自动化学报
ACTA AUTOMATICA SINICA

面向示意图问答的前瞻性多视角视觉推理框架

张新宇 吴艳瑞 张玲玲 杨泽晟 董宇轩 武亚强 郑庆华

A Look-ahead Multi-perspective Visual Reasoning Framework for Diagram Question Answering

ZHANG Xin-Yu, WU Yan-Rui, ZHANG Ling-Ling, YANG Ze-Sheng, DONG Yu-Xuan, WU Ya-Qiang, ZHENG Qing-Hua

在线阅读 View online:
http://www.aas.net.cn/article/current/c17d1126907901407a7c6d93e37114dd3c2bc1f6d62a7ae324627ba371f525a2

您可能感兴趣的其他文章

基于全局覆盖机制与表示学习的生成式知识问答技术
Generative Knowledge Question Answering Technology Based on Global Coverage Mechanism and Representation Learning
自动化学报. 2022, 48(10): 2392-2405 https://doi.org/10.16383/j.aas.c190785

视觉语言导航研究进展
Recent Advances in Vision-and-language Navigation
自动化学报. 2023, 49(1): 1-14 https://doi.org/10.16383/j.aas.c210352

面向智能生化实验室的机器人感知、规划与控制技术
Robot Perception, Planning, and Control Technologies for Intelligent Biochemical Laboratories
自动化学报. 2025, 51(9): 1899-1921 https://doi.org/10.16383/j.aas.c240714

面向具身操作的视觉语言动作模型综述
Survey of Vision-Language-Action Models for Embodied Manipulation
自动化学报. 2026, 52(1): 18-51 https://doi.org/10.16383/j.aas.c250394

基于语言视觉对比学习的多模态视频行为识别方法
Multi-modal Video Action Recognition Method Based on Language-visual Contrastive Learning
自动化学报. 2024, 50(2): 417-430 https://doi.org/10.16383/j.aas.c230159

视觉语言动作模型综述: 从前史到前沿
Vision-Language-Action Models: From the Early Foundations to the State-of-the-Art
自动化学报. 2025, 51(9): 1922-1950 https://doi.org/10.16383/j.aas.c250417

面向示意图问答的前瞻性多视角视觉推理框架

张新宇 [1,2] 吴艳瑞 [1,3] 张玲玲 [1,2] 杨泽晟 [1,3] 董宇轩 [1,3] 武亚强 [4] 郑庆华 [5]

摘 要 针对多模态大模型在抽象度示意图 (如几何、电路、科学图表) 理解中面临的视觉感知静态化挑战, 提出一种基于前瞻性机制的多视角视觉推理框架. 该方法旨在模拟人类"慢思考"的认知机理, 通过引入相对注意力与信息增益门控, 实现对关键视觉线索的主动搜索与去噪; 利用逻辑置信度与视觉相关性进行双维质量评估, 构建"观察–假设–验证"的动态闭环推理系统. 实验表明, 该方法在多学科的示意图推理数据集上显著优于现有主流方法, 在并未引入过多计算开销的同时, 大幅提升了复杂示意图问答的准确率与鲁棒性.

关键词 视觉语言模型; 视觉推理; 示意图问答; 前瞻性机制

引用格式 张新宇, 吴艳瑞, 张玲玲, 杨泽晟, 董宇轩, 武亚强, 郑庆华. 面向示意图问答的前瞻性多视角视觉推理框架. 自动化学报, xxxx, **xx**(x): x−xx

DOI 10.16383/j.aas.c260129 **CSTR** 32138.14.j.aas.c260129

A Look-ahead Multi-perspective Visual Reasoning Framework for Diagram Question Answering

ZHANG Xin-Yu[1,2] WU Yan-Rui[1,3] ZHANG Ling-Ling[1,2] YANG Ze-Sheng[1,3] DONG Yu-Xuan[1,3]
WU Ya-Qiang[4] ZHENG Qing-Hua[5]

Abstract To address the static visual perception challenge faced by multimodal large models in understanding abstract diagrams (such as geometry, circuits, and scientific charts), a multi-perspective visual reasoning framework based on a look-ahead mechanism is proposed. Aiming to simulate the cognitive mechanism of human "slow thinking", this method achieves the active search and denoising of key visual clues by introducing relative attention and information gain gating; Furthermore, it utilizes a dual-dimensional quality evaluation of logical confidence and visual relevance to construct a dynamic "observation-hypothesis-verification" closed-loop reasoning system. Experiments demonstrate that this method significantly outperforms existing mainstream approaches on multidisciplinary diagram reasoning datasets, substantially improving the accuracy and robustness of complex diagram question answering without introducing excessive computational overhead.

Keywords vision-language models; visual reasoning; diagram question answering; look-ahead mechanism

Citation Zhang Xin-Yu, Wu Yan-Rui, Zhang Ling-Ling, Yang Ze-Sheng, Dong Yu-Xuan, Wu Ya-Qiang, Zheng Qing-Hua. A look-ahead multi-perspective visual reasoning framework for diagram question answering. *Acta Automatica Sinica*, xxxx, **xx**(x): x−xx

当前视觉语言模型 (vision-language models, VLMs) 在通用自然场景 (如图像理解、自动驾驶) 下表现优异[1−3], 但在面对涵盖多学科的复杂推理任务时仍显力不从心[4−5]. 这种局限性在数理科学等核心教育场景中尤为突出[6], 因为此类学科知识主要以高度抽象的示意图 (diagram) 形式呈现 (如几何证明、力学模型等). 这类任务不仅要求模型具备视觉感知能力, 还需进行严谨的跨模态逻辑推演[7−8], 因此现有模型难以满足智能教育场景对精准解题与辅导的实际需求.

这一性能瓶颈的根源, 在于现有模型的主流架构难以适应教育数据中"稀疏视觉语义"与"高密度逻辑"的深度耦合. 与现实世界中语义直观的自然图像截然不同, 教育示意图往往在空白背景中仅由线条与符号构成, 其关键解题信息高度浓缩于微小

收稿日期 2026-02-14 录用日期 2026-05-13
Manuscript received February 14, 2026; accepted May 13, 2026
教育部基础学科和交叉学科突破计划 (JYB2025XDXM116), 国家自然科学基金 (62137002, 62293553, 62477036) 资助
Supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM116) and National Natural Science Foundation of China (62137002, 62293553, 62477036)
本文责任编委 黄华
Recommended by Associate Editor HUANG Hua
1. 西安交通大学计算机科学与技术学院 西安 710049 2. 西安交通大学智能网络与网络安全教育部重点实验室 西安 710049 3. 西安交通大学陕西省大数据知识工程重点实验室 西安 710049 4. 联想人工智能技术中心 北京 100094 5. 同济大学 上海 200092
1. School of Computer Science and Technology, Xi′an Jiaotong University, Xi′an 710049 2. Ministry of Education Key Laboratory of Intelligent Networks and Network Security, Xi′an Jiaotong University, Xi′an 710049 3. Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, Xi′an Jiaotong University, Xi′an, 710049 4. Lenovo AI Technology Center, Beijing 100094 5. Tongji University, Shanghai 200092

的视觉线索之中 (如化学角标、几何切点、矢量指向等)[9]. 人类在求解此类问题时, 会进行动态的“看图–思考”循环, 注意力随推理步骤在图像细节与整体结构间灵活切换; 相比之下, 现有方法多采用“一次性编码 (one-pass encoding)”策略, 在完成初步视觉感知后便完全依赖文本推理. 这种静态的感知模式使得模型在长链推理中无法回溯视觉信息, 极易产生“逻辑自洽但视觉失实”的幻觉[10–11], 导致最终答案缺乏可靠的视觉依据.

如图 1 所示, 现有的主流推理范式应用于示意图问题求解时均存在本质局限. 思维链 (chain-of-thought, CoT) 范式 (如图 1 中的 1)) 依赖模型的直觉式响应 (System 1)[12], 由于推理过程缺乏中间步的视觉校准, 其错误往往在推理早期便开始积累且无法被纠正. 为克服这一缺陷, 蒙特卡洛树搜索 (Monte Carlo tree search, MCTS) 等搜索增强范式 (如图 1 中的 2)) 被相继提出[13–15], 它们试图通过扩大搜索空间或预测未来输出来增强推理的鲁棒性. 然而, 这类方法本质上仍局限于文本概率空间的优化, 尽管 MCTS 在扩展过程中成功涵盖包含正确答案的推理路径 (如: 65°), 但在最终决策时, 其评估机制依赖于后验的逻辑自洽性而非图像证据匹配度, 导致正确答案在搜索过程中被“看见”却未被“选中”. 模型本质上是在包含大量“视觉误判分支”的噪声空间中进行低效试错, 这不仅会造成严重的算力冗余, 且依旧难以从根本上剔除那些“逻辑通顺但视觉错误”的幻觉路径[16].

针对现有单次编码范式在教育示意图推理中存在的“感知–推理协同鸿沟 (perception-reasoning synergy gap)”, 本文提出一种基于前瞻性机制的多视角视觉推理框架 (look-ahead multi-perspective framework, LaMP). 该方法旨在模拟人类学习者进行深度思考 (System 2) 的认知机理[17], 将传统的线性单向推理重构为一个“观察–假设–验证”的动态闭环系统. 与被动接收全局特征的传统方法不同, LaMP 的核心理念在于建立一种“推理驱动 (reasoning-driven)”的主动感知模式. 该框架填补了视觉感知与逻辑推演无法同步互动的协同鸿沟, 引入“前瞻性视觉验证机制”, 使得模型能够预演多种可能的推理路径, 并依据这些前瞻假设, 动态地从复杂的示意图背景中解耦出关键的细粒度视觉线索 (无论是几何切点还是地理标识) 以进行一致性校验. 通过这种迭代式交互, LaMP 确保了生成的每一步推理不仅在逻辑上自洽, 更在视觉上拥有确凿的证据支撑, 从而显著提升了模型在多学科教育综合任务上的准确率与可解释性.

