中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-15

论文引用格式： Feng Renshuai, Liu Xilin, Xue Yuhao, Li Xinjing, Tan Yulin, Zhao Cairong. Contextual chain-of-thought driven few-shot risk recognition for industrial safety[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260154.(冯仁帅, 刘希林, 薛宇浩, 李鑫婧, 谭玉林, 赵才荣. 上下文思维链驱动的工业安全小样本风险识别[J/OL]. 中国图象图形学报,XXXX:1-15. DOI: 10.11834/jig.260154.)[DOI: 10.11834/jig.260154]

上下文思维链驱动的工业安全小样本风险识别

冯仁帅[1],刘希林[2],薛宇浩[3],李鑫婧[2],谭玉林[1],赵才荣[3*]

1. 星电集团,湖南省长沙市 410118; 2. 国网长沙供电公司,湖南省长沙市 410000; 3. 同济大学,上海市 200092

摘 要：目的 针对工业安全场景中风险类别长尾、标注样本稀缺、现有视觉模型对未知风险泛化不足且推理过程不透明的问题,研究小样本条件下的可解释风险识别方法。**方法** 构建核心思维链训练集与Few-shot样例库,提出上下文思维链学习框架。在模型层面,引入层级化视觉编码器、语义一致性分类头以及主动感知与迭代精炼机制,以增强多粒度视觉感知能力并约束推理文本与最终判断的一致性。在训练层面,采用“两阶段训练”策略:首先利用结构化思维链监督注入工业安全推理模式;随后通过上下文归纳训练强化模型从少量样例中适应未见风险类别的能力,并结合对比式图文基底损失提升视觉依据约束。**结果** 在14类未见危险评估集(unseen hazards-14,UH-14)3-shot设置下,所提模型F1-score为68.56%,较ChatCH-SFT提高12.81个百分点;F2-score由57.83%提升至70.81%,召回率达到72.40%。消融实验表明,上下文学习与思维链学习均对性能提升具有积极作用。**结论** 所提方法适用于长尾、少样本且对高召回与可解释性要求较高的工业安全识别场景,并可为其他复杂视觉风险识别任务提供参考。

关键词： 危险预警;视觉语言模型;多模态特征融合;上下文学习;小样本识别

Contextual chain-of-thought driven few-shot risk recognition for industrial safety

Feng Renshuai[1], Liu Xilin[2], Xue Yuhao[3], Li Xinjing[2], Tan Yulin[1], Zhao Cairong[3*]

1. Xingdian Group, Changsha 410118, China; 2. State Grid Changsha Power Supply Company, Changsha 410000, China; 3. Tongji University, Shanghai 200092, China

Abstract: Objective Industrial safety monitoring often involves rare but high-risk events that exhibit a pronounced long-tail distribution. In real deployment scenarios, these critical hazard categories occur infrequently, are expensive to annotate at scale, and are often defined by subtle visual cues that must be interpreted together with contextual evidence from the surrounding scene. As a result, conventional vision-based safety monitoring methods, which typically rely on large annotated datasets and closed-set classification assumptions, often struggle to generalize to unseen hazard categories in practical industrial environments. Although recent vision-language models have demonstrated strong cross-modal understanding and reasoning capabilities, their direct application to industrial safety recognition still faces several fundamental challenges. First, existing models are often insufficiently sensitive to small but decisive local details, such as missing protective equipment, improper operation near machinery, or subtle environmental anomalies. Second, their few-shot adaptation ability is

收稿日期： 2026-03-26;**修回日期：** 2026-04-27

* **通信作者：** 赵才荣 Email: zhaocairong@tongji.edu.cn

基金项目： 国家自然科学基金(项目编号: No. U25A20527,62473286)

Supported by: National Natural Science Fund of China

unstable when the support examples are limited and the target category has never appeared during training. Third, the generated reasoning text may not be well aligned with the final decision, which reduces the reliability and interpretability of the output. Finally, these models may generate fluent but weakly grounded explanations that are not fully supported by the visual evidence. To address these issues, this study investigates a few-shot industrial risk recognition framework that aims to improve both generalization to unseen hazard categories and the interpretability, factual consistency, and structural reliability of the decision process. **Methods** A contextual chain-of-thought driven framework is developed for few-shot industrial safety risk recognition. The proposed method is built on three tightly coupled components. First, a dual-database data organization strategy is adopted to support different stages of learning. A core chain-of-thought training set is constructed to teach the model a structured industrial reasoning pattern through observation, analysis, and conclusion-oriented supervision, while an external few-shot example database is designed to provide task-specific contextual examples for unseen hazard categories. This separation allows the model to learn general reasoning ability from structured supervision and then apply it to new tasks through contextual adaptation. Second, the model architecture is enhanced by a hierarchical vision encoder, a semantic consistency classification head, and an active perception with iterative refinement mechanism. The hierarchical vision encoder fuses shallow, middle, and deep visual features so that both global scene semantics and fine-grained local details can be utilized during reasoning. The semantic consistency classification head predicts the final risk label from the semantic representation of the generated reasoning text rather than from an isolated parallel branch, thereby encouraging the final decision to remain structurally aligned with the explanation. The active perception mechanism allows the model to trigger local refinement when the initial global observation is insufficient, ambiguous, or affected by clutter, occlusion, or small target scale, so that difficult samples can receive additional focused inspection on critical regions. Third, a two-stage training strategy is adopted. In the first stage, structured chain-of-thought supervision is used to inject an industrial safety reasoning pattern into the model by jointly optimizing reasoning generation, final risk classification, and image-text semantic grounding. In the second stage, meta-samples composed of contextual examples and a query image are used to explicitly train in-context generalization, so that the model learns to infer a new risk concept from only a few support examples rather than relying on memorized category names. A contrastive image-grounding loss is further introduced to strengthen the factual alignment between visual evidence and generated text, reduce hallucination, and improve the faithfulness of the reasoning process. **Results** Experiments are conducted on a dedicated evaluation protocol for unseen industrial hazard recognition. The core supervised training set contains three common industrial risk categories with structured reasoning annotations, while the UH-14 benchmark is used to evaluate generalization to 14 unseen hazard categories under few-shot settings. Quantitative experiments are reported under 1-shot, 3-shot, and 5-shot settings. Under the 3-shot setting on UH-14, the proposed method achieves an F1-score of 68.56%, outperforming ChatCH-SFT by 12.81 percentage points over its 55.75% result. The F2-score improves from 57.83% to 70.81%, and recall reaches 72.40%, indicating a substantially stronger ability to reduce missed detections in safety-critical scenarios where recall is particularly important. Under the 1-shot setting, the proposed model still achieves an F1-score of 56.92%, demonstrating that the framework remains effective even when only a single support example is available for each unseen class. Under the 5-shot setting, the model reaches an F1-score of 74.49% and an F2-score of 76.27%, showing that the method can continue to benefit from additional contextual examples. Ablation experiments further verify the contribution of the major components. Removing the contextual learning stage causes a significant drop in performance, with the F1-score decreasing by 19.72 percentage points and the F2-score decreasing by 19.51 percentage points, which confirms that explicit contextual generalization training is central to the framework. Removing the chain-of-thought learning stage also leads to clear degradation, with the F1-score and F2-score dropping by 4.76 and 5.42 percentage points, respectively, indicating that structured reasoning supervision provides an important foundation for downstream few-shot transfer. Additional analyses show that the proposed design is beneficial not only for recognition accuracy but also for improving explanation consistency and strengthening the robustness of decisions on visually complex or ambiguous samples. **Conclusion** The proposed framework improves few-shot recognition of unseen industrial risk categories while providing a more interpretable, structurally consistent, and visually grounded decision process. By coupling contextual learning, hierarchical visual perception, semantic consistency constraints, and conditional iterative refinement, the method is particularly suitable for industrial safety scenarios character-

ized by long-tail hazards, scarce annotations, strong dependence on contextual cues, and strict recall requirements. The framework does not merely improve classification performance; it also strengthens the factual reliability and logical coherence of generated explanations, which is important for human verification, expert review, and practical safety deployment in real industrial environments. In this sense, the study contributes not only a recognition model, but also a pathway toward more trustworthy multimodal reasoning for safety-critical applications. At the same time, the current study is still mainly focused on static-image scenarios. Future work may extend the method toward video-based temporal reasoning, deployment-oriented efficiency optimization, larger-scale open-category evaluation protocols, and more systematic construction of industrial reasoning data, so as to further improve practical applicability in real-world safety monitoring systems.

