MC-SEARCH：通过结构化长推理链评估和增强多模态智能体搜索

Xuying Ning[1*], **Dongqi Fu**[2*], **Tianxin Wei**[1*], **Mengting Ai**[1], **Jiaru Zou**[1], **Ting-Wei Li**[1], **Hanghang Tong**[1], **Yada Zhu**[3], **Hendrik Hamann**[4], **Jingrui He**[1†]伊利诺伊大学厄巴纳-香槟分校, [2]Meta, [3]IBM 研究院, [4]石溪大学
[*]共同贡献。[†]通讯作者。{xuyingn2, jingrui}@illinois.edu https://mc-search-project.github.io

arXiv:2603.00873v1 [cs.AI] 1 Mar 2026

摘要

随着对逐步、跨模态和以知识为基础的推理需求日益增长，多模态大语言模型（多模态大语言模型）正逐渐超越传统的固定先检索后生成范式，转向更复杂的智能体多模态检索增强生成（多模态检索增强生成）。然而，现有基准主要关注检索链较短的简化问答，对自适应规划和多模态推理的探索不足。我们提出**MC-SEARCH**，这是首个针对智能体多模态检索增强生成的基准，包含跨越五种代表性推理结构的长篇、逐步标注推理链。每个样例均明确列出子问题、检索模态、支持事实和中间答案，并通过**HAVE**（证据逐跳归因与验证）确保保真度，最终构建出 3,333 个高质量样例，平均跳数为 3.7。除答案准确率外，MC-SEARCH 还引入了用于衡量推理质量、逐步检索和规划准确率的新过程级指标。通过开发统一的智能体多模态检索增强生成流程，我们对六种领先的多模态大语言模型进行了基准测试，并揭示了过度检索、检索不足和模态错配规划等系统性问题。最后，我们提出**SEARCH-ALIGN**，这是一种利用经验证的推理链进行过程监督微调的框架。结果表明，我们的数据不仅支持忠实评估，还能提升开源多模态大语言模型的规划和检索保真度。

1 引言

多模态大语言模型（MLLMs）在推理能力方面迅速进步（Wang 等人，2025b；Huang 等人，2025b；Li 等人，2024年），从纯文本输入拓展到交错的跨模态上下文。为确保事实性和鲁棒性，检索增强生成（RAG）已成为一种关键范式，可借助外部文本证据为模型输出提供事实依据（Gao 等人，2023年；Zhao 等人，2024年；Fan 等人，2024年）。在此基础上，多模态检索增强生成（MM-RAG）赋予多模态大语言模型视觉搜索能力，并将基于事实性从纯文本证据扩展至多模态上下文，从而促进针对知识密集型查询的跨模态推理。

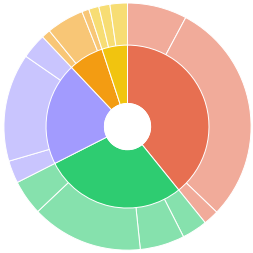
实际上，面向多模态大语言模型的查询往往含糊且复杂（图 1）。它们远非通过固定的最少检索即可解决的浅层问题，而是需要多步、跨模态且知识密集型的推理。随着多模态大语言模型向长推理模型发展（Wang 等人，2024年），它们日益被定位为满足此类问题求解需求的骨干模型。然而，要实现这一潜力，就需要更丰富的智能体行为（Li 等人，2025a；b），例如迭代式任务分解、自适应跨模态检索和多模态证据整合。这一转变要求超越经典检索增强生成固定的先检索后生成范式，转向多模态智能体搜索增强推理（多模态智能体检索增强生成）（Wang 等人，2025a；Li 等人，2025a；Huang et al.，2025a；Jin et al.，2025年；Zheng 等人，2025b；Chen等人，2025年）。因此，我们需要新的基准和评估流程，以准确捕捉并评估这些能力。



图1：一个需要多模态智能体搜索和长推理链的查询示例。

表 1：**左侧**：现有多模态检索增强问答数据集的比较。**右侧**：MC-SEARCH 中五种推理拓扑的分布，外环展示跳数多样性（2–5 跳）。

基准	知识模态	典型 Hops	长链（≥ 4跳数）	逐步标注	推理拓扑
OK-VQA（Marino等，2019年）	Text	1 跳	✗	✗	✗
ViQuAE（Lerner等，2022年）	Text	1 跳	✗	✗	✗
WebQA（Chang等，2022年）	文本/图注	≤2 跳数	✗	✗	✗
InfoSeek (Chen等人,2023年)	Text	1 跳	✗	✗	✗
MMSearch (Jiang等人,2024年)	文本/图像	1 跳	✗	✗	✗
Dyn-VQA (Li 等人,2025b)	文本/图像	≤2 跳数	✗	✗	✗
MRAG (Hu 等人,2025年)	图像	1 跳	✗	✗	✗
M²检索增强生成（Liu 等人，2025c）	文本/图像	1 跳	✗	✗	✗
MC-SEARCH	**文本/图像**	**≥4 跳数**	✓	✓	✓



尽管近期有若干基准（Hu 等人， 2025年； Jiang等人， 2024年； Liu 等人， 2025c； Li 等人， 2025b）对多模态检索增强生成进行了有价值的评估，但它们主要关注简单的视觉证据：所研究的问题仅需 1–2 个检索步骤，因此无法充分检验实践中所需的自适应搜索增强推理行为。具体而言，现有数据集仍存在三方面的关键局限：（i）大多数采用简单的问答格式和固定的先检索后生成流水线，将多模态证据归并到主要的文本通道中（Hu 等人， 2025年）；（ii）它们将评估限制在短程的 1–2 跳检索，缺乏长程、自适应的推理轨迹（Li 等人， 2025b）；（iii）它们缺少逐步标注和明确的推理拓扑，无法阐明不同模态在推理过程中的作用（Hu 等人， 2025年； Liu 等人， 2025b；Li 等人， 2025b）。如表 1 （左）所示，这些局限使人难以判断多模态大语言模型是否真正能够基于多模态证据进行长程、结构化推理。

为填补这一空白，我们推出了**MC-SEARCH**，这是首个面向具有长结构化Chain 的**Search**增强推理的**M**多模态智能体 RAG 基准。每个样例均配有黄金逐步轨迹，标明子问题序列、检索模态、支持事实和中间答案，从而支持细粒度评估，并为过程奖励建模奠定基础。

MC-SEARCH 旨在实现长链、多样且不冗余。如图 2 所示，它涵盖五种具有代表性的多跳推理结构，包括纯文本链、图像发起链、文本发起链、并行图文分叉和多图像分叉，展现跨模态的串行与并行推理模式。为确保每一跳都不可或缺且具有结构意义，我们提出 HAVE（逐跳证据归因与验证），用于筛除无效或冗余步骤，最终得到 3,333 个高质量标注样例，平均链长度为 3.7 跳，优于现有基准（Hu 等人，2025年；Liu 等人，2025b；Li 等人，2025b）。表 1（右）报告了不同推理拓扑下的链长度分布。

除答案级准确性外，我们在 MC-SEARCH 中提出了三**个过程级评估指标**：(i) *LLM* 评审 ，用于评估开放式推理质量；(ii) 结构感知逐步命中率 ，用于评估逐步检索保真度；(iii) 展开偏差 ，用于量化执行偏移。为评估这些能力，我们进一步构建了统一的智能体式多模态检索增强生成流程，以开展公平的基准测试，并使用六种领先的多模态大语言模型进行广泛实验，其中包括专有模型和开源模型。我们的分析揭示了过度检索、检索不足和模态特定规划错误等关键弱点，凸显了改进多模态证据上的自适应规划与检索的必要性。

除了作为评估资源，我们的数据还为提升模型能力提供训练信号。为此，我们提出 **SEARCH-ALIGN**，这是一种基于经 HAVE 验证的推理链构建的对话级对齐框架。该框架将监督微调扩展至最终答案之外，利用逐步轨迹提供过程级监督，从而增强开源模型跨模态规划和检索的能力。

我们的主要贡献总结如下：

- **基准：** 我们提出 MC-SEARCH，这是首个面向智能体多模态 RAG 的基准，涵盖五种代表性结构，并包含较长的、逐步标注的推理链；这些推理链经 HAVE 程序验证其必要性和非冗余性。
- **指标：** 我们提出新的过程级指标，不再局限于答案准确率，而是精准评估推理质量、逐步检索和规划保真度。
- **评估：** 我们构建了统一的智能体式多模态检索增强生成流程，并对六种领先的多模态大语言模型进行了广泛的基准测试，揭示了系统性问题，包括检索过度与检索不足、规划与模态不匹配，以及八种典型错误类型。

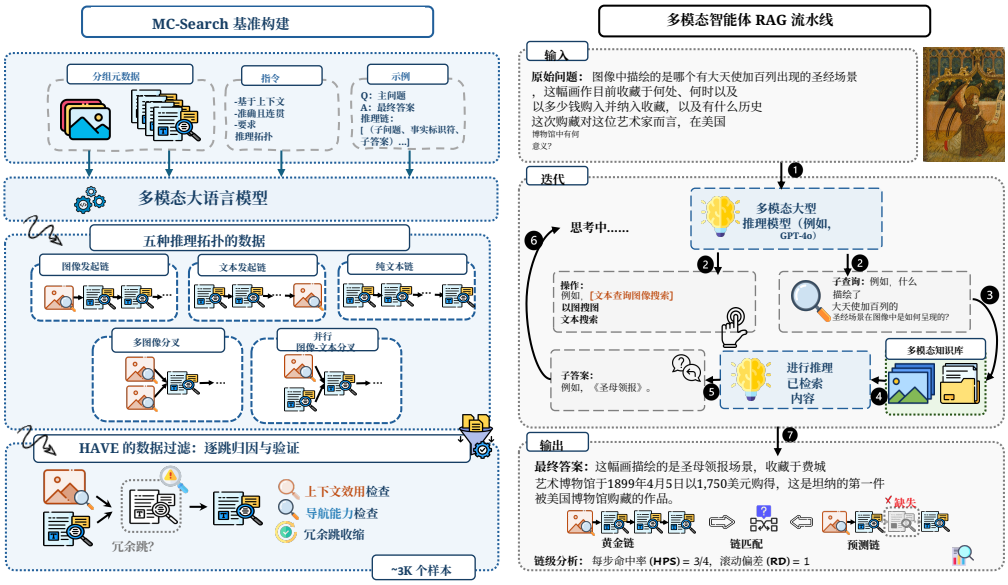


图 2：**MC-SEARCH** 基准与评估概览。**左侧**：涵盖五种推理拓扑的基准，并通过逐跳证据归因与验证（HAVE）流程进行筛选。**右侧**：多模态智能体 RAG 流水线，其中多模态大语言模型（MLLM）迭代生成子查询和操作，检索多模态证据，对检索到的信息进行推理，并将其整合以生成最终答案。我们的框架还将预测的推理链与标准轨迹对齐，以评估链级检索和规划。

• **方法：** 我们提出了 SEARCH-ALIGN，利用经验证的推理链对多模态大语言模型进行过程监督微调。结果表明，我们的数据能够有效用于训练，并提升智能体多模态检索增强生成中的规划能力和检索忠实度。

2 MC-SEARCH 基准

我们提出了 MC-SEARCH，这是一个用于评估智能体多模态 RAG 的大规模基准。不同于以浅层视觉问答任务为重点的先前基准，MC-SEARCH 提供跨多种推理拓扑的长程、逐步验证的推理链，从而支持细粒度的过程级分析和模型对齐。本节将详细介绍基准构建、数据集组成和过程级评估指标，整体流程如图 2所示。

2.1 基准构建与组成

搜索增强推理拓扑结构。 为捕捉现实世界中智能体多模态检索增强生成工作流的多样性，我们设计了 五种不同的推理拓扑结构 ，用于刻画多模态知识在长期搜索增强推理过程中如何相互作用并相互依赖。每种拓扑结构均表示为一张推理图（或推理链） $\mathcal{G}$（Q、 A），并与一个总体问题 Q 及其最终答案 A相关联。该图由子问题 q_t、检索模态$m_t \in$ {文本、图像}、检索到的证据 r_t以及 a_t 在跳 t （图 2）中根据证据推导出的中间答案组成。我们将 $\mathcal{R}$ 定义为检索函数，该函数给定一个子问题 q_t 和模态m_t，返回相应的证据 r_t。形式化地，

$$\mathcal{G}(Q, A) = \{(q_t, m_t, r_t, a_t)\}_{t=1}^{T}, \quad r_t = \mathcal{R}(q_t, m_t), \quad A = f(\{a_t\}_{t=1}^{T}), \tag{1}$$

其中， $f(\cdot)$ 表示将中间答案聚合为总体问题 Q的最终答案A 的推理过程。

具体而言，五种推理图定义如下：(i)图像发起链，推理以第一步的图像检索为锚点，随后基于文本检索展开；适用于包含图像和纯文本查询；(ii) 文本发起链，推理从检索事实知识的文本检索开始，之后再纳入图像检索以进行视觉验证；(iii)并行图文分叉，图像检索和文本检索并行推进，彼此不依赖特定跳数，因此互换此类跳数的检索顺序不会改变最终答案，但需要跨模态协调；(iv)多图像分支，其中必须检索多张图像以进行视觉比较，然后转向文本证据

