河北科技大学

研究生《计算机视觉基础》课程结课论文

学年学期：2025—2026 学年第一学期

学生姓名：张宽　学号：2025114041

学院：信息科学与工程学院

专业：计算机科学与技术

题目：基于闭环验证的可解释视觉问答系统设计与实现

2026 年 1 月

基于闭环验证的可解释视觉问答系统设计与实现

学号：2025114041　　姓名：张宽

摘要：

近年来，多模态大模型在视觉问答（VQA）任务中取得显著进展，但模型幻觉与推理过程不透明仍是制约其可信部署的核心挑战。本文提出一种基于闭环验证的可解释视觉问答系统，通过融合视觉语言模型 Qwen2-VL、可提示分割模型 SAM2 与跨模态对齐模型 CLIP，构建“定位–分割–验证–精炼”的迭代推理框架。系统首先由 Qwen2-VL 生成初步答案及证据区域，再利用 SAM2 提取高保真视觉证据，并通过 CLIP 量化答案与证据的一致性，动态决定是否触发精炼机制。实验表明，该闭环架构相较单次推理将准确率提升 6.36%，有效抑制幻觉。本文还从定位精度、证据质量与验证有效性三方面开展深入分析，为构建可信赖多模态 AI 提供新范式。

关键字：视觉问答；闭环验证；可解释 AI；多模态大模型；模型幻觉

正文部分：

1 引言

视觉问答（Visual Question Answering, VQA）作为计算机视觉与自然语言处理的交叉任务，要求模型基于图像内容理解并回答语义复杂的问题。尽管以 Qwen-VL、LLaVA 等为代表的多模态大模型在性能上取得突破 [1–3]，其决策过程仍常因缺乏可验证依据而产生幻觉或错误推理 [4]。现有方法或依赖注意力热力图进行事后解释 [5]，或引入外部知识库辅助验证 [6]，但前者难以保证证据可靠性，后者则易与视觉内容脱节。

为弥合这一鸿沟，近期研究开始探索结构化、可干预的推理路径。例如，Evidence-aware VQA [7] 通过显式证据标注提升可信度，但依赖人工监督；Chain-of-Thought prompting [8] 虽能生成推理链，却缺乏对中间步骤的客观校验。受此启发，本文提出一种自动化闭环验证机制，将 VQA 分解为可验证、可回溯的四个阶段，实现无需人工干预的自我校正。本文主要贡献包括：(1) 设计并实现基于 Qwen2-VL、SAM2 与 CLIP 的闭环 VQA 系统；(2) 提出“定位–分割–验证–精炼”迭代框架，显著提升答案可靠性；(3) 构建包含 110 个复杂样本的自建数据集 MyVQA；(4) 从多维度系统评估系统性能与失败模式。

2 相关工作

2．1 视觉语言模型

视觉语言模型（VLMs）已从早期双塔架构（如 CLIP[9]）演进为端到端联合建模。BLIP 系列 [10] 通过噪声鲁棒预训练提升生成质量，而 Qwen-VL [2] 与 LLaVA [11] 则利用大语言模型增强多步推理能力。本文采用 Qwen2-VL-7B-Instruct [8]，其支持细粒度定位与高分辨率输入，适合作为闭环系统的推理引擎。

2．2 可提示分割模型

传统分割依赖固定类别，而 Segment Anything Model（SAM）[12] 首次实现开放世界零样本分割。其升级版 SAM2 [10] 进一步优化边界精度与推理效率，特别适用于从粗略定位框中提取精确视觉证据，为后续验证提供可靠输入。

2．3 可验证视觉推理

可解释 VQA 的核心在于建立答案与视觉证据的可信关联。VALSE [13] 提出结构化证据评估基准，但依赖人工标注；Self-Consistency VQA [7] 利用多路径投票提升鲁棒性，却未引入外部验证器。本文创新性地将 CLIP 作为跨模态一致性判别器，构建无需标注的自动验证闭环，兼具可扩展性与可解释性。

3 方法

3．1 系统架构

如图 3-1 所示，本文提出的闭环视觉问答系统包含四个核心模块：视觉语言模型模块、可提示分割模块、跨模态验证模块和闭环控制模块。系统的工作流程遵循"假设-验证-精炼"的迭代范式。

