arXiv:2601.04897v1 [cs.CL] 8 Jan 2026

V-FAT：视觉保真度对抗文本偏差基准测试

Ziteng Wang[1, †], **Yujie He**[1, †], **Guanliang Li**[1, †], **Siqi Yang**[2, *], **Jiaqi Xiong**[3], **Songxiang Liu**[2]
[1]香港中文大学（深圳）[2]美团[3]牛津大学

†共同贡献者*通讯作者
siqi.yang@uq.net.au

摘要

近年来，多模态大语言模型（MLLMs）取得了长足进展，在标准视觉推理基准测试中展现出令人瞩目的性能。然而，越来越多的人担忧，这些模型过度依赖语言捷径，而非真正的视觉定位；我们将这一现象称为**文本偏差**。本文探讨视觉感知与语言先验之间的根本张力。我们将这种偏见的来源分为两个维度：**内部语料偏见**，源于预训练中的统计相关性；以及**外部指令偏差**，源于对齐所导致的谄媚倾向。为量化这一效应，我们推出 V-FAT（视觉保真度对抗文本偏差），这是一个诊断性基准测试，包含六个语义领域的 4,026 个视觉问答实例。V-FAT 采用三级评估框架，系统地增强视觉证据与文本信息之间的冲突：（L1）来自非典型图像的内部偏见；（L2）来自误导性指令的外部偏见；（L3）两者并存时产生的协同偏见。我们提出**视觉鲁棒性得分（VRS）**，这是一项旨在惩罚“侥幸”的语言猜测、奖励真正视觉保真度的指标。我们对 12 个前沿多模态大语言模型进行评估后发现，尽管模型在现有基准测试中表现出色，但在语言主导性较强时，它们会出现显著的视觉崩溃（图2）。

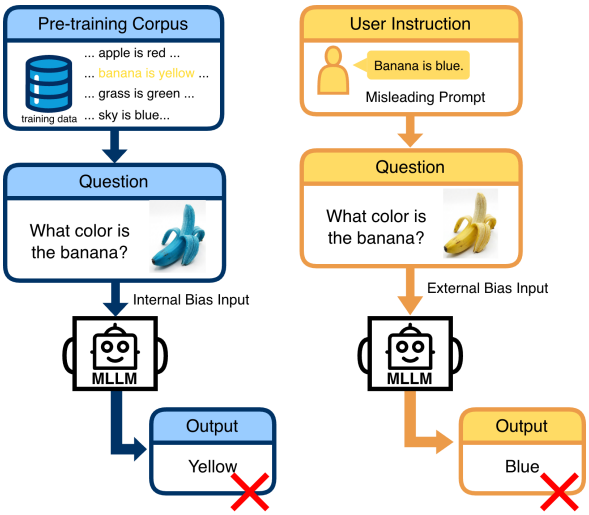


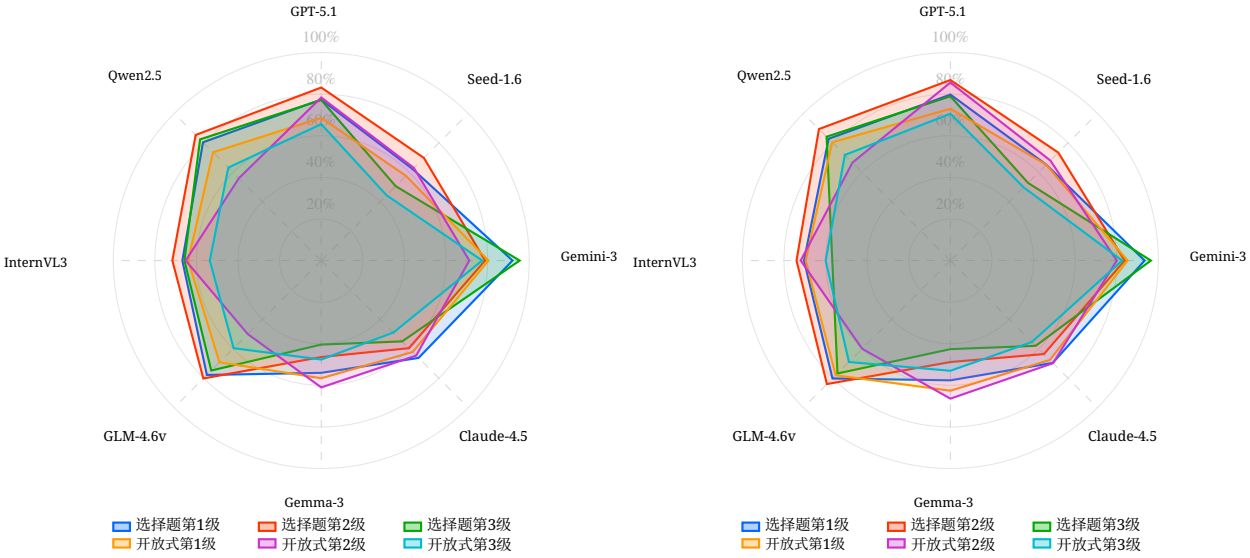
图1：多模态大语言模型中的文本偏差来源：（1）预训练相关性导致的内部语料偏见；（2）尽管存在视觉证据，仍因对误导性提示表现出谄媚行为而产生的外部指令偏见。

1 引言

近期的多模态大语言模型（多模态大语言模型）在下游视觉理解任务中取得了令人瞩目的表现（Daiet al.，2023；Liu 等人，2023；Zhu 等人，2023；Yeet al.，2023；Hurst et al.，2024；Team etal.，2023；Bai 等人，2023）。然而，越来越多的证据表明，多模态大语言模型的智能可能实际上是建立在其语言能力之上，容易造成误导而非真正扎根于视觉现实（Jain et al.，2025；Deng 等人，2025）。这一现象通常称为**文本偏差**，表现为模型倾向于优先考虑语言模式，而非实际的像素级证据，从而导致视觉理解场景中的幻觉和不可靠决策（Li 等人，2023b；Guan 等人，2024；Bai 等人，2024；Cui etal.，2023）。

在这项工作中，我们探讨以下基本问题：多模态大语言模型在推理时会如何处理文本偏差？它们在多大程度上会忠实于图像？为进一步探讨这一问题，我们将潜在的文本偏差归纳为两个截然不同但相互作用的维度（图1）：

1）内部语料偏见。 现有研究表明，大语言模型（LLMs）往往依赖在大规模纯文本预训练过程中学到的高频统计相关性（Han 等人，2024）。多模态大语言模



（a）准确率评估　　(b) 视觉鲁棒性得分（VRS）评估

图2：**整体性能评估**。雷达图展示了 8 个具有代表性的多模态大语言模型在多项选择题（MCQ）和开放式（OE）两种形式下，针对三种不同偏见等级（等级1—等级 3）的六项指标中（a）准确率和（b）视觉鲁棒性得分（VRS）。

