arXiv:2603.01050v1 [cs.CV] 1 Mar 2026

MM-DeepResear 中文：一个简单而有效的多模态智能体搜索基线 e

Huanjin Yao[1] **Qixiang Yin**[3] **MinYang**[1], ✉ **Ziwang Zhao**[1] **Yibo Wang**[4]
HaotianLuo[5]**Jingyi Zhang**[4] **Jiaxing Huang**[2], ✉

摘要

我们的目标是开发一种多模态研究智能体，使其能够进行显式推理与规划、调用多种工具并综合跨模态信息，从而执行深度研究任务。然而，我们发现开发此类智能体面临三大挑战：（1）搜索密集型多模态问答数据匮乏；（2）缺乏有效的搜索轨迹；（3）使用在线搜索 API 进行训练的成本高昂。为应对这些挑战，我们首先提出**Hyper-Search**，这是一种基于超图的问答生成方法，通过对模态内部及模态之间的视觉与文本节点进行建模和关联，生成需要调用各种搜索工具才能解决的搜索密集型多模态问答对。其次，我们提出**DR-TTS**，该方法首先根据搜索工具类型将涉及搜索的任务分解为若干类别，并针对每种工具分别优化专门的搜索工具专家。随后，它重新组合这些工具专家，通过树搜索协同探索搜索轨迹，从而生成能够使用各种搜索工具成功解决复杂任务的轨迹。第三，我们构建了一个支持多种搜索工具的离线搜索引擎，使智能体强化学习无需使用成本高昂的在线搜索 API。基于这三项设计，我们开发了**MM-DeepResearch**，这是一种强大的多模态深度研究智能体，广泛的实验结果表明它在各项基准测试中表现优越。代码见

h
t
t
p
s://
github.com/HJYao00/MM-DeepResearch

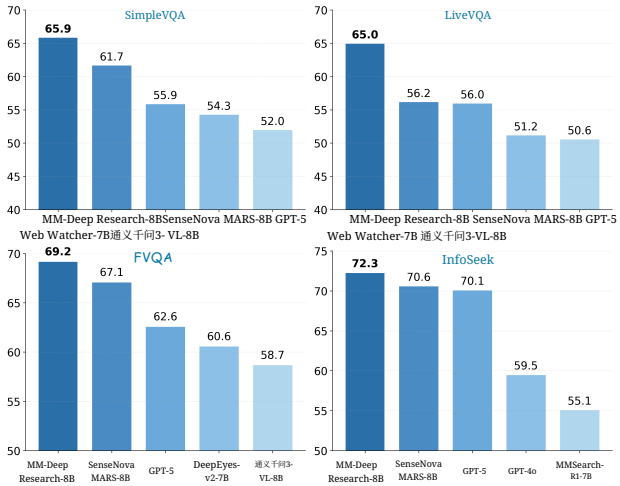


图 *1.* MM-DeepResearch-8B 在四个基准测试上的整体表现与其他模型的比较。

1. 引言

近年来，具备推理能力的多模态大语言模型（Jaech 等，2024年；Comanici 等，2025年）通过显式生成中间推理过程，在解决复杂任务方面展现出了卓越的能力

尽管取得了这些进展，现有模型仍受到容量受限的参数及固定且有限的知识的根本制约，这限制了它们处理超出自身固有知识范围的问题的能力，尤其是在信息密集型和开放世界任务中。

为缓解这一局限，早期方法主要采用预定义的信息检索工作流，例如检索增强生成（RAG）（Chen 等，2024年）或基于提示的方法（Jiang 等，2024年），它们在预定义流程中检索外部信息作为语境，然后基于补充语境开始推理。然而，这些方法将检索与推理过程相分离，无法随着模型推理状态的变化而迭代调整检索策略，因而搜索能力和泛化能力有限。

我们的目标是开发一种具备智能体搜索能力的多模态深度研究模型，包括显式推理与规划、多工具调用以及跨模态信息综合，使其能够执行深度研究任务。我们发现，开发此类智能体面临若干根本性挑战：（1）搜索密集型多模态问答数据稀缺。 公开可用的搜索密集型多模态问

✉通讯作者。 [1] 字节跳动； [2] 香港理工大学； [3] 中关村学院； [4] 南洋理工大学； [5] 四川大学。

答数据集非常有限，且缺乏多轮搜索和多工具调用，导致无法提供充足的监督信号，有效激励多模态大语言模型具备智能体搜索能力。（2）缺乏有效的搜索轨迹。传统的基于提示的搜索轨迹合成方法主要针对单轮搜索设计，难以适用于多轮场景；此类场景需要在迭代推理过程中与各种搜索工具进行多轮交互。（3）使用在线搜索 *API* 进行训练的成本高昂。目前大多数方法在训练期间依赖在线搜索 API（例如 SerpAPI 和 Jina），每次训练都可能轻易花费数千美元，极大限制了大规模实验和系统性探索。

为应对这些挑战，我们提出了三种互补技术：Hyper-Search、DR-TTS 和离线多模态搜索引擎。首先，我们提出**Hyper-Search**，这是一种基于超图的方法，用于建模并连接模态内部及不同模态之间的视觉与文本节点，从而生成需要调用各种搜索工具才能解答的搜索密集型多模态问答对。具体而言，Hyper-Search 将网络中的视觉和文本内容共同建模为节点，并通过超边捕捉这些节点之间固有的关系，从而以结构化方式表示复杂的多源、多模态信息依赖关系。基于这一超图，Hyper-Search 通过关联一个或多个超边上的两个或更多超图节点来构建搜索密集型多模态问答对，确保每个生成的问题都必须通过多轮、多类型的搜索工具调用才能解答。

其次，我们引入分解-重组工具树搜索（**DR-TTS**），以有效合成利用各种搜索工具完成任务的搜索密集型推理轨迹。DR-TTS 首先根据搜索工具类型，将涉及搜索的任务分为若干类别，并分别通过强化学习训练专门的搜索工具专家，每位专家精通一种工具。这种分解设计简化了搜索工具的学习难度，并提升了各工具的熟练度。随后，DR-TTS 将这些工具专家重新组合，通过树搜索协同探索并筛选有效的搜索轨迹，从而生成高质量的监督微调数据。与由单个模型直接生成轨迹相比，DR-TTS 能够减轻工具使用偏差并提升各工具的熟练度，从而在工具内部及工具之间实现更均衡、更有效的探索，并显著提高轨迹探索的多样性和成功率。

第三，基于上述两项设计，我们构建了一个离线搜索引擎，同时支持基于信息的搜索和基于知识的搜索，以应对各种复杂的现实世界任务。基于信息的搜索工具从预先构建的多模态语料库中检索视觉和文本语境，以支持搜索密集型任务，而基于知识的搜索工具则接受复杂的搜索查询，并返回模型生成的知识密集型内容。与在线搜索引擎相比，我们的离线搜索引擎能够提供更快的工具响应，并缩短工具交互训练时间。更重要的是，它避免了昂贵的在线搜索 API 费用，每次训练运行的费用可能高达数千美元。

基于这些设计，我们训练了一个强大的多模态深度研究智能体，命名为**MM-DeepResearch**。该模型使用 Hyper-Search 生成的密集搜索问答数据、通过 DR-TTS 合成的搜索工具轨迹以及离线搜索引擎进行训练。在使用在线搜索 API 进行评估时，MM-DeepResearch 在多个基准测试中均表现出色，优于此前使用成本高昂的在线搜索 API 训练的智能体。

本研究的主要贡献有四点。首先，我们提出 Hyper-Search，引入超图来建模并连接模态内部及模态之间的节点，从而生成高质量的密集搜索多模态问答对。据我们所知，这是首项将超图概念用于密集搜索问答数据合成的研究。其次，我们设计了 DR-TTS，将涉及搜索的任务分解，以训练专门的搜索工具专家；并将工具专家重新组合，共同探索并合成高质量的搜索轨迹。第三，我们构建了一个包含多种搜索工具的离线搜索引擎，使智能体强化学习无需使用成本高昂的在线搜索 API。第四，大量实验表明，我们的方法和模型在各种多模态深度研究任务中均具有优越性。

2. 相关工作

2.1. 智能体多模态大语言模型

早期工作利用大规模多模态预训练和指令微调，构建用于视觉语言理解任务的通用多模态大语言模型（Liu 等人，2024年；An 等人，2025年），并采用后训练技术，激励模型在复杂推理任务中具备长链推理能力（Huang 等人，2025年；Zhang 等人，2025a）。近期工作进一步强调智能体行为（Yao 等人，2025c），使多模态大语言模型能够自主且迭代地选择并调用工具，以完成更复杂的任务。近期的智能体基础多模态大语言模型（Team 等人，2026b），例如 Qwen3-VL（Bai 等人，2025年），通过在智能体预训练中纳入多样化的工具交互轨迹，具备了原生的工具使用能力。本文旨在开发一种多模态深度研究智能体，以提升搜索工具调用和多模态信息综合能力。

2.2. 基于工作流的搜索智能体

基于工作流的搜索智能体采用预定义的静态信息获取流程，检索外部知识以进行推理。这些方法大致可分为基于检索增强生成的方法和基于提示词的方法。（1）基于检索增强生成的方法（Wang等人，2025b；Li 等人，2025c）通过基于相似度的检索从数据库中获取外部知识，并将检索到的内容注入模型输入，从而增强多模态大语言模型。（2）基于提示词的方法（Zheng 等人，2025年；Wu 等人，2025b）通过编码在提示中的人工设计工作流，显式编排搜索行为，其中工具使用模式在推理期间预先设定。这些方法通过引入外部知识，提升了事实依据以及模型在知识密集型任务上的性能。然而，基于工作流的搜索智能体将搜索与推理解耦，并受到固定且僵化流程的限制，容易导致搜索不足或过度搜索，进而限制其泛化性能。这些局限促使我们探索多模态深度研究智能体，使其以紧密耦合的推理—搜索方式主动且迭代地调用搜索工具。

2.3. 多模态深度研究智能体

近期研究提出了深度研究智能体，使大语言模型具备智能体搜索能力，从而能够通过与外部文本搜索工具交互，迭代地获取信息并综合证据，代表性工作包括 OpenAI Deep-Research（OpenAI，2025年b）和 MiroThinker（Team 等人，2025年b）。除了仅处理文本的智能体外，此前的一些研究（Yang 等人，2025年；Liu 等人，2025a；Xiao 等人，2025年）将深度研究范式扩展到了多模态领域。MMSearch-R1（Wu 等人，2025a）率先采用端到端强化学习，使多模态大语言模型能够使用图像和文本搜索工具。WebWatcher（Geng 等人，2025年）合成密集搜索问答对，并将其转换为视觉问答数据，用于训练多模态深度研究智能体。然而，现有大多数工作并未公开密集搜索视觉问答数据集和搜索工具轨迹，且严重依赖在线搜索 API，这给多模态深度研究智能体的开发带来了重大挑战。为应对这些挑战，本文提出了用于生成多模态密集搜索问答数据的 Hyper-Search、用于合成搜索轨迹的 DR-TTS，以及用于强化学习的离线搜索引擎。这些组件共同使我们能够训练深度研究智能体，并取得出色的实证表现。