以获取事实支持；(v) 纯文本链，作为基线，其推理完全通过文本进行。附录C中提供了每种推理图的数据示例。

数据生成。我们通过从维基百科收集多模态知识，并将其整理为连贯的文本与图像簇（通常取自相同或密切相关的页面），来构建 MC-SEARCH 数据集，方法遵循Chang等(2022年)。基于这些簇，我们筛选出重要的文本事实及相关图像，作为生成多跳问答的基础，并补充与各推理拓扑相匹配的示例。随后，我们从筛选后的素材池中大规模生成问题，同时确保五种推理结构的覆盖度均衡。具体而言，我们提示 Gemini-2.5-Flash 选择一种拓扑，并生成结构化的多跳问题及其答案，以及以支持事实为依据的连贯推理链。该流程生成了约 2.1 万个与五种搜索增强推理拓扑结构相对应的示例。更多构建细节和提示模板请参见附录N。

数据过滤。构建高质量数据集的一项关键挑战是，多模态大语言模型经常会生成包含幻觉步骤的长推理链，也就是说，生成看似合理、却缺乏依据且没有证据支持的跳数，以及冗余步骤，这些步骤对最终答案没有帮助。为解决这一问题，我们提出**HAVE**（逐跳证据归因与验证），这是一种数据过滤机制，用于判定推理链中的每一跳是否必要且不冗余。令 $\mathcal{C} = \{r_1, r_2, \ldots, r_T\}$ 表示推理图中每一跳检索到的证据集合。我们将步骤 r_t 的**上下文效用** 定义为移除 r_t 后答案准确率的下降幅度（以标准答案与生成答案之间的 F1 分数衡量）：

$$\text{Util}(t) = \text{F1}(\mathcal{C}) - \text{F1}(\mathcal{C} \setminus r_t). \tag{2}$$

如果 Util(t) 超过阈值，则认为该步骤是必要的。不过，有些步骤可能不会直接影响答案准确率，却仍能连接子问题并引导后续推理。例如，在一个两跳链中，第一步可能会将输入图像中的建筑识别为“基督堂座堂”，而第二步则查询有关该实体的事实知识。尽管第一步本身的上下文效用较低，但它提取出的实体对第二跳至关重要，因此对于推理过程不可或缺。为涵盖这类情况，我们还会通过检查步骤的中间答案 a_t 中的实体是否出现在后续子问题中，来评估该步骤的**导航作用**：

$$\text{Nav}(t) = \begin{cases} 1, & \text{if Ent}(a_t) \cap \text{Ent}(q_{t+1:T}) \neq \emptyset, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

whe其中 Ent(·) 表示从子问题或中间答案中提取的实体集合。如果 a step Util(t)低于阈值且 Nav(t) $= 0$ 时，将其标记为冗余。

我们首先使用开源模型 Qwen2.5-VL-7B 对生成数据集进行筛选，移除推理链中冗余跳数超过两次的样本，从而大规模应用 HAVE，得到 4.8k 条候选链。为进一步减轻模型特定偏差，我们随后使用 Gemini-2.5-Pro 进行跳数缩减和样本验证，同时从标准答案中提取关键知识实体，以便后续进行答案评估。最后，我们使用数据生成模型 Gemini-2.5-Flash，通过 HAVE 进行冗余检查，最终得到 3,333 条高质量样本，其中包含所有必要跳数和连贯的推理轨迹。最终数据集的构成和统计数据见表2 和附录E。

表 2：基准统计数据。

统计项	数量
并行图文分叉	680
图像发起链	1,306
文本链	945
文本发起链	169
多图像分叉	233
总样本	**3,333**
知识库图像	389,750
知识库文档	784,473
平均跳数	3.79

此外，为确保后续逐步检索准确率评估的可靠性，我们要求每个问题的推理图都必须唯一。为此，我们使用 Gemini-2.5-Flash 进行自动验证，在知识库中检测并筛除混淆性知识片段，确保同一主题簇中的任何未使用文本事实或图像都无法独立解答任何子问题。这为智能体多模态检索增强生成的逐步评估奠定了坚实基础。

质量验证。为确保数据集在推理连贯性、基于事实性和问答准确性方面的质量，我们采用 Gemini-2.5-Pro，按 1–5 分量表从四个维度评估每条链：事实正确性、步骤必要性、清晰度和多模态对齐（附录 F）。该数据集

总体平均分达到 4.87，低分样本进一步通过针对性的答案编辑加以完善，从而形成用于搜索增强型多模态推理的高质量基准。

2.2 评估协议

为了评估多模态搜索增强推理，我们认为，仅衡量最终答案准确率是不够的，尤其是对于较长的推理轨迹而言，因为检索或规划中的错误可能不会反映在最终输出中。为了实现更细粒度的评估，我们引入了同时刻画答案正确性以及逐步检索和规划保真度的指标。

答案准确率。 我们采用四种互补指标：**(i) F1**，即生成的最终答案与标准答案之间标准的词元级重叠分数；**(ii) ΔF1**，用于衡量智能体多模态检索增强生成相对于模型参数化知识所带来的性能增益，定义为使用检索与不使用检索时性能之差： $\Delta\text{F1} = \text{F1} - \text{F1}_{\text{w/o } \mathcal{R}}$；**(iii) 黄金标准 F1**，即向模型提供标准推理链和标准检索内容后得到的上界；以及 **(iv) LLM 评审（LJ）**，由强大的推理模型（Gemini-2.5-Pro）从四个维度将预测的推理链与标准推理链进行评估：答案准确性、推理连贯性、关键知识实体覆盖率和步骤对齐（提示词见附录 P）。

ˆ链对齐。 我们评估预测推理图 $\mathcal{G}$ 与黄金推理图 $\mathcal{G}$的对齐程度。每个推理图都是由检索增强推理步骤组成的序列 ˆ，索引为 t，其中 $r_t \in \mathcal{G}$ 表示在黄金步骤 t中检索到的证据，而 $\hat{r}_{t'} \in \mathcal{G}$ 表示在预测步骤 t'中检索到的证据。

为评估步骤级检索保真度，我们将 **(i) 每步命中率（HPS）**指标定义为黄金步骤中其证据在预测图中成功找回的比例：

$$\text{HPS}(\hat{\mathcal{G}}, \mathcal{G}) \;=\; \frac{1}{|\mathcal{G}|}\Big|\{\,(t,t') \;\mid\; r_t \in \mathcal{G},\; \hat{r}_{t'} \in \hat{\mathcal{G}},\; \hat{r}_{t'} = r_t\}\Big|, \tag{4}$$

其中 $|\mathcal{G}|$是黄金步骤的数量，右侧的 $|\cdot|$是匹配步骤的数量。每个“命中”对应一个预测步骤，其证据与某个黄金步骤的证据完全匹配，重复项 ˆ仅计一次。除精确匹配外，我们还通过最大权二分图匹配对齐 $\mathcal{G}$ 和 $\mathcal{G}$，其中节点对应两个图中的步骤，边权重反映预测步骤与黄金步骤之间的证据相似度。这种对齐支持细粒度的逐步比较，并能更好地处理涉及并行推理步骤的情况。

为刻画结构保真度，我们测量预测推理图与黄金推理图之间的步骤长度差。较大的差值表明存在检索不足或过度检索，而较小的值则意味着推理复杂度更为接近。我们将该指标称为**(ii) 展开偏差（RD）**：

$$\text{RD}(\hat{\mathcal{G}}, \mathcal{G}) = \big|\,|\hat{\mathcal{G}}| - |\mathcal{G}|\,\big|. \tag{5}$$

3 AGENTICMM-RAG流水线与过程级对齐

为系统评估并改进多模态大语言模型在搜索增强推理中的表现，我们提出了一种统一的智能体多模态检索增强生成流程，用于自适应规划、多模态证据检索以及基于证据的推理；同时提出SEARCH-ALIGN这一过程级对齐框架，利用经验证的推理链来增强开源模型。

3.1 AGENTICMM-RAG流水线

如图2（右）所示，我们将多模态搜索增强推理建模为一个包含三个模块的迭代过程：**(i)** 子查询与动作生成、**(ii)** 证据获取以及**(iii)**迭代推理与综合。在每次迭代中，智能体制定一个子目标，选择一个模态感知检索动作，整合检索到的证据，并决定继续搜索还是以最终答案结束。该设计不仅支持有效推理，还支持针对智能体RAG的细粒度步骤级评估。

（i）子查询与动作生成。给定一个复杂问题，多模态大语言模型首先预测下一个子目标，并生成一个聚焦的子查询（图 2 中的步骤 1–2）。然后，它会自适应地从三种检索操作中选择一种，这些操作对应于获取多模态证据的不同方式：（a）使用文本查询进行文本搜索；（b）使用文本查询进行图像搜索；或（c）使用输入图像进行图像搜索。

（ii）证据获取。 所选操作会在我们的本地多模态知识库中执行。子查询由多模态编码器嵌入，以便在相应模态中进行密集检索，并返回文本或视觉证据。对于每个子查询，我们根据查询-答案相似度保留排名第一的证据（步骤 3），并让多模态大语言模型在推理后生成子答案（步骤 4–5）。为进行过程评估，系统会记录所选模态和检索到的证据。

(iii) 迭代推理与综合。 将子答案及其证据反馈给模型，以指导下一步规划（步骤 6），从而形成规划、检索和推理的循环。智能体会动态检查累积的证据是否充分；如果充分，则输出最终答复（步骤 7）。否则，它会继续提出另一个子查询。这种迭代式设计展现了完整的推理轨迹，使模态感知执行和精确的链级评估成为可能。

3.2 SEARCH-ALIGN：过程级对齐

除了评估之外，我们的数据集还是提升模型能力的资源。我们提出 **SEARCH-ALIGN**，一种基于经 HAVE 验证的推理链、面向开源多模态大语言模型的对话级对齐框架。与仅监督最终答案的传统监督微调不同，SEARCH-ALIGN 利用包含子问题、检索操作、证据和中间答案的逐步标注轨迹，提供过程级监督。

训练数据构建。 每个推理图都通过 Gemini-2.5-Flash 生成的显式推理思考进行增强，这些思考解释了如何依据证据进行推理，以及如何连接相邻跳数。这将逐步标注转化为连贯的对话轨迹：助手提出子问题并依据证据进行推理，而用户执行检索操作并返回结果。

监督微调（SFT）。随后，我们使用这些对话式轨迹对多模态大语言模型进行微调，使其不仅学会生成正确答案，还能规划、选择检索模态，并跨步骤整合证据。这种过程级对齐提供了更丰富的训练信号，使模型更好地为长程多模态推理做好准备。

4 实验

在本节中，我们在统一的智能体多模态检索增强生成流程下，使用 MC-SEARCH 基准评估领先的多模态大语言模型，重点关注五个研究问题（RQ）：**研究问题1**：多模态大语言模型在不同推理结构下表现如何？SEARCH-ALIGN 能否提升开源模型的表现？**研究问题2**：推理链长度如何影响推理表现？**研究问题3**：检索过度和检索不足会产生什么影响？**研究问题4**：模型是否表现出内在的模态偏好？**研究问题5**：多模态智能体推理中常见的错误类型有哪些？

4.1 实验设置

We 使用第3.1节介绍的智能体式多模态检索增强生成流程，评估以下多模态大语言模型 ：
• **闭源模型**：GPT-4o-Mini（Achiam 等人，2023年）、Gemini-2.5-Flash（Team 等人，2023年）、Gemini-2.5-Pro（Team 等人，2023年）和 Claude-3.7-Sonnet（Anthropic，2024年，代表了当前商用领域中实力最强的推理型多模态大语言模型。
• **开源模型**：InternVL3.5-8B（Wang 等人，2025b）和 Qwen2.5-VL-7B（Bai 等人，2025年），这两款模型均为先进的视觉语言模型。我们还评估了它们使用 SEARCH-ALIGN 框架后的表现。

评估协议。 所有模型均在相同条件下进行评估。对于多模态检索，我们使用相同的嵌入模型（Zhou 等人，2024年）、提示词（附录O）以及各模型通用的解码参数。每个检索步骤仅限于从本地多模态知识库中获取排名第 1 的结果，并采用相同的最大推理迭代次数。我们使用第 2.2 节中的六项指标报告性能：答案准确率（F1、 ΔF1、黄金标准 F1 和 LLM 评审）以及链级检索与规划保真度（每步命中率和展开偏差）。关于评估以及 SEARCH-ALIGN 训练的更多细节见附录 I。此外，在附录 D.1中，我们还报告了骨干模型的固定步数 RAG 基线（1 跳和 2 跳），以补充主要结果。

表格 3： 在智能体式多模态检索增强生成流程下评估多模态大语言模型在 MC-SEARCH 基准上的表现，结果按五种推理图报告。最佳骨干模型结果以 **粗体**显示，次佳结果加下划线；SEARCH-ALIGN 在开源模型上的改进以 **红色粗体**突出显示。

