DOI: 10.19666/j.rlfd.202605095

基于视觉语言模型语义推理的火电厂安全监测告警复核研究

郭京，李军，高林，王林，周俊波，王明坤，王潇雨，张显荣
（西安热工研究院有限公司，陕西 西安 710054）

[摘　　要]【目的】火电厂智能监控系统中目标检测在视觉模糊边界的场景下误报频发，值班人员普遍面临告警疲劳，本文旨在解决该问题。【方法】作者提出一种融合目标检测与视觉语言模型语义推理的告警复核方法：目标检测模型保持高召回率完成初筛，视觉语言模型依托电力安全规程进行语义复核。针对安全帽佩戴、工作服穿着、烟雾和明火 4 类风险，按照核心词基线、领域定义、思维链推理、多模态示例 4 阶段构建渐进式提示词方案。从某火电厂生产现场采集告警样本，构建了 1 068 张垂直领域评测集，并在 3 款国产视觉语言模型上进行对比验证。【结果】结果显示，最优提示词配置下复核准确率从基线的 46.73%提升至 81.66%，F_1 值从 41.54%升至 86.36%，告警过滤效率由 0.242 升至 0.631，值班终端处置告警总量下降 33.6%，终端侧误报占比由 31.9%降至 12.5%。【结论】该方法可降低火电厂智能监控系统的误报水平，并为类似场景的应用提供参考。

[关　键　词]视觉语言模型；提示词工程；目标检测；智慧电厂

[引用本文格式]（本段作者勿动）作者姓名. 中文标题[J]. 热力发电, 年, 卷(期): 起始页码-终止页码.　作者姓名. 英文标题[J]. Thermal Power Generation, 年, 卷(期): 起始页码-终止页码.

Research on VLM semantic reasoning based alarm re-examination for thermal power plant safety monitoring

Guo Jing, Li Jun, Gao Lin, Wang Lin, Zhou Junbo, Wang Mingkun, Wang Xiaoyu, Zhang Xianrong
(Xi'an Thermal Power Research Institute Co., Ltd., Xi'an 710054, China)

Abstract: [Objective] Object detection in intelligent surveillance systems at thermal power plants generates frequent false alarms in scenes with ambiguous visual boundaries, such as a partially occluded safety helmet or a small flame that is easily confused with reflected light. Because every alarm has to be confirmed by hand, these false positives subject control-room operators to chronic alarm fatigue and gradually increase the risk that a genuine hazard is overlooked. This paper aims to solve this problem. [Methods] This paper proposes an alarm re-examination method that combines an object detector with a vision-language model (VLM) for semantic reasoning. Unlike a detector, which classifies an isolated bounding box, a VLM can weigh the surrounding scene against the wording of the relevant safety rule, which makes it better suited to the borderline cases that produce most false alarms. In the proposed pipeline, the detector preserves a high recall rate for initial screening so that few real hazards are missed, and the alarms it raises are then passed to the VLM, which performs semantic verification against power industry safety codes before any alarm reaches the operator. The study addresses four typical risk categories in plant operation: safety helmet wearing, work-uniform compliance, smoke, and open flame. For these categories, a progressive prompt scheme was developed in four stages—a keyword baseline, domain-specific definitions, chain-of-thought reasoning, and multimodal exemplars—with each stage supplying the model with more task knowledge than the previous one. A domain-specific evaluation set of 1 068 alarm images was collected from the production site of a thermal power plant, and three domestic VLMs were compared on this set so that the influence of both the

收 稿 日 期：2026-05-26　修 回 日 期：2026-07-01　接 受 日 期：2026-07-29　网络首发日期：XXXX-XX-XX
基 金 项 目：中国华能集团有限公司总部科技项目（TW-26-HJK01）
Supported by: Science and Technology Project of China Huaneng Group Co., Ltd. (TW-26-HJK01)
第一作者简介：郭京（2001），男，硕士研究生，主要研究方向为智慧电厂，864066131@qq.com。
通信作者简介：高林（1981），男，正高工，博士，主要研究方向为电站自动控制与智能优化技术，58560930@qq.com。

prompt design and the model could be assessed. [Results] Under the optimal prompt configuration, review accuracy increased from a baseline of 46.73% to 81.66%, the F_1 value from 41.54% to 86.36%, and the alarm-filtering efficiency (AFE) from 0.242 to 0.631. At the operator terminal, the total alarm volume requiring handling decreased by 33.6%, and the terminal-side false-alarm ratio fell from 31.9% to 12.5%. The progressive prompt design accounts for most of this improvement, as accuracy rose stage by stage from the keyword baseline, while the three models behaved consistently under the same prompts. [Conclusion] Overall, pairing a high-recall detector with VLM-based semantic verification effectively reduces false alarms in thermal power plant surveillance systems while preserving detection coverage; because it relies only on domestically available models and a modest amount of on-site data, the method offers a practical reference for similar industrial applications.

