

适用于内容安全审核的图像文本识别质量评估方法

张玉龙^{1,2}, 梁天一³, 张茹¹, 刘功申²

(1. 北京邮电大学网络空间安全学院, 北京 100876;
2. 上海交通大学计算机学院(网络空间安全学院), 上海 200240)
3. 华东师范大学计算机科学与技术学院, 上海 200062;

摘要: 针对内容安全审核场景下图像文本识别质量评估的难题, 提出基于一致性置信(CC)机制的质量评估方法。现有评估方法存在显著不足, 传统方法依赖人工标注且成本高昂, 视觉语言模型作为评判器(VLM-as-Judge)方法存在逻辑悖论和偏见, 不确定性量化方法需要访问模型内部参数。基于CC机制的质量评估方法通过多模型一致性分析实现无监督质量评估, 利用编辑距离量化一致性程度, 并设计动态权重和阈值门控机制识别低质量输出。该方法具有黑盒评估、完全无监督、模型无关性三大核心优势。实验结果表明: 在两个数据集上质量验证F1得分达49.9%~51.61%, 相比基线方法提升42.76%; 在对抗性文本、复杂场景等挑战性场景下均表现最佳; 模型集成使低成本开源模型超越高成本闭源模型, 性能提升达32分。

关键词: 图像文本识别; 质量评估; 内容安全; 无监督

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.AQ25222

Quality assessment method for image text recognition in content security review

Zhang Yulong^{1,2}, Liang Tianyi³, Zhang Ru¹, Liu Gongshen²

1. School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China
2. School of Computer Science (Cyber Science and Engineering), Shanghai Jiao Tong University, Shanghai 200240, China
3. School of Computer Science and Technology, East China Normal University, Shanghai 200062, China

Abstract: A quality assessment method based on the consistency confidence (CC) mechanism was proposed to address the challenge of image text recognition quality assessment in content security auditing scenarios. Significant deficiencies were identified in existing evaluation methods. Traditional methods were found to rely on manual annotation and be cost-prohibitive. The visual language model as judge (VLM-as-Judge) approach was observed to suffer from logical paradoxes and biases. Uncertainty quantification methods were shown to require access to internal model parameters. With the proposed CC-based method, unsupervised quality assessment was achieved through multi-model consistency analysis. Edit distance was utilized to quantify the degree of consistency, and dynamic weighting with threshold gating mechanisms was designed to identify low-quality outputs. This method possesses three major advantages: black-box evaluation, complete unsupervisedness, and model-agnostic property. Experimental results demonstrated that the quality verification F1 score reached 49.9%~51.61%, which represented a 42.76% improvement over baseline methods. The best performance was exhibited in challenging scenarios such as

收稿日期: 2025-08-24; 修回日期: 2026-03-14

通信作者: 刘功申, lgshen@sjtu.edu.cn

基金项目: 国家自然科学基金联合重点项目 (No.U21B2020)

Foundation Item: The Joint Funds of the National Natural Science Foundation of China (No. U21B2020)

adversarial text and complex scenes. Through model ensemble, low-cost open-source models were enabled to surpass high-cost closed-source models, with a performance improvement of 32 points.

Key words: image text recognition, quality assessment, content security, unsupervision

0 引言

在数字化信息时代,图像内容已成为网络信息传播的主要载体之一。据统计,大量网络内容包含图像元素,其中相当比例的图像携带文本信息^[1-3]。随着社交媒体、在线论坛、短视频平台和数字营销的快速增长,图像中的文本内容越来越多地被用于信息传递、观点表达和商业推广。这一趋势使得内容安全审核系统面临前所未有的挑战:如何准确、高效地理解和分析海量图像中的文本信息,成为维护网络空间安全的关键技术问题。

内容安全审核作为数字平台治理的核心环节,其有效性直接关系到网络环境的健康发展^[1-3]。传统的纯文本审核机制已无法应对当前复杂的内容形态,现代审核系统必须具备多模态理解能力,特别是对图像中文本内容的准确识别和理解能力^[4-5]。图像文本识别技术,即光学字符识别(optical character recognition, OCR)技术,作为连接视觉信息与文本理解的桥梁,在整个审核流程中扮演着基础性和决定性的作用(下文统称“图像文本识别”。

当前主流的内容审核解决方案主要采用两种技术路线:一是直接基于视觉模型的端到端理解方案,通过模型直接生成对图像内容的判断;二是基于图像文本提取+文本分析的分步处理方案,先提取图像中的文本内容,再进行后续的语义分析和安全评估^[5-6]。然而,第一种方案往往在文本细节处理上存在不足,容易遗漏关键信息;第二种方案的效果很大程度上取决于图像文本提取环节的质量,这使得图像文本识别的可靠性成为影响整个审核系统性能的关键因素^[4,7-8]。

深入分析内容安全审核场景的特殊需求,本文发现现有图像文本识别系统面临着一系列独特挑战。

第一,高可靠性要求。与一般的文档处理应用不同,内容安全审核对图像文本识别的准确性提出了极高的要求。在安全审核场景下,即使是

微小的文本提取错误也可能导致严重后果:漏检可能使有害内容逃避审查,误检则可能导致正常内容被错误处理。这种对错误的极低容忍度要求远超一般图像文本识别应用场景,需要有针对性的技术解决方案^[9-10]。