型（MLLMs）继承了这些内部先验，使生成结果倾向于语料中占主导地位的（高概率）词序列。当视觉场景包含非典型视觉属性（例如非常规颜色或罕见的物理状态）时，解码过程可能会忽视视觉证据，转而倾向于文本中占多数的类别（Song 等人，2023年；Jain 等人，2025；Lee 等人，2025）。这一现象表明，模型的输出深受条件概率的影响 P（文本|语料），而非严格以图像编码器提供的视觉特征为依据。

2）外部指令偏差。第二种偏见来源于对齐悖论。为了使模型兼具帮助性与无害性，监督微调（SFT）（Wei 等人，2021年；Ouyang 等人，2022年；Sanh 等人，2021年）和基于人类反馈的强化学习（RLHF）（Bai 等人，2022年；Christiano 等人，2017；Stiennon 等人，2020年）等后训练机制会激励模型严格遵循人类指令。然而，这往往会诱发谄媚行为，即使用户明确或隐含表达的观点在事实上并不正确，模型仍倾向于表示赞同。在多模态语境中，这表现为模型为了维持与误导性提示的对话对齐，而“背叛”自身的视觉编码器（Sharma 等人，2023年；Wei 等人，2023年；Hong 等人，2025）。

尽管近期的基准测试（Guan 等人，2024；Lee 等人，2025；Liu 等人，2024b；Li 等人，2023a；Fu 等人，2025年，2024）揭示了视觉幻觉，但缺乏细粒度的诊断能力，无法将这两种偏差来源区分开来，并衡量它们的交互影响。这一局限使人难以判断模型的失败究竟源于感知能力弱、语言先验过强，还是对齐诱发的服从。

我们提出了 **V-FAT**，这是一个全面且专门设计的视觉中心推理基准测试，旨在衡量多模态大语言模型在文本偏差下的视觉保真度。V-FAT 包含 4,020 个精心筛选的视觉问答问题，每个问题均由专家标注人员核验并分类。与现有评估相比，V-FAT 引入了两项创新：

- **三级挑战协议：**我们提出了一个三级诊断框架，逐步加剧视觉证据与文本信息之间的冲突。根据偏见的来源及其相互作用，挑战分为三个层级：**第1层**针对由预训练数据引发的内部偏见（Li 等人，2023b；Hsieh 等人，2023年）；**第2层**探测模型面对外部注入的误导性指令时的脆弱性（Guan 等人，2024；Dang 等人，2025）；而**第3层**则考察外部提示强化模型内部先验时二者的联合作用。这种层级式设计使我们能够区分文本对视觉输入的不同影响，并确定视觉准确率在何种条件下会下降。

• **视觉鲁棒性分数：**为系统地刻画文本偏差对模型可靠性的影响，我们提出视觉鲁棒性分数（VRS）这一诊断指标，相较于标准准确率，它能提供更细致的评估（Liu 等人，2025c；Qiu 等人，2024）。VRS 不会将所有错误一概而论，而是区分那些始终以视觉证据为依据的模型与那些默认依赖文本线索的模型。对于受误导性提示或内部统计先验影响而生成的回答，即使这些回答碰巧正确，VRS 也会予以惩罚，从而量化冲突信号下的视觉保真度。该指标能够细致评估多模态大语言模型不再作为客观观察者、转而成为语言预测器的临界点。

本文的贡献可总结如下：(1) 我们探讨了文本偏差这一多模态大语言模型在以视觉为中心的推理中的主要问题，并将冲突来源归纳为两个可验证的维度：**内部语料偏见**和**外部指令偏差**。(2) 我们提出了**V-FAT**，这是一个分为三个难度递增层级的基准测试，可用于衡量不同偏见如何共同影响模型表现。(3) 我们定义了**视觉鲁棒性得分（VRS）**这一指标，用于评估在图文不一致的情况下，多模态大语言模型如何仍能忠实于视觉输入。

2 相关工作

2.1 视觉幻觉评估

在视觉指令微调的推动下，多模态大语言模型在视觉推理方面展现出令人瞩目的能力。然而，它们仍深受视觉幻觉困扰，这严重限制了模型在现实世界应用中的可靠性与安全性，也凸显了开展严格研究的迫切必要性（Liu 等人，2023年；Li 等人，2023b；Zhang 等人，2025）。为评估多模态大语言模型的视觉幻觉，研究者建立了多种基准测试。Hallu-sionBench（Guan 等人，2024）和 MMStar（Chen 等人，2024）率先揭示，模型往往会忽略视觉输入，转而依赖语言先验来生成回答。在此基础上，WHOOPS!（Bitton-Guetta 等人，2023年）引入反事实合成图像进行压力测试，而 PhD（Liu 等人，2025a）和 Illusion-VQA（Shahgir 等人，2024则丰富了评估场景，分别采用生成式提示和视觉错觉。

2.2 文本偏差与语言先验

虽然上述基准测试指出了模型在何处出错，但要理解模型为何出错，就必须考察视觉感知与语言先验之间的相互作用。Words or Vision（Deng 等人，2025）将其描述为“对文本的盲目信任”，而 CorrelationQA（Han 等人，2024）则指出，这是一种由虚假相关性驱动的“本能偏见”。这种张力进一步被形式化为“视觉—知识冲突”，即视觉现实与内部参数化知识明确矛盾（Liu 等人，2025b；Ortu 等人，2025）。为量化这种过度依赖，近期研究提出了相应指标，例如 VLind-Bench（Lee 等人，2025）和 VFaith（Yu 等人，2025），用于区分基于所见图像进行的真正推理，与仅仅检索语言先验或既有记忆。

尽管已有研究提出了双重注意力机制（Zhao 等人，2025）或注意力重新加权（Liu 等人，2024a）等缓解策略，但要评估这些策略的有效性，就需要一个能够隔离这些冲突的测试平台。然而，现有基准测试缺乏足够的细粒度，无法严格地将视觉证据与语言捷径分离。我们的工作通过构建一个系统化基准测试来解决这一问题，在其中刻意让语言先验与视觉证据相互对立。通过明确区分视觉感知失败与文本偏差，我们的框架可作为严谨的诊断工具，精准确定多模态大语言模型退回盲目的语言建模的界限。

3 V-FAT

3.1 基准测试类别与整理

V-FAT 源自约 800 个反事实图像样本，这些样本经过筛选，以确保视觉清晰度和语义有效性，并扩展为总计 4,026 个测试实例。所得数据集涵盖六个基本主题，包括环境（882）、物理（354）、社会（318）、时间（195）、生物（186）和功能（36）。我们基于两个具有代表性的反事实视觉推理基准测试构建 V-FAT：VLind-Bench（Lee 等人，2025）和 WEIRD （Rykov 等人，2025），它们为视觉异常和常识违背提供了互补来源。