Key words: vision-language model; prompt engineering; object detection; smart power plant

火电厂生产现场设备密集、作业交叉，安全管控压力大。随着智慧电厂建设的推进[1]-[2]，基于深度学习的视觉感知技术已在国内多家火电企业部署应用[3][4]，用于违章作业、防护缺失等风险的自动识别与预警。但为保障高召回率，此类系统输出的告警中夹杂大量误报。某火电厂集控室单日推送告警可达数百条，有效占比有限，误报多源于光照、遮挡及相似目标干扰。告警长期低质，会钝化值班人员、拉长响应，削弱系统防护效能。因此，抑制误报成为系统有效运行的关键。

现有智能监控系统多以单阶段目标检测模型为核心[5][6]，对视频流实时推理并上报原始告警。这类模型在公共数据集上精度良好，但部署到火电厂工业场景后性能明显下降，原因涉及模型能力与数据分布两方面：模型能力上，相似类别边界难以划分，如真实烟雾与冷却塔水汽、输煤粉尘等干扰缺少稳定的视觉界限；违章语义与检测语义存在粒度错位，工作服检测往往只给出“着装/未着装”的粗判断，难以识别外套敞开等细粒度违章。数据分布上，样本采集与标注成本高，前端模型多在通用或有限的电力数据上训练，对现场泛化不足。仅优化检测器自身难以解决，需在系统架构中引入语义推理。

视觉语言模型（vision-language model，VLM）具备跨模态语义理解与推理能力，契合合规复核所需的细粒度语义；将其接在检测器之后对告警逐条复核，有望在前端高召回初筛的基础上进一步滤除误报。但 VLM 在火电厂多类作业风险上的复核能力尚不明确，电力安全规程如何转译为其可执行的判定规则也缺少研究，其表现还受模型选型与提示词设计的影响。这些问题都有待系统评测来厘清。

针对上述不足，本文的主要贡献如下：

1）提出面向火电厂告警复核的“检测初筛—规程语义复核”级联范式，把电力安全规程转译为含告警触发、误报排除与适用场景约束的结构化判定规则，作为后端合规复核的可执行依据。

2）设计 4 阶段渐进式提示词方案，逐层增强 VLM 对火电厂作业风险的领域判定能力。

3）基于某火电厂现场采集、经双盲复核的 1 068 张告警图像，构建覆盖安全帽佩戴、工作服穿着、烟雾、明火 4 类风险的复核评测集；在豆包（Doubao）、智谱（GLM）、千问（Qwen）上进行对比实验。结果表明，在前端高召回初筛之上，级联复核能明显降低终端告警总量与误报占比。

1 相关工作

火电厂安全监控的早期方法多以手工特征结合传统分类器，为安全帽等目标设计颜色、形状或纹理特征，再由支持向量机、Adaboost 等判别，鲁棒性有限[7]。随着卷积神经网络的发展，以 YOLO 系列为代表的单阶段检测器凭借端到端结构与实时性成为主流[8][9]，被用于安全帽佩戴[10][11]、工作服合规[12]、烟火[13]等检测，并经注意力机制、迁移学习与知识蒸馏、轻量化等改进。但这类改进始终围绕检测精度，难以支撑对作业行为合规性的判断，语义复核由此进入研究视野

VLM 通过多模态表示学习将视觉与语言信号对齐到同一语义空间，可支持图像理解等任务[14]。早期工作如 CLIP 以双塔结构通过图文对比学习获得通用匹配能力[15]，近期则演进为以大语言模型为推理中枢的多模态架构，代表性模型有 LLaVA[16]、InternVL[17]、CogVLM[18]等；面向中文场景的千问（Qwen-VL）[19]等国产模型也相继发布。在此基础上，已有研究将检测器与 VLM 组合，由 VLM 对检出区域做语境化判定并调整置信度阈值，应用于车辆火灾[20][21]、电扶梯[21]、建筑工地[22]等安全监测场景。但相关研究多面向单一场景，在火电厂多类作业风险上的复核能力及其与检测系统的协同方式尚缺乏系统评估。

在上述协同架构中稳定输出可靠判定，提示词

设计尤为关键。提示词工程通过设计输入提示、在不更新参数的前提下引导大模型完成任务，按示例数量可分为零样本与少样本提示[23]；思维链提示[24]则在二者之上叠加使用，先推理再作答提升多步推理性能。面向视觉—语言任务，已有研究以建筑缺陷识别为对象，经对照实验归纳出 3 条提示词设计原则[25]：领域专用定义优于通用定义，完整句式优于核心词序列，图文结合优于单一文本。这些原则为工业场景提供了参考，但文献[25]面向建筑等场景，未涉及电力安全规程向判定规则的转译，本文正是从这里切入。

2 方法与实验

2.1 整体架构

火电厂智能监控系统整体架构如图 1 所示，自下而上由数据采集与传输层、边缘感知与初级检测层、多模态大模型级联推理层、云端决策与闭环反馈层、应用服务与展示层构成。其中前 2 层属于已有部署系统，承担多源视频接入、流媒体处理与基于轻量化检测网络的初级告警识别等任务；后 2 层负责告警分发、闭环优化与业务集成。