为了验证 LaMP 的有效性, 本文在多个包含示意图的推理数据集上进行广泛实验. 结果表明, LaMP 在不同参数规模的基座模型上均取得显著优于现有主流 (state-of-the-art, SOTA) 方法的性能; 相较于原始基线模型, 该方法在平均准确率上实现约 4.2% 至 5.2% 的显著提升. 此外, 计算开销分析表明, LaMP

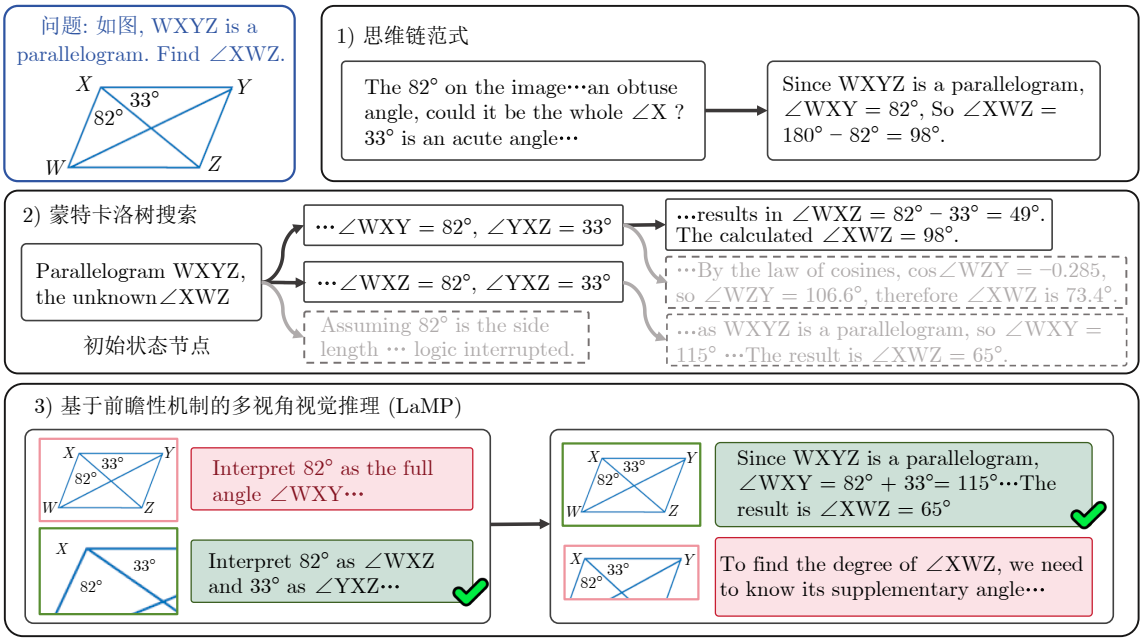


图 1 不同视觉推理机制在求解几何问题上的对比

Fig. 1 Comparison of different visual reasoning mechanisms in solving geometry problem

在大幅增强复杂跨模态任务推理能力的同时, 对比基线模型并未引入过大的额外计算负担, 成功实现推理性能与计算效率的优良平衡.

本文的主要贡献总结如下:

1) 提出一种主动视觉推理新范式: LaMP 框架突破了传统方法"被动看图"的桎梏, 通过前瞻性搜索与动态视觉交互, 将多模态推理从静态的开环过程转变为动态的闭环探索, 有效解决了示意图推理中视觉感知静态化的难题;

2) 构建"视觉–逻辑"双重校准机制: 设计一套包含相对注意力提取、自适应信息增益门控及双维度 (逻辑自信度与视觉相关性) 评分的完整算法流程, 从机理上保证了推理路径在逻辑上的严密性与视觉上的忠实度;

3) 验证多学科场景下的优越性: 在多个示意图复杂推理基准上的实验证明, LaMP 能够有效提升模型的视觉推理性能, 为智能教育应用的落地提供新的理论视角与技术路径.

1 相关工作

1.1 基于大语言模型的推理规划与优化

大语言模型 (large language models, LLMs) 的出现为通用人工智能带来了新的机遇[18]. 其逐步推理能力[12] 通过引入树搜索等更具表达力的算法得到显著增强. 例如, ToT (tree of thoughts)[13] 通过维护推理状态树来探索多种可能性, 在此基础上, 部分工作进一步结合多源知识注入以提升常识问答的可靠性[19]. 而基于蒙特卡洛树搜索 (MCTS) 的框架 (如 RAP[14]) 则进一步将语言模型作为世界模型进行推理规划. 同时, 前瞻性解码 (predictive decoding)[15] 通过预测未来输出来引导当前生成, 增强了推理的全局一致性. 为了进一步提升推理的闭环能力, 研究者引入了自我反思 (self-reflexion)[20]、过程监督 (process supervision)[21] 以及持续学习 (continual learning)[22] 的演进思想, 使模型能够识别并修正自身的逻辑谬误. 此外, 借鉴控制领域中的世界模型[23] 与组合优化思想, 规划 (planning) 算法被广泛应用于解决复杂长程任务. ReAct[24] 通过交错生成推理轨迹与行动来协同处理任务, 这一范式也推动了 LLMs 工具学习 (tool learning) 领域的快速发展[25]. 尽管这些 System 2 范式在纯文本推理中取得了突破, 但将其有效迁移至多模态领域, 仍需解决视觉反馈稀疏与跨模态对齐困难等核心问题.

1.2 多模态思维链与视觉推理

关于多模态思维链 (multimodal CoT) 推理的研究主要沿袭两大路径. 第一类是专用视觉专家模型, 该类方法通过为视觉语言模型 (VLMs) 配备绘图工具并增强空间推理能力来实现. 例如, Visual Sketchpad[26] 引入了辅助的草图绘制步骤来辅助思维过程, 而 CogCoM[27] 则通过操作链 (chain-of-manipulations) 机制使模型能够主动提取视觉证据, 这与视觉–语言–动作 (vision-language-action, VLA) 模型的发展趋势不谋而合[28]. 第二类是改进的提示策略与视觉指令微调, 旨在激活模型的隐含推理能力. T-SciQ[9] 利用大语言模型生成的数据合成策略来增强科学问答能力. 近期, InstructBLIP[29] 等工作通过大规模视觉指令微调, 显著提升了模型对细粒度视觉特征的敏感度. 与上述工作不同, 本文提出的 LaMP 框架不依赖外部专家模型, 而是通过内源性的前瞻性注意力机制, 实现了推理过程中的动态视觉纠错.

2 前瞻性多视角视觉推理方法

2.1 整体框架

不同于自然图像语义在全局空间上的均匀分布, 教育领域的示意图 (如几何证明、电路拓扑或化学分子式) 具有显著的"背景极度稀疏、关键信息高度浓缩"的视觉特性. 传统的静态视觉编码范式往往被大量空白背景主导, 难以捕捉微小但决定性的符号线索 (如垂直标记、电流方向). 为了弥合由此产生的"感知–推理协同鸿沟", LaMP 借鉴了人类学生在处理此类高密度信息时的"系统 2 (System 2)"认知机理, 即并非依赖一次性浏览, 而是基于逻辑推断主动回溯图像以搜寻证据.

据此, 如图 2 所示, 本文将推理过程重构为一个"观察–假设–验证"的动态闭环, 共包含四个层级递进的协同阶段: 模型首先利用相对注意力机制 (第 2.2 节) 抑制无效背景噪声, 根据当前推理上下文动态解耦出关键的视觉线索 (如几何切点或物理受力区域); 随后, 引入信息增益门控 (第 2.3 节) 类比"考点筛选"过程, 自适应过滤低信息量的背景区域, 实现对高密度视觉区域的细粒度感知; 在此基础上, 模型并行生成前瞻性假设 (第 2.4 节), 并通过包含逻辑置信度与视觉相关性的双重校验机制, 进行"图文一致性校验", 剔除符合语言习惯但缺乏图像证据的"伪科学"解释, 从而动态决策出最优推理路径并更新上下文 (第 2.5 节).

上述过程不断迭代 (详见算法 1), 从而构建出一条推理过程严密且视觉信息充分的完整的推理过程.

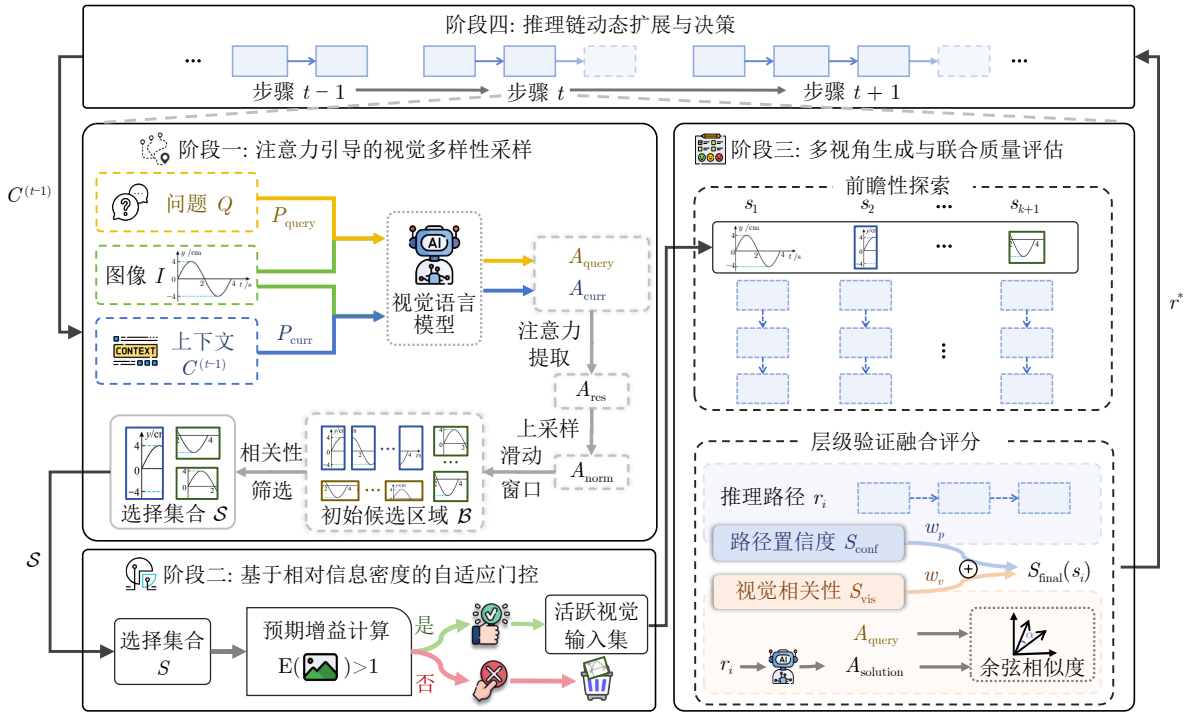


图 2 LaMP 整体架构示意图

Fig. 2 Diagram of the overall architecture of LaMP

算法 1. LaMP 前瞻性推理算法流程

输入. 原始图像 I, 查询问题 Q, 最大迭代步数 $T_{\max}$.

输出. 推理链上下文 C 及最终答案.

1) 初始化上下文 $C^{(0)} \leftarrow \emptyset$
2) **For** $t = 1$ **to** $T_{\max}$;
3) **阶段一: 注意力引导的视觉多样性采样**
4) 构造 P_{query} 和 P_{curr}, 提取相对注意力 A_{res};
5) 上采样并归一化得到 A_{norm};
6) 生成候选区域集合 $\mathcal{S}$.
7) **阶段二: 基于相对信息密度的自适应门控**
8) 初始化活跃输入集 $\mathcal{I}_{active} \leftarrow \{I_{global}\}$;
9) **For** 每个候选区域 $B_k \in \mathcal{S}$ **do**
10) 计算预期信息增益 $E(B_k)$;
11) **if** $E(B_k) \geq 1$ **then**
12) $\mathcal{I}_{active} \leftarrow \mathcal{I}_{active} \cup \{I_{crop}(B_k)\}$
13) **End if**
14) **End for**
15) **阶段三: 多视角生成与联合质量评估**
16) 并行生成候选推理样本集 $\{s_i\}$;
17) 计算路径置信度 S_{conf} 与视觉相关性 S_{vis};
18) 计算归一化融合得分 S_{final}.
19) **阶段四: 推理链动态扩展与决策**
20) 选择最优样本 $s^* = \arg\max_{s \in \mathcal{I}_{active}} S_{final}(s)$;
21) 更新上下文 $C^{(t)} \leftarrow C^{(t-1)} \oplus s^*$;
22) **if** s^* 包含终止标记 **or** $\mathcal{I}_{active} = \emptyset$
23) **break**
24) **end if**
25) **End for**
26) **return** $C^{(t)}$

2.2 阶段一: 注意力引导的视觉多样性采样

为了在每一步推理中精准定位所需的视觉信息, 本阶段致力于从冗余的历史上下文中解耦出当前时刻的新增关注点, 从而动态确定模型所需的视觉输入区域.