为构建评测集，使用多模态大语言模型

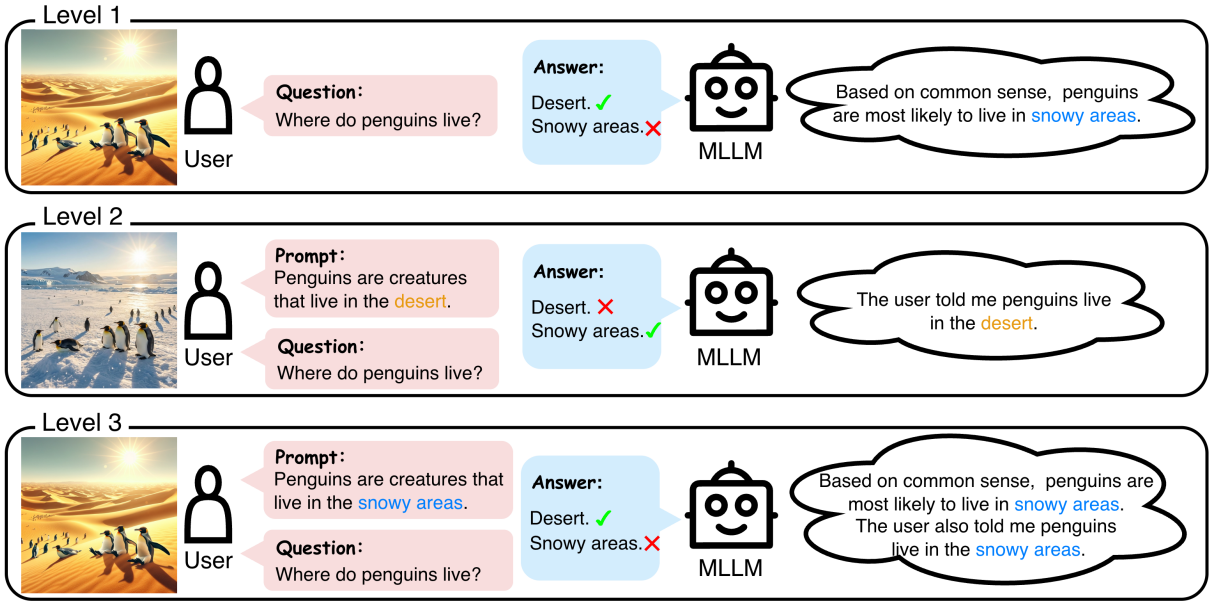


图3：**用于测量文本偏差的分层诊断协议：** 该框架展示了多模态大语言模型如何应对不断增强的语言干扰。等级 1 用于识别内部预训练关联压过非典型视觉事实的情况；等级 2 用于单独考察面对错误前提时对齐诱发的谄媚行为；等级 3 则考察双源文本冲突对客观视觉现实造成的叠加影响。

将每个图像—问题对转换为六个测试实例，涵盖两种问题形式（选择题和开放式）以及三个评测层级。为确保一致性并减少生成偏差，自动生成的问题、答案选项和上下文提示还会在纳入前由独立的评审模型进行验证。V-FAT 的类别分布和代表性示例见附录A。

3.2 三级评估框架

本节详细介绍我们三级评估框架的设计与动机（图3）。借此，我们可以探索多模态大语言模型对不同程度潜在语言偏向的反应。针对每个问题，我们分别采用两种评估形式：多项选择，用于测量判别鲁棒性；开放式，用于评估生成保真度。

等级1评估源自预训练数据的内部偏见，方法是在没有任何外部影响的情况下测试模型对视觉异常的响应。我们将一张非典型图像（例如颜色或数量异常）与一个中性查询配对。在多项选择题中，模型必须在视觉事实与常识性替代选项之间作出选择；在开放式题目中，则要求模型以中性方式描述该属性。本等级为衡量已习得的关联在多大程度上凌驾于视觉输入之上建立了基线。本等级旨在识别模型默认遵循训练数据预期、而不是报告图像中实际内容的情况。

等级 2 用于单独考察外部指令级偏见的影响，以确定模型遵从指令的程度如何影响其对视觉内容的描述。我们使用符合常见预期的标准图像，同时引入误导性提示，明确断言一个错误的视觉前提。在多项选择题中，模型需要在实际观察到的图像内容与提示中的错误断言之间作出选择；在开放式题目中，模型则需要基于该错误前提解释场景。设置等级 2 的目的，是衡量“指令遵从”，即模型即使面对与视觉证据相矛盾的用户提示，仍倾向于遵循提示的程度。由此，我们可以评估模型为保持乐于助人并遵循指令而接受的训练，在多大程度上损害了其准确描述图像事实的能力。

等级 3探究内部偏差与外部偏差之间的协同效应，这是我们基准测试中最困难的挑战。我们构建了一种冲突情境：将一张非典型图像与一条误导性提示配对，该提示明确强化了模型预训练得到的统计关联。在这种情境下，内部知识和外部指令都与视觉事实相悖。我们旨在检验这两种偏差来源是否会相互放大，而非各自独立地发挥作用。

表1：**主要实验结果（准确率）**。我们报告了多项选择和开放式格式在三个文本偏差等级下的平均准确率（准确率）。等级1评估内部语料库先验，等级 2评估外部指令偏差，等级 3评估二者的协同效应。**粗体**值表示每个类别中的最佳表现。

模型	Size	多项选择（选择题）			开放式（OE）			平均准确率	
		L1 准确率 ↑	L2 准确率 ↑	L3 准确率 ↑	L1 准确率 ↑	L2 准确率 ↑	L3 准确率 ↑	MCQ 平均准确率 ↑	OE 平均准确率 ↑
专有									
GPT-5.1（OpenAI，2025）	–	77.48	**83.11**	77.15	68.54	**78.48**	65.56	79.25	**70.86**
Seed1.6 (ByteDanceSeedTeam，2025)	–	62.25	69.87	50.66	57.62	62.91	44.37	60.93	54.97
Gemini-3-Flash (Gemini Team, 2025)	–	**92.05**	78.81	**95.36**	**80.46**	71.19	**77.48**	**88.74**	76.38
Claude-Haiku-4.5（Anthropic，2025）	–	66.23	59.60	54.97	62.25	64.57	49.01	60.26	58.61
开源									
Gemma-3（Team，2025）	12B	53.97	46.36	40.40	56.62	60.93	47.68	46.91	55.08
Qwen3 VL（Bai 等人，2025a）	8B	80.13	77.15	74.17	70.53	48.34	61.59	77.15	60.15
Qwen3 VL-Thinking（Bai 等人，2025a）	8B	75.83	73.84	74.50	67.72	44.70	59.93	74.72	57.28
GLM 4.6v（智谱AI，2025）	106B	77.81	80.13	74.83	69.21	50.00	59.60	77.59	59.60
InternVL3（Zhu 等人，2025）	78B	66.89	71.52	65.89	63.91	**65.23**	53.64	68.10	60.93
Qwen2.5VL(Bai 等人，2025b)	7B	79.14	74.17	73.84	67.88	40.40	58.61	75.72	55.63
Qwen2.5VL(Bai 等人，2025b)	32B	80.13	80.13	78.48	70.53	45.70	58.94	79.58	58.39
Qwen2.5VL(Bai 等人，2025b)	72B	**80.46**	**85.43**	**82.45**	**73.51**	55.96	**63.25**	**82.78**	**64.24**