第二,复杂多变的图像环境。网络图像的多样性和复杂性给图像文本识别带来了巨大挑战。这些图像往往包含复杂的背景、多样化的字体样式、不规则的文本布局、各种视觉特效和人为的干扰因素。与标准化文档相比,网络图像中的文本提取难度呈指数级增长,传统图像文本识别方法的适用性受到严重限制^[4,11-16]。

第三,实时性与规模化处理压力。现代网络平台每天需要处理数以亿计的图像内容,这要求图像文本识别系统不仅要保证高准确率,还要具备强大的并发处理能力和低时延响应特性。如何在保证质量的前提下实现大规模实时处理,是系统设计的重要考量^[6]。

第四,质量评估方法的局限性。在实际应用中,往往缺乏充分的真值标签来评估图像文本识别结果的准确性,这使得质量评估成为内容安全审核系统的关键瓶颈。深入分析现有评估方法,可发现其存在以下不足。传统基于准确率的评估方法依赖大量人工标注数据,不仅成本高昂,而且无法应对海量实时数据的评估需求,更难以适应快速变化的对抗性内容。近年来兴起的视觉语言模型(VLM)作为评判器(VLM-as-Judge)方法使用大模型直接评估文本识别质量^[17-18],但存在根本性逻辑悖论:“如果VLM能够有效识别和评价其他模型的文本识别结果,为什么不直接用该VLM来识别图片?”这种逻辑矛盾不仅削弱了方法的可信度,而且该方法还存在对同架构模型的偏好、对提示敏感、训练数据污染等系统性问题^[19-20]。不确定性量化方法虽然能够提供概率输出,但需要访问模型内部参数(如logits、注意力权重等),这在实际应用中严重限制了其适用范围,特别是对于商业闭源模型和第三方应用程序编程接口(API)服务。更为关键的是,上

述方法均未针对内容安全审核场景的特殊需求进行专门设计，难以有效识别可能导致严重后果的错误输出，使得系统缺乏自我监控和质量保障能力。这些方法层面的不足，而非单纯的系统实现问题，构成了当前质量评估技术的主要挑战。

第五，对抗性挑战。在内容安全审核场景下，部分用户会采用各种技术手段来规避检测，包括使用特殊字体、添加噪声干扰、文本变形和人工智能（AI）对抗生成等反检测技术。这种对抗性环境要求图像文本识别系统不仅要处理自然产生的复杂图像，还要应对人为设计的挑战性输入^[21]。

面对这些挑战，传统的图像文本识别评估方法显得力不从心。现有方法主要关注平均准确率等宏观指标，而在安全审核场景下，本文重点关注以下问题：在没有真值标签的情况下，如何可靠地评估每个图像文本识别输出的质量，并准确识别出可能存在错误的结果；如何构建具有自我监控能力的图像文本识别系统，自动发现并处理不可靠的输出。这些问题的核心在于建立针对安全审核场景的质量评估和不确定性量化机制。

为了深入理解这一问题，本文对多个主流图像文本识别模型在处理相同图像时的行为模式进行了系统性分析。通过对大量真实网络图像的实验观察，本文发现了一个关键现象：当多个模型处理相同图像时，正确识别的结果往往表现出显著的一致性，即不同模型的输出在语义和字符层面都高度相似；而错误或不准确的结果则表现为明显的分歧，不同模型的输出差异较大，显示出较高的不确定性。

这一观察结果为本文提供了重要的启发：模型间的一致性可以作为评估图像文本识别结果可信度的重要指标。基于这一核心洞察，本文提出了一种一致性置信（consistency confidence, CC）机制，专门用于在无监督环境下评估文本识别结果的可信度。CC机制不仅能够识别高质量和低质量的输出，还能够为每个输出提供量化的置信度评估，为后续的处理决策提供重要依据。

本文的主要贡献包括：将图像文本识别质量评估与内容安全审核需求结合，提出一致性置信机制，专门解决安全审核场景下文本识别结果可

信度评估问题；设计了基于多模型一致性的无监督质量评估体系，能够在无真值标签的情况下有效识别高质量和低质量的图像文本识别输出，实现了真正的黑盒评估，可应用于任何OCR模型。大规模实验证实，本文方法在质量验证和结果改进方面的显著优势，相比VLM-as-Judge方法平均提升42.76%，在对抗性文本、复杂场景等挑战性场景下均表现最佳，且能使低成本开源模型超越高成本闭源模型，为实际应用提供了有效的技术方案。

1 研究现状

针对内容安全审核场景下图像文本识别质量评估问题，本节从3个方面回顾相关研究工作：首先分析内容安全审核场景的特殊性及其对质量评估提出的挑战，然后系统梳理现有质量评估方法的成果与不足，最后探讨多模型一致性分析方法的研究进展，为本文提出的基于CC机制的质量评估方法提供理论支撑和对比基础。

