

# VimRAG：通过多模态记忆图在检索增强生成中导航大规模视觉上下文

Qiuchen Wang, Shihang Wang, Yu Zeng, Qiang Zhang, Fanrui Zhang, Zhuoning Guo,  
Bosi Zhang, Wenxuan Huang, Lin Chen, Zehui Chen, Pengjun Xie, Ruixue Ding<sup>†</sup>

Tongyi Lab, Alibaba Group

## Abstract

有效检索、推理和理解多模态信息仍是智能体系统面临的关键挑战。传统的检索增强生成（RAG）方法依赖于线性的交互历史，难以处理长上下文任务，尤其是在迭代推理场景中涉及信息稀疏但 token 密集的视觉数据时表现不佳。

为弥合这一差距，我们提出 VimRAG，一种专为跨文本、图像和视频的多模态检索增强推理设计的框架。受我们系统性研究的启发，我们将推理过程建模为动态有向非循环图，用以组织智能体状态与检索到的多模态证据。在此结构化记忆基础上，我们引入了图调制视觉记忆编码机制，通过节点的拓扑位置评估记忆结点的重要性，使模型能够动态地将高分辨率 token 分配给关键证据，同时对冗余线索进行压缩或丢弃。为实现该范式，我们提出了图引导的策略最优化方法。该方法通过剪枝与冗余动作相关的记忆结点，将步骤级有效性与其轨迹级奖励解耦，从而实现细粒度的信用分配。

大量实验表明，VimRAG 在多种多模态 RAG 基准上均持续达到领先性能。代码已开源，地址见<https://github.com/Alibaba-NLP/VRAG>。

## 1 引言

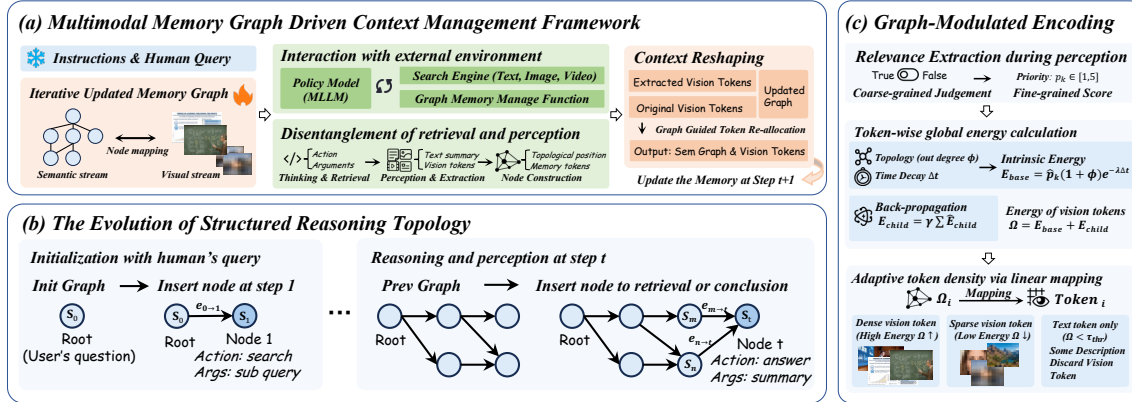


图 1: VimRAG 框架的推理流水线。(a) 由推理、检索和记忆演化组成的循环推理环。(b) 详细描述了结构化推理拓扑的演化，其中每个结点存储特定智能体的记忆，包括动作、动态压缩的多模态观测及其对应的时间与拓扑结构。(c) 展示了图调制视觉记忆编码的逐步过程。该机制通过整合时间、拓扑和语义相关性来模拟人类遗忘，调节视觉 token 密度，有效过滤噪声以保留真正有价值的线索。

近年来，多模态大模型 Bai et al. (2025); Team et al. (2025); Singh et al. (2025); Team et al. (2023) 的进展从根本上拓展了多模态智能体增强型检索生成 (RAG) Cho et al. (2024); Arslan et al. (2024); Wang et al. (2025a); Yu et al. (2024) 的能力。借助由统一嵌入模型 Guo et al. (2025); Li et al. (2026); Meng et al. (2025); Sun et al. (2025b); Faysse et al. (2024) 驱动的搜索引擎，基于多模态大模型的智能体能够对包含交错文本与图像的大规模语料库进行检索和推理。 Su et al. (2025); Wang et al. (2025b); Yu et al. (2025b); Jeong et al. (2025); Yeo et al. (2025); Geng et al. (2025) 然而，与文本不同，视觉数据在 token 层面较为密集，且相对于特定查询往往表现出语义稀疏性 Ma et al. (2024);

Tanaka et al. (2023); Wang et al. (2025c)。由于记忆与上下文管理策略已被证实是优化长上下文任务的有效方法 Zhou et al. (2025); Yu et al. (2025a); Wu et al. (2025); Xu et al. (2025); Chhikara et al. (2025); Chen et al. (2024a); Ye et al. (2025)，将这一范式迁移至高效管理海量视觉上下文的同时不丢失关键信息，是一个极具前景的方向。

受这些进展的启发，我们提出 VimRAG，一种专为通过多模态 Agentic 记忆范式实现迭代检索增强推理而设计的新框架。我们的方法受到关于将上下文管理与基于记忆的方法适配至多模态场景所面临的三个关键观察的驱动：

(i) **动作历史与上下文先验之间的错位**。智能体的实际执行历史与其呈现给模型的重塑提示之间存在根本性不匹配。这种结构上的盲区掩盖了关键的状态参数；具体而言，在 RAG 任务中，这会导致重复查询以及与搜索引擎的无用交互。

(ii) **文本记忆与视觉观察之间的不一致**。虽然将视觉信息压缩为文本记忆显著提升了 token 效率，但固有的细粒度细节丢失造成了语义鸿沟，常使记忆在验证过程中失效。

(iii) **监督不足且效率低下**。当前的拒绝采样策略仅将最终结果奖励广播至每一步，并对长轨迹中的所有响应 token 计算梯度。这在多步 Agentic RAG 任务中造成了误导性监督，其中有效的检索因错误的最终答案而受到惩罚，而低效的查询则仅因正确结果而获得奖励。

为了进一步验证这些见解，我们开展了一项试点研究，重点关注三个关键方面：交互历史的拓扑结构组织、多模态记忆中分辨率与效率之间的权衡，以及基于多步记忆的监督的有效性。这引出了我们的第一个研究问题：1) **如何构建推理过程，以防止在上下文压缩过程中丢失关键信息？**此外，保留视觉细节带来的高昂 token 成本引发了一个新的挑战：2) **智能体在严格的 token 约束下，如何解决视觉记忆分辨率困境？**基于上述两点洞察，我们提出第三个研究问题：3) **如何将中间交互与最终奖励解耦，以实现细粒度的监督？**

针对这些挑战，VimRAG 通过三项对应创新从根本上重构了智能体推理范式：

(i) 为解决结构瓶颈问题，我们提出了 **多模态记忆图**，将推理过程建模为动态的有向非循环图，其中每个节点编码智能体的动作和多模态观测。如图 1 (b) 所示，该拓扑结构不仅捕捉了与智能体相关的细节，还保留了时间与逻辑依赖关系（形式化为状态  $\mathcal{S}$  和有向边  $\mathcal{E}$ ）。作为推理的先验，该结构塑造了对原始问题分解的上下文，使智能体能够区分死胡同分支与新的探究路径，避免了简单记忆追加带来的潜在冗余以及迭代重摘要导致的效率低下。

(ii) 为构建面向下一状态预测的上下文而非单纯存储事实，我们实现了一种称为 **图调制视觉记忆编码**的机制，直接基于图拓扑构建。如图 1 (c) 所示，通过拓扑中心性与递归反馈评估结点重要性，该模块自适应地分配视觉 token 密度。它保留关键证据的高分辨率 token，同时压缩或丢弃次要细节，使推理在紧凑的 token 预算内与有价值的观测保持一致。

(iii) 观察到图拓扑天然适合分步评估，我们提出一种 **图引导策略最优化策略**以实现细粒度监督。如图 4 所示，不同于将稀疏的结果奖励广播至整个轨迹中的样本，我们利用记忆图通过识别从原点到答案结点的 **关键路径**来执行结点剪枝。通过将检索过程与最终结果解耦，在演员更新过程中掩码掉假阳性（尽管答案正确但无关的结点）和假阴性（在错误答案中具有价值的检索项）。该机制通过仅聚焦于有效且有价值的样本进行梯度更新，提升了训练效率与有效性。

我们的主要贡献如下：

- 我们系统地研究了多模态智能体记忆范式在 RAG 任务中的应用，识别出上下文错位和监督稀疏性中的关键瓶颈。
- 我们提出 VimRAG，这是一种新颖的框架，通过将多模态记忆图与图调制的视觉记忆编码相结合，以构建推理拓扑结构。
- 我们提出一种图引导的策略最优化方法，以分离检索有效性与稀疏奖励，从而在训练过程中实现细粒度的信用分配。
- 大量实验表明，VimRAG 始终能带来显著提升，在多模态 RAG 基准上达到了顶尖水平。

## 2 试点研究

在本节中，我们提出了问题并探讨了多模态 RAG 中记忆范式所面临的挑战，从而为第 3 小节中提出的 VimRAG 设计提供了动机。

### 2.1 初步的

**任务定义。**给定一个用户查询  $q$  和一个包含文本文档、视觉丰富的图像以及视频流的大型语料库  $\mathcal{C}$ ，我们的目标是高效地检索、准确地感知并推理复杂的跨模态信息，以生成对查询  $q$  的回答  $a$ 。

**历史累积范式。**标准智能体通常在思维  $\tau$ ，动作  $a$ ，观察  $o$  环中运作：

$$\mathcal{H}_t = [q, \tau_1, a_1, o_1, \dots, \tau_{t-1}, a_{t-1}, o_{t-1}] \quad (1)$$

策略  $\pi_\theta(\cdot|\mathcal{H}_t)$  会根据整个序列生成下一个动作。这会导致关键信息  $\mathcal{O}_{\text{crit}}$  的显著分散，尤其是在稀疏的多模态提示情况下。信息密度  $|\mathcal{O}_{\text{crit}}|/|\mathcal{H}_t| \ll \epsilon$  随着  $t$  的增加而增加。

**基于记忆的智能体范式。**相比之下，基于记忆的方法从被动的历史累积转向主动的上下文管理。模型根据最近的观测  $o_t$  更新记忆状态  $m_t$ ：

$$m_{t-1} \xrightarrow{\pi_\theta(\cdot)} (\tau_t, a_t) \xrightarrow{Env} o_t \xrightarrow{\pi_\theta(\cdot|\tau_t, a_t, m_{t-1})} m_t \quad (2)$$

该机制保持了高注意力集中度，因为信息密度保持稳定（即  $|\mathcal{O}_{\text{crit}}|/|m_t| \approx C$ ）。然而，仅依赖压缩状态  $m_t$  会引入马尔可夫盲区，导致潜在的信息丢失和推理断裂，这给设计稳健的记忆范式带来了挑战。

### 2.2 记忆结构对 Agentic 推理的影响

本小节研究了记忆结构对多模态 RAG 智能体基本能力的影响。

**实验情景。**我们基于当前上下文管理方法，对比三种 Agentic 记忆范式：1) 传统 ReAct Yao et al. (2022)，仅将  $(\tau, a, o)$  连结以形成完整上下文；2) 迭代摘要作为记忆 Zhou et al. (2025)，通过迭代将观察结果压缩至先前的记忆状态；以及 3) 结构化图作为记忆，维护智能体推理状态的结构化拓扑。我们通过提示 Qwen3VL-30B-A3B-Instruct 实现这些范式。详见附录 C.1 中的详细工作流程。