Key words: Danger warning; Visual language model; Multimodal feature fusion; Context learning; Few-Shot Learning

0 引 言

电力系统在社会生产生活中扮演着至关重要的角色。近年来,中国的电力产业发展迅速,电网规模和输电线里程均为世界第一。然而,在电站建设、电力生产和电网维护等相关作业中,人身伤亡事故屡屡发生,给国家,社会和家庭都造成了重大损失。对电力作业人员行为的准确分析和及时预警,是保障电力系统安全运行的关键(Xiang等,2023;Hu等,2023)。精准识别作业人员和作业现场存在的潜在安全风险,能够有效地将隐患排除于无形,是电力系统在复杂多变的环境中稳定运行的技术保障之一。高风险行为的自动化、智能化识别——例如高空作业未佩戴安全带、带电操作未使用绝缘手套、或违规进入受限区域,已成为现代现场安全管理不可或缺的一环。然而,电力行业等风险行为在现实场景中往往具有发生频率低、视觉表征复杂、且高度依赖场景上下文等特点(Chen等,2025;Xiao等,2021)。并且,那些对安全构成最严重威胁的关键事件,其发生频率远低于常规作业活动,但系统必须对其保持极高的检测灵敏度。此外,为每一种罕见风险行为采集并标注大量样本在实践中几无可能。面对这些挑战,研发一种具备强大上下文理解、逻辑推理能力,并能在极少数据样本下实现精准识别的智能系统,已成为业界和学界的迫切需求。

电力场景下的安全检测方法,传统方式主要依靠人工盯防。监控系统将实时监控数据传送到安监部门,由安全员负责监督并给出指示,不仅耗费大量人力,而且受制于人的注意力、精力和观察范围等因素,既容易忽视存在的危险,又无法及时对潜在的风险进行预警。传统视觉检测方法(马莉等,2020;胡正文,2016;曾宪武和冉祯伟,2015),以VGG卷积神经网络作为目标检测器(王碧霄,2019),K最近邻分类算法进行无监督分类(刘玮,2014)。随着基于深度学习的计算机视觉技术的迅猛发展(戴彦,2018;彭明智,2023),目标检测、人体姿态估计、多目标跟踪等技术广泛运用于一线生产环境(闫云凤,2024),例如区域卷积神经网络(region-based convolutional neural network, R-CNN)、YOLO模型(you only look once, YOLO)系列或3D卷积网络。尽管以上方法大大缓解了安监人员的工作压力,且在数据均衡的基准测试中表现优异,但在真实的工业环境中,其性能会因风险行为的稀疏性与复杂性而大打折扣。

近年来,以对比语言-图像预训练(contrastive language-image pre-training, CLIP)(Radford等,2021)、自抽样语言-图像预训练(bootstrapping language-image pre-training, BLIP)(Li等,2022)及Qwen-VL(Wang等,2024)为代表的视觉语言模型(vision-language model, VLM)在跨模态理解和小样本推理方面展现出强大潜力(Chen等,2024)。这些模型能够通过文本提示(Prompt)进行上下文学习,即便只有有限的视觉样例也能实现良好的泛化。例如,Chen等人(2025)将Qwen-VL与参数高效微调技术中的低秩适配(parameter-efficient fine-tuning-low-rank adaptation, PEFT-LoRA)(Hu等,2021)相结合,实现了建筑工地的危害自动识别与报告生成。尽管已有诸多进展(Zhang等,2024;Xiao等,2025;Gu等,2024),但现有方法在工业安全监测的实际应用中仍面临三大核心瓶颈:1)模型浅层、静态的推理范式。现有VLM如同一个只能静态的观察者,对图像进行一次性的、全局性的单遍分析。这种模式与人类专家的认知过程背道而驰,并导致两个致命缺陷:其一,是对关键细节的存在感知盲点。一个危险的关

键要素,可能在图像中占比很小,而单遍推理极易在宏观语义的干扰下忽略这些决定性细节。其二,是无法主动解决不确定性。当模型遇到模糊、遮挡或距离过远的区域时,它无法像人类专家一样做出细致观察的决策,只能进行被动的猜测,这在安全领域是不可接受的。2)对长尾风险的泛化能力与上下文适应性不足:工业安全风险呈现典型的长尾分布,大量高危行为因其罕见而难以收集充足数据。现有模型严重受限于数据不平衡,难以从极少量样本中学到新知识。同时,多数方法仍依赖对单张图片的孤立分析,缺乏一种机制,能让它们像人类一样,通过实时获得的几个新案例(上下文),动态地理解和掌握一种全新的风险定义。3)推理过程的可靠性与忠实性缺失,现有VLM在生成解释时常出现"言行不一"的矛盾——即生成的文本描述与最终的分类判断相悖;同时,它们也无法避免"事实幻觉"(Hallucination)问题,可能生成流畅但与图像细节完全不符的推理,这在安全监测这一零容忍错误的领域是致命的。

为应对上述挑战,本文提出了一种全新的面向工业安全监测的上下文学习框架。本文的核心贡献体现在三个相互关联、层层递进的层面:

首先,在训练范式与数据构建上,本文设计了一种创新的"两阶段"学习流程以赋予模型强大的推理与泛化能力。本文摒弃了单一的微调模式,第一阶段,本文利用一个精心构建的、包含结构化"观察-分析-结论"推理链的核心标注数据库,通过标准监督微调(standard supervised fine-tuning,SFT)向模型注入基础的专家思维模式;第二阶段,本文通过自动化构建的"元样本"(包含上下文示例和待查询图片),专门精调模型从少量示例中归纳新知识并解决未知问题的上下文学习(in-context learning,ICL)能力。

更核心的是,在推理范式上,本文进一步提出了基于主动感知与迭代精炼的全新推理框架。该框架赋予模型模拟人类专家的能力:在进行初步的全局分析后,能够主动识别并"请求聚焦"于模糊或关键的局部区域,再结合放大的细节信息进行二次精炼,最终形成高置信度的判断。为支撑此高级推理过程,本文在模型架构上进行了两项关键创新以提升感知的精度与决策的可信度。第一,本文提出了层级化视觉编码器,通过融合视觉主干网络的多尺度特征,为框架的全局分析和局部精炼阶段提供了动态的、多粒度的视觉信息支撑。第二,本文设计了新颖的语义一致性分类头,该分类头强制最终的风险判断必须从模型经过迭代精炼后生成的最终推理链文本中导出,从结构上根除了"言行不一"的矛盾现象。

