arXiv:2602.04413v1 [cs.CL] 4 Feb 2026

基于历史引导并具备自我纠正能力的迭代视觉推理 n

Xinglong Yang [1] Zhilin Peng [1] Zhanzhan Liu [1] Haochen Shi [*1] Sheng-Jun Huang [*1]

摘要

自一致性方法是提升多模态大语言模型（MLLM）推理可靠性的核心技术。它们通过重复采样生成多个推理结果，并通过投票选出最佳答案，在跨模态任务中发挥着重要作用。然而，现有的大多数自一致性方法都局限于固定的“重复采样与投票”范式，且不会重用历史推理信息。因此，模型难以主动纠正视觉理解错误，也难以在迭代过程中动态调整推理。受人类通过重复验证和动态纠错进行推理的行为启发，我们提出了 H-GIVR 框架。在迭代推理过程中，多模态大语言模型会多次观察图像，并将先前生成的答案作为后续步骤的参考，从而动态纠正错误并提高答案准确率。我们在五个数据集和三种模型上进行了全面实验。结果表明，H-GIVR 框架能够显著提升跨模态推理准确率，同时保持较低的计算成本。例如，在 ScienceQA 数据集上使用 Llama3.2-vision:11b 时，模型平均每个问题需要生成 2.57 次回答，即可达到 78.90% 的准确率，较基线提升了 107%。

1. 引言

近年来，多模态大语言模型在视觉—语言推理任务上取得了显著进展，初步展现出多步推理和跨模态对齐能力（Li 等人，2022；Alayrac 等人，2022；Liu 等人，2023）。然而，单纯增大模型规模以提升性能会导致训练成本呈指数级增长，这表明仅依靠模型扩展并非可持续的发展路径（Kaplan 等人，2020；Hoffmann 等人，2022）。因此，提示工程已成为提升大语言模型推理性能的关键研究方向之一。

以往研究表明，显式提示可以引导模型生成推理步骤（Wei 等人，2022），从而提升模型在复杂推理任务上的表现。此外，重复采样结合多数投票可以增强模型输出的稳定性和可靠性（Wang 等人，2022）。然而，现有的大多数自洽性方法都将多次生成视为相互独立的重复采样，没有充分利用历史输出作为后续推理的参考。尽管提示的复杂度和多样性会显著影响性能（Fu 等人，2022），这种独立重复采样策略难以在多次迭代中纠正动态错误，因此限制了模型推理能力的进一步提升。

人类推理的一项关键优势不在于机械重复，而在于通过迭代验证不断完善结论。例如，考试时，学生会结合先前的作答与问题中的视觉线索来复查答案：简单的问题会快速核对，而较难的问题则会调整策略、反复审视。大语言模型最初旨在模仿人类对话逻辑(Cai 等人，2024)，但当前方法尚未具备这种验证行为。受此启发，我们提出了**基于历史引导并具备自我纠正能力的迭代视觉推理（H-GIVR）**框架。该框架旨在赋予模型类似人类的验证能力，使其能够在推理过程中反复获取视觉信息，并将历史答案融入后续迭代，从而通过迭代一致性得出最终结果。

图1 展示了 H-GIVR 框架回答一个视觉问答问题的详细过程。具体而言，在视觉问答任务中，H-GIVR 首先引导语言模型检查图像并生成相应的图像描述。该描述会追加到问题文本中，作为后续推理的视觉证据。在迭代推理过程中，上一轮生成的答案会与问题一同作为

[*]贡献相同 [1]南京航空航天大学。通讯作者：Haochen Shi<hshi03@nuaa.edu.cn>，Sheng-Jun Huang<huangsj@nuaa.edu.cn>。

预印本。*2026*年*2*月*5*日。

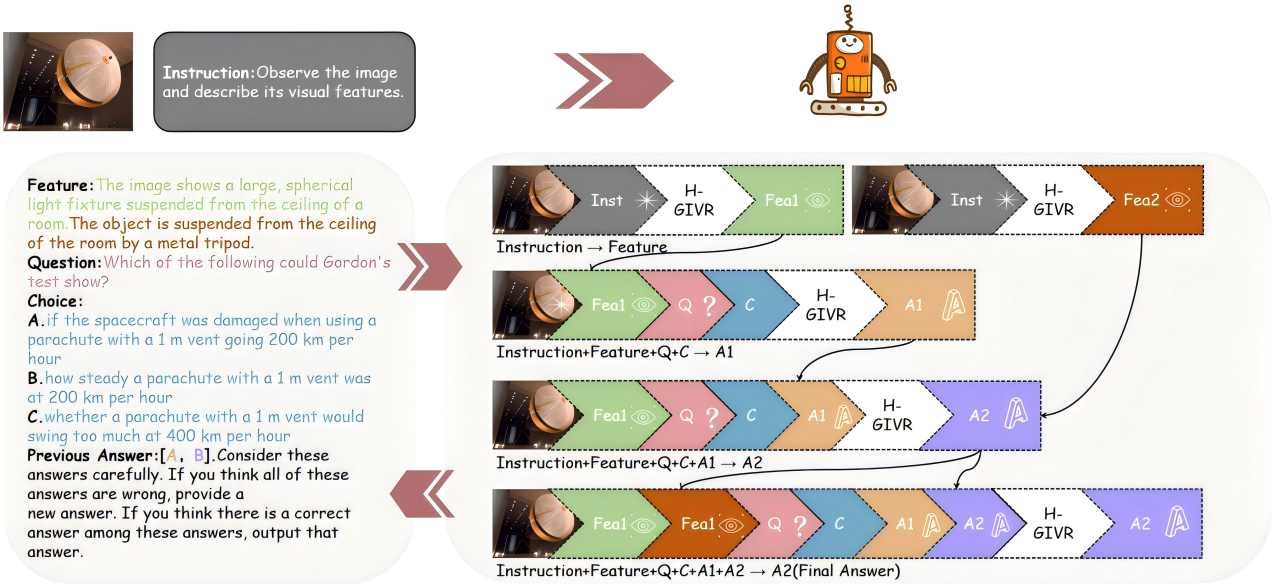


图*1*. H-GIVR 框架回答视觉问答问题的详细过程。

历史上下文输入，模拟人类反复检查和修正答案的习惯。

当迭代索引为偶数时，该框架假定模型尚未对视觉内容形成充分理解。随后，它会触发第二次视觉观察步骤，引导模型重新生成图像描述，以加深并强化其对视觉信息的理解。一旦模型在迭代过程中给出一致的答案，推理即被视为已达到稳定状态，并将该答案作为最终输出。