**观察。**图 2 (a) 表明，基于记忆的范式（摘要和图）相比 ReAct 显著降低了 token 消耗。关于动作鲁棒性（图 2 (b)），随着上下文扩展，ReAct 的性能下降最为剧烈。基于摘要的方法同样存在状态盲视问题：智能体无法追踪其历史检索-感知动作，导致在多跳场景中频繁出现重复查询。相比之下，通过系统性地跟踪智能体的状态，基于图的记忆方法有效减少了冗余搜索和循环错误。

**洞察。**记忆的真正价值在于塑造智能体的未来行为，而不仅仅是存储过去观测中的事实。我们认为，基于图形的结构提供了维持智能体推理状态所需的结构偏差，使其能够连接过去与未来之间的有效路径。

### 2.3 记忆中信息压缩的影响

本小节探讨了观测与记忆之间的语义对齐，以及压缩率与关键信息保留之间的权衡。

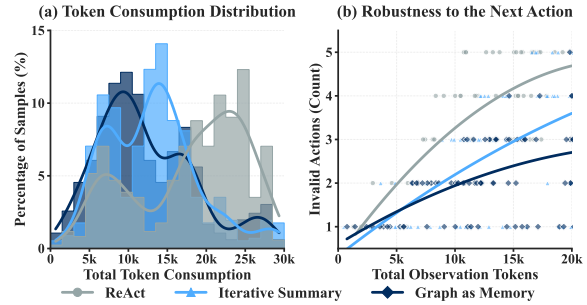


图 2: 记忆结构的定量分析。(a) 完整样本的总 token 消耗分布。(b) 无效检索动作的数量。通过建模智能体的当前状态而非仅仅存储事实，基于图形的范式相较于摘要法有效避免了重复检索。



**实验情景。**我们在记忆范式 Zhou et al. (2025) 上对比四种跨模态压缩策略：1) 预标题化：语料库的视觉部分预先生成标题。智能体仅执行纯文本检索与记忆。2) 视觉观察作为记忆：智能体直接将观测结果以原始多模态 token 形式存储。3) 上下文感知的标题化：智能体以原始多模态数据作为观测结果，但将其以文本摘要形式进行记忆。4) 语义相关视觉记忆：在检索视觉数据后，智能体选择性保留相关的视觉 token，而丢弃无关的 token。请参见附录 C.2 获取实现细节。

**观测。**如表 1 所示，纯文本策略（策略 1）虽然最小化了 token 消耗，但存在文本与视觉表示之间的差距。同时，简单地将所有原始观测值存储在上下文中（策略 2）表现不佳，这是由于信噪比下降所致，与我们在第 2.2 节中的见解一致。选择性视觉策略（策略 4）优于文本摘要（策略 3），这表明为智能体最终验证保留关键视觉特征的必要性。

**洞察。**在多模态任务中，将视觉 token 专门分配给记忆中的关键视觉细节对于验证至关重要，从而保留高价值证据，同时丢弃噪声，以实现最佳的 token 效率。

## 2.4 稀疏奖励信号对记忆范式中信用分配的影响

本小节研究多步智能体轨迹中结果奖励在信用分配中的可靠性。

**实验情景。**设一条轨迹为一系列步骤  $\tau = \{s_1, s_2, \dots, s_T\}$  Zhou et al. (2025)。我们将步骤分解为两个互不相交的子集：1) **证据检索** ( $\mathcal{S}_{evd}$ )，包含捕捉关键线索的步骤，以及 2) **噪声/冗余** ( $\mathcal{S}_{noise}$ )，包含无关的动作。为了评估特定步骤的贡献，我们进行反事实消融研究。我们通过掩码特定的步骤子集并重建剩余步骤为完整历史，构造反事实轨迹  $\hat{\tau}$ ，以重新评估。具体而言，对于正例（其中奖励  $r = 1$ ），我们评估  $\hat{\tau} = \tau \setminus \mathcal{S}_{evd}$  和  $\hat{\tau} = \tau \setminus \mathcal{S}_{noise}$ 。对于包含有效检索步骤 ( $s_t \in \mathcal{S}_{evd}$ ) 的负例 ( $r = 0$ )，我们测试通过降噪是否能够恢复性能，i.e.，仅保留证据集  $\hat{\tau} = \mathcal{S}_{evd}$ 。更多细节请参见附录 C.3。

**观察。**图 3 (a) 揭示了奖励  $r$  与逐步样本之间存在关键的不匹配。具有  $r = 1$  的样本并非纯粹高效；它们经常包含  $s_t \in \mathcal{S}_{noise}$ ，基于结果的监督会错误地为其分配正梯度。具有  $r = 0$  的样本不应被普遍惩罚，因为它们可能包含有效  $s_t \in \mathcal{S}_{evd}$ ，尽管最终失败。图 3 (b) 通过反事实消融实验展示了这一点。对于负例，仅移除冗余步骤即可恢复性能。这表明失败源于对噪声的推理，而非证据不足。对于正例，移除证据步骤 ( $\tau \setminus \mathcal{S}_{evd}$ ) 后仍保持非零性能，证实模型部分依赖于参数化内部知识。

**洞察。**记忆范式自然地智能体推理过程分解为离散状态，使我们能够解耦检索质量，以解决粗粒度结果奖励的不足。这使得我们能够校准信用分配，驱动模型在细粒度步骤级别学习建设性动作的分布。

## 3 VimRAG

在本节中，基于试点研究中获得的洞察与基础思想，我们全面描述了我们的 VimRAG 框架。我们首先详细阐述结构化推理拓扑 (§ 3.1)，作为我们智能体记忆的结构骨干。接着，我们引入图调制视觉记忆编码 (§ 3.2)，以在记忆中实现动态分辨率缩放。最后，我们提出图引导的拒绝采样 (§ 3.3)，通过拓扑信用分配实现细粒度最优化。

### 3.1 结构化推理拓扑

我们将多模态推理过程表述为有向非循环图 (DAG) 的序列演化，记作  $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ ，其中  $t$  表示离散的推理步骤。针对我们 1<sup>st</sup> insight 中指出的线性历史结构局限性，该拓扑结构显式地捕捉了智能体动作之间的逻辑依赖关系。

表 1: 跨模态记忆策略的比较。语义相关视觉记忆在压缩比与关键信息保留之间实现了更优的权衡。

Memory Strategy	Modality Retrieval → Memory	Average Tokens	Performance (%)	
			Image	Video
(1) Pre-Caption	Text → Text	0.9k	14.5%	17.2%
(2) Raw Visual Tokens	Vision → Vision	15.8k	45.6%	30.4%
(3) Context-Aware Caption	Vision → Text	1.5k	52.8%	39.5%
(4) Semantically-Related	Vision → Selective Vision	2.7k	58.2%	43.7%

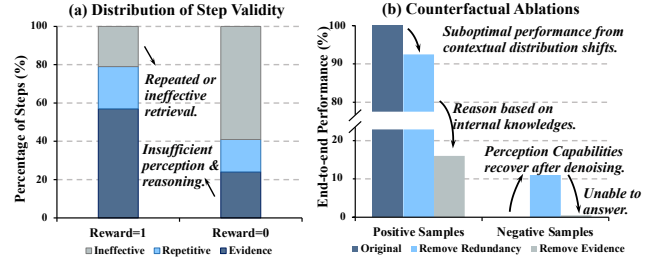


图 3: 结果奖励与步骤有效性之间错位的实证分析。(a) 二元结果奖励下步骤类别的分布。(b) 移除冗余或证据步骤的影响，展示了奖励的粗糙性。

**图结点作为认识状态** 我们将每个结点  $v_i \in \mathcal{G}_t$  定义为智能体认知状态的离散单元。

$$v_i \triangleq (p_i, q_i, s_i, m_i), \quad (3)$$

其中  $v_i$  定义为一个元组,  $p_i$  表示父结点索引集合, 编码局部依赖结构;  $q_i$  代表与搜索动作相关的分解子查询;  $s_i$  作为简洁的文本摘要;  $m_i$  构成多模态情景记忆库 (例如从检索到的文档或帧中获取的视觉 token)。边集  $\mathcal{E}_t = \{(v_j, v_i) \mid j < i\}$  自然地结构化编码了推理流程。因此, 完全图状态是一个有序序列  $\mathcal{G}_t = [v_{root}, \dots, v_t]$ 。

**迭代图演化。**我们将图构建表述为一个部分可观测马尔可夫决策过程 (POMDP)。在每一步  $t$ , 策略  $\pi_\theta$  采样一个动作  $a_t \in \{a^{ret}, a^{mem}, a^{ans}\}$ , 驱动状态转移:

$$a_t \sim \pi_\theta(\cdot \mid \mathcal{G}_{t-1}), \quad \mathcal{G}_t \leftarrow \Psi(\mathcal{G}_{t-1}, a_t), \quad (4)$$

其中  $\Psi$  表示外部环境的算子。

---

**Algorithm 1: Inference Pipeline of VimRAG**

---

Require: Human Query  $Q$ , Policy  $\pi_\theta$ , External Environment  $\mathcal{V}$ .

```

1: Initialize:  $\mathcal{G}_0 \leftarrow \{v_{root} : Q\}$ ,  $t \leftarrow 0$ .
2: while  $t < T_{max}$  do
3:   // 1. Context shaping & Action Generation
4:   Context  $\mathcal{H}_t \leftarrow \mathcal{V}.LinearizeGraph(\mathcal{G}_t)$ , generate next action  $a_t \sim \pi_\theta(\cdot \mid \mathcal{H}_t)$ 
5:   // 2. Topological Expansion (Section 3.1)
6:   if  $a_t = a^{ret}$  then
7:     Initial node  $v'_t$  with query  $q_t$  and parent  $p_t$ , then search multimodal information:
8:      $\mathcal{O}_t \leftarrow \mathcal{V}.Search(q_t)$ 
9:     Multimodal perception & Memory population:
10:     $a^{mem} \sim \pi_\theta(\cdot \mid \mathcal{H}_t, a_t, \mathcal{O}_t)$ ,  $(s_t, m_t) \leftarrow \mathcal{V}.Execute(a^{mem})$ 
11:    Graph updating:  $\mathcal{G}_{t+1} \leftarrow \mathcal{G}_t \cup \{v_t \mid (p_t, q_t, s_t, m_t)\}$ 
12:   else if  $a_t = a^{ans}$  then
13:     Connect terminal node  $v_{ans}$  and return  $a_t.answer$ 
14:   end if
15:   // 3. Dynamic Visual Memory Shaping (Section 3.2)
16:   for each visual node  $v_i \in \mathcal{G}_{t+1}$  do
17:     Calculate Energy:  $\Omega(v_i) \leftarrow \mathcal{V}.Energy(\mathcal{G}_{t+1})$  (Eq. 7)
18:     Allocate Token Budget:  $b_i \leftarrow \mathcal{V}.Scale(\Omega(v_i))$  (Eq. 8)
19:     Compress Memory:  $m_i \leftarrow \mathcal{V}.VisualEncode(m_i, b_i)$ 
20:   end for
21:    $t \leftarrow t + 1$ 
22: end while
```

---

如算法 1 所示, 模型在定义的动作空间内与外部环境进行多轮交互。演化周期分为三个阶段:

- **探索性扩展 ( $a^{ret}$ )**: 当当前证据不足时, 智能体生成一个骨架结点  $v'_t = (p_t, q_t, \emptyset, \emptyset)$ 。对外部语料库执行查询  $q_t$ , 以获取原始的多模态观测结果  $\mathcal{O}_t$ 。
- **多模态感知与记忆填充 ( $a^{mem}$ )**: 获取  $\mathcal{O}_t$  后, 策略调用感知动作, 将高熵信息提炼为结构化记忆:  $\mathcal{O}_t \rightarrow (s_t, m_t)$ 。为实现稳健的噪声抑制, 模型采用由粗到细的过滤策略: 对于每个检索项, 模型生成一个二值显著性掩码  $u \in \{0, 1\}$  和细粒度语义得分  $p \in [1, 5]$ 。对于视频观测  $\mathcal{O}_t^{video}$ , 该机制利用基础模型 (如 Qwen3-VL) 的时间定位能力, 提取与时间戳对齐的关键帧。该操作将原始数据转换为摘要  $s_t$  和视觉 token  $m_t$ , 最终完成结点  $v_t = (p_t, q_t, s_t, m_t)$ 。
- **终端投影 ( $a^{ans}$ )**: 一旦策略确定  $\mathcal{G}_t$  中的推理路径已足够, 便会执行答案动作。从  $v_{root}$  到  $v_{ans}$  的路径构成了任务完成的关键逻辑与语义路径。

**时序定位视觉压缩。**我们利用模型的时序定位能力来提取感兴趣的帧, 将稀疏的原始观测转换为稠密的、语义丰富的表示。输入表示为一个帧与时间戳的序列:

$$\mathcal{O}_t^{video} = [(ts_k, f_k)]_{k=1}^n \quad (5)$$

其中  $ts_k$  表示对应于第  $k$  帧  $f_k$  的时间戳 (格式为 `<%0.1f 秒>`)。通过执行内存动作  $a^{mem}$ , 原始流被提炼为已填充结点  $(s_t, m_t)$  的内容。

### 3.2 图调制的视觉记忆编码

受 2<sup>nd</sup> insight 的启发，我们提出了图调制记忆编码，以解决视觉记忆保真度与 token 预算之间的冲突。我们不采用静态的视觉项分辨率，而是将视觉 token 的分配表述为一个受限资源分配问题，自适应地将高密度 token 分配给关键证据。

**内存能量公式化。** 令  $v_i$  表示推理图  $\mathcal{G}$  中的第  $i$  个结点，且  $\mathcal{M}_i = \{m_{i,k}\}_{k=1}^K$  是该结点内  $K$  个检索到的视觉项的集合。如图 1(c) 所示，能量计算将内在先验与递归强化相结合。

1) 内在能量。首先，我们计算每个项目  $m_{i,k}$  的基准能量，记为内在能量  $\mathcal{E}_{\text{int}}$ ：

$$\mathcal{E}_{\text{int}}(m_{i,k}) = \underbrace{\hat{p}_{i,k} \cdot (1 + \deg_{\mathcal{G}}^+(v_i))}_{\text{Structural-Semantic Relevance}} \cdot \underbrace{\exp(-\lambda(T - t_i))}_{\text{Temporal Decay}}, \quad (6)$$

其中，规范化的  $\hat{p}_{i,k} \in [0, 1]$  表示细粒度语义优先级， $\deg_{\mathcal{G}}^+(v_i)$  表示结点  $v_i$  的出度。为了模拟人类遗忘机制，基于流逝时间  $T - t_i$  应用时间衰减，从而防止过时信息的累积。

2) 递归强化。单纯依赖内在能量是不够的，因为早期证据尽管初始重要性较低，却常常是连接下游洞察的关键桥梁。为解决这一贡献度分配问题，我们通过后续结点的反馈来强化内在能量，计算最终能量  $\Omega(m_{i,k})$ ：

$$\Omega(m_{i,k}) = \mathcal{E}_{\text{int}}(m_{i,k}) + \gamma \sum_{v_j \in \text{Child}(v_i)} \bar{\Omega}(v_j), \quad (7)$$

其中  $\gamma$  控制反馈强度， $\bar{\Omega}(v_j)$  对  $\text{Child}(v_i)$  的平均能量进行聚合。此公式确保了支持高价值未来推理的基础结点能够抵御时间衰减。

**全球选择与分辨率分配。** 如算法 1 所示，我们在记忆塑造过程中根据每个视觉项  $m_{i,k}$  的能量动态分配 token：

$$b_{i,k} = \left\lfloor S_{\text{total}} \cdot \frac{\Omega(m_{i,k})}{\sum_{m' \in \mathcal{M}_{\text{top}}} \Omega(m')} \right\rfloor. \quad (8)$$

其中  $S_{\text{total}}$  表示维持模型性能最优的总 token 预算， $\mathcal{M}_{\text{top}}$  表示根据能量秩保留的前  $K$  个项的集合。该公式确保 ViT 编码器能够捕捉高能量证据的细粒度细节，将计算预算集中于最具信息量的视觉区域。

### 3.3 图引导的策略最优化

受 3<sup>rd</sup> insight 的启发，我们提出了图引导的策略最优化 (Graph-Guided Policy Optimization, GGPO)，其中我们利用图结构来解耦推理路径，并实现细粒度的信用分配，如图 4 所示。

**轨迹分割** 我们首先形式化步骤  $t$  的初始提示  $\mathcal{C}_t$ 。它包含系统指令  $inst$ 、用户的查询  $q$  以及线性化的记忆图  $\mathcal{L}(\mathcal{G}_t)$ ：

$$\mathcal{C}_t = \{inst, q, \mathcal{L}(\mathcal{G}_t)\} \quad (9)$$

在交互滚动过程中，智能体收集一个分解为结点构建单元的结构化轨迹。每个单元对应于图结点的构建  $v_t$ ：

$$\mathcal{H}^{(t)} = (\mathcal{C}_t, \tau_t, a_t^{\text{ret}}, o_t, \tau'_t, a_t^{\text{mem}}) \rightarrow v_t \quad (10)$$

其中  $\tau_t$  代表导致检索动作的推理过程， $\tau'_t$  表示用于合成记忆动作的反思。每个回合以终止推理块  $\mathcal{H}^{(T)} = (\mathcal{C}_T, \tau_{\text{ans}}, a^{\text{ans}})$  结束。

**通过图剪枝进行信用分配。** 轨迹级别的奖励与单个步骤有效性之间的不匹配是一个关键挑战。如图 4 所示，我们通过利用语义图拓扑来评估每个步骤并执行剪枝来解决这一问题：

- 1) 剪枝假阳性 (死胡同状态)。给定一个正例  $(\mathcal{T}, r = 1)$ ，我们通过从答案结点反向遍历来识别关键路径  $\mathcal{P}_{\text{ans}} \subseteq \mathcal{G}$ 。结点  $v \notin \mathcal{P}_{\text{ans}}$  代表死胡同，即冗余的探索或与解无关的逻辑关系。
- 2) 剪枝假阴性 (有价值检索)。给定一条负向轨迹  $(\mathcal{T}, r = 0)$ ，利用每个查询的参考标注，我们识别出检索结果中包含相关信息的步骤。我们将这些有价值检索的动作从负向策略梯度更新中排除，以避免对有效行为进行惩罚。

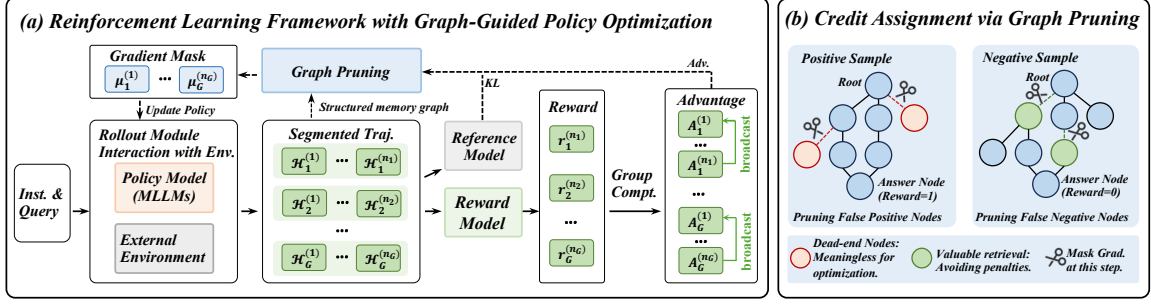


图 4: 图引导策略最优化概述。(a) Agentic 记忆训练框架将回溯轨迹分割为记忆范式内的原子推理循环, 其中基于结果的优势信息被广播, 以实现步骤级的信用分配。(b) 通过图剪枝进行信用分配利用结构化图实现精确的信用分配, 应用梯度掩码以避免强化正例中的低效死胡同, 并防止惩罚负例中的有价值检索。

表 2: 主要结果。最佳性能以粗体标出。基准测试根据参考内容的模态分为三类: 通用文本、图像和视觉文档, 以及大规模长上下文视频语料库。评估在统一的大规模多模态数据集上进行, 引入了更大的挑战, 并更贴近真实应用场景。

Method	General Text		Image & Visual Document			Large-Scale Long-Context Video Corpus				Overall
	HotpotQA	SQuAD	WebQA	SlideVQA	MMLongBench	LVBench	WikiHowQA	SyntheticQA	XVBench	
Qwen3-VL-4B-Instruct										
Vanilla RAG	63.6	63.0	45.3	44.8	15.2	12.0	11.0	32.6	27.2	35.0
ReAct	62.9	63.8	39.3	44.9	13.9	12.2	14.1	29.9	21.3	33.6
UniversalRAG	50.9	64.7	41.2	14.7	5.4	15.5	3.3	23.9	7.5	25.2
VideoRAG	57.2	63.2	41.8	34.0	16.2	19.4	20.1	45.8	29.8	36.4
MemAgent	67.3	70.4	46.5	43.1	12.4	18.9	19.2	34.3	24.4	37.4
Mem1	70.8	67.5	44.1	49.8	27.1	16.8	14.3	42.9	31.9	40.6
VimRAG (Ours)	75.1	73.7	47.6	52.8	28.1	22.8	21.8	51.0	34.2	45.2
Qwen3-VL-8B-Instruct										
Vanilla RAG	64.0	64.2	48.1	48.5	16.2	14.8	15.7	37.0	29.7	37.6
ReAct	70.8	65.5	40.0	50.0	15.4	15.9	23.0	35.0	24.0	37.7
UniversalRAG	55.9	65.3	45.8	17.2	6.6	19.1	12.0	25.0	9.6	28.5
VideoRAG	62.0	62.2	42.1	35.5	18.2	23.8	25.7	49.5	30.7	38.9
MemAgent	71.1	74.8	47.1	45.3	14.7	22.2	23.1	37.5	26.9	40.3
Mem1	73.0	68.4	44.5	55.7	32.6	22.4	19.9	43.4	32.2	43.6
VimRAG (Ours)	79.1	76.4	53.9	62.4	33.4	24.5	29.7	54.5	37.1	50.1