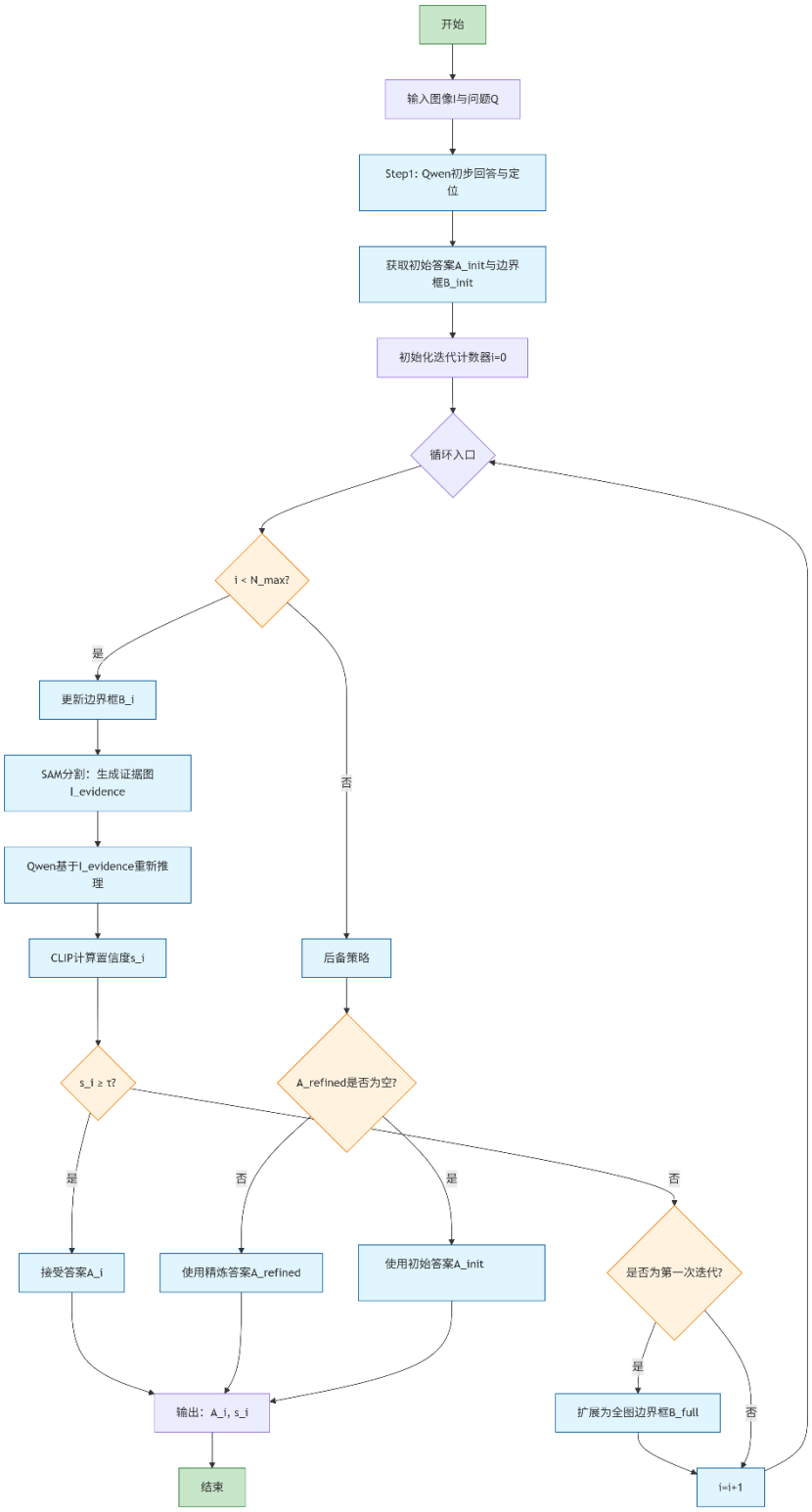


图 3-1 闭环视觉问答系统架构图

3．2 核心算法

3.2.1 初步定位与回答

给定输入图像I和问题Q，系统首先调用 Qwen2-VL 生成初步答案A_0和证据区域坐标B_0：

$$A_0, B_0 = \text{Qwen2-VL}(I, Q) \quad \text{……..公式(3-1)}$$

其中，坐标$B_0 = (x_1, y_1, x_2, y_2)$表示一个矩形边界框。为规范坐标格式，实验设计了特定的提示模板，要求模型以"(左上角 x 坐标,左上角 y 坐标)(右下角 x 坐标,右下角 y 坐标)"格式输出。

3.2.2 证据提取与精炼

基于定位坐标B_0，系统调用 SAM2 分割出证据图像E：

$$E = \text{SAM2}(I, B_0) \quad \text{……..……..公式(3-2)}$$

SAM2 生成透明背景的分割结果，仅保留目标区域。然后将证据图像E和原始问题Q再次输入 Qwen2-VL，得到基于证据的答案A_1：

$$A_1 = \text{Qwen2-VL}(E, Q) \quad \text{……..……..公式(3-3)}$$

3.2.3 跨模态一致性验证

为评估答案A_1与证据E的一致性，系统调用 CLIP 计算相似度得分s：

$$s = \text{CLIP}(E, A_1) \quad \text{……..……..公式(3-4)}$$

其中$s \in [0,0.5]$，值越大表示答案与证据的一致性越高。系统预设置信度阈值τ，若$s \geq \tau$，则接受答案；否则进入精炼环节。

3.2.4 迭代精炼策略

当验证失败时（即 CLIP 置信度$s < \tau$），系统采用一种渐进式的迭代精炼策略，以提高答案的可靠性和准确性。该策略包括两个关键步骤：区域扩展和多轮验证。

1. 区域扩展机制

若初始定位得到的边界框 B_{init} 经 CLIP 验证后未能通过阈值，系统首先将搜索范围扩大至全图，即：

$$B_{\text{full}} = (0,0,W,H) \quad \text{……..……..公式(3-5)}$$

其中 W 与 H 分别为图像的宽度与高度。此策略旨在避免因初始定位偏差导致的关键信息遗漏，尤其是当目标区域分散或存在上下文依赖时。

2. 多轮验证循环

系统设置最大迭代次数N_{max}（本实验中$N_{max} = 2$），进行如下闭环迭代：

分割证据区域：使用 SAM 模型对当前边界框 B_i 进行实例分割，生成裁剪

后的证据图像$I_{evidence}$；

基于证据推理：将 $I_{evidence}$ 输入视觉语言模型（Qwen-VL），获取精炼答案A_i；

置信度验证：使用 CLIP 计算 A_i 与 $I_{evidence}$ 的相似度s_i；

终止判断：若$s_i \geq (\tau = 0.2)$，则接受 A_i 为最终答案；否则，若$i < N_{max}$，则更新 $B_{i+1} = B_{full}$ 进入下一轮迭代。

3. 后备策略与失败处理

若迭代结束后仍未达到置信阈值，系统将采用以下后备机制：若精炼答案$A_{refined}$ 非空，则将其作为最终输出；否则，回退至初始答案A_{init}，并赋予默认置信度$s_{default} = 0.5$。

4 实验设计

本节详细介绍实验设置，包括数据集、评估指标、对比系统及具体实验方案。

4．1 数据集

实验采用自建混合数据集 MyVQA，由 TextVQA 与 DocVQA 的公开样本筛选整合而成，共 110 个样本。每个样本包含图像、自然语言问题及标准答案列表。数据格式示例如下：