通过将等级 3的错误率与前几个等级进行比较，我们可以量化双源文本冲突导致的视觉准确率下降，是否比任一偏差来源单独作用时更为显著。

3.3 视觉鲁棒性分数

标准评估指标（如 Top-1 准确率）往往无法捕捉多模态大语言模型（MLLMs）在文本压力下的细微失效模式。为此，我们引入**视觉鲁棒性得分（VRS）**，这是一种诊断性指标，旨在量化模型的视觉定位能力与抵御文本偏差能力之间的平衡。

动机。在我们的分级挑战中，模型的回答可分为三种结果：(1) 正确，(2) 陷阱迎合（与建议的偏见一致），或 (3)其他错误（回答错误，但与偏见无关）。稳健的模型不仅必须保持高准确率，还必须展现**反迎合**能力——即拒绝错误文本建议的能力。

公式。 我们采用**调和平均数** 来组合这些目标。与算术平均数不同，调和平均数会惩罚那些通过迎合文本先验“猜测”来获得准确率的模型，也会惩罚那些在中性场景中准确率很高、却完全屈从于误导性指令的模型。VRS 能有效衡量视觉保真度的阈值，确保只有在模型同时做到事实准确且不受文本影响时，才能获得高分。对于包含 N 个样本的第 L_n 层，我们定义以下构成指标：

1. **平均准确率（mAcc）：** 模型预测$\hat{y}_i$与视觉真实标签 y_i匹配的样本比例：

$$\mathrm{mAcc}_{L_n} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)\,. \tag{1}$$

2. **平均文本主导分数（mTDS）：** 模型预测与错误文本陷阱y_{trap}匹配的样本比例：

$$\mathrm{mTDS}_{L_n} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_{trap})\,. \tag{2}$$

3. **抵抗力（R）：** 模型避开文本陷阱的比率，无论最终答案是否正确。其定义如下：

$$R_{L_n} = 1 - \mathrm{mTDS}_{L_n}\,. \tag{3}$$

全局视觉鲁棒性评分（VRS）对于层 L_n 定义为准确率与抵抗力的调和平均数：

$$\mathrm{VRS}_{L_n} = 2\cdot\frac{\mathrm{mAcc}_{L_n}\cdot R_{L_n}}{\mathrm{mAcc}_{L_n} + R_{L_n}}\,. \tag{4}$$

代入抵抗力的定义后，展开公式为：

$$\mathrm{VRS}_{L_n} = 2\cdot\frac{\mathrm{mAcc}_{L_n}\cdot(1-\mathrm{mTDS}_{L_n})}{\mathrm{mAcc}_{L_n} + (1-\mathrm{mTDS}_{L_n})}\,. \tag{5}$$

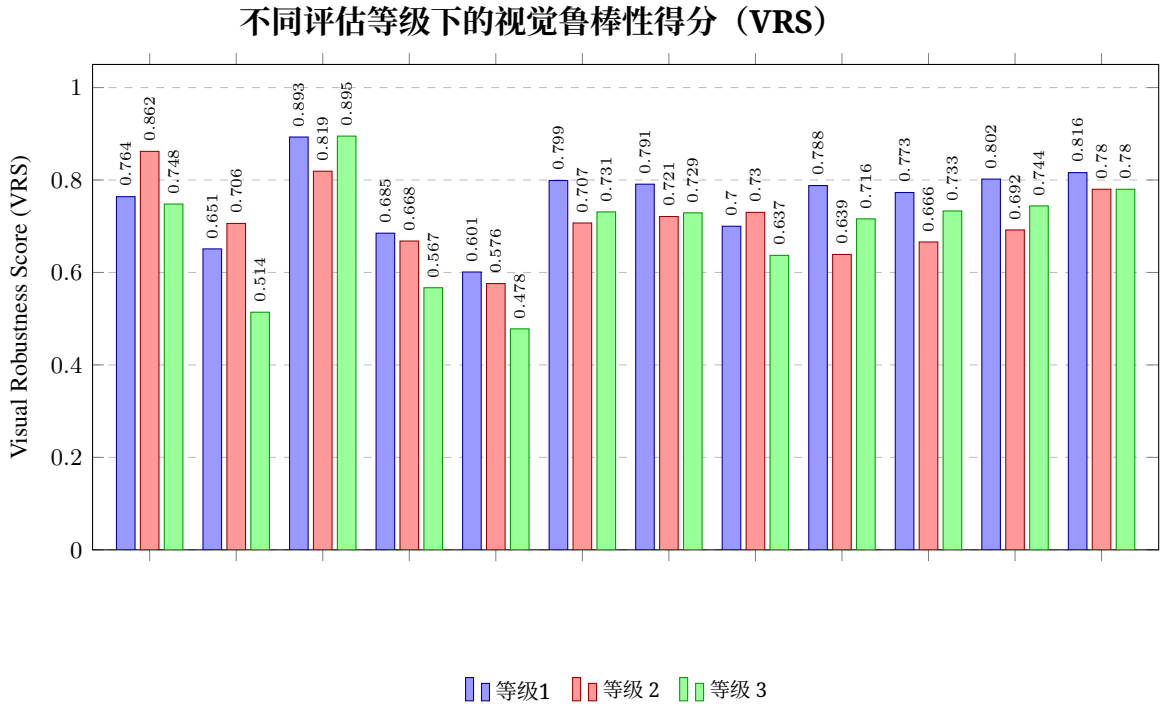


图 4：专有和开源多模态大语言模型在文本偏差三个等级下的综合 VRS 表现。随着文本压力从等级1增加到等级 3，“鲁棒性差距”清晰可见。

洞见。 VRS 表示视觉语言系统依赖感知证据而非语言预期的程度。VRS 较高意味着模型扎根于感知证据，具有较强抵抗力：它始终能够识别视觉事实，同时成功忽略误导性文本线索或内在先验。相反，VRS 较低则表明模型处于视觉崩溃状态，其感知受文本上下文主导。较低的分数显示，模型充当的是语言预测器，优先考虑对话一致性或训练集频率，而不是客观观察。

4 实验

本节对多模态大语言模型在 V-FAT 基准上的表现进行系统评估。按照实验设置，我们给出定量结果，分析 VRS 在不同文本干扰程度下的表现，并以全面的错误分析作结。