综上所述，我们的贡献如下：

1. 我们提出了一种新颖的框架 H-GIVR，它超越了基于“重复作答与多数投票”的传统自一致性方法。2. H-GIVR 框架有效利用历史信息与视觉信息，以模拟人类的解题行为。3. H-GIVR 使用两种验证机制，以促使模型反复进行图像检查，并判断模型何时已得出稳定的最终答案。

2. 相关工作

2.1. 大语言模型的自我纠正

赋予大语言模型（LLMs）自我纠错能力，是提升其可靠性的关键途径。根据干预机制，当前方法主要分为两类：参数调整和推理阶段的交互式引导（Pan 等，2024）。参数调整方法，例如基于 SCoRe 框架的强化学习方法（Kumar 等，2024），会在训练阶段通过算法修改模型参数。这些方法使用模型自身生成的数据进行训练，有效解决了传统监督微调中常见的分布偏移和行为崩溃问题。然而，这些方法涉及复杂流程，且计算成本高昂。

相比之下，RISE 等推理时引导方法（Qu 等人，2024）和 S^2R（Ma 等人，2025a）旨在通过设计提示词，激发模型内在的反思潜力。研究表明，即使是简单的提示词，也能触发预训练模型明确的反思行为（Shah 等人，2025）。我们的工作沿着第二条研究路径展开并对其进行拓展，重点探索如何通过多模态情境中的结构化交互，实现有效的推理时自我纠正。

2.2. 自一致性与反思

单轮纠正往往不足以修复复杂错误。因此，多轮迭代式自我改进机制便成为自然的延伸。现有研究提出的 RISE 框架（Qu 等人，2024）使模型能够递归地反思并修订此前各轮的输出。此外，探索式迭代（Jiang 等人，2025）强调，迭代过程中保持发散思维对于避免陷入局部最优至关重要。这些研究的共同核心是追求自洽性。相关研究表明，对于

模型要实现有效的自我改进，需要展现验证和反馈等关键行为（Gandhi 等人，2025）。这揭示了一项重要原则：有效的自我纠正并非简单地“重新作答”，而是一种涉及错误检测、假设回溯和逻辑重构的认知过程。我们的工作设计了一种多轮反馈结构，系统地引导模型完成这一认知过程。通过反复修正答案，并以图像描述作为回溯和验证的锚点，我们增强了跨模态推理的自洽性。

2.3. 多模态模型的认知局限

尽管多模态模型已在众多任务中展现出出色的性能，但其底层认知理解能力仍然令人质疑。近期研究开始系统评估大型多模态模型（MLLM）的基本缺陷。现有研究表明，MLLM 通常缺乏物体恒存性和基本物理直觉等核心知识，而这一缺陷无法仅通过扩大模型规模来解决（Li 等人，2024b）。由于模型存在固有的认知“盲点”，为其设计外部的交互式“反思—纠正”机制，成为弥补此类缺陷并提升推理可靠性的有效方法。在模型适配领域，采用反思来分析和改进模型性能已成为一种趋势（Ming&Li，2024）。

2.4. 交互与对齐

“如何实现准确的跨模态对齐”是视觉语言任务的核心挑战。现有研究大多聚焦于训练阶段的模态融合与表征对齐。目前，将自我纠正机制拓展至视觉语言多模态场景的研究仍处于早期阶段，且现有工作大多沿着参数调整的路径展开。例如，一些研究通过微调提升视觉定位的自洽性（He 等人，2024；Sun 等人，2025），另一些研究则增强模型抵御扰动的鲁棒性（Ma 等人，2025b）。这些研究表明，在多模态情境中追求一致性是一条可行且重要的研究方向。

然而，这些方法并未赋予模型在推理阶段动态评估并纠正自身跨模态理解的能力。关于无需复杂训练、完全通过推理时交互实现的多模态自我纠正框架，相关研究仍然稀少。我们的工作补充了上述研究：我们并不试图一次性训练出一个“完美”的跨模态对齐模型，而是承认现有模型存在内在的对齐误差。随后，我们设计了一种轻量级的推理时交互框架，使模型能够独立反思并纠正这些误差。

3. 方法

图2 展示了 H-GIVR 的整体工作流程。该框架包含两个核心组成部分：**（i）视觉描述**，在整个作答过程中持续生成图像描述；以及**（ii）一致性迭代推理**，即将模型上一轮的答案附加到下一轮的问题中，作为进一步推理的参考。此外，H-GIVR 还包含两种关键判断机制：**（i）图像重新观察机制**，当迭代次数为偶数时启动，提示模型重新生成图像描述，以减少记忆丢失对答案正确性的影响；以及**（ii）答案确认机制**，当推理过程中出现两个相同的答案时，停止迭代并将结果作为最终答案输出。

3.1. 问题定义

视觉问答（VQA）数据集表示为 $D = \{(x_i, v_i, y_i)\}_{i=1}^{N}$，其中 x_i 表示问题，v_i 表示相关图像，y_i 表示真实标注，N 表示测试样本总数。对于给定的输入 x_i 和 v_i，模型生成的图像描述定义为 $f_k(v_i)$，模型生成的答案定义为 $p_k(x_i, v_i)$。经过多次迭代后，这些生成结果最终整理为图像描述集 $F(v_i)$ 和答案集 $P(x_i, v_i)$：

$$\begin{aligned} F(v_i) &= \{f_1(v_i), f_2(v_i), ..., f_n(v_i)\}, \\ P(x_i, v_i) &= \{p_1(x_i, v_i), p_2(x_i, v_i), ..., p_m(x_i, v_i)\}. \end{aligned} \tag{1}$$

如果存在两个不同的索引 α 和 β ($1 \leq \alpha, \beta \leq m$ & $\alpha \neq \beta$)，使得 $p_\alpha(x_i, v_i) = p_\beta(x_i, v_i)$，则将该答案作为最终输出。

3.2. 视觉描述

在视觉问答任务中，图像所承载的视觉信息是生成正确答案的关键依据。如果模型未能充分理解图像内容，其推理过程可能会逐渐偏离正确方向，从而导致答案错误。为解决这一问题，我们提出了一种分阶段处理方法，在推理之前分别处理图像 v_i 和问题x_i。首先，引导模型聚焦于图像 v_i 并生成相应的视觉描述$f(v_i)$。随后，在答案生成阶段，模型通过整合图像 v_i 和描述 $f(v_i)$进行联合推理。这一机制旨在增强