最后,为确保模型在每一轮推理中生成的文本都真实可靠,本文在训练目标中引入了对比式图文基底损失(Contrastive Image-Grounding Loss)。该损失函数通过对比学习机制,有效抑制了模型的"事实幻觉"问题,迫使模型生成的每一句分析都必须有其视觉依据(Visual Grounding),从而显著提升了整个迭代推理过程的忠实性与可靠性。

本文的主要贡献如下:首先,在训练范式上,本文构建了"基础思维模式注入—上下文归纳能力强化"的两阶段训练框架,使模型不仅能够学习结构化的工业安全推理方式,还能够在少量样例条件下对未见风险类别进行快速适应。其次,在模型设计上,本文引入层级化视觉编码器与语义一致性分类头,前者用于增强模型对全局场景与局部关键细节的联合感知能力,后者用于缓解视觉语言模型中解释文本与最终判断不一致的问题,从结构上提升决策过程的一致性与可解释性。最后,在实验层面,本文构建了面向工业安全推理学习与小样本评测的双层数据体系,并通过对比实验、消融实验以及案例分析验证了所提方法在14类未见危险评估集(unseen hazards-14,UH-14)上的有效性,特别是在强调降低漏检的评价指标上表现出更明显优势。

1 相关研究

1.1 计算机视觉方法在工业安全检测的应用

长期以来,基于计算机视觉技术实现自动化安全监测一直是学术界与工业界的研究热点。早期的工作主要依赖于传统的图像处理和机器学习技术,例如基于K最近邻分类算法进行无监督分类(刘玮,2014)。随着深度学习的兴起,以卷积神经网络(convolutional neural network,CNN)为基础的方法成为主流。该类方法主要可分为两大范式:一是基于目标检测的方法,如采用YOLO系列(Redmon等,2016)、R-CNN等模型(Ren等,2017),通过在图像中定位工作人员并分类其行为或着装状态(如是否佩戴安全帽、是否穿着反光衣)来实现风险识别。二是

基于图像分类或分割的方法，通过对整个场景或关键区域进行分析，判断是否存在如烟火、危险区域入侵等特定风险。在相关视觉任务中，徐晗等人(2025)提出了融合上下文感知注意力的Transformer目标跟踪方法，通过跨尺度整合浅层与深层特征，并在编解码过程中引入上下文感知注意力，有效提升了复杂场景下目标表征和抗干扰能力。该研究表明，在复杂视觉场景中显式引入上下文信息和多层特征交互，有助于增强模型对关键目标的辨识精度与鲁棒性，这也为本文在工业安全识别任务中引入层级化视觉编码与上下文建模提供了相关启发。

尽管这些方法在许多预定义场景下取得了显著成功，但其根本性的局限也日益凸显。首先，它们遵循一种“封闭世界”的学习范式，其性能高度依赖于大规模、精细标注的训练数据集。这导致模型难以泛化至训练数据中未包含的新型或罕见风险，在面对动态变化的工业环境时适应性差。其次，这些模型通常作为“黑盒”运行，仅能输出一个分类标签或边界框，而无法提供其决策背后的逻辑解释。这种可解释性的缺失严重制约了系统在需要高可靠性和人工复核的严肃场景中的可信度与应用深度。本研究旨在通过引入视觉语言模型的推理能力，从根本上突破上述数据依赖和可解释性的瓶颈。

1.2 视觉语言模型

视觉语言模型(vision-language model，VLM)旨在通过联合训练来对齐图像和文本两种模态的表示，从而赋予模型强大的跨模态理解与生成能力。其发展历程大致可分为两个阶段。早期以CLIP(Radford等，2021)和图像文本对齐模型(Jia等，2021)为代表的模型，通过在大规模图文对数据上进行对比学习，成功地学习到了鲁棒的、可泛化的多模态表示空间，在零样本图像分类等任务上表现出惊人潜力。

随后的发展趋势是将视觉编码器与大型语言模型(large language model，LLM)相结合，催生了以Flamingo(Alayrac等，2022)、BLIP-2(Li等，2023)以及开源的Qwen-VL(Wang等，2024)等为代表的大型视觉语言模型(large vision-language model，LVLM)。这些模型不仅能理解图像内容，更能遵循复杂的自然语言指令，进行多轮对话、详细的图像描述以及复杂的视觉推理。它们强大的能力为解决传统视觉模型的局限性提供了新的思路：利用其丰富的世界知识和推理能力，模型有望理解抽象的安全概念，并生成人类可读的解释。然而，将现有VLM直接应用于工业安全等高风险领域仍面临挑战。其一，标准VLM的视觉编码器通常只利用最后一层的高度抽象特征，这可能导致对决定安全状况的局部关键细节(如设备上的微小裂纹)感知不足。其二，“事实幻觉”问题依然存在，模型可能生成与图像内容不符的文本。其三，模型生成的解释与其最终判断之间可能存在逻辑矛盾。本文的模型架构创新(层级化视觉编码器、语义一致性分类头)与损失函数设计(对比式图文基底损失)正是为了解决上述VLM在严肃应用中的核心短板。

1.3 大语言模型上下文学习

上下文学习(in-context learning，ICL)(Dong等，2024)最初由GPT-3等大型语言模型展现，指的是模型在不进行任何梯度更新的情况下，仅通过在提示(Prompt)中提供少量任务样例(Few-shot examples)，就能在推理时快速适应并执行新任务的能力。这种能力被认为是大型模型从海量数据中涌现出的一种隐性元学习或模式识别能力。

近年来，ICL的能力已成功扩展至多模态领域。研究者发现，通过构建包含(图像，文本描述)样例对的序列作为上下文，大型视觉语言模型也能够快速理解新的视觉概念或任务范式。例如，用户可以提供几张“消防通道堵塞”的正反例图片及其标签，模型便能据此判断一张新的图片是否存在该风险。这为实现对未知风险的快速适应提供了极具吸引力的途径。此外，赵珞君等人(2026)在少样本连续教学行为识别任务中提出了可伸缩思维链引导方法，通过对行为标签进行思维链扩展、三元组知识凝练以及多层次跨模态匹配，增强了模型对新型行为类别的语义理解与少样本识别能力。该研究表明，思维链不仅能够用于解释生成，还能够作为少样本场景下类别语义扩展与泛化建模的有效手段。不过，该方法主要围绕行为标签语义挖掘展开，尚未涉及工业安全场景中基于上下文示例的动态任务归纳、局部细节主动感知以及推理文本与视觉证据一致性约束。然而，ICL的性能在很大程度上是一种“涌现”而非“习得”的能力，其表现可能不稳定，且高度依赖于上下文样例的质量和组织形式。直接将预训练VLM用于ICL，并未在训练阶段就该能力进行显式优化。与此不同，本研究的核心创新之一在于提

出了一种专门针对ICL的训练阶段——“上下文归纳能力精调”。本文通过构建包含上下文和查询的“元样本”并利用损失屏蔽技术，显式地训练和强化模型从上下文中归纳规律并解决新问题的能力，使其从一种不稳定的涌现能力，转变为一种稳健、可靠的核心技能。