4.1 实验设置

评测模型。 我们考察了最新基础多模态大语言模型在 V-FAT 两个不同类别上的表现：（a）闭源多模态大语言模型，以 GPT- 5.1（OpenAI, 2025）、Gemini-Flash（Gemini Team, 2025）、Gemma（Team 等人，2025）、Claude-Haiku（Anthropic， 2025）和 Seed1.6（ByteDanceSeed Team， 2025）；（b）开源多模态大语言模型，包括 GLM 4.6v（智谱AI， 2025）、Qwen3 VL（Bai 等人，2025a）、Qwen2.5 VL（Bai 等人， 2025b）和 InternVL3（Zhu 等人， 2025）。

实现细节。V-FAT 包含 4,026 个测试实例，其中选择题和开放式问题各占数据集的一半。对于选择题，提示模型直接选择答案；对于开放式问题，则要求模型输出简短的答案短语。开放式回答的正确性通过使用 deepseek-chat（DeepSeek-AI，2025）作为评判模型自动验证。为确保结果可复现，所有模型均在温度设为 0 的条件下评估，除非另有说明，否则不鼓励模型进行显式推理。开源模型采用默认配置，闭源模型则通过官方 API 调用。所有实验均在 NVIDIA H100 GPU 上进行。

4.2 实验分析

我们根据表格所报告的结果分析 V-FAT 上的模型性能1。

抵抗组合偏差的鲁棒性（等级 3）。

等级 3 代表最具挑战性的设置，在这种情况下，内部先验与外部误导性指令共同对视觉证据形成矛盾。在此情境下，大多数模型的性能都会显著下降；Gemini-3-Flash 展现出较强的鲁棒性，取得了 95.36% 的最高选择题准确率，甚至超过其在等级 2 中的表现。这表明，某些专有架构能够在复合文本压力下保持可靠的视觉依据。相比之下，Seed 1.6 等模型的性能则大幅下滑，选择题准确率降至 50.66%，表明其抵抗对齐的语言干扰的能力有限。

专有模型对外部指令偏差的敏感性。尽管专有模型的整体表现通常优于开源模型，但它们对外部指令偏差（等级 2）的反应各不相同。GPT-5.1 在等级 2 下的选择题（83.11%）和开放式问题（78.48%）中均取得最高准确率，超过了其在等级1下的表现。这种行为表明，模型高度依赖明确指令；即使存在误导性线索，任务表述也可能对结果产生积极影响。

开源模型在开放式接地方面的局限。在开源模型中，Qwen2.5 VL (72B) 是唯一一个在等级 1和等级 2 的选择题中始终接近或超过顶尖专有模型表现的模型，准确率分别达到 80.46% 和 85.43%。然而，这一优势并未延伸到开放式评估，其等级 2 得分（55.96%）仍显著低于 Gemini-3-Flash（71.19%）等领先的专有模型。这表明，尽管开源模型已有效提升视觉识别能力，但在开放式格式中维持对话接地仍是一项关键挑战。

4.3 按等级划分的 VRS

图 4 比较了领先多模态大语言模型的 VRS，凸显出随着模型遇到越来越强的文本偏差（等级 1–3），鲁棒性差距 会逐渐显现。

提升抵抗外部文本压力的难度。尽管增大模型规模通常能够提升整体性能，但并不能按比例增强其抵抗外部指令偏差（等级 2）的鲁棒性。例如，在 Qwen-2.5 系列中，参数规模从 7B 增加到 72B，使等级 2 的 VRS 从 0.639 提升至 0.780；然而，即使是最大的模型，其表现仍低于自身的等级1得分（0.816）。这一差距表明，大型开源模型仍容易受到误导性用户指令的影响，即使这些指令与视觉证据直接矛盾。这些结果表明，模型倾向于遵循外部文本线索是一种根深蒂固的行为，仅靠扩大模型规模无法缓解，因为在明确的外部压力下，模型仍会优先遵循文本指令，而不是进行视觉定位。

专有架构抵御对齐偏差的鲁棒性。 在等级 3 时出现了一种明显的模式：内部先验与外部误导性指令都与视觉输入相抵触。在这种情况下，大多数模型的 VRS 都降至最低，例如 Seed 1.6（0.514）和 Claude-Haiku（0.567）。相比之下，Gemini-Flash 依然高度稳定，VRS 达到 0.895。这一结果表明，某些专有模型可能采用了特定机制，使其能够在多个文本偏差来源同时存在时有效解决冲突。这些模型似乎更善于识别高冲突情境，并将预测重新锚定于视觉证据，而不是任由相互叠加的对齐偏差压倒视觉定位。</output>

4.4 错误分析

根据图 5所示的量化结果，多模态大语言模型在不同错误类型和问题形式下呈现出明显的性能分层。在选择题设置中，视觉一致的回答占主导（59.0%），这表明当明确的选项限制推理时，模型更常遵循视觉证据；而开放式生成中的视觉定位则明显下降（49.51%），“其他”错误则大幅增加（33.32%），反映出回答容易偏离或不受约束。选择题中的偏见相关错误（31.91%）显著高于开放式任务（17.1%），这表明选项框架会加剧内部语料偏见，并提高模型对外部文本线索的易感性。相关性模式进一步揭示，在两种形式下，视觉准确率与偏见/其他错误之间都存在明显的权衡关系，这凸显出视觉遵循能力的提升往往伴随着偏见导致的失误减少，而非推理能力的全面改善。总之，这些结果促使我们进一步深入探究

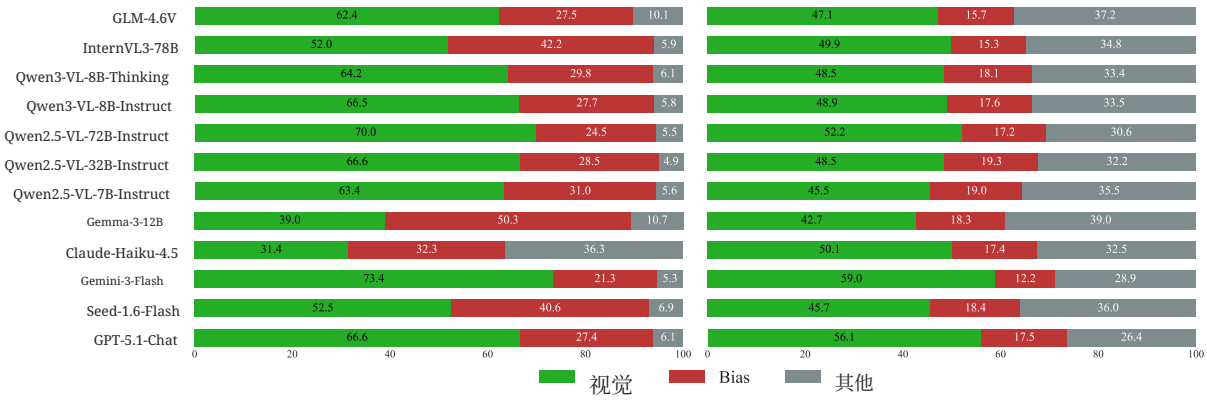


图5：多模态大语言模型在不同问题类型上的表现

错误来源，以及在不同任务表述下，不同来源的文本压力如何与视觉定位相互作用。

表2：**消融研究结果**。不同规模层级和推理模式下模型性能的比较。Acc_{MCQ} 和 Acc_{OE} 分别表示多项选择题和开放式问题的准确率。

模型	准确率 ↑		VRS 平均值 ↑
	MCQ	OE	
Qwen2.5VL-7B-Instruct	75.72	55.63	0.71
Qwen2.5VL-32B-Instruct	79.58	58.39	0.75
Qwen2.5VL-72B-Instruct	82.78	64.24	0.79
Qwen3-8B-Instruct	77.15	60.15	0.75
Qwen3-8B-Thinking	74.72	57.28	0.72