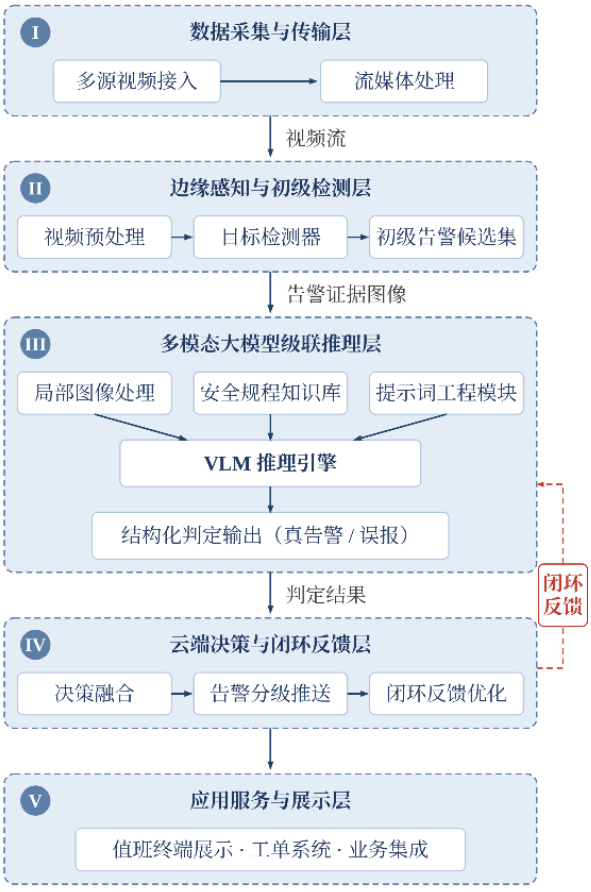


图 1 火电厂智能监控系统整体架构

Fig.1 Overall architecture of the thermal power plant intelligent monitoring system

本文工作聚焦于中间的多模态大模型级联推理层，针对前端检测层输出的告警结果引入基于 VLM 的语义复核环节，将原有的一次性告警推送过程重构为“实时检测—语义复核—决策推送”的处理流程。

受现场工程条件约束，本文所用数据仅包括前端系统自动生成的告警图像，图 2 示出灰库区域作业现场告警图像的典型示例。该类图像由前端检测器在图像上标注了红色矩形检测框，框内为其认定的疑似风险目标。本文直接以带检测框的告警图像作为复核环节的输入。

图 2 带检测框的告警图像示例

Fig.2 Example of an alarm image with detection bounding boxes

复核环节由 2 层级联构成。前端目标检测模型对视频流推理并输出告警图像；为保证召回率，前端模型通常设置较低的判定阈值，告警集中夹杂较多疑似误报。后端语义复核层将带检测框的告警图像与对应类别的提示词输入 VLM 进行推理，输出“真告警”或“误报”判定；真告警事件推送至值班终端，误报事件由系统抑制并归档。复核过程可表示为：

$$d_{final} = \mathrm{VLM}(I_{overlay}, P_k), k \in \{1,2,3,4\} \quad (1)$$

式中：$I_{overlay}$ 表示带检测框的告警图像；P 表示第 k 类风险对应的提示词，k 取值 1 至 4 分别对应安全帽佩戴、工作服穿着、烟雾、明火；$d_{final} \in$ {真告警,误报}为最终判定结果。

判定规则是连接电力安全规程与 VLM 推理的中间层。本文以相关规程[26]及电厂安全管理细则为依据，结合火电厂现场作业实际场景，对安全帽佩戴、工作服穿着、烟雾和明火 4 类典型风险设计了结构化判定规则，详见表 1。

表 1 4 类典型风险的结构化判定规则
Tab.1 Structured judgment rules for the four typical risk categories

风险类别	告警触发条件	误报排除条件	适用场景约束
安全帽	头部未佩戴安全帽；安全帽提在手中或悬挂于身边；帽檐反戴	人员已佩戴（任何颜色，含黑色）；堆积物、倒影等类人形误检；画面边缘不完整人形残影	强制佩戴区域内的作业人员
工作服	未着工装；未穿反光马甲；上衣明显缺失或被便装替代；衣袖卷起	已穿反光马甲（即使部分遮挡）；驾驶室内司机；搬运物遮挡无法确认情形	正式作业场地
烟雾	明显扩散性烟雾；有源头；烟团向上或向外飘动	背景干扰（云朵、远景模糊区域）；视觉相似物（水雾、蒸汽、粉尘、灰色设备或工服）；运动残影	工艺背景下合理蒸汽排放区域
明火	开放性明火；跳动火焰；焊接或切割产生的飞溅火花	静态高亮物体（黄色警示物、照明灯具、车灯、信号灯）；非燃烧性反光（金属反光、夕阳光晕）；节日烟花	动火作业区的合规焊接与气割火花

为评估提示词对模型效果的影响，本文设计渐进式提示词方案（P1—P4），逐级验证提示词的增益。P1 为基线提示词，P2 增加完整句式，P3 引入思维链推理，P4 进一步加入多模态示例。表 2 列出每种提示词方案的具体构成要素。

表 2 渐进式提示词 P1-P4 方案的构成要素
Tab.2 Component elements of the progressive prompt schemes P1——P4