推理图		模型	答案准确率 F1 (↑)	ΔF1 (↑)	LJ (↑)	链对齐 HPS (↑)	展开偏差 (↓)	黄金标准 F1 (↑)
图像发起链	CLOSE	GPT-4o-Mini	36.49	34.18	2.63	27.51	1.46	68.29
		Gemini-2.5-Flash	44.10	37.38	3.01	**31.46**	2.91	72.39
		Gemini-2.5-Pro	**47.61**	**42.76**	**3.18**	25.90	**1.05**	69.83
		Claude-3.7-Sonnet	37.80	33.09	2.60	27.31	1.18	**72.62**
	OPEN	InternVL3.5-8B	39.11	29.49	2.27	22.59	1.58	63.86
		+ SEARCH-ALIGN	**42.27**	**32.65**	**2.53**	**32.49**	**0.94**	
		Qwen2.5-VL-7B	26.30	8.65	1.34	16.51	4.04	60.95
		+ SEARCH-ALIGN	**45.70**	**28.05**	**2.23**	**33.59**	**0.70**	
多图 Fork	CLOSE	GPT-4o-Mini	24.00	19.16	2.13	16.04	1.94	59.46
		Gemini-2.5-Flash	36.80	31.89	2.35	13.45	1.66	**64.40**
		Gemini-2.5-Pro	**40.37**	**36.58**	**2.76**	18.68	**1.40**	61.29
		Claude-3.7-Sonnet	29.78	27.85	2.07	**19.43**	1.45	62.41
	OPEN	InternVL3.5-8B	33.08	20.98	2.13	16.02	2.25	57.65
		+ SEARCH-ALIGN	**36.53**	**24.43**	**2.49**	**39.33**	**1.67**	
		Qwen2.5-VL-7B	26.13	7.53	1.41	12.58	3.80	57.04
		+ SEARCH-ALIGN	**38.01**	**19.41**	**2.11**	**38.01**	**1.16**	
文本发起链	CLOSE	GPT-4o-Mini	30.11	18.46	2.41	**27.18**	1.76	49.47
		Gemini-2.5-Flash	43.55	26.34	3.30	25.20	1.20	**66.27**
		Gemini-2.5-Pro	**45.30**	**29.89**	**3.62**	19.51	**0.95**	55.94
		Claude-3.7-Sonnet	36.10	17.36	2.63	25.09	1.12	57.13
	OPEN	InternVL3.5-8B	38.34	25.48	**2.60**	19.06	1.97	38.70
		+ SEARCH-ALIGN	**40.87**	**28.01**	2.59	**37.82**	**1.44**	
		Qwen2.5-VL-7B	31.47	14.34	1.66	17.02	3.80	39.86
		+ SEARCH-ALIGN	**45.07**	**27.94**	**2.42**	**33.59**	**0.87**	
并行图像-文本分叉	CLOSE	GPT-4o-Mini	21.98	21.65	2.20	15.00	1.71	53.98
		Gemini-2.5-Flash	29.92	27.94	2.58	11.43	2.70	57.99
		Gemini-2.5-Pro	**34.83**	**34.19**	**2.99**	**16.74**	**1.29**	53.46
		Claude-3.7-Sonnet	26.43	24.65	1.95	14.32	1.37	58.51
	OPEN	InternVL3.5-8B	29.74	23.05	2.13	14.22	1.75	**58.82**
		+ SEARCH-ALIGN	**30.72**	**24.03**	**2.29**	**23.55**	**1.45**	
		Qwen2.5-VL-7B	22.94	14.32	1.33	10.09	3.93	55.76
		+ SEARCH-ALIGN	**32.73**	**24.11**	**1.78**	**25.07**	**1.05**	
纯文本链	CLOSE	GPT-4o-Mini	30.96	30.94	2.58	21.94	1.13	61.90
		Gemini-2.5-Flash	34.03	33.96	2.49	20.59	**1.04**	**67.72**
		Gemini-2.5-Pro	34.47	**34.46**	**2.66**	21.59	1.07	62.42
		Claude-3.7-Sonnet	27.37	26.75	2.08	**29.64**	3.14	66.38
	OPEN	InternVL3.5-8B	**34.82**	27.84	2.45	**21.81**	2.15	65.72
		+ SEARCH-ALIGN	**38.45**	**31.47**	**2.68**	20.54	**1.28**	
		Qwen2.5-VL-7B	23.54	14.89	1.36	13.35	4.63	61.57
		+ SEARCH-ALIGN	**37.54**	**28.89**	**1.99**	**19.04**	**0.98**	

4.2 主要结果（RQ1）

骨干网络性能与搜索对齐收益。 如表3所示，Gemini-2.5-Pro 等专有模型在大多数推理拓扑上实现了最高的总体准确率。在开源模型中，InternVL3.5-8B 是更强的骨干模型，而 Qwen2.5-VL-7B 则始终落后。然而，通过 SEARCH-ALIGN，两个骨干模型都取得了显著提升：InternVL 的 F1 平均提高约 +2.8，HPS 提高 +12.0，同时 RD 降低 0.6；QwenVL 的提升幅度更大，F1 提高 +13.7，HPS 提高 +16.0，RD 下降 3.1。对齐后，Qwen2.5-VL-7B 在文本中心链上的表现几乎达到 Gemini-2.5-Pro 的水平，而 InternVL 在所有拓扑结构上都展现出稳定提升。这些结果表明，开源模型的主要弱点在于检索规划和逐步推理对齐，而非基础的感知或语义理解；同时也表明，通过使用高质量的搜索增强推理数据进行训练，可以有效提升这些能力。

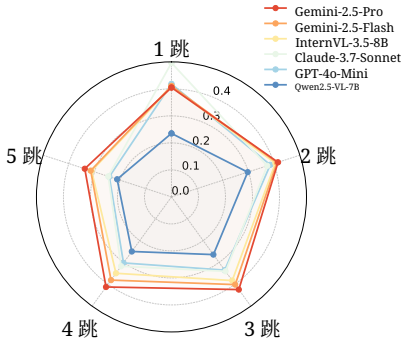


图 3：不同链长度样本的模型 F1 分数，显示随着链长度增加，性能持续下降。

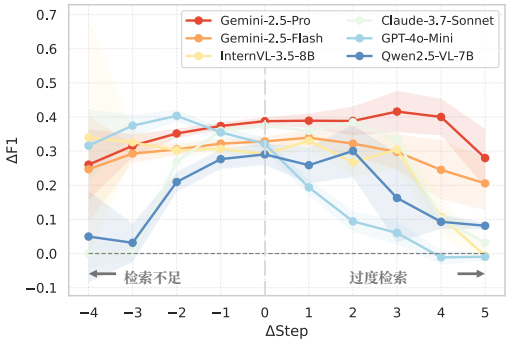


图 4：模型 ΔF1 与 Δ 步数（生成的推理链与标准推理链之间的长度差）的关系，其中较大的正向 Δ步数表示过度检索。

拓扑特定挑战。 在各种推理拓扑中，我们发现并行图文分叉最具挑战性，所有骨干模型在此拓扑上的 F1 和 HPS 均达到最低。这是因为需要同时规划并覆盖文本和图像分支，遗漏任一分支都会直接降低 HPS；而在 top-1 检索约束下，模型往往无法获取完整的证据集。缺乏严格顺序的并行证据对齐，进一步增加了推理不完整或不一致的风险。多图像分支也颇具难度，主要因为需要将多个视觉证据与相应的文本事实对齐；模型往往能检索到正确的图像，却会错误归因，或无法连贯地整合这些信息。相比之下，文本发起和纯文本链相对容易，因为它们主要依赖文本检索和线性推理，这与模型的优势相符。图像发起链也较易处理：一旦初始视觉步骤正确，后续的文本验证便是线性的，也比多分支推理更稳定。

4.3 分析

RQ2：性能与链长度。图3 展示了不同链长度样本上的模型 F1 分数，表明随着链长度增加，性能会持续下降。对于较短的推理链（1–3 跳），模型仍相对稳健，各骨干模型之间的性能差异适中；但 Qwen2.5-VL-7B 是个例外，即使在较早的跳数上，其表现也始终落后。然而，当链长度延伸至 4 或 5 跳时，所有模型的性能都急剧下降，反映出持续进行长时程推理的难度会不断累积。

在闭源系统中，Gemini-2.5-Pro 最为稳健，在较长跳数下仍能保持最高性能。Gemini-2.5-Flash 和 InternVL3.5-8B 的性能逐渐下降，而 GPT-4o-Mini 和 Claude-3.7-Sonnet 在超过 3 跳后则急剧下滑。Qwen2.5-VL-7B 在所有链长度下的表现都最差。总体而言，主要挑战在于延伸的多步链，其中不断累积的检索错误和不稳定的规划会降低准确性。

RQ3：过度检索与检索不足分析。图4 展示了 Δ不同 Δ步数下的 F1，其中负值表示检索不足，正值表示过度检索。结果表明，这两种极端情况都会造成负面影响，但方式各不相同。当模型严重检索不足（Δ步 < -2）时，由于跳过了关键证据，性能会急剧下降，从而产生后续推理无法弥补的信息缺口。相反，适度的过度检索（Δ步 $= 1$–2）往往能提高准确性：Gemini-2.5-Pro、Gemini-2.5-Flash 和 InternVL3.5-8B 在此处取得了最高的 ΔF1，这表明多进行一两轮检索有助于弥补规划不完善的问题。

然而，过度检索过多（Δ步 ≥ 4）时，噪声和无关上下文会导致所有模型的表现大幅下降。Gemini-2.5-Pro 的抗干扰能力最强，Gemini-2.5-Flash 和 InternVL3.5-8B 保持相对稳定，而 Claude-3.7-Sonnet 和 GPT-4o-Mini 则极易受到影响。这凸显了一个核心矛盾：既要获取足够的证据，又要避免过度检索。

研究问题4：模态覆盖率分析。 如表 4所示，该表报告了在有图像输入和无图像输入的查询下，图像与文本两种模态的步骤级检索覆盖率（已覆盖步骤与金标准步骤的对比）。两个模型在所有设置下均保持较高的文本覆盖率（>78%），但图像覆盖率明显更弱，且高度依赖于显式图像输入。例如，Gemini-2.5-Pro 在有图像提及时为 87.35%，无图像提及时降至 29.50%；而 InternVL3.5-8B 则从 63.84% 降至 0.66%。

这些发现揭示了显著的模态鸿沟：模型往往默认进行文本检索，并在缺乏显式图像线索时失去视觉grounding。

表4： 逐步检索模态覆盖率（Cov.）。 红色 =表示高覆盖度；蓝色 = 表示低覆盖度。

Type	模态	Gemini-2.5-Pro		InternVL-3.5-8B	
		覆盖数/金标准数	覆盖率（%）	覆盖数/金标准数	覆盖率（%）
带图像的查询	图像	1,339/1,533	87.35	977/1,533	63.84
	Text	3,230/4 ,109	78.61	3,395/4 ,109	82.67
不含图像的查询	图像	300/1,017	29.50	7/1,017	0.66
	Text	2,097/2 ,510	83.55	2,251/2 ,510	89.78
总体	图像	1,639/2,550	64.27	861/2,550	33.69
	Text	5,327/6 ,619	80.48	5,688/6 ,619	85.96

研究问题5：错误分析。 图5展示了由 Gemini-2.5-Pro 作为评审标注的八类错误分布。最常见的错误是 检索失败 (84.7%)、实体/属性幻觉 (75.8%) 和 步骤遗漏 (74.3%)，这表明当前模型往往无法以相关证据为推理提供依据、会引入无依据内容，或跳过关键步骤。

Mid-f频率类错误（例如，不匹配、冗余步骤）dality mi

反映了规划的不稳定，而结构性错误（顺序/依赖关系错误、多跳失败）则较为少见，但鉴于检索有限，仍值得未来进一步研究。证据误读并不常见，这表明模型能够处理检索到的上下文；主要失败更早发生在规划和证据获取阶段。总体而言，这些模式结合我们的 HPS 以及过度/检索不足分析，表明检索保真度（考虑模态的检索、充分的规划/跳数覆盖，以及

校准停止）是可靠链级推理的核心瓶颈。

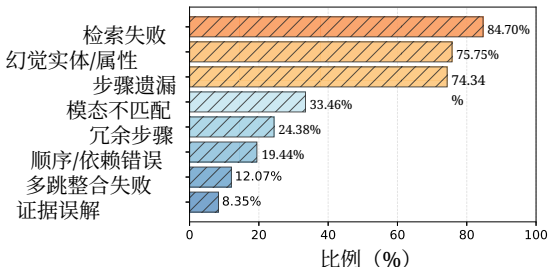


图5：八类错误的比例（比例越高表示情况越差），由 Gemini-2.5-Pro 标注。

5 相关工作

我们回顾现有的多模态检索增强生成基准和智能体RAG相关工作，完整详情见附录 B。

多模态检索增强生成基准。 多模态检索增强生成通过纳入图像证据来支持跨模态推理（Xia等，2024年；Chang等，2022年）。现有基准仍有局限：大多数采用固定的先检索后生成流程（Marino等，2019年；Hu 等人，2025年），将图像简化为图像说明（Lerner等，2022年；Chen等人，2023年），将推理限制在 1–2 跳（Li 等人，2025b；Liu 等人，2025c），且缺少逐步标注，因此难以评估多模态搜索增强推理。我们推出 MC-SEARCH，其中包含经 HAVE 验证、覆盖多种拓扑结构的长推理轨迹，可支持细粒度归因与分析。

智能体RAG。 近期工作将检索重新定义为序贯决策，其中智能体式检索增强生成系统分解查询、自适应检索并综合知识（Wang 等人，2025a；Li 等人，2025a；Zheng 等人，2025b）。尽管有效，大多数仍停留在纯文本层面，并聚焦于多跳问答或科学领域（Trivedi 等人，2022年；Rein 等人，2024年）。我们的基准通过将智能体RAG扩展到多模态场景，并提供多样化且逐步经验证的推理链，填补了这一空白。

6 结论

我们提出了 MC-SEARCH，这是一个用于结构化、逐步多模态检索增强推理的基准，涵盖五种不同的推理机制。借助细粒度标注、逐跳归因与验证，以及新的链级指标，MC-SEARCH 能够全面诊断检索规划和推理质量。我们的分析凸显了在智能体多模态检索增强生成系统中采用自适应、具备模态意识的搜索策略的重要性。展望未来，我们希望 MC-SEARCH 能成为推动多模态智能体发展的基础，并促进建立智能体推理的原则性评估标准。我们将把评估扩展到更强大的推理模型，并将基准拓展至科学和数学等更多领域。

致谢

本研究获得了美国国家科学基金会（NSF）（2134079）和 IBM-Illinois Discovery Accelerator Institute 的支持。本文件所含信息的内容不一定反映美国政府的立场或政策，也不应据此推断获得了官方认可。尽管本文载有任何版权声明，美国政府仍有权为政府用途复制和分发转载件。