2.2.1 相对注意力提取与空间映射

给定图像 I、问题 Q 和当前推理过程的上下文 $C^{(t-1)}$, 为了从视觉层面区分"题目要求的关注区域"与"当前已推理的关注区域", 首先构造问题提示 P_{query} 与当前推理过程提示 P_{curr}.

$$P_{query} = Q, \quad P_{curr} = C^{(t-1)} \tag{1}$$

这两组文本提示分别代表了问题的原始语义约束与当前的推理轨迹. 本文所使用的基座模型 (如 Qwen2.5-VL) 均采用统一的 Decoder-Only Transformer 架构, 视觉编码器提取的图像 Token 经线性投影后与文本 Token 拼接为同一序列, 在解码器中进行因果自注意力 (causal self-attention) 计算. 因此, 下文所提取的注意力图来源于自注意力权重中文本查询末位 Token 对所有图像 Token 的注意力

分布, 反映的是模型在生成决策时对图像不同区域的关注程度. 选择最后一层全头平均的设计动机在于: 1) 由于残差连接的逐层聚合效应, 最后一层的注意力分布天然反映了模型最终决策时刻对图像区域的聚焦, 与 LaMP "推理驱动感知"的目标最为一致; 2) 深层注意力更多编码高级语义与跨模态对齐关系, 相比浅层侧重局部纹理特征的注意力, 更适合捕捉"问题语义–图像区域"的关联; 3) 多头平均可有效缓解单头的随机噪声与偏特异性问题, 提供更稳健的空间分布估计. 相关对比实验详见参数敏感性分析部分. 为了将其转化为对应的空间分布, 将图像 I 分别与这两个提示输入 VLMs, 提取 Transformer 最后一层 H_{heads} 个注意力头的平均权重. 具体而言, 对于第 $k \in \{\text{query}, \text{curr}\}$ 种情况, 本文将 H_{heads} 个注意力头的视觉注意力权重进行平均, 得到聚合注意力图 $A_k \in \mathbf{R}^{N_{\text{img}}}$.

$$A_k = \frac{1}{H}\sum_{h=1}^{H}\mathcal{A}_k^{(h)} \tag{2}$$

其中, N_{img} 为图像 Token 总数; $\mathcal{A}_k^{(h)}$ 表示第 h 个注意力头从文本查询末位 Token 到所有图像 Token 的注意力分布; H 为注意力头总数. 由此得到的 A_{query} 和 A_{curr} 分别表示基于全局初始问题与当前推理步骤所提取的聚合注意力图.

为了精准捕捉那些"问题需要关注但当前尚未充分关注"的区域, 本文提出了相对注意力提取机制. 通过计算二者的比值, 获得相对注意力 A_{res}:

$$A_{\text{res}} = \frac{A_{\text{query}} + \epsilon}{A_{\text{curr}} + \epsilon} \tag{3}$$

其中, ϵ ($\epsilon = 10^{-8}$) 为数值稳定常数. 该操作利用 A_{curr} 作为分母, 不仅抑制了历史关注点, 更针对示意图"前景稀疏、背景主导"的视觉特性, 有效过滤了大量无信息的空白背景噪声, 同时保留 A_{query} 中未被覆盖的高响应区域. 这能够显著高亮显示下一步推理所需的新增视觉线索, 有效避免模型陷入重复关注的局部极值. 在基于示意图的问题求解中, 这一机制尤为关键. 例如, 在几何推理过程中, 相对注意力机制能够抑制当前求解过程的注意力占用, 引导模型主动将注意力聚焦到所需要的特定的几何元素 (如切线、角标) 上, 从而实现对微小视觉线索的精准捕捉.

为了将尺寸相对较小的注意力特征映射回原始图像空间并进行标准化度量, 首先对 A_{res} 进行双三次插值 (bicubic interpolation) 上采样得到 A_{up}, 随后应用空间归一化处理得到 A_{norm}.

$$\begin{cases} A_{\text{up}} = \text{BicubicUpsample}(A_{\text{res}},\ (h,\ w)) \\ A_{\text{norm}} = \dfrac{A_{\text{up}} - \min(A_{\text{up}})}{\max(A_{\text{up}}) - \min(A_{\text{up}}) + \delta} \end{cases} \tag{4}$$

其中, h 为图像的高; w 为图像的宽; δ ($\delta = 10^{-8}$) 为防止分母为零的微小量. 这一过程确保了注意力分布被规范化至区间 [0, 1], 为后续的跨尺度比较与区域积分提供了统一的度量基准. 基于 A_{norm}, 进一步采用基于多阈值分割与形态学处理的候选区域生成策略生成初始候选区域集合 $\mathcal{B} = \{B_i\}_{i=1}^{M}$. 具体流程如下: 1) 在注意力图上设置 $M = 10$ 个均匀分布的阈值 $\{\tau_i\}_{i=1}^{M}$, $\tau_i \in [0.5,\ 0.9]$; 2) 对每个阈值生成二值化图并应用 5×5 核的形态学闭运算; 3) 提取最大轮廓的最小外接矩形作为候选区域; 4) 过滤宽或高小于 $\rho_{\min} \times \min(w,\ h)$($\rho_{\min} = 0.3$) 的区域; 5) 若有效候选不足, 补充中心裁剪和全图区域. 关于初始候选数 M 和最小裁剪比例 $\rho_{\min}$ 的参数敏感性分析详见实验部分.

2.2.2 基于最大边界相关性的区域选择

在获得大量候选区域后, 为了从集合 $\mathcal{B}$ 中筛选出 K 个最具代表性的裁剪区域, 必须解决贪心选择可能带来的视觉信息冗余问题. 为此, 将区域选择建模为最大边界相关性优化问题.

定义两个候选区域 B_i、B_j 的空间互补度如下:

$$D(B_i,\ B_j) = 1 - \text{IoU}(B_i,\ B_j) \tag{5}$$

其中, IoU (intersection over union) 表示交并比. 采用迭代策略构建选择集合 $\mathcal{S}$ (初始为空). 在每一步迭代中, 依据以下目标函数从剩余集合中选出最优候选区域 B^*:

$$B^* = \arg\max_{B \in \mathcal{B} \setminus \mathcal{S}} \left[\lambda U(B) + (1-\lambda)\min_{B' \in \mathcal{S}} D(B,\ B')\right] \tag{6}$$

其中, $U(B)$ 为区域内的注意力积分; λ 设置为 0.7. 选出 B^* 后, 将其动态加入集合 $\mathcal{S}$ (即 $\mathcal{S} \leftarrow \mathcal{S} \cup \{B^*\}$), 并进入下一轮迭代直至满足数量要求. 该策略通过显式建模"信息质量"与"空间多样性"的权衡, 确保选出的视觉输入既能覆盖关键线索, 又能在空间上保持最大的互补性.

2.3 阶段二: 基于相对信息密度的自适应门控

经过阶段一的多样性采样, 模型已获得了一组在空间上互补的候选视觉区域. 然而, 并非所有的候选区域都包含对推理有价值的信息. 为了避免将计算资源浪费在低信息量的背景区域, LaMP 引入了预期增益门控 (expected gain gating) 机制, 核心在于仅当局部区域的信息密度显著高于全局

平均水平时, 才对其进行独立的细粒度推理. 定义区域 B_k 的预期信息增益 $E(B_k)$ 为其相对显著性密度比.

$$E(B_k)=\frac{\frac{1}{|B_k|}\sum\limits_{(x,\,y)\in B_k}A_{\rm norm}(x,\,y)}{\frac{1}{|I|}\sum\limits_{(x,\,y)\in I}A_{\rm norm}(x,\,y)} \tag{7}$$

该指标客观衡量了区域内的注意力集中程度. 从物理含义上看, $E(B_k)\geq 1$ 等价于"仅保留信息密度高于全图平均水平的区域", 是一个基于相对密度的无参数自适应准则, 而非需要人工调整的经验阈值. 该准则具有内在的跨数据集自适应性: 对于信息高度集中的示意图 (如几何图形), 注意力分布峰值更尖锐, 候选区域中仅少量能超过均值, 门控趋于严格; 对于信息较分散的图表 (如科学流程图), 注意力分布更均匀, 门控则相应放宽. 相关敏感性实验详见参数敏感性分析部分. 这一筛选机制高度契合示意图"信息密度分布极不均匀"的本质特性, 即关键语义往往高度浓缩于局部 (如函数拐点、电路元件), 而存在大量无效空白. 与自然图像不同, 教育场景下的视觉推理不需要全图扫描. 该门控机制类似于学生审题时的"考点筛选"过程: LaMP 能够自动忽略低密度的背景区域 ($E(B_k)<1$), 强制模型将计算资源集中在可能包含解题线索的"高价值区域". 基于此, 构建最终的活跃输入集 $\mathcal{I}_{\rm active}$, 仅保留那些优于平均增益的区域.

$$\mathcal{I}_{\rm local}=\{I_{\rm crop}(B_k)\mid B_k\in\mathcal{S},\ E(B_k)\geq\theta\} \tag{8}$$

$$\mathcal{I}_{\rm active}=\mathcal{I}_{\rm local}\cup\{I_{\rm global}\} \tag{9}$$

其中, $\mathcal{I}_{\rm local}$ 为局部视图集合; $I_{\rm crop}$ 为由原图裁剪得到的区域; $I_{\rm global}$ 为全局视图; θ 为信息增益门控阈值, 取值为 1.

值得注意的是, 全局视图始终被保留, 以维持推理过程中的整体语义连贯性. 至此, $\mathcal{I}_{\rm active}$ 包含了 1 个提供整体信息的全局视图与 K 个聚焦细粒度信息的局部视图, 对应 K 条不同的推理路径. 这种"宏观 + 微观"的互补组合, 使得模型能够从不同认知层面并行出发, 最终形成 $K+1$ 条侧重点各异的推理路径, 从而有效避免了单一视角的推理盲区.

2.4 阶段三: 多视角生成与联合质量评估

在获得具备多视角特征的活跃视觉输入集 $\mathcal{I}_{\rm active}$ 后, 模型进入前瞻性探索阶段. 对于 $\mathcal{I}_{\rm active}$ 中的每个视图, 模型并行生成候选推理步骤. 值得说明的是, 本文采用单视图独立推理策略, 即为每个视觉视图独立生成推理链, 而非将多图联合输入模型. 这是因为将原图与局部裁剪配对输入时, 模型自注意力需在显著增长的图像 Token 序列上进行分配, 导致局部关键细节的注意力权重被摊薄 (注意力稀释); 同时, 局部裁剪本身为原图子区域, 配对输入会造成同一视觉内容以不同分辨率重复出现的冗余干扰. 单视图独立推理使模型在每次前向传播中专注于单一视角, 随后通过融合评分在推理层面而非特征层面进行综合判断, 实验表明该策略性能优于多图联合输入 (详见实验部分分析). 为了全面评估这些生成路径的质量, 提出了结合逻辑自洽性与视觉忠实度的双维度评分指标.

