arXiv:2602.01561v1 [cs.CV] 2 Feb 2026

多模态反常识：从奇特到寻常，再从寻常到奇特

Yejin Son[1*], **Saejin Kim**[1*], **Dongjun Min**[1], **YoungjaeYu**[2†] 首尔大学 [2] 首尔大学**通信作者：** yejinhand@yonsei.ac.kr, jerry0110@yonsei.ac.kr, youngjaeyu@snu.ac.kr

摘要

多模态常识推理仍是人工智能领域的一项基础性挑战。我们提出了**多模态反常识（MUN）**基准测试，旨在评估模型处理偏离典型视觉或文本预期的场景的能力。MUN 将视觉场景与用自然语言描述的令人惊讶或不太可能的结果配对，促使模型运用日常逻辑解释看似奇怪的图像，或发掘普通场景中的意外解读。为支持这项任务，我们提出了一种基于检索的上下文学习（R-ICL）框架，无需额外训练即可将大型模型的推理能力迁移至小型模型。借助新颖的**多模态集成检索器（MER）**，即使图像与文本对被有意设置为不一致，我们的方法仍能识别语义相关的示例。实验表明，与基线ICL方法相比，该方法平均提升了 8.3%，凸显了 R-ICL 在低频、非典型场景中的有效性。MUN 为评估和提升视觉语言模型在真实世界、文化多样化及非原型场景中的鲁棒性与适应性开辟了新方向。

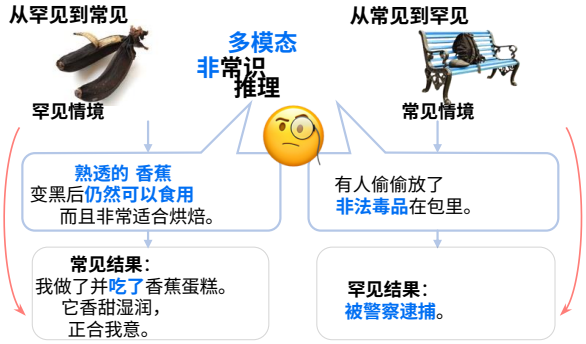


图 1：多模态反常识推理旨在生成解释，使给定结果显得合理。例如，过熟香蕉（罕见背景）仍可用于烘焙香甜湿润的香蕉蛋糕（常见结果）；而长椅上的包（常见背景）却会导致有人被捕（罕见结果）。这凸显了将视觉线索与逻辑推理相结合的挑战，我们的多模态反常识（MUN）数据集正是针对这一挑战。

1 引言

在日常生活中，常识如同宇宙中的“暗物质”，构成一个无形的框架。尽管我们无法直接感知它，常识却会潜移默化地影响我们的决策，例如辨识社会规范或理解模糊不清的情境（Bosselut 等人，2019年；Tao 等人，2024）。

尽管这种常识往往能让模型在熟悉的背景下进行稳定推理，但当遇到超出典型经验或文化熟悉度范围的视觉或文本线索时（Wang 等人，2023）。然而，现有的大多数常识推理基准测试，都使用大型英语语料库和标准视觉语言数据集中覆盖充分的常见或典型案例来评估模型（Brown 等人，2020；Raffel 等人，2020；Hendricks 等人，2016年；Agrawal 等人，2017；Li 等人，2019年）。因此，面对训练分布之外的罕见、含糊或具有文化特异性的现象时，当前人工智能系统表现出明显的脆弱性。

为弥补这一空白，我们推出**多模态 Uncommonsense（MUN）**基准测试：这是一个由人类整理的数据集，专门用于评估模型在视觉语境下推理**罕见或反直觉结果**的能力。不同于以典型常识为重点的既有数据集，MUN 聚焦于违背通常预期的视觉情境，例如过熟的香蕉更适合烘焙，或具有特定文化含义的手势，例如印度式摇头动作表示一致性。

*贡献相同

视觉语境	文本上下文	结果	解释
	香蕉上的黑斑。	人物喜欢吃这根香蕉 没有任何健康的面包 问题。	这些过熟香蕉很可能 被用于烘焙香蕉面包，而该面包 在食用时没有任何健康问题。
	新鲜的气泡 浇筑的混凝土。	人物观察混凝土 凝固良好，提供了 坚实的基础，供 施工使用。	新浇筑混凝土中的气泡 混凝土，这是一种常见现象 在搅拌过程中，并未影响 其完整性，混凝土 充分凝固，形成坚实的地基。
	一个蓝色背包 放在.....上 一张木制长椅。	被警察逮捕。	趁主人去洗手间时， 有人放入了非法毒品 包里 而他们对此毫不知情。
	一台笔记本电脑和咖啡 咖啡馆桌子上的咖啡杯。	因欺诈而损失了钱财。	有人试图通过 一个未设防的 Wi-Fi连接， 窃取个人信息并 从事欺诈行为。

图2：MUN 示例。前两个示例来自 MUN-vis，后两个示例来自 MUN-lang；解释由人工标注员撰写。请注意，文本背景仅用于数据集生成。

这些案例要求模型将视觉异常与基于逻辑或文化的解释相协调。

MUN 的必要性在于它聚焦于**低频、多模态推理**，这是现实应用中的关键方面，因为视觉输入与常识预期往往相互冲突。通过构建并评估此类示例，MUN 成为对现有数据集的补充性基准测试，将常识评估的范围拓展至传统界限之外。

我们为每个案例收集了人工撰写和大语言模型生成的解释，揭示了两者在可解释性和多样性方面存在显著差异。正如先前研究所指出的，尽管大语言模型解释通常较为准确，人工标注却提供了更广泛的视角（Zhao 等人，2023）。我们通过数据增强利用两者，构建了一个高质量基准测试，以支持丰富的监督和评估。

为了提升模型在视觉和背景层面均不典型的场景中的推理能力，我们采用基于检索的上下文学习（R-ICL）方法（Lin 等人，2022年，2023），通过利用大型模型生成的语义相关示例来提升较小模型的表现。具体而言，为了在存在视觉与文本不一致的场景中有效识别这些示例，我们提出了一种创新的检索框架，称为多模态集成检索器 （MER）。MER 对每种模态的相似度进行独立评分，并通过可调加权机制将其融合，从而在有意设置不一致的 MUN 图文对时实现灵活检索。

与假设跨模态对齐程度较高的传统检索器不同，MER 能够适应我们基准测试中未对齐的特性。据我们所知，这是首次在视觉信号与文本信号被刻意设置为不一致的情境中应用 R-ICL，从而能够对未对齐的多模态输入进行溯因推理。该方法的胜率平均比随机基线高出 8.3%，证明了 R-ICL 在提升细致入微的多模态推理方面的有效性。

通过在视觉上罕见的场景中，将直观印象与潜在真相联系起来，MUN 为构建能够跨越不同文化、背景和预期，进行超越显而易见之事推理的可信赖的人工智能系统奠定了基础。

2 相关工作

溯因推理。溯因推理是常识的核心，涉及从不完整的观察中推断出最合理的解释。尽管已有多种研究探索文本和多模态溯因推理，但现有方法各有局限（表1）。例如，Abductive-NLI（Bhagavatula 等人，2020）仅关注日常场景中的文本输入，缺乏视觉锚定。Sherlock（Hessel 等人，2022年）融合了真实图像和文本，但仍局限于常见情境，且推理方向单一。Uncommonsense（Zhao 等人，2023）面向罕见情境，但缺少视觉信号；而 NL-Eye（Ventura 等人，2024）采

数据集	模态	Real 图像	罕见	双向推理?
Abductive-NLI	T	✗	✗	✗
Sherlock	I+T	✓	✗	✗
Uncommonsense	T	✗	✓	✗
NL-Eye	I+T	✗	✗	✗
MUN（我们的方法）	I+T	✓	✓	✓

表 1：与溯因推理基准的比较。（Bhagavatula 等人，2020；Hessel 等人，2022年；Zhao 等人，2023；Ventura 等人，2024）“I”代表图像，“T”代表文本。MUN 独有地支持“双向非常识推理”，结合罕见背景、结果以及细致的视觉场景。

用合成图像，却未能充分处理反常识场景或双向推理。因此，现有方法都无法同时融合真实图像、处理罕见情境并支持双向溯因推断。相比之下，我们提出的 MUN（多模态反常识）数据集融合了真实图像和文本，积极纳入罕见场景，并支持双向推理，从而弥补了这些不足，并提供了更全面、更细致的溯因推理基准测试。

检索增强与上下文学习大语言模型（LLM）（Brown 等，2020；Chowdhery 等，2022；Achiam 等，2023）和视觉语言模型（VLMs）（Alayrac 等，2022；Li 等，2023a）的最新进展展现出了卓越的零样本和少样本学习能力。然而，它们的推理往往仍然受限于训练数据中根深蒂固的模式。检索增强范式（Thoppilan 等，2022）和上下文学习（ICL）技术（Wei 等，2022；Zhou 等，2022）代表了将模型能力扩展到记忆知识之外的有前景策略。通过动态融入外部文档、示例或背景线索，模型能够处理更复杂的推理任务，并适应新的领域。在视觉领域，多源检索（Zhu 等，2020；Shao 等，2024）和基于检索的图像定位展现出了潜力。我们的工作顺应这一趋势，采用基于检索的上下文学习方法。我们从 MUN 场景中检索文本和视觉示例，引导模型进行推理，并将复杂的溯因推理和文化逻辑提炼为易于理解的形式。这种方法帮助较小的视觉语言模型应对罕见场景，并生成连贯且具有丰富背景信息的解释。

3 多模态反常识（MUN）

为了推进视觉非常识推理领域的研究，我们构建了基准测试**多模态反常识（MUN）**，旨在通过偏离标准视觉或背景预期的场景来挑战模型。受此前关于Zhao 等人，2023）反常识推理工作的启发，尤其是针对异常情境的溯因推理研究，我们的数据集采用了结构化的背景-结果-解释范式。在这一框架中，模型需要解读基于图像的背景以及文本场景（结果），然后生成能够调和二者的解释。