**强化学习的实现** 为了实现结构化的信用分配, 我们为每个分段轨迹执行一个二值剪枝掩码  $\mu = [\mu_1^{(1)}, \dots, \mu_G^{(n_G)}]$ , 其中  $\mu = 1$  表示该步骤应从更新中排除。令  $\mathcal{P}_{ans}$  表示正确解中的关键路径结点,  $\mathcal{R}_{val}$  表示错误解中产生有价值检索的结点。我们将  $\mu_t$  定义为:

$$\mu_t = \underbrace{\mathbb{I}(r=1) \cdot \mathbb{I}(v_t \notin \mathcal{P}_{ans})}_{\text{Dead-Ends in Positive}} + \underbrace{\mathbb{I}(r=0) \cdot \mathbb{I}(v_t \in \mathcal{R}_{val})}_{\text{Valuable Retrieval in Negative}} \quad (11)$$

其中  $\mathbb{I}(\cdot)$  为指示函数。第一项掩码正回合中的冗余步骤, 第二项掩码负回合中的有效检索动作, 以避免对其进行惩罚。最优化目标表述为:

$$\max_{\pi_\theta} \mathbb{E}_{q \sim \mathcal{D}, \{\mathcal{H}_g^{(i)}\}_{g=1, i=1}^{G, n_g} \sim \pi_\theta} \left[ \frac{1}{\sum_{g=1}^G n_g} \sum_{g=1}^G \sum_{i=1}^{n_g} (1 - \mu_{g,i}) \cdot \min \left( r_{g,i}(\theta) \hat{A}_{g,i}, \text{clip}(r_{g,i}(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_{g,i} \right) \right] \quad (12)$$

其中,  $\{\mathcal{H}_g^{(i)}\}_{i=1}^{n_g}$  表示与第  $g$  次采样轨迹对应的  $n_g$  序列,  $r_{g,i}(\theta)$  表示轨迹  $g$  中第  $i$  段的概率比。

## 4 实验

### 4.1 实验情景

**基准。** 我们对比了当前先进的 RAG 与上下文管理方法: (1) **原始 RAG** 使用原始问题作为查询输入搜索引擎, 随后大模型直接进行推理。(2) **ReAct** Yao et al. (2022): 模型采用“思考-行动”范式进行推理与检索。(3) **VideoRAG** Jeong et al. (2025) 通过帧选择提取信息以支持推理。(4) **UniversalRAG** Yeo et al. (2025) 将跨模态语料库中的 RAG 视为一个路由问题。(5) **MemAgent** Yu et al. (2025a): 我们



通过依次输入检索结果来实现该方法。(6) Mem1 Zhou et al. (2025) 通过循环的检索-记忆过程更新其记忆。

**基准与指标。** 我们在涵盖多种任务的综合性基准上评估了我们的方法：通用文本任务 HotpotQA Yang et al. (2018) 和 SQuAD Rajpurkar et al. (2016)；基于图像的任务 WebQA Chang et al. (2022)；视觉丰富的文档基准 SlideVQA Tanaka et al. (2023) 和 MMLongBench Ma et al. (2024)；长视频基准 LVBench Wang et al. (2025c)；以及视频语料库理解基准 WikiHowQA 和 SyntheticQA Jeong et al. (2025)。此外，我们构建了 XVBench 以解决跨视频理解基准缺失的问题。我们采用基于二值模型的评估指标（0 或 1）来衡量智能体在这些任务中答案的正确性。有关数据集和环境设置的更多细节，请参见附录 E。

## 4.2 结果

**主要结果。** 如表2所示，传统的范式如 ReAct 在处理包含大量 token 的视觉数据时，容易出现上下文耗尽的问题。与此同时，针对特定任务设计的多模态 RAG 方法（如 VideoRAG 和 UniversalRAG）由于其固定的推理流水线，通常表现出有限的泛化能力。此外，与我们在初步研究中观察到的状态盲视现象（见第2节）一致，当前基于摘要的记忆范式无法追踪历史检索动作，导致冗余的推理环路。通过解决这些结构性局限，VimRAG 能够有效管理大规模多模态上下文，并在近期基准模型（如 MemAgent 和 Mem1）上取得显著优势。具体而言，VimRAG 在 Qwen3-VL-8B-Instruct (43.6 → 50.1) 和 Qwen3-VL-4B-Instruct (40.6 → 45.2) 上均实现了显著提升。这证实了显式建模推理拓扑结构的重要性，相较于被动积累历史信息，前者对于充分释放大模型在多模态密集任务中的潜力至关重要。

**方法消融实验。** 如表 3 所示，我们对 VimRAG 的关键组件进行了分解，以检验其影响。在引入模块后逐步提升的表现验证了我们范式的优势。从图拓扑中获得的显著提升表明，显式建模逻辑依赖关系能够缓解状态盲区，这 **证明了在长时程推理中结构化状态追踪的必要性。**

此外，采用基于能量分配的视觉记忆通过为关键节点优先分配高分辨率 token，实现了更高的准确率，这 **证明了我们的图调制视觉记忆编码在细节与效率之间权衡优化方面的有效性。**最后，与图 5 中展示的稳定性一致，我们的图引导拒绝采样将检索有效性与结果奖励解耦，这 **证明了细粒度拓扑信用分配对于鲁棒训练的重要性。**总体而言，“图即记忆”不仅结构化智能体的推理轨迹，还通过结构解耦促进了更优的模型最优化。

表 3: 消融研究

Memory Structure			Memory Shaping			Acc.
Iter.	Graph	Multi.	Std.	Graph	Energy	
✓			✓			43.6
	✓			✓		47.1
	✓	✓		✓		48.9
	✓	✓		✓	✓	50.1

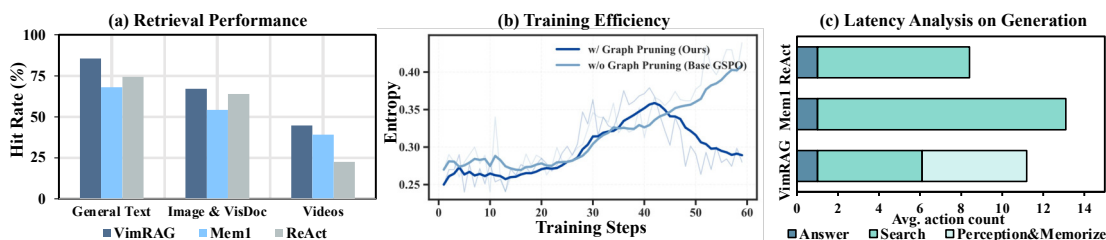


图 6: 鲁棒性与效率分析。(a) 不同模态下的检索命中率。(b) 训练熵曲线，展示了图剪枝带来的更快收敛。(c) 推理步骤的分解，突出了 VimRAG 减少的冗余。

## 4.3 分析

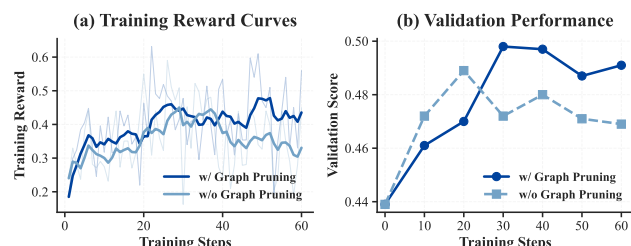


图 5: GGPO 的消融实验。我们的方法在不进行剪枝的情况下，比基准方法 GSPO 更具鲁棒性。

**鲁棒性检索是高质量生成的基础。** 高质量的生成严重依赖于检索到的上下文的准确率。如图 6 (a) 所示，不同记忆范式在检索鲁棒性方面存在显著性能差距。传统的线性或基于摘要的方法常常面临状态盲区问题，即随着上下文扩展，智能体无法追踪历史执行路径，导致重复查询和与搜索引擎的冗余交互。



**记忆模态对齐的必要性。** 第2.3节的结果表明,记忆模态与语料库类型具有更好的匹配性。我们的研究证实,保留相关视觉 token 显著优于文本压缩。VimRAG 通过自

适应地保留关键证据的高分辨率 token,同时丢弃噪声,解决了这一问题。这种对视觉信息的高效处理,促成了表2中展示的优异性能。

**结构解纠缠加速策略最优化。** 图6 (b) 表明,与基准方法相比,我们的策略实现了更快的收敛。这一观察为 Agentic RL 带来了两条关键洞见:(1) 最优化的稳定性取决于确保正梯度的正确性,同时消除来自负例的模糊更新;(2) rollout 样本的质量至关重要,使用具有明确偏好对齐的样本,相较于单纯扩大训练集,在性能和训练效率方面更具决定性作用。

**生成延迟** Figure 6 (c) 表明, VimRAG 尽管引入了感知步骤,仍显著降低了相比 ReAct 和 Mem1 的整体轨迹长度。线性方法通常会表现出由重复重读和无效搜索引起的 token 使用的长尾现象。相比之下, VimRAG 中的结构化记忆避免了冗余的环,并收敛到总动作更少的解决方案。

**案例研究** 附录中的案例研究G 强调了 VimRAG 如何成功识别一个死胡同结点  $v_1$  并回溯到一个新的查询结点  $v_2$ 。这种定性分析进一步验证了我们的图拓扑结构赋予智能体类人的自我修正能力。

## 5 相关工作

**多模态检索增强生成** 当前的检索增强生成方法在解决知识密集型问题方面展现出显著优势 Riedler & Langer (2024); Fang et al. (2025); Wang et al. (2025b); Chen et al. (2024a); Arslan et al. (2024); Geng et al. (2025); Bonomo & Bianco (2025); Han et al. (2025); Asai et al. (2024); Huang et al. (2026); Chen et al. (2024b)。随着多模态嵌入的发展,构建统一的多模态 RAG 智能体已成为主流趋势 Guo et al. (2025); Yu et al. (2024); Li et al. (2026); Faysse et al. (2024); Meng et al. (2025)。目前,越来越多的研究将 RAG 应用于长视频理解、文档理解等挑战性任务,已超越简单的文本问答范畴 Fan et al. (2024); Jeong et al. (2025); Wang et al. (2025a); Zeng et al. (2025b;a); Shi et al. (2024); Wang et al. (2025d); Li et al. (2024a;b)。我们的工作基于这些进展,实现了一个交织文本、图像和视频的 RAG 流系统,在统一框架中融合了检索、感知与理解。

**智能体的上下文管理与记忆** 基于大语言模型的智能体中最广泛使用的情境管理方法是 ReAct 的追加全部历史策略 Yao et al. (2022)。随着对长上下文推理需求的增长,研究人员正越来越多地关注情境管理的优化 Xu et al. (2025); Zhou et al. (2025); Yu et al. (2025a); Chhikara et al. (2025); Wu et al. (2025); Ye et al. (2025); Chen et al. (2024a); Sun et al. (2025a); Zhang et al. (2025)。然而,视觉数据通常具有高 token 密度且稀疏,相较于文本需要更高效的处理方法。我们的方法引入了结构化记忆图来处理这些视觉特征,在有效解决冗余问题的同时保留了关键细节。

## 6 结论与未来工作

我们提出了 VimRAG,该方法利用动态记忆图来处理海量视觉上下文。这一方法能够对关键信息进行细粒度感知,从而在复杂的多模态任务中取得更优的表现。在未来的工作中,我们将尽力训练一个统一的模型,以实现多任务和多模态推理。

## 影响声明

本研究通过解决 RAG 任务中导航大规模视觉上下文这一关键挑战,显著提升了多模态智能体系统的能力。通过引入结构化记忆图和基于能量的 token 分配机制,我们的框架不仅提高了推理准确率,还大幅增强了计算效率,推动了多模态大语言模型在资源受限环境中的可持续部署。此外,对推理路径的显式建模增强了智能体行为的可靠性,为开发能够处理长时程任务的可信多模态 AI 系统奠定了坚实基础。