2.4.1 路径置信度与视觉相关性建模

1) 路径置信度 (path confidence): 该指标用于衡量模型生成文本的内部逻辑确定性, 给定生成序列的 Token 对数概率, 计算其平均值.

$$S_{\rm conf}(s_i)=\frac{1}{L_s}\sum_{j=1}^{L_s}\log p(t_j|t_{<j},\ I,\ C^{(t-1)}) \tag{10}$$

其中, s_i 表示活跃输入集中独立生成的第 i 个候选推理样本, 在此过程中设定前瞻深度为 L, 即每个候选样本 s_i 包含前瞻生成的 L 个句子; L_s 表示该生成序列 s_i 的总 Token 长度; t_j 为当前预测的第 j 个 Token; $t_{<j}$ 则表示在预测 t_j 之前所有已生成的历史 Token 序列.

2) 视觉相关性 (visual relevance): 该指标用于验证生成的推理内容是否忠实于图像信息, 防止模型生成符合语言逻辑但与图像事实 (如物理方向、几何数值) 相悖的幻觉. 将生成的推理过程文本 r_i 作为新的提示信息, 提取其对应的视觉注意力图 $A_{\rm solution}$, 并计算其与基于问题 Q 得到的注意力图 $A_{\rm query}$ 的余弦相似度.

$$S_{\rm vis}(s_i)=\cos({\rm vec}(A_{\rm query}),\ {\rm vec}(A_{\rm solution})) \tag{11}$$

其中, ${\rm vec}(\cdot)$ 表示矩阵的向量化 (vectorization) 操作, 即把注意力图矩阵按列或按行展开成一个一维向量, 以便计算两个向量之间的余弦相似度.

2.4.2 基于层级验证的融合评分

最终评分 $S_{\rm final}$ 采用加权融合机制. 为了消除不同指标量纲带来的影响, 首先对两项指标进行归一化. 对于当前的候选样本集 $\{s_i\}_{i=1}^N$, 归一化计算如下:

$$\begin{aligned}\hat{S}_{\rm conf}^{(i)}&=\frac{S_{\rm conf}^{(i)}-\min\limits_j S_{\rm conf}^{(j)}}{\max\limits_j S_{\rm conf}^{(j)}-\min\limits_j S_{\rm conf}^{(j)}+\delta}\\ \hat{S}_{\rm vis}^{(i)}&=\frac{S_{\rm vis}^{(i)}-\min\limits_j S_{\rm vis}^{(j)}}{\max\limits_j S_{\rm vis}^{(j)}-\min\limits_j S_{\rm vis}^{(j)}+\delta}\end{aligned} \tag{12}$$

其中, j 的取值范围为 $j \in [1, N]$ (即 $1 \le j \le N$), 表示在当前共包含 N 个推理样本的候选集中进行全局的极大值与极小值搜索比对. 随后, 通过加权组合得到最终得分.

$$S_{\text{final}}^{(i)} = w_v \hat{S}_{\text{vis}}^{(i)} + w_p \hat{S}_{\text{conf}}^{(i)} \tag{13}$$

设定 w_p、w_v (例如 $w_p = 0.7$, $w_v = 0.3$), 这种双重校准机制直击教育推理中的核心痛点: 传统的 CoT 方法容易在长链推理中产生"逻辑自洽但视觉失实"的幻觉 (例如, 模型通过文本逻辑推导出一个公式, 但忽略了图像中隐含的定义域限制). 通过引入 S_{vis}, LaMP 实际上在每一步推理中都进行了一次检查, 模拟了人类在得出中间结论后"查阅图像以验证假设"的认知习惯, 确保生成的每一步数学推导都有确凿的图像证据支持.

2.5 阶段四: 推理链动态扩展与决策

基于上述多维评估, 算法进入决策更新阶段. 选择候选集综合得分最优的样本 s^*:

$$s^* = \arg \max_{s \in \mathcal{I}_{\text{active}}} S_{\text{final}}(s) \tag{14}$$

将该最优步骤产生的推理内容 r^* 动态追加至当前推理链上下文 $C^{(t)}$ 中. 这一过程循环迭代, 直至模型生成终止标记或活跃输入集为空, 从而构建出一条逻辑严密且视觉依据充分的完整推理链.

3 实验

3.1 数据集

为了全面评估模型面对不同类型示意图时的推理性能表现, 本文选取了多个代表性数据集进行系统测试:

- 数理推理 (math & physics reasoning). 该类数据集主要考察模型在数学和物理领域的视觉推理能力. 本文选用 MathVista[4]、MathVision[30] 和 PhysReason[31]. 这些任务要求模型不仅能精准理解视觉信息, 还需结合数学公理或物理定律进行多步逻辑推导.
- 多学科综合 (multi-discipline). 聚焦于跨学科知识融合及对各类示意图的理解能力. 数据集包括 AI2D[32]、SciQA-IMG[33]、MMStar[34] 和 M3CoT[35]. 此类任务侧重于评估模型在复杂综合场景下的推理泛化性与鲁棒性.

3.2 实验设置

基线模型与对比方法. 实验基于当前主流的视觉语言模型 (VLMs) 展开, 基座模型包括 Qwen2.5-VL-Instruct (7B、32B) 与 InternVL2.5-Instruct (8B). 基线 (Baseline) 设定为基座模型本身在不依赖任何推理增强技术的情况下, 采用标准自回归解码直接生成答案, 以此作为性能对比的基础. 为验证 LaMP 的有效性, 选取了四类代表不同技术路线的推理范式进行对比: 1) MCTS[12]: 代表基于蒙特卡洛树搜索算法的复杂逻辑推理增强策略; 2) Predictive Decoding (Pred. Dec.)[15]: 一种基于未来输出来引导当前生成, 从而增强全局一致性的前瞻性解码方法; 3) DyFo[36]: 专注于动态视觉搜索, 强调在推理开始之前完成注意力聚焦; 4) ICoT[37]: 利用交错图文上下文的多模态思维链提示方法.

评估指标与实施细节. 采用单次生成正确率 (即对每个问题仅生成一个候选答案, 该答案与标准答案一致的比例, 即大模型评测中的 Pass@1 指标) 作为所有基准测试中衡量准确率的主要性能指标, 记为准确率 (Acc.). 此外, 为了精确量化计算资源消耗与性能之间的权衡 (trade-off), 计算推理过程中的浮点运算次数 (floating point operations, FLOPs), 近似公式为 FLOPs $\approx 6nP$, 其中 P 为模型参数量, n 为生成 Token 的总数量. FLOPs 数值越低, 计算成本越低, 对应的部署效率越高. 为确保样本多样性并测试鲁棒性, 温度参数的取值范围为 0.6 至 1.0, 间隔为 0.1, 每次生成样本时均随机选择. 所有实验均在 4 张 NVIDIA A100 (80 GB) GPU 上并行进行. 对于 Predictive Decoding 和本文提出的 LaMP 方法, 当模型输出"REASONING_COMPLETE"标记时视为推理结束.

3.3 主要实验结果

如表 1 和表 2 所示, 在七个不同的数据集上, 基于三种不同规模的 VLMs, 将本文提出的 LaMP 方法与当前最先进的推理增强策略进行了全面对比, 表 1 中 SciQA 为 SciQA-IMG 的缩写, Avg. 为模型在不同任务上的平均准确率的缩写. 实验结果的具体分析如下:

综合性能的显著优势. LaMP 在所有测试的基座模型 (从 7B 到 32B) 及各类任务上的平均准确率均取得了约 4.2% 至 5.2% 的显著提升. 与基线模型及其他推理增强方法相比, LaMP 在平均准确率上始终保持领先. 值得注意的是, 即便是在参数量较大的 Qwen2.5-VL-32B-Instruct 模型上, LaMP 依然能进一步挖掘模型潜力, 超越了计算代价高昂的 MCTS 和 Predictive Decoding. 这证明了 LaMP 在增强视觉感知与逻辑推理方面的通用性, 且该优势不依赖于特定的模型架构.

视觉理解与逻辑推理的动态平衡. 多模态大模

表 1 不同方法在两大类视觉推理基准上的性能对比 (Acc. (%))
Table 1 Performance comparison of different methods on two major categories of visual reasoning benchmarks (Acc. (%))

模型方法	数理推理			多学科综合				
	MathVista	MathVision	PhysReason	AI2D	SciQA	MMStar	M3CoT	Avg.
Qwen2.5-VL-7B-Instruct								
Baseline	68.20	18.09	14.35	80.51	74.42	63.87	59.62	54.16
MCTS	69.60	18.75	16.27	81.87	78.09	66.20	60.87	55.95
Pred. Dec.	69.90	19.73	18.65	82.67	78.68	65.67	61.34	56.67
ICoT	47.50	10.53	15.73	81.35	77.24	42.07	61.17	47.95
DyFo	68.40	16.78	14.75	80.93	76.95	64.47	61.26	54.80
LaMP	70.80	24.34	21.76	84.07	81.95	69.80	62.68	59.34
InternVL2.5-8B-Instruct								
Baseline	64.40	22.00	11.58	81.61	76.15	60.47	57.16	53.34
MCTS	65.00	24.67	14.21	82.97	79.67	62.00	58.24	55.25
Pred. Dec.	65.40	25.00	17.16	83.74	80.37	62.40	58.76	56.12
ICoT	42.90	16.12	13.57	82.25	78.63	39.40	58.50	47.34
DyFo	64.70	20.39	12.08	81.96	77.89	61.47	58.58	53.87
LaMP	67.10	28.62	19.35	84.84	82.10	65.67	59.32	58.15
Qwen2.5-VL-32B-Instruct								
Baseline	74.70	25.33	19.42	83.29	78.83	69.47	62.81	59.13
MCTS	76.20	27.31	21.38	84.68	80.76	70.60	64.15	60.73
Pred. Dec.	76.60	27.96	22.76	85.17	81.95	71.07	64.62	61.45
ICoT	58.70	21.05	20.76	84.03	79.77	54.60	63.93	54.69
DyFo	75.60	24.67	20.35	83.52	78.93	70.07	64.32	59.64
LaMP	77.70	30.59	25.43	85.87	84.33	72.87	66.35	63.31

表 2 不同推理方法在各基座模型上的计算成本 (FLOPs) 对比
Table 2 Comparison of the computational cost (FLOPs) of different reasoning methods across different base models

基座模型	Baseline	MCTS	Pred. Dec.	ICoT	DyFo	LaMP
Qwen2.5-VL-7B	8.35e+12	4.05e+14	1.85e+14	1.88e+13	1.98e+13	1.54e+14
InternVL2.5-8B	1.02e+13	4.58e+14	2.09e+14	2.21e+13	2.38e+13	1.87e+14
Qwen2.5-VL-32B	3.59e+13	1.79e+15	8.17e+14	8.07e+13	8.63e+13	6.38e+14

型在处理复杂任务时, 必须在图像理解与文本推理之间取得平衡. 以 MCTS 和 Predictive Decoding 为代表的语言侧增强方法虽然在部分综合推理任务上带来了性能提升, 但在 MathVista 和 MathVision 等高度依赖视觉细节解析的数学几何任务中, 性能提升相对有限. 这表明, 单纯依靠文本进行多步推理而忽视视觉信息的实时验证并不十分可靠. 相反, 侧重于视觉搜索的方法 (如 DyFo) 虽然增强了局部感知, 但在需要长链条逻辑推演的任务中表现不尽如人意. LaMP 通过引入"前瞻性"机制, 在推理过程中动态交替进行视觉验证与路径规划, 有效解决了这一矛盾, 因而在各类任务中均取得了最佳平衡.