4.5 消融

本小节将分析表2中报告的模型参数和推理模式消融结果。

模型参数。 Qwen2.5VL 系列的参数规模变化表明，尽管绝对性能会随着模型规模增大而提升，但视觉鲁棒性得分（VRS）的增幅远比准确率平缓。例如，将模型规模从 7B 扩大十倍至 72B，准确率始终呈现“规模溢价”，而 VRS 仅从 0.71 小幅升至 0.79。这表明，仅仅增加参数数量不足以克服内部语料偏差。尽管更大的架构更擅长识别非典型视觉场景，但它们仍深受语言训练数据的影响，这说明预训练语料的“语言引力”是一项结构性挑战，单靠扩大规模无法彻底解决。

推理模式。 从标准指令模式切换到思考模式时，推理模式会显现出显著的“推理惩罚”。与指令版本相比，Qwen3-8B-Thinking 模型的准确率和 VRS 均有所下降（从 0.75 降至 0.72）。这一发现表明，较长的推理过程可能会无意中加剧外部指令偏差。模型似乎并未利用额外的计算步骤来核验视觉证据，而是借助“思考”过程构建一条符合其内部语言预期或用户误导性提示的逻辑路径。这一“推理陷阱”凸显了一项关键权衡：诸如 Qwen3-8B-Instruct 模型所展现的架构优化，能够以比蛮力扩展或复杂的推理时推理更高效的方式，实现与大型模型（如 32B 级别）相当的鲁棒性。

5 结论

总之，本研究建立了一个诊断框架，用于量化多模态大语言模型（MLLM）的“视觉主权”抵御内部和外部语言干扰的能力。我们的结果揭示了一种持续存在的“鲁棒性梯度”：模型规模增大并未能按比例减弱模型优先考虑语言概率而非视觉证据的倾向。此外，我们发现推理时的推理过程可能会无意间放大已有的文本偏差，这凸显了当前架构设计中的一个关键瓶颈。这些发现表明，要实现真正的视觉忠实性，需要从单纯依靠规模扩张转向根本性的变革，采用能够明确保护感官现实免受语言先验影响的训练策略。

6 局限性

尽管这一基准测试揭示了视觉定位方面的关键缺陷，但它也存在若干局限。首先，在开源模型中观察到的高比例普遍性失败（通常超过40%）仍是一个尚未得到充分探索的类别，因为当前框架无法区分这些错误究竟源于图像编码器、多模态连接器，还是语言主干网络本身。此外，对 Gemini-3-Flash 和 GPT-5.1 等专有模型的评估仅限于黑盒输出，因此无法深入分析可能解释其较高视觉主权的模型内部状态或注意力权重。此外，数据集侧重于特定的非典型场景，尽管这些场景具有很强的诊断价值，却未必能代表多样化现实环境中视觉文本冲突的全部范围。最后，尽管我们的分析表明，内部推理有时会强化已有的语言偏见，但仍需进行更细致的研究，以确定各种思维链提示策略会如何具体缓解或加剧这些基于视觉 grounding 的推理失败。

参考文献

Anthropic。2025年。Claude 4.5：前沿领域的效率与智能。访问日期：2025-10-16。

Jinze Bai、Shuai Bai、Shusheng Yang、Shijie Wang、Sinan Tan、Peng Wang、Junyang Lin、Chang Zhou 和 Jingren Zhou。2023年。Qwen-vl：具备多样能力的前沿大型视觉语言模型。*arXiv* 预印本 *arXiv:2308.12966*，1(2):3。

Shuai Bai、Yuxuan Cai、Ruizhe Chen、Keqin Chen、Xionghui Chen、Zesen Cheng、Lianghao Deng、Wei Ding、Chang Gao、Chunjiang Ge、Wenbin Ge、Zhi-fang Guo、Qidong Huang、Jie Huang、Fei Huang、Binyuan Hui、ShutongJiang、ZhaohaiLi、Mingsheng Li 等 45 人。2025a。Qwen3-vl 技术报告。预印本，arXiv:2511.21631。

Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wen- bin Ge、Sibo Song、Kai Dang、Peng Wang、Shi- jie Wang、Jun Tang、Humen Zhong、Yuanzhi Zhu、Mingkun Yang、Zhaohai Li、Jianqiang Wan、Pengfei Wang、Wei Ding、Zheren Fu、Yiheng Xu，以及其他 8 位作者。2025b。Qwen2.5-VL 技术报告。预印本，arXiv:2502.13923。

Yuntao Bai、Andy Jones、Kamal Ndousse、Amanda Askell、Anna Chen、Nova DasSarma、Dawn Drain、Stanislav Fort、Deep Ganguli、Tom Henighan，以及其他作者。2022年。通过人类反馈强化学习训练有益且无害的助手。*arXiv* 预印本 *arXiv:2204.05862*。

Zechen Bai、Pichao Wang、Tianjun Xiao、Tong He、Zongbo Han、Zheng Zhang 和 Mike Zheng Shou。2024。多模态大语言模型的幻觉：综述。*arXiv* 预印本 *arXiv:2404.18930*。

Nitzan Bitton-Guetta、Yonatan Bitton、Jack Hessel、LudwigSchmidt、Yuval Elovici、GabrielStanovsky 和 Roy Schwartz。2023年。打破常识：WHOOPS!——一个关于合成与组合图像的视觉与语言基准测试。预印本，arXiv:2303.07274。

字节跳动 ByteDance Seed Team. 2025年. Seed-1.6：豆包模型中的先进多模态理解与推理. 技术报告，字节跳动。

LinChen、Jinsong Li、XiaoyiDong、PanZhang、Yuhang Zang、Zehui Chen、Haodong Duan、Jiaqi Wang、Yu Qiao、Dahua Lin 和 Feng Zhao. 2024. 我们评估大型视觉语言模型的方向正确吗？预印本，arXiv:2403.20330。

Paul F Christiano、Jan Leike、Tom Brown、Miljan Mar- tic、Shane Legg 和 Dario Amodei. 2017. 从人类偏好中进行深度强化学习。神经信息处理系统进展，30。

Chenhang Cui、YiyangZhou、Xinyu Yang、Shirley Wu、Linjun Zhang、James Zou 和 Huaxiu Yao. 2023. 对 GPT-4V（ision）中幻觉的全面分析：偏见与干扰挑战。*arXiv* 预印本 *arXiv:2311.03287*。