2 本文方法

为实现一个能够模拟人类专家、进行动态深度推理的工业安全监测系统，本文提出了一套以“主动感知与迭代精炼”范式为核心的端到端方法。该范式旨在将视觉语言模型从被动的单遍观察者，转变为能够主动识别不确定性、聚焦关键细节并迭代优化其结论的智能感知体。这一高级认知过程的实现，依赖于三个相互支撑的关键支柱：第一，一个为教授模型迭代式推理而设计的双功能增强数据库，它通过结构化的“初步-精炼”式标注，为模型学习提供了必要监督；第二，一个为支撑多粒度信息处理而定制的创新模型架构，它融合了能够动态切换感知尺度的层级化视觉编码器与保证最终决策可靠的语义一致性分类头；第三，一套为注入并泛化此高级能力的多目标、两阶段训练范式，它引导模型先掌握迭代推理的基础，再将其泛化应用于从上下文中学习到的未知风险。本章将对这三个核心组成部分进行详细阐述。

2.1 双功能联动数据集构建

本研究的基石在于一个精心设计的、由两个功能互补的数据库构成的协同系统，它为本文创新的两阶段训练范式提供了坚实的数据支撑。

首先，本文构建了核心标注数据库（Core Annotation Database），如图1所示。其主要目标是为模型注入基础的、可泛化的“专家思维模式”。该数据库的构建借助了大型模型进行初步标注生成，并辅以严格的人工审核与修正，以确保数据的高质量与逻辑严谨性。为提升结构化思维链（chain-of-thought，COT）标注的可靠性，本文在初步标注生成后增加了人工复核环节。具体而言，研究人员重点围绕“视觉观察是否与图像事实一致”“风险分析是否与工业安全常识相符”“结论判断是否能够由前述分析自然推出”三个维度进行审核，并对存在歧义、跳步推理或视觉依据不足的样本进行修正。对于少数争议样本，进一步通过回看图像细节与交叉讨论的方式完成统一修订。通过上述流程，尽可能保证结构化思维链标注在事实性、逻辑性与任务相关性上的一致。数据库中的每一个样本单元均由一张图片和一个结构化的对象组成。该对象不仅包含了用于监督分类任务的最终风险判断标签（ground_truth_judgment），更核心的是，它包含了一段模仿人类专家思考过程的推理链文本。该文本被设计为“观察-分析-结论”的三段式结构，旨在显式地教会模型如何从视觉证据出发，进行逻辑推演，最终得出结论。在内容覆盖上，本文刻意将风险类别限定在3种常见的基础风险（包括未戴安全帽，未佩戴口罩，人员进入机械作业区），但对每个类别都构建了大量的“最小差异对”（Minimal Pairs）与“硬负样本”（Hard Negatives），旨在通过这些高挑战性的数据来锻炼模型精细化的视觉辨识与逻辑分辨能力。

图1 核心标注数据集与Few-shot样例数据集示意

Fig. 1 Illustration of the core annotated dataset and the few-shot example dataset

与之相辅相成的是一个轻量级的小样本样例数据库。该数据库的设计目标并非用于大规模训练，而是在第二阶段的上下文学习和最终的推理阶段，为模型提供动态的、任务相关的即时知识。因此，它追求的是“小而精”。 该数据库覆盖了核心标注数据库之外的多类风险样例，用于为第二阶段上下文

归纳训练和推理阶段提供任务相关的上下文示例。为避免数据泄漏，第二阶段训练所使用的小样本样例数据库与UH-14评估集在类别上严格隔离。其数据形式被简化为直接的(图片，简明文本标签)对，例如一张消防通道被堵塞的图片会配对一个简洁的标签“风险：消防通道堵塞”。这种简洁的设计使其能高效地作为上下文注入到模型的提示(Prompt)中。总而言之，核心标注数据库通过深度和结构化的标注教会模型“如何思考”，而Few-shot样例数据库则通过广度覆盖和简洁的样例，为模型在面对新任务时提供“思考的素材”，二者协同作用，共同驱动模型实现从掌握基础能力到触类旁通的跃迁。需要强调的是，本文关注的并非类别名称本身的记忆，而是从少量示例中归纳“风险判定所依赖的视觉线索—上下文关系—结论输出”这一通用识别范式。第一阶段训练通过结构化思维链监督使模型掌握工业安全场景中的基本观察与分析方式，第二阶段则通过上下文归纳训练强化模型依据少量新样例动态建立类别判别规则的能力。因此，模型能够将在基础类别上习得的“观察—分析—判断”模式迁移到未见风险类别上，实现对新类别的快速适应，而非依赖对特定类别标签的直接记忆。

此外，为专门赋予模型进行主动感知与迭代精炼的能力，本文对核心标注数据库中的一部分样本进行了深度增强标注。具体而言，对于包含关键但细微的风险细节、或存在视觉模糊区域的样本，在其JSON对象中额外增补了三个关键字段：1) refinement_bbox：一个定义了需要被聚焦放大的感兴趣区域的边界框坐标；2) preliminary_cot：一段初步思维链文本，它反映了模型仅基于全局观察得出的、可能包含不确定性描述的初步分析；3) final_cot：一段最终思维链文本，它是在结合了由refinement_bbox指定区域的放大细节后，得出的更精确、更高置信度的最终结论。这种结构化的、从“初步”到“精炼”的数据标注模式是至关重要的，它将训练数据从静态的“问题-答案”对，转变为动态的“过程-监督”样本，从而为模型学习从全局扫描到局部聚焦、再到结论修正这一完整的高级认知过程提供了可能。

2.2 模型架构

为构建一个既能精确感知视觉细节，又能进行可靠逻辑推理的系统，本文对现有的VLM基座进行了深度定制与创新。本文选用业界先进的Qwen2.5-VL系列模型作为基座，其强大的基础跨模态理解能力为后续工作提供了高起点，提出了上下文思维链驱动(contextual chain-of-thought，COCOT)模型架构。在此基础上，本文的核心架构创新主要体现在层级化视觉编码器与语义一致性分类头两个模块，旨在分别解决视觉感知的“精度”与推理决策的“信度”两大核心问题。

整体模型架构如图2所示，其核心由层级化视觉编码器、主动感知与局部精炼机制以及语义一致性分类头三部分组成。其中，层级化视觉编码器负责提供多尺度视觉信息，主动感知机制负责在局部证据不足时触发二次观察，而语义一致性分类头则用于将最终判断与生成推理文本进行结构绑定。传统的VLM通常仅采用其视觉编码器最后一层输出的特征向量序列。这些特征富含高级全局语义，但在逐层抽象的过程中不可避免地丢失了对精细化局部细节的表征，而这些细节在工业安全场景下往往是决定性的。为克服此缺陷，本文提出了层级化视觉编码器。该设计不再丢弃宝贵的中间层信息，而是从视觉Transformer(vision transformer，ViT)的浅层、中层和深层网络中并行提取特征图。这些多尺度特征随后被送入一个轻量级多尺度融合模块，该模块通过上采样与特征加权等操作，将它们融合成一个统一的、兼具多尺度信息的视觉表征$\boldsymbol{H}_{\mathrm{v}}$：

$$\boldsymbol{H}_{\mathrm{v}}=\boldsymbol{\Phi}_{\mathrm{f}}\left(\{\boldsymbol{F}_{\mathrm{s}},\boldsymbol{F}_{\mathrm{m}},\boldsymbol{F}_{\mathrm{d}}\}\right)\#(1)$$