3. MM-DeepResearch 数据

本节聚焦多模态深度研究在数据层面面临的挑战。首先，我们在第3.1节介绍一种基于超图的方法（Hyper-Search），用于构建搜索密集型问答。3.1。随后，我们将在第3.2节介绍 DR-TTS，该方法可联合探索并寻找有效的搜索工具轨迹。

3.1. 用于问答数据生成的 Hyper-Search

如图 2所示，通过 Hyper-Search 构建多模态、密集搜索的问答数据包含三个阶段：(1) 超图构建，对网络图像、网页内容及其关系进行建模；(2) 多模态问答生成，基于构建的超图生成问答对；(3) 数据过滤，旨在移除低质量和免搜索多模态问答数据。

3.1.1. 搜索超图构建

节点定义。 在 Hyper-Search 中，我们使用由两种基本节点类型组成的超图来表示网络信息，即*i.e.*，图像节点 $\mathcal{I}$ 和文本节点 $\mathcal{T}$。每个图像节点$i_{d,k} \in \mathcal{I}$ 对应于从在线来源收集的一幅图像，而每个文本节点 $t_{d,k} \in \mathcal{T}$表示一个网页的完整内容，其中 d表示扩展深度， k 表示该深度下节点的索引。对于每个节点，我们利用多模态大语言模型为图像节点生成图注，并为文本节点生成摘要，作为后续多模态问答生成过程中的信息提示。这些表示支持针对不同模态的节点扩展与互连，从而合成以多样化信息来源为依据的多模态问答对。

节点扩展。我们为文本节点和图像节点设计了不同的扩展策略，以检索更多相关信息并丰富超图。(a)图像节点扩展。对于每个图像节点 $i_{d,\ k}$，我们定义了两种扩展操作，分别引入新的文本节点和图像节点。第一种操作是图像反向搜索，通过查询图像来检索 K 相关网页内容，从而生成文本节点 $\{t_{d+1,1}^{i_{d,k}}, \ldots, t_{d+1,K}^{i_{d,k}}\}$。第二种操作是图像视觉搜索，返回 K 视觉上相似的图像，从而生成图像节点 $\{i_{d+1,1}^{i_{d,k}}, \ldots, i_{d+1,K}^{i_{d,k}}\}$。(b) 文本节点扩展。对于每个文本节点 $t_{d,\ k}$，我们同样设计了两种扩展操作，以便通过文本节点和图像节点扩展超图。对于文本节点扩展，我们使用多模态大语言模型从原始网页内容中提取信息量最高的前K 个网页网址，每个网址都会引入一个对应于检索到的网页 $\{t_{d+1,1}^{t_{d,k}}, \ldots, t_{d+1,K}^{t_{d,k}}\}$的新文本节点。对于图像节点扩展，我们采用类似方法，由大语言模型直接从原始网页内容中识别最相关的前K 个图像链接，并据此生成新的图像节点 $\{i_{d+1,1}^{t_{d,k}}, \ldots, i_{d+1,K}^{t_{d,k}}\}$。上标表示新节点所扩展自的父节点。

超边连接。 每次节点扩展都会形成一条超边 $e_{d,\ k}$ 将父节点与所有新生成的节点连接起来，从而捕捉多模态交互。例如，给定一个图像节点 $i_{d,\ k}$，其扩

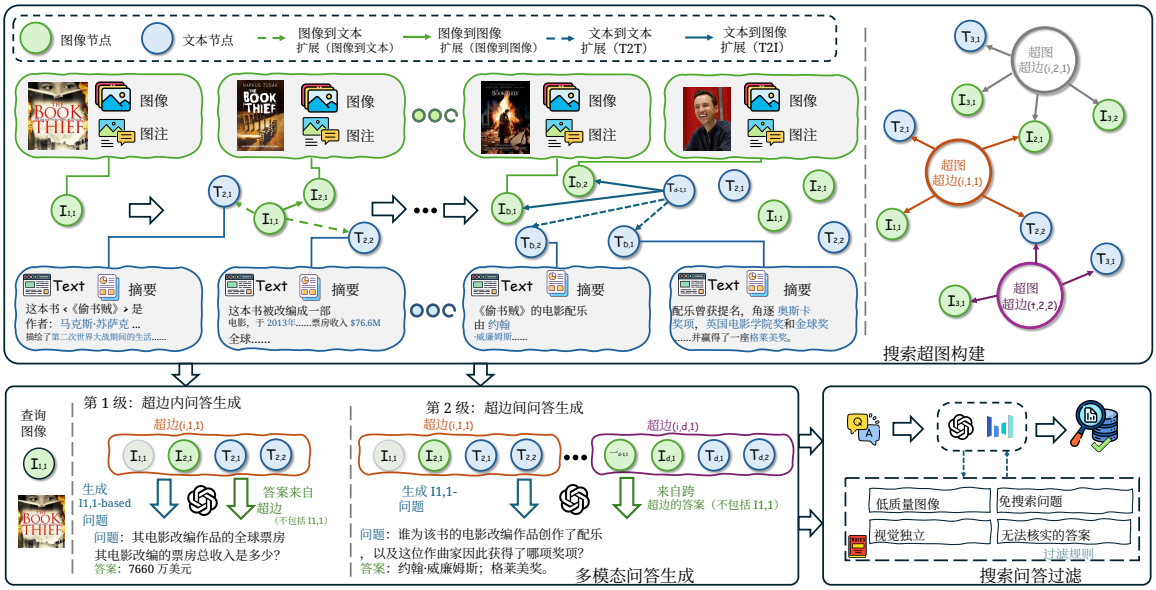


图 *2.* Hyper-Search 概述：通过超图构建、问答生成和过滤，生成密集搜索问答数据。

展会生成新的图像节点 $\{i_{d+1,1}^{i_{d,k}},\ldots,i_{d+1,K}^{i_{d,k}}\}$和文本节点 $\{t_{d+1,1}^{i_{d,k}},\ldots,t_{d+1,K}^{i_{d,k}}\}$。这些节点被归入一条定义为 $e_{d,k}=\{i_{d,k}^{i_{d-1,k}},\ i_{d+1,1}^{i_{d,k}},\ldots,i_{d+1,K}^{i_{d,k}},\ t_{d+1,1}^{i_{d,k}},\ldots,t_{d+1,K}^{i_{d,k}}\}$. 的超边。这种构造明确地对相连节点之间的跨模态检索关系进行建模。

超图。 我们从一个图像节点开始构建超图。通过持续扩展，Hyper-Search 逐步构建深度为 D 的超图，用于组织多模态网络信息，并为下游多模态问答生成提供结构化关系单位。

3.1.2. 多模态搜索问答生成

问答生成在超边层级进行，利用超边内部及超边之间的跨模态和跨来源信息，生成多模态密集搜索问答数据。根据聚合超边的范围，我们定义了两种生成策略：超边内 和超边间 问答生成。

第 1 级：超边内问答生成。在第 1 级，我们在单个超边 $e_{d,\ k}$内执行问答生成。具体而言，我们选择一个图像节点 $i_{d,\ k}$ 作为查询图像，并以 $e_{d,\ k}$内其余图像图注和网页内容作为证据，生成问答数据。这些生成的搜索密集型问题明确依赖于 $i_{d,\ k}$中的视觉信息，而相应答案必须根据所提供的证据得出。由于证据涵盖多种模态，这一级的数据能有效促使模型主动使用图像与文本工具进行搜索。

级别 2：超边间问答生成。在级别 2 中，我们扩大证据范围，以增加搜索深度和难度；问答生成会整合多个超边中的信息 $\{e_{d_1,k_1},\ldots,e_{d_m,k_m}\}$。具体而言，选取一个图像节点作为查询图像，并基于这些超边中的聚合图像和网页内容生成问答数据。此级别通常需要多轮搜索，从而促使智能体搜索行为进行更深入的探索和更广泛的信息聚合。

3.1.3. 搜索问答过滤

为确保数据质量，我们使用多模态大语言模型（即：例如，GPT-5（OpenAI，2025a）和 Seed1.8（Seed））来筛选低质量的多模态问答实例。此流程会剔除低质量图像、重复问答对、视觉上无关或无需搜索的问题，以及无法通过证据验证的答案。最终得到的数据集包含 3K 个搜索密集型多模态问答对，称为 **Hyper-Search-3K**。

3.2. 用于搜索轨迹合成的 DR-TTS

我们提出了分解-重组工具树搜索（DR-TTS），这是一种用于合成搜索工具轨迹的树搜索方法。DR-TTS 首先根据所需的搜索工具对搜索相关任务进行分类，并将模型训练分解为针对每种工具分别优化专用专家。随后，它重新组合这些工具专家，以执行联合树搜索，旨在发现有效轨迹，同时提高探索成功率和轨迹多样性。

搜索工具定义。 我们定义了基于信息的搜索工具和基于知识的搜索工具，使智能体能够获取有依据的事实信息和专门知

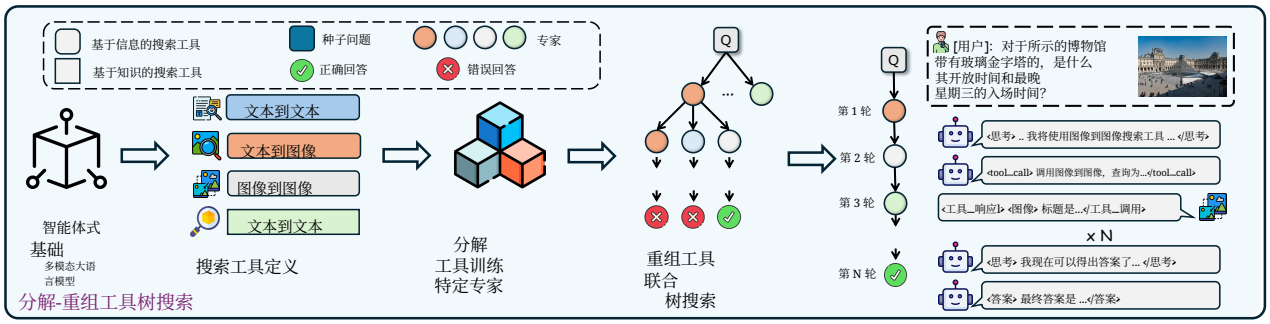


图 *3.* 用于探索搜索轨迹的分解-重组工具树搜索概述。

识，从而进行更准确的轨迹搜索。*(1)* 基于信息的搜索工具 通过将多模态查询与外部网络来源进行匹配，检索现实世界的信息。我们设计了三种基于信息的搜索工具（即，文本到文本、文本到图像和图像到图像检索），根据不同的查询类型获取所需的文本和视觉信息。具体而言，文本到文本检索根据文本查询返回相关网页；文本到图像检索根据文本查询返回相关网络图像；图像到图像检索则根据图像查询返回视觉上相似的图像。这些工具提供有依据的文本和视觉信息，使模型能够应对搜索密集型推理任务。*(2)* 基于知识的搜索工具 旨在通过使用复杂查询询问语言模型，提供知识密集型信息。我们将这些工具构建为文本到文本交互：智能体提出自然语言查询，并接收由领域特定专家模型生成的回复。这类工具适用于领域特定推理任务，尤其是那些需要获取无法通过网络检索直接访问的信息的任务。