3.1 任务设置

我们聚焦于两种互补的任务设置，强调视觉线索与文本推理之间微妙的相互作用。

MUN-vis：罕见图像（背景）→ 常见结果在这项任务中，模型会看到一幅图像，它乍看之下在视觉上显得奇特或“罕见”，呈现的是发生频率或概率较低的情境。尽管表面上看起来怪异，任务目标是生成连贯的解释，使这一情境显得合乎常理，并说明它实际上很常见或完全合理。例如，在图2的第一行中，一张发黑的香蕉照片乍看之下可能显得不同寻常。然而，结果表明：“人物毫无健康顾虑地享用了香蕉面包”，这意味着需要通过诸如“香蕉熟透了，因此被用来烘焙香蕉面包，食用后也没有任何健康问题”这样的解释，衔接背景与结果。这项任务需要生成解释，将看似奇特的视觉输入与熟悉且合乎逻辑的日常情境联系起来。

MUN-lang：常见图像（背景）→ 不常见结果在这一情境中，模型会看到一幅看似完全普通的图像，但必须解释与之相关的不寻常或“不常见”的文本结果。在图2中，第三行所描绘的背景展示了一个看似典型的场景：一个蓝色背包放在木制长椅上，而结果却是“被警察逮捕”，这与给定背景并不自然相符。解释必须通过揭示不太明显的细节来弥合这一差距，例如

“当包的主人在洗手间时，有人在其不知情的情况下偷偷将非法毒品放入包中”，从而提供一个出人意料但合理的理由，使这一不协调的情境变得合乎情理。

3.2 数据集构建

我们通过多步流程构建了 MUN 数据集，生成了多样的“罕见常识”场景，以挑战多模态推理模型。

场景生成 我们使用 GPT-4o 生成了丰富多样的文本场景（背景）。对于 MUN-vis，我们指示模型描绘乍看之下视觉上显得古怪、但最终却很正常的场景。对于 MUN-lang，我们要求模型生成外表平常、却隐藏着令人惊讶理由的场景。我们的提示策略鼓励模型分析假设性的图文对，将其归类为“正常”或“异常”，并提供简短解释。通过改变视觉和背景线索、突出潜在原因，我们获得了充满文化背景、感官细节和概念反转的场景。这种方法引导 GPT-4o 生成结构清晰、逻辑扎实的解释。在 MUN-vis 条目中，通过揭示合理的背景故事，原本看似奇怪的图像变得合乎情理。在 MUN-lang 条目中，则通过隐藏的惊喜或非传统做法，重新诠释平淡无奇的表象。

通过筛选确保多样性 为确保数据集具有多样性，我们在生成大量候选场景后实施了全面的筛选流程。观察到许多相似场景后，我们优先将其移除，以促进多样性并减少冗余。我们使用 Dedupe 库，[1] 这是一款专门用于数据去重的工具，能够有效消除重复项。

受多样性筛选方法的启发（Han 等人，2023），我们进一步通过确定一系列特定关键词来丰富背景的多样性。如果图像的语言描述中包含任何此类关键词，就会将相应示例筛除。为保持平衡，我们确保这些关键词在背景中的出现次数低于 20。

图像配对与选择流程 对于每个文本场景（背景），我们首先使用必应网页搜索 API[2],检索五张候选图像然后对这些图像进行人工审核，选择最能体现该场景独特性或寻常性的图像。如果通过自动搜索未找到合适的图像，我们会进一步进行人工搜索，以确定合适的选项。

通过融入真实世界的图像，模型能够获得更稳定且更具泛化性的推理能力，ALBEF（Li 等人，2021年）、BLIP（Li 等人，2022年）和 LLaVA（Liu 等人，2023a）的研究以及 LAION-5B（Schuh- mann 等人，2022年）等大规模、多样化图像资源都证明了这一点。构建以真实视觉内容为基础的数据集，有助于拓展覆盖范围，纳入罕见情境和文化细微差异。通过迭代优化，这种方法克服了现有局限，并支持更细致的跨领域推理。

人工解释生成我们招募了 26 名专攻计算机科学和人工智能的研究生担任标注员，参与主要的解释撰写任务。所有参与者都精通英语，界面和指令也均以英语提供。为确保工作流程公平高效，任务被拆分为多个小批次，并在标注员之间平均分配工作量。这种做法避免了个别标注员负担过重，从而保证了数据集的一致性和质量。此外，为增强数据集在背景方面的可靠性，我们要求标注员不要为他们认为无关或不恰当的情境撰写解释。这一措施避免纳入不必要或非必要的解释。此外，我们鼓励标注员根据提供的视觉语境，进行合乎逻辑的推断，并阐述那些看似不匹配的结果背后的原因。

大语言模型增强的人工撰写解释正如后续分析所示，这也与以往研究（Zhao 等人，2023）的结论一致：人工撰写的解释展现出多样性和广泛的理解，而大语言模型生成的回答往往相对狭窄且具体。我们的目标是结合这些互补优势，进一步完善人工标注。具体而言，我们使用精心设计的提示词引导 GPT-4o 改进人工撰写的解释，使其能够更清晰、具体地呈现视觉场景与给定罕见结果之间的逻辑联系。这个过程在保留人工解释多样性和细微差别的同时，

[1] https://github.com/dedupeio/dedupe。[2]必应网页搜索 API：https://www.microsoft.com/en-us/bing/apis/bing-web-search-api。

LV	人类(%)	大语言模型(%)	人类+大语言模型(%)
1	30.5	0.3	1.3
2	40.1	8.9	9.3
3	8.6	35.4	20.5
4	11.9	55.0	62.6
5	8.9	0.3	6.3
Avg.	**2.29**	**3.46**	**3.63**

表 2：人类撰写的解释（人类）、大语言模型生成的解释（大语言模型），以及经大语言模型增强的人类撰写解释（大语言模型+人类）的具体性比较。表中每个数值表示被评为各具体性层级（1 至 5）的解释所占百分比。

利用大语言模型的精确性，从而得到一组经过改进的解释，为比较提供信息更丰富的基线。

3.3 数据分析

MUN 数据集包含两个子任务：MUN-vis 包含 515 个视觉上罕见的背景和常见结果实例；MUN-lang 包含 500 个视觉上常见的背景和罕见结果实例，共计 1,015 组视觉语境—结果配对。我们从 MUN-vis 中的 143 个实例和 MUN-lang 中的 156 个实例收集了人工解释，并为所有配对生成了大语言模型生成的解释。

MUN的多样性 MUN 数据集涵盖各种类别中广泛的场景，每个示例都包含详细的文本解释，将视觉语境与结果联系起来。尽管某些类别可能得到重点强调，但单个示例仍然呈现出复杂的多层次场景。有关主题及其组合频率的详细报告，请参见附录 I。t-SNE（Maaten和 Hinton，2008）可视化结果（图 3）显示，文本情境聚集成不同的群组，涵盖广泛的主题。

3.4 解释的比较分析

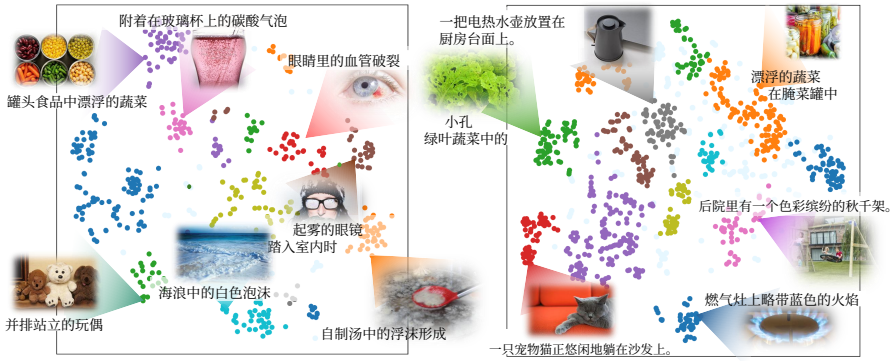
与 Uncommonsense（Zhao 等人，2023）的研究结果一致，大语言模型、人类以及人类+大语言模型生成的解释，在长度和词汇多样性方面都存在明显差异。图4展示了解释长度的分布。在**MUN-vis**任务中，人类的解释相对较长且差异较大，平均为32.0 ± 38.5 个词元。相比之下，大语言模型的解释较短且更稳定，为 25.1 ± 4.4个词元，而人类+大语言模型的解释较长，为 50.7 ± 37.2 个词元，内容也更为详细。在**MUN-lang** 任务中，人类给出的解释较短（16.3 ± 7.9 个词元），而大语言模型的输出更长且更加一致（44.5 ± 6.4 个词元）。这一模式表明，在 MUN-lang 这类更开放式的任务中，大语言模型会生成内容更丰富、篇幅更长的解释；而在 MUN-vis 这类结构更明确的任务中，人类往往会提供更长的描述。人类+大语言模型的解释达到 44.6 ± 13.1 个词元，长度接近大语言模型的水平，同时兼具人类的创造力与大语言模型的稳定性。

为了量化词汇多样性，我们测量了 n-元熵（$n \in \{1, \ldots, 5\}$），如图 5所示，并进行了1,000次自助法迭代[3]。在**MUN-vis**中，人工解释显示出比大语言模型解释更高的n-元熵，而人类+大语言模型的熵超过了人类的熵，反映出一种协同效应：人类变异性与大语言模型的精确性得以结合。大语言模型解释显示出较低的熵，这可能是由于任务具有结构化特征。在**MUN-lang** 中，大语言模型的熵与人类 + 大语言模型的熵相近，甚至更高，并且显著超过人类的熵，这表明在开放式任务中，大语言模型采用了更加多样的措辞，而人类语言使用则更加受限。人类+大语言模型仍保持较高的熵，有效融合了人类创造力与大语言模型严谨性。

近期研究（Stiennon 等人，2020；Liu 等人，2023b）指出，只要提示词结构良好，大语言模型就能可靠地评估文本的定性特征，例如具体性。遵循这一方法，我们使用 GPT-4o 对具体性进行评估（评分范围为 1 至 5）。

表2 显示，人工解释中低具体性（评分为 1 至 2）的比例较高（70.6%），而高具体性（评分为 4 至 5）的比例相对较低（20.8%）。大语言模型的解释通常保持中等至较高的具体性（评分为 3 至 4），其中 4 分的比例较高（55.0%），但达到最高具体性（5 分）的极少（0.3%）。相比之下，人类+与大语言模型结合的方法中，4 分的比例更高（62.6%），5 分的比例也有所提升（6.3%），因此整体具体性最高。这表明大语言模型能够对人类输入进行完善和扩展，达到更高的细节层次和具体性；人类+与大语言模型结合的方法可以兼顾双方优势，同时弥补各自的不足。根据这些结果，人类+与大语言模型结合的方法