式中，$\boldsymbol{\Phi}_{\mathrm{f}}$表示由多尺度特征序列到统一视觉表征的受特征金字塔网络(feature pyramid networks，FPN)启发的轻量级多尺度融合映射，$\boldsymbol{F}_{\mathrm{s}}\in\mathbf{R}^{N_{\mathrm{s}}\times D}$，$\boldsymbol{F}_{\mathrm{m}}\in\mathbf{R}^{N_{\mathrm{m}}\times D}$和$\boldsymbol{F}_{\mathrm{d}}\in\mathbf{R}^{N_{\mathrm{d}}\times D}$分别代表浅层(shallow)、中层(middle)和深层(deep)的特征序列，N_{s}，N_{m}，N_{d}分别代表浅层、中层和深层的序列长度，D为特征维度。这个增强后的视觉表征$\boldsymbol{H}_{\mathrm{v}}$随后被送入语言模型的交叉注意力层，使得语言模型在生成文本时，能够根据当前推理上下文的需要，自适应地从最相关的视觉尺度中获取信息。该设计的关键价值在于，它使得模型在全局感知阶段能侧重于深层语义特征以理解整体场景，而在局部精炼阶段则能利用浅层细节特征对高分辨率的“感兴趣区域”进行精细审查，为整个迭代过程提供了动态的、多粒度的视觉信息支撑。

为了从根本上解决VLM中普遍存在的“言行不

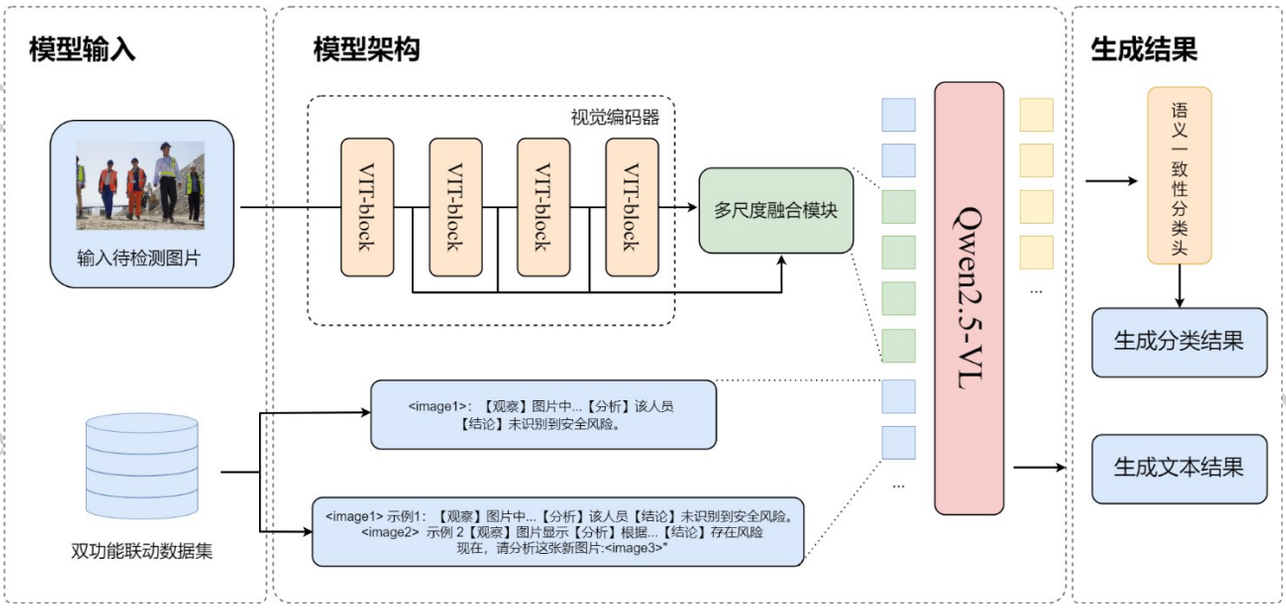

图2 COCOT模型架构

Fig. 2 Structure of COCOT model

一"(即生成的解释与最终判断相矛盾)的问题,本文设计了一种新颖的语义一致性分类头,其核心思想是强制模型的最终判断必须且只能从其自己生成的思维链文本中推导。在本文的架构中,模型首先依据融合后的视觉表征$\boldsymbol{H}_{\mathrm{v}}$和输入指令,自回归地生成完整的分析文本$\boldsymbol{T}_{\mathrm{c}}$。随后,提取代表整个生成文本序列语义的最终标记所对应的顶层隐藏状态向量$\boldsymbol{h}_{\mathrm{E}} \in \mathbf{R}^{D_{\mathrm{lm}}}$,式中$D_{\mathrm{lm}}$是语言模型的隐藏层维度。此向量被视为对整个推理过程的高度浓缩的语义总结,并被送入一个简单的多层感知机(MLP)以输出最终的分类概率分布$P\left(y_{\mathrm{c}}\right)$:

$$P\left(y_{\mathrm{c}} \mid \boldsymbol{T}_{\mathrm{c}}\right)=S\left(\boldsymbol{W} \boldsymbol{h}_{\mathrm{E}}+\boldsymbol{b}\right) \#(2)$$

式中,y_{c}表示模型对当前输入图像的最终风险类别预测结果,用于表征样本属于各候选类别的判别输出,$\boldsymbol{T}_{\mathrm{c}}$表示生成完整的分析文本,$P(\cdot)$代表"安全"或"危险"等离散的风险判断类别,$\boldsymbol{W}$和$\boldsymbol{b}$分别表示分类层的权重矩阵和偏置向量,S(·)表示softmax,用于将输入向量映射为各类别的概率分布。由于分类结果直接由推理文本的语义表示导出,因此最终判断与生成解释之间形成了更强的结构一致性约束。这一架构上的强制约束确保了模型的解释与其结论之间存在强耦合与逻辑自洽性。分类任务的损失信号会通过反向传播,直接优化整个语言模型的生成过程,迫使模型必须生成在语义上能够明确支持其最终结论的文本,从而使模型从一个黑盒判断器转变为一个逻辑闭环、过程透明的安全监控模型。

实现上述两个模块高效联动的关键,在于设计的主动感知机制(Active Perception Mechanism)。其核心在于赋予语言模型生成"聚焦请求"的能力。本文在模型的词汇表中引入了一个特殊的功能标记,记为s_{A}。在第一遍推理(全局感知)时,模型被训练来识别不确定性,并自回归地生成此标记,随后紧跟一个用于回归边界框坐标的预测。这个由模型自主生成的请求,将触发系统从原始高清图像中裁剪出相应的"感兴趣区域",并构建用于第二遍推理的增强上下文,从而构成了整个"从全局扫描到局部精探"的主动感知与迭代精炼的闭环。

2.3 两阶段训练范式

为将所设计的模型架构优势充分转化为实际能力,本文提出了一种创新的两阶段训练范式,并辅以一个多目标的复合损失函数。该范式旨在循序渐进地引导模型,在第一阶段掌握基础的单遍分析能力以及更高级的、基于不确定性的主动迭代精炼能力;而后,在第二阶段将这种习得的高级推理范式泛化至从上下文中快速学习到的未知风险之上。

训练与推理流程如图3所示。第一阶段以核心思维链训练集为基础进行基础思维模式注入,第二阶段利用由上下文样例和查询图像构成的元样本进行上下文归纳能力精调,而推理阶段则在少量支持样例条件下对查询图像进行判断,并在必要时进入