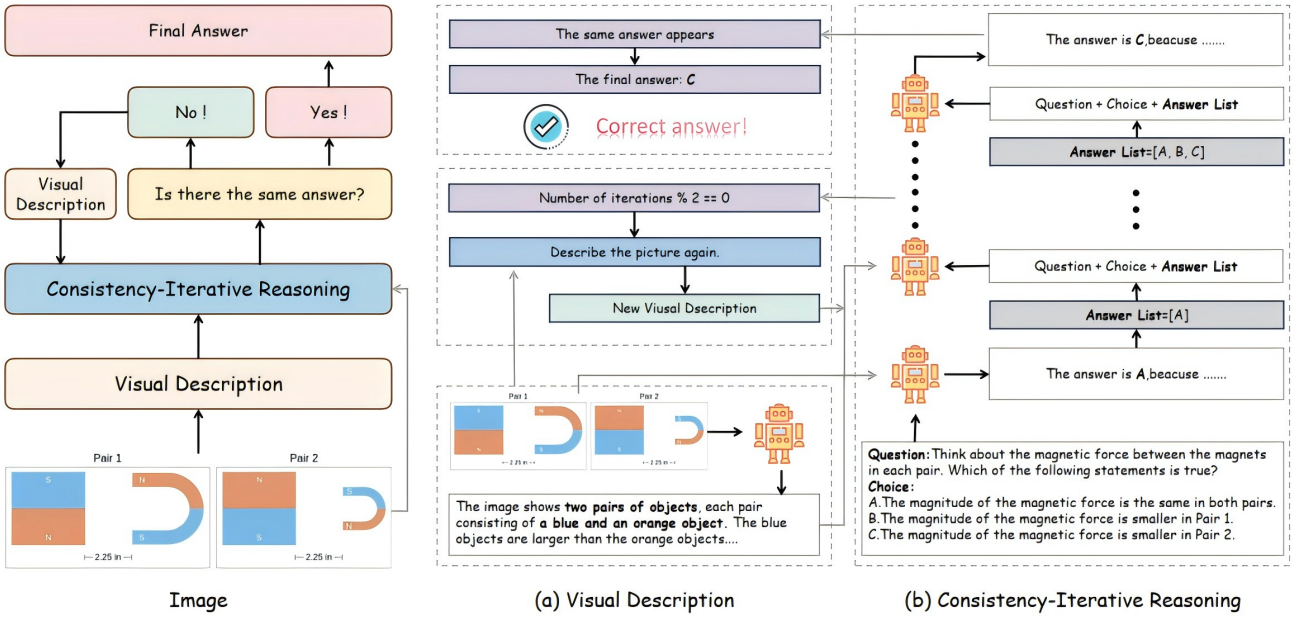


图 2. H-GIVR 框架示意图。 **视觉描述** 和 **一致性迭代推理** 是两个核心组成部分，旨在加深模型对图像的理解，并模拟人类在作答过程中的复核行为：图像重新观察机制和答案确认机制是两个关键决策机制，旨在减少模型对视觉信息的误解并确定最终答案。

模型对视觉信息的感知和运用能力，从而加深其对图像内容的理解，并提高答案生成的准确性。

3.3. 图像重新观察机制

在视觉问答任务中，人们通常会持续观察图像，反复查看以捕捉可能遗漏的细节，并纠正理解错误。这种动态观察过程有助于在推理过程中逐步加深对视觉信息的把握。然而，由于上下文长度等方面的限制，现有模型难以在单次推理过程中多次获取图像信息，这限制了它们回看图像以进行自我纠正的能力。

为弥补这一不足，我们在模型的迭代推理过程中引入了一种结构化的图像重新观察机制。具体而言，当迭代次数达到偶数时，模型会独立地重新处理原始图像，并生成更新后的视觉描述。这一设计旨在模拟人类在解题过程中定期回看图像以加深理解的行为。通过定期强化视觉信息，模型能够在多个推理步骤中持续留意图像内容中的关键信息，减少因记忆丢失或初始理解不足而导致的错误，从而提升视觉推理的鲁棒性和准确性。

3.4. 一致性迭代推理

在检查某个问题的答案时，人类并非只是机械地重复之前的回答，而是会带着明确的意图和目的审视已有答案。为模拟这一认知过程，我们不再局限于基于重复采样的传统自一致性方法，而是提出一种迭代一致性推理组件。具体而言，H-GIVR 框架会动态更新答案列表，并将其作为上下文的一部分附加到问题文本中。每轮推理结束后，都会将新生成的答案添加到该列表中：

$$Answer\ List = \{a_1, a_2, ..., a_n\} \tag{2}$$

其中，a_i 表示历史答案，n 表示当前迭代步骤。随后，指示模型在后续推理过程中仔细考虑列表中的所有历史答案。这一设计明确赋予模型反思并验证自身输出的能力，使其更接近人类解决问题时深思熟虑地检查答案的行为。

同时，如表1所示，在后续推理中将错误答案作为参考，不仅不会误导模型，还能促使其逐渐收敛到正确答案。**!!!! 在实际应用中，由于通常无法预先判断答案是否错误，这种设置主要用于方法论验证，并非直接面向实际部署。**

表1. **错误的历史答案是否会误导模型后续推理**。标准表示直接调用模型，False 表示使用错误答案作为后续推理的参考，而 H-GIVR 指我们提出的框架。

方法	标准	False	H-GIVR
准确率	38.08	83.33	78.90

3.5. 答案确认机制

在迭代过程中，当出现两个相同的答案时，H-GIVR 会将该答案视为最终输出。此外，实验表明，模型在处理较难的问题时可能会生成“None”。为模拟考试环境中的时间限制并提高计算效率，我们将最大迭代次数设为 10。根据实验统计，H-GIVR 框架所需的平均迭代次数最多约为 4，远低于预设的 10 次上限。因此，将最大迭代次数设为 10，不会对框架的实际效率或性能产生明显影响。此设置为迭代推理提供了充足的空间，同时避免不必要的计算开销。此外，如果模型连续生成若干个“None”，随后首次给出有效答案，则直接将该答案作为最终结果。H-GIVR 框架的伪代码见算法1。