[3]每轮迭代中，我们从每个子集中针对每个上下文-结果对随机选取一条解释。



(a) MUN-vis（正常文本，异常图像） (b) MUN-lang（异常文本，正常图像）

图3：基于 SimCSE 的 MUN-vis (a) 和 MUN-lang (b) 在不同类别上的 t-SNE 可视化（Gao 等人，2021年）。

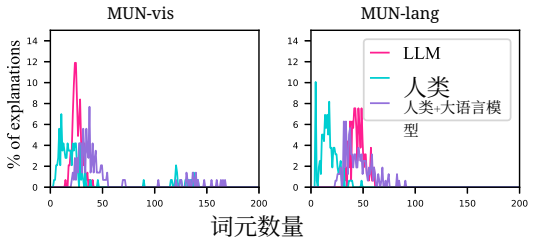


图4：MUN 中解释的词元长度分布：左侧表示 MUN-vis，右侧表示 MUN-lang，数据由各数据子集开发集的计算结果得出。

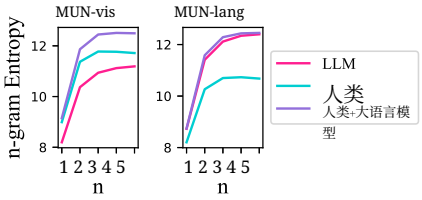


图5：MUN-vis（左）和 MUN-lang（右）的 n 元语法分布熵根据各数据子集的开发集计算得出。

被认为是最佳方案，因此被选为评估基线。[4]

4 基于检索式上下文学习的视觉常识推理

MUN 数据集的核心目标是通过非典型、低频的视觉文本场景来挑战模型，这些场景难以用传统常识来解释。与标准基准测试不同，MUN 专注于“Uncommonsense”推理，在这种推理中，模型必须推断出细致入微且往往具有溯因性质的理由，以解释不同寻常的结果。这超越了大型视觉语言模型（VLM）通常凭借在统计上占主导地位的模式上进行的大量预训练，在这方面表现出色。

然而，面对这类非典型场景时，模型往往会退回到预训练期间学到的高概率模式，生成字面化或过于笼统的图像描述，而非溯因解释（Hessel 等人，2022年）。为了在这些陌生场景中实现更贴合背景的推理，我们采用基于检索的上下文学习（ICL），引入相关但语义上并不显而易见的示例。

传统检索方法假设不同模态之间具有很强的对齐关系，因此往往难以应对 MUN 中图文对有意呈现的差异。为此，我们提出了一种多模态集成检索器（MER），分别计算图像和文本的相似度，并通过可调节的融合机制将二者结合起来。即使视觉和文本线索指向不同甚至相互冲突的常识预期，这种方法也能帮助 MER 检索到语义连贯的示例。

具体而言，MER 使用 CLIP 风格的图像编码器和基于 BERT 的文本编码器对（图像、文本）对进行嵌入，并分别计算查询与数据集条目在各模态上的余弦相似度。随后，使用加权系数 α 对这两个相似度分数进行整合，以平衡各模态的贡献。这种分开但对齐的检索策略使 MER 能够灵活适应特定模态的信号，为弥合视觉基础溯因推理中的概念鸿沟提供了一种原则性机制。

据我们所知，这是首次在视觉与文本模态被刻意设置为不一致的场景中应用此类双重评分检索框架。完整的公式

[4]关于人类如何感知人类撰写的回答与大语言模型生成或增强的回答之间差异的进一步分析，见附录D。

数据集	模型	零样本	单样本		三样本		五样本	
			随机	R-ICL	随机	R-ICL	随机	R-ICL
MUN vis	*Gemma3*	0.335	0.428	0.457	0.422	**0.491**	0.393	0.382
	InternVL 2.5	0.243	0.121	0.254	0.208	0.312	0.387	**0.434**
	LLaVA OV	0.301	0.439	0.405	0.434	0.376	**0.474**	0.445
	Phi 3.5v	0.283	0.324	0.335	0.387	0.387	0.428	**0.445**
	Phi 4mm	0.410	0.312	0.272	0.393	0.387	**0.630**	0.618
	Qwen2.5 VL	0.364	0.428	0.387	0.422	0.428	0.439	**0.538**
	Qwen2 VL	0.225	0.399	0.405	0.283	0.376	0.272	**0.486**
MUN lang	*Gemma3*	0.257	0.341	0.418	0.430	0.454	0.498	**0.546**
	InternVL 2.5	0.325	0.273	0.293	0.361	0.357	0.430	**0.470**
	LLaVA OV	0.285	0.301	0.325	0.333	0.369	0.365	**0.369**
	Phi 3.5v	0.337	0.353	0.329	0.410	0.390	0.430	**0.442**
	Phi 4mm	0.357	0.502	0.582	0.534	0.554	0.651	**0.655**
	Qwen2.5 VL	0.422	0.353	0.329	0.357	0.357	0.410	**0.426**
	Qwen2 VL	0.349	0.349	0.321	0.365	0.341	**0.365**	0.357

表3：在不同的示例数量设置下，模型相对于人工辅助解释的胜率比较（越高越好）。“随机”表示随机选择的示例，“R-ICL”表示检索得到的上下文学习示例。模型输出与人类+大语言模型解释进行比较，并由大语言模型进行评判。

和算法实现见附录 B。

5 实验

我们使用 MUN 数据集，评估所提出的检索式上下文学习（ICL）方法在多模态反常识推理中的有效性。实验研究围绕两个核心问题展开：

RQ1. 上下文示例（示例数）的数量如何影响模型性能？**RQ2.** 与随机选择的示例相比，基于检索的示例有何影响？

为建立可靠的基线并确保全面评估，我们对多种最先进的视觉语言模型（VLMs）进行基准测试，并在检索式上下文学习方法中采用多模态集成检索器。

5.1 模型选择与检索机制

我们评估了七种交错式视觉语言模型，涵盖不同的架构系列和规模：Qwen2-VL（Wang 等人，2024）、Qwen2.5-VL（Bai 等人，2025）、Phi-3.5-vision（Abdin 等人，2024）、Phi-4-multimodal（Abouelenin 等人，2025）、InternVL-2.5（Chen 等人，2024）、Gemma3（Team 等人，2025）、LLaVA-Onevision（Li 等人，2024）。为支持检索式上下文学习，我们使用一个结合文本和视觉输入的多模态集成检索器。基于 BERT 的文本编码器（Xiao 等人，2023）根据查询结果对文本示例进行编码和检索，而基于 CLIP 的图像

模型	LR	LC	LE	CS
Gemma3	3.16 (+0.12)	3.59 (+0.15)	4.12 (+0.14)	3.59 (+0.14)
InternVL 2.5	2.83 (+0.45)	3.40 (+0.52)	3.79 (+0.63)	3.33 (+0.54)
LLaVA-OV	3.21 (+0.07)	3.77 (+0.07)	4.13 (+0.04)	3.84 (+0.11)
Phi 3.5v	3.23 (+0.14)	3.75 (+0.09)	4.16 (+0.11)	3.77 (+0.15)
Phi 4mm	3.31 (+0.21)	3.85 (+0.29)	4.11 (+0.25)	3.82 (+0.28)
Qwen2.5 VL	3.29 (+0.08)	3.87 (+0.13)	4.25 (+0.07)	3.89 (+0.12)
Qwen2 vl	3.28 (+0.11)	3.86 (+0.14)	4.22 (+0.11)	3.94 (+0.15)

表 4：基于检索的上下文选择对基于烧瓶的技能指标的影响（越高越好）。LR 表示逻辑稳健性，LC 表示逻辑正确性，LE 表示逻辑效率，CS 表示常识。每个单元格显示 R-ICL 分数，括号内为相较于随机基线的增益。

编码器（Radford 等人，2021年）处理图像。集成模型将两种模态的相似度分数合并，超参数 α 设定为 0。4。在实验中，我们创建了一个数据库，其中包含来自 MUN-vis 和 MUN-lang 的 372 个和 344 个图像-场景对。这些数据集缺少人工标注解释，且未用于测试。为了比较基线，我们实现了标准上下文学习（ICL）提示词，其中的示例从 MUN 数据集中随机选取，不考虑它们与查询的相关性。

5.2 实验设置

改变上下文示例的数量。为了探究上下文示例的数量如何影响模型性能，我们改变提供给模型的检索示例数量（从 1、3 到 5）。这一设置使我们能够评估上下文学习方法的可扩展性，并确定有效推理所需的最佳示例数量。

基于检索与随机选择的示例。为了评估检索质量的重要性，我们将基于检索的上下文学习与一种基线上下文学习方法进行比较，后者使用从 MUN 数据集中随机选择的示例。

关于评估指标，我们采用 Alpaca-Eval 框架（Li 等人，2023b）通过将生成的解释与人类和大语言模型生成的解释进行比较，来评估其质量。具体而言，我们提示 GPT-4o 对不同模型生成的解释进行排名，并将其与人类+大语言模型解释进行比较。该排名评估模型生成的解释在连贯性、相关性和溯因推理方面的质量。

5.3 结果

表3 总结了所选视觉语言模型在不同上下文示例配置和检索策略下的表现。我们报告了 MUN-vis 和 MUN-lang 两项任务的结果，以评估视觉推理和语言推理的质量。

RQ1. 示例数扩展的影响。随着上下文示例数量从零增加到 1、3 和 5 个示例，所有模型的表现通常都会提升，尤其是在使用 R-ICL 时。在七个视觉语言模型中，视觉任务的中位数提升为 **+6.**1 个百分点，语言任务为 **+7.**2 个百分点。虽然 MUN-lang 也能从增加示例和 R-ICL 中受益，但 MUN-vis 的提升通常更为显著，这凸显出视觉推理能从有效筛选和增加上下文示例中获得更多收益。RQ2. 检索与随机选择的比较。比较表3中每种示例数下的两列可以发现，在 14 个模型—数据集组合中，有 12 个组合的 R-ICL 优于随机选择。视觉任务上的提升（+0.040）高于语言任务（+0.023），这证实了提供语义对齐的图像示例对视觉推理尤其有帮助。显著的例子包括 InternVL 2.5 在视觉任务上的表现（单样本时提升 +13 个百分点），以及 Qwen2.5-VL（五样本时提升 +9.9 个百分点）；语言任务上的提升虽为正，但幅度较小（例如 Phi-4mm 在三样本时提升 +3.4 个百分点）。附录 C 的“定性结果”部分对这些模式和示例进行了详细的定性分析。