伦理声明。 我们的基准完全基于公开来源构建（例如维基百科文本和图像），不涉及私人数据或敏感内容。我们采用逐跳归因与验证（HAVE）流程来过滤幻觉或冗余步骤，确保基于事实性，并尽量降低误导性信息的风险。该数据集仅用于研究目的，以推动多模态推理模型和智能体RAG的发展，并非用于敏感或高风险应用。

可复现性声明。 我们在论文和附录中详细记录了基准构建流程、筛选流程、评估协议和实验设置。为便于完全复现，我们已在论文被接收后公开发布数据集和代码，从而支持对所有报告结果进行独立验证和复现。

参考文献

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

Shakiba Amirshahi、Amin Bigdeli、Charles LA Clarke 和 Amira Ghenai。评估检索增强生成在健康领域面对对抗性证据时的鲁棒性。 *arXiv* 预印本 *arXiv:2509.03787*，2025年。

Anthropic。推出新一代 Claude，2024年。网址 https://www.anthropic.com/news/claude-3-family。

Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。Qwen2. 5-vl 技术报告。 *arXiv* 预印本 *arXiv:2502.13923*，2025年。

Yingshan Chang、Mridu Narang、Hisami Suzuki、Guihong Cao、Jianfeng Gao 和 Yonatan Bisk。WebQA：多跳与多模态问答。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 16495–16504 页，2022年。

Mingyang Chen、Tianpeng Li、Haoze Sun、Yijie Zhou、Chenzheng Zhu、Haofen Wang、Jeff Z Pan、Wen Zhang、Huajun Chen、Fan Yang 等。通过强化学习学习让大语言模型利用搜索进行推理。*arXiv* 预印本 *arXiv:2503.19470*，2025年。Yang Chen、Hexiang Hu、Yi Luan、Haitian Sun、Soravit Changpinyo、Alan Ritter 和 Ming-Wei Chang。预训练视觉语言模型能否回答视觉信息寻求问题？*arXiv* 预印本 *arXiv:2302.11713*，2023年。

Wenqi Fan、Yujuan Ding、Liangbo Ning、Shijie Wang、Hengyun Li、Dawei Yin、Tat-Seng Chua 和 Qing Li。《关于检索增强生成与大语言模型相结合的综述：迈向检索增强型大语言模型》。载于 Ricardo Baeza-Yates 和 Francesco Bonchi（编），《第*30*届 *ACM SIGKDD* 知识发现与数据挖掘会议论文集》，*KDD 2024*，西班牙巴塞罗那，*2024*年*8*月*25*日至*29*日，第6491–6501页。ACM，2024年。doi：10.1145/3637528.3671470。网址https://doi.org/10.1145/3637528.3671470。

Yanlin Feng、Xinyue Chen、Bill Yuchen Lin、Peifeng Wang、Jun Yan 和 Xiang Ren。《面向知识感知问答的可扩展多跳关系推理》。*arXiv* 预印本 *arXiv:2005.00646*，2020年。Robert Friel、Masha Belyi 和 Atindriyo Sanyal。《Ragbench：用于检索增强生成系统的可解释基准》。*arXiv* 预印本 *arXiv:2407.11005*，2024年。

Yunfan Gao、Yun Xiong、Xinyu Gao、Kangxiang Jia、Jinliu Pan、Yuxi Bi、Yi Dai、Jiawei Sun、Qianyu Guo、Meng Wang 和 Haofen Wang。面向大语言模型的检索增强生成：综述。*CoRR*，abs/2312.10997，2023年。doi：10.48550/ARXIV.2312.10997。网址 https://doi.org/10.48550/arXiv.2312.10997。

Wenbo Hu、Jia-Chen Gu、Zi-Yi Dou、Mohsen Fayyaz、Pan Lu、Kai-Wei Chang 和 Nanyun Peng。MRAG-bench：面向检索增强多模态模型的以视觉为中心的评估。载于第十三届国际学习表征会议，2025年。网址https://openreview.net/forum?id=Usklli4gMc。

Jerry Huang、Siddarth Madala、Risham Sidhu、Cheng Niu、Hao Peng、Julia Hockenmaier 和 Tong Zhang。Rag-rl：通过强化学习和课程学习推进检索增强生成。*arXiv* 预印本 *arXiv:2503.12759*，2025a。Wenxuan Huang、Bohan Jia、Zijie Zhai、Shaosheng Cao、Zheyu Ye、Fei Zhao、Zhe Xu、Yao Hu 和 Shaohui Lin。Vision-r1：激发多模态大语言模型的推理能力。*arXiv* 预印本 *arXiv:2503.06749*，2025b。

Dongzhi Jiang、Renrui Zhang、Ziyu Guo、Yanmin Wu、Jiayi Lei、Pengshuo Qiu、Pan Lu、Zehui Chen、Chaoyou Fu、Guanglu Song 等。Mmsearch：评测大型模型作为多模态搜索引擎的潜力。*arXiv* 预印本 *arXiv:2409.12959*，2024年。

Bowen Jin、Hansi Zeng、Zhenrui Yue、Jinsung Yoon、Sercan Arik、Dong Wang、Hamed Zamani 和 Jiawei Han。Search-r1：通过强化学习训练大语言模型进行推理并利用搜索引擎。*arXiv* 预印本 *arXiv:2503.09516*，2025年。

Pankaj Joshi、Aditya Gupta、Pankaj Kumar 和 Manas Sisodia。面向包含文本、表格和图像的文档的稳健多模型检索增强生成流程。载于 *2024*年第*3*届应用人工智能与计算国际会议（*ICAAIC*），第 993–999 页。IEEE，2022年。

Meng-Chieh Lee、Qi Zhu、Costas Mavromatis、Zhen Han、Soji Adeshina、Vassilis N Ioannidis、Huzefa Rangwala 和 Christos Faloutsos。Agent-g：用于图检索增强生成的智能体框架。

Paul Lerner、Olivier Ferret、Camille Guinaudeau、Hervé Le Borgne、Romaric Besançon、José G Moreno 和 Jesús Lovón Melgarejo。ViQuAE：一个关于命名实体的基于知识的视觉问答数据集。载于 第*45*届*ACM SIGIR*信息检索研究与开发国际会议论文集，第 3108–3120 页，2022年。

Feng Li、Renrui Zhang、Hao Zhang、Yuanhan Zhang、Bo Li、Wei Li、Zejun Ma 和 Chunyuan Li。Llava-next-interleave：在大型多模态模型中应对多图像、视频和三维问题。*arXiv* 预印本 *arXiv:2407.07895*，2024年。

Xiaoxi Li、Guanting Dong、Jiajie Jin、Yuyao Zhang、Yujia Zhou、Yutao Zhu、Peitian Zhang 和 Zhicheng Dou。Search-o1：搜索增强型大型推理模型中的智能体式搜索。*arXiv* 预印本 *arXiv:2501.05366*，2025a。

Yangning Li、Yinghui Li、Xinyu Wang、Yong Jiang、Zhen Zhang、Xinran Zheng、Hui Wang、Hai-Tao Zheng、Fei Huang、Jingren Zhou 和 Philip S. Yu。使用动态视觉问答数据集和自适应规划智能体对多模态检索增强生成进行基准测试。载于 第十三届国际学习表征会议，2025b年。网址 https://openreview.net/forum?id=VvDEuyVXkG。

Weizhe Lin 和 Bill Byrne。利用外部知识进行检索增强视觉问答。*arXiv* 预印本 *arXiv:2210.03809*，2022年。

Weizhe Lin、Jinghong Chen、Jingbiao Mei、Alexandru Coca 和 Bill Byrne。用于检索增强视觉问答的细粒度后期交互多模态检索，2023年。网址https://arxiv.org/abs/2309.17133。

Ben Liu、Jihai Zhang、Fangquan Lin、Cheng Yang、Min Peng 和 Wotao Yin。Symagent：一种用于在知识图谱上进行复杂推理的神经符号自学习智能体框架。发表于*ACM Web Conference 2025*论文集，第98–108页，2025a。

Zhenghao Liu、Xingsheng Zhu、Tianshuo Zhou、Xinyi Zhang、Xiaoyuan Yi、Yukun Yan、Yu Gu、Ge Yu 和 Maosong Sun。在多模态语境下对检索增强生成进行基准测试。*CoRR*，abs/2502.17297，2025b。doi：10.48550/ARXIV.2502.17297。网址 https://doi.org/10.48550/arXiv.2502.17297。

Zhenghao Liu、Xingsheng Zhu、Tianshuo Zhou、Xinyi Zhang、Xiaoyuan Yi、Yukun Yan、Ge Yu 和 Maosong Sun。多模态语境下检索增强生成的基准测试。*arXiv* 预印本 *arXiv:2502.17297*，2025c。

Kenneth Marino、Mohammad Rastegari、Ali Farhadi 和 Roozbeh Mottaghi。OK-VQA：一项需要外部知识的视觉问答基准。在*IEEE/CVF* 计算机视觉与模式识别会议论文集中，第 3195–3204 页，2019年。

Zhenyu Pan、Haozheng Luo、Manling Li 和 Han Liu。Chain-of-Action：通过大语言模型实现忠实的多模态问答。*arXiv* 预印本 *arXiv:2403.17359*，2024年。

David Rein、Betty Li Hou、Asa Cooper Stickland、Jackson Petty、Richard Yuanzhe Pang、Julien Dirani、Julian Michael 和 Samuel R Bowman。《Gpqa：研究生级谷歌防作弊问答基准》。载于首届语言建模会议，2024年。

Zhili Shen、Chenxin Diao、Pavlos Vougiouklis、Pascual Merita、Shriram Piramanayagam、Enting Chen、Damien Graux、Andre Melo、Ruofei Lai、Zeren Ji ang 等。《Gear：用于检索增强生成的图增强智能体》。arXiv 预印本 arXiv:2412.18431，2024年。Aditi Singh、Abul Ehtesham、Saket Kumar 和 Tala Talaei Khoei。《智能体检索增强生成：智能体RAG综述》。arXiv 预印本 arXiv:2501.09136，2025年。Gemini Team、Rohan Anil、Sebastian Borgeaud、Jean-Baptiste Alayrac、Jiahui Yu、Radu Soricut、Johan Schalkwyk、Andrew M Dai、Anja Hauth、Katie Millican 等。《Gemini：一系列能力强大的多模态模型》。arXiv 预印本 arXiv:2312.11805，2023年。Harsh Trivedi、Niranjan Balasubramanian、Tushar Khot 和 Ashish Sabharwal。《♫musique：通过单跳问题组合构建多跳问题》。计算语言学协会会刊，10：539–554，2022年。Liang Wang、Haonan Chen、Nan Yang、Xiaolong Huang、Zhicheng Dou 和 Furu Wei。《检索增强生成链》。ArXiv，abs/2501.14342，2025a。网址 https://api.semanticscholar.org/CorpusID:275906944。Weiyun Wang、Zhangwei Gao、Lixin Gu、Hengjun Pu、Long Cui、Xingguang Wei、Zhaoyang Liu、Linglin Jing、Shenglong Ye、Jie Shao 等。《Internvl3. 5：提升开源多模态模型的通用性、推理能力和效率》。arXiv 预印本 arXiv:2508.18265，2025b。Wenya Wang 和 Sinno Pan。《面向多跳阅读理解的深度归纳逻辑推理》。载于计算语言学协会第60届年会论文集（第1卷：长篇论文），第4999–5009页，2022年。Yiqi Wang、Wentao Chen、Xiaotian Han、Xudong Lin、Haiteng Zhao、Yongfei Liu、Bohan Zhai、Jianbo Yuan、Quanzeng You 和 Hongxia Yang。《探索多模态大语言模型（MLLM）的推理能力：多模态推理新兴趋势综述》。arXiv 预印本 arXiv:2401.06805，2024年。Peng Xia、Peng Xia、Kangyu Zhu、Haoran Li、Haoran Li、Weijia Shi、Sheng Wang、Linjun Zhang、James Zou 和 Huaxiu Yao。《Mmed-rag：面向医学视觉语言模型的多功能多模态检索增强生成系统》。ArXiv，abs/2410.13085，2024年。网址 https://api.semanticscholar. org/CorpusID:273403866。Zhilin Yang、Peng Qi、Saizheng Zhang、Yoshua Bengio、William W Cohen、Ruslan Salakhutdinov 和 Christopher D Manning。《Hotpotqa：用于多样化、可解释多跳问答的数据集》。arXiv 预印本 arXiv:1809.09600，2018年。Han Zhang、Langshi Zhou 和 Hanfang Yang。《通过主动自我反思学习在知识图谱上进行检索与推理》。arXiv 预印本 arXiv:2502.14932，2025年。Penghao Zhao、Hailin Zhang、Qinhan Yu、Zhengren Wang、Yunteng Geng、Fangcheng Fu、Ling Yang、Wentao Zhang、Jie Jiang 和 Bin Cui。《面向人工智能生成内容的检索增强生成：综述》。arXiv 预印本 arXiv:2402.19473，2024年。Xu Zheng、Ziqiao Weng、Yuanhuiyi Lyu、Lutao Jiang、Haiwei Xue、Bin Ren、Danda Paudel、Nicu Sebe、Luc Van Gool 和 Xuming Hu。《视觉领域的检索增强生成与理解：综述及新展望》。arXiv 预印本 arXiv:2503.18016，2025a。Yaowei Zheng、Richong Zhang、Junhao Zhang、Yanhan Ye、Zheyan Luo、Zhangchi Feng 和 Yongqiang Ma。《Llamafactory：100+ 语言模型的统一高效微调》。载于计算语言学协会第62届年会论文集（第3卷：系统演示），泰国曼谷，2024年。计算语言学协会。网址 http://arxiv.org/abs/2403.13372。Yuxiang Zheng、Dayuan Fu、Xiangkun Hu、Xiaojie Cai、Lyumanshan Ye、Pengrui Lu 和 Pengfei Liu。《Deepresearcher：通过真实环境中的强化学习扩展深度研究》。arXiv 预印本 arXiv:2504.03160，2025b。Junjie Zhou 、 Zheng Liu 、 Ze Liu 、 Shitao Xiao 、 Yueze Wang、Bo Zhao、Chen Jason Zhang、Defu Lian 和 Yongping Xiong。《Megapairs：用于通用多模态检索的大规模数据合成》。arXiv 预印本 arXiv:2412.14475，2024年。