工具特定专家的分解训练。 直接使用预训练的智能体式基础多模态大语言模型进行树搜索，往往会导致次优性能，因为模型经常产生无效工具调用，并且整合检索信息的能力有限，这严重阻碍了有效探索。为解决这一问题，我们首先根据所需的搜索工具对搜索相关任务进行分类，并分解模型训练，为每种工具优化专用专家。具体而言，我们使用 GPT 按工具类型对搜索相关训练数据进行分类。随后，我们使用第 4.2 节所述的强化学习策略训练每个工具专属子集对应的专家，无需冷启动。此过程会得到一组 M 专用工具专家，每位专家都精通一种搜索工具，从而降低学习复杂度并提升各工具上的熟练度。

用于联合树搜索的重组工具专家。 随后，我们对这些工具专家进行重组，使其协作探索并寻找能够通过树搜索成功解决任务的搜索轨迹。从根节点（*i.e.*，输入问题 $Q = \{T, I\}$）开始，我们扩展 M 个子节点，每个子节点均由一名工具专家生成。树中的每个节点记为 s_i^j，由一个推理与规划步骤 t_i^j、一次工具调用 c_i^j 以及工具响应 r_i^j 组成，其中，i 表示节点深度，j 表示工具专家的索引。在每个深度切换不同的工具专家进行分支后，搜索会递归地遍历后续树层，探索由不同专家驱动的多样化决策，并构建候选轨迹。当某位专家选择在不调用工具的情况下生成最终答案 a 时，我们会使用大语言模型评估该答案的正确性。如果验证答案正确，则整个树搜索终止；否则，生成错误最终答案的节点将被视为终端节点并从后续扩展中剪枝，同时搜索继续处理其余节点。

搜索轨迹提取。 对于成功得出正确答案的搜索轨迹，我们从树中提取相应的探索路径，并构建用于监督微调（SFT）的轨迹数据。每条轨迹都表示为由推理步骤和工具交互组成的有序序列，如 $\tau = \{t_1, c_1, r_1, t_2, c_2, r_2, ..., t_n, a_n\}$。通过 DR-TTS，在多样化工具专家上进行树结构搜索，并在探索过程中平衡工具使用，从而能够发现有效的推理轨迹。最终，我们使用 DR-TTS 收集了 10K 条 SFT 轨迹，并将其称为 DR-TTS-10K。

4. MM-DeepResearch 智能体

我们利用生成的问答数据和搜索工具轨迹，通过两阶段训练方案训练 MM-DeepResearch：（i）将搜索工具轨迹上的监督微调用作冷启动；（ii）使用离线搜索引擎进行智能体多轮强化学习。

4.1. 使用多轮搜索工具调用进行监督微调

利用 DR-TTS 探索的搜索轨迹，我们进行多轮监督微调作为冷启动，训练模型学习工具调用模式，并提升跨模态信息整合能力。训练目标最大化似-

然，即模型生成的推理与规划步骤 t_i、搜索工具调用 c_i以及最终答案 a_n的概率，同时从损失中掩蔽工具响应 r_i 。监督微调目标为：

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{\tau\sim\mathcal{D}} \sum_{i=1}^{n} \log p_\theta(t_i, c_i, a_n \mid Q, t_{<i}, c_{<i}, r_{<i}) \tag{1}$$

最终答案 a_n 仅在最后一步生成并接受监督，而工具响应 r_i 则被视为用于条件化下一步预测的观测。通过这种方式，模型能够学习恰当的工具调用模式，并提升长上下文信息综合能力。

4.2. 使用离线搜索引擎进行强化学习

与先前在强化学习期间使用在线搜索 API 的方法不同，此类方法的成本可能高得令人难以承受，尤其是 Guo 等人（2025b）提出的 GRPO 需要进行多次轨迹采样。我们构建了一个离线搜索引擎，用于模拟真实世界的搜索环境，并支持视觉和文本检索。

离线搜索引擎构建。 为使离线搜索引擎能够在强化学习期间调用第 3.2 节定义的工具，我们预先收集了一个大规模多模态语料库，其中包含多样化的图像和文本。具体而言，对于文本信息，我们使用 GPT 生成多个候选搜索查询，并预先获取对应的网页内容。为进一步扩充语料库，我们还纳入了维基百科数据。对于图像相关信息，我们同样预先收集可能相关的图像，并使用多模态稠密检索模型为其建立索引。通过避免训练期间反复调用在线 API，这种语料库构建方式大幅降低了搜索 API 成本，同时仍能有效支持智能体搜索和工具调用行为的学习。离线语料库构建和检索实现的详细信息见附录。

多轮强化学习。 为了进一步激发模型的深度研究能力，我们使用多种工具，通过多轮强化学习（即，GRPO）对其进行优化。具体而言，给定一个问题 Q，策略模型会生成一组多轮搜索工具轨迹 $\{\tau_1, \tau_2, \ldots, \tau_G\}$。每条轨迹由一系列推理与规划步骤、工具调用、工具响应以及最终答案组成。

奖励计算。随后，我们为每条轨迹计算奖励 R_i ，τ_g 该奖励由搜索工具格式奖励R_i^{format} 和结果级准确率奖励构成 R_i^{acc}。(1)格式奖励。格式奖励通过要求符合预期的交互格式，鼓励结构化的多轮推理和有效的工具调用。符合第 τ 3.2 节中定义的推理格式的轨迹将获得 1 分，而违反格式要求或生成无法解析的工具调用查询的轨迹将获得 0 分。 (2) 准确率奖励。准确率奖励用于评估模型生成的最终答案是否正确。不同于采用基于规则的评估的以往方法，我们会提取模型的最终答案，并通过将其与标准答案进行比较，使用能力强大的大语言模型评估其正确性。最终奖励定义为() R_i ，其中用于平衡准确率奖励和格式奖励。随后，根据奖励计算组相对优势：

0。(2) 准确率奖励。准确率奖励用于评估模型生成的最终答案是否正确。不同于采用基于规则的评估的以往方法，我们会提取模型的最终答案，并通过将其与标准答案进行比较，使用能力强大的大语言模型评估其正确性。最终奖励定义为$R = \alpha R_{acc} +(1-\alpha)$ R_{format}，其中 α 用于平衡准确率奖励和格式奖励。随后，根据奖励计算组相对优势 A_g：

$$A_g = R_g - \frac{1}{G}\sum_{k=1}^{G} R_k. \tag{2}$$

优化。 最后，GRPO 使用带有 KL 惩罚项的裁剪目标来优化模型：

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{(I,T)\sim p_{\mathcal{D}}, \tau\sim\pi_{\theta_{\mathrm{old}}}(\cdot|I,T)} \left[\frac{1}{G}\sum_{i=1}^{G} \min\left(\frac{\pi_\theta(\tau_i \mid I,T)}{\pi_{\theta_{\mathrm{old}}}(\tau_i \mid I,T)} A_i, \mathrm{clip}\left(\frac{\pi_\theta(\tau_i \mid I,T)}{\pi_{\theta_{\mathrm{old}}}(\tau_i \mid I,T)}, 1-\epsilon, 1+\epsilon \right) A_i - \beta D_{\mathrm{KL}}\left(\pi_\theta || \pi_{\mathrm{ref}}\right) \right) \right]. \tag{3}$$

4.3. 评估协议

在测试时，我们将离线搜索工具切换为在线搜索工具。与离线搜索相比，在线搜索通常会返回冗长且包含大量冗余内容的网页内容。因此，我们使用一个辅助大语言模型，在将工具响应纳入推理过程之前对其进行核验和总结，避免过长的输出超出模型的上下文长度限制并导致推理失败。

5. 实验

5.1. 数据集。

Hyper-Search 3K 的来源。 为构建全面的问答数据集，我们从 7 个不同类别（即 艺术、体育、教育、历史、电影、地点和技术）收集视觉来源，作为 Hyper-Search 的起始图像节点，从而确保涵盖广泛领域。

训练数据集。 我们按如下方式准备监督微调和强化学习阶段的训练数据。(1) 对于监督微调，我们选用 InfoSeek（Chen 等，2023年）来生成搜索工具轨迹。在树搜索之前，我们筛选出对 Qwen3-VL-32B 来说较有挑战性的问答对，最终通过 DR-TTS 合成 10K 条有效搜索轨迹，用于监督微调。(2) 对于强化学习，我们使用 Hyper-Search 中的 3K 个样本，以及从 FVQA（Wu 等人，2025a）中额外筛选出的 3K 个搜索密集型样本。

评估数据集。我们在六个需要调用搜索工具才能解决的信息密集型基准测试上评估并比较 MM- DeepResearch

表*1.* **主要结果.** 为考察我们方法的有效性，我们将 MM-DeepResearch 与其他深度研究智能体，以及在直接推理和检索增强生成工作流范式下表现强大的多模态大语言模型进行比较。† 表示由我们自行评估的结果。