算法1 H-GIVR

输入： 问题 x_i，图像 v_i
初始化： $AnswerList \leftarrow \emptyset$, $FeatureSet \leftarrow \emptyset$, $iteration \leftarrow 0$, $max_iterations \leftarrow 10$
视觉描述： $feature \leftarrow$ 多模态大语言模型(v_i)
$FeatureSet \leftarrow FeatureSet \cup \{feature\}$
当 $iteration < max\ iterations$**时执行**
  $iteration \leftarrow iteration + 1$
  $prompt \leftarrow x_i + FeatureSet$
  如果 $AnswerList \neq \emptyset$，**则**
    $prompt \leftarrow prompt + AnswerList$
  结束如果
  $current\ answer \leftarrow$多模态大语言模型($prompt$)
  如果 $current\ answer \in AnswerList$**则**
    $A_{final} \leftarrow current_answer$
    中断
  结束如果
  $AnswerList \leftarrow AnswerList \cup \{current_answer\}$
  如果 $iteration$ 取模 $2 = 0$ **则**
    $new\ feature \leftarrow$ 多模态大语言模型(v_i)
    $FeatureSet \leftarrow FeatureSet \cup \{new_feature\}$
  结束如果
结束循环
如果 $iteration \geq max\ iterations$**则**
  $A_{final} \leftarrow AnswerList[-1]$
  返回最新答案
结束条件判断
返回： A_{final}

4. 实验

在本节中，我们进行了广泛的实验，以评估 H-GIVR 的有效性和计算效率。我们的实验主要围绕以下关键问题展开：

问题一。 H-GIVR 在多种场景下的表现如何?

问题二。 H-GIVR 是否需要大量的计算资源?

问题3. 与现有基线方法相比，H-GIVR 能否提高准确率?

问题4. 视觉描述和一致性迭代推理是否都能有效提高准确率?

问题5. H-GIVR 在特定领域中是否仍然有效?

4.1. 数据集与指标

遵循多模态大语言模型推理研究中的标准评估设置，我们在五个常用基准测试上开展实验：ScienceQA (Lu 等人，2022)、A-OKVQA (Schwenk 等人，2022)、OK-VQA (Marino 等人，2019)、VQAv2 (Goyal 等人，2017) 和 TextVQA (Singh 等人，2019)。在这些数据集中，ScienceQA 和 A-OKVQA 是多项选择数据集，而 OK-VQA、VQAv2 和 TextVQA 是开放式答案数据集。我们遵循相关研究的做法 (Liu 等人，2024；Li 等人，2024a；Bai 等人，2023)，采用准确率作为评估指标。我们使用 Ollama 上下载量均超过一百万次的三个模型开展实验：`Llama3.2-vision:11b`、`Qwen2.5vl:7b`和 `Gemma3:12b`。

4.2. 实验结果

针对问题 1：H-GIVR 在多种场景下的表现如何? 我们在多模态大语言模型的三种典型应用设置下开展实验：标准设置，即直接调用模型；简单思维链，使用“让我们一步一步地思考”这一简单提示词；以及复杂思维链，采用少样本提示。实验结果如表 2所示。

结果表明，H-GIVR 在全部五个数据集、三种模型和三种应用设置中均取得了显著的性能提升，在多项选择数据集上的提升尤为明显。对于 Llama3.2-vision,

表2. 在五个数据集、三种模型和三种应用设置下的总体结果（%）。对于每种设置，最佳结果以粗体突出显示，提升幅度以粉色显示。 ✓ 符号表示使用了 H-GIVR，而 × 符号表示未使用。

模型	提示词	H-GIVR	数据集					平均值
			ScienceQA	A-OKVQA	OK-VQA	VQAv2	TextVQA	
Llama3.2-vision:11b	标准	×	38.08	38.34	28.74	56.94	46.16	41.65
		✓	**78.90**	**76.42**	**48.62**	**64.99**	**57.48**	**65.28**
			↑ 40.82	↑ 38.08	↑ 19.88	↑ 8.05	↑ 11.32	↑ 23.63
	简单思维链	×	40.23	38.42	28.86	60.46	50.16	43.63
		✓	**75.10**	**78.52**	**47.99**	**65.26**	**59.12**	**65.20**
			↑ 34.87	↑ 40.10	↑ 19.13	↑ 4.80	↑ 8.96	↑ 21.57
	复杂思维链	×	64.47	40.34	40.11	62.78	54.56	52.45
		✓	**78.14**	**68.56**	**42.49**	**64.56**	**58.44**	**62.44**
			↑ 13.67	↑ 28.22	↑ 2.38	↑ 1.78	↑ 3.88	↑ 9.99
Qwen2.5vl:7b	标准	×	40.16	42.97	23.50	58.52	35.10	40.05
		✓	**85.12**	**81.83**	**55.43**	**73.22**	**68.18**	**72.76**
			↑ 44.96	↑ 38.86	↑ 31.93	↑ 14.70	↑ 33.08	↑ 32.71
	简单思维链	×	40.75	42.70	31.20	38.31	50.32	40.66
		✓	**85.03**	**81.57**	**55.35**	**73.19**	**65.96**	**72.22**
			↑ 44.28	↑ 38.87	↑ 24.15	↑ 34.88	↑ 15.64	↑ 31.56
	复杂思维链	×	73.83	39.56	40.24	58.96	61.70	54.86
		✓	**84.93**	**78.08**	**52.20**	**70.25**	**66.08**	**70.31**
			↑ 11.10	↑ 38.52	↑ 11.96	↑ 11.29	↑ 4.38	↑ 15.45
Gemma3:12b	标准	×	40.16	41.57	33.74	29.19	**56.38**	40.21
		✓	**86.14**	**83.93**	**50.09**	**65.89**	**55.76**	**68.36**
			↑ 45.98	↑ 42.36	↑ 16.35	↑ 36.70	↓ 0.62	↑ 28.15
	简单思维链	×	40.75	44.19	35.00	38.17	54.34	42.49
		✓	**86.39**	**83.49**	**51.31**	**66.27**	**55.72**	**68.64**
			↑ 45.64	↑ 39.30	↑ 16.31	↑ 28.10	↑ 1.38	↑ 26.15
	复杂思维链	×	61.90	41.92	45.20	59.03	41.88	49.99
		✓	**84.58**	**81.92**	**50.56**	**66.37**	**55.22**	**67.73**
			↑ 22.69	↑ 40.00	↑ 5.36	↑ 7.34	↑ 13.34	↑ 17.74

通义千问2.5vl 和 Gemma3 模型，在标准和简单思维链设置下，性能提升均超过 50%；在复杂思维链设置下，最低提升也达到 19%。在更具挑战性的开放式数据集上，H-GIVR 同样带来了显著收益，提升幅度为 2.8% 至 69.2%。

此外，结果表明，即使向 H-GIVR 提供额外的提示词信息，无论这些信息简单与否或经过精心设计，该框架仍然有效，但其准确率低于不提供提示词时的表现。这表明，H-GIVR 无需任何精心编写的提示词即可直接使用，从而有助于降低多模态大语言模型的使用门槛。