附录

目录

A 大语言模型的使用 15

B 相关工作 15
B.1 多模态检索增强生成基准.15 B.2 智能体式搜索增强推理.15 B.3 结构感知基准.15

C D数据示例 16
C.1 图像输入 Trained Chain.17 C.2 文本发起链.17 C.3 文本链.17 C.4 并行视觉-文本分叉.18 C.5 多图像分支.19 6

D Baseline RAG性 erf 20
能D.1 单步 RAG 基线.20 D.2 传统多模态 RAG 基线.21

E 数据集统计 22

F 数据集质量验证 22

G 跨模型与人工一致性分析 23G.1 跨模型 LLM 评审一致性.23 G.2 人工与 LLM 评审的一致性.23 G.3 跨模型真实标注一致性.24

H SEARCH-ALIGN 的细粒度改进分析 24

I SEARCH-ALIGN 微调细节 25

J 基于阈值语义匹配的软HPS评估 26

K 在 Top-k 检索下的评估 26K.1 Retrieval@K 下的检索不足与过度检索.26 K.2 Retrieval@K 下的 SEARCH-ALIGN.27

L AgenticMM-RAG 案例研究 28L.1 成功案例.28 L.2 失败案例.29

M 拓扑设计的理论依据 30
M.1 问题形式化.30 M.2 穷举证明.30 M.3 结论.30

N 数据构建提示 31
N.1 图像发起链.31 N.2 并行图文分叉.31 N.3 文本发起链.32 N.4 多图像分支.32 N.5 文本链.32

O AgenticMM-RAG流水线的提示模板 34

P LLM 评审的提示模板 35

A 大语言模型的应用

在撰写本文期间，我们对大语言模型，特别是 ChatGPT，进行了受控使用，将其作为辅助写作工具。大语言模型仅用于润色文风，即改善作者最初撰写段落的流畅度、语法和可读性。重要的是，本文的科学内容、方法论、实验设计和主要叙述均由作者独立构思、撰写并验证，未依赖大语言模型。因此，大语言模型纯粹用于辅助润色作者撰写的文本，其贡献尚不足以达到共同署名的程度。

B 相关工作

我们从三个互补的视角组织相关工作：(1) 将检索范围扩展到文本之外的多模态检索增强生成基准，(2) 将检索视为序贯决策过程的智能体RAG系统，以及 (3) 结构感知基准。

B.1 多模态检索增强生成基准测试

多模态检索增强生成通过纳入图像证据，突破文本范畴，为知识密集型跨模态推理提供支持（Xia 等人，2024年；Chang等，2022年；Pan 等人，2024年）。现有基准虽迈出了重要的第一步，但仍有局限：多数采用简单的问答格式和固定的先检索后生成流程（Marino等，2019年；Hu 等人，2025年；Chang等，2022年；Amirshahi 等人，2025年；Friel 等人，2024年），视觉搜索能力有限，将图像信息简化为图像说明（Chang等，2022年；Lerner等，2022年；Chen等人，2023年），在视觉问答式任务中将推理限制为 1–2 跳检索（Li 等人，2025b；Liu 等人，2025c），且缺少逐步标注或显式推理结构。这些限制妨碍了我们判断多模态大语言模型究竟是在进行结构化多模态推理，还是仅仅利用捷径。为此，我们提出 MC-SEARCH，它提供经 HAVE 验证、涵盖多种推理拓扑结构的长程轨迹，从而能够对多模态大语言模型的长时程搜索增强推理进行细粒度归因和全面分析。
</output>

B.2 搜索智能体增强推理

近期研究将检索重新定义为一种序贯决策过程，由此催生了智能体式检索增强生成系统。这类系统会分解查询、自适应地检索证据，并从检索到的上下文中综合知识（Wang 等人，2025a；Li 等人，2025a；Zheng 等人，2025b；Singh 等人，2025年；Pan 等人，2024年）。监督微调（Li 等人，2025a）和强化学习方法（Chen等人，2025年；Jin 等人，2025年）进一步使模型行为与真实推理轨迹对齐，以增强可解释性和依据忠实度。然而，这些进展仍主要局限于纯文本，通常针对多跳问答（Trivedi 等人，2022年）或科学领域（Rein 等人，2024年），因此缺乏对多模态信号如何影响多模态大语言模型中的检索规划和逐步监督，以及不同搜索增强推理拓扑结构如何影响模型行为的系统研究。为弥补这一空白，我们推出了一个多模态基准，涵盖多样的搜索增强推理结构、经逐步验证的长推理链及过程级指标，揭示了模态错配规划、检索不足与过度检索以及错误分类体系等挑战，并提出 SEARCH-ALIGN，通过逐步监督提升规划能力和检索保真度。

B.3 结构感知基准

另一项相关研究方向考察了在符号知识库或图上运行的推理模型，这类模型通常使用模式约束查询或专用工具（Lee 等人；Shen 等人，2024年；Zhang 等人，2025年；Wang 和 Pan，2022年；Feng 等人，2020年；Liu 等人，2025a）。这些研究增强了符号落地能力，但主要关注检索什么。相比之下，对智能体多模态检索增强生成进行基准测试，需要对检索链的结构本身进行监督，涵盖启动模态、跳转顺序以及并行与串行模式。据我们所知，尚无多模态数据集提供此类结构化的逐步标注。我们的 Search-MM 填补了这一空白，提供标准推理链及模态特定检索步骤，从而实现细粒度归因、对检索过度或不足的分析，以及对多种多模态推理结构的评估。

C 数据示例

C.1 图像发起链

示例推理链 — 图像发起链



图像说明：《圣母领报》中的大天使加百列 - Google 艺术计划

主问题。图像中描绘的是哪一幕以大天使加百列为主题的圣经场景？这幅画目前收藏在哪里？它于何时、以多少钱购入并纳入收藏？此次购藏对于这位艺术家在美国博物馆界的地位具有怎样的历史意义？

最终答案。这幅画描绘了圣母领报场景，现藏于费城艺术博物馆，于*1899*年*4*月*5*日以*1,750*美元购得，是坦纳的第一幅被美国博物馆购藏的作品。

推理链（4 步）：

1. **（图像）**图像中描绘了哪个有大天使加百列出现的圣经场景？图像证据 *ID*：30332773 答案：《圣母领报》。 2. **（文本）**哪家博物馆收藏了这幅画作？文本证据 *ID*：d5be2ae00dba11ecb1e81171463288e9 0_ 答案：费城艺术博物馆。 3. **（文本）**这幅画作于何时、以什么价格购入并纳入收藏？文本证据 *ID*：d5be2ae00dba11ecb1e81171463288e9 1_ 答案：1899年4月5日，以1,750美元购入。 4. **（文本）**这次购藏对于这位艺术家在美国博物馆领域有何历史意义？文本证据 *ID*：d5be2ae00dba11ecb1e81171463288e9 8_答案：这是坦纳的第一件被美国博物馆购藏的作品。

C.2 文本发起链

推理链示例——文本发起链



图像说明： 索尔万加利福尼亚州风车。

主问题。图像是否从视觉上证实了文本中所描述的索尔万独特的文化特色?

最终答案。 是的，图像中突出显示了一座大型风车和一些具有鲜明欧洲/丹麦建筑风格的建筑，这与将索尔万描述为加利福尼亚州的“丹麦一隅”相符。

推理链（2 个步骤）：

1. **(文本)**根据文本，索尔万具有怎样的文化特色? 文本证据 *ID*: d5bd142a0dba11ecb1e81171463288e9 7_答案: 索尔万被描述为 “ 坐落在这片美丽的加利福尼亚州中的一小块丹麦”。 2. **(图像)** 图像是否直观地印证了这一文化特色? 如果是，是如何印证的? 图像证据 *ID*：30326437答案： 是的，图像中醒目地展示了一座大型风车和具有鲜明欧洲／丹麦建筑风格的建筑，印证了对其文化特色的描述。

C.3 文本链

示例推理链 — 文本链

主问题。 宾利慕尚有哪些关键特征，包括其生产时期、发动机规格、名称由来和著名特别版?

最终答案。 宾利慕尚是一款于*2010*年至*2020*年生产的全尺寸豪华轿车，其名称源自勒芒赛道的慕尚弯。它搭载*6.75*升宾利*L*系列*V8*发动机，其中一个著名特别版是*“W.O.*特别版*”*，该版本配有一块取自W.O.宾利私人座驾曲轴的碎片。

推理链（3 步）：

1. **(文本)** 宾利慕尚是什么类型的汽车? 它于何时生产? 其名称由来是什么? 文本证据 *ID*：d5bd6ace0dba11ecb1e81171463288e9 15_答案： 宾利慕尚是一款全尺寸豪华轿车，生产于2010年至2020年，其名称源自勒芒赛道的慕尚弯。 2. **(文本)** 宾利慕尚的发动机规格是什么? 文本证据 *ID*：d5bd6ace0dba11ecb1e81171463288e9 6_ 答案： 慕尚搭载一台6.75升（6750立方厘米/411立方英寸）的宾利L系列V8发动机，并经过改装以符合欧V排放法规。 3. **(文本)** 慕尚推出了哪款值得关注的特别版? 它有什么独特之处? 文本证据 *ID*：d5bd6ace0dba11ecb1e81171463288e9 11_ 答案： 慕尚“W.O. Edition”特别版正式亮相，其独特之处是在扶手中展示了一块取自W.O.宾利私人座驾曲轴的碎片。

C.4 并行视觉-文本分叉

推理链示例——并行视觉-文本分叉



图像说明： 曼陀林1。

主问题。 鉴于图中曼陀林的外观，Pasquale Vinaccia于1835年在弦材方面的创新如何促使乐器结构发生变化，并对曼陀林设计产生了持久影响?

最终答案。 Pasquale Vinaccia 于 1835年 创新性地采用钢丝弦，这些琴弦在曼陀林上清晰可见，因此必须加固琴体 并加深琴碗，以增强共鸣；这种上弦方式最终成为曼陀林的主流方式。

推理链（4 个步骤）：

1. （**图像**）图中所示的曼陀林上可以看到哪种琴弦? 这一视觉细节与 Pasquale Vinaccia 对乐器的历史性改良有何关联? 图像证据编号：30204742答案：曼陀林上醒目地装有金属／钢弦。” 2.（**文本**）由于采用了这些新型金属弦，曼陀林琴身随后需要进行哪些结构改动? 文本证据编号：d5be68020dba11ecb1e81171463288e9 15_答案：金属弦使琴身必须加固，琴碗也需要加深。 3.（**文本**）由于这些结构变化，加深曼陀林琴碗增强了哪种具体的声学特性? 文本证据编号：d5be68020dba11ecb1e81171463288e9 3_答案：加深琴碗增强了音色共鸣。 4.（**文本**）综合这些改良来看，Vinaccia 采用钢弦这一决定对曼陀林设计与制作产生了怎样的长期影响? 文本证据编号：d5be68020dba11ecb1e81171463288e9 5_答案：他采用钢弦的做法成为曼陀林上弦的主流方式。

C.5 多图像分支

示例推理链 — 多图像分支



图像说明： 加拿大馆



图像说明： 摩洛哥馆

主问题。图中所示的未来世界展馆中，加拿大馆和摩洛哥馆哪一座以高耸纤细的塔楼为醒目的建筑中心？这座展馆在加入世界之窗一事上有何历史意义？

最终答案。 摩洛哥馆以一座醒目的高耸纤细塔楼为特色，在历史上也具有重要意义，因为它是世界之窗增建的第一个扩建展馆，于1984年9月7日开放。

推理链（4 步）：

1. **（图像）**主建筑加拿大馆的屋顶结构最显著的特点是什么？图像证据 *ID*：30383576答案：陡峭的多层绿色屋顶，顶部有尖塔。 2. **（图像）**摩洛哥馆建筑天际线最显著的建筑特色是什么？图像证据 *ID*：30021436答案：一座高耸的矩形塔楼，顶部有一个小圆顶。 3. **（图像）**对比加拿大馆的陡峭屋顶与摩洛哥馆的塔楼这两种主要垂直建筑元素，这两座展馆中，哪一座更适合描述为以醒目的塔楼而非陡峭屋顶为特色？图像证据 *ID*：30021436 答案：摩洛哥馆。 4. **（文本）**根据所提供的文本，关于摩洛哥馆作为世界之窗新增展馆的历史地位，有什么重要史实？文本证据 *ID*：d5bef66e0dba11ecb1e81171463288e9 7_ 答案：它是世界之窗新增的第一个扩建展馆，于1984年9月7日开放。

D BaselineRAG 性能

D.1 单步 RAG 基线

表 5 报告了**单步检索**设置下固定步数 RAG 基线的结果。在所有实验中，我们使用 `BAAI/bge-large-en-v1.5`（Zhou 等人，2024年）嵌入模型进行密集检索。如果输入问题包含图像，系统会执行一次图像检索以获取相应的图像说明，然后将其与查询拼接后输入模型。如果输入内容纯为文本，则使用查询本身进行一次文本检索。在这两种情况下，系统都只选取排名第一的检索段落作为上下文，并将其附加到输入提示词中。
为确保与智能体多模态 RAG 进行公平比较，我们采用相同的评估策略：只有在模型有足够把握时，才允许其输出答案；否则必须弃答。这一限制避免通过强迫模型猜测来人为提高准确性，也反映了模型在知识依据场景中的实际可靠性。