## 参考文献

- Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. A survey on rag with llms. *Procedia computer science*, 246:3781–3790, 2024.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. 2024.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.
- Valeriia Bolotova-Baranova, Vladislav Blinov, Sofya Filippova, Falk Scholer, and Mark Sanderson. Wikihowqa: A comprehensive benchmark for multi-document non-factoid question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5291–5314, 2023.
- Mirco Bonomo and Simone Bianco. Visual rag: Expanding mllm visual knowledge without fine-tuning. *arXiv preprint arXiv:2501.10834*, 2025.
- Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16495–16504, 2022.
- Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. Mindsearch: Mimicking human minds elicits deep ai searcher. *arXiv preprint arXiv:2407.20183*, 2024a.
- Zehui Chen, Kuikun Liu, Qiuchen Wang, Wenwei Zhang, Jiangning Liu, Dahua Lin, Kai Chen, and Feng Zhao. Agent-flan: Designing data and methods of effective agent tuning for large language models. *arXiv preprint arXiv:2403.12881*, 2024b.
- Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.
- Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024.
- Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.
- Yixiong Fang, Tianran Sun, Yuling Shi, and Xiaodong Gu. Attentionrag: Attention-guided context pruning in retrieval-augmented generation. *arXiv preprint arXiv:2503.10720*, 2025.
- Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.
- Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, et al. Webwatcher: Breaking new frontier of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*, 2025.
- Zhuoning Guo, Mingxin Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Xiaowen Chu. Towards universal video retrieval: Generalizing video embedding via synthesized multimodal pyramid curriculum. *arXiv preprint arXiv:2510.27571*, 2025.
- Haoyu Han, Li Ma, Harry Shomer, Yu Wang, Yongjia Lei, Kai Guo, Zhigang Hua, Bo Long, Hui Liu, Charu C Aggarwal, et al. Rag vs. graphrag: A systematic evaluation and key insights. *arXiv preprint arXiv:2502.11371*, 2025.

- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. ICLR, 1(2):3, 2022.
- Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, et al. Vision-deepresearch: Incentivizing deepresearch capability in multimodal large language models. arXiv preprint arXiv:2601.22060, 2026.
- Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. Videorag: Retrieval-augmented generation over video corpus. arXiv preprint arXiv:2501.05874, 2025.
- Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibao Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720, 2026.
- Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhao Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. arXiv preprint arXiv:2501.00574, 2024a.
- Xinze Li, Yixin Cao, Yubo Ma, and Aixin Sun. Long context vs. rag for llms: An evaluation and revisits. arXiv preprint arXiv:2501.01880, 2024b.
- Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. Advances in Neural Information Processing Systems, 37:95963–96010, 2024.
- Rui Meng, Ziyang Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. arXiv preprint arXiv:2507.04590, 2025.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 2630–2640, 2019.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250, 2016.
- Monica Riedler and Stefan Langer. Beyond text: Optimizing rag with multimodal inputs for industrial applications. arXiv preprint arXiv:2410.21943, 2024.
- Yuling Shi, Songsong Wang, Chengcheng Wan, Min Wang, and Xiaodong Gu. From code to correctness: Closing the last mile of code generation with hierarchical debugging. arXiv preprint arXiv:2410.01215, 2024.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. arXiv preprint arXiv:2601.03267, 2025.
- Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. arXiv preprint arXiv:2506.23918, 2025.
- Weiwei Sun, Miao Lu, Zhan Ling, Kang Liu, Xuesong Yao, Yiming Yang, and Jiecao Chen. Scaling long-horizon llm agent via context-folding. arXiv preprint arXiv:2510.11967, 2025a.
- Yubo Sun, Chunyi Peng, Yukun Yan, Shi Yu, Zhenghao Liu, Chi Chen, Zhiyuan Liu, and Maosong Sun. Visrag 2.0: Evidence-guided multi-image reasoning in visual retrieval-augmented generation. arXiv preprint arXiv:2510.09733, 2025b.

- Sijun Tan, Michael Luo, Colin Cai, Tarun Venkat, Kyle Montgomery, Aaron Hao, Tianhao Wu, Arnav Balyan, Manan Roongta, Chenguang Wang, et al. rllm: A framework for post-training language agents, 2025.
- Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pp. 13636–13645, 2023.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.
- Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. arXiv preprint arXiv:2504.07491, 2025.
- Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. arXiv preprint arXiv:2502.18017, 2025a.
- Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. arXiv preprint arXiv:2505.22019, 2025b.
- Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22958–22967, 2025c.
- Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haiyan Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386, 2025d.
- Xixi Wu, Kuan Li, Yida Zhao, Liwen Zhang, Litu Ou, Huifeng Yin, Zhongwang Zhang, Xinmiao Yu, Dingchu Zhang, Yong Jiang, et al. Resum: Unlocking long-horizon search intelligence via context summarization. arXiv preprint arXiv:2509.13313, 2025.
- Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. arXiv preprint arXiv:2502.12110, 2025.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 conference on empirical methods in natural language processing, pp. 2369–2380, 2018.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In The eleventh international conference on learning representations, 2022.
- Rui Ye, Zhongwang Zhang, Kuan Li, Huifeng Yin, Zhengwei Tao, Yida Zhao, Liangcai Su, Liwen Zhang, Zile Qiao, Xinyu Wang, et al. Agentfold: Long-horizon web agents with proactive context management. arXiv preprint arXiv:2510.24699, 2025.



- Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. Universal-rag: Retrieval-augmented generation over corpora of diverse modalities and granularities. arXiv preprint arXiv:2504.20734, 2025.
- Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, et al. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. arXiv preprint arXiv:2507.02259, 2025a.
- Runpeng Yu, Xinyin Ma, and Xinchao Wang. Introducing visual perception token into multimodal large language model. arXiv preprint arXiv:2502.17425, 2025b.
- Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. arXiv preprint arXiv:2410.10594, 2024.
- Yu Zeng, Wenxuan Huang, Shiting Huang, Xikun Bao, Yukun Qi, Yiming Zhao, Qiuchen Wang, Lin Chen, Zehui Chen, Huaian Chen, et al. Agentic jigsaw interaction learning for enhancing visual perception and reasoning in vision-language models. arXiv preprint arXiv:2510.01304, 2025a.
- Yu Zeng, Yukun Qi, Yiming Zhao, Xikun Bao, Lin Chen, Zehui Chen, Shiting Huang, Jie Zhao, and Feng Zhao. Enhancing large vision-language models with ultra-detailed image caption generation. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 26703–26729, 2025b.
- Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. ACM Transactions on Information Systems, 43(6):1–47, 2025.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. arXiv preprint arXiv:2403.13372, 2024.
- Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. arXiv preprint arXiv:2506.15841, 2025.

## A 数据集构建

整体的数据构建流水线如图 7 所示。我们从 Howto100M 数据集构建初始视频语料库  $\mathcal{V}$ 。为确保数据多样性，我们显式地平衡了视频时长的分布。对于一个较长的视频  $v \in \mathcal{V}$ ，我们将其划分为一系列片段  $\{s_1, s_2, \dots, s_n\}$ 。我们方法的一个关键特征是所选片段之间的可变时间间隔。具体而言，我们以短时间和长时间间隔采样片段。该策略分别捕捉稠密的局部上下文和稀疏的全局关系。接着，我们使用多模态大模型（MLLM）为这些片段生成详细的描述  $C$ 。基于  $C$ ，语言模型（LLM）合成一个特定查询  $q$  及其相应的推理步骤。然后，我们对生成的查询应用语义过滤器。至关重要，该过滤器确保  $q$  依赖于大规模多模态语料库，而非一般性问题。具体的提示设计见附录 16。此外，我们从生成的数据中采样一部分以构建一个基准，命名为 XVbench。该基准解决了大规模语料库中跨视频理解缺乏评估标准的问题。

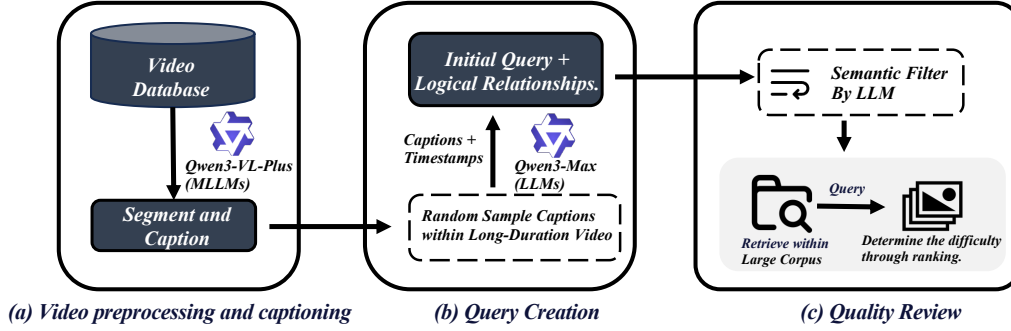


图 7: 数据构建流水线概述。该过程包含三个阶段：(a) 视频预处理，将长视频分割并由大模型生成字幕；(b) 查询生成，基于采样的字幕，大模型生成复杂的查询和逻辑步骤；(c) 质量审查，包括语义过滤和难度秩排序，以确保数据质量和挑战水平。

## B 环境与实验设置

**训练与推理配置。** 我们使用 Llama-Factory Zheng et al. (2024) 进行了 LoRA Hu et al. (2022) 的 SFT，使用 rLLM Tan et al. (2025) 进行了 RL。所有实验均在 NVIDIA H20-3e 141G GPU 上进行。为了优化长上下文多模态任务的训练效率，我们采用了梯度累积策略，总批量大小为 64。SFT 阶段持续 3 轮，学习率为  $1.0e-4$ ，而 RL 阶段则对演员模型采用更保守的学习率  $1.0e-6$ ，以确保策略收敛稳定。详细超参数请参见附录 F。

**搜索引擎的构建** 在训练和推理过程中，我们基于约 20 万个多模态条目的语料库构建了一个搜索引擎。数据集的详细构成如表 4 所示。我们使用 GVE-7B Guo et al. (2025) 作为嵌入模型，因为它支持文本到文本、图像和视频的检索。这使我们能够根据嵌入结果测量条目之间的距离。对于视频，我们将其分割为 1 分钟的片段，然后生成嵌入，遵循该模型的最佳实践。

表 4: 语料库的统计量。我们的情景将来自不同模态的参考信息合并为一个统一的语料库，这对 RAG 系统提出了更高的挑战，并更符合现实世界的应用场景。

Dataset	Domain / Category	Corpus Scale	Query Type
HotpotQA	General Text	3-10 paragraphs per question	Multi-hop reasoning
SQuAD	General Text	Single/Multiple passages	Span extraction
WebQA	Image-Text	Web-scale snippets & images	Multi-hop multimodal
SlideVQA	Visually Rich Doc	52k+ slide images	Multi-hop & Numerical
MMLongBench	Visually Rich Doc	6,492 documents	Long-context understanding
LVBench	Long-context Video	103 long videos (~68m avg.)	Temporal grounding & Reasoning
WikiHowQA	Large Video Corpus	~500 videos (HowTo100M subset)	Retrieval & Generation
Synthetic QA	Large Video Corpus	~500 videos (HowTo100M subset)	Retrieval
XVBench	Cross-Video	Fine-grained Segments from HowTo100M	Cross-video reasoning
Merged (Ours)	Interleaved Multimodal RAG	~200k multimodal items containing text, images, and videos	Complex Long-context Interleaved Reasoning

