前问题的推理过程，被称为因果推理路径。在这一背景下，我们采用了一种启发式规则基础方法，从关联实体开始，逐步生成因果理由，直至达到正确答案。这些生成的因果理由可以用来验证多模态大模型的推理过程是否正确，从而提供可解释性。它们也可以用于微调专门的多模态大模型，以增强其多跳推理能力。因果理由还可以通过如 ChatGPT 这样的大型语言模型（LLMs）进一步打磨和精炼，我们将这一点留给未来的工作。

四、在MORE上评估多模态大模型

1. 实验设置

数据集 我们使用MORE的所有测试数据进行评估。我们采用了两种不同的设置：

- 开放式。要求MLLM基于输入的图像和问题生成答案。
- 多选。为MLLM提供四个选项，让它从中选择正确答案。后一种设置有一个随机基线（准确率为25%）。

基准 我们以零样本（zero-shot）的方式在我们的MORE数据集上评估各种领先的多模态大模型，包括两个有限访问的多模态大模型：GPT-4V和Gemini Pro，以及四个开源的多模态大模型：BLIP-2 (6.7B)，InstructBLIP (13B)，mPLUG-Owl (7B)，和LLaVA (v1.5, 13B)。就评估指标而言，我们采用VQA准确率对所有模型进行公平比较。

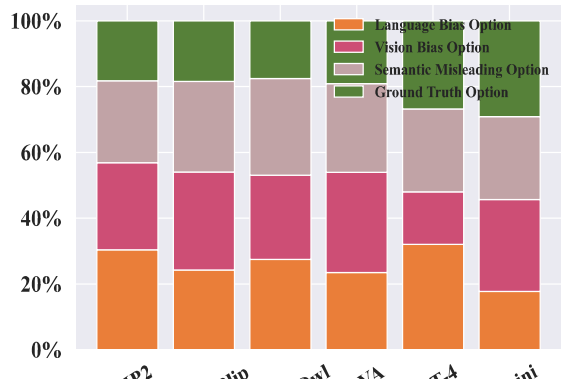

图 4　多模态模型的选项分布

2. 评估结果

我们在表2中分别展示了多模态大模型在MORE数据集的两跳、三跳和所有数据上的结果。我们观察到：1）所有基线在MORE上的性能都较差（例如，在``多选''设置下，只有Gemini Pro超过了随机基线，准确率为28.9%），这表明MLLM对语言和视觉偏见的脆弱性。2）在MORE上，开源模型与有限访问模型之间仍存在差距，尤其是在“开放式”设置下。3）大多数模型在两跳数据上的表现优于三跳数据（Gemini Pro在两跳数据上表现尤为出色，准确率达到33.5%），这表明当问题变得更加复杂时，多模态大模型的推理能力受到挑战。4）GPT-4V在开放式''设置下表现最佳，但在多选''设置下相比Gemini Pro略显不足，可能是因为在构造语言偏见选项时，我们使用了同源的ChatGPT生成的干扰项，这对GPT-4V的判断构成了更大的挑战。这一点也在后续的分析中得到验证。

3. 对 VQA 偏见的因果分析

在本小节中，我们通过因果视角分析多模态大模型的性能。

选项分布 图4显示了在“多选”设置下各种MLLM的选项分布。我们观察到：（1）BLIP2和GPT-4V经常错误选择表明语言偏见的选项，这与我们对GPT-4V的先前分析一致。（2）在所有模型中，语言或视觉偏见的比例超过了40%，显示了单模态偏见对它们预测的显著影响。（3）在一定程度上，模型选择语义上误导的选项表明了一些结合视觉和文本信息的能力，尽管并未完全掌握问题。这突出了我们的MORE数据集对当前MLLM所提出的挑战。请注意，这里呈现的正确选项的比例与表2中报告的准确率值之间可能存在差异，因为某些模型（例如，mPLUG-Owl）的输出可能不符合提供的选项，这影响了有效答案的计数。

图像和问题的因果效应 为了进一步分析视觉偏见和语言偏见对模型预测的影响，我们根据第二章提供的定义评估了因果效应。具体而言，我们随机选择100个样本进行干预，然后测量所有实例的效果平均值，以计算TCE（对应于模型的敏感性）和DCE（对应于模型的鲁棒性）。总的来说，较高的TCE是可取的，表示更好的敏感性，而较低的DCE表示更好的鲁棒性。

从图5可以看出：1）当前的多模态大模型展现出高敏感性（高TCE），一个可能的原因是指令调整使得模型对输入的变化更为敏感。2）然而，鲁棒性相对较低（高DCE），显示出即使在固定的答案值下，预测也会随着输