对于问题二：H-GIVR 是否需要大量的计算资源？ 我们测量了平均数量

H-GIVR 在三个模型、五个数据集和三种应用设置下处理每个问题所需的模型调用次数，如图3所示。结果表明，H-GIVR 每个问题需要进行 2.57 至 4.04 次调用，这说明 H-GIVR 在计算上高效，并且具有良好的成本控制能力。

我们还观察到，与多项选择题相比，开放式问题通常需要更多模型调用。这表明，这类问题的决策不确定性更高，模型也更难收敛到明确且稳定的答案。由于开放式任务允许更广泛的有效回答，模型往往需要进行额外推理，以完善输出并得出一致的结论。此外，与不使用提示的标准设置相比，使用提示的简单思维链和复杂思维链设置通常需要

表3。 两个模型在五个数据集和六种方法上的总体结果（%）。最佳结果以粗体突出显示，且 $\Delta Gains$ 行报告了 H-GIVR 相较于最佳基线准确率的提升。

模型	提示词	数据集					平均值
		ScienceQA	A-OKVQA	OK-VQA	VQAv2	TextVQA	
Llama3.2-vision:11b	标准	38.08	38.34	28.74	56.94	46.16	41.65
	FS-CoT	64.47	40.34	40.11	62.78	54.56	52.45
	Auto-CoT	39.07	59.82	48.13	63.57	54.40	53.00
	Active-Pro	40.44	42.62	42.62	60.12	56.04	48.37
	Self-Con	50.00	54.41	43.56	**66.04**	54.85	53.77
	Ours	**78.90**	**76.42**	**48.62**	64.99	**57.48**	**65.28**
	$\Delta Gains$	↑ 14.43	↑ 16.60	↑ 0.49	↓ 1.05	↑ 1.44	↑ 11.51
Qwen2.5vl:7b	标准	40.16	42.97	23.50	58.52	35.10	40.05
	FS-CoT	73.83	39.56	40.24	58.96	61.70	54.86
	Auto-CoT	74.09	69.78	39.46	66.79	63.64	62.75
	Active-Pro	77.79	71.26	43.94	57.77	64.10	62.97
	Self-Con	74.70	64.45	54.10	40.43	43.82	55.50
	Ours	**85.12**	**81.83**	**55.43**	**73.22**	**68.18**	**72.76**
	$\Delta Gains$	↑ 7.33	↑ 10.57	↑ 1.33	↑ 6.43	↑ 4.08	↑ 9.79

表4. 在五个基准测试上的消融研究结果（%）。**视觉**表示视觉描述模块，**迭代**表示一致性迭代推理模块。符号表示使用了相应模块，粉色表示与标准设置相比的提升。

	模块		数据集					Avg.
	视觉	迭代	ScienceQA	A-OKVQA	OK-VQA	VQAv2	TextVQA	
			Qwen2.5vl:7b					
(a)	标准		40.16	42.97	23.50	58.52	35.10	40.05
(b)	✔		82.79 ↑ 42.63	80.96 ↑ 37.99	53.36 ↑ 29.86	72.36 ↑ 13.84	65.20 ↑ 30.10	70.93 ↑ 30.88
(c)		✔	83.28 ↑ 43.12	81.79 ↑ 38.82	54.92 ↑ 31.42	72.45 ↑ 13.93	68.14 ↑ 33.04	72.11 ↑ 32.06
(d)	✔	✔	**85.12** ↑ 44.96	**81.83** ↑ 38.86	**55.43** ↑ 31.93	**73.22** ↑ 14.70	**68.18** ↑ 33.08	**72.76** ↑ 32.71

更少的调用，但准确率往往低于标准设置。这表明，引入显式提示后，模型可能更容易固守先前生成的错误答案，而不是继续深入推理并进行纠正。

对于问题3：与现有基线方法相比，H-GIVR 能否提高准确率？ 我们选取了几种经典的单模态大语言模型提示方法作为基线，包括少样本思维链提示词（FS CoT）、Auto-CoT（Zhang 等人，2022）、主动提示（Diao 等人，2024）和自洽性（Wang 等人，2022）。如表3所示，在五个数据集上，

与其他基线方法相比，Llama3.2-vision 和 Qwen2.5vl 模型在大多数情况下使用 H-GIVR 时表现更好。性能提升幅度为 0.49 至 16.60。

尽管自洽性在少数情况下表现略好（例如，Llama3.2-vision 在 VQAv2 数据集上），但其模型调用次数固定为 5 次，而 H-GIVR 平均仅需调用 3.58 次。因此，H-GIVR 在保持竞争力的同时，在计算效率方面具有明显优势。

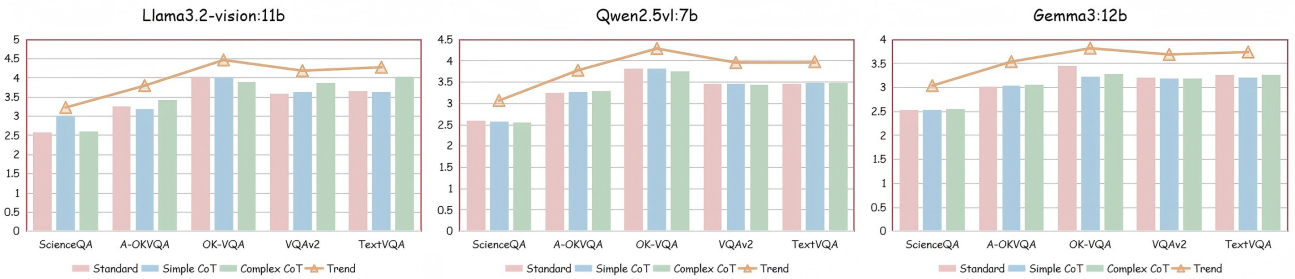


图*3.* H-GIVR 在三个模型、五个数据集和三种应用设置下处理每个问题所需的平均模型调用次数。柱状图显示了平均模型调用次数，折线图则展示了平均调用次数在五个数据集之间的变化情况。

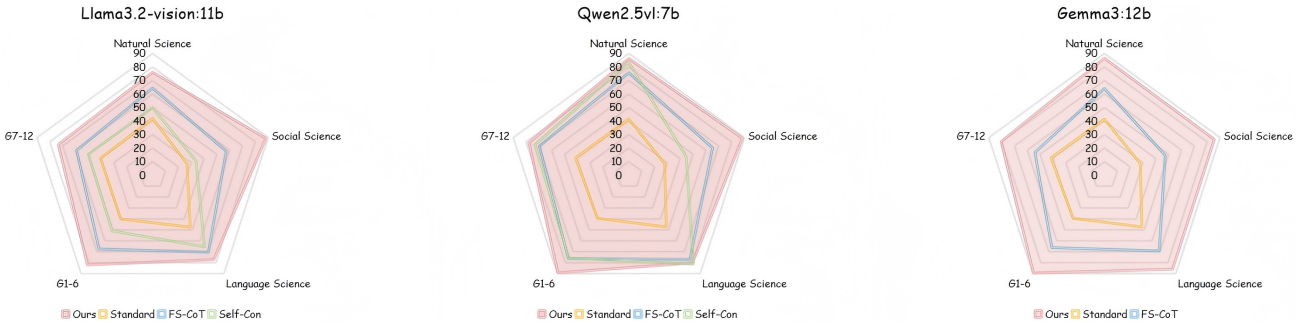


Figure 4. H-GIVR 与其他基线方法在细粒度领域中的性能差异

ns.