方法	SimpleVQA	MMSearch	LiveVQA	FVQA-test	InfoSeek	Browsecomp-VL	平均
直接推理							
GPT-4o（OpenAI，2024年）	51.7	18.7	28.1	48.0	52.9	5.5	34.1
GPT-5（OpenAI，2025a）	54.1	35.1	44.4	54.4	61.7	48.6	49.7
Qwen3-VL-8B（Bai 等人，2025年）	47.1	11.7	23.1	24.2	23.1	24.1	25.5
Qwen3-VL-32B（Bai 等人，2025年）	58.0	19.8	45.5	34.1	28.0	30.8	36.0
检索增强生成工作流							
GPT-4o（OpenAI，2024年）	63.6	49.1	40.1	66.3	59.5	13.4	48.7
GPT-5（OpenAI，2025a）	55.9	52.6	56.0	62.6	70.6	54.9	58.8
Qwen3-VL-8B（Bai 等人，2025年）	62.3	47.3	39.3	53.6	46.1	29.3	46.3
智能体搜索							
DeepMMSearch-R1-7B（Narayan 等人，2025年）	55.8	–	–	–	47.5	–	–
Visual-ARFT-7B (Liu 等人,2025b)	42.4	34.5	25.4	41.7	37.9	16.5†	33.1
MMSearch-R1-7B (Wu 等人,2025a)	57.4	53.8	48.4	58.4	55.1	20.9†	49.0
DeepEyes-v2-7B (Hong 等人,2025年)	59.4	63.7	–	60.6	51.1	–	–
WebWatcher-7B (Geng 等人,2025年)	54.3	49.1	51.2	–	–	21.2	–
MM-DeepResearch-7B	62.0	61.4	60.0	61.9	58.7	32.8	56.1
Qwen3-VL-8B（Bai 等人，2025年）	52.0	37.4	50.6	58.7	50.3	27.9	46.2
SenseNova-MARS-8B（Chng 等人，2025年）	61.7	67.4	56.2	67.1	70.1	35.1†	59.6
MM-DeepResearch-8B	65.9	67.8	65.0	69.2	72.3	37.9	63.0
Qwen3-VL-32B（Bai 等人， 2025年）	58.7	44.4	45.5	60.2	58.5	35.1	50.4
WebWatcher-32B（Geng 等人，2025年）	59.0	55.3	58.7	–	–	27.0	–
MM-DeepResearch 32B	**67.6**	**69.0**	**68.0**	**70.1**	**73.9**	**43.0**	**65.3**

。详细的数据集说明见附录。

5.2. 实现细节

搜索引擎。 本文详细介绍在线和离线搜索工具的实现。（1）在线搜索。我们采用 SerpAPI [1] 作为超图搜索和 DR-TTS 的在线搜索引擎，以从网络中检索文本和视觉信息，这得益于其强大而灵活的搜索功能。对于图像检索，我们直接从提供的图像 URL 下载相应的图像文件。对于文本检索，由于 SerpAPI 会返回网页 URL 列表，我们进一步使用 Jina Reader [2] 获取并解析相关网页的内容。（2）离线搜索。我们预先收集多模态网络信息，并构建支持基于文本和多模态检索的离线搜索引擎。基于文本的检索使用纯文本 E5 嵌入模型（Wang等人，2022），而多模态检索则通过 Jina-CLIP 嵌入（G¨unther 等人，2025年）实现。检索索引使用 FlashRAG（Jin 等人，2025b）构建。

基础模型。 为了检验我们密集搜索问答数据、搜索轨迹以及基于离线搜索引擎的训练的有效性，我们选取了两类多模态大语言模型：（1）不具备原生智能体工具使用能力的多模态大语言模型（即。，Qwen2.5-VL-7B-Instruct），用于评估我们的方法能否从零开始赋予模型智能体搜索能力；（2）具备原生智能体工具使用能力的多模态大语言模型（即。，Qwen3-VL-8B 和 32B-Instruct），用于检验我们的方法能否在现有基线的基础上进一步提升性能。

训练设置。对于监督微调，我们使用 LLaMA-Factory（Zheng 等人，2024年）以批大小 128、学习率 5e-6 对模型进行全量微调，训练 3 轮。对于强化学习，我们使用 VeRL（Sheng 等人，2024年），全局批大小为 128，每个提示进行 5 次轨迹采样，学习率为 1e-6。每条轨迹最多允许调用工具 5 次。为处理较长的多模态工具响应并增强长上下文信息整合能力，我们将最大训练上下文长度扩展至 70,000 个词元。对于较小的多模态大语言模型（例如，7B 和 8B），监督微调和强化学习均在 8 张 NVIDIA H100 GPU 上进行；较大的模型（例如，32B）则在 32 张 NVIDIA H100 GPU 上训练。此外，我们分配 8 张 NVIDIA H100 GPU 来部署离线搜索引擎和评判模型（即，通义千问*3*-Next-80B-A3B-Instruct（团队，2025年））。

5.3. 主要结果

为评估我们的方法，我们在不具备和具备原生智能体工具使用能力的模型上进行实验，并将 MM-DeepResearch 与多种最先进方法进行比较，包括表1中的非智能体和智能体搜索多模态大语言模型。

[1]SerpAPI：https://serpapi.com/ [2]Jina Reader：https://jina.ai/reader/

表 *2.* **Hyper-Search 的消融研究**。

方法	平均工具调用	MMSearch
InfoSeek	1.6	62.1
基于图的方法	1.7	64.8
Hyper-Search	2.3	67.8

表 *3*。**MM-DeepResearch 的消融研究**。

方法	DR-TTS（监督微调）	Hyper-Search（强化学习）	MMSearch
通义千问3-VL（基线）	–	–	37.4
MM-DeepResearch-SFT	✔	–	52.3
MM-DeepResearch-RL	–	✔	65.2
MM-DeepResearch	✔	✔	67.8

与非智能体模型的比较。我们首先在缺乏原生工具调用能力的 Qwen2.5-VL-7B 上进行实验。通过在我们合成的搜索轨迹和生成的问答数据上训练，MM-DeepResearch-7B 学会了智能体搜索能力，能够进行多步推理并协调使用多种工具来检索所需信息、综合得出最终答案，从而显著提升其在深度研究任务上的表现。在相同的基础模型下，MM-DeepResearch-7B 的表现优于此前的智能体搜索多模态大语言模型 Visual-ARFT 和 MMSearch-R1-7B，在六项基准测试上的平均提升分别为 23% 和 7.1%。此外，在 SimpleVQA 和 MM-Search 上，其表现分别超过此前 7B 最先进方法 WebWatcher 7.7% 和 12.3%。这些结果证明，我们的方法能够有效地从零开始激发智能体搜索能力。</final>

与智能体式模型相比。我们进一步在具备基本工具使用能力的智能体基础多模态大语言模型（即 Qwen3-VL-8B 和 Qwen3-VL-32B）上评估我们的方法，以考察我们的方法能否进一步提升其性能。如表 1所示，MM-DeepResearch-8B 的平均表现比基线通义千问3-VL-8B 高 17%。与深度研究智能体 SenseNova-MARS-8B 相比，MM-DeepResearch-8B 的平均得分高出 3.4 分，在 SimpleVQA 上提升了 4.2%。此外，将模型扩展至 32B 后，MM-DeepResearch-32B 相较于基线通义千问3-VL-32B 仍取得了 14.9% 的稳定性能提升，进一步验证了其提升智能体基础多模态大语言模型性能的有效性。

5.4. 消融研究

Hyper-Search 的消融研究。我们分析了不同数据生成方法如何影响搜索深度（*i*。*e*.，即平均工具调用次数）以及表 2 中的性能。我们比较了使用不同数据集进行强化学习训练的结果，即 InfoSeek、基于图的方法和 Hyper-Search；每个数据集均包含 3K 个样本，并以我们的监督微调模型为初始模型。结果表明，使用 Hyper-Search 构建的数据进行强化学习训练，能够促使模型展现更深入的探索行为，增加搜索轮次，并基于更全面的信息给出更准确的答案。

MM-DeepResearch 的消融研究。我们在表格中研究了 MM-DeepResearch 中不同组件的贡献。3。以 Qwen3-VL 基线（在 MMSearch 上得分为 37.4）为起点，使用 DR-TTS 探索得到的搜索轨迹进行训练

表 *4.* **离线搜索工具的消融研究**。

基于信息			基于知识的	MMSearch
T2T	T2I	I2T	T2T	
–	–	–	–	11.7
✔	–	–	–	55.9
✔	✔	–	–	64.7
✔	✔	✔	–	66.9
✔	✔	✔	✔	67.8

用于监督微调的轨迹显著提升了性能，得分达到 52.3，这表明我们的搜索轨迹数据有效增强了搜索行为和信息整合能力。仅使用 Hyper-Search 数据进行强化学习训练时，MM-DeepResearch-RL 进一步将性能提升至 65.2，说明该数据集能够激励多轮智能体搜索。通过联合利用 DR-TTS 轨迹数据进行监督微调，以及 Hyper-Search 问答数据进行强化学习，MM-DeepResearch 达到了最佳性能，在 MMSearch 上得分为 67.8，验证了我们所提方法的有效性。

搜索工具消融研究。表4展示了离线搜索工具的消融研究。启用基于信息的搜索后，MMSearch 的性能逐步提升，其中文本到文本、文本到图像和图像到文本搜索使得分从 11.7 提升至 66.9。加入基于知识的文本到文本搜索后，性能进一步提升至 67.8，这表明基于信息的搜索与基于知识的搜索在多模态推理中具有互补作用。

6. 结论

本文介绍了 MM-DeepResearch，这是一种面向深度研究任务的智能体，集推理与规划、多轮搜索工具调用以及长上下文跨模态信息整合于一体。为构建 MM-DeepResearch，我们引入了三个核心组件：用于搜索密集型问答生成的 Hyper-Search、用于合成有效搜索轨迹的 DR-TTS，以及用于可扩展强化学习训练的离线搜索引擎。这些设计共同支持从零开始训练搜索智能体，并显著增强其搜索和推理能力。大量实验表明，我们的方法对于具备和不具备原生智能体能力的模型均有效。我们希望这项工作能为未来多模态深度研究智能体的研究提供一个简单而有效的基线。

参考文献

An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., Wu, C., Tan, H., Li, C., Yang, J., Yu, J., Wang, X., Qin, B., Wang, Y., Yan, Z., Feng, Z., Liu, Z., Li, B., and Deng, J. Llava-onevision-1.5: 面向普惠化多模态训练的完全开放框架，2025年。URLhttps://arxiv.org/abs/2509.23661。

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl技术报告。*arXiv* 预印本*arXiv:2511.21631*，2025年。

Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., and Chang, M.-W. 预训练的视觉与语言模型能否回答视觉信息查询问题？*arXiv* 预印本 *arXiv:2302.11713*，2023年。

Chen, Z., Xu, C., Qi, Y., and Guo, J. 多模态大语言模型是强大的重排序器：通过知识增强重排序和注入噪声的训练，推动多模态检索增强生成。*arXiv* 预印本 *arXiv:2407.21439*，2024年。

Cheng, X., Zhang, W., Zhang, S., Yang, J., Guan, X., Wu, X., Li, X., Zhang, G., Liu, J., Mai, Y., 等。SimpleVQA：面向多模态大语言模型的多模态事实性评估。发表于*IEEE/CVF* 国际计算机视觉大会论文集，第 4637–4646 页，2025年。