**基于模型的奖励** 我们采用基于模型的奖励来评估生成回复的质量和相关性。具体而言，我们使用 Qwen3-Max Yang et al. (2025) 作为我们的奖励模型。用于奖励模型的提示如图 12（附录 J.1）所示。给定输入查询、参考答案和生成的回答，奖励模型评估生成回答的正确性，并输出一个二值（0 或 1）以表示答案的准确率。

## C 关于试点研究的更多详情

### C.1 记忆结构的实验情景

为评估不同记忆结构对智能体推理与上下文管理的影响，我们针对多模态语料库的一个代表性子集开展了一项初步研究。在此情景下，我们仅使用视频语料库构建搜索引擎，以便更清晰地观察智能体状态与动作序列之间的交互。

我们实现了并比较了三种结构：(1) **传统 ReAct** 遵循标准的思考-动作-观察环，如附录 J.4 所述，其中整个交互历史被线性连结。(2) **迭代摘要作为记忆** 如附录 J.5 所述，持续将观测结果压缩到先前的记忆状态中，以保持上下文窗口的效率。(3) **结构化图作为记忆** 如附录 J.6 所示，维护一个动态的有向非循环图，其中每个结点显式存储分解后的查询及其对应的提取语义摘要。

结果证实，ReAct 存在状态盲区，导致随着上下文扩展，出现重复查询和无用交互。通过采用基于图形的结构化拓扑来保存智能体状态，跟踪每一次查询及其对应的信息提取，基于图形的记忆机制显著减少了冗余搜索动作，并更有效地管理大规模视觉上下文。

### C.2 记忆模态的实验情景

本部分研究观测与记忆模态之间的语义对齐，以解决压缩比与关键信息保留之间的权衡问题。为系统评估不同模态对智能体验证的影响，我们在记忆范式下采用四种不同的跨模态策略进行实验：(1) **预生成描述** 代表一种仅文本的基准方法，其中语料库的全部视觉成分在检索前被转换为文本描述。智能体仅基于这些预先生成的描述进行推理，虽可最大限度减少 token 使用量，但会丢失细粒度的视觉特征。

(2) **视觉观测作为记忆** 通过直接在上下文窗口中存储原始多模态 token 保持最高保真度。尽管该方法保留了所有视觉细节，但显著降低了信噪比，并在长时程任务中常导致上下文耗尽。

(3) **上下文感知的描述生成** 涉及检索原始多模态数据，但将其以动态文本摘要的形式记忆。该方法试图以压缩的文本形式捕捉任务相关的视觉信息，但在复杂验证任务中常因表示差距而表现不佳。

(4) **语义相关视觉记忆** 实现了一种选择性保留机制。在检索多模态数据后，智能体评估视觉区域的重要性，仅保留相关视觉 token 而丢弃噪声。

### C.3 监督中的信用分配实验设置

我们探讨了稀疏结果奖励在多步智能体轨迹中用于信用分配的可靠性。主要目标是确定轨迹级奖励是否准确反映了单个检索与感知步骤的有效性。遵循 Mem1 Zhou et al. (2025) 中的协议，我们将智能体动作分解为两个互不重叠的子集：捕获关键线索的证据检索步骤，以及代表无关动作或使用相同查询重复检索的噪声或冗余步骤。具体而言，我们重新收集整个推理过程中所有的观察结果，并采用直接推断方法来评估每个观察结果对最终结果的具体贡献。为了最小化上下文长度变化对模型性能的混淆影响，我们使用 Qwen3-VL-Plus 作为该反事实评估的骨干模型。我们在两个维度上进行分析：(1) **对正例 ( $r = 1$ )**：我们在移除证据与移除噪声后评估轨迹。此对比旨在验证冗余步骤的无贡献性，判断在移除非必要动作后，正确的最终答案是否仍然保持。(2) **对负例 ( $r = 0$ )**：我们测试通过降噪轨迹仅保留有效证据集，模型性能是否能够恢复。该过程旨在验证这些证据步骤的固有价值，识别初始失败是否源于累积噪声的推理，而非关键信息的缺失。

## D 对比基准

本文详细介绍了我们对比的基准方法以及我们的复现细节。

1. **原始 RAG**。在检索阶段，它直接使用原始问题来搜索相关文本、图像和视频，然后将这些内容插入上下文以回答问题。请参阅附录 J.3 以获取详细的提示。
2. **ReAct RAG** Yao et al. (2022)。该方法使用思维-动作-观察环格式提示 RAG 智能体。详细提示请参见附录 J.4。

3. VideoRAG [Jeong et al. \(2025\)](#). 该方法执行帧选择以提取推理所需的资讯。我们使用 GVE [Guo et al. \(2025\)](#) 计算帧与查询之间的相似度。尽管此方法专为视频设计，但嵌入模型使我们能够将相同的粗粒度到细粒度的策略应用于文本和图像，作为性能的参考。
4. UniversalRAG [Yeo et al. \(2025\)](#). 通过将任务表述为路由问题，在跨模态语料库中引入了 RAG。我们使用 Qwen3VL-8B (4B) 作为路由器以对齐不同情景，提示词则来自原始代码，以确保公平比较。
5. MemAgent [Yu et al. \(2025a\)](#). 我们通过将长上下文搜索结果依次输入模型的上下文来实现该方法。具体而言，我们直接使用原始问题检索相关的文本、图像和视频，将检索结果的理解视为一个长上下文多模态理解任务，然后利用 MemAgent 处理这一扩展后的上下文，使模型能够在比原始 RAG 更广的有效上下文值域内运行。
6. Mem1 [Zhou et al. \(2025\)](#). 该方法通过循环的检索-记忆过程来更新其记忆。这是一种天然适用于 RAG 任务的上下文管理范式。该方法与第 2.2 节中的初步研究高度相似，并遵循一种迭代摘要范式。通过参考附录 J.5 可以实现该效果的近似版本。我们使用原始的 Mem1 提示来复现该方法。

## E 基准信息

我们在涵盖多种任务的综合性基准上评估了我们的方法：

1. HotpotQA [Yang et al. \(2018\)](#) 是一个专注于多跳问答的大规模数据集，要求在多个文档之间进行推理。该数据集包含约 113,000 个基于维基百科的问题-答案对。与受限于预定义知识库的数据集不同，它包含多样化的自然语言问题，并提供句子级别的支持事实，以使系统能够生成可解释的预测。该数据集还引入了比较类问题，要求模型比较两个实体的属性以推断答案。
2. SQuAD [Rajpurkar et al. \(2016\)](#) 是一个大规模的阅读理解数据集，包含超过 10 万个由众包工作者在一组维基百科文章上创建的问题。与以往依赖选择题答案或填空式任务的数据集不同，SQuAD 要求模型从阅读段落中选择特定的文本片段（跨度）作为答案。该数据集提供了多样化的答案类型，包括日期、实体和从句，并挑战模型处理问题与相应段落之间显著的句法差异。
3. WebQA [Chang et al. \(2022\)](#) 是一个用于模拟开放领域网络搜索场景的多模态数据集。它包含需要在文本片段和图像之间进行多跳推理才能找到正确答案的问题。与标准 VQA 任务中图像作为主要上下文不同，WebQA 将图像和文本均视为有效的知识来源，需要被检索并整合。
4. SlideVQA [Tanaka et al. \(2023\)](#) 是一个专注于理解幻灯片的文档视觉问答数据集。该数据集包含超过 2,600 个幻灯片演示文稿，涵盖 52,000 多张幻灯片图像和 14,500 个问题，这些问题需要复杂的推理能力，如单跳、多跳和数值推理。该数据集旨在支持多种推理类型，并为数值类问题提供了标注的算术表达式，以增强推理能力。
5. MMLongbench [Ma et al. \(2024\)](#) 是一个旨在评估视觉语言模型在长上下文、多模态文档上的文档理解能力的数据集，这些文档由文本、图像、图表、表格和版式结构组成。
6. LVBench [Wang et al. \(2025c\)](#) 是一个专门设计用于评估长视频理解能力的基准。与专注于短片段的数据集不同，它包含 103 个公开获取的视频，平均时长约 68 分钟，涵盖电影、纪录片和体育等多种类别。该数据集包含 1,549 个手动标注的问题-答案对，用于测试六项核心能力，包括时间定位、推理和实体识别。其构建旨在挑战多模态模型展现长期记忆和复杂推理能力，以应对理解长时间上下文的需求。
7. WikiHowQA 与 HowTo100M [Bolotova-Baranova et al. \(2023\)](#); [Miech et al. \(2019\)](#); [Jeong et al. \(2025\)](#) 是一个综合基准，旨在评估基于视频的检索与生成任务。该基准将 WikiHowQA 数据集中的高质量、人工撰写的教学类问答与 HowTo100M 语料库相结合，后者包含来自 YouTube 的数百万个教学视频。通过将文本查询与相关视频进行关联，该数据集评估系统在大规模语料库中搜索正确视频以及生成准确且具有视觉依据的回答的能力。
8. 基于 HowTo100M 的合成问答 [Jeong et al. \(2025\)](#); [Miech et al. \(2019\)](#) 是一个自动生成的数据集，旨在解决检索增强生成 (RAG) 系统中查询-视频-答案三元组训练数据不足的问题。该



数据集基于 HowTo100M 语料库构建，利用先进的大视觉-语言模型，在特定视频基础上生成多样化的问答对。问题设计具有一定的通用性，适用于检索任务——避免过度依赖帧级别的细节，同时仍需理解视频内容才能作答，从而能够全面评估检索与生成两个组件的表现。

9. XVBench 是一个旨在解决跨视频理解评估标准缺失问题的基准。我们使用图 7 中所示的综合性流水线构建该数据集，该流水线基于 Qwen3-Max 实现细粒度视频分割、详细描述生成以及推理图构建。为确保基准的质量和适当的难度，我们采用嵌入距离对样本进行秩排序并有效筛选。

## F 超参数

我们在训练过程中使用的详细超参数如表5和表6所示。问答数据与轨迹的构建过程见附录A。我们采用  $\lambda = 0.1$  和  $\gamma = 0.3$  来表示能量动态，而总资源预算定义为  $S_{\text{total}} = 5 \times 256 \times 32 \times 32$ ，以支持高分辨率特征的保留。在 SFT 和 RL 训练期间，我们对多模态记忆库中的像素进行平均处理，仅在推理阶段启用动态分配。

表 5: SFT 的关键超参数。

Name	Value
Finetuning type	LoRA
LoRA Rank	32
Freeze vision tower	True
Freeze multi-modal projector	True
Freeze language model	False
Cutoff len	16384
Epochs	3
Batch size	8
Gradient accumulation steps	8
Learning rate	1.0e-4
LR scheduler type	cosine

表 6: 强化学习的关键超参数。

Name	Value
Number of agent groups	8
Batch Size	32
Mini batch size	32
Loss Mode	GSPO
Learning rate (Actor)	1.0e-6
KL loss coefficient	0.001 (optional)
Tensor model parallel size	4
Total epochs	2
Max prompt length	20240
Max response length	512
GPU memory utilization	0.6

## G 案例研究

在图 8 和图 9 中，我们描述了 VimRAG 的推理路径，以展示我们的框架如何构建和更新多模态记忆图。这些案例揭示了长序列多模态检索任务中的两个主要挑战。第一个挑战是解决状态盲区问题，以在不处理无关数据的情况下高效搜索大规模视频数据集。第二个挑战涉及动态分辨率分配机制，在满足 token 限制的同时保持详细的视觉信息。