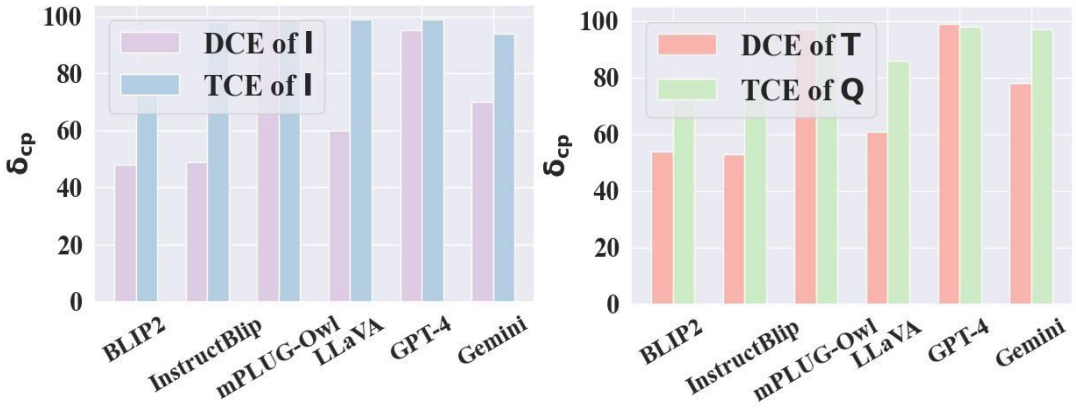

图 5　比较图像和问题对多模态大模型预测的因果效应

入的变化而改变，这表明依赖于伪造路径而非真实的因果特征。

五、总结

本文提出了一种全面的方法来量化和评估多模态大模型中的单模态偏见。通过我们的因果推理框架，我们深入分析了这些偏见对 VQA 问题中模型预测的因果效应。我们引入的 MORE 数据集要求多模态大模型进行多跳推理，并克服语言和视觉偏见，从而扩展了它们的推理能力边界。一系列定量与定性的分析实验为未来的工作提供了见解。

责任编委 **魏秀参**

参考文献

[1] Chen, M., Cao, Y., Zhang, Y., Lu, C., Quantifying and Mitigating Unimodal Biases in Multimodal Large Language Models: A Causal Perspective. *arXiv:2403.18346.*

[2] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *NeurIPS*, *35*, 27730-27744.

[3] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv:2302.13971*.

[4] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv:2303.08774*.

[5] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... & Ahn, J. (2023). Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*.

[6] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2023). A survey on multimodal large language models. *arXiv:2306.13549*.

[7] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. *NeurIPS*, *36*.

[8] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669-688.

[9] Zhu, Y., Groth, O., Bernstein, M., & Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. *CVPR,* pp. 4995-5004.

[10] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *CVPR,* pp. 6904-6913.

[11] Wang, P., Wu, Q., Shen, C., Dick, A., & Van Den Hengel, A. (2017). Fvqa: Fact-based visual question answering. *IEEE T-PAMI*, *40*(10), 2413-2427.

[12] Marino, K., Rastegari, M., Farhadi, A., & Mottaghi, R. (2019). Ok-vqa: A visual question answering benchmark requiring external knowledge. *CVPR,* pp. 3195-3204.

[13] Jain, A., Kothyari, M., Kumar, V., Jyothi, P., Ramakrishnan, G., & Chakrabarti, S. (2021, July). Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. *ACM SIGIR,* pp. 2491-2498.

[14] Schwenk, D., Khandelwal, A., Clark, C., Marino, K., & Mottaghi, R. (2022, October). A-okvqa: A benchmark for visual question answering using world knowledge. *ECCV,* pp. 146-162.

[15] Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., & Chang, M. W. (2023). Can pre-trained vision and language models answer visual information-seeking questions? *arXiv:2302.11713*.

[16] Stolfo, A., Jin, Z., Shridhar, K., Schölkopf, B., & Sachan, M. (2022). A causal framework to quantify the robustness of mathematical reasoning with language models. *arXiv:2210.12023*.

[17] Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., ... & Wen, J. R. (2022). Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, *13*(1), 3094.

[18] Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., & Tang, J. (2021). KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, *9*, 176-194.

陈美琪

北京大学智能学院 2020 级博士研究生，导师为张岩教授，主要研究方向为自然语言处理和大模型。

Email:meiqichen@stu.pku.edu.cn

曹艺馨

新加坡管理大学助理教授，清华大学计算机科学博士，曾担任南洋理工大学的研究助理教授和新加坡国立大学 NExT++研究中心的研究员。研究领域涵盖自然语言处理、知识图谱和推荐系统，各项工作已发表在包括 ACL、EMNLP、COLING 和 WWW 在内的顶级会议上，共获得超过 4,000 次引用。

Email: caoyixin2011@gmail.com

张岩

张岩，北京大学智能学院教授，博士生导师。研究兴趣是信息检索、自然语言处理、数据挖掘、网络科学和大数据分析，近年来在这些领域的国际期刊和会议上发表论文 100 多篇，作为项目负责人和技术骨干承担和参与国家自然科学基金项目、核高基重大专项、科技支撑计划重点项目、973 项目、北京市基金、教育部项目、粤港合作项目等二十余项。

Email:zhyzhy001@pku.edu.cn

陆超超

上海人工智能实验室青年科学家，博士生导师，因果智能团队负责人，主要从事因果推理方面的理论研究及其在机器学习等相关领域的应用。

Email: luchaochao@pjlab.org.cn