5.4 模型响应分析

为了严格量化 R-ICL 对视觉语言模型逻辑推理能力的贡献，我们进行了两项互补分析：自动评估和人工分析。

自动评估 在 FLASK 提出的自动化技能评估框架中（Ye 等人，2023），评估涵盖四项互补标准：逻辑稳健性（LR）、逻辑正确性（LC）、逻辑效率（LE）和常识理解（CS）。
GPT-4o 根据 FLASK 框架中的评分标准，以 1 至 5 分进行评分。表4总结显示，基于检索的上下文选择使所有模型在每项指标上均取得稳定提升。最大绝对增幅出现在*InternVL-2.5* (+0.63 LE； +0.54 CS），而即使是最小的增幅，在四项技能上也都为正。这些发现表明，提供语义相关的示例不仅能增强演绎推理，还能提升常识

设置	MUN-vis (%)	MUN-lang (%)
零样本	16	36
随机5样本	12	44
检索式五样本	32	56

表5：**人类偏好胜率（%）**。在每种模态的50个样本中，人工标注员更倾向于模型生成的解释而非大语言模型+人类解释的样本比例。

意义推断，进一步凸显了上下文质量在上下文学习中的作用。

人工评估表5 展示了人工评估结果，该评估比较了*phi-4-mm*在不同样本设置下的推理质量。对于视觉和语言两种模态，我们随机选取了50个样本，并请人工标注员将模型在零样本、随机5样本和检索式五样本设置下生成的回答，与人工辅助生成的解释进行比较。对于每个样本，标注员选出他们认为更连贯、更有说服力的回答。报告的数值代表胜率，即模型输出优于人工辅助解释的案例比例。尽管样本量有限会带来一定波动，但总体趋势，尤其是人工评估结果，表明增加上下文示例的数量，特别是通过检索增加示例，通常能提升推理性能。

6 结论

我们推出了多模态反常识（MUN）数据集，用于评估视觉语言模型处理挑战常识推理的非典型场景的能力。大量实验表明，与随机选择的上下文学习（ICL）示例相比，检索式上下文学习示例能提升模型性能。通过将意料之外的视觉线索与逻辑解释相结合，我们成功引导模型生成更加连贯、符合背景的推理。这种方法能够构建更具适应性和可靠性的多模态人工智能系统，使其在真实环境中更善于理解罕见事件、文化细微差异和低频现象。

7 致谢

本研究得到现代汽车公司和起亚公司的支持，此外还获得了

由韩国政府（MSIT）资助的信息与通信技术规划与评估研究院（IITP）项目（编号：RS-2020-II201361，延世大学人工智能研究生院项目）以及由韩国政府（MSIT）资助的韩国国家研究基金会（NRF）项目（编号：RS-2024-00354218 和 RS-2024-00353125）的资助。我们还使用人工智能助手润色文风，并协助进行初步编码。

8 局限性

尽管 MUN 为评估多模态反常识推理提供了有价值的基准测试，但它并非没有不足。首先，尽管该数据集得益于细致的人工筛选，从而提升了单样本质量，但这是以规模为代价的，可能限制其对现实场景中更广泛变异性的表征，也可能无法涵盖文化、环境或特定领域复杂性的全部广度。

其次，我们基于检索的上下文学习方法虽然有效，但依赖于可用示例的质量与多样性；过于特定领域或同质化的检索集合可能会限制结果的可泛化性。

此外，当前方法依赖语言模型进行事后评估来衡量解释质量，这可能引入偏见，或对推理能力给出不完整的指标。后续工作还可以整合多轮交互式推理过程，使模型在给出最终解释之前澄清歧义。自动评估指标的进步可为溯因推理质量提供更客观的评估。

此外，将基于检索的技术与模型微调或参数高效适配策略相结合，可能产生更稳健且可跨领域迁移的推理系统。最终，沿着这些方向推进，可以进一步增强多模态人工智能模型在处理复杂和非典型场景时的实用性、公平性与韧性。

9 道德考量

数据集构建。该数据集由网络来源的图像构建而成，并经过仔细筛选，以尽量减少不当和敏感内容。所有图像均由标注人员依据一套严格的准则进行审核，以确保该数据集不会传播偏见、刻板印象或有害的文化描绘。

文化与语境推理。MUN 中所呈现的推理任务鼓励模型生成以文化和上下文知识为基础的溯因解释。这引发了一种可能性：模型可能会无意中生成反映隐性偏见或文化不敏感叙事的内容。我们强调，使用多样化的评估者群体和检索语料库对于降低这些风险、提升公平性与包容性至关重要。我们鼓励研究人员、开发者和用户开展对抗性测试并进行持续监测，以识别和处理任何非预期的伤害。

负责任的应用与保障措施。最后，尽管我们追求的推理能力提升可能带来有益的现实应用——从医疗保健领域更准确的图像分析，到更好地理解全球文化现象——但它们也为更加复杂的图像与文本操纵打开了大门。开发者必须实施完善的护栏机制、透明度措施和用户同意机制，以确保这些先进的推理技术能够以负责任的方式服务公共利益，同时尊重隐私、文化价值观和知识产权。

参考文献

Marah Abdin、Jyoti Aneja、Hany Awadalla、Ahmed Awadallah、Ammar Ahmad Awan、Nguyen Bach、Amit Bahree、Arash Bakhtiari、Jianmin Bao、Harkirat Behl 等。2024。Phi-3 技术报告：可在手机本地运行的高性能语言模型。*arXiv* 预印本 *arXiv:2404.14219*。

Abdelrahman Abouelenin、Atabak Ashfaq、Adam Atkinson、Hany Awadalla、Nguyen Bach、Jianmin Bao、Alon Benhaim、Martin Cai、Vishrav Chaudhary、Congcong Chen 等。2025。Phi-4-mini 技术报告：通过 LoRA 混合实现紧凑而强大的多模态语言模型。*arXiv* 预印本 *arXiv:2503.01743*。

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。2023。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*。

Aishwarya Agrawal、Dhruv Batra 和 Devi Parikh。2017年。《别只凭假设；看一看再回答：克服视觉问答中的先验偏见》。发表于计算机视觉与模式识别会议（*CVPR*）。

Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson、Karel Lenc、Arthur Mensch、Katherine Millican、Malcolm Reynolds 等。2022年。Flamingo：用于少样本学习的视觉语言模型。神经信息处理系统进展，35：23716–23736。Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wen-bin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。2025年。Qwen2.5-vl 技术报告。*arXiv* 预印本 *arXiv:2502.13923*。

Chandra Bhagavatula、Ronan Le Bras、Chaitanya Malaviya、Keisuke Sakaguchi、Ari Holtzman、Han- nah Rashkin、Doug Downey、Wen-tau Yih 和 Yejin Choi。2020。溯因常识推理。载于国际学习表征会议。

Antoine Bosselut、Hannah Rashkin、Maarten Sap、Chai- tanya Malaviya、Asli Celikyilmaz 和 Yejin Choi。2019。COMET：用于自动知识图谱构建的常识变换器。*arXiv* 预印本 *arXiv:1906.05317*。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020。语言模型是少样本学习者。神经信息处理系统进展，33:1877–1901。

Zhe Chen、Weiyun Wang、Yue Cao、Yangzhou Liu、Zhangwei Gao、Erfei Cui、Jinguo Zhu、Shenglong Ye、Hao Tian、Zhaoyang Liu 等。2024。借助模型、数据和测试时扩展，拓展开源多模态模型的性能边界。*arXiv* 预印本 *arXiv:2412.05271*。

Aakanksha Chowdhery 等。2022年。PaLM：通过路径扩展语言建模。载于*arXiv* 预印本 *arXiv:2204.02311*。

Matthijs Douze、Alexandr Guzhva、Chengqi Deng、Jeff Johnson、Gergely Szilvasy、Pierre-Emmanuel Mazaré、MariaLomeli、LucasHosseini 和 Hervé Jégou。2024。FAISS 库。*arXiv* 预印本 *arXiv:2401.08281*。

Tianyu Gao、Xingcheng Yao 和 Danqi Chen。2021年。SimCSE：句子嵌入的简单对比学习。载于*2021*年自然语言处理实证方法会议论文集，第6894–6910 页，Online 和多米尼加共和国蓬塔卡纳。计算语言学协会。Seungju Han、J unhyeok Kim、Jack Hessel、Liwei Jiang、Jiwan Chung、Yejin Son、Yejin Choi 和 Youngjae Yu。2023。读书固然好，但开车时可不行！关于可违背常识规范的视觉基础推理。载于*2023*年自然语言处理实证方法会议论文集，第894–914 页，新加坡。计算语言学协会。

Lisa Anne Hendricks、Subhashini Venugopalan、Marcus Rohrbach、Raymond Mooney、Kate Saenko 和 Trevor Darrell。2016年。生成视觉解释。载于欧洲计算机视觉会议（*ECCV*），第 3–19 页。Jack Hessel、Jena D Hwang、Jae Sung Park、Rowan Zellers、Chandra Bhagavatula、Anna Rohrbach、Kate Saenko 和 Yejin Choi。2022年。夏洛克·福尔摩斯的溯因推理：一个用于视觉溯因推理的数据集。载于欧洲计算机视觉会议，第 558–575 页。施普林格。

Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph Gonzalez、Hao Zhang 和 Ion Stoica。2023。借助 PagedAttention 高效管理大型语言模型服务的内存。载于第*29*届操作系统原理研讨会论文集，第611–626页。

Bo Li、Yuanhan Zhang、Dong Guo、Renrui Zhang、Feng Li、Hao Zhang、Kaichen Zhang、Yanwei Li、Ziwei Liu 和 Chunyuan Li。2024。LLaVA-OneVision：简易视觉任务迁移。*arXiv*预印本 *arXiv:2408.03326*。

Junnan Li、Dongxu Li、Xiaohua Xie、Mark Yatskar 和 Steven C.H. Hoi。2023a。BLIP-2：利用冻结图像编码器和大语言模型引导语言-图像预训练。载于国际机器学习大会。

Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi。2022年。BLIP：面向统一视觉语言理解与生成的语言-图像预训练引导。载于国际机器学习大会，第12888–12900页。PMLR。