方案	提示表述	领域知识	思维链（CoT）	多模态示例(Few-shot)
P1	关键词	—	—	—
P2	完整自然语句	电力规程领域定义	—	—
P3	完整自然语句	电力规程领域定义	思维链推理	—
P4	完整自然语句	电力规程领域定义	思维链推理	多类别正负配对示例

2.2 数据集构建

本文实验数据来源于某火电厂智能视频监控系统在实际运行环境下的告警记录。该电厂已部署基于深度学习的目标检测告警系统，覆盖主厂房、输煤栈桥、化学水处理区及升压站等典型作业区域。经电厂授权后，采集一段连续时间窗口内由系统自动推送至值班终端的全部告警图像。所收集图像均带有目标检测系统标注的红色检测框，其中涉及的敏感信息均已脱敏处理，以满足数据合规与隐私保护要求。

同一风险事件在持续触发期间会生成多帧告警推送，原始数据中存在一定比例的重复或近似重复样本，为降低样本冗余对评测结果的影响，本文按以下步骤进行清理与筛选：结合时间戳与摄像头编号对连续告警进行事件级归并，同一事件内仅保留画面完整、目标清晰的一帧；随后对筛选结果进行人工复核，剔除图像质量较差、目标尺度过小或严重过曝等不适于判读的样本。经处理后，共获得高质量告警图像 1 068 张，构成本文数据集。

为获得可靠的真值标签，本文邀请 2 名具有现场安全管理经验的人员对全部图像进行双盲独立标注，将每张图像标注为“真告警”或“误报”。为评估标注一致性，引入 Cohen κ 系数，其定义为：

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (2)$$

式中：p_o 为观察一致率，p_e 为偶然一致率。两位标注人员的 Cohen κ 系数为 0.87，表明标注结果具有较高一致性。对于存在分歧的样本（占比 5.6%），由第三方专家进行仲裁以确定最终标签。

按约 20%：80%的比例将 1 068 张图像划分为开发集（212 张）与评测集（856 张）。开发集用于提示词方案的迭代验证，并为 P4 方案提供多模态示例图像；评测集用于最终对比实验。考虑到不同风险类别及其真伪标签分布存在不均衡，本文以“风险类别×真值标签”为分层依据进行抽样，使开发集与评测集在类别构成及标签比例上均与全集保持一致，数据分布如表 3 所示。

表 3 数据集分布统计
Tab.3 Dataset distribution statistics

风险类别	告警样本（开发/评测/合计）	误报样本（开发/评测/合计）	小计	误报占比
安全帽佩戴	55/222/277	37/149/186	463	40.2%
工作服穿着	12/48/60	6/26/32	92	34.8%
烟雾	43/171/214	12/49/61	275	22.2%
明火	35/142/177	12/49/61	238	25.6%
合计	145/583/728	67/273/340	1068	31.9%

2.3 实验设置与评价指标

选取 3 款国产 VLM 作为评测对象，具体信息见表 4。3 款模型在底层架构、训练数据规模与多模

态对齐策略上各具特点，所有模型均以推理服务的形式调用，temperature 统一设为 0.1、top_p 统一设为 0.9，以确保对比条件的一致；为抑制随机性，对每张评测图像独立调用 3 次并按多数投票确定最终判定，全部实验在统一时间窗口内完成。实验分组上，模型 A 承担 P1—P4 渐进式提示词方案的全流程评测，用于考察各方案的增益，模型 B 与模型 C 仅参与 P3 与 P4 方案的对比，用于评测思维链推理与多模态示例策略在不同模型上的差异。

表 4 评测中使用的 3 款国产 VLM
Tab.4 Three domestic VLMs used in the evaluation

编号	模型名称	厂商	模型版本	参与方案
A	豆包视觉大模型	字节跳动火山引擎	doubao-seed-2-0-pro-260215	P1, P2, P3, P4
B	GLM-4.6V	智谱 AI	glm-4.6v	P3, P4
C	Qwen3-VL-Plus	阿里云通义千问	qwen3-vl-plus	P3, P4

告警复核任务可形式化为“真告警”与“误报”的二分类判定问题。考虑到工程场景中误报与漏报的代价存在非对称性，本文采用 A、P、R、F_1、R_{FA} 与 E_{AF} 等 6 项指标对模型性能进行综合评估。设 N_{TP}、N_{FP}、N_{FN}、N_{TN} 分别表示真告警的正确识别、误报的错误识别、真告警的漏判以及误报的正确识别，则各指标定义如下：

$$A=\frac{N_{TP}+N_{TN}}{N_{TP}+N_{FP}+N_{FN}+N_{TN}} \tag{3}$$

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}} \tag{4}$$

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}} \tag{5}$$

$$F_1=\frac{2PR}{P+R} \tag{6}$$

$$R_{FA}=\frac{N_{TN}}{N_{TN}+N_{FP}} \tag{7}$$

$$E_{AF}=R_{FA}\times R \tag{8}$$

式中：R_{FA}（FARR, False Alarm Recognition Rate）用于衡量模型对误报的识别能力；E_{AF}（AFE, Alarm Filtering Efficiency）用于综合反映模型在过滤误报与保留真告警两方面的平衡性能。上述指标均基于评测集 856 张图像的逐样本预测结果汇总计算。