细粒度视觉细节的精准捕捉. LaMP 在处理包含复杂几何图形和科学图例的数据集时表现突出. 得益于相对注意力机制和信息增益门控, LaMP 能够主动聚焦于对当前推理步骤最具价值的局部区域, 而非被动接收全局模糊特征. 相较于 Baseline, LaMP 在这些任务上的性能增幅明显高于其他对比方法, 验证了其在细粒度视觉信息提取上的优越性.

良好的模型扩展性 (scaling properties). LaMP 的性能增益表现出良好的可扩展性. 随着基座模型参数量的增加, LaMP 带来的绝对性能提升依然稳健. 这表明该方法并非仅是对小模型能力的修补, 而是能够有效激发并利用更大规模基础模型潜在的强认知与推理能力. 随着基座模型的持续演

进, LaMP 的增益效果有望进一步放大.

计算效率与性能的权衡. 尽管 LaMP 相比原始 Baseline 引入了额外的计算开销, 但其效率显著优于同类复杂推理策略. 与 MCTS 相比, LaMP 通过"前瞻性"机制和信息增益门控设置大幅减少了无效路径的探索, 运算量 (FLOPs) 降低了约 62%, 同时在准确率上实现了反超. 这表明 LaMP 在计算资源消耗与推理性能提升之间取得了更优的 trade-off, 是一种高效且高性能的推理增强框架.

为更全面地反映实际部署效率, 进一步补充了 wall-clock runtime 对比实验. 如表 3 所示, 在统一硬件条件 (A100 80 GB) 下, LaMP 在单卡上的推理时间约为 68 s, 在 4 × A100 样本级并行配置下可降至约 31 s (见表 4), 不到 Predictive Decoding 的一半, 仅为 MCTS 的约 1/5. 若按"单位 FLOPs 倍率所带来的准确率回报"衡量, LaMP 相对 Baseline 的 FLOPs 倍率为 ~ 18.4×, 对应效率比为 +5.18%/18.4× ≈ 0.28%/×, 远优于 MCTS (+1.79%/48.5× ≈ 0.04%/×) 和 Predictive Decoding (+2.51%/22.2× ≈ 0.11%/×). 综合来看, LaMP 的延迟水平在云端教育平台等延迟容忍度为数十秒至数分钟的核心应用场景中完全可以接受, 具备良好的实际可部署性.

表 3　各推理方法的计算成本、推理时间与准确率综合对比 (Qwen2.5-VL-7B-Instruct, A100)

Table 3　Comprehensive comparison of computational cost, inference time and accuracy for different reasoning methods (Qwen2.5-VL-7B-Instruct, A100)

方法	FLOPs	时间开销 (s)	Avg. (%)	准确率增益
Baseline	8.35e+12	~ 3	54.16	—
ICoT	1.88e+13	~ 8	47.95	−6.21
DyFo	1.98e+13	~ 12	54.80	+0.64
LaMP	**1.54e+14**	**~ 68**	**59.34**	**+5.18**
Pred. Dec.	1.85e+14	~ 81	56.67	+2.51
MCTS	4.05e+14	~ 170	55.95	+1.79

表 4　LaMP 各阶段计算开销与时间分解 (Qwen2.5-VL-7B-Instruct, 4 样本/迭代 × 5 轮有效迭代)

Table 4　Computational cost and time breakdown for each stage of LaMP (Qwen2.5-VL-7B-Instruct, 4 samples/iteration × 5 effective iterations)

阶段	FLOPs	FLOPs 占比	单卡 (s)	四卡 (s)
阶段一: 视觉多样性采样	1.48e+13	9.6%	~ 7	~ 7
阶段二: 自适应门控	—	—	< 1	< 1
阶段三: 多视角生成与质量评估	1.39e+14	90.4%	~ 58	~ 21
阶段四: 推理链动态扩展与决策	—	—	< 1	< 1
I/O 与图像预处理	—	—	~ 2	~ 2
合计	**1.54e+14**	**100.0%**	**~ 68**	**~ 31**

此外, 本文对 LaMP 内部各阶段的时间开销进行了细粒度分解 (见表 4). 结果显示, 阶段三 (多视角生成与联合质量评估) 是绝对的计算瓶颈, 占据了 90.4% 的 FLOPs 和约 85.3% (58/68) 的实际耗时, 但该阶段天然支持多卡样本级并行, 同一迭代内的候选样本彼此完全独立, 不存在跨样本的数据依赖关系. 在 4 × A100 配置下, 阶段三的耗时从 ~ 58 s 降至 ~ 21 s (加速比 ~ 2.8×), 端到端推理时间从 ~ 68 s 降至 ~ 31 s (全局加速比 ~ 2.2×). 这一近线性的加速特性表明, 随着可用 GPU 数量的增加, LaMP 的端到端延迟有望进一步降低. 阶段二虽几乎不产生 GPU FLOPs, 但通过过滤低信息量的背景区域从源头减少了阶段三的视觉输入数量, 对维持整体计算效率起到了关键的调控作用.

3.4　消融实验

为了深入探究 LaMP 框架中各核心组件对最终性能的贡献, 本文在 MathVista 和 M3CoT 数据集上开展了多层次的消融实验. 首先通过对比 LaMP 完整模型与三种组件级变体的表现 (如图 3 所示) 验证各基础模块的有效性, 随后通过变体级组合消融与跨方法组件替换实验, 进一步剖析各阶段的独立贡献与协同效应.

3.4.1　组件级消融

视觉主动搜索的影响 (active search). 首先, 为了验证细粒度视觉信息的重要性, 移除了多阶段的视觉采样与门控机制 (即"w/o active search"), 仅保留全局视图 I_{global} 作为输入. 实验结果显示, 该变体在两个数据集上的性能均出现了明显的下滑. 这表明, 仅依赖全局图像的粗粒度特征难以支撑复杂

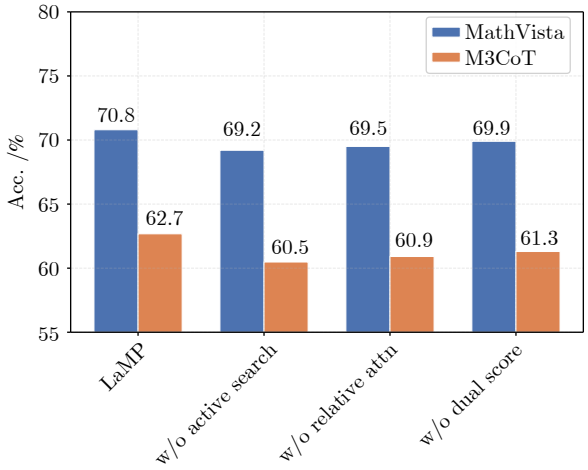


图 3　消融实验结果

Fig. 3　Results of ablation study

的示意图推理, 而本文提出的主动搜索机制能够有效补充关键的局部视觉线索.

相对注意力机制的贡献 (relative attention). 其次, 用标准的原始注意力图直接替换了相对注意力 A_{res} (即"w/o relative attn"). 结果表明, 模型性能受到了一定程度的抑制, MathVista 和 M3CoT 上的准确率均出现下降. 这证实了原始注意力由于受当前推理状态的影响, 往往容易过度关注已处理区域; 相比之下, 相对注意力机制通过抑制对已处理区域的关注, 成功引导模型聚焦于未探索的新增信息点, 从而优化了推理过程中的关注区域分布.

双维度评分策略的必要性 (dual score). 最后, 本文探究了联合评分函数 S_{final} 的合理性. 当移除视觉相关性评分而仅依靠路径置信度 S_{conf} (即"w/o dual score") 时, MathVista 和 M3CoT 上的性能均出现下降. 虽然该变体生成的逻辑链条语义通顺, 但缺乏视觉特征的强约束导致了部分与图像事实不符的幻觉现象. LaMP 通过融合视觉相关性与逻辑置信度, 在保证视觉忠实度的同时维持了逻辑的严密性, 从而在两个数据集上均取得了最佳性能.

3.4.2 变体级组合消融

为系统分析各阶段的独立贡献与协同关系, 将 LaMP 的四个阶段在功能上归纳为两大协同变体: 变体 A (主动视觉感知), 涵盖阶段一 (相对注意力引导的视觉采样) 与阶段二 (信息增益门控), 负责"看什么"即根据当前推理进展动态确定所需的视觉输入区域, 移除后模型仅使用全局视图 I_{global} 作为视觉输入; 变体 B (前瞻性推理评估), 对应阶段三 (多视角并行生成与双维度融合评分), 负责"怎么选", 即从候选推理路径中筛选最优步骤, 移除后模型仅生成单一路径并以置信度进行贪心选择, 阶段四 (推理链动态扩展与决策) 作为迭代框架始终保留. 这种功能性分组方式能清晰地区分视觉增强和推理增强各自的贡献边界. 基于 Qwen2.5-VL-7B-Instruct 的实验结果如表 5 所示.

实验结果表明: 1) 两条变体均具有显著且不可替代的贡献. 相较于两者均被移除的基础配置 (68.30/59.80), 单独启用变体 B 可带来 +1.50/+2.02 的提升, 单独启用变体 A 可带来 +1.10/+1.11 的提升, 二者联合后达到最优 (70.80/62.68). 其中变体 B 的独立贡献更为突出, 说明前瞻性推理评估对最终性能的驱动力更大, 这与主实验中语言侧增强方法整体优于纯视觉增强方法的趋势相吻合. 2) 两条变体的贡献高度互补, 功能冗余极小. 以 MathVista 为例, 变体 A 和变体 B 的独立贡献之和 (1.10% + 1.50% = 2.60%) 与两者联合使用的总增益 (2.50%) 近似相当, 二者之间几乎不存在重叠的性能增益来源. 每条变体在另一条已激活的条件下仍保持显著的边际贡献, 充分确认了两条变体均不可省略.

表 5 变体级组合消融实验结果 (Qwen2.5-VL-7B-Instruct, Acc. (%))

Table 5 Results of variant-level combinatorial ablation study (Qwen2.5-VL-7B-Instruct, Acc. (%))

变体 A	变体 B	MathVista	M3CoT
✓	✓	**70.80**	**62.68**
✓	×	69.40	60.91
×	✓	69.80	61.82
×	×	68.30	59.80

3.4.3 跨方法组件替换

为进一步验证 LaMP 各变体设计的必要性与不可替代性, 将 LaMP 的两条变体分别替换为对比方法中功能对应的组件进行交叉替换实验. 具体而言, 变体 A 可被替换为 DyFo 的静态区域选择策略, 变体 B 可被替换为 Predictive Decoding 的文本概率路径评估机制. 实验结果如表 6 所示.