1.1 内容安全审核场景的特殊性与评估挑战

内容安全审核场景对图像文本识别的质量评估提出了特殊挑战。尽管多模态大模型和新近的视觉-文本联合建模方法在文档处理方面取得显著进展^[5-6,11-16]，但这些技术主要关注识别准确率的提升，而在质量评估方面缺乏有效手段。内容安全审核场景的特殊性在于：即使微小的文本提取错误也可能导致严重后果，现代网络平台每天需要处理数以亿计的图像内容，恶意用户往往采用特殊字体、噪声干扰、文本变形等对抗性手段来规避检测^[21-23]。这种对错误的极低容忍度要求、大规模实时处理压力和对抗性挑战，使得现有以标准场景为主的评测方法难以满足实际需求，迫切需要针对安全审核场景设计专门的质量评估方法^[4,7-8]。

1.2 图像文本识别质量评估方法

传统的质量评估方法主要依赖人工标注数据计算准确率，成本高昂且难以实时进行^[4,7-8]。值得注意的是，图像文本识别领域的研究工作主要集中在提高识别准确率本身，而非识别结果的质量评估。例如，Qwen2-VL、GPT-4o、InternVL等视觉语言模型虽然在图像文本识别任务上表现优异，但它们本质上是识别模型而非评估方法，

不具备对其他模型识别结果进行质量评估的能力。目前,针对OCR输出的无监督质量评估方法研究十分稀缺。

现有的评估方法主要包括以下几类。

(1) 传统基于准确率的方法。该方法需要大量人工标注数据,成本高昂且难以实时进行^[4,7-8]。

(2) VLM-as-Judge 方法。近年来随着视觉语言模型的发展,该方法使用大模型直接评估文本识别质量^[17-18]。然而,这种方法存在根本性的逻辑悖论。此外,还存在对提示敏感、训练数据污染、格式偏见等问题^[19-20]。

(3) 不确定性量化方法:虽然能够提供概率输出,但需要访问模型内部参数(如logits、注意力权重等),限制了其应用范围^[9-10]。

综上所述,现有方法要么需要人工标注或模型内部参数,要么存在逻辑悖论或偏见问题,缺乏能够在黑盒条件下无监督地评估任意OCR模型输出质量的通用方法。

1.3 多模型一致性分析方法

基于一致性的分析方法逐渐受到重视,其核心假设是:多个独立训练的模型产生一致输出时,该输出更可能是正确的^[24]。一致性分析方法在语言模型领域取得了显著成功,为多模型集成提供了重要启发。然而,现有集成方法通常采用简单的投票机制或加权平均方法,缺乏对模型间一致性的深入分析^[7],且未能解决以下3个关键问题:在图像文本识别任务中有效量化模型间的一致性;将一致性分析与安全审核的特殊需求相结合;在黑盒条件下构建即插即用的质量评估框架。

针对上述不足,本文提出基于CC机制的质量评估方法。基于CC机制的质量评估方法是唯一不需要模型内部参数的黑盒方法,具备完全无监督能力和模型无关性,在对抗性文本、复杂场景等挑战性场景下均表现最佳。通过分析多个图像文本识别模型在处理相同图像时的行为模式,本文发现正确识别的结果往往表现出显著的一致性,而错误结果则表现为明显的分歧。基于这一核心洞察,本文方法强调在黑盒、多模型、无监督前提下,输出一致性构建即插即用的质量评估方法,专门服务于内容安全审核场景。

2 基于一致性置信机制的质量评估方法

基于CC机制的质量评估方法的整体流程如图1所示,首先使用 N 个OCR模型对输入图像进行识别,计算输出对之间的相似性矩阵 S ;然后基于平均相似性计算一致性置信权重 W ;通过加权得到最终一致性置信度 δ_{cc} ;最后根据阈值 θ 判定输出质量等级(高质量或低质量)。该方法面向内容安全审核场景设计,核心目标是在无真值标签的情况下,准确评估图像文本识别结果的质量并识别可能存在错误的输出。该方法采用多模型共识策略,通过CC机制量化文本识别结果的可信度。该机制通过分析多个图像文本识别模型输出之间的相似性/一致性来评估结果的可信度,并特别考虑了安全审核场景下的分布特征。