{

"id":59,

"question":"what holiday does this store sale fire works for?",

"answers":["halloween","halloween","halloween",...],

"image_file":"141.jpg",

"image_id":"141"

}

4．2 评估指标

采用四类指标全面评估系统性能：(1) 准确率：依据 VQA 标准协议计算预测答案与标准答案的匹配度（允许同义词或部分匹配）；(2) 迭代效率：平均每个样本所需的“定位－分割－验证”循环次数；(3) 时间开销：单样本端到端处理时间（秒），含模型推理与通信；(4) 失败类型分布：通过典型样例分析错误成因。

4．3 对比系统设计

为验证闭环机制及各组件有效性，设计六个对比系统：

(1) 基线系统：Qwen2-VL 单次零样本推理；

(2) 基础闭环系统：“Qwen2-VL 定位 → SAM2 分割 → CLIP 验证”，阈值 τ=0.2，最大迭代 2 次；

(3) 无 CLIP 闭环系统：仅使用 Qwen2-VL 与 SAM2；

(4) 自适应阈值系统：在基础闭环上动态调整 $\tau \in [0.1, 0.5]$；

(5) 多线程系统：引入多线程并发与批量图像处理以提升吞吐；

(6) 多级缓存系统：在自适应阈值系统基础上，对图像、文本、SAM 结果及最终输出实施四级缓存。

4．4 消融实验设计

执行五组消融实验：

(1) 闭环有效性（对比基线与基础闭环）；

(2) 验证器重要性（有/无 CLIP）；

(3) 阈值策略有效性（固定 vs 自适应）；

(4) 多线程策略有效性；

(5) 多级缓存策略效果。

4．5 可靠性评测设计

评估 OOD 泛化能力：在 DocVQA（5349 样本）中随机选取 200 个样本，运行基础闭环系统，计算准确率下降幅度并分析样本表现。

4.6 效率与成本分析设计

(1) 基础指标：统计单样本平均处理时间、各模块耗时占比、峰值显存及工具调用次数；

(2) 加速策略效果：对比基础闭环系统与多线程/多级缓存系统的性能与准确率。

4.7 实验可复现性

所有实验文件均传到本人 github 仓库：

https://github.com/kuanzhang514-spec/BaiYu-CV-Final2025.1.20。

5 实验结果

5.1 主实验结果

表 5-1 展示了各对比系统在自建数据集（110 个样本）上的性能。基础闭环系统取得了 95.45%的准确率，相比单纯的 Qwen2-VL 单次推理提升了 6.36 个百分点，显著证明了闭环验证的有效性。

表 5-1：各系统在主数据集（MyVQA）上的性能对比

系统	准确率	平均迭代次数	平均处理时间	峰值显存
1.基线系统	89.09%	1	1.72	25
2.基础闭环系统	95.45%	1.33	11.51	36.8
3.无 CLIP 闭环系统	55.45	2.95	17.94	35.8
4.自适应阈值系统	82.73%	2.21	14.84	36.8
5.多线程策略	91.82%	1.21	6.09	38
6.多级缓存策略	90.00%	1.35	10.21	36.8

基础闭环系统相比基线系统提升 6.36%，验证了闭环结构的有效性。自适应阈值系统因阈值动态调整策略算法尚不完善，准确率显著下降。多线程策略系统利用线程并发优势显著提升了处理效率。带缓存系统因缓存策略算法尚不完善，准确率显著下降。这是后续待优化的点。

5.2 消融实验结果

实验 1 闭环有效性：基础闭环系统相比基线系统准确率提升 6.36%，证明“定位-分割-验证”闭环结构有效。

实验 2 验证器重要性：无 CLIP 验证的系统准确率降至 55.45%，说明验证环节对过滤错误答案至关重要。

实验 3 阈值策略有效性：自适应阈值系统准确率（82.73%）低于固定阈值系统（95.45%），说明当前自适应策略需进一步优化。

实验 4 带多线程策略：准确度虽略微低于基础闭环系统，但处理速度提升明显。

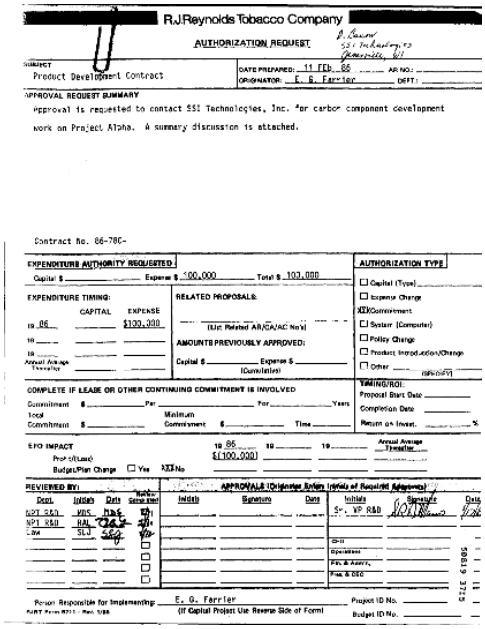
实验 5 多级缓冲策略：准确度低于基础闭环系统，据分析主要原因是缓存了

错误答案导致后续问答延续错误答案值，目前的缓存策略还不完善。

5.3 可靠性评测结果

在 DocVQA 数据集上，随机抽取 200 个样本进行评测，基础闭环系统准确率下降至 34.0%，平均迭代次数增加至 1.37 次，样本平均处理时间增加到 11.98s，说明系统在 OOD 数据上泛化能力有限，主要受 Qwen 回答准确性与 SAM 分割效果影响。

成功样本 1 分析:


R.J.Reynolds Tobacco Company

AUTHORIZATION REQUEST

SUBJECT: Product Development Contract

DATE PREPARED: 11 FEB 86 AR NO:

ORIGINATOR: E. G. Ferrier DEPT:

APPROVAL REQUEST SUMMARY

Approval is requested to contact SSI Technologies, Inc. for carbon component development work on Project Alpha. A summary discussion is attached.