2.4 实验结果与分析

在评测集上测试渐进式提示词方案对 VLM 复核性能的影响，并对比 3 款模型的响应特性，综合结果见表 5，消融轨迹与模型间横向对比分别见图 3、图 4。

表 5 3 款 VLM 在 P1—P4 方案下的综合性能对比
Tab.5 Overall performance comparison of three VLMs under P1 – P4 schemes

模型	方案	准确率/%	精确率/%	召回率/%	F_1/%	FARR/%	AFE
A 豆包	P1	46.73	82.23	27.79	41.54	87.18	0.242
A 豆包	P2	68.93	87.83	63.12	73.45	81.32	0.513
A 豆包	P3	75.12	87.30	74.27	80.26	76.92	0.571
A 豆包	P4	81.66	87.50	85.25	86.36	73.99	0.631
B GLM-4.6V	P3	72.78	84.45	73.58	78.64	71.06	0.523
B GLM-4.6V	P4	68.69	79.55	72.73	75.99	60.07	0.437
C 千问 VL	P3	59.81	94.10	43.74	59.72	94.14	0.412
C 千问 VL	P4	66.47	95.68	53.17	68.36	94.87	0.504

由模型 A 的完整消融轨迹可见，提示词要素的逐步叠加带来了综合性能的单调递增。P1 基线下准确率仅 46.73%、F_1 仅 41.54%，模型在任务描述信息不足的条件下倾向保守判定，大量真告警被误判为误报；引入完整句式与电力规程领域定义后，P2 准确率跃升至 68.93%，F_1 提升至 73.45%，这一增

益表明领域专用定义与完整句式能显著促进 VLM 对任务的理解。P3 在 P2 基础上引入思维链推理，准确率、F_1 与 AFE 均进一步提升。P4 在 P3 基础上引入多类别正负配对示例后，准确率达到 81.66%，F_1 达到 86.36%，召回率提升至 85.25%，综合指标 AFE 由基线的 0.242 提升至 0.631。

3 款模型在 P3、P4 方案下的响应表现明显的模型异质性。模型 A 由 P3 到 P4 稳步提升，与多模态示例的预期增益方向一致；模型 C 同期亦有提升，但整体指标受限于较为保守的判定，精确率高达 95.68%而召回率仅 53.17%；模型 B 则出现回落，P4 准确率较 P3 下降 4.09 百分点，FARR 由 71.06% 降至 60.07%，约 30 张原被正确识别为误报的样本在 P4 下被改判为告警。上述差异表明，多模态示例对 VLM 性能的影响因模型架构与训练策略而异。

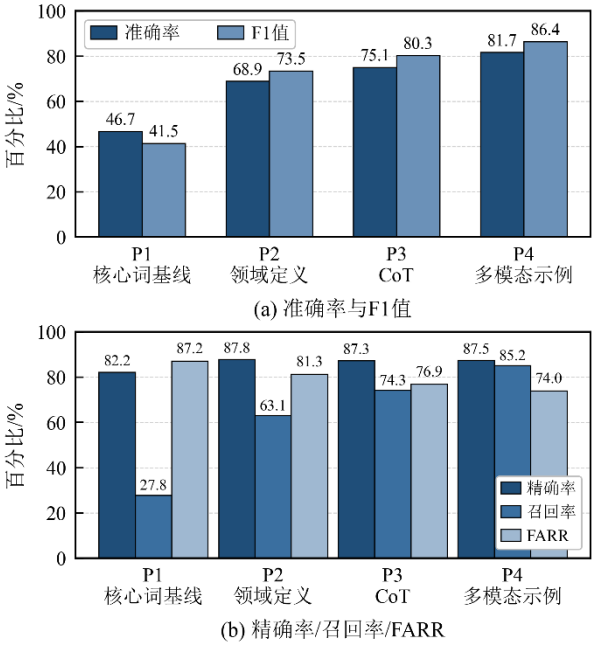
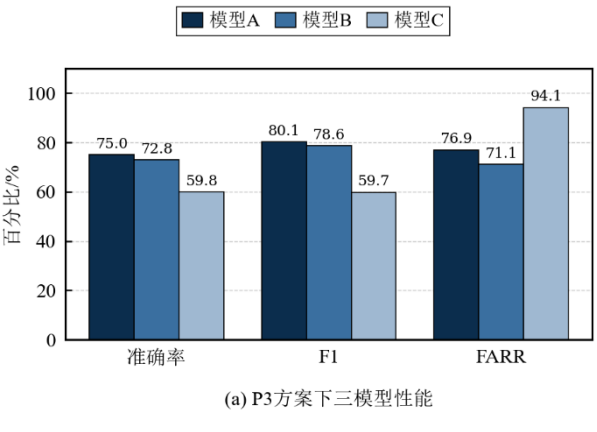


图 3 模型 A 在 P1—P4 方案下的综合性能对比
Fig.3 Overall performance of model A under P1 – P4 schemes