实验结果揭示了以下关键发现: 1) DyFo + Predictive Decoding (69.00/60.96) 为所有跨方法组合中表现最弱的配置. DyFo 的区域选择在推理开始前即一次性确定, 无法随推理进展动态调整; Predictive Decoding 仅依赖文本概率空间评估路径质量, 两个组件之间缺乏信息反馈通路, 形成了"视觉静态 + 文本单维"的双重瓶颈. 2) DyFo + LaMP 变体 B (69.70/61.73) 为跨方法组合中表现最优的配置, 但与 Full LaMP 仍存在 1.10/0.95 的差距, 原因在于 DyFo 的静态裁剪区域与变体 B 的视觉相关性评分在空间上可能不对齐, 削弱了双维评分的协同效果. 3) LaMP 变体 A + Predictive Decoding (69.40/61.30) 的性能低于 DyFo + LaMP 变体 B, 进一步印证了路径评估机制相较于视觉输入增强对最终性能的贡献更为关键. 所有跨方法组合与 Full LaMP 之间的持续差距确认了完整框架中各变体设计的必要性与协同紧密性.

表 6 跨方法组件替换实验结果 (Qwen2.5-VL-7B-Instruct, Acc. (%))

Table 6 Results of cross-method component replacement experiments (Qwen2.5-VL-7B-Instruct, Acc. (%))

组合方式	MathVista	M3CoT
DyFo + Predictive Decoding	69.00	60.96
DyFo + LaMP 变体 B (前瞻评估)	69.70	61.73
LaMP 变体 A (视觉感知)+ Predictive Decoding	69.40	61.30
Full LaMP (变体 A + 变体 B)	**70.80**	**62.68**

3.5　参数敏感性分析

LaMP 的性能与计算成本受多个关键超参数的共同影响. 为系统地找到最优的性能–效率平衡点, 本节依次对前瞻深度 L 与推理路径数量 K、评分权重 w_p/w_v、候选区域生成参数 M 与 $\rho_{\min}$、信息增益门控阈值 θ, 以及注意力提取策略和多视图输入策略展开敏感性分析. 除特别说明外, 所有实验均基于 Qwen2.5-VL-7B-Instruct 模型.

3.5.1　前瞻深度 L 与推理路径数量 K

前瞻推理的最大深度 L 与每一步采样的候选区域数量 K 是影响 LaMP 推理质量与计算效率的两个最核心超参数. 基于 MathVista 和 M3CoT 数据集对二者分别进行了控制变量实验, 在研究 L 参数时, K 固定为 3; 在研究 K 参数时, L 保持为 5. 结果如表 7 所示.

前瞻深度 L 的影响. 固定采样数 $K=3$, 将推理步数 L 从 3 增加至 7. 观察发现, 随着步数的增加, 模型性能呈现先上升后趋于饱和的趋势. 当 $L=3$ 时, 由于推理链过短, 模型往往难以完成多跳逻辑推演; 当 L 增加至 5 时, 模型在 MathVista 上的准确率达到峰值. 然而, 继续增加步数至 $L=7$ 并未带来显著的性能提升, 反而导致 FLOPs 呈线性增长, 且过长的推理链引入了误差累积的风险.

推理路径数量 K 的权衡. 固定 $L=5$, 研究推理路径数量 K (从 1 到 5) 的影响. 结果显示, 当 K 较小 (如 $K=1$) 时, 模型可能遗漏关键的视觉线索; 随着 K 增加至 3, 性能显著提升. 但当 K 进一步增加至 5 时, 性能增益边际递减, 同时计算开销 (FLOPs) 大幅增加, 且过多的背景区域可能引入无关噪声干扰推理.

综合考虑计算效率与推理性能, 在最终的实验设置中选取了 $L=5$ 和 $K=3$ 作为默认配置, 以在可接受的计算成本下实现最优的推理准确率.

3.5.2　评分权重 w_p/w_v

为探讨式 (13) 中 w_p 与 w_v 的分配比例对不同学科任务的影响, 固定 $w_p+w_v=1.0$, 将 w_v 从 0.1 变化至 0.5 (步长 0.1), 在三个代表不同学科特点的数据集上进行了参数敏感性分析. 该实验旨在揭示逻辑置信度与视觉校验在不同类型任务中的最优配比关系, 实验结果如表 8 所示.

实验结果表明: 1) $w_v=0.3$、$w_p=0.7$ 在所有三个数据集上均取得了最佳或接近最佳的性能, 该配置具有良好的跨学科稳定性. 这一现象的合理解释是: 无论何种学科, 教育示意图推理的核心仍然是逻辑推导能力 (因此 w_p 应占主导), 而视觉校验主要起“纠偏”与“锚定”作用, 用于约束推理路径不偏离图像事实 (因此 w_v 不宜过高). 2) 当 w_v 较大时, 所有数据集的性能均开始下降, 这是因为过度依赖视觉相似度可能导致模型倾向于选择视觉“保守”的推理路径, 从而抑制了深层逻辑推演的探索.

3.5.3　候选区域生成参数 M 与 $\rho_{\min}$

为验证候选区域生成参数设置的合理性, 在 MathVista 上对初始候选数 M 和最小裁剪比例 $\rho_{\min}$ 进行了控制变量实验, 结果如表 9 所示. 对于初始候选数 M, 固定最小裁剪比例 $\rho_{\min}=0.3$, 当 M 从 3 增至 10 时准确率稳步提升 (69.3% → 70.8%), 更多阈值级别能够捕获不同粒度的注意力聚焦区域; 当 M 继续增至 15 和 20 时性能趋于饱和, 过密的阈值采样产生了大量高度重叠的候选区域, 信息增益边际递减. 当固定初始候选数 M 为 10, 对于最小裁剪比例 $\rho_{\min}$, 过小的值 ($\rho_{\min}=0.1$) 导致候选区域仅覆盖极小的图像碎片, 丢失了必要的空间上下文信息; 过大的值 ($\rho_{\min}=0.5$) 则使方法难以聚焦于关键局部细节. 综合来看, $M=10$、$\rho_{\min}=0.3$ 在性能与效率间取得了最佳平衡, 且方法对两个参数的敏感度均较为温和, 准确率波动不超过 1.5 个百分点, 表明 LaMP 在参数选择上具有良好的鲁棒性.

3.5.4　信息增益门控阈值 θ

为验证门控阈值 $E(B_k)\geq\theta$ 中 $\theta=1$ 的合理性, 系统对比了不同阈值对性能与效率的影响, 实验中固定 $K=3$, 结果如表 10 所示. 实验结果揭示了清

表 7　前瞻深度 L 和推理路径数量 K 对实验结果的影响分析

Table 7　Analysis of the impact of lookahead depth L and reasoning path number K on the experimental results

前瞻步数	MathVista Acc. (%)	M3CoT Acc. (%)	FLOPs	采样数	MathVista Acc. (%)	M3CoT Acc. (%)	FLOPs
$L=3$	69.40	60.65	9.50e+13	$K=1$	68.90	60.48	8.70e+13
$L=4$	70.20	61.91	1.25e+14	$K=2$	70.10	61.73	1.22e+14
$L=5$	70.80	62.68	1.54e+14	$K=3$	70.80	62.68	1.54e+14
$L=6$	71.00	63.07	1.79e+14	$K=4$	71.20	63.55	1.83e+14
$L=7$	71.10	63.42	2.02e+14	$K=5$	71.50	64.06	2.10e+14

表 8 逻辑置信度 w_p 与视觉相关性 w_v 在不同数据集上的敏感性分析 (Qwen2.5-VL-7B-Instruct, Acc. (%))

Table 8 Sensitivity analysis of logical confidence w_p and visual relevance w_v on different datasets (Qwen2.5-VL-7B-Instruct, Acc. (%))

w_v	w_p	MathVista (数学)	PhysReason (物理)	SciQA-IMG (多学科)
0.1	0.9	70.20	20.21	80.48
0.2	0.8	70.50	21.05	81.23
0.3	**0.7**	**70.80**	**21.76**	**81.95**
0.4	0.6	70.60	21.28	81.61
0.5	0.5	70.10	20.47	80.83

表 9 候选区域生成核心参数敏感性分析 (Qwen2.5-VL-7B-Instruct, MathVista) (%)

Table 9 Sensitivity analysis of core parameters for candidate region generation (Qwen2.5-VL-7B-Instruct, MathVista) (%)

候选数 M	3	5	**10**	15	20
Acc.	69.3	69.9	**70.8**	70.6	70.3
裁剪比例 $\rho_{\min}$	0.1	0.2	**0.3**	0.4	0.5
Acc.	69.5	70.2	**70.8**	70.1	69.4

表 10 信息增益门控阈值 θ 的敏感性分析 (Qwen2.5-VL-7B-Instruct, $K = 3$)

Table 10 Sensitivity analysis of information gain gating threshold θ (Qwen2.5-VL-7B-Instruct, $K = 3$)

θ	MathVista Acc. (%)	M3CoT Acc. (%)	FLOPs
0.6	70.10	61.95	1.76e+14
0.8	70.50	62.34	1.64e+14
1.0	**70.80**	**62.68**	**1.54e+14**
1.2	70.30	62.12	1.38e+14
1.4	69.50	61.52	1.15e+14

晰的“倒 U 型”趋势: 当 θ 过低 (如 0.6) 时, 候选区域几乎全部通过门控, 门控机制实质上失效, 低信息量的背景区域被纳入推理流程, 不仅增加了约 14% 的计算开销, 引入的视觉噪声还干扰了推理路径的选择, 导致准确率下降; 当 θ 过高 (如 1.4) 时, 候选中平均不足 1 个能通过筛选, 模型几乎退化为仅依赖全局视图 I_{global} 的单视角推理, 丧失了多视角互补的核心优势. $\theta = 1.0$ 在准确率与计算效率之间取得了最优平衡, 且该结论在两个不同类型的数据集上保持一致, 进一步说明该阈值具有良好的泛化稳定性.

3.5.5 注意力提取策略

本文系统对比了六种注意力提取策略, 结果如表 11 所示. 实验结果表明: 1) 性能随提取层级加深而持续提升 (前 1/3 → 中间 1/3 → 后 1/3), 验证了深层注意力在跨模态语义对齐上的优势, 这与 Transformer 模型中“浅层捕捉低级特征、深层建模高级语义”的一般规律一致; 2) 全体层平均 (70.00%/61.82%) 反而不及后 1/3 层平均 (70.30%/62.17%), 说明浅层注意力中的低级纹理响应对该任务构成了干扰, 简单的全层聚合会稀释深层的高级语义信号; 3) 最后一层 Top-4 头 (70.50%/62.34%) 略逊于最后一层全头平均 (70.80%/62.68%), 说明各注意力头承载了互补的空间信息, 最后一层全头平均是更稳健的聚合策略. 最后一层全头平均在两个数据集上均取得最佳性能, 验证了当前设计的合理性.