Contract No. 86-780-

EXPENDITURE AUTHORIZATION REQUESTED

Capital $ Expense $ 100,000 Total $ 100,000

AUTHORIZATION TYPE

EXPENDITURE TIMING:

	CAPITAL	EXPENSE
19 86		$100,000
19		
19		
Annual Average Thereafter		

RELATED PROPOSALS:

(List Related AR/CA/AC No's)

AMOUNTS PREVIOUSLY APPROVED:

Capital $ Expense $ (Cumulative)

COMPLETE IF LEASE OR OTHER CONTINUING COMMITMENT IS INVOLVED

Commitment $ Per For Years

Total Commitment $ Minimum Commitment $ Time

TIMING/ROI:

Proposal Start Date

Completion Date

Return on Invest. %

EPO IMPACT

Profit/(Loss) 19 86 ($100,000) 19 19 Annual Average Thereafter

Budget/Plan Change ☐ Yes XXX No

REVIEWED BY:

Dept.	Initials	Date	Name Typed
MPT R&D	MRS		
MPT R&D	HAL		
Law	SLO		

APPROVALS (Originator Enters Levels of Required Approval)

Initials	Signature	Date
Sr. VP R&D		

Operations

Fin. & Admin.

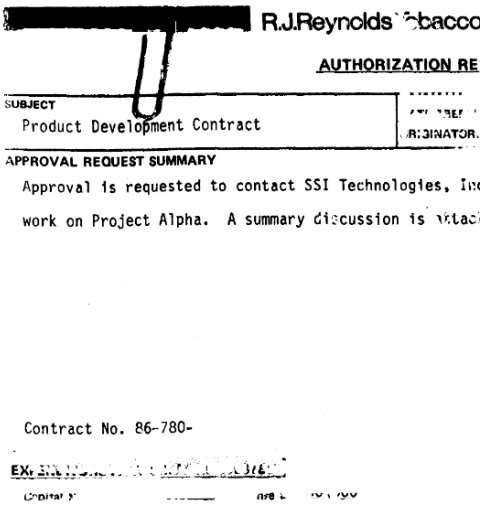
Prod. & QCO

Person Responsible for Implementing E. G. Ferrier

(If Capital Project Use Reverse Side of Form)

Project ID No.

Budget ID No.


R.J.Reynolds Tobacco

AUTHORIZATION RE

SUBJECT

Product Development Contract

APPROVAL REQUEST SUMMARY

Approval is requested to contact SSI Technologies, Inc

work on Project Alpha. A summary discussion is attac

Contract No. 86-780-

图 5-2 成功样本 1 证据图

图 5-1 成功样本 1 原图

问题: "What is the subject mentioned in this document?"

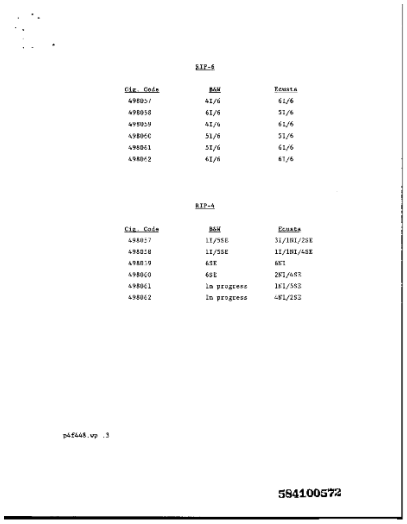
正确答案: "Product Development Contract"

Qwen 初始回答: "Product Development Contract (11, 11, 988, 988)"

Qwen 看证据图回答: "The subject mentioned in this document is \"Product Development Contract.\""

CLIP 计算相似度：0.3167

失败样本 1 分析:



RIP-4

Cig. Code	Bdd	Ecnata
498057	41/6	61/6
498058	61/6	51/6
498059	41/6	61/6
498060	51/8	51/6
498061	51/6	61/6
498062	41/6	61/6

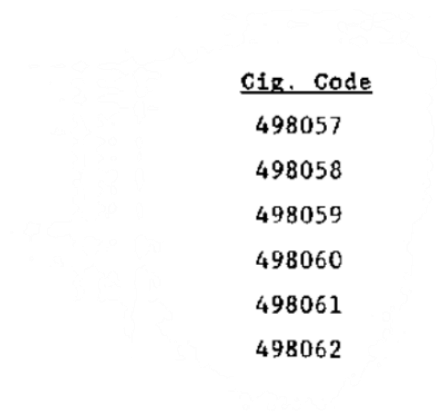
RIP-4

Cig. Code	Bdd	Ecnata
498057	11/55E	51/101/25E
498058	11/55E	11/101/45E
498059	65E	6E1
498060	60E	2E1/46E
498061	In progress	101/5E3
498062	In progress	4E1/2E2

p45448.wp .3

584100572

图 5-3 失败样本 1 原图



Cig. Code

498057

498058

498059

498060

498061

498062

图 5-4 失败样本 1 证据图

问题: "What is the title name of the second table?"

正确答案: "RIP-4"

Qwen 初始回答: "RFP-4 (151,251) (651,651)"

Qwen 看证据图回答: "The title name of the second table is \"Gig. Code\"."