## H 局限性

尽管我们尽了最大努力，本文仍存在一些局限性。首先，提升基础模型的能力有助于系统整体性能的改善。其次，当前的多轮交互可能无法满足高实时性应用的需求。最后，当前检索器的准确率仍有待提高，以更好地支持 RAG 系统。

## I 伦理声明

本研究致力于提升多模态智能体系统的效率与准确率。我们实验中使用的数据集，如 HowTo100M 和 HotpotQA，均为公开数据集，不包含任何私有或敏感的个人敏感信息。此外，通过优化 token 分配并减少不必要的计算，我们的方法有助于构建更加节能的人工智能系统，契合可持续计算的目标。

## J 提示词

在本节中，我们展示了论文中使用的所有提示。

### J.1 基于模型的评估提示

见图 12。

### J.2 问题验证提示

参见图 16

### J.3 香草 RAG 提示

见图 11。

### J.4 ReAct 提示

见图 10。

### J.5 摘要作为记忆提示

见图 13。内存由汇总信息直接更新。

### J.6 图示记忆提示

见图 14。内存中仅包含语义图。

### J.7 VimRAG 提示

见图 15。

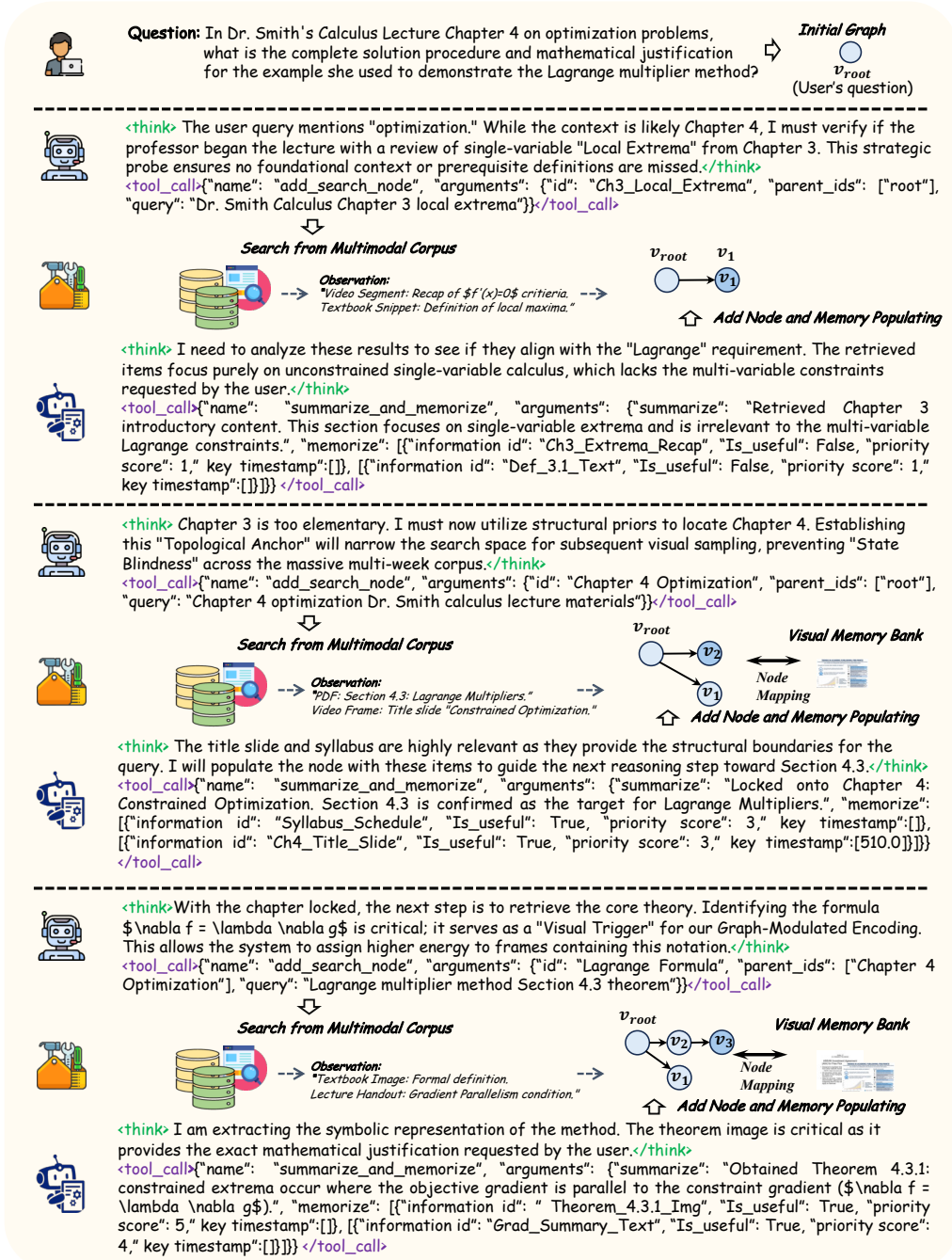


图 8: 案例研究 (第一部分) . 智能体初始化多模态记忆图, 以解决关于微积分讲座的复杂查询。

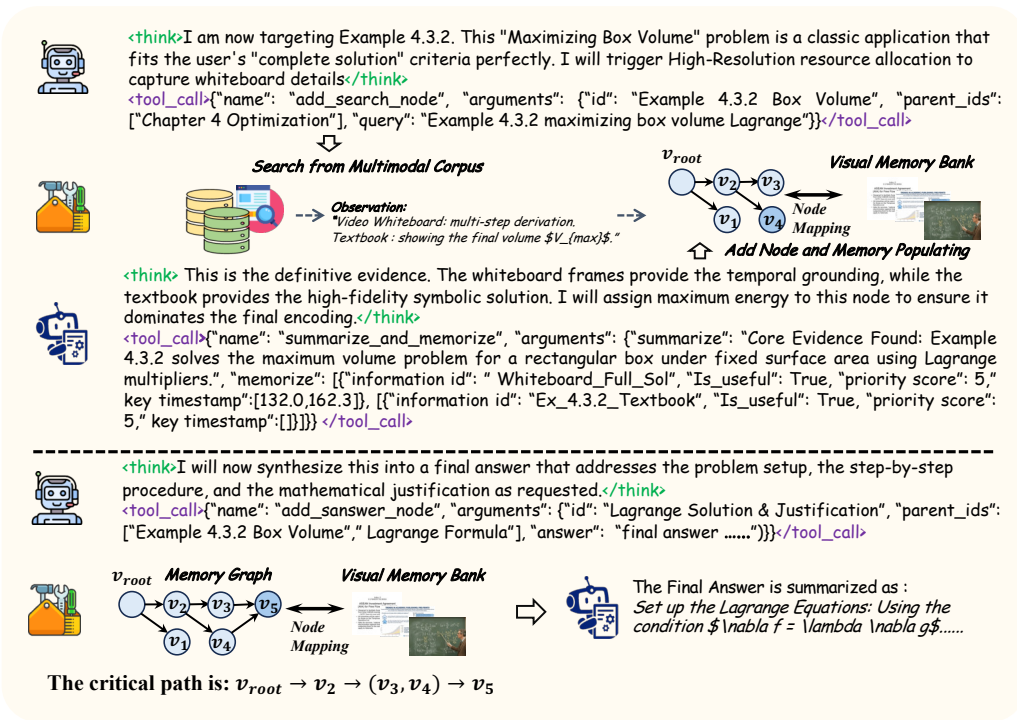


图 9: 案例研究（第二部分）。最终答案通过遍历关键路径 ( $v_{root} \rightarrow v_2 \rightarrow (v_3, v_4) \rightarrow v_5$ ) 得出。



## ReAct Prompt

**系统提示：**

我是一个能够使用工具调用解决任何问题的专家。为此，我已获得搜索引擎和各种工具的访问权限。

**可用工具**

1. **查询：**根据查询收集相关信息。(参数：关键词；返回：查询结果)。
2. **答案：**根据搜索结果直接回答用户的问题。(参数：response)

**工作流**

1. 理解用户的查询并识别询问的核心要点。
2. 使用 `search` 函数制定清晰且具体的搜索字符串。
3. 根据所获信息，撰写一份清晰简洁的最终答复。

**严格禁止的行为**

1. 不得使用未通过指定工具获取的信息来提供答案。
2. 不得编造或超出工具返回内容进行推断。
3. 不要输出模糊的总结或未经证实的猜测。
4. 不要调用搜索引擎，同一回复中给出答案。

**回复格式****必须**

选项 1：搜索

`<thinking>` 你的推理过程 `</thinking>`

`<search>` 您的查询 `</search>`

选项 2：回答

`<thinking>` 你的推理过程 `</thinking>`

`<answer>` 您的详细回答 `</answer>`

## User Prompt:

**执行说明**

`<thinking>` 已理解任务要求，将对提供的 LaTeX 代码进行翻译。翻译仅针对文本内容，不包括 LaTeX 命令和数学表达式。确保输出为有效的 LaTeX 格式，并且不包含任何解释性文字。`</thinking>`

2. 如果你缺乏知识，请使用 `<search>` 标签调用搜索引擎。

您可以随意进行多次搜索。

4. 收集到足够信息后，请将最终答案放在 `<answer>` 标签内。

**所需响应格式**

搜索时：

您的推理过程

您的查询

作答时：

您的推理过程

您的详细回答

**用户查询**

`{查询描述}`

图 10: ReAct 的提示

## Vanilla RAG Prompt

**系统提示：**

<thinking> The question requires translating LaTeX code into Simplified Chinese while preserving LaTeX commands and mathematical expressions. The instruction specifies that only the text should be translated, not the commands or math content. Since no specific LaTeX code was provided in the user's input, there is no content to translate. Therefore, the response must remain empty or indicate that no translation is needed due to missing input.  
</thinking> <answer></answer>

User Prompt:

查询

{查询描述}

检索到的多模态信息

{检索到的视频 / 图像 / 文本 token}

图 11: Vanilla RAG 的提示。

## Reward Model Prompt

**系统提示：****角色介绍**

你是一个用于评估问答聊天机器人的专家评价系统。

你得到了以下信息：

- 查询
- 生成的答案

参考答案

你的任务是评估生成答案的正确性。

**响应格式**

True or False

(无原文可译，翻译结果为空) judge: False

请注意，生成的答案可能包含超出参考答案的额外信息。

User Prompt:

查询：{查询描述}

参考答案：{参考答案}

{生成的答案}

图 12: 基于模型的奖励提示。

## Iterative Summarization as Memory Prompt

**系统提示：**

请提供具体问题和之前记忆的内容，以便我为您解答。

为此，你已获得搜索引擎的使用权限。

**### 可用工具**

搜索：

- 根据查询收集相关信息。
- 参数：要搜索的关键词或问题。

返回：查询的搜索结果。

**### workflow**

1. 理解用户的查询并识别询问的核心要点。
2. 使用 `search` 函数检索相关信息时，制定清晰且具体的搜索字符串。
3. 每次调用搜索引擎时，都需要根据搜索结果和当前记忆更新记忆。
4. 如果信息充分，请给出清晰简洁的最终答复。

**### 要求**

1. 确保工具使用准确，查询问题表述清晰。
2. 准确且条理清晰地回答用户问题。
3. 如果初始结果不足，重复搜索尝试。
4. 你只能提供最终答案或使用搜索引擎，不能在同一回复中同时使用两者。
5. 您必须至少调用一次搜索引擎以获取搜索结果。
6. 按照响应格式。