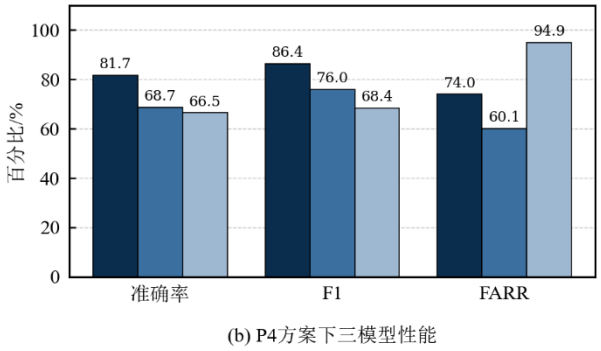


图 4 3 款 VLM 在 P3/P4 方案下的横向性能对比
Fig.4 Cross-model performance comparison of three VLMs under P3/P4

图 5 以 Recall 为横轴，FARR 为纵轴展示了 3 款模型的演进轨迹，散点大小正比于 AFE 值，虚线为 AFE 等值线，右上方为性能理想区域。模型 A 沿 P1 至 P4 的轨迹持续向右下方展开，召回率提升 57.46 百分点的同时 FARR 仅下降 13.19 百分点，AFE 单调递增，表明渐进式提示词方案能够在维持较高误报过滤能力的同时持续提升真告警的识别率；模型 B 呈现与模型 A 相反的演进方向，P3 至 P4 轨迹同时向更低 FARR 与更低 AFE 区域移动，印证了表 5 中模型 B 在引入多模态示例后性能回落的定量结果；模型 C 始终位于高 FARR、低召回的左上区域，P3 至 P4 轨迹近乎水平右移，提示该模型对多模态示例的响应主要体现在召回率的小幅提升，对误报过滤能力影响有限。

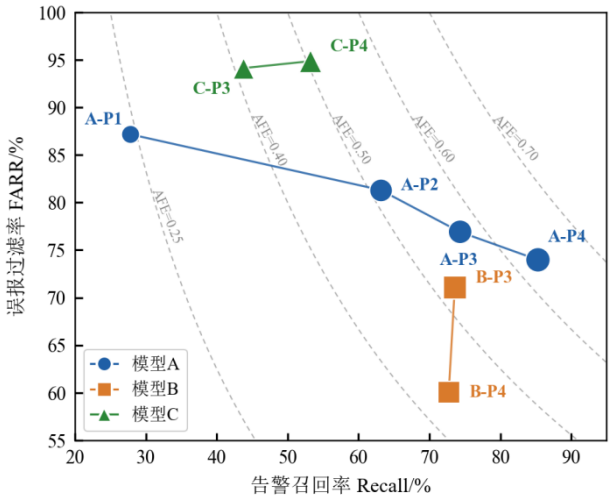


图 5 3 款模型在 FARR-Recall 工程权衡平面上的分布轨迹
Fig.5 Trajectories of the three models on the FARR-Recall trade-off plane

为进一步呈现 3 款模型在边界样本上的分化行为，本文选取一例典型误报案例，展示 3 款模型在 P4 方案下的判定结果及其关键推理文本，图 6 为带检测框的图像，3 款模型在该案例下的判定细节与

匹配规则汇总于表 6。

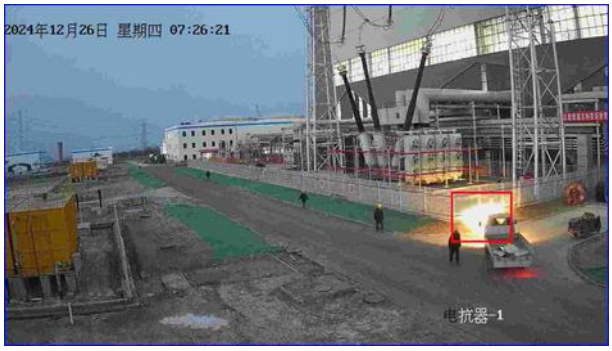


图 6 典型误报案例
Fig.6 A typical false-alarm case

如表 6 所示，3 款模型在相同输入下因视觉证据抽取重点与判据匹配策略的不同而出现最终判定分化，与图 4、图 5 反映的指标层面差异相互印证。

为刻画所提方法的能力边界，围绕模型 A 在 P4 方案下的最优配置，从风险类别维度开展分项统计，结果如图 7 所示。

表 6 3 款 VLM 在典型案例下的判定结果与匹配规则对比
Tab.6 Decision results and matching rules of three VLMs on a typical case

项目	模型 A	模型 B	模型 C
真值	误报（车辆车灯）		
视觉证据	红框内为车辆的车头灯发出的高亮光线，属于交通工具的车灯，形态固定无火焰跳动特征	红框内存在明亮火光，呈现跳动、闪烁的橙红色火焰形态，伴随有人员作业场景	红框内为白色小型货车尾部，车后方有强烈橙黄色光亮区域，呈不规则团状，边缘模糊、有向上扩散趋势
匹配规则	火焰类误报条件-交通工具的车灯、信号灯	火焰类告警条件：框内存在明亮的火光，具有跳动、闪烁、橙红色或黄白色火焰形态	火焰告警条件：存在明亮火光，具有跳动、闪烁、橙红色或黄白色火焰形态；参考示例 7 中动火作业飞溅火花情形
判定结果	误报	告警	告警