表 11 不同注意力提取策略的性能对比 (Qwen2.5-VL-7B, Acc. (%))

Table 11 Performance comparison of different attention extraction strategies (Qwen2.5-VL-7B, Acc. (%))

注意力提取策略	MathVista	M3CoT
前 1/3 层平均	69.30	61.04
中间 1/3 层平均	69.80	61.65
后 1/3 层平均	70.30	62.17
全体层平均	70.00	61.82
最后一层 Top-4 头	70.50	62.34
最后一层全头平均	**70.80**	**62.68**

3.5.6 多视图输入策略

为评估多图联合输入与单视图独立推理策略的效果差异, 将二者进行了系统对比, 结果如表 12 所示. 将原图与局部裁剪配对输入在两个数据集上均出现明显的性能下降. 原因主要有两方面: 1) 注意力稀释, 原始图像包含大量与当前局部区域无关的视觉内容, 局部区域中的关键细节在全局 Token 的“竞争”下获得的注意力权重被摊薄, 模型难以精准定位推理所需的信息; 2) 视觉冗余干扰, 局部裁剪本身即为原图的子区域, 配对输入导致同一视觉内容以不同分辨率重复出现, 增加了模型对齐两个视图中相同语义元素的负担, 反而引入了额外的认知冗余. LaMP 采用的单视图独立推理策略通过“先

表 12 多视图输入策略对比 (Qwen2.5-VL-7B-Instruct, $K = 3$, $\theta = 1.0$, Acc. (%))

Table 12 Comparison of multi-view input strategies (Qwen2.5-VL-7B-Instruct, $K = 3$, $\theta = 1.0$, Acc. (%))

策略	MathVista	M3CoT
多图联合输入	69.70	61.73
单视图独立推理	**70.80**	**62.68**

独立思考、再综合判断”的范式有效规避了上述问题, 取得了更优的推理性能.

3.6　案例分析

如图 4 所示, 在案例 1) 几何问题求解中, LaMP 充分发挥了多视角探索的优势. 面对模糊的标注, 模型并未盲从单一的全局直觉 (即误判斜边为 10, 如图 4 中红色分支), 而是并行生成了基于不同局部关注点的假设路径. 这种“多视角竞争 + 局部验证”机制有效规避了因粗略浏览导致的视觉误判, 进而推导出正确的几何方程. 在案例 2) 电路分析题中, LaMP 成功避免了繁琐的回路方程计算. 通过聚焦于中心电流表的局部视角, 验证了电桥平衡的关键条件, 从而利用简单的对角线乘积迅速推导出正确结果.

为了进一步反映模型现阶段存在的局限性, 图 5 展示了 LaMP 在极端困难场景下的三类典型失败案例. 第一类是模型本身存在的领域知识幻觉问题: 模型在处理特定领域问题时, 可能产生幻觉并错误引入外部假设 (如错误类型一中的“生态补偿”), 导致推理链偏离正确方向, 最终得出错误结论. 这一现象暴露了当前 VLMs 在专业领域知识储备上的不足, 仅靠推理策略的改进无法从根本上弥补基座模型的知识缺陷. 第二类是由示意图中存在的图文矛盾引发的层级验证融合评分失效问题: 如错误类型二所示, 尽管模型具备识别视觉特征的能力 (候选路径 1 正确识别了相交弦), 但由于其结论与具有误导性的文本提示相悖, 导致置信度被压低; 基于错误文本提示生成的路径 (候选路径 2) 反而获得较高分数. 这说明 LaMP 的双维评分机制在遭遇图文信息矛盾时, 可能因过度依赖高置信度的文本对齐而发生失效, 致使模型陷入“文本误导”陷阱. 第三类是复杂条件下的逻辑推理错误: 如错误类型三中的谜题所示, 问题要求基于几何约束与交叉点数量进行抽象推断. 尽管模型在初步思考阶段试图进行严谨的数学推演 (计算组合交点数), 但面对复杂的图文映射时, 模型在长链逻辑推理过程中发生断裂, 退化为依赖视觉直觉的猜测 (即寻找“最明显的 S 型曲线”). 这表明模型在处理需要深度几何拓扑与抽象逻辑相耦合的问题时, 推理稳定性仍有待提升.

这些案例表明, LaMP 具备处理多种教育领域复杂示意图问题的鲁棒性, 能够通过视觉注意力的动态调整与逻辑验证的交互, 引导模型构建出精准且高效的推理路径. 这种引导模型构建可解释、高质量推理路径的能力, 不仅证明了其在解决真实世界视觉任务中的鲁棒性, 也为其在个性化教育辅导等领域的落地应用奠定了基础. 同时, 对失败案例的分析也为后续研究指明了改进方向, 即需要引入更强健的图文冲突检测机制, 并进一步强化模型在复杂长链逻辑推理中的抗干扰能力.

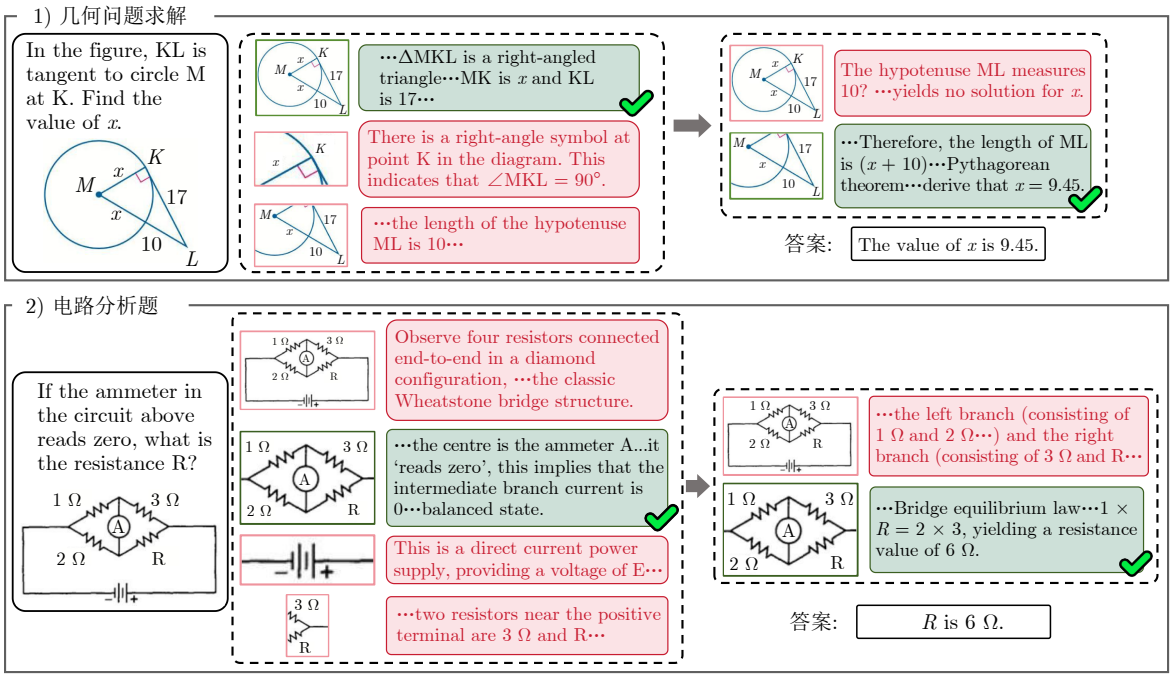


图 4　LaMP 方法在示意图问题求解中的多视角探索与局部验证案例分析

Fig. 4　A case analysis on the LaMP method for multi-perspective exploration and local verification in diagram problem-solving

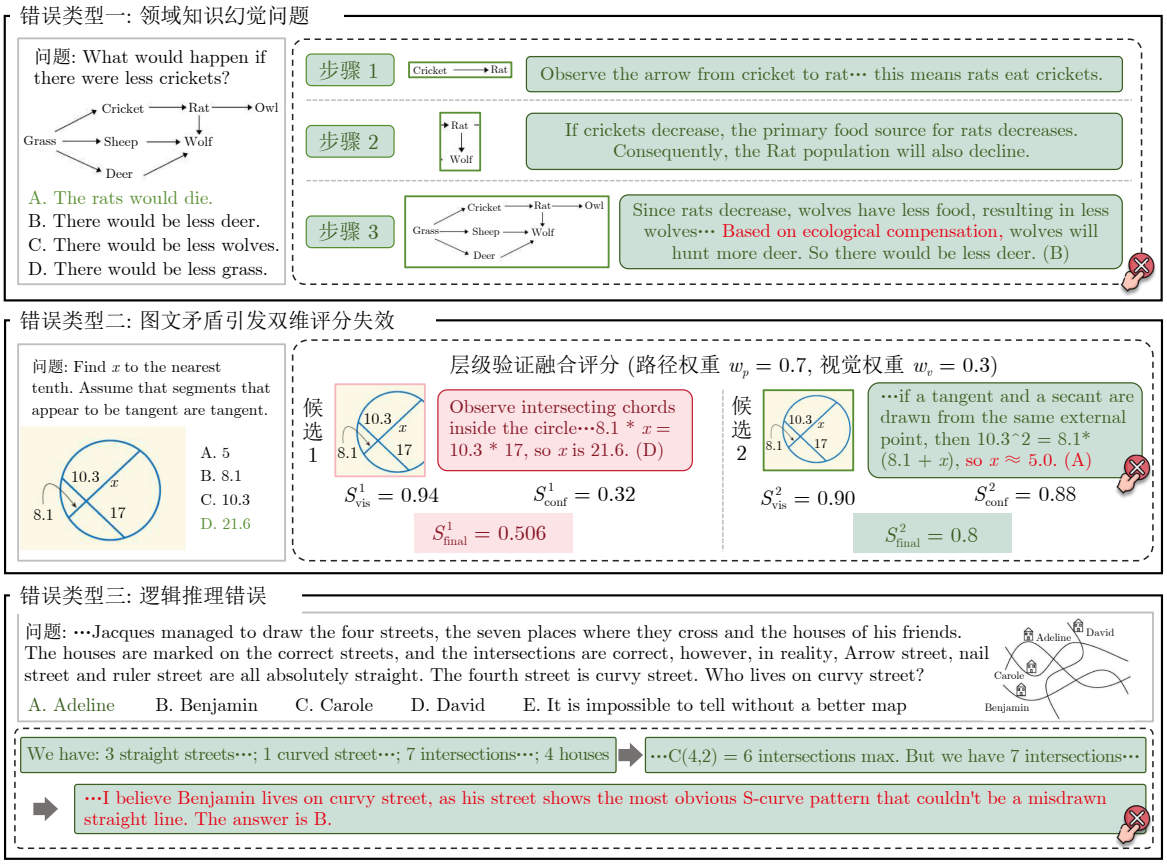


图 5 LaMP 方法在示意图问题求解中的局限性分析

Fig. 5 Analysis of limitations of the LaMP method in diagram problem-solving

4 结束语

本文针对多模态大模型在示意图问答中存在的“感知–推理协同鸿沟”, 提出了一种基于前瞻性机制的多视角视觉推理框架 (LaMP). 受启发于人类“慢思考 (System 2)”的认知机理, 本文通过引入相对注意力机制与信息增益门控, 克服了示意图背景噪声干扰, 实现了从被动全局编码到主动细粒度感知的范式转变. LaMP 构建了动态的“观察–假设–验证”闭环系统, 利用视觉相关性与逻辑置信度进行双重校准, 有效抑制了长链推理中的逻辑幻觉. 在 MathVista 与 M3CoT 等多个权威基准上的实验表明, LaMP 不仅在准确率上显著优于 MCTS 等现有主流方法, 更在计算开销与推理性能之间取得了极佳的平衡. 未来, 也将进一步探索该机制在更复杂的开放世界动态视觉任务中的泛化能力.