**### 严格禁止的行为：**

1. 使用未通过指定工具获取的信息提供答案。
2. 伪造或超出工具返回内容进行推断。
3. 输出模糊的总结、假设性判断或未经证实的猜测。
4. 在调用搜索引擎时，反复使用语义相似的查询。
5. 不要调用搜索引擎，同一回复中给出答案。

**### 回复格式**

**必须回答以下格式：**

当您需要查询时，必须以下列格式提供查询内容：

`<think>` 你的推理过程 `</think>`

您的查询

当您需要更新记忆时，您必须按照以下格式提供摘要：

`<update_memory>` 更新的记忆 `</update_memory>`

当你收集到足够的信息来回答问题时，请立即提供你的回答。

`<think>` 你的推理过程 `</think>`

您的详细回答

User Prompt: Question: {查询} 记忆: {memory}

Memory Update Prompt:

请立即提供更新后的内存信息：

`<update_memory>` 更新的记忆 `</update_memory>`

检索到的多模态信息：{检索到的视频 / 图片 / 文本 token}

图 13: 迭代摘要作为记忆的提示。

## Graph as Memory Prompt

### 系统提示：

你是一个智能体，旨在通过直接回答或迭代搜索信息来解决用户查询。你的目标是构建一个有向非循环图 (DAG)，以表示单个用户查询的搜索过程，其中每个结点对应推理或信息收集的一个步骤。

### 图结点

该图包含三种类型的结点：

- **root**: 表示用户原始问题的初始结点。
- **search**: 代表向外部搜索引擎发出的查询的结点。每个搜索结点必须具有唯一且高度概括的标题，以捕捉查询的意图，并且必须与之前的查询有显著不同。
- **answer**: 最终结点，您在此处提供对用户问题的完整回答。此结点没有 ID。

### 规则

1. 每回合只能添加一个结点。
2. 每个 **search** 结点必须满足：(a) 拥有一个唯一的 id (标题)，该标题为简短且具有描述性的短语，用于总结查询意图；(b) 通过有向边与父结点相连 (指定 **parent\_id**)；(c) 包含一个 **query** 字段，其中包含实际的搜索字符串；(d) 该查询必须与之前的查询有显著差异。
3. 发出查询后，您将收到结果。随后，您必须将这些结果中的相关内容总结成一个简洁的 **summary** (该内容将在外部添加到结点中)。
4. 您必须在每一步决定是：直接回答 (输出一个 **answer** 结点)，还是搜索 (输出一个带有新查询的 **search** 结点)。
5. 查询必须与先验查询有实质性差异——避免冗余或对同一观点的换种表述。
6. 在生成一个 **search** 结点时，使用 **add\_search\_node** 函数。
7. 收到搜索结果后，可以使用 **summary\_search\_node** 函数对结果进行总结。
8. 一旦您认为已有足够的信息来回答问题，请输出一个 **add\_answer\_node** 函数调用。

### 可用工具

#### 1. **add\_search\_node**

**描述**：在图中创建一个新的搜索结点。此工具应用于向外部引擎发出搜索查询。每个结点必须具有唯一且概括性的 ID，以反映其意图。

**参数**：

- **id**: 用于捕捉查询意图的唯一、简短且具有描述性的结点标题。
- **parent\_ids**: 此搜索所源自的前驱结点的 ID。
- **query**: 要执行的实际查询。必须与所有先验查询有显著不同。

#### 2. **add\_answer\_node**

**描述**：创建图中的最终结点，包含用户原始问题的完整且最终的答案。

**参数**：

- **parent\_ids**: 提供生成此最终答案所需信息的结点的 ID。
- **answer**: 用户问题的全面且完整的最终答案。

#### 3. **总结并记忆**

**描述**：必须在每次调用 **'add\_search\_node'** 后无一例外地调用的工具。它作为搜索结点的最终步骤，将原始数据过滤为高密度内存。

**参数**：

- **summarize**: 用 1-3 句话对信息进行总结，严格聚焦于直接回应用户意图的事实。如果搜索未返回相关信息，请明确说明。

### 流程

- 以用户的查询作为根结点。
- 每回合时，评估你现在是否能够作答，还是需要更多信息。
- 如果进行搜索：发出一个带有新 **search** 结点的 **add\_node** MCP 命令。





图 14: 图示记忆提示

## VimRAG Prompt

**系统提示：**

你是一个智能体，旨在通过直接回答或迭代搜索信息来解决用户查询。你的目标是构建一个有向非循环图 (DAG)，以表示单个用户查询的搜索过程，其中每个结点对应推理或信息收集的一个步骤。

**图结点**

该图包含三种类型的结点：

- **root**: 表示用户原始问题的初始结点。
- **search**: 代表向外部搜索引擎发出的查询的结点。每个搜索结点必须具有唯一且高度概括的标题，以捕捉查询的意图，并且必须与之前的查询有显著不同。
- **answer**: 最终结点，您在此处提供对用户问题的完整回答。此结点没有 ID。

**规则**

1. 每回合只能添加一个结点。
2. 每个 **search** 结点必须满足：(a) 拥有一个唯一的 **id** (标题)，该标题为简短且具有描述性的短语，用于总结查询意图；(b) 通过有向边与父结点相连 (指定 **parent\_id**)；(c) 包含一个 **query** 字段，其中包含实际的搜索字符串；(d) 该查询必须与之前的查询有显著差异。
3. 发出查询后，您将收到结果。随后，您必须将这些结果中的相关内容总结成一个简洁的 **summary** (该内容将在外部添加到结点中)。
4. 您必须在每一步决定是：直接回答 (输出一个 **answer** 结点)，还是搜索 (输出一个带有新查询的 **search** 结点)。
5. 查询必须与先验查询有实质性差异——避免冗余或对同一观点的换种表述。
6. 在生成一个 **search** 结点时，使用 **add\_search\_node** 函数。
7. 收到搜索结果后，可以使用 **summary\_search\_node** 函数对结果进行总结。
8. 一旦您认为已有足够的信息来回答问题，请输出一个 **add\_answer\_node** 函数调用。

**可用工具**1. **add\_search\_node**

**描述：**在图中创建一个新的搜索结点。此工具应用于向外部引擎发出搜索查询。每个结点必须具有唯一且概括性的 ID，以反映其意图。

**参数：**

- **id**: 用于捕捉查询意图的唯一、简短且具有描述性的结点标题。
- **parent\_ids**: 此搜索所源自的前驱结点的 ID。
- **query**: 要执行的实际查询。必须与所有先验查询有显著不同。

2. **add\_answer\_node**

**描述：**创建图中的最终结点，包含用户原始问题的完整且最终的答案。

**参数：**

- **parent\_ids**: 提供生成此最终答案所需信息的结点的 ID。
- **answer**: 用户问题的全面且完整的最终答案。

3. **总结并记忆**

**描述：**必须在每次调用 'add\_search\_node' 后无一例外地调用的工具。它作为搜索结点的最终步骤，将原始数据过滤为高密度内存。即使检索到的信息与用户查询完全无关，也必须执行此工具以正式关闭当前搜索周期。

**参数：**

- **summarize**: 用 1-3 句话对信息进行总结，严格聚焦于直接回应用户意图的事实。如果搜索未返回相关信息，请明确说明。
- **memorize**: 包含所有项目 (文本、图像、视频) 的完整列表。每个项目包括：
  - **information\_id**: 唯一标识符 (例如, 'Text 1')。

- `is_useful`: 值的布尔判断。
- `key_timestamp`: 秒数数组（用于视频）或空。
- `priority_score`: 1（次要）到 5（关键）。

### 流程

- 以用户的查询作为根结点。
- 每回合时，评估你现在是否能够作答，还是需要更多信息。
- 如果进行搜索：发出一个带有新 `search` 结点的 `add_node` MCP 命令。
- 搜索返回后：发出包含结果洞察的 `summary` MCP 命令。
- 重复直到你能回答为止。

**重要：**每回合只能执行一个动作。不要合并动作。必须严格遵循 MCP 格式。

### 回复格式

对于每个函数调用，返回一个包含函数名称和参数的 JSON 对象，并将其放在 `<tool_call></tool_call>` XML 标签内，同时将你的推理过程放在 `<thinking></thinking>` XML 标签内：

思考

你的推理过程……

`</思考 >`

`< 工具 _ 调用 >`

`{"name": < 函数名 >, "arguments": <args-json 对象 >}`

`</工具调用 >`

User Prompt:

### 用户查询

{[查询描述](#)}

### 智能体动作图

{[语义动作图](#)}

### 多模态记忆库

{[视觉 token](#)}

### 记忆提示：

查询结果已返回，请仔细分析并提供与原始用户问题直接相关的信息的简洁、客观摘要；若信息足以回答问题，请给出答案结点。

**您的总结应包含：**

简洁明了（1-3 句话）。

仅关注结果中的关键见解或事实。

避免冗余或猜测。

突出说明此信息如何有助于解决问题或缩小答案范围。

- 不重复先前结点已知或涵盖的内容。

**检索到的多模态信息**

{[检索到的视频 / 图像 / 文本 token](#)}

图 15: VimRAG 的提示。模型根据系统提示和用户提示执行检索或生成答案。如果触发了检索，它将根据记忆提示的指导向图中添加结点。

## Question Verifier Prompt

**系统提示：**

我这里有一些问答数据，你可以观察到问题可分为两类：

**类别 #A：**当您仅看到此问题而没有给出文档时，您确定能在语料库中找到唯一一个文档来提供唯一答案。该问题包含一些关键词，帮助您从语料库中定位到相关文档。

**类别 #B：**当您单独看到这个问题而没有给定文档时，你会发现很难为这个问题定位到一个文档以给出确定性答案，因为您在语料库中会发现多个候选文档，这些文档可能导致该问题出现不同的答案。这个问题没有任何特殊关键词可以帮助你从语料库中定位文档。

**示例：**

右侧边距旁边的数字？ #B

第二个表格中提到的日期是什么？ #B

PUF 的完整形式是什么？ #A

页面底部的粗体数字是多少？ #B

谁在商用飞机客舱空气质量研究中报告了结果？ #A

该公司的名称是什么？ #B

致谁？ #B

来源是什么？ #B

文档的标题是什么？ #B

主题是什么？ #B

里士满小组主观抽吸的万宝路薄荷烟是什么型号？ #A

这是何种类型的通信/信件？ #B

根据所列要求，女性吸烟者的年龄组必须是？ #A

在原型制作和环形翻转过程中，观察到一些香烟的纸张上出现了烧穿孔。 #A

有多少种不同的机制似乎在烟柱破裂成多维流场的过程中起作用？ #A

会议是在哪里举行的？ #B

这封信中抄送的人是谁？ #B

在加粗的条件下，混合物 #24 的初级生产将在哪一天完成？ #A

从纺纱筒子到织造之间的纬纱准备有哪些步骤？ #A

织造技术有三种类型：机织、针织和非织造。

介于中层管理者和非管理类员工之间的层级是什么？ #A

组织中詹姆斯担任英国数字战略领导角色的协作模型的六个部分是什么？ #A

CONCERN 的协作模型包含六个部分： #A

土地征收的透明化过程是以下哪一项的实例？ #A

旧法第 4 条在新法中对应哪一条？ #A

更新或活动的持续时间较短且强度高于项目，并且经常通过离线媒体支持，是否具有较低的强度？ #A

在市场中，如果很少有人对这类内容感兴趣，但又能使你与众不同，那么这种差异化是高还是低？ #A

**User Prompt:**

根据以下标准，将下列问题标记为 #A 或 #B：

{question}

图 16: 问题验证器的提示