关于 Q4：视觉描述和一致性迭代推理都能有效提升准确率吗？为验证框架中两个核心组件的有效性，我们开展了系统的消融实验，结果如图所示4。实验采用标准设置下的 Qwen2.5vl 7B 模型，该模型在五个数据集上的平均准确率为 40.05%。仅引入视觉描述模块时，平均表现显著提升至 70.93%。仅使用一致性迭代推理模块时，平均准确率进一步提升至 72.11%。最后，同时包含这两个模块的完整 H-GIVR 取得了最佳表现，平均准确率达到 72.76%，并在所有数据集上都实现了稳定且明显的提升。

这些结果证实了视觉描述模块和一致性迭代推理模块的有效性。结果还表明，这两个组件具有互补性，将它们结合起来能够共同提升多模态大语言模型的推理能力和答案准确率。

针对Q5：H-GIVR在特定领域中仍然有效吗？为了更系统地评估 H-GIVR 在不同学科类型和难度级别上的表现，我们在 ScienceQA 数据集上进行了细粒度分析。该数据集可沿两个维度划分。就学科类别而言，它涵盖三大领域：语言科学、社会科学和自然科学。就问题难度而言，它被分为两个级别，即 G1–6 和 G7–12。实验结果如图 4所示。

结果表明，H-GIVR在大多数子领域中表现良好。只有在语言科学类别中，它在少数情况下略逊于自一致性方法。在其他学科领域，尤其是社会科学领域，H-GIVR显著提高了答案准确率，展现出较强的学科适应性。此外，H-GIVR在不同难度级别下均保持稳定的性能优势。这些发现表明，H-GIVR适用于不同的学科背景，并具有良好的泛化能力。

5. 结论

H-GIVR 创新性地突破了“重复作答与多数投票”的传统自一致性范式。它并非机械地重复推理，而是将历史推理信息作为后续决策的关键参考。与此同时，它引导模型在推理过程中反复观察视觉内容，从而精确纠正视觉理解错误。该框架在各项视觉问答基准上显著提升了推理准确率，同时将计算成本保持在可控且相对较低的水平。

影响声明

本文提出了历史引导的迭代视觉推理与自我纠正（H-GIVR）框架，其目标是推动多模态大语言模型推理领域的发展。通过降低对精细提示工程和大规模模型扩展的依赖，H-GIVR 降低了多模态系统的部署门槛，使先进的跨模态推理更易于非专业用户使用。在伦理方面，该框架的轻量化设计将计算成本降至最低，从而减少运行大规模多模态模型所带来的碳足迹。就社会影响而言，H-GIVR 有望在教育、无障碍工具等现实应用中提升视觉问答系统的可靠性。目前尚无可以预见的负面伦理后果，且这项工作与开发高效、用户友好的人工智能技术这一更广泛目标相一致。

参考文献

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., 等。Flamingo：一种用于少样本学习的视觉语言模型。《神经信息处理系统进展》，35:23716–23736，2022。Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., 等。通义千问技术报告。*arXiv*预印本 *arXiv:2309.16609*，2023。Cai, Z., Duan, X., Haslett, D., Wang, S., 和 Pickering, M. 大语言模型在语言使用方面是否与人类相似？载于认知建模与计算语言学研讨会论文集，第37–56页，2024。Diao, S., Wang, P., Lin, Y., Pan, R., Liu, X., 和 Zhang, T. 面向大语言模型的思维链主动提示。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第1330–1350页，2024。Fu, Y., Peng, H., Sabharwal, A., Clark, P., 和 Khot, T. 面向多步推理的复杂度提示。*arXiv*预印本 *arXiv:2210.00720*，2022。Gandhi, K., Chakravarthy, A., Singh, A., Lile, N., 和 Goodman, N. D. 使推理者能够自我提升的认知行为，或高效明星的四个习惯。*arXiv*预印本 *arXiv:2503.01307*，2025。Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., 和 Parikh, D. 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。载于*IEEE*计算机视觉与模式识别会议论文集

*p*模式识别 *g*，pp。6904–6913，2017。

He, R., Cascante-Bonilla, P., Yang, Z., Berg, A. C. 和 Ordonez, V. 通过自洽解释改进视觉定位。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第13095–13105 页，2024 年。Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., 等。训练计算最优的大语言模型。*arXiv*预印本 *arXiv:2203.15556*，2022 年。Jiang, M., Lupu, A. 和 Bachrach, Y. 为自我改进引导构建任务空间。*arXiv*预印本 *arXiv:2509.04575*，2025 年。Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. 和 Amodei, D. 神经语言模型的缩放定律。*arXiv*预印本 *arXiv:2001.08361*，2020 年。Kumar, A., Zhuang, V., Agarwal, R., Su, Y., Co-Reyes, J. D., Singh, A., Baumli, K., Iqbal, S., Bishop, C., Roelofs, R., Zhang, L. M., McKinney, K., Shrivastava, D., Paduraru, C., Tucker, G., Precup, D., Behbahani, F. 和 Faust, A. 通过强化学习训练语言模型进行自我纠正。*arXiv*预印本 *arXiv:2409.12917*，2024 年。Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., 等。Llava-onevision：轻松实现视觉任务迁移。*arXiv*预印本 *arXiv:2408.03326*，2024a。Li, J., Li, D., Xiong, C. 和 Hoi, S. BLIP：通过引导式语言-图像预训练，实现统一的视觉语言理解与生成。载于国际机器学习会议，第 12888–12900 页。PMLR，2022 年。Li, Y., Gao, Q., Zhao, T., Wang, B., Sun, H., Lyu, H., Hawkins, R. D., Vasconcelos, N., Golan, T., Luo, D., 等。多模态语言模型的核心知识缺陷。*arXiv*预印本 *arXiv:2410.10855*，2024b。Liu, H., Li, C., Wu, Q. 和 Lee, Y. J. 视觉指令微调。《神经信息处理系统进展》，36:34892–34916，2023 年。Liu, H., Li, C., Li, Y. 和 Lee, Y. J. 通过视觉指令微调改进基线。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第26296–26306 页，2024 年。Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P. 和 Kalyan, A. 学会解释：通过思维链进行多模态推理，以解答科学问题。《神经信息处理系统进展》，35:2507–2521，2022 年。