参考文献

1 Miao Qing-Hai, Wang Xing-Xia, Yang Jing, Zhao Yong, Wang Yu-Tong, Chen Yuan-Yuan, et al. From foundation intelligence to general intelligence: The state-of-art and perspectives of GenAI and AGI based on foundation models. *Acta Automatica Sinica*, 2024, **50**(4): 674–687 (缪青海, 王兴霞, 杨静, 赵勇, 王雨桐, 陈圆圆, 等. 从基础智能到通用智能: 基于大模型的 GenAI 和 AGI 之现状与展望. 自动化学报, 2024, **50**(4): 674–687)

2 Liu H T, Li C Y, Wu Q Y, Lee Y J. Visual instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). Vancouver, BC, Canada: Curran Associates, 2024. 34892–34916

3 Cui C, Ma Y S, Cao X, Ye W Q, Zhou Y, Liang K Z, et al. A survey on multimodal large language models for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024. 958–979

4 Lu P, Bansal H, Xia T, Liu J C, Li C Y, Hajishirzi H, et al. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In: Proceedings of the International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. 23439–23554

5 Yue X, Ni Y S, Zhang K, Zheng T Y, Liu R Q, Zhang G, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024. 9556–9567

6 Guan T R, Liu F X, Wu X Y, Xian R Q, Li Z X, Liu X Y, et al. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024. 14375–14385

7 Zhang X Y, Zhang L L, Wu Y R, Wang S W, Wu W J, Huang M Y, et al. Memory-enriched thought-by-thought framework for

complex diagram question answering. *Computer Vision and Image Understanding*, 2025: Article No. 104608

8 Wang S W, Zhang L L, Zhu L J, Qin T, Yap K H, Zhang X Y, et al. CoG-DQA: Chain-of-guiding learning with large language models for diagram question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024. 13969−13979.

9 Wang L, Hu Y, He J B, Xu X, Liu N, Liu H, et al. T-SciQ: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Vancouver, BC, Canada: AAAI Press, 2024. 19162−19170

10 Li Y F, Du Y F, Zhou K, Wang J P, Zhao W X, Wen J R. Evaluating object hallucination in large vision-language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore: Association for Computational Linguistics, 2023. 249−260

11 Yin S K, Fu C Y, Zhao S R, Xu T, Wang H, Sui D H, et al. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 2024, **67**(12): Article No. 220105

12 Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). New Orleans, LA, USA: Curran Associates, 2022. 24824−24837

13 Yao S Y, Yu D, Zhao J, Shafran I, Griffiths T, Cao Y, et al. Tree of thoughts: Deliberate problem solving with large language models. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). New Orleans, LA, USA: Curran Associates, 2023. 11809−11822

14 Hao S B, Gu Y, Ma H D, Hong J, Wang Z, Wang D, et al. Reasoning with language model is planning with world model. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore: Association for Computational Linguistics, 2023. 8154−8173

15 Ma C, Zhao H T, Zhang J L, He J X, Kong L P. Non-myopic generation of language models for reasoning and planning. In: Proceedings of the International Conference on Learning Representations (ICLR). Singapore: OpenReview.net, 2025. 35031−35058.

16 Wang X Z, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. In: Proceedings of the International Conference on Learning Representations (ICLR). Kigali, Rwanda: OpenReview.net, 2023.

17 Kahneman, D. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

18 Lu Jing-Wei, Guo Chao, Dai Xing-Yuan, Miao Qing-Hai, Wang Xing-Xia, Yang Jing, et al. The ChatGPT after: Opportunities and challenges of very large scale pre-trained models. *Acta Automatica Sinica*, 2023, **49**(4): 705−717
(卢经纬, 郭超, 戴星原, 缪青海, 王兴霞, 杨静, 等. 问答 ChatGPT 之后: 超大预训练模型的机遇和挑战. 自动化学报, 2023, **49**(4): 705−717)

19 Zhu Jia-Jun, Bao Mei-Kai, Zhang Kai, Liu Ye, Liu Qi. A commonsense question answering method based on multi-source knowledge infusion. *Computer Engineering and Science*, 2025, **47**(2): 349−360
(朱嘉骏, 包美凯, 张凯, 刘烨, 刘淇. 基于多源知识注入的常识问答方法研究. 计算机工程与科学, 2025, **47**(2): 349−360)

20 Shinn N, Cassano F, Gopinath A, Narasimhan K, Yao S Y. Reflexion: Language agents with verbal reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). New Orleans, LA, USA: Curran Associates, 2023. 8634−8652

21 Lightman H, Kosaraju V, Burda Y, Edwards H, Baker B, Lee T, et al. Let′s verify step by step. In: Proceedings of the International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. 39578−39601.

22 Wang Quan-Zi-Ang, Wang Ren-Zhen, Meng De-Yu, Xu Zong-Ben. The evolution of continual learning methodologies in the era of large models. *Acta Automatica Sinica*, 2025, **52**(8): 1493−1519
(王全子昂, 王仁振, 孟德宇, 徐宗本. 面向大模型时代的持续学习方法论演变. 自动化学报, 2025, **52**(8): 1493−1519)

23 Ha D, Schmidhuber J. Recurrent world models facilitate policy evolution. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). Montréal, QC, Canada: Curran Associates, 2018. 2455−2467

24 Yao S Y, Zhao J, Yu D, Du N, Shafran I, Narasimhan K R, et al. ReAct: Synergizing reasoning and acting in language models. In: Proceedings of the International Conference on Learning Representations (ICLR). Kigali, Rwanda: OpenReview.net, 2023.

25 Zheng Yi-Ning, Yu Zhen, Li Bu-Fan, Yang Jie, Yin Lin-Qi, Yin Zhang-Yue, et al. Survey of tool use in large language models. *Acta Automatica Sinica*, 2025, **51**(11): 2371−2386
(郑逸宁, 余镇, 李不凡, 杨捷, 殷林琪, 印张悦, 等. 大语言模型的工具使用综述. 自动化学报, 2025, **51**(11): 2371−2386)

26 Hu Y S, Shi W J, Fu X Y, Roth D, Ostendorf M, Zettlemoyer L, et al. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). Vancouver, BC, Canada: Curran Associates, 2024, 139348−139379

27 Qi J, Ding M, Wang W H, Bai Y S, Lv Q S, Hong W Y, et al. CogCoM: A visual language model with chain-of-manipulations reasoning. In: Proceedings of the International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview. net, 2024. 11090−11110.

28 Zhang Hui, Liang Shu-Tong, Li Ming-Xuan, Tian Yong-Lin, Ge Jing-Wei, Yu Hui, et al. Vision-language-action models: From the early foundations to the state-of-the-art. *Acta Automatica Sinica*, 2025, **51**(9): 1922−1950
(张慧, 梁姝彤, 李明轩, 田永林, 葛经纬, 于慧, 等. 视觉–语言–动作模型综述: 从前史到前沿. 自动化学报, 2025, **51**(9): 1922−1950)

29 Dai W L, Li J N, Li D X, Tiong A, Zhao J Q, Wang W S, et al. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). New Orleans, LA, USA: Curran Associates, 2023. 49250−49267.

30 Wang K, Pan J T, Shi W K, Lu Z M, Ren H X, Zhou A J, et al. Measuring multimodal mathematical reasoning with MathVision dataset. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). Vancouver, BC, Canada: Curran Associates, 2024. 95095−95169

31 Zhang X Y, Dong Y X, Wu Y R, Huang J X, Jia C Y, Fernando B, et al. PhysReason: A comprehensive benchmark towards physics-based reasoning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). Vienna, Austria: Association for Computational Linguistics, 2025. 16593−16615.

32 Kembhavi A, Salvato M, Kolve E, Seo M, Hajishirzi H, Farhadi A. A diagram is worth a dozen images. In: Proceedings of the European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: Springer, 2016. 235−251

33 Lu P, Mishra S, Xia T L, Qiu L, Chang K W, Zhu S C, et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). New Orleans, LA, USA: Curran Associates, 2022. 2707−2721

34 Chen L, Li J S, Dong X Y, Zhang P, Zang Y H, Chen Z H, et al. Are we on the right way for evaluating large vision-language

models? In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). Vancouver, BC, Canada: Curran Associates, 2024. 27056–27087

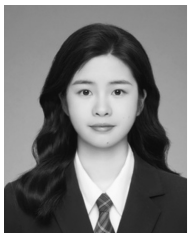
35 Chen Q G, Qin L B, Zhang J, Chen Z, Xu X, Che W X. M3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). Bangkok, Thailand: Association for Computational Linguistics, 2024. 8199–8221

36 Li G, Xu J L, Zhao Y Z, Peng Y X. DyFo: A training-free dynamic focus visual search for enhancing LMMs in fine-grained visual understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2025. 9098–9108

37 Gao J, Li Y Q, Cao Z Q, Li W J. Interleaved-modal chain-of-thought. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2025. 19520–19529

张新宇 西安交通大学计算机科学与技术学院博士研究生. 主要研究方向为视觉推理, 视觉语言模型. E-mail: zhang1393869716@stu.xjtu.edu.cn

(**ZHANG Xin-Yu** Ph.D. candidate at the School of Computer Science and Technology, Xi′an Jiaotong University. His research interests include visual reasoning and vision-language models.)



吴艳瑞 西安交通大学计算机科学与技术学院硕士研究生. 主要研究方向为示意图理解和逻辑推理.
E-mail: yanrui.wu@stu.xjtu.edu.cn

(**WU Yan-Rui** Master student at the School of Computer Science and Technology, Xi′an Jiaotong University. Her research interests include diagram understanding and logical reasoning.)

张玲玲 西安交通大学计算机科学与技术学院副教授. 主要研究方向为智慧教育, 知识图谱和大模型推理. 本文通信作者.
E-mail: zhanglling@xjtu.edu.cn

(**ZHANG Ling-Ling** Associate professor at the School of Computer Science and Technology, Xi′an Jiaotong University. Her research interests include smart education, knowledge graphs, and LLM reasoning. Corresponding author of this paper.)

杨泽晟 西安交通大学计算机科学与技术学院硕士研究生. 主要研究方向为自然语言处理.
E-mail: 2204215061@stu.xjtu.edu.cn

(**YANG Ze-Sheng** Master student at the School of Computer Science and Technology, Xi′an Jiaotong University. His main research interest is natural language processing.)

董宇轩 西安交通大学计算机科学与技术学院硕士研究生. 主要研究方向为多模态大模型和强化学习. E-mail: yuxuandong@stu.xjtu.edu.cn

(**DONG Yu-Xuan** Master student at the School of Computer Science and Technology, Xi′an Jiaotong University. His research interests include multimodal large models and reinforcement learning.)

武亚强 联想人工智能技术中心高级总监, 正高级工程师. 主要研究方向为企业级人工智能包括评估、推理与编排.
E-mail: wuyqe@lenovo.com

(**WU Ya-Qiang** Senior director and professor-level senior engineer at the Lenovo AI Technology Center. His research interests include enterprise artificial intelligence, covering evaluation, reasoning, and orchestration.)

郑庆华 中国工程院院士, 同济大学教授. 主要研究方向为大数据知识工程.
E-mail: qhzheng@mail.xjtu.edu.cn

(**ZHENG Qing-Hua** Academician of the Chinese Academy of Engineering, and professor of Tongji University. His main research interest is big data knowledge engineering.)