Chng, Y. X., Hu, T., Tong, W., Li, X., Chen, J., Yu, H., Lu, J., Guo, H., Deng, H., Xie, C., 等。Sensenova-mars：通过强化学习赋能多模态智能体推理与搜索。*arXiv* 预印本 *arXiv:2512.24330*，2025年。

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 等。Gemini 2.5：以高级推理、多模态、长上下文及新一代智能体能力推动前沿发展。*arXiv* 预印本 *arXiv:2507.06261*，2025年。

Compagnoni, A.、Morini, M.、Sarto, S.、Cocchi, F.、Caffagni, D.、Cornia, M.、Baraldi, L. 和Cucchiara, R. Reag：面向基于知识的视觉问题回答的推理增强生成。*arXiv* 预印本 *arXiv:2511.22715*，2025年。

Fu, M.、Peng, Y.、Liu, B.、Wan, Y. 和 Chen, D. LiveVQA：实时视觉知识寻求。*arXiv* 预印本 *arXiv:2504.05288*，2025年。Geng, X.、Xia, P.、Zhang, Z.、Wang, X.、Wang, Q.、Ding, R.、Wang, C.、Wu, J.、Zhao, Y.、Li, K. 等。WebWatcher：开创视觉语言深度研究智能体的新前沿。*arXiv* 预印本 *arXiv:2508.05748*，2025年。

G¨unther, M., Sturua, S., Akram, M. K., Mohr, I., Ungureanu, A., Wang, B., Eslami, S., Martens, S., Werk, M., Wang, N., 等。jina-embeddings-v4：用于多模态多语言检索的通用嵌入。载于多语言表征学习第五届研讨会（*MRL 2025*）论文集，第 531–550 页，2025年。

Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., 等。Seed1. 5-vl 技术报告。*arXiv* 预印本 *arXiv:2505.07062*，2025a。

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., 等。Deepseek-r1：通过强化学习激励大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*，2025b。

Hong, J., Zhao, C., Zhu, C., Lu, W., Xu, G. 和 Yu, X. Deepeyesv2：迈向智能体多模态模型。*arXiv* 预印本 *arXiv:2511.05271*，2025年。

Huang, A., Yao, C., Han, C., Wan, F., Guo, H., Lv, H., Zhou, H., Wang, J., Zhou, J., Sun, J. 等。Step3-vl-10b 技术报告。*arXiv* 预印本 *arXiv:2601.09668*，2026年。

Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., 和 Lin, S。Vision-r1：激励多模态大语言模型的推理能力。*arXiv* 预印本 *arXiv:2503.06749*，2025年。

Jaech, A.,Kalai, A., Lerer, A., Richardson,A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Car- ney, A. 等。OpenAI o1 系统卡。*arXiv* 预印本 *arXiv:2412.16720*，2024年。

Jiang, D., Zhang, R., Guo, Z., Wu, Y., Lei, J., Qiu, P., Lu, P., Chen, Z., Fu, C., Song, G., 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。

Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H., 和 Han, J. 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*，2025a。

Jin, J., Zhu, Y., Dou, Z., Dong, G., Yang, X., Zhang, C., Zhao, T., Yang, Z., 和 Wen, J.-R. Flashrag：一个用于高效检索增强生成研究的模块化工具包。载于*ACM 2025*年万维网大会配套论文集，第 737–740 页，2025b。

Li, S., Bu, X., Wang, W., Liu, J., Dong, J., He, H., Lu, H., Zhang, H., Jing, C., Li, Z., 等。Mm-browsecomp：面向多模态浏览智能体的综合基准。*arXiv* 预印本 *arXiv:2508.13186*，2025a。

Li, X., Dong, G., Jin, J., Zhang, Y., Zhou, Y., Zhu, Y., Zhang, P., and Dou, Z. Search-o1：搜索增强型智能体大型推理模型。*arXiv*预印本 *arXiv:2501.05366*，2025b。

Li, Y., Zhang, W., Yang, Y., Huang, W.-C., Wu, Y., Luo, J., Bei, Y., Zou, H. P., Luo, X., Zhao, Y., 等。迈向具备深度推理能力的智能体检索增强生成：大语言模型中检索增强生成推理系统综述。*arXiv*预印本 *arXiv:2507.09477*，2，2025c。

Liang, J., Lin, H., Wu, Y., Zhao, R., Li, Z., 等。通过系统1或系统2进行推理型检索增强生成：面向行业挑战的智能体检索增强生成推理综述。载于第*14*届国际自然语言处理联合会议暨第*4*届计算语言学协会亚太分会会议论文集，第1954–1966页，2025年。Liu, H., Li, C., Li, Y., and 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。Liu, Z., Zang, Y., Zou, Y., Liang, Z., Dong, X., Cao, Y., Duan, H., Lin, D., and Wang, J. 视觉智能体强化微调。*arXiv*预印本 *arXiv:2505.14246*，2025a。Liu, Z., Zang, Y., Zou, Y., Liang, Z., Dong, X., Cao, Y., Duan, H., Lin, D., and Wang, J. 视觉智能体强化微调。*arXiv*预印本 *arXiv:2505.14246*，2025b。Narayan, K., Xu, Y., Cao, T., Nerella, K., Patel, V. M., Shiee, N., Grasch, P., Jia, C., Yang, Y., and Gan, Z. DeepMMSearch-R1：赋能多模态大语言模型进行多模态网络搜索。*arXiv*预印本 *arXiv:2510.12801*，2025年。OpenAI。GPT-4o。 https://openai.com/index/hello-gpt-4o/，2024年。OpenAI。推出GPT-5。 https://openai.com/index/introducing-gpt-5/，2025a。OpenAI。推出深度研究。 https://openai.com/index/introducing-deep-research/，2025b。

Seed, B. Seed1. 8 模型卡：迈向通用的真实世界智能体能力。技术报告，2025a. 技术报告。

Shang, N., Liu, Y., Zhu, Y., Zhang, L. L., Xu, W., Guan, X., Zhang, B., Dong, B., Zhou, X., Zhang, B., 等。rstar2-agent：智能体推理技术报告。*arXiv* 预印本 *arXiv:2508.20722*，2025年。

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. HybridFlow：一种灵活高效的基于人类反馈的强化学习框架。*arXiv* 预印本 *arXiv: 2409.19256*，2024年。

Team, K., Bai, Y., Bao, Y., Chen, G., Chen, J., Chen, N., Chen, R., Chen, Y., Chen, Y., Chen, Y., 等。Kimi K2：开放智能体智能。*arXiv* 预印本 *arXiv:2507.20534*，2025a。

Team, M., Bai, S., Bing, L., Chen, C., Chen, G., Chen, Y., Chen, Z., Chen, Z., Dong, X., 等。MiroThinker：通过模型、语境和交互式扩展，推动开源研究智能体的性能边界。*arXiv* 预印本 *arXiv:2511.11793*，2025b。

Team, M. L., Gui, A., Li, B., Tao, B., Zhou, B., Chen, B., Zhang, C., Gao, C., Zhang, C., Han, C., 等。Longcat-flash-thinking-2601 技术报告。*arXiv* 预印本 *arXiv:2601.16725*，2026a。Team, Q. Qwen3 技术报告，2025年。网址https://arxiv.org/abs/2505.09388。Team, V., Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., Duan, S., Wang, W., Wang, Y., Cheng, Y., He, Z., Su, Z., Yang, Z., Pan, Z., Zeng, A., Wang, B., Chen, B., Shi, B., Pang, C., Zhang, C., Yin, D., Yang, F., Chen, G., Li, H., Zhu, J., Chen, J., Xu, J., Xu, J., Chen, J., Lin, J., Chen, J., Wang, J., Chen, J., Lei, L., Gong, L., Pan, L., Liu, M., Xu, M., Zhang, M., Zheng, Q., Lyu, R., Tu, S., Yang, S., Meng, S., Zhong, S., Huang, S., Zhao, S., Xue, S., Zhang, T., Luo, T., Hao, T., Tong, T., Jia, W., Li, W., Liu, X., Zhang, X., Lyu, X., Zhang, X., Fan, X., Huang, X., Xue, Y., Wang, Y., Wang, Y., Wang, Y., An, Y., Du, Y., Huang, Y., Niu, Y., Shi, Y., Wang, Y., Wang, Y., Yue, Y., Li, Y., Liu, Y., Zhang, Y., Wang, Y., Zhang, Y., Xue, Z., Du, Z., Hou, Z., Wang, Z., Zhang, P., Liu, D., Xu, B., Li, J., Huang, M., Dong, Y., and Tang, J. GLM-4.5V 和 GLM-4.1V-Thinking：借助可扩展强化学习实现通用多模态推理，2026b。网址 https://arxiv.org/abs/2507.01006。Wang, B., Wang, B., Wan, C., Huang, G., Hu, H., Jia, H., Nie, H., Li, M., Chen, N., Chen, S., 等。Step-3 虽然规模庞大，但价格实惠：面向高性价比解码的模型-系统协同设计。*arXiv* 预印本 *arXiv:2507.19427*，2025a。Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. 基于弱监督对比预训练的文本嵌入。*arXiv* 预印本 *arXiv:2212.03533*，2022年。

Wang, P., Wu, Q., Shen, C., Dick, A., and Van Den Hengel, A. FVQA：基于事实的视觉问答。*IEEE*《模式分析与机器智能汇刊》，40(10):2413–2427，2017年。

Wang, Q., Ding, R., Zeng, Y., Chen, Z., Chen, L., Wang, S., Xie, P., Huang, F., and Zhao, F. Vrag-rl：通过结合强化学习的迭代推理，赋能基于视觉感知的检索增强生成，以理解视觉丰富的信息。*arXiv* 预印本 *arXiv:2505.22019*，2025b。

Wang, Y., Wang, L., Deng, Y., Wu, K., Xiao, Y., Yao, H., Kang, L., Ye, H., Jing, Y., and Bing, L. Deepresearcheval：一个用于深度研究任务构建和智能体评估的自动化框架。*arXiv* 预印本 *arXiv:2601.09688*，2026年。