2.1 一致性置信理论

CC机制建立在统计学习理论和集成学习理论的数学基础之上。

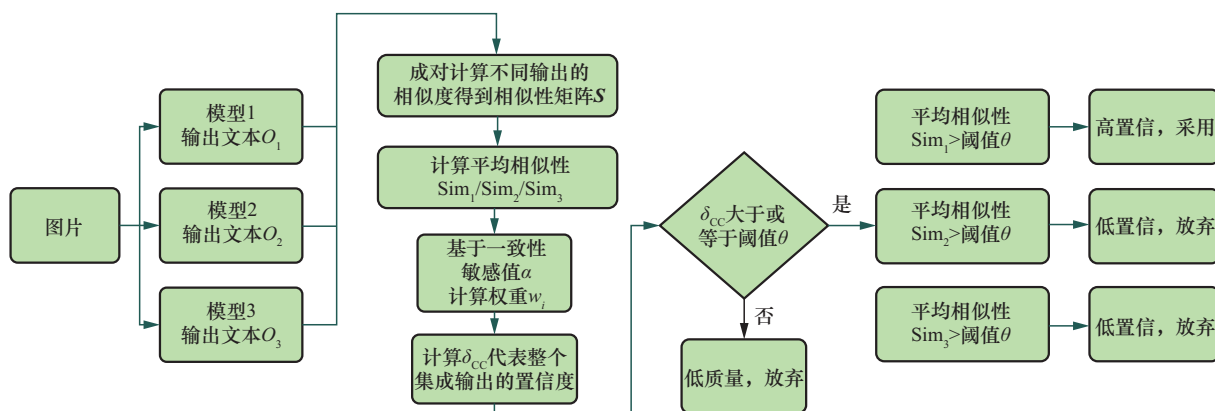


图1 基于CC机制的质量评估方法的整体流程

(1) 理论依据：在统计学习框架下，假设有 N 个独立训练的模型 $M_1, M_2, \dots, M_N$ ，对于输入图像 x ，每个模型产生预测 $y_i = M_i(x)$ 。根据统计决策理论，当多个独立模型对同一输入产生一致预测时，该预测的期望误差可以表示为：

$$E((y - \bar{y})^2) = \frac{1}{N} \sum_{i=1}^N E((y - y_i)^2) - \text{Var}(\bar{y}) \quad (1)$$

其中， $\bar{y}$ 是模型预测的平均值， y 是真实标签。这表明当模型输出方差较小（即一致性较高）时，集成预测的误差会降低。

(2) 一致性假设：在图像文本识别任务中，本文假设模型间的一致性程度与预测正确性存在正相关关系。数学上，对于任意输入 x ，定义模型间的一致性度量为 $C(x)$ ，质量评估函数 $Q(x)$ 满足：

$$Q(x) = f(C(x)), \quad f'(C) > 0 \quad (2)$$

即一致性越高，质量评估得分越高。对于清晰、标准的文本图像，由于任务难度较低，不同模型通常都能够收敛到正确答案，表现出高度一致性；而对于模糊、变形或故意干扰的文本，任务难度较高，模型更容易产生不同的错误输出，表现为低一致性。从贝叶斯角度，模型一致性可以视为对预测不确定性的间接估计。

(3) 在内容安全审核场景中的应用：恶意内容往往采用各种技术手段来规避检测，包括文本变形、字体特殊化、背景干扰等，这些对抗性因素会增加文本识别的难度和不确定性，导致不同模型产生不一致的输出结果。因此，通过量化分析模型间的一致性程度，本文可以在无监督条件下有效识别出可能存在问题的图像内容，为内容安全审核提供质量保障。

2.2 多模型体系架构设计

本文方法采用多模型共识策略进行质量评估。假设模型集合为 $M = \{M_1, M_2, \dots, M_N\}$ ，对于输入图像 I ，每个模型独立生成文本输出 $y_i = M_i(I)$ 。为确保一致性分析的有效性，模型集合的设计需要满足以下原则。

独立性与异构性：不同模型应具有不同的架构特点和训练数据来源，以确保模型间具有一定的独立性。实验发现，模型异构性有助于提升 CC 性能。模型集合的多样性可从架构、训练数

据、参数规模等维度体现。

互补性与多样性：模型集合应在不同场景下具有互补能力。实际应用中可以包括：通用型模型（如基于 Transformer 的大规模预训练模型）、专用型模型（针对特定字体或语言优化）、鲁棒型模型（抗噪声和变形能力强）、轻量型模型（计算效率高）等。这种多样性设计使得 M 能够覆盖更广泛的输入空间。

灵活可配置性：本文方法提供的是一种抽象的设计框架而非固定架构，用户可以根据应用场景（多语言、特殊字体等）、计算资源、成本预算、实时性要求等因素灵活配置模型集合 M 。这种适应性是本文方法的核心优势之一，使本文方法能够广泛应用于不同的实际场景。

通过这种异构且灵活的设计，不同模型对同一图像的敏感性和处理能力形成互补，本文方法能够从多个视角分析识别结果的一致性，从而获得更加全面和可靠的质量评估。

2.3 一致性量化方法

为了准确量化模型间的一致性，本文设计了多层次的相似性分析方法。首先，计算所有模型输出对 (O_i, O_j) 之间的成对相似性。考虑到图像文本提取任务的特点，采用编辑距离作为基础相似性度量，因为它能够有效捕捉字符级别的差异。

具体地，对于两个文本输出 O_i 和 O_j ，它们的相似性值定义为：

$$\text{Sim}(O_i, O_j) = 1 - \frac{\text{EditDistance}(O_i, O_j)}{\max(|O_i|, |O_j|)} \quad (3)$$

其中， $|O_i|$ 表示文本 O_i 的长度。这种归一化处理确保了相似性值的度量在 $[0, 1]$ ，便于后续分析。

其次，基于成对相似性，本文构建相似性矩阵 S ，其中 $S_{ij} = \text{Sim}(O_i, O_j)$ 。最后，本文计算每个模型输出与其他所有输出的平均相似性。

$$\overline{\text{Sim}}_i = \frac{1}{n-1} \sum_{j \neq i} \text{Sim}(O_i, O_j) \quad (4)$$