局部精炼过程。训练的第一阶段，基于大模型标准监督微调（Standard Supervised Fine-tuning，SFT）训练，对模型进行基础思维模式注入，其目标是让模型学会“专家思维模式”。在此阶段，训练数据均从核心标注数据库中抽取。对于不包含迭代标注的标准样本，输入模型的形式为一个（视觉表征$\boldsymbol{H}_{\mathrm{v}}$，指令$\boldsymbol{T}$）对。模型需要执行两个并行的任务：语言模型主干需自回归地生成完整的推理链文本$\boldsymbol{T}_{\mathrm{c}}$，同时语义一致性分类头需根据生成的文本内容，输出对该图片的风险判断。为了同时优化这两个任务并确保推理的忠实性，本文设计了一个由三部分构成的复合损失函数L_{t1}。

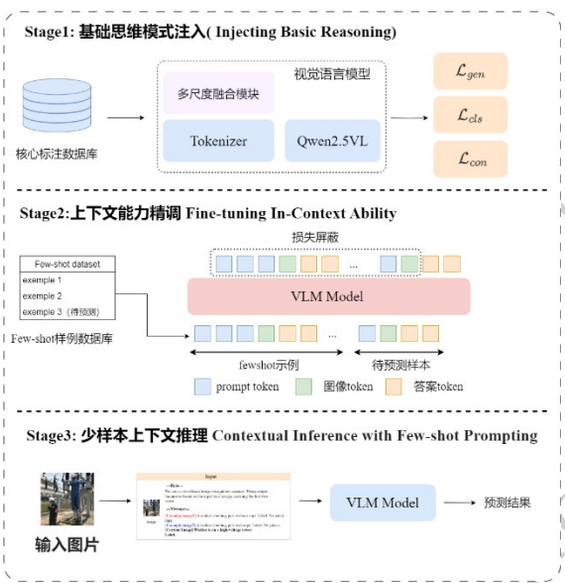

图3　模型两阶段训练流程与推理流程

Fig. 3　Two-stage training and inference pipeline of the model

第一部分是**分类损失**L_{c}，为监督语义一致性分类头输出与真实风险标签之间的一致性，本文采用交叉熵损失定义分类目标：

$$L_{\mathrm{c}}=-\sum_{c\in C}y_{\mathrm{c}}^{(c)}\log\left(P\left(y_{\mathrm{c}}^{(c)}|\boldsymbol{T}_{\mathrm{c}}\right)\right)\#(3)$$

式中，$\boldsymbol{C}$是风险类别的集合（如`Safe`，`Danger`），y_{c}是独热（one-hot）编码的真实标签，$P\left(y_{\mathrm{c}}|T_{\mathrm{c}}\right)$是模型根据生成文本$T_{\mathrm{c}}$预测的概率分布。

第二部分是生成损失L_{g}，它采用标准的自回归语言模型损失，用于监督推理链文本的生成。为约束模型生成符合标注推理链的文本内容，本文采用标准自回归语言建模损失：

$$L_{\mathrm{g}}=-\sum_{i=1}^{|T_{\mathrm{c}}|}\log P\left(t_i|t_{<i},\boldsymbol{H}_{\mathrm{v}}\right)\#(4)$$

式中，$\boldsymbol{T}_{\mathrm{c}}$是真实的推理链文本，$|\boldsymbol{T}_{\mathrm{c}}|$是$\boldsymbol{T}_{\mathrm{c}}$的token数量，$t_i$是其第$i$个token，$\boldsymbol{H}_{\mathrm{v}}$是层级化视觉编码器输出的视觉表征。

第三部分，也是确保推理忠实性的关键，是对比式图文基底损失L_{o}。该损失旨在将模型生成的文本$\boldsymbol{T}_{\mathrm{c}}$与其对应的源图像在语义表示空间中进行强绑定。在一个训练批次（batch）内，图像和模型为其生成的文本T_i构成一个正样本对，而该图像I_i与批次内所有其他文本T_j（其中$j\neq i$）构成负样本对。本文采用与CLIP方法相同的信息噪声对比估计（information noise-contrastive estimation，InfoNCE）损失来计算该项，其公式为：

$$Z_i=\sum_{j=1}^{B}\exp\left(\frac{s\left(E_{\mathrm{I}}\left(\boldsymbol{I}_i\right),E_{\mathrm{T}}\left(\boldsymbol{T}_j\right)\right)}{\tau}\right)\#(5)$$

$$L_{\mathrm{o}}=\log\left(Z_i\right)-\frac{s\left(E_{\mathrm{I}}\left(\boldsymbol{I}_i\right),E_{\mathrm{T}}\left(\boldsymbol{T}_i\right)\right)}{\tau}\#(6)$$

式中，Z_i表示第i个图像样本对应的归一化项，E_{I}和E_{T}分别是图像和文本的编码器，用于提取其语义表示向量；$\boldsymbol{I}_i$表示第i个图像样本；$\boldsymbol{T}_i$表示与$\boldsymbol{I}_i$对应的文本；$\boldsymbol{T}_j$表示批次中第j个文本样本；$\mathrm{s}(\cdot,\cdot)$是余弦相似度函数；B是批次大小，τ是温度系数。

最终，第一阶段的总损失由这三部分加权求和得到：

$$L_{\mathrm{t1}}=\alpha L_{\mathrm{c}}+\beta L_{\mathrm{g}}+\gamma L_{\mathrm{o}}\#(7)$$

其中α，β，γ是用于平衡不同任务重要性的超参数。

对于标准样本，训练目标主要包括推理文本生成、最终风险分类以及图文事实对齐三部分；而对于包含主动感知与迭代精炼标注的增强样本，训练目标还需进一步覆盖初步推理、局部区域定位以及精炼后推理结果，从而形成更完整的多目标联合优化过程。对于包含迭代标注的增强样本，模型被训练执行完整的“主动感知与迭代精炼”流程。该多步骤过程的监督被封装在一个更全面的损失函数L_{it}中，它联合优化了多个关键目标。这包括监督初步思维链T_{p}生成的损失L_{p}，监督最终精炼思维链T_{f}生成的损失L_{f}，以及应用于最终输出的分类损失L_{fc}和对比损失L_{fo}，其中，L_{p}和L_{f}分别表示作用于初步思维链与最终精炼思维链的自回归生成损失；L_{fc}表示作用于最终风险判断的分类损失；L_{fo}表示作用于最终推理

文本与输入图像之间的对比式图文基底损失，其具体形式与前文定义保持一致。此外，一个关键的组成部分是监督模型所预测的边界框 $\hat{\boldsymbol{b}}$ 与真实标注 $\boldsymbol{b}$ 之间差距的回归损失 L_{b}，采用SmoothL1损失来保证回归的稳定性：

$$L_{\mathrm{b}} = S1\left(\hat{\boldsymbol{b}} - \boldsymbol{b}\right)\#(8)$$

式中，$\hat{\boldsymbol{b}}$ 表示模型预测的边界框；$\boldsymbol{b}$ 表示真实边界框；$S1(\cdot)$ 表示SmoothL1损失。这些子目标共同构成了迭代样本的总损失函数：

$$L_{\mathrm{it}} = w_1 L_{\mathrm{p}} + w_2 L_{\mathrm{b}} + w_3 L_{\mathrm{f}} + \alpha L_{\mathrm{fc}} + \gamma L_{\mathrm{fo}}\#(9)$$