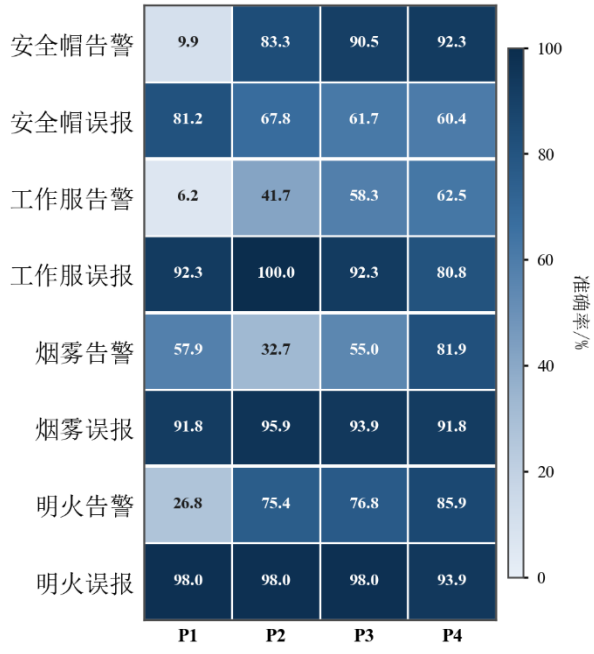


图 7 模型 A 在 4 类风险×8 类子集上的准确率分布
Fig.7 Accuracy distribution of model A on four risk categories × eight subsets

由图 7 可见，告警类子集准确率随 P1 至 P4 显著提升，其中安全帽告警由 9.9%提升至 92.3%，明火告警由 26.8%提升至 85.9%，烟雾告警由 57.9%提升至 81.9%，体现了渐进式提示词工程对真告警识别能力的直接增益；误报类子集准确率在 P1 阶段即普遍较高，系模型偏向保守判定所致，随着提示词阶段推进略有下降，是告警判定信心提升后假阳性增多的必然代价，但告警侧收益显著大于误报侧损失。工作服告警在 4 类风险中表现最弱，P4 准确率仅 62.5%，反映出模型对反光条、工装细节等小像素比例视觉概念的识别存在固有瓶颈。

为定位上述瓶颈的具体来源，本文对模型 A 在 P4 方案下的 157 个误判样本进行人工归因分析，并将其归入预定义的 6 类错误模式，结果如图 8 所示。

烟雾与水汽、粉尘混淆占 32.1%，为最大单一误判类别，主要发生在冷却塔顶部水汽与输煤栈桥粉尘等干扰场景；明火与合规火源、光源混淆占 24.8%，主要来源于动火作业现场对合规电焊火花与违章明火的边界判定以及夜间黄色灯光被误识别为火焰的情形；细粒度违章漏判占 19.6%，集中于工作服反光条、安全帽帽带等小区域视觉细节，是工作服告警表现偏低的直接来源；其余光照与遮挡相关误判、画面边缘人形残影及其他类别合计占 23.6%。上述归因表明，视觉相似边界与细粒度视觉概念仍是当前 VLM 在火电厂安全监控任务上的主要性能瓶颈，相关挑战难以仅通过提示词工程化解。

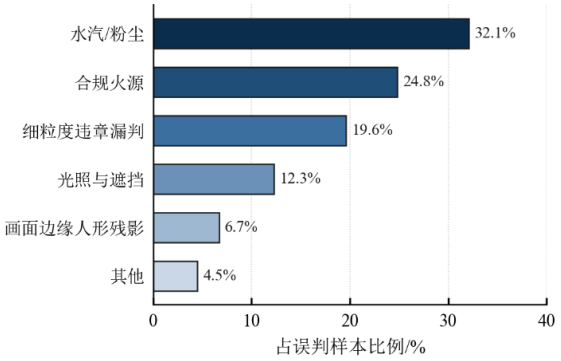


图 8 最优配置下误判样本的错误分布与能力边界分析
Fig.8 Error distribution of misjudged samples under optimal configuration

将无大模型语义复核作为基线情形，与最优配置在告警召回与误报过滤 2 个核心指标上进行对照。基线情形下，评测集被全量推送至值班终端，由人工逐一判断。引入最优配置后，Recall 为 85.25%，FARR 为 73.99%。从值班终端处置负担看，处置的告警总量下降 33.6%，终端侧误报占比由 31.9%降至 12.5%，每 10 条告警中有效告警占比由约 7 条提升至 9 条，从数量与质量 2 个维度缓解了告警疲劳。

3 讨论

面向火电厂安全监测系统误报率偏高的问题，本文所提框架将前端检测的实时初筛能力与 VLM 的语义推理能力相结合，在某火电厂真实生产数据上验证了后端语义复核对告警误报的有效抑制。结果表明，不同 VLM 对同一渐进式提示词方案呈现出差异化响应；因而在部署前，须将模型与提示词方案作为耦合整体进行评估。渐进式提示词方案的消融结果进一步表明，领域专用定义与完整句式是激活 VLM 垂直任务理解能力的关键要素，而思维链推理与多模态示例则在保障工程可用性方面发挥了不可忽视的作用。

本文存在以下局限。其一，数据集来源于单一电厂的特定时间窗口，所反映的场景代表性有限，框架在跨厂、跨场景条件下的泛化能力尚待进一步验证；其二，对于漏报代价极高的场景存在安全隐患，如何在召回率与误报识别率之间实现自适应权衡是后续工作的核心方向；其三，对视觉相似边界与细粒度视觉概念的识别仍是当前 VLM 的瓶颈，其突破有赖于基础模型能力的演进。

[参 考 文 献]

[1] 王翔宇,陈武晖,郭小龙,等.发电系统数字化研究综述[J].发电技术,2024,45(01):120-141.