这个平均相似性反映了该输出与集体共识的一致性程度，较高的值表示该输出与其他输出高度一致，较低的值则表示存在显著分歧。

2.4 一致性置信权重机制

考虑到内容安全审核对可靠性的特殊要求，

本文引入一致性置信权重机制, 该机制能够根据输出的一致性程度动态调整其在最终决策中的影响力。一致性置信权重的计算式为:

$$w_i^{\text{cons}} = \frac{(\overline{\text{Sim}}_i)^\alpha}{\sum_{j=1}^n (\overline{\text{Sim}}_j)^\alpha} \quad (5)$$

其中, α 是一致性敏感参数, 其作用是放大高一致性输出的权重。当 α 较大时, 系统更加保守, 倾向于信任那些与集体共识高度一致的输出; 当 α 较小时, 系统更加宽容, 允许更大的输出多样性。

本文通过在 OCRBench 中的少量标注样本上进行测试发现, $\alpha=2.0$ 时效果较好, 能够有效平衡准确率和低质量识别率。不同应用场景可根据实际需求, 使用少量标注样本进行参数调优, 一般来说要求越高, 设置值越大。模型输出要求有一定随机性时, 可以设置一致性置信权重。

2.5 最终一致性置信度计算

基于一致性置信权重, 本文使用加权平均方法计算最终的一致性置信度 δ_{CC} 。

$$\delta_{\text{CC}} = \sum_{i=1}^n w_i^{\text{cons}} \cdot \overline{\text{Sim}}_i \quad (6)$$

其中, n 为平均相似性的数量。这个最终的一致性置信度具有明确的实际意义, 当 δ_{CC} 接近 1 时, 表示所有模型的输出高度一致, 可信度很高; 当 δ_{CC} 较低时, 表示模型间存在显著分歧, 输出的可信度较低, 需要进一步验证或处理。

为了便于实际应用, 本文将 δ_{CC} 归一化到 $[0,1]$, 使其能够直接作为置信度指标使用。归一化过程考虑了模型数量、输出长度等因素的影响, 确保了指标的稳定性和可比性。

2.6 阈值门控机制

基于计算得到的一致性置信度 δ_{CC} , 本文设计阈值门控机制来评估结果质量。

$$\text{Quality}(\delta_{\text{CC}}, \theta) = \begin{cases} \text{High}, & \delta_{\text{CC}} \geq \theta \\ \text{Low}, & \delta_{\text{CC}} < \theta \end{cases} \quad (7)$$

其中, θ 是质量阈值, 可根据具体应用场景的质量要求进行调整。当一致性置信度高于阈值时, 模型输出被认为是高质量 (High), 接受; 当一致性置信度低于阈值时, 模型输出被认为是低质量 (Low), 需要进一步验证或人工审核。这种二元分类机制为内容安全审核提供了清晰的质量

判断标准, 有效区分了可信和不可信的文本提取结果。

3 实验与结果分析

为验证本文方法的有效性, 在多个数据集上进行了系统的实验。本文旨在通过实验验证以下 3 个关键问题: CC 机制能否有效识别低质量的文本识别结果; 本文方法在内容安全审核的挑战性场景下的性能表现; 本文方法能否通过模型集成提升整体性能。

3.1 实验设置

本文使用仅推理的评估方式, 采用多样化的视觉语言模型, 遵循 vlmevalkit^[6] 的默认设置。编辑距离计算使用 Levenshtein^[25] 算法。所有实验模型运行设置 Temperature (模型推理时的温度设置) 为 0.0 (单次运行)。主要在 OCRBench 数据集^[7]上评估, 为了评估与人工判断图像文本识别质量的一致性, 本文人工标注了 1 000 个来自真实场景的图像文本识别结果 (使用 Qwen2.5-VL-72B 模型对不同类型的图片进行文本识别), 根据文本识别的准确性进行人工评分。

本文采用以下基线方法: VLM-as-Judge 分别使用 GPT-4o 和 Qwen2-VL-7B 两个 VLM 对人工标注数据集进行质量评估; 单模型直接使用单个 VLM 预测。

3.2 低质量结果筛选实验

本文首先在 OCRBench 数据集上进行实验, 以验证 CC 机制在无监督环境下识别低质量文本识别结果的有效性。对每个测试样本, 本文使用多个模型进行文本识别, 计算每个模型输出的 CC 阈值, 然后根据不同的 CC 阈值筛选输出结果, 观察被保留结果的平均准确率变化。

不同 CC 阈值下模型在 OCRBench 数据集上的平均准确率如图 2 所示。由图 2 可以看出, CC 阈值与模型准确性之间存在明显的正相关关系。随着 CC 阈值的提升, 被保留输出的平均准确率显著提高。这一趋势在本文测试的前 100 个 VLM 中普遍成立, 即使对于 GPT-4o 等高性能模型也同样有效。这表明 CC 机制能够在完全无监督的条件下, 通过分析模型间的一致性来有效评估文本识别质量, 为内容安全审核系统提供可靠的质量监控手段。