式中，w_i 是用于平衡迭代流程中不同任务的权重超参数，L_{p}，L_{b}，L_{f}，L_{fc}，L_{fo} 分别表示初步推理损失、边界框回归损失、精炼推理损失、最终分类损失和最终对比损失。这种条件化的训练策略，能够有效地教会模型根据场景的复杂性，灵活地决定何时以及如何调用其高级的迭代推理能力。

训练的第二阶段，即上下文归纳能力精调（In-Context Training），旨在将这些能力泛化至模型从未见过的未知风险。此阶段的训练数据是通过自动化构建的“元样本”（Meta-Sample）来组织的，每个元样本包含 N 个上下文示例和一张查询图片。在此阶段，本文采用一种关键的损失屏蔽（Loss Masking）机制。模型需要处理包含多个示例的整个长序列，但其损失函数（根据查询图片的复杂性，条件化地选择 L_{st} 或 L_{it}）仅针对模型对最终查询图片的预测输出进行计算。所有前序上下文示例部分的预测损失在反向传播时被完全屏蔽，不会产生梯度。这种设计强制模型将上下文示例纯粹作为一种用于理解新任务定义的条件信息，而不是试图去“记忆”它们。通过这种方式，本文显式地、端到端地优化了模型进行高级上下文学习（Advanced In-Context Learning）的能力，将其包括迭代精炼在内的复杂推理技能，从一种不稳定的涌现现象，转变为一个稳健、可靠的核心泛化能力。

3 实验设计和结果

为了系统性地验证本文提出的COCOT框架的有效性，特别是其在数据高效性和小样本泛化方面的优越性，本章设计并执行了一系列全面的定量与定性实验。本章将首先详细阐述实验所依赖的数据集、具体的评估指标以及模型训练的参数配置，随后将呈现核心的小样本性能对比实验、关键的消融研究，并以一个直观的案例分析来展示模型的实际推理能力，从而全方位地论证方法的先进性。

3.1 实验数据与实验设置

为了有效评估模型在面对未知任务时的学习到的思考推理能力，而机械的记忆SFT过程中的数据，本文构建了两套功能明确、相互独立的特制数据集。首先是核心思维链训练集（IS-CoT-3），该数据集是本文方法论的基石，专门用于COCOT的第一阶段监督微。其核心目标并非让模型记住特定危险，而是通过高质量的“思维链”标注，向模型注入一种通用的、结构化的“工业安全专家思维模式”。该数据集包含了3个生产场景中常见且具有代表性的危险类别：H01-未戴安全帽，H02-未佩戴口罩，H03-人员进入机械作业区。数据集总规模为8，326张图像，其中7，526张用于训练，800张用于验证。其最关键的特征在于，每张图像均配有一段模仿人类专家“观察现象→分析依据→得出结论”思考过程的analysis_cot文本标注。

其次，本文构建了小样本泛化评估集UH-14作为检验模型泛化能力的“试金石”。该数据集用于模拟真实世界中不断出现新危险类型的场景，专门评估模型在训练阶段完全未见过的危险类别上的小样本学习性能。其构成上，本文从Chen等人（2025）的危险分类体系中精心挑选出14个与IS-CoT-3数据集无任何类别重叠的危险类别，覆盖了从工人行为到设备、环境等多个方面。该评估集所包含的14类未见风险覆盖了个人防护缺失、危险区域闯入、机械防护缺失、现场警示缺失、通道占用及作业环境异常等多个方面，从而尽可能模拟真实工业现场中新风险类别不断出现的应用场景。需要说明的是，小样本样例数据库仅用于构造第二阶段训练所需的元样本，其类别集合与UH-14评估集严格不重叠；UH-14中的14个类别仅用于测试阶段，不参与任一训练阶段。该数据集每个类别精确包含10张高质量图像，共计140张。在进行K-shot评估时，对于每个类别，本文随机抽取K张图像作为上下文支持集（Support Set），模型需对剩余的10-K张查询集（Query Set）图像进行预测。为保证结果的稳健性，所有小样本实验结果均为5轮随机抽样的平均值。

为确保实验的公平性与可复现性，所有模型均

在统一的环境下进行。实验选用Qwen2.5-VL-7B作为所有实验的基础模型，以保证初始能力的对等。在微调阶段，实验采用参数高效的低秩适配（Low-Rank Adaptation，LoRA）技术，其rank和alpha分别设置为8和32。所有模型的训练均采用AdamW优化器，学习率恒定为5e-5，批处理大小为16，并进行10个epoch的训练，最终选用在验证集上表现最佳的模型检查点进行评估。所有实验均在一台配备NVIDIA两卡RTX4090 GPU（24GB显存）的服务器上完成。

在评估指标的选择上，实验充分考虑了工业安全监测领域的特殊性，即对危险的“漏报”所带来的后果远比“误报”严重。因此，指标体系不仅包括标准的精确率（Precision）、召回率（Recall）及综合性能指标F1-Score，还引入了F2-Score。F2-Score给予召回率比精确率高4倍的权重，能更准确地衡量模型在避免“漏报”方面的能力，这与安全监测的实际需求高度契合。其中，召回率（Recall）是评估模型性能时最为关注的核心指标之一。

3.2 小样本性能对比实验

本节旨在定量评估COCOT在面对未知危险类别时的核心泛化能力。为此，本文按照3.1节给出的K-shot评估设置，在UH-14数据集上，将COCOT与基础模型Qwen2.5-VL及对比方法ChatCH-SFT进行比较。其中ChatCH-SFT选用Qwen2.5-VL-7B作为基础模型，按照ChatCH-SFT论文中描述的数据构造方式构建中文训练数据集，进行一阶段的微调训练，实验分别在1-shot、3-shot和5-shot的设置下进行，其详细结果呈现在表1中。

表1 小样本设置下的性能对比结果

Table 1 Performance comparison under few-shot settings

模型	K-shot	精确率	召回率	F1-Score	F2-Score
Qwen2.5-VL	1-shot	18.40%	24.20%	20.91%	22.76%
ChatCH-SFT	1-shot	48.20%	51.50%	49.80%	50.80%
COCOT (Ours)	1-shot	53.50%	60.80%	56.92%	59.18%
Qwen2.5-VL	3-shot	21.10%	26.50%	23.49%	25.21%
ChatCH-SFT	3-shot	52.60%	59.30%	55.75%	57.83%
COCOT (Ours)	3-shot	65.10%	72.40%	68.56%	70.81%
ChatCH-SFT	5-shot	57.10%	62.20%	59.54%	61.11%
COCOT (Ours)	5-shot	**71.70%**	**77.50%**	**74.49%**	**76.27%**

注：所有模型参数均为7B，黑色字体表示最优结果。

如表1所示，实验结果清晰地展示了COCOT方法的优势。首先，在所有K-shot设置和所有评估指标上，COCOT的性能均以显著的幅度超越了两个基线模型。以3-shot场景为例，COCOT的F1-Score达到了68.56%，相比于代表当前主流领域微调范式的ChatCH-SFT模型（55.7%），实现了高达12.8个百分点的性能飞跃。这证明了两阶段训练范式在提升模型泛化能力方面具有根本性的优越性。

其次，本文重点关注与安全应用直接相关的召回率（Recall）指标，因为它关系到对潜在危险的“零遗漏”目标。在3-shot设置下，COCOT的召回率达到72.40%，明显高于ChatCH-SFT的59.30%；相应地，F2-Score也由57.83%提升至70.81%。这表明所提方法在降低漏报方面具有更明显优势，更符合工业安全监测对高召回率的实际需求。这种高召回率的特性，成功地将本文注入的“专家思维模式”转化为了模型在面对不确定性时更强的风险识别倾向。