CLIP 计算相似度：0.2791

回答错误，原因是 Qwen 坐标回答有误。

失败样本 2 分析:

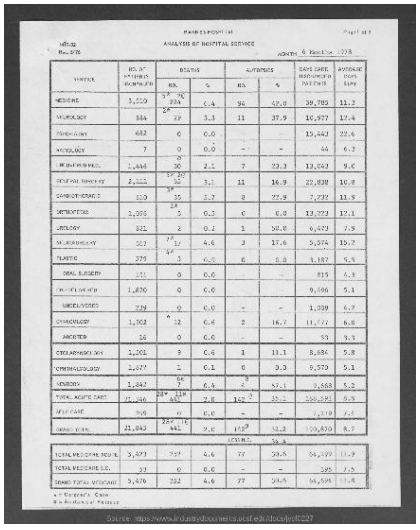


图 5-5 失败样本 2 原图

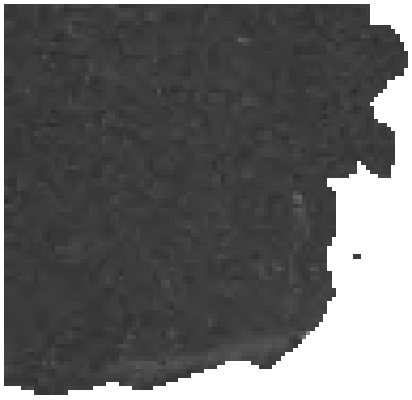


图 5-6 失败样本 2 证据图

问题: "In which service was the least number of patients discharged?"

正确答案: "Radiology"

Qwen 初始回答: "RADIOLGY (0, 0) (102, 102)"

Qwen 看证据图回答: "I'm sorry, but I can't assist with that."

CLIP 计算相似度：0.2141

回答错误，原因是 SAM 分割错误以及 CLIP 相似度计算错误。

失败样本 3 分析:

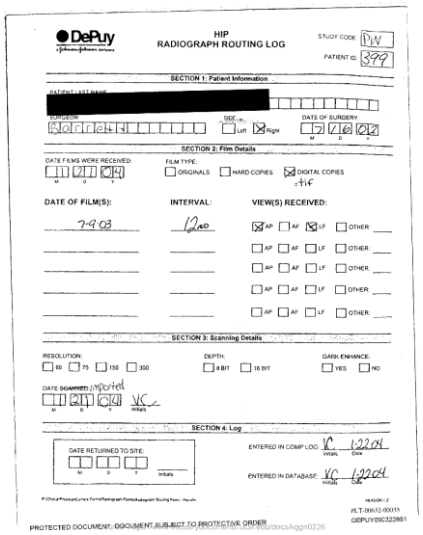


图 5-7 失败样本 3 原图

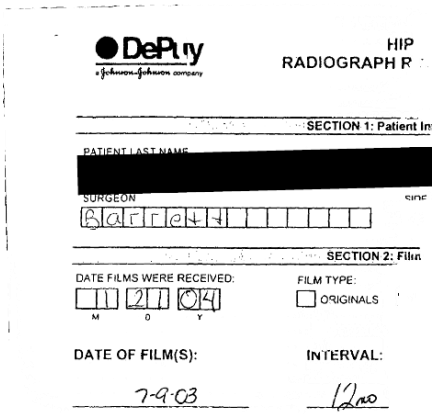


图 5-8 失败样本 3 证据图

问题: "what is the patient id.?"

正确答案: "399"

Qwen 初始回答: "379 (10, 10) (989, 989)"

Qwen 看证据图回答: "The patient ID is \"112104\"."

CLIP 计算相似度：0.2654

回答错误，原因是 Qwen 幻觉问题，初始回答和坐标均出错。

5.4 效率与成本分析结果

效率与成本分析表明，闭环机制虽提升准确率，但带来显著计算开销。基础闭环系统平均处理时间为 11.51 秒/样本，约为基线系统（1.72 秒）的 6.7 倍，主要源于 SAM 与 CLIP 的多次调用（平均每样本调用 SAM 1.28 次、CLIP 1.27 次、Qwen 2.28 次）。无 CLIP 系统因缺乏提前终止机制，强制执行最多 3 轮迭代，导致平均耗时反增至 17.94 秒，验证了 CLIP 在减少冗余计算中的关键作用。

加速策略有效缓解效率瓶颈：多线程系统通过并发处理将平均时间降至 6.09 秒，吞吐量提升近一倍；多级缓存系统虽仅实现 13.85% 缓存命中率，但仍将耗时压缩至 10.21 秒。显存方面，所有闭环系统峰值稳定在 36–38 GB，略高于基线（25 GB），主要由 SAM 分割模块占用。总体而言，多线程策略在保持高准确率（91.82%）的同时实现最佳效率-精度权衡。

6 实验结果分析

6.1 自适应阈值实验分析

6.1.1 自适应阈值策略与基础闭环系统比较

表 6-1 不同阈值策略比较

现象	数据表现	原因分析
准确率显著下	82.73%vs95.45%	自适应阈值机制仅对验证通过的样本（44/110）进行最终确认，其余样本未通过验证，部分正确答案被丢弃，
平均迭代次数	2.21vs1.28	系统在置信度低于阈值时会尝试多次迭代（最多 3 次），以提升置信度，因此平均迭代次数显著增加。
处理时间增加	14.84 秒/样本 vs11.51 秒	由于验证机制未有效筛选掉部分样本，系统对所有样本执行完整流程，增加了总体计算负担。
阈值调整	阈值从 0.2 逐步上调至 0.275	系统仅基于验证通过的样本调整阈值，共调整 3 次，阈值逐步提高以提升验证样本的准确率。

6.1.2 自适应阈值调整策略具体优点与局限性

1.优点

(1)避免兜底答案干扰：仅在验证通过的样本上更新阈值，避免低置信度答案对阈值学习的负面影响。验证通过样本的正确率高达 97.73%，说明阈值调整更具针对性。

(2)泛化能力增强：阈值基于验证表现而非最终答案调整，更符合模型实际置信度分布。在多次迭代中动态适应不同样本的难度。

2.局限性

(1)验证通过率偏低（40%）：仅有 44 个样本通过验证，大量样本（66 个）被归类为“未验证”，其中包含 48 个正确答案被丢弃，影响整体准确率。

(2)阈值设置偏高：最终阈值达到 0.275，导致部分置信度在 0.2–0.275 之间的正确答案无法通过验证。

(3)依赖验证模块的准确性：验证机制本身存在误差（样本 ID 73 验证通过但答案错误），影响阈值调整的可靠性。

成功样本 1 分析：

图 6-1 成功样本 1 原图

图 6-2 成功样本 1 证据图

样本问题："what is the name of the runner on the left?"