Junnan Li、Ramprasaath Selvaraju、Akhilesh Gotmare、Shafiq Joty、Caiming Xiong 和 Steven Chu Hong Hoi。2021年。先对齐再融合：基于动量蒸馏的视觉与语言表征学习。神经信息处理系统进展，34:9694–9705。

Liunian Harold Li、Mark Yatskar、Da Yin、Cho-Jui Hsieh、Kai-Wei Chang 和 Zhenyu Chi。2019年。Visu-alBERT：一种简单且性能出色的视觉与语言基线。*arXiv* 预印本 *arXiv:1908.03557*。Xuechen Li、Tianyi Zhang、Yann Dubois、Rohan Taori、Ishaan Gulrajani、Carlos Guestrin、Percy Liang 和 Tatsunori B. Hashimoto。2023b。Alpacaeval：一种用于评估指令遵循模型的自动评估器。https://github.com/tatsu-lab/alpaca_eval。Bill Yuchen Lin、Abhilasha Ravichander、Ximing Lu、Nouha Dziri、Melanie Sclar、Khyathi Chandu、Chandra Bhagavatula 和 Yejin Choi。2023。基础大语言模型上的解锁咒语：通过上下文学习重新思考对齐。在第十二届国际学习表征会议。Bill Yuchen Lin、Kangmin Tan、Chris Miller、Beiwen Tian 和 Xiang Ren。2022年。通过检索增强实现无监督跨任务泛化。进展

神经信息处理系统，35:22003– 22017。

Haotian Liu、ChunyuanLi、Qingyang Wu 和 Yong Jae Lee。2023a。LLaVA：大型语言和视觉助手。预印本，arXiv:2302.06675。

Yang Liu、Dan Iter、Yichong Xu、Shuohang Wang、Ruochen Xu 和 Chenguang Zhu。2023b。G-eval：使用 GPT-4 进行自然语言生成评估，获得更好的人类一致性。载于*2023*年自然语言处理实证方法会议论文集，第 2511–2522 页，新加坡。计算语言学协会。

Laurens van der Maaten 和 Geoffrey Hinton。2008。使用 t-SNE 可视化数据。机器学习研究期刊，9（11月）：2579–2605。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021年。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第 8748–8763 页。PMLR。

Colin Raffel、Noam Shazeer、Adam Roberts、Kather- ine Lee、Sharan Narang、Niki Matena、Niki Parmar、Yanqi Liu 和 Alex Jolicoeur-Martineau。2020。《利用统一的文本到文本转换器探索迁移学习的极限》。机器学习研究期刊，21:1–67。

Christoph Schuhmann、Lyonel Beaumont、Romain Vencu、Cade W. Gordon、Ross Wightman、Mehdi Cherti、Theo Coombes、Kartikay Katta、Chris Mullis、Roman Kaczmarczyk 等。2022年。LAION-5B：用于训练新一代图文模型的大规模开放数据集。预印本，arXiv:2210.08402。

Dustin Schwenk、Apoorv Khandelwal、Christopher Clark、Kenneth Marino 和 Roozbeh Mottaghi。2022年。A-OKVQA：一项利用世界知识进行视觉问答的基准测试。载于欧洲计算机视觉会议。

Yunfan Shao、Zhichao Geng、Yitao Liu、Junqi Dai、Hang Yan、Fei Yang、Zhe Li、Hujun Bao 和 Xipeng Qiu。2024。CPT：一种用于中文语言理解与生成的预训练非平衡转换器。中国科学：信息科学，67(5):152102。

Nisan Stiennon、Long Ouyang、Jeffrey Wu、Daniel M Ziegler、Ryan Lowe、Casey Voss、Alec Radford、Dario Amodei 和 Ilya Sutskever。2020。利用人类反馈学习摘要生成。载于神经信息处理系统进展。

Yan Tao、Olga Viberg、Ryan S Baker 和 René F Kizil-cec。2024。大语言模型的文化偏见与文化对齐。*PNAS Nexus*，3(9):pgae346。

Gemma团队、Aishwarya Kamath、Johan Ferret、Shreya Pathak、Nino Vieillard、Ramona Merhej、Sarah Perrin、Tatiana Matejovicova、Alexandre Ramé、Morgane Rivière等。2025。Gemma 3技术报告。*arXiv*预印本 *arXiv:2503.19786*。

Romal Thoppilan、Daniel De Freitas、Jamie Hall、Noam Shazeer、Apoorv Kulshreshtha、Heng-Tze Cheng、Alicia Jin、Taylor Bos、Leslie Baker、Yu Du等。2022年。LaMDA：面向对话应用的语言模型。*arXiv*预印本 *arXiv:2201.08239*。

Mor Ventura、Michael Toker、Nitay Calderon、Zorik Gekhman、Yonatan Bitton 和 RoiReichart。2024。NL-Eye：针对图像的溯因自然语言推理。*arXiv* 预印本 *arXiv:2410.02613*。

Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhihao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Ji alin Wang、Wenbin Ge等。2024。Qwen2-VL：增强视觉语言模型在任意分辨率下对世界的感知。*arXiv*预印本 *arXiv:2409.12191*。

Xingbo Wang、Renfei Huang、Zhihua Jin、Tianqing Fang和Huamin Qu。2023。CommonsenseVis：可视化并理解自然语言模型的常识推理能力。*IEEE*可视化与计算机图形学汇刊。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou等。2022年。思维链提示激发大语言模型中的推理。神经信息处理系统进展，35:24824–24837。

Shitao Xiao、Zheng Liu、Peitian Zhang 和 Niklas Muennighoff。2023 年。C-Pack：推进通用中文嵌入的资源包。预印本，arXiv:2309.07597。

Seonghyeon Ye、Doyoung Kim、Sungdong Kim、Hyeonbin Hwang、Seungone Kim、Yongrae Jo、James Thorne、Juho Kim 和 Minjoon Seo。2023 年。FLASK：基于对齐技能集的细粒度语言模型评估。*arXiv* 预印本 *arXiv:2307.10928*。

Wenting Zhao、Justin T Chiu、Jena D Hwang、Faeze Brahman、Jack Hessel、SanjibanChoudhury、Yejin Choi、Xiang Lorraine Li 和 Alane Suhr。2023 年。反常识推理：针对罕见情境的溯因推理。*arXiv* 预印本 *arXiv:2311.08469*。

Denny Zhou、Nathanael Schärli、Le Hou、Jason Wei、Nathan Scales、Xuezhi Wang、Dale Schuurmans、ClaireCui、OlivierBousquet、QuocLe 等人。2022年。由少到多提示使大语言模型能够进行复杂推理。*arXiv* 预印本 *arXiv:2205.10625*。

Yixin Zhu、Tao Gao、Lifeng Fan、Siyuan Huang、Mark Edmonds、Hangxin Liu、Feng Gao、Chi Zhang、Siyuan Qi、Ying Nian Wu 等。2020。《Dark, beyond deep: A paradigm shift to cognitive AI with humanlike common sense》。《Engineering》，6(3):310–345。

A 实验详情和超参数

表 6 展示了我们在实验中使用的模型的超参数；实验中使用的确切模型检查点见表7。除涉及 GPT-4o 的实验外，所有实验均使用两块 NVIDIA A6000 GPU 进行。

超参数	配置
文本嵌入模型	BAAI/bge-large-en
图像嵌入模型	clip-vit-base-patch16
图像分辨率	512×512
集成比例 α	0.4
检索库	langchain (https://python.langchain.com/docs/introduction/)
向量数据库库	FAISS(Douze 等人，2024)
vLLM 库。	VLLM(Kwon 等人，2023)

Tabl实验主要超参数配置

。

开源模型	
Gemma 3	google/gemma-3-4b-it
InternVL 2.5	OpenGVLab/InternVL2_5-8B
LLaVA-OneVision	llava-hf/llava-onevision-qwen2-7b-ov-hf
Phi 3.5-Vision	microsoft/Phi-3.5-vision-instruct
Phi 4-Multimodal	microsoft/Phi-4-multimodal-instruct
Qwen 2.5-VL	Qwen/Qwen2.5-VL-7B-Instruct
Qwen 2-VL	Qwen/Qwen2-VL-7B-Instruct
闭源模型	
GPT-4o	gpt-4o-2024-11-20（通过 OpenAI API）

表 7：实验中使用的模型检查点。开源模型通过 Hugging Face 获取，闭源模型（GPT-4o）通过 OpenAI API 获取。

B 双编码器检索机制

为了检索用于反常识推理的相关上下文示例，我们采用双编码器检索策略，计算并融合模态特定相似度分数。首先，分别使用 CLIP 风格的图像编码器 E_I 和基于 BERT 的文本编码器 E_T，对数据集 $D_{(i,t)}$ 中存储的（图像、文本）对进行嵌入。给定用户查询 $q = (q_i, \ q_t)$，我们计算查询向量 $(v_{q_i}, \ v_{q_t})$ 与存储向量 $(v_i, \ v_t)$ 之间的余弦相似度。最终相似度分数通过可调系数 α 对图像和文本相似度加权得出，该系数用于控制各模态的相对贡献。

算法 1 集成检索方法（双编码器）。该方法在视觉和文本嵌入空间中计算余弦相似度，使用 α融合这些相似度，并返回排名前 k 的匹配项。

1：输入： $q = (q_i, q_t)$；检索次数 k，权重比例 α2：输出：检索到的前 k 个（d_i, d_t）配对列表 3：包含（图像，文本）配对的向量数据库 $D_{(i,t)}$ 4：图像编码器 E_i 和文本编码器 E_t5：将查询转换为向量： $v_q = (v_{q_i}, v_{q_t}) = (E_i(q_i)\ E_t(q_t))$ 6：初始化结果 $\leftarrow$ []、索引 $\leftarrow$ []7：对于 $D_{(i,t)}$ 中索引为 j 的每个（v_i, v_t），执行 8：计算相似度：9：$s = \alpha \cdot \text{cos_sim}(v_{q_i}, v_i) + (1 - \alpha) \cdot \text{cos_sim}(v_{q_t}, v_t)$ 10：将 s 追加到结果 11：将 j 追加到索引 12：结束循环 13：将结果和索引组合成配对，并按相似度降序排序 14：选择前 k 个索引作为 TopKIndices 15：初始化 TopK $\leftarrow$ []16：对于 TopKIndices 中的每个 j，执行 17：将（d_{i_j}, d_{t_j}）追加到 TopK 18：结束循环 19：返回 TopK