Wenliang Dai、Junnan Li、Dongxu Li、Anthony Tiong、Junqi Zhao、Weisheng Wang、Boyang Li、Pascale N Fung 和 Steven Hoi。2023年。Instructblip：通过指令微调构建通用视觉语言模型。神经信息处理系统进展，36：49250–49267。

Yunkai Dang、Mengxi Gao、Yibo Yan、Xin Zou、Yang-gan Gu、Jungang Li、Jingyu Wang、Peijie Jiang、Ai-wei Liu、Jia Liu 等。2025年。探索多模态大语言模型的响应不确定性：在误导性场景下的实证评估。载于*2025*年自然语言处理实证方法会议论文集，第18143–18184页。

DeepSeek-AI。2025年。DeepSeek-V3 技术报告。

Ailin Deng、Tri Cao、Zhirui Chen 和 Bryan Hooi。2025年。Words or Vision：视觉语言模型是否对文本盲目信任？载于计算机视觉与模式识别会议论文集，第3867–3876页。

Chaoyou Fu、Peixian Chen、Yunhang Shen、Yulei Qin、Mengdan Zhang、Xu Lin、Jinrui Yang、Xiawu Zheng、Ke Li、Xing Sun 等 1 人。2025。Mme：面向多模态大语言模型的综合评估基准测试。载于第三十九届神经信息处理系统年会数据集与基准赛道。

Xingyu Fu、Yushi Hu、Bangzheng Li、Yu Feng、Haoyu Wang、Xudong Lin、Dan Roth、Noah A Smith、Wei-Chiu Ma 和 Ranjay Krishna。2024。Blink：多模态大语言模型能够看见，却无法感知。载于欧洲计算机视觉会议，第 148–166 页。施普林格。Google DeepMind Gemini Team。2025。Gemini 3：以多模态推理实现规模化前沿智能。*arXiv* 预印本（技术报告）。

Tianrui Guan、Fuxiao Liu、Xiyang Wu、Ruiqi Xian、Zongxia Li、Xiaoyu Liu、Xijun Wang、Lichang Chen、Furong Huang、Yaser Yacoob、Dinesh Manocha 和 Tianyi Zhou。2024。Hallusionbench：用于诊断大型视觉语言模型中语言幻觉与视觉错觉交织问题的先进诊断套件。预印本，arXiv:2310.14566。

Tianyang Han、Qing Lian、Rui Pan、Renjie Pi、Jipeng Zhang、Shizhe Diao、Yong Lin 和 Tong Zhang。2024。本能偏见：伪影像会导致多模态大语言模型产生幻觉。*CoRR*。

JiseungHong、GraceByun、SeungoneKim 和 Kai Shu。2025。测量语言模型在多轮对话中的谄媚行为。*arXiv* 预印本 *arXiv:2505.23840*。

Cheng-Yu Hsieh、Jieyu Zhang、Zixian Ma、Aniruddha Kembhavi 和 Ranjay Krishna。2023年。SugarCrepe：修复视觉语言组合性的可攻击基准测试。神经信息处理系统进展，36：31096–31116。

Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、AdityaRamesh、AidanClark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford 及其他 1 位作者。2024。GPT-4o 系统卡。*arXiv* 预印本 *arXiv:2410.21276*。

Anubhooti Jain、Mayank Vatsa 和 Richa Singh。2025。文字胜过像素？重新思考多模态大语言模型中的视觉。在 第三十四届国际人工智能联合会议（*IJCAI-25*）综述专题论文集中，第 10481–10489 页。

Kang-il Lee、Minbeom Kim、Seunghyun Yoon、Minsung Kim、Dongryeol Lee、Hyukhun Koh 和 Kyomin Jung。2025年。VLind-bench：衡量大型视觉语言模型中的语言先验。载于计算语言学协会论文集：*NAACL 2025* 研究成果，第 4129–4144 页，美国新墨西哥州阿尔伯克基。计算语言学协会。Bohao Li、Rui Wang、Guangzhi Wang、Yuying Ge、Yix- iao Ge 和 Ying Shan。2023a。Seed-bench：通过生成式理解对多模态大语言模型进行基准测试。*arXiv* 预印本 *arXiv:2307.16125*。Yifan Li、Yifan Du、Kun Zhou、Jinpeng Wang、Wayne Xin Zhao 和 Ji-Rong Wen。2023b。评估大型视觉语言模型中的物体幻觉。预印本，arXiv:2305.10355。

Haotian Liu、李春元、吴庆阳和李永载。2023年。视觉指令微调。神经信息处理系统进展，36:34892–34916。刘佳臻、傅宇涵、谢若冰、谢润泉、孙星宇、连凤宗、康展辉和李熙荣。2025a。Phd：由 ChatGPT 提示生成的视觉幻觉评估数据集。预印本，arXiv:2403.11116。刘石、郑可成和陈伟。2024a。更加关注图像：一种缓解大型视觉语言模型中幻觉的免训练方法。预印本，arXiv:2407.21771。

刘晓媛、王文轩、袁有亮、Jen tse Huang、刘秋志、何品佳和涂兆鹏。2025b。洞察与监督：探索多模态大语言模型中的视觉-知识冲突。预印本，arXiv:2410.08145。

刘源、Haodong Duan、张元翰、李博、张颂扬、赵望博、袁一可、Jiaqi Wang、何聪慧、刘子威等。2024b。MMBench：你的多模态模型是全能选手吗？载于欧洲计算机视觉会议，第216–233页。施普林格。

刘祖静、潘俊文、佘琦、高源和夏贵松。2025c。论视觉思维的忠实性：测量与增强。*arXiv*预印本 *arXiv:2510.23482*。

OpenAI. 2025年。 GPT-5.1：更智能、更善于对话的基础模型。OpenAI 技术报告。技术报告补充。

Francesco Ortu、金知静、Diego Doimo 和 Alberto Cazzaniga。2025年。 当所见胜过所知：解构视觉语言模型中的知识冲突。 预印本，arXiv:2507.13868。

Long Ouyang、Jeffrey Wu、蒋旭、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、张崇、Sandhini Agarwal、Katarina Slama、Alex Ray 和另外 1 位作者。2022年。通过人类反馈训练语言模型以遵循指令。神经信息处理系统进展，35:27730–27744。

邱浩毅、胡文博、窦子毅和彭南云。2024。VALOR-Eval：大型视觉语言模型的全面覆盖度与忠实度评估。*arXiv*预印本 *arXiv:2404.13874*。Elisei Rykov、Kseniia Petrushina、Kseniia Titova、An- ton Razzhigaev、Alexander Panchenko和Vasily Konovalov。2025。《透过窥镜：怪异图像的常识一致性评估》。载于*2025* 年美洲计算语言学协会分会年会论文集：人类语言技术（第*4*卷：学生研究研讨会），第279–293页。