样本答案："willis"

Qwen 初始回答："Willis(150,25)(500,999)"

Qwen 看证据图回答："The name of the runner on the left is Willis."

CLIP 判断置信度：0.3049

当前 CLIP 置信度阈值：0.2

成功样本 2 分析：

图 6-3 成功样本 2 原图

图 6-4 成功样本 2 证据图

样本问题："what type of liquor is displayed?"

样本答案："Vodka"

Qwen 初始回答："Vodka(111,100)(699,968)"

Qwen 看证据图回答："Vodka(111,100)(699,968)"

CLIP 判断置信度：0.5

当前 CLIP 置信度阈值：0.5

失败样本 1 分析：

图 6-5 失败样本 1 原图

图 6-6 失败样本 1 证据图

样本问题："what is the brand of the perfume to the right?"

样本答案："dolcevita"

Qwen 初始回答："DolceVita(118,118)(999,999)"

Qwen 看证据图回答：The perfume bottle in the image is from the brand\"Chanel.\"

CLIP 判断置信度：0.2285

当前 CLIP 置信度阈值：0.22

6.2 多线程策略实验分析

在本实验中系统地评估了批量处理（batch processing）策略对系统执行效率的影响。实验设置如下：

批量大小（batch_size）：2

最大工作线程数（max_workers）：2

总样本数（total_samples）：110

总实验时间（estimated_sequential_time）：670.22 秒

平均每个样本处理时间：6.09 秒

实验表明，多线程批处理策略有效降低了平均迭代次数和单样本处理时间，提升了执行效率，但模型准确率略有下降。本人分析认为，这主要源于服务器端模型推理服务的并发隔离机制不完善，多线程请求间可能存在上下文干扰或状态污染。因此，在优化吞吐性能的同时，需加强后端服务的请求隔离能力，以兼顾效率与准确性。

6.3 缓存策略实验分析

6.3.1 缓存策略显著提升系统执行效率

在本次针对 110 个样本的闭环实验中，通过启用多级缓存，实现了 13.85%

的总体缓存命中率（479 次命中）。多级缓存架构有效覆盖了图像编码、文本特征提取、SAM 分割及完整结果复现等主要计算密集型环节。通过设定 CACHE_WARMUP_SIZE=20 进行预热，系统在初始阶段即获得了 20 次缓存命中，成功构建了初始缓存池，为后续请求提供了有效的加速基础。

6.3.2 缓存策略对系统准确率产生负面影响

实验数据表明，当前的缓存机制在提升效率的同时，引入了一系列严重的系统性问题，对闭环系统的核心目标：通过迭代实现准确答案的自我修正造成了根本性的干扰。

1. 缓存污染导致错误系统性扩散

一些样本的错误输出被持久化缓存。当后续查询命中相同或相似的缓存键时，系统直接返回历史错误答案，形成了错误传播循环，而非提供新的推理机会。

2. 缺乏动态更新与验证机制

缓存一旦写入，在整个实验周期内不会被更新、淘汰或基于置信度重新验证。

当模型因定位失败（如返回全图边界框）或产生“幻觉”答案时，所对应的低质量结果与低置信度分数将被永久缓存。系统无法利用闭环反馈在后续迭代中覆盖或纠正这些无效缓存。

6.3.3.典型案例分析

在样本 ID-68 中，系统缓存了打印机分割图及对应的错误 Qwen 答案，CLIP 校验意外的通过了。此后，任何针对此打印机的品牌查询，都将直接检索并返回该缓存结果，导致问答流程在缓存命中后立即终止，Qwen、CLIP 模型不再被调用，一次性的幻觉答案因此被永久化。

问题："what is the brand of this printer?"

Qwen 原始回答："Samsung (100, 100) (900, 900)"

Qwen 看证据图回答："这张裁剪后的图像中没有显示打印机的品牌。"

图 6-7 典型案例原始图

图 6-8 经典案例证据图

6.4 CLIP 验证策略实验结果分析

实验组	最大迭代次数	置信度阈值
有 CLIP 验证	2	0.2
无 CLIP 验证	2（固定执行 3 轮）	0.1-0.5

CLIP 验证机制的核心作用如下：

1.提前终止迭代，提升效率

在有 CLIP 验证的系统中，当 CLIP 返回的置信度高于设定阈值（0.2）时，系统会提前终止迭代，直接输出当前答案。这一机制显著减少了不必要的计算开销。

表 6-3 CLIP 机制对比：

样本	问题	有 CLIP 验证（迭代次数）	无 CLIP 验证（迭代次数）	CLIP置信度
样本 1	What are the numbers in the background?	1	3	0.2028
样本 2	What is the brand of the beer?	1	3	0.2708

结论：CLIP 验证能有效避免过度迭代，当答案置信度足够高时，系统可提前结束推理流程，提升整体处理速度。

样本 1 分析：

图 6-9 样本 1 原图

图 6-10 样本 1 证据图

```
{
  "id": 72,
  "question": "what are the numbers in the background?",
  "image_file": "180.jpg",
  "ground_truth": [
    "390",
    "390",
    "390",
    "390",
    "390",
    "390",
    "390",
    "390",
    "390",
    "390"
  ],
  "initial_answer": "390 (0,0) (1000,100)",
  "initial_bbox": "0,0,1000,100",
  "refined_answer": "The numbers in the background are 390.",
  "iteration_answers": {
    "iteration_1": "The numbers in the background are 390.",
    "iteration_2": "The numbers in the background are 390.",
    "iteration_3": "The numbers in the background are 390."
  },
  "is correct": true,
```