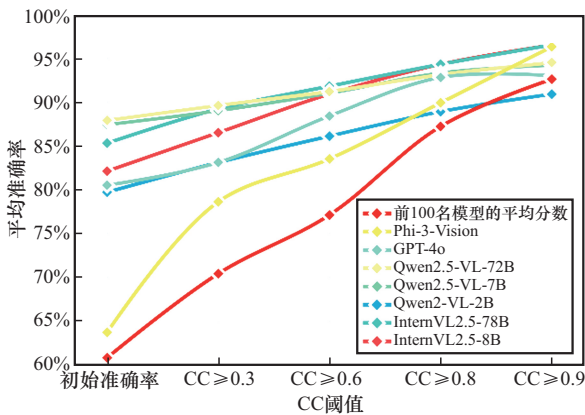


图2 不同CC阈值下模型在OCRBench数据集上的平均准确率

3.3 图像文本识别质量评估实验

为评估与人工判断的一致性，本文构建了1 000个真实场景样本的标注数据集。样本从中英文杂志、书籍、学术论文等来源随机抽取，使用Qwen2.5-VL-72B进行文本识别后由人工标注。4档评分标准：完美（0.9~1.0，与图像文本完全匹配）、良好（0.7~0.8，仅有小错误但不影响理解）、中等（0.4~0.6，包含明显错误或仅捕获部分文本）、差（0~0.3，大部分均错误或与图像文本无关）。判定依据为字符准确率和部分结构信息的保留程度。标注采用4人独立标注方式以确保可靠性。

VLM-as-Judge方法和本文方法在人工评分数据集上的F1得分比较见表1。从表1可以看

出，本文方法在所有质量区间都取得了更高的F1得分。在GPT-4o与Qwen2-VL-7B集成模型上本文方法的F1得分为49.9%~51.61%，而VLM-as-Judge方法为36.15%~39.86%，平均提升了15.46个百分点（42.76%）。特别地，在处理低质量文本（评分0~0.3）时，本文方法的F1得分达到67.76%~70.63%，显著高于VLM-as-Judge方法的F1得分40.00%~54.74%。这种在困难样本上的稳定表现，符合内容安全审核场景的实际需求。

3.4 多模型集成选择

本文进一步探讨所提方法通过多模型集成提升整体识别性能的有效性，特别是验证低成本开源模型组合能否超越高成本的单一闭源模型。本文在OCRBench数据集上设计了4组不同的模型组合方案，每组包含3~5个不同规模和架构的模型（参数规模为1~72 B），使用CC机制进行集成。

基于CC的多模型集成在OCRBench数据集上的性能见表2。由表2可知，最突出的是组合1，通过集成5个模型，将最高单模型得分从926分提升到958分，提升了32分。值得注意的是组合2，全部使用小于10 B参数的开源模型，集成后得分达到933分，超越了多个大规模闭源模型的单独表现。即使是组合4这样使用7B-8B参数量较小模型的方案，也通过CC集成得分达到897分，提升了23分。这些结果说明，本文方

表1 VLM-as-Judge 和本文方法在人工评分数据集上的F1得分（满分为100%）比较

文本质量人工评分	GPT-4o			Qwen2-VL-7B		
	VLM-as-Judge 方法	本文方法	提升	VLM-as-Judge 方法	本文方法	提升
完美	72.26%	67.90%	-6.04%	70.05%	74.42%	+6.24%
良好	24.82%	35.20%	+41.85%	19.82%	30.84%	+55.6%
中等	22.37%	28.75%	+28.52%	0	30.53%	+∞
差	40.00%	67.76%	+69.40%	54.74%	70.63%	+29.04%
平均	39.86%	49.90%	+25.19%	36.15%	51.61%	+42.76%

表2 基于CC的多模型集成在OCRBench数据集上的性能

组合序号	集成模型列表	单个模型得分/分	CC集成分数/分	提升分数/分
1	Ovis2-1B, Qwen2.5VL-72B, SenseChat, Step1V, Step1o	891, 879, 894, 886, 926	958	+32
2	Ovis2-1B, Ovis2-4B, Qwen2VL-7B, Qwen2.5VL-7B	890, 909, 843, 874	933	+24
3	Ovis2-4B, Qwen2.5VL-7B, Step1o	909, 874, 926	938	+12
4	InternVL2.5-8B, Qwen2VL-7B, Qwen2.5VL-7B	821, 843, 874	897	+23