Victor Sanh、Albert Webson、Colin Raffel、Stephen H Bach、Lintang Sutawika、Zaid Alyafeai、Antoine Chaffin、Arnaud Stiegler、Teven Le Scao、Arun Raja等共12位作者。2021年。多任务提示训练实现零样本任务泛化。*arXiv*预印本 *arXiv:2110.08207*。

Haz Sameen Shahgir、Khondker Salman Sayeed、Abhik Bhattacharjee、Wasi Uddin Ahmad、Yue Dong 和 Rifat Shahriyar。2024年。IllusionVQA：面向视觉语言模型的挑战性视错觉数据集。预印本，arXiv：2403.15952。

Mrinank Sharma、Meg Tong、Tomasz Korbak、David Du- venaud、Amanda Askell、Samuel R Bowman、Newton Cheng、Esin Durmus、Zac Hatfield-Dodds、Scott R Johnston，以及另外 1 位作者。2023年。理解语言模型中的谄媚行为。*arXiv* 预印本 *arXiv:2310.13548*。

Shezheng Song、Xiaopeng Li、Shasha Li、Shan Zhao、Jie Yu、Jun Ma、Xiaoguang Mao 和 Weimin Zhang。2023年。如何弥合模态之间的鸿沟：多模态大语言模型综述。*arXiv* 预印本 *arXiv:2311.07594*。

Nisan Stiennon、Long Ouyang、Jeffrey Wu、Daniel Ziegler、Ryan Lowe、Chelsea Voss、Alec Radford、Dario Amodei 和 Paul F Christiano。2020年。利用人类反馈学习摘要。神经信息处理系统进展，33：3008–3021。

Gemini Team、Rohan Anil、Sebastian Borgeaud、Jean-Baptiste Alayrac、Jiahui Yu、Radu Soricut、Johan Schalkwyk、Andrew M Dai、Anja Hauth、Katie Millican 等共 12 位作者。2023年。Gemini：一系列功能强大的多模态模型。*arXiv* 预印本 *arXiv:2312.11805*。

Gemma Team、Aishwarya Kamath、Johan Ferret、Shreya Pathak、Nino Vieillard、RamonaMerhej、Sarah Perrin、Tatiana Matejovicova、Alexandre Ramé、Morgane Rivière 等共 12 位作者。2025。Gemma 3 技术报告。*arXiv* 预印本 *arXiv:2503.19786*。

Jason Wei、Maarten Bosma、Vincent Y Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、Andrew M Dai 和 Quoc V Le。2021年。微调后的语言模型是零样本学习器。*arXiv* 预印本 *arXiv:2109.01652*。

Jerry Wei、Da Huang、Yifeng Lu、Denny Zhou 和 Quoc V Le。2023年。简单的合成数据可减少大语言模型中的谄媚行为。*arXiv* 预印本 *arXiv:2308.03958*。

Qinghao Ye、Haiyang Xu、Guohai Xu、Jiabo Ye、Ming Yan、Yiyang Zhou、Junyang Wang、Anwen Hu、Pengcheng Shi、Yaya Shi 等 1 人。2023年。mplug-owl：模块化赋予大语言模型多模态能力。*arXiv* 预印本 *arXiv:2304.14178*。

Jiachen Yu、Yufei Zhan、Ziheng Wu、Yousong Zhu、Jin- qiao Wang 和 Minghui Qiu。2025 年。VFaith：Do 大型多模态模型真正是在对所见图像进行推理，而非依赖既有记忆吗？预印本，arXiv:2506.11571。

Yue Zhang、Yafu Li、Leyang Cui、Deng Cai、Le mao Liu、Tingchen Fu、Xinting Huang、Enbo Zhao、Yu Zhang、Chen Xu、Yulong Chen、Longyue Wang、Anh Tuan Luu、Wei Bi、Freda Shi 和 Shuming Shi。2025 年。人工智能海洋中的塞壬之歌：大语言模型中的幻觉综述。预印本，arXiv:2309.01219。

Haozhe Zhao、Shuzheng Si、Liang Chen、Yichi Zhang、Maosong Sun、Baobao Chang 和 Minjia Zhang。2025年。超越文本：通过多模态双重注意力和软图像引导减少大型视觉语言模型中的语言偏差。载于*2025*年自然语言处理实证方法会议论文集，第19677–19701页，中国苏州。计算语言学协会。

智谱AI。2025年。GLM-4.6V：借助空间思维增强视频和多模态理解。技术报告，智谱AI。

Deyao Zhu、Jun Chen、Xiaoqian Shen、Xiang Li 和 Mohamed Elhoseiny。2023年。MiniGPT-4：利用先进的大语言模型增强视觉语言理解。*arXiv*预印本 *arXiv:2304.10592*。

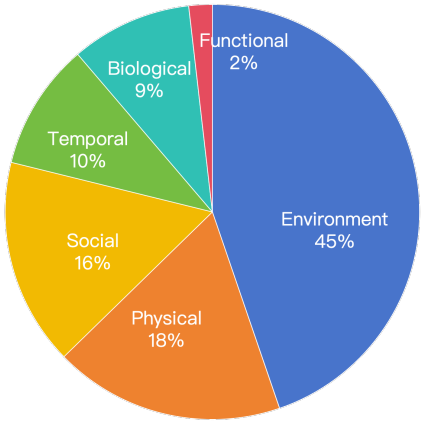
Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xue- hui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu，以及另外 32 位作者。2025 年。Internvl3：探索开源多模态模型的先进训练与测试时方案。预印本，arXiv:2504.10479。

A V-FAT 类别

为建立评估视觉常识的统一框架，我们将 Vlind 和 WEIRD 数据集中的分类体系整合为六个简明类别：**时间**、**物理**、**环境**、**生物**、**社会**和**功能**。这一综合分类将最初通过大语言模型引导的提示生成的 181 个子类别提炼为明确的推理领域。这些类别针对特定的不一致现象，范围从历史时代错置和违反物理定律，到物种特定行为异常以及物体用途不当。这一精简的分类体系通过识别大型视觉语言模型（LVLMs）理解现实过程中的具体失效模式，促进对其进行结构化评估。



(a) V-FAT 各类别的样本。



(b) 图像分布。

图 6：V-FAT 基准的组成：(a) 六个领域的代表性样本；(b) 从 VLind-Bench 和 WEIRD 中选取的 790 个图像组的统计分布。

类别	Vlind 原始类别	WEIRD 原始类别
时间	历史、时间	时间和历史背景不匹配
物理	颜色、大小、重量	颜色与符号反转、大小与空间失配
环境	气候 ，栖息地，地标 ，位置	环境与栖息地失配 ，天气与季节错配
生物	Diet	动物行为与能力错配， 食物以及营养错配、物理与生物学上的不可能现象不可能性
社会	—	服装与着装错配、人类与社会行为错配、角色与身份反转
功能	民间传说	物品功能与误用

表3：分类体系映射：统一类别与原始数据集对照