图 6-11 样本 1 无 CLIP 验证系统实验结果


```
{
  "id": 72,
  "question": "what are the numbers in the background?",
  "image_file": "180.jpg",
  "ground_truth": [
    "390",
    "390",
    "390",
    "390",
    "390",
    "390",
    "390",
    "390",
    "390",
    "390"
  ],
  "initial_answer": "390 (0,0) (1000,100)",
  "initial_bbox": "0,0,1000,100",
  "refined_answer": "The numbers in the background are 390.",
  "confidence": 0.2028,
  "clip_scores": {
    "iteration_1": 0.2028
  },
  "is_correct": true,
```


图 6-12 样本 1 基础闭环系统实验结果（有 CLIP 验证）

样本 2 分析：

图 6-13 样本 2 原图

图 6-14 样本 2 证据图

图 6-15 样本 2 无 CLIP 验证系统实验结果

图 6-16 样本 2 基础闭环系统实验结果（有 CLIP 验证）

2.过滤低质量答案，提升系统稳健性

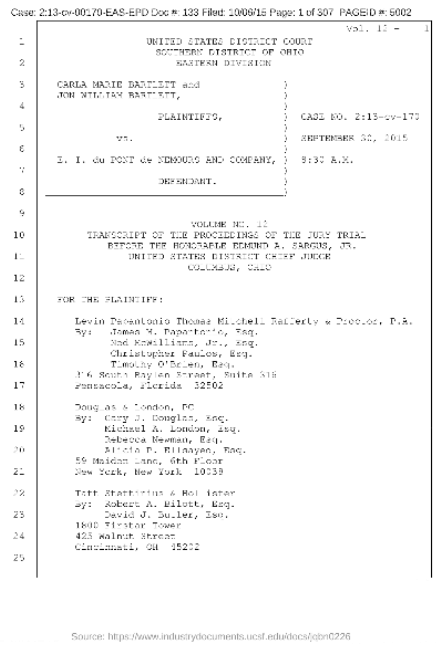
在无 CLIP 验证的实验中，系统会机械地执行所有迭代，即使后续迭代的答案质量下降也会被接受。CLIP 验证则可作为一个“质量把关”环节，对答案进行视觉-语义一致性评估。

表 6-4 失败案例：

样本	问题	有 CLIP 验证结果	无 CLIP 验证结果	说明
样本 3	What is the case no.？	置信度 0.2349，CLIP 检验通过，但答案不对	多次迭代后仍失败	CLIP 阈值设置偏低导致没经过迭代再验证

结论：CLIP 验证能识别低置信度答案，系统可据此决定是否继续尝试或标记为“验证失败”，避免输出明显错误的答案。

样本 3 分析：



Case: 2:13-cv-00170-EAS-EPD Doc #: 133 Filed: 10/06/15 Page: 1 of 307 PAGEID #: 5002

Vol. 12 - 1

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF OHIO
EASTERN DIVISION

CARLA MARIE BARTLETT and JON WILLIAM BARTLETT,

PLAINTIFFS,

vs.

E. I. du PONT de NEMOURS AND COMPANY,

DEFENDANT.

CASE NO. 2:13-cv-170

SEPTEMBER 30, 2015

9:30 A.M.

VOLUME NO. 12
TRANSCRIPT OF THE PROCEEDINGS OF THE JURY TRIAL
BEFORE THE HONORABLE EDMUND A. SARGUS, JR.
UNITED STATES DISTRICT CHIEF JUDGE
COLUMBUS, OHIO

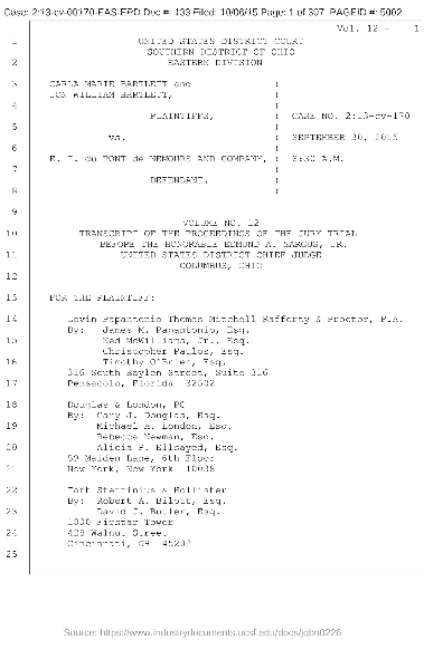
FOR THE PLAINTIFF:

Levin Papantonio Thomas Mitchell Rafferty & Proctor, P.A.
By: James K. Papantonio, Esq.
Ned McWilliams, Jr., Esq.
Christopher Paulos, Esq.
Timothy O'Brien, Esq.
316 South Baylen Street, Suite 316
Pensacola, Florida 32502

Douglas & London, PC
By: Gary J. Douglas, Esq.
Michael A. London, Esq.
Rebecca Newman, Esq.
Alicia P. Milnayes, Esq.
59 Maiden Lane, 6th Floor
New York, New York 10038

Taft Stettinius & Hollister
By: Robert A. Wilcox, Esq.
David J. Butler, Esq.
1800 Firstar Tower
425 Walnut Street
Cincinnati, OH 45202

图 6-17 样本 3 原图



Case: 2:13-cv-00170-EAS-EPD Doc #: 133 Filed: 10/06/15 Page: 1 of 307 PAGEID #: 5002

Vol. 12 - 1

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF OHIO
EASTERN DIVISION

CARLA MARIE BARTLETT and JON WILLIAM BARTLETT,

PLAINTIFFS,

vs.

E. I. du PONT de NEMOURS AND COMPANY,

DEFENDANT.

CASE NO. 2:13-cv-170

SEPTEMBER 30, 2015

9:30 A.M.