C 定性结果



图 6：MUN-vis 定性结果。

图 6 展示了模型在 MUN-vis 中检索上下文知识并给出精确且富有解释性的回答的能力。对于第一行，当被问及冷藏橄榄油中形成的雾状现象时，模型将其类比为橙汁中出现的白色斑点。在这两种情况下，低温都会导致成分凝结并聚集：脂肪在橄榄油中凝固，而富含果肉和果胶的颗粒则在橙汁中结团。当液体恢复至室温后，它们会变得清澈，表明两种产品的质量都未受影响。这个例子展示了系统如何通过并置不同背景下罕见却相似的现象，增强其解释能力。

图 7展示了模型在检索上下文知识并以 MUN-lang 生成精准且具有解释性的回答方面的能力。对于第一行，虽然插线板乍看之下状况正常，但“烟雾和紧急疏散”这一结果促使模型通过溯因推理推断插线板存在“固有过热风险”。为促进这一推断，检索到的少样本示例包括“居民因舒适的壁炉冒出的烟雾而疏散”等场景。尽管这些示例描绘的是

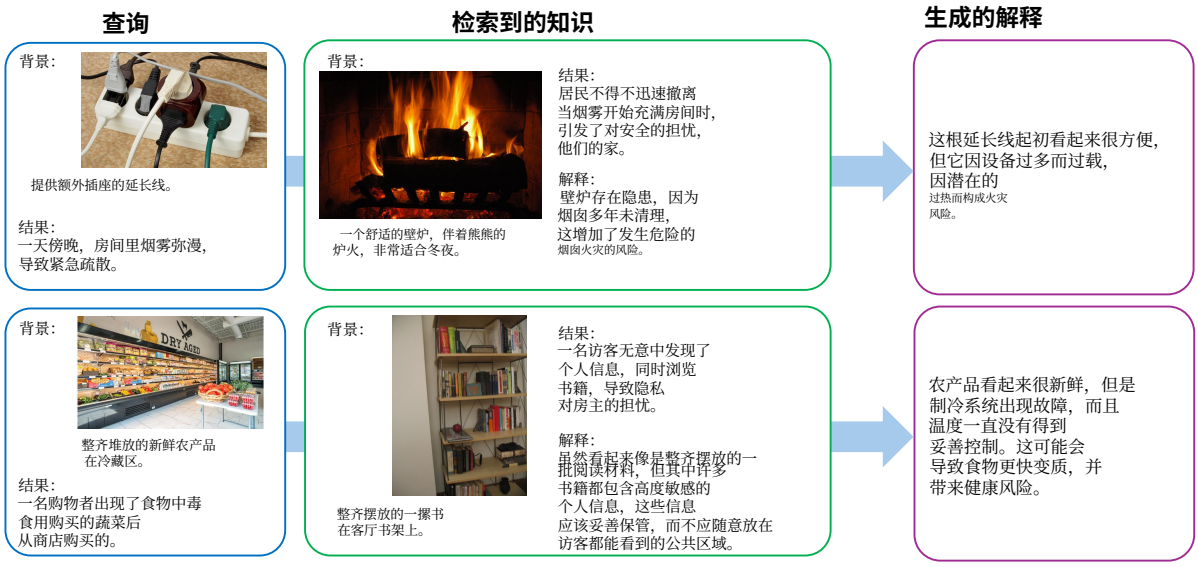


图 7：MUN-lang 定性结果。

不同的视觉主体，却引发了共同的因果模式：“看似正常的物体中潜藏着火灾隐患”。这与基于日常经验的溯因推理相联系，即“会产生热量的电器存在火灾隐患”，从而帮助模型推断视觉表象之外的潜在危险。

定性示例清楚地表明，MUN-vis 和 MUN-lang 探究了反常识推理中两个相互独立却又互补的方面。在 MUN-vis 中，必须借助特定的常识知识来对罕见的视觉线索进行常态化处理（例如，“浑浊的冷藏橄榄油” $\leftrightarrow$ “冰镇橙汁中的白色斑点”）；而 MUN-lang 则反转了这一挑战：常见的视觉场景掩盖了异常结果，因此需要通过溯因重构来推断隐藏风险（例如，“看似无害的插线板” $\rightarrow$ “隐蔽的火灾隐患” $\rightarrow$ “疏散”）。这两项任务共同构成了平衡的评估，考察模型能否将醒目的图像与日常事实联系起来，并推断意外事件背后不可见的因果机制。通过将这两个维度整合到单一基准测试中，并利用 MER 有针对性地检索具体类比或抽象因果模板，MUN 提供了一个全面的模型评估框架，涵盖从具体的常识锚定到抽象的因果推断。

D 人类对解释的一致性比较

我们开展了人工评估，将人类撰写的解释与两类模型生成的解释进行比较：LLM（由模型直接生成）和 HLLM（基于人类撰写的内容，由模型生成或由模型增强）。如表 12 所总结，LLM 生成的回答在感知质量方面没有显著下降。具体而言，70.8% 的大语言模型生成的解释与人类撰写的解释达到了高于中等水平的一致性，而 71.6% 的 HLLM 解释达到了这一水平的一致性。这些结果表明，就回答质量而言，模型生成的解释和模型增强的解释能够与人类撰写的解释高度匹配。

E 在 MUN 数据集上对 GPT4o 的评估

我们采用与第 5 节类似的设置，测试了 GPT-4o 在我们数据集上的表现。结果显示，GPT-4o 在 MUN-vis 和 MUN-lang 上均表现出色，其性能提升趋势总体上与开源模型相似。不过，由于有充分记录的“自我偏好偏差”现象——大语言模型倾向于偏好自己生成的回答——我们在初始实验中排除了 GPT-4o，并将相关结果附于附录中。

数据集	单样本		三样本		五样本	
	随机	R-ICL	随机	R-ICL	随机	R-ICL
MUN vis	0.572	0.597	0.604	0.610	0.673	0.704
MUN语言	0.678	0.636	0.671	0.664	0.650	0.692

表8：GPT4o 在不同样本设置下的评估结果，以对比人工辅助解释的胜率衡量（越高越好）。“随机”表示随机选择的示例，“R-ICL”表示为上下文学习检索的示例。模型输出与人类+大语言模型解释进行了比较，并由大语言模型进行评判。

α	0.3	0.4	0.5	0.6	0.7
胜率	0.572	0.618	0.611	0.611	0.603

表9：使用 Phi4-mm 模型对 MUN 视觉进行超参数 alpha 的消融研究。

数据集	模型	零样本	单样本		三样本		五样本	
			随机	R-ICL	随机	R-ICL	随机	R-ICL
MUN lang	*Gemma3*	0.259	0.364	**0.399**	0.294	0.301	0.259	0.287
	Phi3.5v	0.273	0.329	**0.336**	0.287	0.329	0.315	0.287
	Phi4mm	0.371	**0.497**	0.448	0.455	0.427	0.448	**0.497**
	Qwen2.5VL	0.364	0.322	0.357	0.287	**0.378**	0.273	0.322
MUN vis	*Gemma3*	0.333	0.233	0.289	0.226	**0.365**	0.233	0.314
	Phi3.5v	0.075	0.126	0.176	0.113	**0.201**	0.170	0.157
	Phi4毫米	0.195	0.239	0.214	0.132	**0.277**	0.101	0.170
	Qwen2.5VL	0.170	0.164	**0.358**	0.176	0.327	0.208	0.302

表10：不同样本设置下模型的比较，以相对于人工辅助解释的胜率衡量，并由开源大语言模型进行评判。“随机”表示随机选择的示例，“R-ICL”表示为上下文学习检索的示例。模型输出与人类+大语言模型的解释进行比较，并由开源大语言模型（Llama-4-Scout）评判。

F 超参数 alpha 的评估

表9 展示了不同超参数 α 对 Phi4-mm 模型在 MUN-vis 子集上 MER 性能的影响。根据表9的结果，我们在主要实验中使用了 α 值 0.4。

G 使用开源评判模型进行评估

为验证我们的 R-ICL 方法带来的性能优势具有稳健性，且不依赖于单一的专有评估器，我们使用先进的开源 Llama-4-Scout 模型作为评判模型，对比模型输出与人类+大语言模型解释。正如表10 所证实的，主要实验中观察到的核心趋势依然成立。尽管由于新评估器的偏好不同，绝对胜率有所差异，但在大多数模型和设置中，我们的检索增强上下文学习（R-ICL）始终优于零样本和随机少样本基线，或与其相比仍极具竞争力（以粗体突出显示的成果）。

H 在其他开源基准测试上评估 MER

为了提供关于我们的多模态实体识别框架泛化能力的实证证据，我们在 A-OKVQA 基准（Schwenk 等人，2022年）上进行了一项初步实验。我们测试了 Qwen-2.5-VL 在验证集随机选取的 500 道多项选择题上的准确率，并使用训练集中的 5000 个样本作为上下文学习上下文。表11 显示，R-ICL 的准确率高于零样本和随机样本基线。在 A-OKVQA 上取得的性能提升——这项任务需要同时具备视觉

采样模式	准确率
零样本	0.818
随机单样本	0.832
R-ICL 单样本	**0.842**

表 11：使用 Qwen-2.5-VL 在 A-OKVQA 数据集上对 MER 方法进行的消融研究。

理解能力和外部知识——有力表明，MER 检索相关上下文的能力是一项可泛化的原则。

层级	LLM		HLLM	
	Cnt	%	Cnt	%
1	136	13.6	123	12.3
2	162	16.2	161	16.1
3	194	19.4	204	20.4
4	221	22.1	255	25.5
5	287	28.7	257	25.7
Avg.	**3.36**		**3.36**	

表12：大语言模型与人类以及HLLM与人类回答的人类一致性等级分布（每组各抽取1000个样本）。平均分的计算假设等级1至5分别对应1至5分。

I 数据集类别

类别	MUN-视觉	MUN-语言
家居用品和家具	100	300
饮料	82	22
水果和蔬菜	80	8
工具和设备	57	143
乳制品和鸡蛋	54	0
健康与个人护理	44	15
罐装、包装和加工食品	36	5
商品		
肉类和海鲜	22	2
调味品和酱料	21	0
谷物 ，面包，和烘焙食品	19	5
总计	**515**	**500**