Wei, J., Zhou, H., Zhang, X., Zhang, D., Qiu, Z., Wei, W., Li, J., Ouyang, W., and Sun, S. Alignrag：利用批判学习实现证据敏感的检索增强推理。*arXiv*预印本 *arXiv:2504.14858*，2025年。Wu, J., Deng, Z., Li, W., Liu, Y., You, B., Li, B., Ma, Z., and Liu, Z. MMSearch-R1：激励多模态大语言模型进行搜索。*arXiv*预印本 *arXiv:2506.20670*，2025a。Wu, J., Zhu, J., Liu, Y., Xu, M., and Jin, Y. 智能体式推理：一种利用智能体工具增强大语言模型推理能力的精简框架。载于计算语言学协会第 *63* 届年会论文集（第 *1* 卷：长篇论文），第 28489–28503 页，2025b。Xia, Y., Zhou, J., Shi, Z., Chen, J., and Huang, H. 通过自我推理改进检索增强语言模型。载于*AAAI* 人工智能会议论文集，第 39 卷，第 25534–25542 页，2025年。Xiao, B., Xia, B., Yang, B., Gao, B., Shen, B., Zhang, C., He, C., Lou, C., Luo, F., Wang, G., 等。Mimo-v2-flash 技术报告。*arXiv*预印本 *arXiv:2601.02780*，2026年。Xiao, Z., Yu, Q., Li, B., Chen, G., Chen, C., and Zhang, W. M2io-r1：一种用于多模态检索增强多模态生成的高效强化学习增强推理框架。*arXiv*预印本 *arXiv:2508.06328*，2025年。Yang, Z., Pan, B., Wang, H., Wang, Y., Liu, X., Weng, L., Feng, Y., Feng, H., Zhu, M., Zhang, B., 等。多模态深度研究者：从零开始生成文本与图表交错报告的智能体框架。*arXiv*预印本 *arXiv:2506.02454*，2025年。Yao, H., Huang, J., Wu, W., Zhang, J., Wang, Y., Liu, S., Wang, Y., Song, Y., Feng, H., Shen, L., 等。Mulberry：通过集体蒙特卡洛树搜索，赋能多模态大语言模型类 o1 推理与反思。*arXiv*预印本 *arXiv:2412.18319*，2024a。

Yao, H., Wu, W., Yang, T., Song, Y., Zhang, M., Feng, H., Sun, Y., Li, Z., Ouyang, W., and Wang, J. 面向多模态大语言模型的稠密连接器。神经信息处理系统进展，37:33108–33140，2024b。

Yao, H., Huang, J., Qiu, Y., Chen, M. K., Liu, W., Zhang, W., Zeng, W., Zhang, X., Zhang, J., Song, Y., et al. Mm-reason：一个面向通用人工智能的开放式多模态大语言模型多模态多步推理基准。发表于*IEEE/CVF*国际计算机视觉会议论文集，第 273–283 页，2025a。

Yao, H., Yin, Q., Zhang, J., Yang, M., Wang, Y., Wu, W., Su, F., Shen, L., Qiu, M., Tao, D., 等。R1-sharevl：通过 share-grpo 激励多模态大语言模型的推理能力。载于第三十九届神经信息处理系统年会，2025b。Yao, H., Zhang, R., Huang, J., Zhang, J., Wang, Y., Fang, B., Zhu, R., Jing, Y., Liu, S., Li, G., 等。智能体多模态大语言模型综述。*arXiv* 预印本 *arXiv:2510.10991*，2025c。Zhang, J., Huang, J., Yao, H., Liu, S., Zhang, X., Lu, S. 和 Tao, D.。R1-vl：通过逐步组相对策略优化，学习如何运用多模态大语言模型进行推理。*arXiv* 预印本 *arXiv:2503.12937*，2025a。Zhang, W., Li, Y., Bei, Y., Luo, J., Wan, G., Yang, L., Xie, C., Yang, Y., Huang, W.-C., Miao, C., 等。从网络搜索迈向智能体深度研究：利用推理智能体激励搜索。*arXiv* 预印本 *arXiv:2506.18959*，2025b。Zhang, Y., Zeng, Y., Li, Q., Hu, Z., Han, K. 和 Zuo, W.。Tool-r1：面向智能体工具使用的高样本效率强化学习。*arXiv* 预印本 *arXiv:2509.12867*，2025c。Zhao, S., Zhang, H., Lin, S., Li, M., Wu, Q., Zhang, K. 和 Wei, C.。Pyvision：借助动态工具调用实现智能体视觉。*arXiv* 预印本 *arXiv:2507.07998*，2025。Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z. 和 Ma, Y.。LlamaFactory：统一高效地微调 100+ 语言模型。载于计算语言学协会第六十二届年会论文集（第三卷：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。Zheng, Y., Fu, D., Hu, X., Cai, X., Ye, L., Lu, P. 和 Liu, P.。Deepresearcher：在真实世界环境中通过强化学习扩展深度研究。*arXiv* 预印本 *arXiv:2504.03160*，2025年。

附录

A. 多模态问答数据集

为了构建全面的问答数据集，我们从七个不同类别（*i.e.*, 艺术、体育、教育、历史、电影、地点和技术）收集视觉素材，作为 Hyper-Search 的起始图像节点，确保广泛覆盖各个领域。图 4 ~ 展示了我们的多模态问答数据集中的一些示例。10,



图 *4.* 历史类别的数据示例。



图 *5.* 艺术类别的数据示例。



图 *6.* 教育类别的数据示例。

电影



节点1
问题： 这里出现的金发角色叫什么名字，是一个咧嘴微笑的表情符号角色
来自这部以表情符号为主题的动画电影？
答案： 微笑者（《表情奇幻冒险》）



节点2
问题： 金发的欢快角色
表情符号角色
海报中展示的角色——在《表情奇幻冒险》中，
她的角色是什么？由谁为她配音？
答案： 主要反派；配音演员是
玛娅·鲁道夫



Node3 问题： 电影海报上显示的标题为“《微笑》”，官方海报宣布的美国影院上映日期是什么？ **答案：** 9月30日



节点4 问题： 图像展示了所描绘电影的海报；在该片制作期间，曼哈顿的哪栋建筑被用作外景拍摄地？
答案： 安索尼亚酒店

图7. 电影类别的数据示例。

地点



节点1 问题： 图中带有玻璃金字塔的博物馆，星期三的开放时间和最晚入场时间分别是什么？
答案： 星期三：9:00–21:00；最晚入场时间为20:00



节点2
问题： 在卢浮宫展厅中
展出雅克·路易·大卫的作品
在这里看到的画作中，哪两件重要的
借展作品在这场
展览中展出？
答案： 网球厅宣誓
（凡尔赛宫）和《马拉之死》原作
（皇家美术博物馆
比利时，布鲁塞尔）



Node3
问题： 在这幅骑马肖像画中，
一位骑在扬起前蹄的白马上的将军
正穿越阿尔卑斯山，哪些历史
领袖是这幅图像旨在
根据艺术史分析，将他与
联系起来？
答案： 汉尼拔和查理曼



节点4
问题： 这座发光玻璃
金字塔是这里展示的主要
这座博物馆的入口——
由谁设计，以及它是在哪项
扩建计划下建成的？
答案： 贝聿铭；大卢浮宫
项目（1980年代至1990年代）

图8. 地点类别的数据示例。

技术



节点1 问题： 图中这款可同时看到大尺寸内屏和小尺寸外盖屏幕的可折叠智能手机。其封面显示屏支持多高的刷新率？ **答案：** 165Hz



节点2
问题： 这款青绿色的摩托罗拉手机
展示着一个 2×2 四摄模组
类似 Edge 60 Pro；对于美国的购物者来说，
这款机型在
未出售，哪两款摩托罗拉手机
是推荐的替代机型？
答案： Motorola Razr Ultra 2025；
Moto G75 5G



Node3
问题： 蓝色的摩托罗拉
所示智能手机有一个方形摄像头
模组，带有四个镜头；哪些官方
耐用性认证该机型
具备？
答案： IP69和MIL-STD-810H



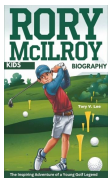
节点4
问题： 在菲律宾
市场，所展示的这款电动轿车的官方整车保修期限（年数和总
公里数）是多少？这款车是 2026年后
轮驱“Advance”版
吗？
答案： 6年 / 150,000公里

图9. 技术类别的数据示例。

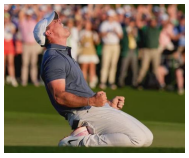
体育



节点1 问题： 图像中，一位职业高尔夫球手在海滨球场手持水晶奖杯；根据他的官方赛程，他计划于2026年何时何地参加 AT&T 圆石滩职业业余配对赛？ **答案：** 2026年2月12日至15日，美国加利福尼亚州圆石滩



Node2 问题： 这本书籍封面介绍了一位高尔夫球手的传记。根据 PGA TOUR Studios 的视频《The Long Game》，这位球手取得了什么值得瞩目的职业生涯里程碑？ **答案：** 第六位完成职业生涯大满贯的球手



Node3 问题： 对于在果岭上庆祝的这位高尔夫球手，根据同款高尔夫球服的相关图像，他通常佩戴的白色高尔夫手套上是什么品牌标志？ **答案：** 耐克勾形标志



节点4 问题： 图像展示了一张宣传图，宣布两名美国职业棒球大联盟球员为汉克·阿伦奖得主；这些获奖者是在哪一年获得认可的？
答案： 2024年

图10. 体育类别的数据示例。

B. 更多设置

离线搜索引擎实现。 为支持离线语料库搜索，我们预先收集多模态语料库，并采用基于 FlashRAG（Jin等人，2025b）的检索框架。对于文本语料库，我们提示 GPT 针对每个问题生成多个候选搜索查询，并使用 SerpAPI[3] 和 Jina[4] 检索相应的网页内容。此外，我们纳入大规模英文维基百科（2018年），以提高语料库的覆盖范围和多样性，从而模拟网络上可获取的大规模信息。这些网络内容被切分为附有标题的 512 词段落，并采用 e5-base-v2（Wang等人，2022年）作为文本检索器。对于图像语料库，我们还提示 GPT 生成多个候选图像搜索查询，并使用 SerpAPI 检索相关图像 URL。我们采用 jina-embeddings-v4（G¨unther等人，2025年）作为多模态检索器，以实现跨模态检索。

搜索引擎的超参数。在训练期间，我们会为每个查询检索相似度最高的前 3 个文本段落。对于图像检索，仅保留嵌入相似度大于 0.7 的图像；对于每次工具调用查询，我们只保留相似度最高的单张图像，以避免推理语境过长和内存消耗过高。在评估期间，我们采用在线检索，将文本搜索扩展至前 5 个结果，并为每个查询最多检索三张图像。

C. 评估基准

我们在六个基准测试上评估 MM-DeepResearch，具体如下所述。

- **SimpleVQA**。SimpleVQA（Cheng 等人，2025年）是一个新提出的多模态基准，旨在评估多模态大语言模型在回答以图像为依据的简短自然语言问题时的事实性。该基准涵盖多样化的任务和场景，包含高质量且具有挑战性的查询，并围绕多个主题进行对齐。该基准提供静态且可验证的参考答案，并采用严格的质量控制，因此能够以极小的差异便捷地评估模型输出。SimpleVQA 结合人工验证和大语言模型评判打分系统，促进了对领先多模态大语言模型和纯文本大语言模型进行全面的实证比较，从而深入了解图像理解、生成准确性以及常见的失败模式。遵循 WebWatcher（Geng 等人，2025年），我们在从 1,013 个英文问答对中抽样的同一批 300 个样例上进行评估。