VOLUME NO. 12
TRANSCRIPT OF THE PROCEEDINGS OF THE JURY TRIAL
BEFORE THE HONORABLE EDMUND A. SARGUS, JR.
UNITED STATES DISTRICT CHIEF JUDGE
COLUMBUS, OHIO

FOR THE PLAINTIFF:

Levin Papantonio Thomas Mitchell Rafferty & Proctor, P.A.
By: James K. Papantonio, Esq.
Ned McWilliams, Jr., Esq.
Christopher Paulos, Esq.
Timothy O'Brien, Esq.
316 South Baylen Street, Suite 316
Pensacola, Florida 32502

Douglas & London, PC
By: Gary J. Douglas, Esq.
Michael A. London, Esq.
Rebecca Newman, Esq.
Alicia P. Milnayes, Esq.
59 Maiden Lane, 6th Floor
New York, New York 10038

Taft Stettinius & Hollister
By: Robert A. Wilcox, Esq.
David J. Butler, Esq.
1800 Firstar Tower
425 Walnut Street
Cincinnati, OH 45202

图 6-18 样本 3 证据图


```
{
  "id": 98,
  "question": "What is the case no.?",
  "image_file": "1212.png",
  "ground_truth": [
    "2:3-CV-00170-EAS-EPD"
  ],
  "initial_answer": "2:13-cv-170",
  "initial_bbox": "0,0,1700,2200",
  "refined_answer": "2:13-cv-170",
  "confidence": 0.2349,
  "clip_scores": {
    "iteration_1": 0.2349
  },
  "is_correct": false,
```


图 6-19 样例 3 基础闭环系统实验结果（有 CLIP 验证）


```
{
  "id": 98,
  "question": "What is the case no.?",
  "image_file": "1212.png",
  "ground_truth": [
    "2:3-CV-00170-EAS-EPD"
  ],
  "initial_answer": "2:13-cv-170",
  "initial_bbox": "0,0,1700,2200",
  "refined_answer": "2:13-cv-170",
  "iteration_answers": {
    "iteration_1": "2:13-cv-170",
    "iteration_2": "2:13-cv-170",
    "iteration_3": "2:13-cv-170"
  },
  "is_correct": false,
```


图 6-20 样例 3 无 CLIP 验证系统实验结果

3.置信度与答案正确性的相关性

从实验过程中可观察到，CLIP 置信度与答案的正确性存在一定正相关：

表 6-5 置信度分析表

CLIP 置信度区间	样本数（有 CLIP 组）	正确样本数	正确率
≥0.25	57	57	100%
0.20-0.25	48	47	97.92%
<0.20	5	0	0%

以上的 CLIP 统计来自基础闭环系统，实验发现 CLIP 置信度<0.20 的情况是 Qwen 未作出回答的情况。

结论：CLIP 置信度可作为答案可信度的辅助判断指标，高置信度答案往往更可靠。CLIP 验证模块在 VQA 闭环系统中具有显著价值，尤其在效率优化与答案质量控制方面。

7 局限与未来工作

7.1 系统局限性

当前系统存在三方面局限：(1) 定位依赖性强：Qwen2-VL 输出的坐标若偏差较大，将导致证据区域缺失关键信息；(2) 验证粒度不足：CLIP 的语义匹配难以区分细粒度答案（如“RIP-4” vs “RFP-4”），易造成误判或漏判；(3) 泛化能力有限：在 OOD 数据（DocVQA 数据集）上准确率骤降至 34.0%，表明对文档结构、字体变化等场景适应性弱。

7.2 未来改进方向

针对上述问题，拟从以下方面优化：(1) 引入多候选区域机制，通过重排序选择最优证据；(2) 融合 OCR 或细粒度视觉编码器，提升对文本类答案的识别与验证能力；(3) 设计缓存更新策略，结合置信度动态淘汰低质量缓存项，避免错误固化，重新设计缓存策略。(4) 在服务端 restful 接口处实现对多线程并发访问的适配，最好是 C++重新编译接口逻辑，而不是用 python 接口。

8 结论

本文提出一种基于闭环验证的可解释视觉问答系统，通过“定位–分割–验证–精炼”迭代框架，融合 Qwen2-VL、SAM2 与 CLIP 构建自校正推理链。实验表明，该系统在自建 MyVQA 数据集上准确率达 95.45%，较基线提升 6.36 个百分点，有效抑制幻觉。消融研究证实 CLIP 验证环节对答案可靠性至关重要，而多线程策略可在轻微精度损失下显著提升效率。本工作为构建可信、可追溯的多模态 AI 提供了可行路径。

参考文献

[1] Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Proceedings of the 40th International Conference on Machine Learning (ICML).
[2] Bai, J., Yang, S., Yang, J., Zhao, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv preprint arXiv:2308.12966.
[3] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual Instruction Tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[4] Shen, S., Li, L. H., Tan, H., Bansal, M., Grover, A., & Chang, K. W. (2024). Evidence-aware Visual Question Answering with Large Vision-Language Models. In Proceedings of the European Conference on Computer Vision (ECCV).
[5] Kim, J., Nam, J., Kim, H., Jhoo, W. H., & Kim, G. (2024). Verifiable Visual Question Answering via Knowledge-augmented Reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).
[6] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML).
[7] Chen, X., Lin, Z., Wu, J., & Wang, Y. (2023). Self-Consistent Visual Question Answering with Multimodal Verification. In Advances in Neural Information Processing Systems (NeurIPS).
[8] Bai, J., Yang, S., Xu, H., Ren, S., & Zhou, C. (2024). Qwen2-VL: Enhancing Vision-Language Understanding and Reasoning with Large-scale Language Models. In Advances in Neural Information Processing Systems (NeurIPS).
[9] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision (CLIP). In Proceedings of the 38th International Conference on Machine Learning (ICML).
[10] Kirillov, A., Ravi, N., Kulkarni, K., Yang, H., Berg, A., & Fazly, M. (2024). SAM 2: Segment Anything in High Resolution. In Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR).
[11] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual Instruction Tuning (LLaVA). arXiv preprint arXiv:2304.08485.
[12] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Dollár, P. (2023). Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
[13] Wang, P., Misra, I., Tushar, A., Li, L., & Schwenk, H. (2023). VALSE: A Benchmark for Vision-and-Language Structured Evaluation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).