[2] 韩旭东,王仲.火电机组智能辅助诊断系统研究[J].热力发电,2021,50(7):8-14.DOI:10.19666/j.rlfd.202012285.
HAN Xudong, WANG Zhong. Research on intelligent assistant diagnosis system for thermal power units[J]. Thermal Power Generation, 2021, 50(7): 8-14.

[3] 赵振兵,冯烁,席悦,等.大模型时代：电力视觉技术新起点[J].高电压技术,2024,50(5):1813-1825.

[4] 金鑫,程凌森,赵亮,等.基于改进 YOLOv5 模型的智能变电站目标违规行为检测方法研究[J].电测与仪表,2024,61(8):179-185.

[5] 邵延华,张铎,楚红雨,张晓强,饶云波.基于深度学习的 YOLO 目标检测综述[J].电子与信息学报,2022,44(10):3697-3708.
SHAO Yanhua, ZHANG Duo, CHU Hongyu, ZHANG Xiaoqiang, RAO Yunbo. A Review of YOLO Object Detection Based on Deep Learning[J]. Journal of Electronics & Information Technology, 2022, 44(10): 3697-3708.

[6] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 779-788.

[7] 高腾,张先武,李柏.深度学习在安全帽佩戴检测中的应用研究综述[J].计算机工程与应用,2023,59(6):13-29.

[8] Zou Z, Chen K, Shi Z, et al. Object detection in 20 years: A survey[J]. Proceedings of the IEEE, 2023, 111(3): 257-276.

[9] 曹家乐,李亚利,孙汉卿,等.基于深度学习的视觉目标检测技术综述[J].中国图象图形学报,2022,27(6):1697-1722.
Cao Jiale, Li Yali, Sun Hanqing, et al. A survey on deep learning based visual object detection[J]. Journal of Image and Graphics, 2022, 27(6): 1697-1722.

[10] 刘雅洁,伊力哈木·亚尔买买提,席凌飞,等.改进 YOLOv5s 的安全帽佩戴检测算法研究[J].计算机工程与应用,2023,59(20):184-191.

[11] 张国鹏,周金治,马光岑,贺浩洋.改进 YOLOv8 的轻量化安全帽佩戴检测算法[J].电子测量技术,2024,47(17):147-154.

[12] 王广乐,周亚同,王钊. 面向电力作业不规范穿戴检测的轻量化模型[J]. 激光与光电子学进展,2025,62(6):197-206.

[13] Geng X, Su Y, Cao X, et al. YOLOFM: An improved fire and smoke object detection algorithm based on YOLOv5n[J]. Scientific Reports, 2024, 14: 4543.

[14] Zhang J, Huang J, Jin S, et al. Vision-language models for vision tasks: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 5625-5644.

[15] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning (ICML). 2021: 8748-8763.

[16] Liu H, Li C, Wu Q, et al. Visual instruction tuning[C]//Advances in Neural Information Processing Systems (NeurIPS). 2023, 36: 34892-34916.

[17] Chen Z, Wu J, Wang W, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 24185-24198.

[18] Wang W, Lv Q, Yu W, et al. CogVLM: Visual expert for pretrained language models[J]. arXiv preprint, 2023, arXiv:2311.03079.

[19] Wang P, Bai S, Tan S, et al. Qwen2-VL: Enhancing vision-

language model's perception of the world at any resolution[J]. arXiv preprint, 2024, arXiv:2409.12191.
[20] Kim J, Lee Y, Yoon D, et al. An integrated YOLO and VLM system for fire detection in enclosed environments[C]//Proceedings on "I Can't Believe It's Not Better" Workshop at ICLR. PMLR, 2025, 296: 151-162.
[21] Wang, Q., Wang, D., Lu, J.et al. SAL-YOLO-DeepSeek: A lightweight real-time detection and LLM-driven decision framework for intelligent escalator safety monitoring[J]. Scientific Reports, 2025, 15: 24260.
[22] Adil M, Lee G, Gonzalez V A, et al. Using vision language models for safety hazard identification in construction[J]. arXiv preprint, 2025, arXiv:2504.09083.
[23] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems (NeurIPS). 2020, 33: 1877-1901.
[24] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Advances in Neural Information Processing Systems (NeurIPS). 2022, 35: 24824-24837.
[25] YONG G, JEON K, GIL D, et al. Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model[J]. Computer-Aided Civil and Infrastructure Engineering, 2023, 38(11): 1536-1554.
[26] 国家能源局。电力安全工作规程 发电厂和变电站电气部分: DL/T 408—2023 [S]. 北京：中国电力出版社，2023.

（责任编辑　李园）