- **MMSearch**。 MMSearch（Jiang 等，2024年）是一个多模态浏览基准，旨在对必须整合视觉与文本信号，并结合外部检索来回答复杂问题的模型提出严格要求。不同于以往显著物体可能很容易辨认的多模态任务，该基准要求模型对局部视觉线索进行全面推理，有条不紊地跨模态收集证据，在充满噪声的搜索结果中核验来源，并开展漫长且借助工具的推理链。其中，171 个样本包含图像；为与以往方法保持一致，我们在这 171 个基于图像的样本上进行评估。

- **LiveVQA**。 LiveVQA（Fu等人，2025年）是一个大规模视觉问答数据集，聚焦于最新的视觉知识。该数据集涵盖多个现实世界类别，包含超过 107,000 个样本，来源包括近期新闻、视频和学术平台，因此适合作为基准，用于评估多模态大语言模型如何处理超出其训练截止时间的实时视觉信息。该数据集支持对当前模型的性能差距，以及兼顾新视觉知识整合与核心感知和推理能力的参数高效微调策略进行研究。我们还在从 WebWatcher 中抽取的 300 个样本子集上进行了评估。

- **FVQA**。FVQA（Wang等人，2017年）是一个基础性的视觉问答数据集。它扩展了传统的视觉问答任务，要求利用外部常识知识正确回答视觉问题。在 FVQA 中，每个图像–问题–答案三元组都附有一个支持事实，该事实以结构化知识三元组的形式表示，取自常识知识库。这一设计促使模型超越视觉感知进行推理，并检索和整合事实性知识，因此成为基于知识的视觉问答研究中涉及显式外部知识落地的核心基准。我们在此使用完整的 1,800 个高质量样例进行测试。

- **InfoSeek**。 InfoSeek（Chen 等，2023年）是一个视觉问答数据集，专注于信息检索问题——这类问题无法仅凭常识或图像中呈现的视觉内容解答。该数据集结合了人工标注和大规模自动生成的问题–答案对，给模型带来了挑战

[3]SerpAPI：https://serpapi.com/ [4]Jina Reader：https://jina.ai/reader/

，要求模型利用超越典型视觉问答任务的细粒度知识。使用 InfoSeek 进行评估，凸显了当前预训练视觉语言模型在回答知识密集型问题方面的局限性，也表明通过针对明确的信息检索目标进行微调能够带来性能提升。在此，我们使用 MMSearch-R1（Wu 等人，2025a）抽取的 2,000 个样本进行评估。

• **Browsecomp-VL**。 BrowseComp-VL（Geng 等人，2025年）是最近推出的一个基准，将 BrowseComp 风格的任务扩展到多模态领域，要求在文本、图像和网络内容之间进行策略规划与跨模态检索。BrowseComp-VL 旨在评估复杂的多模态智能体，要求其执行深度搜索、序贯证据综合和多模态推理。通过将视觉和文本线索嵌入长期信息搜寻任务中，该基准评估智能体在复杂现实场景下协调感知、检索与规划的能力，并作为整合视觉语言深度研究的评估前沿。我们在完整的 BrowseComp-VL 数据集上进行评估。

D. 更多讨论

D.1. 关于离线搜索引擎中不同嵌入模型的讨论

我们研究了不同嵌入模型对离线文本检索的影响，结果见表 5。总体而言，两种嵌入模型之间的性能差异很小，表明离线文本检索对嵌入模型的选择相对稳健。尽管 jina-embeddings-v4（G¨unther 等人，2025年）的性能略好，但由于其采用高维嵌入且推理速度较慢，因此成本更高。综合考虑效率和可扩展性，我们因此采用 e5-base-v2（Wang等人，2022）进行文本嵌入；在大规模离线索引和检索方面，它能够提供相当的检索质量，同时效率显著更高。

表 *5.* **关于离线搜索引擎所用不同嵌入模型的讨论**。

嵌入模型	MMSearch
e5-base-v2	55.9
jina-embeddings-v4	56.1

D.2. 评估时在线搜索与离线搜索性能的比较

我们在评估过程中比较了在线与离线搜索引擎的性能，如表6所示。总体而言，在线搜索始终取得更好的性能，这得益于其能够获取最新且大规模的网络信息。尽管如此，离线搜索引擎在SimpleVQA和MMSearch上均取得了具有竞争力的结果，表明其能够检索到信息丰富的相关证据。这些结果表明，尽管离线搜索无法完全达到在线搜索的性能，但它足以提供有意义的检索信号，并能够充分促使模型学习搜索行为。这使得离线搜索成为训练中一种实用且成本高效的替代方案。

表*6.* **离线文本搜索引擎的讨论**。

搜索引擎	SimpleVQA	MMSearch
离线	63.4	62.7
在线	65.9	67.8

D.3. 关于训练期间Top-k 文本检索影响的讨论

我们分析了训练期间检索到的文本段落数量（k）变化所产生的影响，结果如表 7所示。总体而言，适中的 k 能够带来更好的性能，而将 k 增加到一定程度后，收益便会递减。具体来说，检索五个段落时，MMSearch 的性能最佳，优于较小（$k = 3$）和较大（$k = 10$）的设置。当 k 过大时，纳入过多且可能含有噪声的信息，可能会稀释相关证据并增加推理难度，进而导致性能下降。

表7. **离线文本搜索引擎讨论**。

嵌入模型	Top-k	MMSearch
e5	3	55.9
e5	5	56.3
e5	10	54.8

D.4. 在线与离线搜索成本讨论。

表 8 比较了强化学习期间在线搜索与离线搜索的训练成本。在线搜索成本高昂（每百步约 640 美元），这是因为每次轨迹采样中都会反复调用外部 API，同时还会带来较高延迟：每个样本的平均响应时间约为 60 秒。相比之下，离线搜索无需额外的搜索相关成本，延迟也显著更低（每个样本约 1 秒）。这些结果表明，离线搜索是一种兼具成本效益且可扩展的替代方案，尤其适用于需要多次轨迹采样和大规模训练的 GRPO 设置。

表8. **在线与离线搜索的训练成本和效率讨论**。

方法	训练成本（$）	训练时间（秒/样本）
在线搜索	640	60
离线搜索	0	1

E. 案例研究

我们在图11中展示了 MM-DeepResearch 的推理案例。在推理轨迹中，MM-DeepResearch 会交替进行思考和工具调用。在工具使用方面，流程首先使用图像到图像搜索工具，从原始图像中检索对应的网址及相关元数据。随后使用图像到文本搜索工具，收集补充视觉数据，以增强推理能力。最后，调用文本到文本搜索工具，检索相关文本信息。最终，通过思考整合所有收集到的信息，得出最终答案。

User



‹图像›请根据系统指令回答以下问题。问题：这座建筑的建筑师是谁?

步骤1：思考

‹思考› 用户提供了一座建筑的图像，并询问其建筑师。要确定建筑师，我需要先通过分析其视觉特征以及可能的位置来识别该建筑。一旦获得该建筑的名称或描述，我就可以搜索其建筑师。我将使用图像_搜索_按_镜头工具，根据图像识别该建筑。我将提供一个描述性查询，以帮助搜索引擎准确识别该建筑。 ‹/思考›

步骤2：工具调用

‹工具_调用› {"name": "图像_至_图像_搜索", "arguments":{"query_查询列表": ["有石墙和红色屋顶的历史要塞", "绿意盎然、树木环绕的景观中的建筑", "由多座建筑组成的欧洲风格防御工事"]}} ‹/工具_调用›"

工具响应





‹图像›‹图像›‹图像›‹工具_响应›{'result': '[图像搜索成功] 已成功检索到相关图像。相关标题为：File:兹巴拉日城堡 RB.jpg - 维基共享资源；历史悠久的兹巴拉日城堡石墙红色库存照片 2638312599 | Shutterstock；位于乌克兰西部捷尔诺波尔州兹巴拉日的中世纪堡垒，该城堡建于波兰立陶宛联邦时期 库存照片 - Alamy。检索到的视觉证据现在可用于下游多模态推理。'}‹/工具_响应›

步骤3：思考

‹think› 图像搜索已成功识别出该建筑为兹巴拉日城堡，位于乌克兰西部的捷尔诺波尔州。据描述，这是一座建于波兰立陶宛联邦时期的中世纪堡垒。既然我已经识别出这座建筑，接下来需要确定其建筑师。由于兹巴拉日城堡是一座历史堡垒，因此它可能并没有一位有明确记载姓名的建筑师，而是由当地建造者或军事工程师在贵族或地方当局的指导下，于17世纪建造的。为确认这一点，我可以搜索有关兹巴拉日城堡及其建筑师的历史记录或学术资料。 ‹/think›

步骤4：工具调用

‹工具_调用› {"name": "图像_转_文本_搜索", "arguments":{"query_列表": ["兹巴拉日城堡建筑师","谁在乌克兰建造了兹巴拉日城堡","兹巴拉日城堡17世纪的历史建造"]}} ‹/工具_调用›

工具响应











"‹图像›‹图像›‹图像›‹图像›‹图像›‹图像›‹图像›‹图像›‹工具_响应›[图像搜索成功] 已成功检索到相关图像。相关标题为：兹巴拉日城堡的蜕变：从要塞到博物馆；兹巴拉日城堡 - 维基百科；Mainly Museums -

博物馆城堡之谜；兹巴拉日城堡 - 维基百科；兹巴拉日城堡的蜕变：从要塞到博物馆；Mainly Museums - 博物馆城堡之谜；兹巴拉日城堡的蜕变：从要塞到博物馆；兹巴拉日城堡（2026年）- 出发前你需要知道的一切（含……；Mainly Museums - 博物馆城堡之谜。检索到的视觉证据现在可用于下游多模态推理。‹/工具_响应›"

步骤5：思考

‹think› 图像搜索已确认该建筑为位于乌克兰的兹巴拉日城堡。搜索结果还指向维基百科和其他可能包含其建造及建筑师信息的来源。为了确定建筑师，我将搜索提及兹巴拉日城堡建造过程的具体历史记录或学术资料，尤其关注其17世纪最初建造阶段的相关信息。现在，我将搜索“兹巴拉日城堡 建筑师”，以查找权威来源。 ‹/think›