我们的结果揭示了不同模型家族之间的鲜明差异。更稳健的智能体模型（例如专有模型 Gemini-2.5-Pro）往往较为谨慎：当仅提供一个检索到的知识片段时，它们通常会判定证据不足，并选择不作答。相比之下，大多数开源模型（例如 InternVL-3.5-Pro、Qwen2.5-VL-7B）则表现得过于自信，即使检索到的上下文十分有限，也会尝试生成完整答案。尽管这种行为在固定步数评估下带来了表面上更高的准确性，却也暴露了固定式 RAG 范式的弱点：依赖单一证据片段往往无法支撑复杂推理，因而产生浅显或缺乏依据的答案。

表 5：在单步 RAG 设置下对 MC-Search 基准上的各种模型进行评估。

推理图	模型	答案准确率			链对齐		黄金标准 F1 (↑)
		F1 (↑)	ΔF1 (↑)	LJ (↑)	HPS (↑)	展开偏差 (↓)	
纯文本链	GPT-4o-Mini	27.50	**27.48**	1.69	19.22	3.00	61.90
	Gemini-2.5-Flash	10.88	10.81	0.75			**67.72**
	Gemini-2.5-Pro	11.11	11.10	0.70			62.42
	Claude-3.7-Sonnet	22.71	22.09	1.69			66.38
	InternVL3.5-8B	30.98	24.00	1.47			65.72
	Qwen2.5-VL-7B	**35.26**	26.61	**1.81**			61.57
图像发起链	GPT-4o-Mini	29.37	**27.06**	0.83	21.52	3.00	68.29
	Gemini-2.5-Flash	3.63	-3.09	0.22			72.39
	Gemini-2.5-Pro	2.34	-2.51	0.13			69.83
	Claude-3.7-Sonnet	8.29	3.58	1.18			**72.62**
	InternVL3.5-8B	21.10	11.48	0.77			63.86
	Qwen2.5-VL-7B	**40.31**	22.66	**1.62**			60.95
多图像分支	GPT-4o-Mini	25.53	20.69	1.27	15.38	3.00	59.46
	Gemini-2.5-Flash	6.49	1.58	0.56			**64.40**
	Gemini-2.5-Pro	7.26	3.47	0.51			61.29
	Claude-3.7-Sonnet	12.70	10.77	1.36			62.41
	InternVL3.5-8B	33.11	**21.01**	1.59			57.65
	Qwen2.5-VL-7B	**36.96**	18.36	**1.89**			57.04
并行图文分叉	GPT-4o-Mini	15.20	14.87	0.77	15.06	3.00	53.98
	Gemini-2.5-Flash	4.09	2.11	0.30			57.99
	Gemini-2.5-Pro	4.04	3.40	0.32			53.46
	Claude-3.7-Sonnet	9.61	7.83	1.22			58.51
	InternVL3.5-8B	20.63	13.94	0.99			**58.82**
	Qwen2.5-VL-7B	**29.70**	**21.08**	**1.57**			55.76
文本发起链	GPT-4o-Mini	24.67	13.02	1.43	26.86	3.00	49.47
	Gemini-2.5-Flash	16.78	-0.43	1.30			**66.27**
	Gemini-2.5-Pro	14.51	-0.90	1.22			55.94
	Claude-3.7-Sonnet	17.63	-1.11	1.80			57.13
	InternVL3.5-8B	28.82	15.96	1.38			38.70
	Qwen2.5-VL-7B	**42.97**	**25.84**	**2.48**			39.86

D.2 传统多模态检索增强生成基线

为提供更有力的固定步数基线，我们将单步 RAG 设置扩展为两跳变体。对于多模态查询，我们遵循标准的多模态 RAG 流程（Lin 等人，2023年；Joshi 等人，2024年；Lin 和 Byrne，2022年；Zheng 等人，2025a）：首先，我们执行图像检索以获取输入图像的图像说明。随后，将检索到的图像说明与原始查询拼接起来，并使用这段组合文本进行第二步文本检索。对于纯文本查询，我们直接连续执行两轮文本检索。在这两种情况下，最终检索到的段落（top-1）都会作为上下文附加到多模态大语言模型的输入中。

与单步基线一样，我们要求模型仅在有足够把握时才输出答案，以确保比较公平，并避免因强制作答而导致准确性虚高。表6展示了固定两跳 RAG 基线的结果。与单步检索（表5）相比，两跳检索在各种推理结构中均带来了稳定提升，尤其是在多图像分支和文本发起链等多模态情形下。这证实了，增加检索步骤能够通过提供更多上下文，部分缓解单跳检索的信息瓶颈。

然而，其性能仍落后于智能体多模态 RAG。虽然固定的两跳检索能提供更多证据，但它无法灵活调整检索次数，也无法有策略地整合跨模态知识。相比之下，智能体方法会有选择地规划多个跳，并决定何时停止，从而生成更忠实的推理链，并显著提高总体准确率。这些结果既凸显了固定两跳 RAG 的效用，也揭示了它的局限：它优于单步检索，但仍不如自适应智能体多模态 RAG 灵活可靠。

表6：各模型在**MC-Search**基准下采用**传统多模态检索增强生成**时的评估

推理图	**模型**	**答案准确率**			**链对齐**		**黄金标准 F1 (↑)**
		F1 (↑)	ΔF1 (↑)	LLJ (↑)	HPS (↑)	展开偏差 (↓)	
图像发起链	GPT-4o-Mini	35.61	**33.30**	1.58	21.11	2.42	68.29
	Gemini-2.5-Flash	7.42	0.70	0.32			72.39
	Gemini-2.5-Pro	6.06	1.21	0.25			69.83
	Claude-3.7-Sonnet	21.72	21.10	1.72			**72.62**
	InternVL3.5-8B	31.26	21.64	1.24			63.86
	Qwen2.5-VL-7B	**39.59**	**21.94**	**1.79**			60.95
多图像分支	GPT-4o-Mini	27.88	**23.04**	1.04	19.35	2.76	59.46
	Gemini-2.5-Flash	10.29	5.38	0.56			**64.40**
	Gemini-2.5-Pro	11.03	7.24	0.57			61.29
	Claude-3.7-Sonnet	19.96	15.25	1.67			62.41
	InternVL3.5-8B	34.55	22.45	1.58			57.65
	Qwen2.5-VL-7B	**36.61**	18.01	**1.97**			57.04
并行图文分叉	GPT-4o-Mini	19.68	19.35	1.53	10.87	3.18	53.98
	Gemini-2.5-Flash	5.19	3.21	0.26			57.99
	Gemini-2.5-Pro	4.03	3.39	0.21			53.46
	Claude-3.7-Sonnet	15.80	13.87	1.45			58.51
	InternVL3.5-8B	23.83	17.14	1.14			**58.82**
	Qwen2.5-VL-7B	**30.16**	**21.54**	**1.61**			55.76
文本发起链	GPT-4o-Mini	28.23	16.58	1.53	26.72	1.90	49.47
	Gemini-2.5-Flash	20.51	3.30	1.19			**66.27**
	Gemini-2.5-Pro	17.60	2.19	1.03			55.94
	Claude-3.7-Sonnet	25.55	23.77	2.20			57.13
	InternVL3.5-8B	30.80	17.94	1.47			38.70
	Qwen2.5-VL-7B	**41.80**	**24.67**	**2.47**			39.86

E 数据集统计

表 7 详细列出了 MC-Search 基准的构成。该数据集共包含 3,333 个多模态推理样本，涵盖五种具有代表性的推理结构。其中，图像发起链占比最大（1,306 个样本），其次是文本链（945 个样本）和并行视觉-文本分叉（680 个样本）。尽管规模较小，文本发起链（169 个样本）和多图像分叉（233 个样本）也带来了颇具挑战性的案例，需要显式进行跨模态或多图像整合。

该基准以一个大规模知识库为基础，其中包含近 39 万张图像和 78 万篇文本文档，确保涵盖两种模态的广泛内容。问题平均长度为 33 个词元，而答案更长且包含丰富的实体（平均 47 个词元、 8 个实体），体现了任务的复杂性。

在推理深度方面，不同结构的平均跳数有所差异，从文本发起链的 3.17 跳到并行视觉-文本分叉的约 4 跳不等，体现了推理复杂度的多样性。在所有样本中，大多数跳依赖文本检索（79.7%），而图像检索约占五分之一（20.2%），凸显了模态分布均衡但以文本为主的特点。

表 7：MC-Search 基准统计数据。

统计项	**数量**
并行视觉-文本分叉样本	680
图像发起链样本	1,306
文本链样本	945
文本发起链样本	169
多图分叉样本	233
总样本	3,333
知识库中的图像总数	389,750
知识库中的文档总数	784,473
平均问题长度	33.23
平均答案长度	47.40
平均答案实体数	8.03
并行视觉-文本分叉中的平均跳数	4.01
图像发起链中的平均跳数	3.82
文本链中的平均跳数	3.65
文本发起链中的平均跳数	3.17
多图分叉中的平均跳数	3.97
文本检索跳数	10,063 (79.72%)
图像检索跳数	2,550（20.20%）

F 数据集质量验证

为评估数据集质量，我们使用 **Gemini-2.5-Pro** 从四个维度为每条标准推理链评分：(i) 事实正确性、(ii) 步骤必要性、(iii) 清晰度，以及 (iv) 多模态对齐。评分为 1（较差）至 5（完美）的整数，总分为四项评分的平均值。表8总结了结果。

表8：Gemini-2.5-Pro 平均评估分数。

维度	**平均分**
事实正确性	4.93
步骤必要性	4.70
清晰度	4.88
多模态对齐	4.95
总体（均值）	4.87

G 跨模型与人类一致性分析

我们还开展了一项一致性研究，以评估 LLM 评审（大语言模型裁判评分，LLJ）评估的稳定性。异构模型系列与人类专家之间的一致性表明，LLJ 评分能够捕捉客观且普遍适用的质量标准，且未观察到其依赖于我们实验中使用的主要评审模型（Gemini-2.5-Pro）偏好的推理模式。

G.1 跨模型 LLM 评审一致性

我们从表 3 中由 Gemini-2.5-Pro 主干模型生成的预测结果中，随机抽取五个互不重叠的子集（每个占 10%）。每个子集均由三个独立的**异构模型系列**重新评分：Gemini-2.5-Pro、GPT-5 和 Qwen-VL-Plus。对于每一对评审器，我们计算皮尔逊相关系数、斯皮尔曼秩相关系数、Bland–Altman 平均偏差以及一致性界限（LoA）。表 9 报告了五个子集中一致性得分的均值和方差。

表 9：跨模型 LLJ 一致性得分，并报告均值（方差）。较高的相关性和较小的平均偏差表明不同评估器系列之间具有较强的一致性。

评估器对	皮尔逊相关系数	斯皮尔曼相关系数	平均偏差	LoA
Gemini / GPT-5	**0.9330** (0.00076)	**0.8720** (0.00120)	+0.0288（0.00051）	0.8438 (0.07937)
Gemini / Qwen	**0.9147** (0.00116)	**0.8145** (0.00360)	−0.0083 (0.00082)	0.9295 (0.08834)
GPT-5 / Qwen	**0.9147** (0.00131)	**0.8259** (0.00374)	−0.0371 (0.00212)	0.9341 (0.09565)

所有评估器配对的皮尔逊相关系数均高于 0.91，斯皮尔曼相关系数均高于 0.81。平均偏差的幅度较小（< 0.04），且一致性界限范围窄而稳定。这些结果表明，LLJ评估器能够针对具有不同训练数据和架构的模型家族给出一致的评估。值得注意的是，尽管 Gemini-2.5-Pro 是我们主要评估中的首要评审模型，但评分并未表现出偏向其生成答案的倾向。

为进一步说明这一点，我们提供了 Gemini-2.5-Pro 与 GPT-5 在所有评分维度（准确性、对齐、覆盖度、连贯性和实体）上的六边形分箱相关图。分布紧密集中在对角线周围，证实 LLJ 评分遵循**客观的、可泛化评估标准**，而非特定于模型的模式。

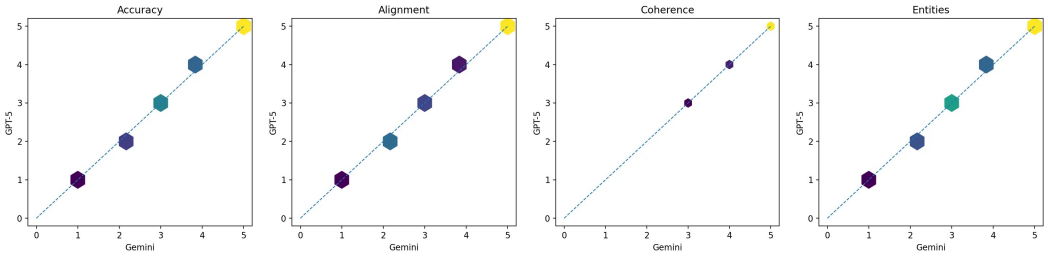


图6：Gemini-2.5-Pro 与 GPT-5 在四个 LLJ维度上的评分六边形分箱图。

G.2 人类与 LLM 评审的一致性

此外，我们还将 LLJ 分数与计算机科学博士专家对 100 个随机抽取的预测所做的人工标注进行比较。人类依据四个 LLJ维度为每个样本评分。表格 10 报告了由 Gemini-2.5-Pro 主干模型生成的预测中，人工评分与 Gemini-2.5-Pro 评分之间的相关性和 Bland–Altman统计量。