表13：视觉语境文本描述中物体类别数量的比较。每个数据集的总数量列于最后一行。

我们根据 MUN-vis 和 MUN-lang 的文本上下文选取了出现频率最高的 30 个类别。如表13所示，MUN-vis 更侧重于与食物相关的元素，而 MUN-lang 则更强调家居用品和家具物品。不过，在示例层面，这些子集仍涵盖多样的子类别和语境丰富的场景，如图8所示。

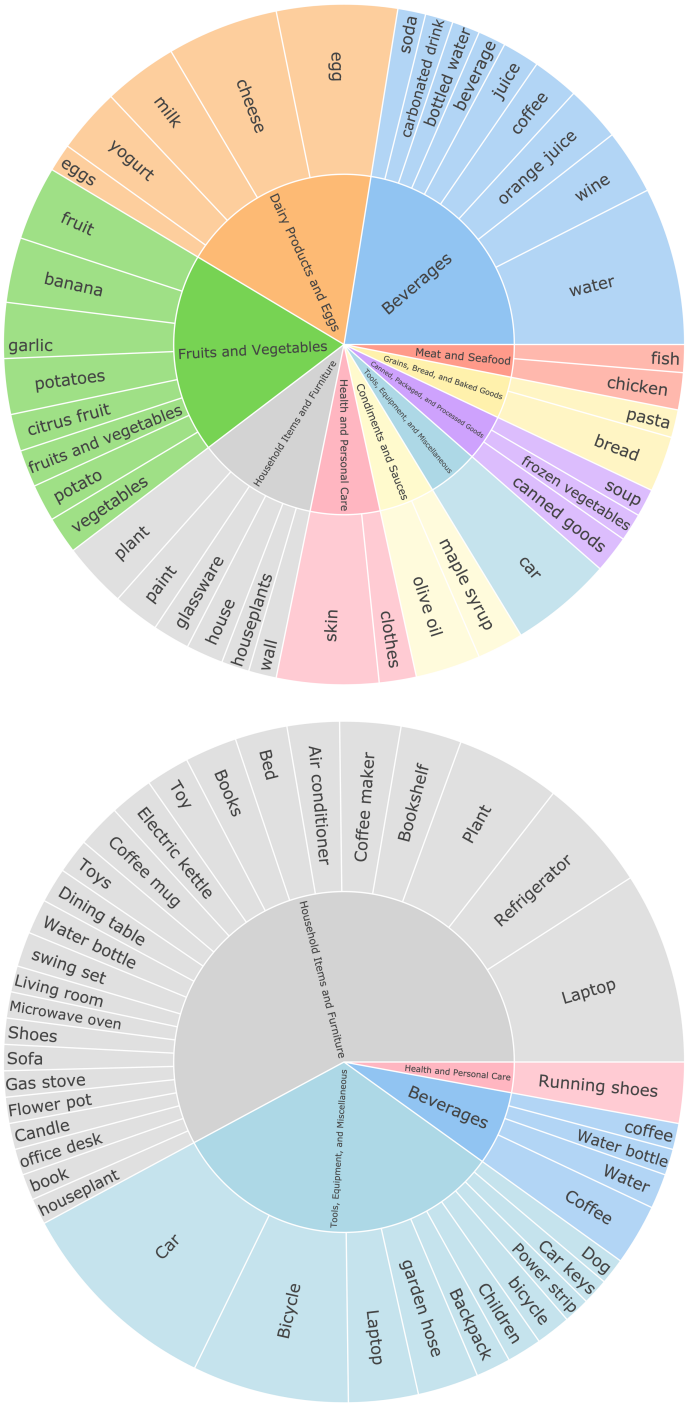


图 8：MUN-vis（上）和 MUN-lang（下）中的文本上下文分布。

J 实验中使用的提示词

Figu第 12 至 17 页展示了数据集生成和评估期间使用的各种提示词 n.

K 人工标注详情

K.1 人类数据集构建

为构建人工撰写的数据集，我们招募了 26 名研究生，生成语境化解释，以逻辑连贯的方式衔接给定的两个片段。我们要求标注人员遵循标准化界面，以引导他们撰写流畅连贯的衔接句。每条解释都参考了周围的背景，以确保叙事连贯性。用于收集人工解释的界面如图 9所示。

K.2 人类评估协议

我们通过 Prolific 平台开展了两项人工评估研究[5]，，招募母语为英语的参与者。

（1）人工一致性评估。为评估人类与模型生成的输出之间的对齐情况，我们要求标注人员针对随机抽取的500个样本中的每个样本，在两种比较中比较两个匿名化回答：（a）大语言模型 vs. 人类，以及（b）HLLM vs. 人类。每个样本由两名独立标注员进行评估，共产生2,000项评判。我们为此任务共招募了141名不同的参与者，工作人员的报酬标准为每小时C7.50。用于收集人类对解释的一致性评估的界面如图10所示。

(2) 胜率比较。我们进一步使用胜率评估设置，评估不同少样本提示变体（零样本、随机5样本、检索五样本）之间的相对回复质量。对于50个代表性样本中的每一个，我们构建了3组成对比较（例如，检索五样本与零样本），每种模态共进行150组成对比较。此项评估分别针对 MUN-LANG 和 MUN-VIS开展，共进行300组成对比较。每组比较由一名标注员评判。本任务共招募了20名不同的参与者，报酬为每小时 C7.71。用于收集胜率评判的界面如图 11所示。

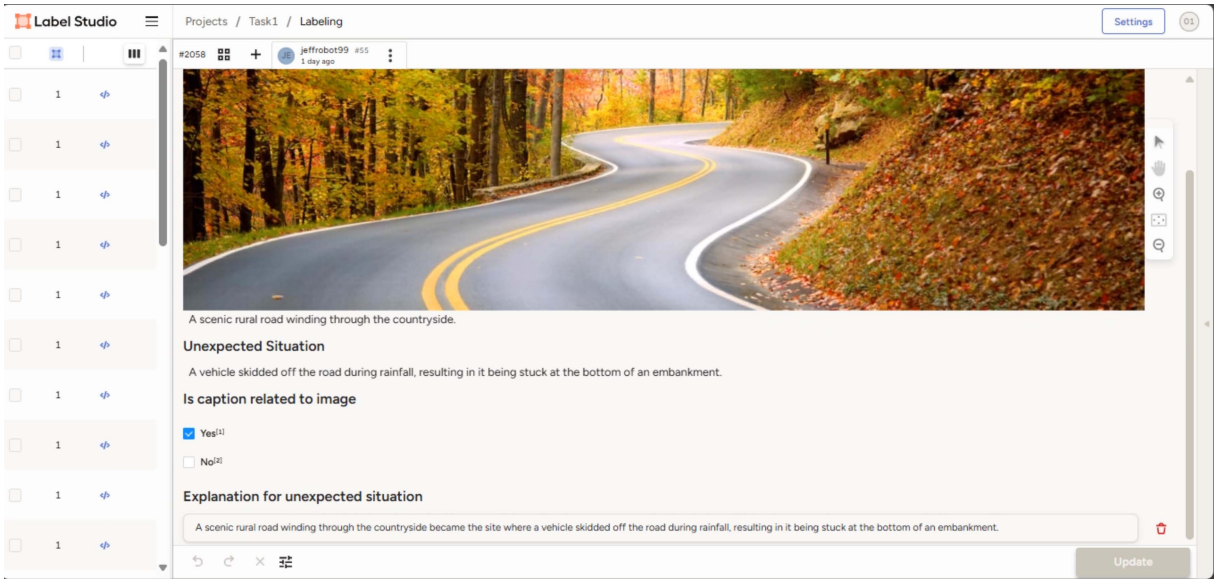


图9：用于人工标注任务的用户界面，旨在促进在MUN-vis和MUN-lang中创建详细且与背景相关的解释。

[5]https://www.prolific.com

通用说明（适用于所有问题）：请评估以下解释与典型人类推理（逻辑和质量）的契合程度。解释 A 和解释 B 分别随机分配给机器生成的解释和人类+ 大语言模型解释（由人工撰写并经人工智能增强），其确切来源均已隐藏。请评价每个解释（A 和 B）与人类推理的相似程度。



图像描述：食品储藏室里存放着各种食物和零食。

图像结果：一名人士食用了食品储藏室里的食物后，出现了食物中毒症状。

人类解释：那里的空气状况很差，里面又潮湿又热。里面的食物没几天就变质了。**A：**食品储藏室的空气状况不佳，湿度过高且温度过热，导致存放在里面的食物很快变质，进而使这名人士食用后出现食物中毒症状。

B：这间食品储藏室看起来物资充足，但存在问题，因为许多食物都已过保质期，食用后会带来健康风险。

以下是关于图像及其结果的几种可能解释。请根据逻辑和质量，评价你在多大程度上同意每种解释**与人工解释相似**。

	1 非常低一致性	2 低一致性	3 中等一致性	4 高一致性	5 非常高一致性
A	○	○	○	○	○
B	○	○	○	○	○

图10：用于人工一致性评估的界面。标注人员会看到两个匿名化回答，其中一个由人类撰写，另一个要么由大语言模型直接生成，要么由大语言模型根据人类版本修订而成，并被要求选出更合适的回答。