法不仅能有效整合多个模型的优势，还能使低成本开源模型组合在性能上超越高成本闭源单一模型，这对实际应用的经济性和灵活性具有重要意义。

3.5 在挑战性场景下的表现

在内容安全审核场景中，挑战性场景下的性能表现尤为重要。本文基于 OCRBench v2 和网络收集的数据，构建了 4 类测试场景：对抗性文本场景（如 ASCII 艺术分类、文本检测定位，模拟恶意用户故意设计的对抗性内容）、复杂场景（如细粒度文本识别、手写文本提取、应用截图理解，模拟真实社交媒体场景）、标准文档场景（作为基准参考），以及违规词检测场景（在复杂场景中随机标注违规词，直接测试对内容安全审核任务的改善效果）。本文将所提方法与 GPT-4o、Gemini-Pro、InternVL2.5 等主流方法以及识别器输出投票错误减少（recognizer output voting error reduction, ROVER）集成方法进行了对比，不同方法对比结果见表 3。其中 ROVER 本身就是基于单个字符位置投票的方法，而其他传统的集成学习方法目前并不适用于大型视觉语言模型（large vision-language model, LVLM）。

表 3 不同方法对比结果

方法类型	异构集成	质量评估	黑盒	抗对抗
本文方法	是	有	支持	强
VLM-as-Judge 方法	否	有	支持	弱
单模型 OCR	否	无	—	弱
相对熵等	否	无	否	弱
ROVER	是	无	支持	弱

挑战性场景下不同方法的性能对比结果见表 4。由表 4 可知，本文方法在所有场景下都取得了最佳性能。在对抗性文本场景中得分 40.9 分，明显优于 GPT-4o 的 37.3 分；在复杂场景中得分 48.3 分；在违规词检测场景中得分 63.1 分，直接验证了对内容安全审核任务的改善；在标准文档场景中得分 62.4 分。这些结果表明，本文方法在处理对抗性文本、复杂场景和违规内容检测等挑战性场景时具有更强的适应能力和鲁棒性。

表 4 挑战性场景下不同方法的性能对比

场景类型得分/分	GPT-4o	Gemini-Pro	ROVER	本文方法
对抗性文本得分/分	37.3	38.4	36.7	40.9
复杂场景得分/分	35.5	50.4	38.2	48.3
标准文档得分/分	50.6	61.3	41.4	62.4
违规词检测得分/分	54.7	61.7	38.7	63.1

3.6 消融研究和实验分析

基于上述实验结果，本文对所提方法的性能进行了深入分析，探讨了影响系统性能的关键因素。

3.6.1 一致性阈值的影响分析

从图 2 可以看出，CC 阈值与模型准确性之间存在明显的正相关关系。随着 CC 阈值提升，模型输出的平均准确率显著提高，这一现象在测试的性能最优前 100 个 VLM（arg_top-100）中普遍存在。这表明 CC 机制能够有效识别和过滤低质量的文本识别结果，为无监督质量评估提供了可靠的技术手段。

3.6.2 多模型集成的效果分析

表 2 的实验结果展示了 CC 集成策略的显著优势。通过组合不同类型的模型，CC 不仅超越了单个高性能模型的性能，还实现了显著的性能提升。例如，组合 1 中通过集成 5 个模型，将最高得分从 926 分提升到 958 分，提升了 32 分。这证明了异构多模型设计在提升系统鲁棒性和准确性方面的重要作用。

3.6.3 质量评估的鲁棒性分析

表 1、表 4 的结果显示，本文方法在各种质量区间都表现出了良好的鲁棒性。特别是在处理低质量文本时，本文方法相比 VLM-as-Judge 方法有显著提升，平均提升幅度达到 42.76%。这表明 CC 机制在处理复杂和挑战性场景时具有更强的适应能力。

3.7 计算效率说明

本文提出的 CC 计算是直接在模型输出的文本上进行计算，计算时间在毫秒级别。根据图 2 和表 1 数据集的测试，单次编辑距离计算平均耗时 0.000 2 s，最高耗时 0.000 5 s（最高 5 个模型集成，输出字符长度 1 000），可在中央处理器

(CPU) 上运行, 相比模型推理时间可以忽略不计。

本文方法本身几乎没有成本(计算时间与算力消耗), 模型数量的增加对处理速度的影响极小, 图像处理时间只跟模型本身推理有关, 所提方法本身的额外成本极少。

4 结束语

本文针对内容安全审核场景下图像文本识别质量评估的关键挑战, 提出了基于CC机制的质量评估方法。该方法通过多模型一致性分析实现无监督质量评估, 具有黑盒评估、完全无监督、模型无关性三大核心优势。实验结果表明, 本文方法在质量验证方面相比基线方法平均提升42.76%, 在对抗性文本、复杂场景等挑战性场景下均表现最佳, 且能使低成本开源模型超越高成本闭源模型, 为内容安全审核提供了可靠且高效的技术解决方案, 具有重要的理论意义和实际应用价值。

参考文献:

- [1] 刘帆, 王凤美. 多模态内容安全审核系统构建思考[J]. 中国传媒科技, 2023(4): 149-153.
Liu F, Wang F M. Thoughts on the construction of multimodal content security audit system[J]. Media Science and Technology of China, 2023(4): 149-153.
- [2] 乔喆. 人工智能生成内容技术在内容安全治理领域的风险和对策[J]. 电信科学, 2023, 39(10): 136-146.
Qiao Z. Risks and countermeasures of artificial intelligence generated content technology in content security governance[J]. Telecommunications Science, 2023, 39(10): 136-146.
- [3] 李亚玲, 覃缘琪, 魏阙. 人工智能生成内容的潜在风险及治理对策[J]. 智能科学与技术学报, 2023, 5(3): 415-423.
Li Y L, Qin Y Q, Wei Q. Potential risks and governance strategies of artificial intelligence generated content technology[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(3): 415-423.
- [4] 王润民, 桑农, 丁丁, 等. 自然场景图像中的文本检测综述[J]. 自动化学报, 2018, 44(12): 2113-2141.
Wang R M, Sang N, Ding D, et al. Text detection in natural scene image: a survey[J]. Acta Automatica Sinica, 2018, 44(12): 2113-2141.
- [5] Zhang D Z, Yu Y H, Dong J H, et al. MM-LLMs: recent advances in MultiModal large language models[C]//Proceedings of the Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg: ACL, 2024: 12401-12430.
- [6] Duan H D, Yang J M, Qiao Y X, et al. VLMEvalKit: an open-source ToolKit for evaluating large multi-modality models[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM Press, 2024: 11198-11201.
- [7] Liu Y L, Li Z, Huang M X, et al. OCRBench: on the hidden mystery of OCR in large multimodal models[J]. Science China Information Sciences, 2024, 67(12): 220102.
- [8] Poznanski J, Rangapur A, Borchardt J, et al. olmOCR: unlocking trillions of tokens in PDFs with vision language models[PP]. arXiv (2025-07-02)[2025-08-24]. arXiv: arXiv.2502.18443.
- [9] Zhang R Y, Zhang H, Zheng Z D. VL-uncertainty: detecting hallucination in large vision-language model via uncertainty estimation[PP]. arXiv(2024-11-28)[2025-08-24]. arXiv: arXiv. 2411.11919.
- [10] Liu X O, Chen T J, Da L C, et al. Uncertainty quantification and confidence calibration in large language models: a survey[C]//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2. New York: ACM Press, 2025: 6107-6117.
- [11] Blecher L, Cucurull G, Scialom T, et al. Nougat: neural optical understanding for academic documents[PP]. V1. (2023-08-25)[2025-08-24]. arXiv: arXiv.2308.13418.
- [12] Wang B, Xu C, Zhao X M, et al. MinerU: an open-source solution for precise document content extraction[PP]. V1. (2024-09-27) [2025-08-24]. arXiv: arXiv.2409.18839.
- [13] Shen Z J, Lo K, Wang L L, et al. VILA: improving structured content extraction from scientific PDFs using visual layout groups[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 376-392.
- [14] Huang Y P, Lyu T C, Cui L, et al. LayoutLMv3: pre-training for document AI with unified text and image masking[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM Press, 2022: 4083-4091.
- [15] Zhong X, Tang J B, Jimeno Yepes A. PubLa_yNet: largest dataset ever for document layout analysis[C]//Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE Press, 2019: 1015-1022.
- [16] Liu C L, Wei H R, Chen J Y, et al. Focus anywhere for fine-grained multi-page document understanding[PP]. arXiv(2024-05-23)[2025-08-24]. arXiv: arXiv.2405.14295.
- [17] Chiang W L, Gonzalez J, Li D C, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena[C]//Proceedings of the Advances in Neural Information Processing Systems 36. California: Neural Information Processing Systems Foundation, Inc., 2023: 46595-46623.
- [18] Zhu L H, Wang X G, Wang X L. JudgeLM: fine-tuned large language models are scalable judges[PP]. arXiv(2023-10-26) [2025-08-24]. arXiv: arXiv.2310.17631.
- [19] Gu J W, Jiang X H, Shi Z C, et al. A survey on LLM-as-a-judge[J]. The Innovation, 2026: 101253.
- [20] Chang Y P, Wang X, Wang J D, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1-45.
- [21] Rando J, Zhang J, Carlini N, et al. Adversarial ML problems are getting harder to solve and to evaluate[PP]. arXiv(2025-02-04)

[2025-08-24]. arXiv: arXiv. 2502.02260.

- [22] Albladi A, Islam M, Das A, et al. Hate speech detection using large language models: a comprehensive review[J]. IEEE Access, 2025, 13: 20871-20892.

- [23] Lu X Y, Zhang T K, Meng C, et al. VLM as policy: common-law content moderation framework for short video platform[C]//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2. New York: ACM Press, 2025: 4682-4693.

- [24] Wang X Z, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models[PP]. arXiv (2022-03-21)[2025-08-24]. arXiv: arXiv.2203.11171.

- [25] Li Y J, Liu B. A normalized levenshtein distance metric[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1091-1095.

[作者简介]



张玉龙 (1999-), 男, 北京邮电大学硕士生, 主要研究方向为多媒体信息内容安全、人工智能安全、大模型。



梁天一 (1998-), 男, 华东师范大学博士生, 主要研究方向为多模态生成式模型。



张茹 (1976-), 女, 博士, 北京邮电大学教授、博士生导师, 主要研究方向为信息隐藏、人工智能安全。



刘功申 (1974-), 男, 博士, 上海交通大学教授、博士生导师, 主要研究方向为人工智能安全、自然语言处理、信息安全。