总体而言，LLJ 分数与人类在所有基于内容的维度上的判断都高度吻合，在准确性、实体覆盖度和堆叠分数上的皮尔逊相关系数均高于 0.92。即使对于主观性更强的连贯性维度，排名一致性仍然很高（斯皮尔曼相关系数 $= 0.82$），这证实 LLJ 评估能够提供可靠且与人类判断一致的质量评估。

表格 10：人类与 LLJ 的一致性得分。

维度	皮尔逊相关系数	斯皮尔曼相关系数	平均偏差	LoA
准确性	**0.9632**	**0.9556**	+0.0800	0.6606
实体覆盖度	**0.9375**	**0.8691**	+0.1200	0.9309
连贯性	0.6061	**0.8200**	−0.0200	0.6778
对齐	**0.9275**	0.7925	+0.2750	0.9982
堆叠分数	**0.9323**	**0.8675**	+0.1138	0.8562

G.3 跨模型真实标签一致性

为考察智能体式多模态检索增强生成流程是否依赖参考推理链中特定于模型的措辞或句法模式，我们使用**由不同模型系列生成的真实标注**构建了一个替代基准。对于五种图拓扑中的每一种，我们随机抽取 50 个样本，并使用 GPT-5 按照第 2 节中的构建流程重新生成所有中间子答案和最终答案。这些 GPT-5 推理链构成另一组真实标注。

随后，我们在完整的智能体式多模态检索增强生成流程下评估 Gemini-2.5-Pro 和 GPT-4o-mini。如表 11 所示，两种模型在 GPT 生成和 Gemini 生成的真实标注上均表现出高度**一致的性能** 。在所有拓扑结构中，F1 和 LLJ 分数彼此密切吻合，模型之间的性能排序也保持稳定。

这些结果表明，该流程对于生成参考推理链所用源模型的变化具有稳健性。评估反映的是真正的多跳推理能力，而非依赖可能使特定源模型系列占优的风格或句法特征。

表 11：Gemini-2.5-Pro 和 GPT-4o-mini 在评估 GPT-5 生成的真实推理链（GPT 列）或原始的基于 Gemini 构建的推理链（Gemini 列）时的表现。

图类型	模型	F1 分数（↑）		LLJ 分数（↑）	
		GPT	Gemini	GPT	Gemini
文本链	Gemini-Pro	32.52	34.42	3.44	3.46
	GPT-4o-mini	26.05	26.98	3.01	3.05
图像发起链	Gemini-Pro	41.37	46.57	3.82	3.81
	GPT-4o-mini	35.01	38.96	3.24	3.20
多图像分支	Gemini-Pro	35.49	38.04	3.48	3.66
	GPT-4o-mini	17.97	18.77	2.69	2.80
并行图文分叉	Gemini-Pro	36.59	38.44	3.64	3.67
	GPT-4o-mini	21.09	21.97	2.58	2.54
文本发起链	Gemini-Pro	34.91	43.71	3.42	3.82
	GPT-4o-mini	22.95	26.01	2.95	3.03

H SEARCH-ALIGN细粒度改进分析

为考察 SEARCH-ALIGN 如何改变多模态智能体的过程行为，我们使用涵盖规划、检索和模态选择行为的细粒度错误分类体系，分析 InternVL-3.5-8B 在对齐前后的表现。对于每个样本，外部 LLM 会标注四种二元错误类型：(i)检索失败、(ii)步骤遗漏、(iii)顺序/依赖错误和 (iv)模态不匹配。我们报告 MC-SEARCH 验证集划分中出现各类错误的样本比例。

图7总结了结果。SEARCH-ALIGN使四类错误均持续减少，其中模态选择和多跳

完整性方面的改进最为显著。模态不匹配从 65.75% 降至 34.72%，表明模型学会了更恰当地选择视觉证据或文本证据。步骤遗漏也显著减少（从 86.15% 降至 73.66%），这表明智能体能够遵循更完整的推理链，较少跳过中间步骤。检索失败和顺序/依赖错误虽降幅较小，但减少趋势稳定，反映出证据获取和跳数结构化方面有所改善。

总体而言，这一过程级分析表明，SEARCH-ALIGN并非仅仅鼓励对标注推理链的模仿；相反，它在规划与多模态决策的多个维度上带来了可衡量的改进。这些结构化错误模式的减少，与答案准确率和对齐指标上报告的定量提升相辅相成。

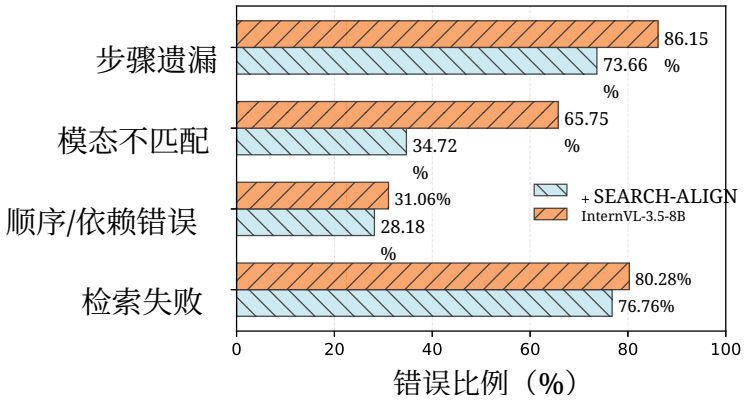


图7：SEARCH-ALIGN前后四个过程级类别的错误比例。

I SEARCH-ALIGN 微调详情

为使多模态大语言模型（多模态大语言模型）的推理轨迹与黄金搜索链保持一致，我们使用 `Llama-factory` 框架进行监督微调（Zheng 等人，2024年）。所有实验均在 4 块 NVIDIA A100 GPU（每块显存 80GB）上进行。我们报告了本研究所用两种开源骨干模型的详细超参数配置，分别为 InternVL3.5-8B 和 Qwen/Qwen2.5-VL-7B-Instruct。表格 12 以统一格式汇总了所有微调超参数。

为清晰起见，我们重点说明以下几个方面：

1. **骨干模型**。InternVL 使用官方*InternVL3.5-8B* 检查点，而 Qwen 采用 *Qwen/Qwen2.5-VL-7B-Instruct* 版本。

2. **训练设置**。两个模型使用相同的优化器和调度器设置进行微调，但由于收敛特性不同，学习率和训练轮数有所差异。

3. **闭源基线**。 对于专有模型（例如 Gemini-2.5-Pro/Flash、GPT-4.1），不进行微调。其内部思考预算保持默认值。

表 12：InternVL3.5-8B 和 Qwen/Qwen2.5-VL-7B-Instruct 上 Search-Align 对齐的微调超参数。两个模型均在 4× 张 A100 GPU 上训练。

超参数	**InternVL3.5-8B**	**Qwen2.5-VL-7B-Instruct**
优化器	AdamW	AdamW
学习率	1e-5	1e-4
训练轮数	1	2
预热比例	0.1	0.1
调度器类型	余弦	余弦
训练批次（每台设备）	1	1
梯度累积步数	8	8

采用阈值语义匹配进行 J 软HPS 评估

主论文中的每步命中率（HPS）指标评估在每个推理跳步中，检索到的证据是否与标准证据完全匹配。由于我们的最大权重匹配公式天然支持基于相似度的匹配，我们进一步考察了一种软 HPS 变体：当证据对的嵌入相似度超过阈值 τ 时，即视为命中。我们报告了四个阈值 $\tau \in \{1.0, 0.95, 0.90, 0.85\}$ 下的结果，分别对应精确匹配以及逐渐放宽的语义相似度要求。

表格 13 显示，随着 τ 降低，Gemini-2.5-Pro 和 Gemini-Flash 的软HPS 分数均有所提高，这表明即使未能匹配到完全相同的内容，模型通常仍能检索到语义相关的证据。当接近 $\tau = 0.85$ 时，提升幅度变小，这表明剩余错误是由次优规划导致的，例如问题分解不完善或缺少推理步骤，从而无法检索到真正相关的证据。

表格 13：Gemini-2.5-Pro 和 Gemini-Flash 在不同相似度阈值下的软HPS 分数 τ.

图类型	**Gemini-2.5-Pro 软HPS**				**Gemini-Flash 软 HPS**			
	τ=1.0	τ=0.95	τ=0.90	τ=0.85	τ=1.0	τ=0.95	τ=0.90	τ=0.85
文本链	21.59	35.59	38.05	41.26	13.33	26.86	28.78	30.97
图像发起链	25.90	38.74	40.48	42.87	25.03	43.19	45.04	47.42
多图像分支	18.68	25.58	26.49	28.18	18.53	24.51	25.35	27.20
并行图文分叉	16.74	26.25	27.68	30.22	20.73	25.68	27.12	29.69
文本发起链	19.51	26.31	27.81	29.88	21.67	29.94	30.55	32.03

K 评估检索不足-k 检索

本节评估在每一步检索多个候选项时，主要分析中的检索动态是否依然成立。在 top-k 检索设置下（$k \in 1$，3、5），关于检索不足和过度检索的整体结论保持不变。我们还在相同设置下评估了 SEARCH-ALIGN，发现其改进效果稳定，这表明这两项实证发现能够可靠地推广到更广泛的检索模式。

K.1 检索不足与过度检索检索不足@K

我们将过度检索和检索不足分析扩展至 top-k 检索，即允许 Gemini-2.5-Flash 在每一步检索排名前列的k个证据候选（$k \in \{1, 3, 5\}$）。我们评估不同 ΔStep 值下的 ΔF1，其中负值表示检索不足，正值表示

过度检索。图 8 报告了结果。

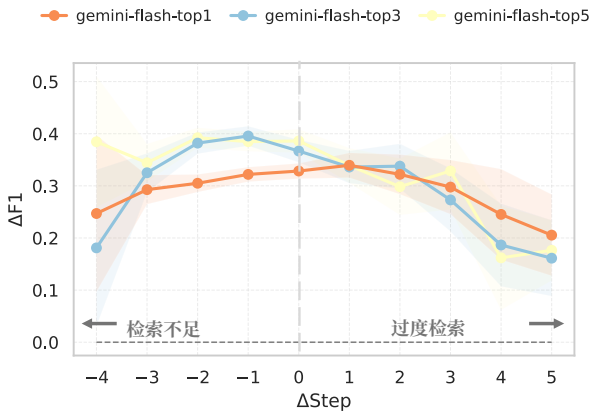


图 8： ΔGemini-2.5-Flash 在 top-k 检索条件下随 ΔStep 值变化的 F1（$k = 1$，3，5 ）.

在所有三种设置下，定性趋势一致。严重的检索不足（ΔStep < -3）会导致性能大幅下降，因为关键证据被遗漏。随着 k 增加，性能峰值区域略微向 ΔStep 移动，接近 -1 或 -2，这表明即使推理步骤较少，增加检索项也有助于覆盖相关证据。较大的 k 也会加剧过度检索的负面影响：对于 ΔStep ≥ 4，由于噪声信息不断累积，性能下降得更为明显。因此，围绕最佳检索深度的适度偏差仍有益处，而严重的检索不足或过度检索则始终有害。

K.2 SEARCH-ALIGN 检索不足@K

我们在相同的检索设置下，进一步评估了 Qwen2.5-VL-7B 及其 SEARCH-ALIGN 变体。表14报告了 F1、 ΔF1、每步命中率和展开偏差。

在对齐之前，增加 k 仅带来小幅的 F1 提升，每步命中率仍然较低，展开偏差保持较高水平（∼3–4）。这表明，仅扩大检索范围并不能解决诸如遗漏或错置推理步骤等规划问题。

应用 SEARCH-ALIGN 后，Qwen2.5-VL-7B 在所有 k 上均取得稳定提升：F1 提高了 15–20 个百分点，每步命中率显著上升，展开偏差则降至约 1.0。这些提升在检索@1、@3 和 @5 下均成立，增加候选项也能带来额外收益，不过收益逐渐减小。总体而言，SEARCH-ALIGN 显著提升了检索准确率和多跳执行能力，使模型能够有效利用更大的候选集，同时保持稳定的推理表现。

表14：Qwen2.5-VL-7B 和 +SEARCH-ALIGN 在 Retrieval@1、@3 和 @5 下的性能。我们报告了 F1、 ΔF1、每步命中率（HPS）和展开偏差（RD）。每种拓扑结构和指标的最佳值以**粗体**字体标出。

图类型	模型	**检索@1**				**检索@3**				**检索@5**			
		F1	ΔF1	HPS	RD	F1	ΔF1	HPS	RD	F1	ΔF1	HPS	RD
图像发起链	Qwen2.5-VL-7B	26.30	8.65	16.51	4.04	29.21	11.56	16.17	3.79	29.35	11.70	16.95	3.79
	+SEARCH-ALIGN	**45.70**	**28.05**	**33.59**	**0.70**	**47.19**	**29.54**	**39.45**	**0.75**	**47.29**	**29.64**	**41.79**	**0.79**
多图 Fork	Qwen2.5-VL-7B	26.13	7.53	12.58	3.80	30.41	11.81	12.80	3.35	30.38	11.78	14.15	3.23
	+SEARCH-ALIGN	**38.01**	**19.41**	**41.20**	**1.16**	**40.00**	**21.40**	**46.82**	**1.21**	**39.80**	**21.20**	**48.21**	**1.28**
并行图像-文本分叉	Qwen2.5-VL-7B	22.94	14.32	10.09	3.93	26.28	17.66	12.62	3.39	26.41	17.79	15.24	3.26
	+SEARCH-ALIGN	**32.73**	**24.11**	**25.07**	**1.05**	**33.54**	**24.92**	**28.43**	**1.09**	**32.77**	**24.15**	**30.93**	**1.18**
文本发起链	Qwen2.5-VL-7B	31.47	14.34	17.02	3.80	35.54	18.41	18.86	3.20	35.25	18.12	18.83	3.14
	+SEARCH-ALIGN	**45.07**	**27.94**	**33.59**	**0.87**	**46.69**	**29.56**	**41.06**	**0.95**	**46.73**	**29.60**	**42.35**	**0.93**