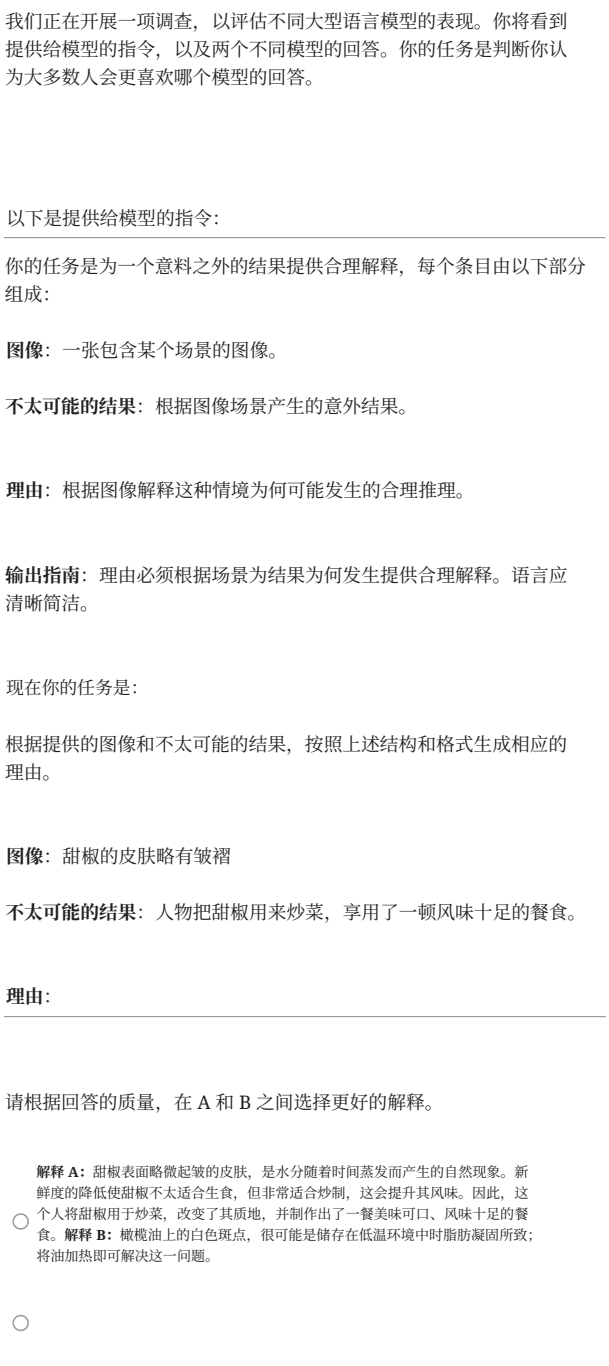


图11：用于人工标注员评估模型回答成对胜率的界面（例如，零样本与检索五样本的比较）。标注人员会看到两个匿名化输出，并被要求根据质量选择较好的一个。

你的任务是 with 生成一个数据集，其中每个 条目 由以下 组成 组件：

说明文字： A 简短描述 一个物体 或场景 在一张 图像中。
理由： A 合理推理 解释 why 物体或场景可能导致 to an 问题。
情境： A 潜力 结果 基于 根据说明文字和 理由 ， 不含 明确地 提及 原因。

指南 for 输出：
- 该 情境必须描述 结果，但不能直接 将其与 联系起来 理由。
- 使用清晰且 简洁的 语言。
- 将每个 条目格式化为 如下形式， 括在花括号中 括号 {} 以便 make 更容易 to 解析：

{ 说明文字： "<说明文字 文本 >"} { 理由： "<理由 文本 >"} {情境： "< 情境 文本 >"}

示例：

示例 1：
{ 说明文字： " 牛排包装中的红色液体 "} {理由： " 这 红色液体 发现于 牛排 包装 常被误认为 for 血。 它实际上是.....的混合物 水和 a 称为蛋白质的蛋白质 肌红蛋白 that 天然 存在 于肌肉 组织。 This 液体是 完全 正常的 并不表示 肉是 不安全的。"} {情境： " 人物烹制了 and 享用了 这份牛排 没有 健康 问题。"}

示例 2：
{ 说明文字： " 沉淀 液体在 酸奶中的 "} {理由: " 当 你打开一盒 of 酸奶， 你可能会看到一层 ..... 透明的 液体 上面， 这 有些人可能认为意味着变质。 This 液体只是在 whey 分离 from the 酸奶 固体， 一种天然的 过程 不会影响 酸奶的 质量。 搅拌 乳清 回到 the 酸奶 will 恢复 它的奶油般 质地 ."} { 情境： " 人物享用了 the 酸奶 as part 他们的 早餐。"}

示例 3：
{ 说明文字： " 绿色 铜制炊具上的铜锈 "} {理由: " 铜制炊具可能 形成一层 称为 铜锈的绿色层。Some 人们会误以为 this for 有害的 腐蚀， but 铜锈 是天然的 and 实际上可以保护铜 from 免受进一步氧化。 炊具 仍然是 清洁后可使用 妥善清洁。"} { 情境： " 人物使用了 铜制炊具 to 准备一份 美味的餐食。"}

示例 4：
{ 说明文字： " 发黄的叶子 叶片上 室内 植物 "} {理由: " 室内 植物 叶片 可能会开始变黄，这是其自然 生长过程的一部分 周期或由于 轻微胁迫 因素 例如浇水过多。 几片黄 叶片 并不 一定 表明 那株 植物 is 正在枯死。”} {情境： " 人物 继续照料 这株 植物， 它 grew 健康 new 叶片随时间变化。"}

示例 5：
{ 说明文字： " 皮肤 脱皮 晒伤后 "} {理由： “晒伤后， 皮肤 may 开始 脱皮。 This 脱皮是 is 自然愈合过程的一部分 在这里 the body 脱落 受损的 皮肤细胞。 虽然 它看起来可能 令人担忧， 这是 不是一个 一种正常的 对皮肤 损坏 由紫外线照射造成，且 原因 令人担忧的事。”} {情境： ” 人物涂抹了 保湿霜并 支持 皮肤的 愈合过程 舒适地。”}

Now 你的任务：
基于 根据 给定说明文字和理由， 生成 相应的 情境 遵循 上述结构和格式。

{ 说明文字： "{ 在此输入说明文字 }"}{ 理由： "{在此输入理由 }"

图12：用于生成MUN-vis场景的提示模板

你的任务是 with 生成一个数据集，其中每个 条目 由.....组成的 以下 组成部分：

说明文字： A 对.....的简短描述 一个物体 或场景 图像中的 图像。
理由： A 合理推理 解释 why 物体或场景可能导致 to an 问题。
情境： A 潜力 结果 基于 说明文字和 理由 ，没有 明确地 提及 原因。

指南 for 输出：
- 该 情境必须描述 结果，但不要直接 将其与 理由 联系起来。
- 使用清晰且 简洁 语言。
- 按以下格式整理每个 条目 as 如下， 用花括号括起来 括号 {} 以便 于轻松 to 解析：

{ 说明文字： "<说明文字 文本 >"} {理由： "<理由 文本 >"} { 情境： "< 情境 文本 >"}

示例：

示例 1：
{ 说明文字： "A 咖啡机已准备就绪 to brew the 完美 杯子。"} { 理由： " 虽然 咖啡机 看起来 功能正常， its 内部部件已腐蚀， 导致 to 潜力 污染的 冲泡的 咖啡。”} {情境： "A 客户 出现了 喝咖啡后胃部不适 由.....冲泡 机器 ."}

示例 2：
{ 说明文字： "A 流线型 跑车 停在 the 车道。"} { 理由： " 这辆 跑车 有问题 因为 它有一个未确诊的 机械故障， 使得 它很危险 to 驾驶。"} {情境： "该 驾驶员遭遇 a 突如其来的 失控 控制 当 驾驶时， 导致 轻微碰撞。”}

示例 3：
{ 说明文字： "A 丰富多彩的 toy 准备好了 for 玩耍时间。"} {理由： " 这 is 存在问题 因为该玩具 因存在安全 隐患而成为召回产品 可能 造成窒息 风险。"} {情境： "A 儿童短暂地 嗜住 当 演奏 with the 玩具，需要立即干预。”}

示例 4：
{ 说明文字： “一台桌面 计算机已就绪 可用于工作。"} {理由： "这台计算机 看起来 功能正常，但已严重 感染 with 恶意软件 可能会 危及敏感信息 。"} { 情境： " 用户 面临 未经授权的 访问他们的私人账户 计算机 for Online 在线交易。"}

Now 你的任务：
基于 所 提供的说明文字和理由， 生成 相应的 情境 遵循 上述结构和格式：

{ 说明文字： "{ 在此输入说明文字 }"}{ 理由： "{在此输入理由 }"

图13：用于为 MUN-lang 生成场景的提示模板

你能否 改进这一解释，使其更具体地针对 背景，并使结果更 可能 发生?

背景：{ INPUT CONTEXT HERE}
结果：{ INPUT OUTCOME HERE}
解释 对应结果的：{ INPUT EXPLANATION HERE}

图14：用于改进人工解释的提示模板

系统提示词

```
你是一个 有帮助的 助手, that 对模型进行排名 by the 质量 他们的回答
    .

提示词
我想 你创建一个 排行榜 不同的大型语言模型。 要执行
    so, I will give you the 指令（提示词） 给定 to 模型, 以及
    两个模型的回答。 请为模型排名 根据 哪些
    会 更受青睐 人类。 所有输入和 输出应为 回答
    字典。 Python

Here is the 提示词:
{
    "指令": """{指令}"""
}

以下是模型的输出:
[
    {
        "模型": "模型_1",
        "答案": """{输出_1}"""
    },
    {
        "模型": "模型_2",
        "答案": """{输出_2}"""
    }
]

Now 请对模型进行排名 by 质量 他们的回答, 以便模型
    with 排名1具有 the best 输出。 然后返回一个 模型列表 名称和
    排名, 即. 生成 the 以下 输出:
[
    {"模型": "模型_1", "排名": 1},
    {"模型": "模型_2", "排名": 2}
]

你的回答必须是有效的 Python 字典 and 应 不包含任何内容
    else 因为我们将直接 执行 它 Python。 请 提供 the
    排名 大多数 人类会 给出。
```

图15：用于评估胜率的提示模板

```
你的任务是 with 评估 具体性 给定文本在 1
    to 5.
1 （非常 Low 具体性）： 极其 模糊且 一般。
2 （低 具体性）： 有限的 细节, 大多 一般。
3 （中等 具体性）： 包含一些 细节 but 仍然 总体而言, 部件。
4 （高 具体性）： 包含 清晰且详细的 信息。
5（非常 高具体性）： 极其 详细 and 精确, 离开 没有房间 for
    歧义。

Only 将分数作为单个数字输出。

输入 文本:
[插入 生成的 文本在此处]

输出 格式:
[评分 (1 -5)]
```

图16：用于评估具体性的提示模板

```
你的任务是 with 评估 具体性 给定文本的具体程度，范围从 1 to 5.
1 （非常 Low 具体）： 极其 模糊且 笼统。
2 （低 具体性）： 有限 细节， 主要 一般。
3 （中等 具体性）： 包括一些 细节 but 仍然 一般在 笼统。
4 （高 具体性）： 包含 清晰且详细的 信息。
5（非常 高具体性）： 极其 详细 and 精确， 留下 没有余地 for 产生歧义。

Only 将分数以单个数字的形式输出。

输入 文本:
[在此插入 生成的 文本]

输出 格式:
[评分 (1-5)]
```

图17：用于评估具体性的提示模板