Ma, R., Wang, P., Liu, C., Liu, X., Chen, J., Zhang, B., Zhou, X., Du, N. 和 Li, J. S^2r：通过强化学习教大语言模型进行自我验证和自我纠正。*arXiv*预印本 *arXiv:2502.12853*，2025a。Ma, X., Xu, L., Fang, L., Zhang, C. 和 Cui, L. 通过原型迭代构建实现可靠的跨模态对齐。载于第*33*届*ACM*国际多媒体会议论文集，第 3847–3855 页，2025b。Marino, K., Rastegari, M., Farhadi, A. 和 Mottaghi, R. OK-VQA：一项需要外部知识的视觉问答基准。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 3195–3204 页，2019 年。Ming, Y. 和 Li, Y. 理解面向视觉语言模型的检索增强任务适应。*arXiv*预印本 *arXiv:2405.01468*，2024 年。Pan, L., Saxon, M., Xu, W., Nathani, D., Wang, X. 和 Wang, W. Y. 自动纠正大语言模型：多样化自动纠正策略概览。《计算语言学协会会刊》，12:484–506，2024 年。Qu, Y., Zhang, T., Garg, N. 和 Kumar, A. 递归内省：教会语言模型智能体如何自我改进。《神经信息处理系统进展》，37:55249–55285，2024 年。Schwenk, D., Khandelwal, A., Clark, C., Marino, K. 和 Mottaghi, R. A-OKVQA：一项运用世界知识的视觉问答基准。载于欧洲计算机视觉会议，第 146–162 页。Springer，2022 年。Shah, D. J., Rushton, P., Singla, S., Parmar, M., Smith, K., Vanjani, Y., Vaswani, A., Chaluvaraju, A., Hojel, A., Ma, A., 等。重新思考预训练中的反思。*arXiv*预印本 *arXiv:2504.04022*，2025 年。Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D. 和 Rohrbach, M. 朝着能够阅读的视觉问答模型迈进。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 8317–8326 页，2019 年。Sun, J., Chen, K., He, X., Liu, X., Li, K. 和 Peng, C. Uni-trans：面向大型多模态基础模型的统一参数高效迁移学习与多模态对齐。《计算机、材料与持续发展》，83(1)，2025 年。Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A. 和 Zhou, D. 自洽性可提升语言模型中的思维链推理。*arXiv*预印本 *arXiv:2203.11171*，2022 年。

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., 等。思维链提示能够激发大语言模型的推理能力。《神经信息处理系统进展》，35:24824–24837，2022。

Zhang, Z.、Zhang, A.、Li, M. 和 Smola, A. 大语言模型中的自动思维链提示。2022。

A. 交互案例

A.1. 针对 ScienceQA 的多项选择

表5 一个由 Llama3.2-vision:11b 作答的 ScienceQA 示例，未使用任何思维链提示。粉色表示多模态大语言模型的答案，蓝色突出了以历史信息作为参考的推理逻辑，绿色表示新增的视觉描述。

图像： 

指令： 观察图像并描述其视觉特征。

多模态大语言模型： 图中，一只雄性北美红雀栖息在一个由链条悬挂的木制喂鸟器上。这只鸟鲜艳的红色羽毛与眼睛周围的黑色面罩和长长的红色尾巴相映成趣。它停在喂鸟器边缘，喂鸟器里装满了黑色葵花籽。背景呈绿色且模糊，说明这张照片是在户外的自然环境中拍摄的。

图像： 

问题： 哪一个问题最适合通过 Chase 的实验来回答?

选项：
主红雀吃更多的种子吗 取食时会吃下更多种子吗?
主红雀会来喂鸟器吗 ，而不是装有亚麻籽的喂鸟器?

图像特征：
图像中，一只雄性北美红雀栖息在一个用链条悬挂的木制喂鸟器上。这只鸟鲜艳的红色羽毛与眼睛周围的黑色面罩及一条长长的红色尾巴相映衬。它栖息在喂鸟器边缘，喂鸟器里装满了黑色葵花籽。背景呈绿色且模糊，表明这张照片是在户外的自然环境中拍摄的。

指令： 请严格按“答案是 ”的格式输出答案。**多模态大语言模型：** 答案是 A。

图像： 

问题： 指出 Chase 的实验最适合回答哪个问题。

选项：
红衣主教鸟从装有葵花籽的喂鸟器每次取食时，吃的种子是否比从装有亚麻籽的喂鸟器取食时更多？红衣主教鸟是否比起装有亚麻籽的喂鸟器，更常造访装有葵花籽的喂鸟器?

图像特征：
图像中，一只雄性北美红雀栖息在一个用链条悬挂的木制喂鸟器上。这只鸟鲜艳的红色羽毛与眼睛周围的黑色面罩及一条长长的红色尾巴相映衬。它栖息在喂鸟器边缘，喂鸟器里装满了黑色葵花籽。背景一片绿色且模糊，表明这张照片是在户外的自然环境中拍摄的。

先前的答案： [A]。请仔细考虑这些答案。如果你认为这些答案都不正确，请给出一个新答案。如果你认为这些答案中有正确答案，请输出该答案。

指令： 严格按照“The answer is ”这一格式输出答案。

多模态大语言模型： 答案是 B。

图像： 

指令： 再次观察图像并描述其视觉特征。**多模态大语言模型：** 图像中有一只红衣主教鸟栖息在悬挂于链条上的喂鸟器上。这只鸟的羽毛呈鲜艳的红色，带有黑色的颈部斑块和鲜红色的喙。它面朝右侧，尾羽朝下。喂鸟器本身由木材制成，内含葵花籽，并配有用于悬挂的金属链条和挂钩。周围环境处于失焦状态，但看起来是带有绿色植被的自然环境。