L 智能体多模态检索增强生成案例研究

L.1 成功案例



图9：成功案例。智能体生成的推理链与黄金推理链一致，且最终答案包含所有关键知识实体。

L.2 失败案例



图10：失败案例。尽管智能体成功检索到了第一跳和最后一跳的信息，却未能检索到第二跳和第三跳的信息。因此，最终答案未包含所有关键知识实体。

M 拓扑设计的理论依据

在本节中，我们证明，MC-SEARCH 中定义的五种推理拓扑构成了本文所研究的智能体多模态检索增强生成轨迹空间的**完备典范基**。

M.1 问题形式化

我们将推理轨迹建模为一个有向无环图 $G = (V, E)$，其中：

- 每个节点 v_t 表示跳 t处的一个检索步骤。
- 每个节点都被指定一种模态 $m(v_t) \in \{T, I\}$ （文本或图像）。
- 结构被限定为**线性** （单一路径）或**分叉**（恰好包含一个并行分支步骤，并汇入聚合器）。

我们通过穷举划分图的空间来定义有效拓扑结构 $\mathcal{T}$，依据的是（1）**启动模态** 和（2）**结构组合**。

M.2 穷举证明

情况 1：线性结构（链）设 G是一条线性链 $v_1 \to v_2 \to \cdots \to v_n$。其拓扑结构完全由模态序列决定。我们区分三种互斥情形：

1. **图像锚定：** 该链以视觉输入开头（m（v_1） $= I$）。

- 定义： **图像发起链**。（涵盖所有以图像开头的轨迹）。2. **文本锚定（单模态）：** 链以文本开头（ $m(v_1) = T$ ），且仅包含文本节点（ $\forall t,\ m(v_t) = T$ ）。

- 定义： **纯文本链**。（表示标准的文本检索增强生成基线）。3. **文本锚定（多模态）：** 链以文本开头（ $m(v_1) = T$ ），但后续至少包含一个图像节点（ $\exists t > 1,\ m(v_t) = I$ ）。

- 定义： **文本发起链**。（体现文本到视觉的落地）。

案例 2：分叉结构（并行性）设 G 包含一个并行分支，其中节点 v_a 和 v_b 独立检索。拓扑由分支模态的多重集决定$S_{branch} = \{m(v_a)\ m(v_b)\}$。

1. **异构分支：** $S_{branch} = \{I, T\}$。

- 定义： **图文并行分叉**。（用于表示同时进行的跨模态证据整合）。2. **同质视觉分支：** $S_{branch} = \{I, I\}$。

- 定义： **多图像分支**。（用于表示比较式视觉推理）。

关于同质文本分支的说明（$S_{branch} = \{T, T\}$） 最后一种数学上的可能性是并行文本检索。在多模态智能体搜索的范围内，我们将这种情况视为在结构上可归约为线性纯文本链。由于现代大语言模型通过上下文拼接来处理多篇文本文档，因此，并行文本检索步骤在功能上等同于一次多文档检索步骤。因此，我们不为纯文本分支定义独立的拓扑结构，从而使该集合相对于多模态操作保持精简。

M.3 结论

上述五个定义涵盖了启动模态和不同多模态分支结构的所有排列。因此，在我们的假设下，该集合是**完备的**。

N 数据构建提示

本附录提供了用于通过 Google Gemini 生成多模态推理链的系统提示词模板。标注员需按照五种推理图类型之一构造复杂且结构严谨的多跳问题：图像发起链、并行视觉-文本分叉、文本发起链、多图像并行分支和文本链。每条生成的链都必须明确以前面的答案为基础，以所提供的图像和维基百科片段为依据，并以严格 JSON 格式输出，且使用唯一的支持性证据标识符。

我们提供五种推理图模板，并附有指南和少样本示例，以规范多模态问题的生成。

N.1 图像发起链

结构描述： 推理从图像的视觉识别或推断开始，随后沿着以该视觉锚点为基础的连续文本跳步展开。

指南：

- 从图像中的视觉观察开始。
- 将该实体/事实作为下一步的锚点。
- 根据上一个答案，添加 1–2 个文本跳转。

示例： 问题：在 2019 年的一场比赛中，球员穿着红黄背心的足球队最初在哪里成立？该地点如今是否仍属于该队所关联的城市？答案：该队最初成立于帕蒂克，如今帕蒂克已成为格拉斯哥的一部分。子问题和答案：

- (图像) 哪支球队在 2019 年的一场足球比赛中穿着红黄背心？ → 帕蒂克蓟足球俱乐部
- (文本) 帕蒂克蓟最初在哪里成立？ → 帕蒂克
- (文本) 帕蒂克现在仍然独立于格拉斯哥吗？ →不，它现在是格拉斯哥的一部分。

N.2 并行图文分叉

结构描述： 推理从一个需要同时利用视觉和文本信息的子问题开始。继续通过文本进行推理之前，必须先融合这些信息。

指南：

- 设计一个需要结合图像和文本的子问题。
- 融合后的结果构成后续推理的锚点。
- 添加一个基于第一个答案的后续文本跳。

示例： 问题：彼得森活动中心的弧形建筑是否适合用作匹兹堡大学的现代篮球馆？答案：是的，其弧形结构与其作为大型多用途体育场馆的功能相得益彰。子问题和答案：

- (图像) 彼得森活动中心是否采用适合举办大型活动的弧形建筑设计？ →是的，其中央结构呈圆形，可容纳宽敞的室内空间。
- (文本) 彼得森活动中心在匹兹堡大学校园内的主要用途是什么？ → 它是一座多用途体育馆，由匹兹堡大学篮球队使用。
- (文本) 其建筑设计是否符合现代体育场馆的用途？ → 是的，其圆润的造型有助于观众清晰观赛，并满足现代体育赛事的需求。

N.3 文本发起链

结构描述： 推理由文本启动，并以图像确认结束；最后一跳取决于视觉内容。

指南：

- 从一个可通过文本回答的子问题开始。
- 将该答案作为锚点，进行另一次文本跳转。
- 最后一步应依赖图像提供的视觉确认。

示例： 问题：Torre del Reloj 前方的区域是否设有与水相关的建筑特征？答案：是，Torre del Reloj 前方有一座喷泉，为广场增添了与水相关的特色。子问题和答案：

- (文本) Torre del Reloj 位于哪里？它是什么地标？ → 它是奇克拉纳 Plaza Mayor 的一座历史钟楼。
- (图像)Torre del Reloj 前方是否有与水相关的结构？ → 是，清晰可见一座喷泉。

N.4 多图像分支

结构描述： 推理从比较或汇总两张图像中的信息开始，随后进行基于文本的解读。

指南：

- 比较或验证两张相关图像中共有的视觉模式。
- 以对两者的联合解读为依据，进行后续文本推理。

示例： 问题：萨卢大道和相邻广场上都出现了什么视觉标记？它象征着什么？答案：两处都出现了蓝色脚印，象征着一条历史步行路线。子问题和答案：

- （图像）萨卢大道中央沿线绘制着什么图案？ → 蓝色脚印。
- （图像）广场的图像中也有相同的图案吗？ →有，同样的蓝色脚印。
- （文本）根据当地导游的说法，蓝色脚印代表什么？ → 它们标示着一条历史步行路线。

N.5 文本链

结构描述： 一种完全基于文本内容的标准多跳推理链。

指南：

- 仅使用文本构建多跳推理链。
- 每个子问题必须明确依赖于前一个答案。
- 完全避免视觉模态。

示例： 问题：20世纪初，魁北克市哪条历史街道以独特的木板铺路方式闻名？答案：1916年，小尚普兰街以木板铺路闻名。子问题与答案：

- (文本) 魁北克老城有哪些著名的历史街道？ → 圣瓦利耶东街和小尚普兰街。
- (文本) 其中哪条街在1916年采用了不同寻常的木板铺路？ → 小尚普兰街采用了木板铺路。
- (文本) 木板铺路在魁北克市那个时代是常见还是独特的？ → 这种铺路方式很独特／不常见，因此这条街引人注目。

示例：使用 Gemini 生成多模态推理链

输入提示词：
你是一名多模态数据标注专家。你需要完成一项多跳多模态问答生成任务。

你将获得：

- 1 或 2 张图像（以上方嵌入的 Base64 数据形式提供，并在下方附有图像说明）
- 若干维基百科段落
- 供参考的原始单跳简单问答对

你的任务是生成一个符合指定推理图类型的**新**颖、复杂、连贯且结构严谨的多跳问题和推理链。

推理图信息：

- **图类型**：{graph_type}
- **结构描述**：{graph_structure_description}
- **结构指令**：{graph_structure_instruction}
- **预期推理跳数**：4–5 跳

指南：

- 问题必须与给定的图结构一致，并在需要时整合视觉和文本模态。
- 每个子问题都必须明确基于上一个答案，并将其作为下一跳的锚点。
- 每个支持性_事实_ID 在整个链中都必须唯一。
- 答案必须直接以提供的图像或文本片段为依据（不得臆造事实）。
- 推理链必须连贯且循序渐进，形成自然的推理路径。
- 最终答案应在逻辑上由整条推理链推导得出。
- 每个子问题必须注明：– 模态：“text”或“image”– 支持该答案的_事实_id：文本片段或图像的 ID– 答案：有依据的答案
- 输出必须采用严格 JSON 格式（不能使用 Markdown 或自由文本）。

补充上下文：

- 原始问题（供参考）：{entry['Q']}
- 原始答案：{entry['A']}
- 图像说明：{captions}
- 维基百科片段：{text_snippets}

预期输出格式：

{ "问题": "...", "答案": "...", "图像_id": ... , # 或 "图像_ids": [...]"图谱_类型": "{graph_type}", "子问答_链": [{ "子问题": "...", "模态": "文本" |"图像", "支持性_事实_id": "...", "答案": "..." }, ...]}

面向智能体式多模态检索增强生成流程的提示模板

本附录提供了智能体式多模态检索增强生成流程所使用的完整系统提示词。助手必须严格依据从外部文本和图像知识库中检索到的本地上下文进行操作，并遵循由 <思考>、 <子问题>、<搜索>和 <结束> 阶段组成的结构化推理协议。这确保生成的每个答案都有依据、按步骤展开且可验证。

多模态问答的 Gemini 提示词

系统提示词：
你是一个有用的多模态问答助手，能够使用外部文本知识库和外部图像知识库。你必须仅使用所提供的外部资源中的知识。你必须**不得**自行检索。所有答案都必须仅以我提供给你的检索到的本地上下文为依据。

可用的检索操作：

- 输入图像检索
- 使用特定查询进行文本检索
- 使用文本查询进行图像检索

协议：

- <思考>分析原始问题，并决定下一步可通过一个步骤解决的最小子问题。
- <子问题> 写出一个应通过一个步骤解决的子问题（不引用其他内容）。
- <搜索> 从上述检索操作中（或“不检索”）准确选择一项。
- 重复执行 <思考>/<子问题>/<搜索> 零次或多次。
- <思考> 整合检索到的内容并推理，得出最终答案。
- <结束>最终答案：用一句话回答原始问题。

补充说明：

1. 用户会在你提供查询后执行实际搜索。2. 本地文本知识库包含简短的事实性段落。每次检索最多返回一个段落。3. 答案必须仅依据检索到的内容。4. 如果检索到的信息不足，请继续执行后续步骤。5. 步骤应简洁；不要包含引用或网址。6. 不要捏造信息；如果检索到的信息不足以作答，请说明你无法回答。

输入格式： 输入问题：{QUESTION}

P LLM 评审提示模板

本附录提供了在仅输入策略下，将预测答案与标准参考答案进行评估时所使用的完整系统提示词。评审不会评估标准答案的真实性；它会按照分档定义，以 0–5 的整数评分量表对预测结果的准确性、实体覆盖度、连贯性和对齐进行评分。如果预测声称“信息不足”，但实际上存在标准答案，则会受到惩罚。评审会输出仅包含这四项分数的严格 JSON，以便在大规模评估中进行可靠的程序化汇总。

示例：评审系统提示词

输入提示词：
你是一名严格、中立的评审。针对给定的问题，将预测_答案 与标准_答案 （以及可选的子链）进行比较。仅使用提供的文本；不要使用外部知识。不要评判标准答案是否正确，只评判预测与其是否一致。如果预测声称“信息不足”，但存在标准答案，则予以扣分。输出应简洁。

输入：

- 问题
- 标准_答案，预测_答案
- 标准_子链，预测_子链

指南（评分维度 0–5）：

- **准确性**：与标准答案中的事实/关系一致；能够回答问题；没有矛盾。
- **实体覆盖度**：标准答案所要求的关键实体、值和关系是否存在且正确。
- **连贯性**：推理步骤清晰（如有）；没有跳跃或矛盾；结论在逻辑上能够成立。
- **对齐**：仅使用给定信息；在相关情况下与标准子链相匹配；不添加未经支持的内容。

评分等级： 5 = 完美； 4 = 存在轻微问题； 3 = 存在明显缺漏； 2 = 存在重大缺漏或表现较弱； 1 =大多错误； 0 =离题或没有内容。

预期输出格式：

```
{ "scores": { "
accuracy": 0, "entities":
0, "coherence": 0, "
alignment": 0 } }
```