步骤6：工具调用

‹工具_调用› {"name": "文本_转_文本_搜索", "arguments":{"query_列表": ["乌克兰兹巴拉日城堡的建筑师是谁"
“兹巴拉日城堡17世纪的建造历史”，“波兰立陶宛联邦时期兹巴拉日城堡的建造者或设计者”]}}‹/tool_调用›

工具响应

"‹工具_响应›{'result': '文档 1：乌克兰兹巴拉日城堡的建筑师是佛兰德建筑师 Hendrik van Peene。他于17世纪初设计了这座城堡，工程于1626年动工，并于1631年完工。这座城堡是一座文艺复兴风格的棱堡，专为波兰权贵克日什托夫和耶日·兹巴拉斯基建造，将军事防御工事与中央的两层宫殿结合在一起。\n文档 2：乌克兰的兹巴拉日城堡由兄弟克日什托夫和耶日·兹巴拉斯基于17世纪上半叶建造。文本中没有提及具体建筑师的姓名，但城堡采用文艺复兴晚期风格设计，具有坚固的防御设施。后来，Dymitr Wiśniowiecki 对城堡进行了重建；他是后来的所有者，而非最初的建筑师。\n\n---\n文档 1：兹巴拉日城堡于1620年至1631年间，在尤里和克里斯托弗——兹巴拉日最后几位亲王——的资助下建成。设计由意大利建筑师 VinchezzoScamozzi 完成。后来，维什涅韦茨基亲王加固了城堡。到了18世纪中叶，在波托茨基家族拥有期间，城堡逐渐衰败。\n文档 2：兹巴拉日城堡于1620年至1631年间，在尤里和克里斯托弗——兹巴拉日最后几位亲王——的资助下建成。该项目由意大利建筑师 Vinchezzo Scamozzi 设计。后来，维什涅韦茨基亲王加固了城堡。到了18世纪中叶，在波托茨基家族拥有期间，城堡逐渐衰败。\n文档 3：兹巴拉日城堡由兄弟克日什托夫和耶日·兹巴拉斯基于17世纪上半叶建造。\n\n---\n文档 1：兹巴拉日城堡由波兰兹巴拉斯基家族的最后几位成员克日什托夫和耶日·兹巴拉斯基，在波兰立陶宛联邦时期建造。\n文档 2：兹巴拉日城堡由兹巴拉斯基兄弟克日什托夫和耶日·兹巴拉斯基委托建造，于1626年至1631年间竣工，作为他们在波兰立陶宛联邦的主要居所。\n\n- **克日什托夫·兹巴拉斯基**于1620年启动该项目，并委托**威尼斯文艺复兴建筑师文森佐·斯卡莫齐**设计。\n- **耶日·兹巴拉斯基**负责监督城堡竣工，并委托**佛兰德军事工程师 Henryk van Peene**修建防御工事，以增强城堡的防御能力。\n\n因此，主要建造者/设计者是建筑师**文森佐·斯卡莫齐**，而**Henryk van Peene**负责防御工事。 \n'}‹/工具_响应›"

步骤7：思考

‹思考› 搜索结果中关于兹巴拉日城堡建筑师的信息相互矛盾。不过，最后一组搜索结果提供的说法最为详尽且前后一致：- 克日什托夫·兹巴拉斯基委托威尼斯文艺复兴建筑师**文森佐·斯卡莫齐**设计。- 耶日·兹巴拉斯基负责监督工程竣工，并委托佛兰德军事工程师**Henryk van Peene**设计防御工事。这表明，**文森佐·斯卡莫齐**是负责整体设计的主要建筑师，而**Henryk van Peene**则参与了军事防御工事的设计。鉴于问题询问的是“建筑师”，而斯卡莫齐被认为负责设计，因此他是最准确的答案。 ‹/思考› ‹答案›兹巴拉日城堡的建筑师是威尼斯文艺复兴建筑师文森佐·斯卡莫齐。。‹/答案›"

图*11.* MM-DeepResearch 案例研究。17

F. 提示词

F.1. 深度研究提示词

MM-DeepResearch 模型在训练和推理过程中使用的完整提示。

系统消息

你是一位乐于助人且无害的深度研究助手。你的任务是认真思考，在必要时查找外部信息，并针对用户的问题提供准确且有充分–支持的答案。

思考准则 1. 逐步推理，解决用户的问题。将原始问题拆解为清晰、易于处理的子–问题。 2. 每次推理循环后，总结目前已确定的内容，并判断是否需要更多子–问题或外部信息。 3. 你的思考过程必须始终保留在内部，并以结构化形式置于 <think>...</think> 中。

工具使用指南 1. 当需要外部信息才能准确回答问题时，请使用工具。2. 工具查询必须具体而明确。避免使用含糊的指代或代词（例如，“it”、“this”、“he”），并使用明确的实体名称、日期、技术术语或唯一标识符。3. 有效的工具使用取决于制定高–质量的查询，以及从工具响应中提取有用信息。4. 将所有工具调用括在 <工具调用>...</工具调用> 内，并将所有工具输出括在 <工具响应>...</工具响应> 内。

答案指南
1. 如果不需要外部信息或详细解释，请始终提供一个具体的最终答案，并将其括在 <答案>...</答案> 内（例如， <答案>北京</答案>）。

格式指南 助手可以按如下方式遵循有效的执行路径：<思考>推理</思考>（如果需要工具使用）<工具调用>工具调用</工具调用><工具响应>工具输出</工具响应>（如有必要，可重复上述步骤）<思考>最终推理</思考><答案>最终答案</答案>

工具
你可以调用一个或多个函数来协助处理用户查询。
以下 <tools></tools> XML 标签中提供了函数签名：
<tools>
{"type": "function", "函数": {"名称": "按文本查询搜索图片", "描述": "搜索根据给定的查询在网络上搜索图像，并返回相关图片结果及其对应的标题。 此工具只能使用一次。", "parameters": {"type": "object", "properties": {"查询列表": {"type": "array", "description": "用于图像搜索】【。该工具会根据此查询检索相关图像。"}}, "required": ["查询列表"]}}}
{"type": "函数", "函数": {"名称": "镜头图像搜索", "描述": "执行图像搜索，使用原始问题中的图像，并通过补充文本查询进行优化，且返回相关图像及其对应的标题。此工具只能使用一次。", "参数": {"类型": "对象", "属性": {"查询列表": {"类型": "数组", "描述": "一个列表用于配合图像搜索的文本查询。该工具会为此图像检索相关图像。"}}, "必需": ["查询列表"]}}}
{"type": "function", "function": {"name": "文本搜索", "description": "根据给定的查询在网络中搜索相关信息。", "parameters": {"type": "object", "properties": {" query list": {"type": "array", "description": " 一组完整 – 成形的语义查询。该工具将针对每个查询返回搜索结果。"}}, "required": ["querylist"]}}}{"type": "function", "function": {\"name\": \"模型搜索\", \"description\": \"查询专家模型 to 答案问题 根据给定查询的 ns。", "parameters": {"type": "object", "properties": {"

查询列表": {"type": "array", "description": "一组完整的–语义查询。该工具将返回每个查询的响应。"}, "required": ["querylist"]}}}</tools>

每次函数调用都要返回一个包含函数名称和参数的 JSON对象，并将其置于 <toolcall></ 工具调用> XML 标签中：<工具调用>{"名称": <函数–名称>, "参数": <参数–JSON–对象>}</工具调用>

提示

请根据系统指令回答以下问题。问题：

F.2. 强化学习奖励判定提示

完整的提示用于强化学习训练阶段、评估答案正确性并计算奖励的提示 .

系统消息

你是一名 AI 助手，负责根据问题和标准答案评估模型回答是否正确。你的判断应遵循以下原则：1. 在评估模型的回答之前，应全面考虑问题和标准答案。

se.

2. 你的判断必须严格为“是”或“否”，依据是模型的回答在事实上是否准确，以及是否与标准答案一致。3. 如果模型回答是标准答案更具体的形式，则视为正确。4. 如果模型回答包含所有关键信息，但增加了少量细节，只要额外细节在事实上正确，就视为正确。5. 如果模型回答与答案矛盾、修改答案或遗漏关键部分，则视为错误。6. 对于数值，即使以不同单位表示，也要确保其正确。7. 对于姓名，检查名与姓是否正确。如果额外提供的中间名正确，也视为正确。8. 对于是非问句，回答必须与“是”或“否”完全一致才算正确。9. 如果模型回答包含拒绝性陈述，且没有直接回答问题，则必须判为错误。10. 如果存在多个候选答案，也可以根据上述规则逐一评估模型回答。如果回答与至少一个候选答案一致，则应视为正确。你的输出必须采用以下格式：“是”或“否”

提示

问题，以及模型回答评估问题： {问题}标准答案： {标准答案}候选答案： {候选答案}模型回答： {模型回答}评估说明 根据问题、标准答案和候选答案，评估模型回答是否正确。遵循预定义的判定规则，并给出明确的是/否答案，不要任何说明。

输出格式 是或否

F.3. 问答生成提示

用于为训练数据生成问答对的完整提示。

问答生成提示

你是一个面向搜索–密集型多模态推理的问题–答案生成智能体。

你的任务是基于查询图像和外部证据，生成一个高–质量的问题–答案对。

重要约束：
– 问题必须直接针对查询图像中所展示的内容。– 问题不得仅凭图像本身即可回答。
– 答案必须仅根据下方提供的外部证据得出，不得来自查询图像本身。

以下是提供给你的信息：

[查询图像]
– 提供了一幅图像作为查询。– 你必须仅使用视觉线索推断图像所描绘的内容。
– 不得假设或编造关于查询图像的任何文字描述。

[外部文本证据]– 通过搜索检索到的网页的多个摘要：{文本
摘要}

[外部视觉证据]– 从网络检索到的其他视觉相关图像的图注：{其他图像图注
}

[问题说明]
1. 问题必须明确提及查询图像中的内容（例如图像中可见的物体、结构、地点、场景或事件）。 2. 问题必须要求提供无法仅凭图像获得的事实性知识、识别结果或语境信息。 3. 问题应自然地源于观察图像后进一步寻求信息。 4. 避免提出是非问句，也避免使用含糊或主观的措辞。 5. 不要明确提及图像图注、摘要、搜索结果或外部证据。

[答案说明]
6. 答案必须真实、简洁，并直接回答问题。 7. 答案必须有提供的外部文本或视觉证据支持。 8. 答案不得依赖假设或仅在查询图像中可见的信息。 9. 优先选择需要综合多条证据才能得出的答案。 10. 答案不得写成完整的语法句子。

[输出格式]
仅准确输出以下JSON对象，不得包含其他内容：

{{"question": "<一个清晰、措辞完整的问题>", "answer": "
<简洁、基于事实的答案>"}}