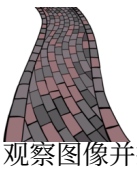
图像： 

问题： 确定 Chase 的实验最能回答哪个问题。
选项：
主红雀从装有葵花籽的喂食器中每次访问吃掉的种子是否比从装有亚麻籽的喂食器中更多？主红雀是否比访问装有亚麻籽的喂食器更频繁地访问装有葵花籽的喂食器？
图像特征：
图像中，一只雄性北美主红雀栖息在一个由金属链条悬挂的木制喂鸟器上。它鲜艳的红色羽毛与眼周的黑色面罩和长长的红色尾巴相映成趣。它停在喂鸟器边缘，喂鸟器里装满了黑色葵花籽。背景呈绿色且模糊，表明照片是在户外自然环境中拍摄的。

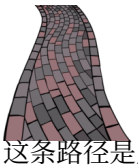
图像中，一只红衣主教鸟栖息在由链条悬挂的喂鸟器上。它的羽毛呈鲜艳的红色，颈部有一块黑色斑块，喙呈鲜红色。它面朝右侧，尾羽向下垂。喂鸟器本身由木材制成，里面装有葵花籽，并用金属链条和挂钩悬挂。周围环境虽未对焦，但看起来是绿意盎然的自然环境。

之前的答案： **[A、B]。请仔细考虑这些答案。如果你认为这些答案全都不正确，请给出一个新答案。如果你认为其中有一个正确答案，请输出该答案。** **指令：** 请严格按照“The answer is ”的格式输出答案。**多模态大语言模型：** 答案是B。

表 6. 一个由 Llama3.2-vision:11b 作答的 ScienceQA 示例，未使用任何思维链提示。 粉色 表示多模态大语言模型的答案，蓝色 突出显示了将历史信息用作参考的推理逻辑。

图像： 

指令： 观察图像并描述其视觉特征。
多模态大语言模型： 所提供的图像似乎是一个三维结构的二维表示，可能是一条由许多小块或砖块构成的道路或路径。这些块似乎沿着一条弯曲的路径排列，从图像顶部蜿蜒延伸至底部。它们排列成某种图案，有些块似乎嵌入其中或以一定角度放置，另一些则似乎与其他块呈90度角。

图像： 

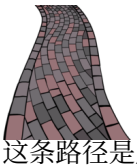
问题： 这条路径是用什么材料制成的?

选项：
塑料砖

图像特征：
提供的图像似乎是某种三维结构的二维表示，可能是一条由许多小块或砖块组成的道路或路径。这些块似乎沿着一条弯曲的路径排列，从图像顶部蜿蜒延伸至底部。这些块按照某种图案排列，其中一些似乎嵌入其中或以倾斜角度放置，另一些则似乎与其他块呈90度角。

指令：请严格按照“答案是”的格式输出答案。**多模态大语言模型：**答案是B。

图像：



问题：这条路径是用什么材料制成的？**选项：**
塑料砖

图像特征：
所提供的图像似乎是三维结构的二维表示，可能是一条由许多小块或砖块铺成的道路或路径。这些块沿着一条弯曲的路径排列，从图像顶部蜿蜒延伸至底部。它们以某种图案排列，有些块似乎嵌入其中或倾斜放置，另一些则与其他块呈90度角。

先前答案： [B]。请仔细考虑这些答案。如果你认为这些答案都不正确，请提供一个新答案。如果你认为其中有正确答案，请输出该答案。**指令：** 请严格按照“答案是 ”的格式输出答案。**多模态大语言模型：**答案是B。

A.2. 对于 OK-VQA，开放式

表7. 一个无需思维链提示、由 Qwen2.5vl:7b 作答的 OK-VQA 示例。粉色表示多模态大语言模型的答案，蓝色突出显示将历史信息作为参考的推理逻辑，而绿色表示新增的视觉描述。

图像：



指令：观察图像并描述其视觉特征。**多模态大语言模型：**图像中，一人背对镜头坐在教堂长椅上。

图像：



问题：这是什么玩具？**图像特征：**图像中，一人背对镜头坐在教堂长椅上。
指令：严格按照“The answer is ”的格式输出答案。**多模态大语言模型：**答案是泰迪熊。

图像：



问题：这是什么玩具?
图像特征：
图像中，一个人背对镜头坐在教堂长椅上。**先前的答案：** [泰迪熊]请仔细考虑这些答案。如果你认为这些答案都不正确，请给出一个新答案。如果你认为其中有一个正确答案，请输出该答案。**指令：**请严格按照“Theanswer is ”这一格式输出答案。**多模态大语言模型：**答案是泰迪熊。

图像： 
指令：再次观察图像并描述其视觉特征。**多模态大语言模型：**这是一个温馨的室内空间，很可能是教堂礼拜堂，沐浴在柔和温暖的光线中。一只浅色毛发的泰迪熊从黑色背包顶部探出头来。背包属于一位穿着红色毛衣的人（只能看到对方深色卷发和背影）。

图像： 
问题：这是什么玩具? **图像特征：**
图像中，一人背对镜头坐在教堂长椅上。这里是一个舒适的室内空间，可能是教堂礼拜堂，沐浴在柔和温暖的光线中。一只浅色毛发的泰迪熊从黑色背包顶端探出头来。背包属于一位穿着红色毛衣的人（只能看到其深色卷发和背影）。**先前的答案：** [泰迪熊，泰迪]请仔细考虑这些答案。如果你认为这些答案都不正确，请给出一个新答案。如果你认为其中有正确答案，请输出该答案。**指令：**严格按照“The answer is ”的格式输出答案。**多模态大语言模型：**答案是泰迪熊。

表 8。一个由 Qwen2.5vl:7b 作答、未使用任何思维链提示的 OK-VQA 示例。粉色表示多模态大语言模型的答案，蓝色突出显示了将历史信息用作参考的推理逻辑。

图像： 
指令： 再次观察图像，并描述其视觉特征。**多模态大语言模型：** 图像中，一名男子在户外的草地上，双手各拿着一个白色飞盘。

图像： 
问题： 如果这种动物在玩那个男子手中物品所用于的游戏，它会用到身体的哪个部位?

图像特征：
图像显示一名男子在户外的草地上，双手各拿着一个白色飞盘。**指令：**严格按照“答案是 ”的格式输出答案。**多模态大语言模型：**答案是 嘴。

图像： 

问题：如果这种动物正在玩那个男子手中物品所用于的游戏，它身体的哪个部位会用到?

图像特征：
图中显示一名男子在户外的草地上，每只手拿着两个白色飞盘。**先前的答案：** [嘴]。请仔细考虑这些答案。如果你认为这些答案全部错误，请提供一个新答案。如果你认为这些答案中有一个正确答案，请输出该答案。
指令： 严格按照“答案是 ”的格式输出答案。**多模态大语言模型：**答案是 嘴。