此外，即使在仅有单个示例的极端1-shot场景下，COCOT依然能够达到56.92%的F1-Score。相比之下，ChatCH-SFT的F1-Score也达到了49.80%，均显著高于未经领域微调的基础Qwen2.5-VL模型。这充分表明针对特定领域任务微调的必要性，同时，随着小样本样例数量的增加，ChatCH-SFT的性能提升不够明显，这表明其并没有从样例数据中学会推理和迁移知识，暴露了传统SFT范式在面对新任务时泛化能力不足的固有局限性。相比之下，模型并

非简单地进行表面特征的模式匹配，而是成功地掌握了从极少量示例中归纳、推理和泛化的“学会学习”能力。

3.3 消融实验

为了深入探究COCOT框架中各个关键组成部分的具体贡献，本文设计了一系列严格的消融实验。本文在UH-14数据集的3-shot任务上进行了消融实验，结果如表2所示。

从表2中可以观察到，当移除专门用于泛化能力训练的第二阶段（Stage 2）后，模型性能出现了断崖式下跌。其F1-Score骤降了19.7个百分点，而与Recall指标相关的F2-Score大幅下降了19.5个百分点。这有力地证明了，本文为小样本泛化设计的上下文学习能力精调是COCOT取得成功的根本原因。仅仅依赖于第一阶段SFT所学到的通用知识，模型无法有效地将知识迁移并应用于未见过的任务中，这凸显了显式地训练模型“学会学习”能力的极端重要性。

与此同时，思维链（CoT）的监督作用同样不可或缺。在移除了CoT监督信号后，尽管模型性能未如移除第二阶段那样急剧恶化，但F1-Score和F2-Score也分别出现了4.8和5.4个百分点的显著下降。这表明，在第一阶段通过CoT向模型注入结构化的“专家思维模式”，能够为其后续的逻辑推理和泛化提供坚实的基础。缺乏这种结构化思考的训练，模型虽然仍能通过两阶段范式学习泛化，但其推理的深度和准确性受到了限制，证明了CoT作为关键助推器的重要价值。

为进一步明确不同模块带来的性能收益，本文将所提框架中的关键设计划分为三个层面进行理解。其一，层级化视觉编码器主要作用于视觉表征质量的提升，使模型在复杂工业场景中能够同时利用全局语义与局部细节信息，从而增强对关键风险线索的感知能力；其二，语义一致性分类头主要作用于“推理文本—最终判断”的对齐，有助于减少解释与结论相互矛盾的情形，提升模型输出的一致性与可信度；其三，主动感知与迭代精炼机制主要作用于复杂样本的处理过程，当模型在初步观察中存在局部不确定性时，该机制能够引导模型进一步聚焦关键区域并修正判断。三者并非彼此割裂，而是分别从感知、决策约束与推理流程三个方面共同提升了模型在工业小样本识别任务中的整体表现。

在推理代价方面，主动感知与迭代精炼机制并非对所有输入样本均执行固定次数的重复推理，而是仅在初步分析阶段检测到局部信息不足、关键细节模糊或存在较强场景干扰时，才进一步触发局部聚焦与结果修正过程。因此，该机制带来的额外计算开销具有条件触发特征，而非对全体样本等比例增加。对于风险特征较明显、场景结构相对清晰的样本，模型通常能够在初步感知阶段直接形成判断；而对于细粒度差异显著、关键证据尺度较小或背景复杂的样本，系统则通过增加一次局部细节审查来提升判断可靠性。从方法结构上看，额外开销主要来源于感兴趣区域的裁剪处理以及后续的补充推理过程，未引入独立的大规模复杂分支，因此整体复杂度增长是可控的。该设计本质上体现了一种面向工业安全任务的动态权衡策略，即在简单样本上优先保持推理效率，在复杂样本上优先保证识别可靠性，以更贴近实际应用需求。

综上，消融实验的结果充分验证了本文两阶段训练范式的完整性和先进性：第一阶段的CoT微调为模型奠定了“像专家一样思考”的基础，而第二阶段的元学习则赋予了模型“将思考应用于未知”的核心泛化能力，两者相辅相成，缺一不可。

3.4 定性实验分析

为了更深入地理解展示COCOT框架的具体效果，并验证本文注入“专家思维模式”的有效性，本节将通过具体的案例进行定性分析，旨在直观地展示

表2 关键模块消融实验结果（3-shot设置）

Table 2 Ablation results of key modules under the 3-shot setting

模型	精确率	召回率	F1-Score	F2-Score
COCOT	**65.10%**	**72.40%**	**68.56%**	**70.81%**
无上下文学习	45.20%	53.10%	48.83%(↓19.72)	51.31%(↓19.51)
无思维链学习	61.30%	66.50%	63.79%(↓4.76)	65.39%(↓5.42)

注：黑色字体表示最优结果。“↓”表示与我们的方法相比下降的百分点。

COCOT在面对复杂场景时，其内部决策过程的逻辑性、可解释性和可靠性。

本文从UH-14数据集中选取了一个具有代表性的挑战性类别：H14-机械设备的旋转部件未安装防护罩。该任务不仅要求模型识别出机械设备，更需要其理解“防护罩缺失”这一隐含的、表示状态异常的关键信息。在此场景下，向模型提供一个1-shot示例，并要求其对一张新的查询图像进行判断和分析。COCOT与ChatCH-SFT基线模型的输出对比，可视化输出结果展示在图4中。

如图4所示，COCOT的表现充分展示了其经过思维链训练后的强大推理能力。面对查询图像，它并未直接输出结论，而是生成了一段逻辑清晰、层次分明的思考结果，结合上下文内容，成功推理出了结果。

图4　COCOT与ChatCH-SFT的定性对比

Fig. 4　Qualitative comparison between COCOT and ChatCH-SFT

相比之下，ChatCH-SFT基线模型的表现则暴露了传统微调范式的局限性。由于缺乏显式的逻辑推理训练，它似乎仅关注到了场景中的“人员”和“机械”的表层视觉特征，最终将危险错误地归类为“非指定区域休息”，未能抓住问题的核心。

这个案例生动地证明了，COCOT不仅仅是一个高精度的分类器，更是一个具备初步可解释性的决策者。其透明的、结构化的推理过程不仅使其判断结果更加可靠，也极大地增强了在实际工业安全应用中的可信度。当安全管理人员收到警报时，可以经由模型输出结果，同时掌握危险行为的具体类型和信息，从而能够更精准、更高效地采取应对方案。

为进一步说明语义一致性分类头的作用，本文从模型输出机制上分析其对“言行不一”问题的缓解效果。传统视觉语言模型在完成风险识别任务时，常出现推理文本与最终判断之间不完全一致的现象，即文本分析强调的风险依据与分类结果之间存在偏差，从而削弱模型输出的可解释性与可信度。本文所设计的语义一致性分类头并非直接依据视觉特征独立给出分类结果，而是以模型生成的思维链文本语义表示为基础完成最终判断，使“观察—分析—结论”之间形成更紧密的结构约束。在这一机制下，最终输出的风险类别不再是与解释文本相分离的并行分支，而是由推理过程的语义表征进一步导出的结果，因此能够在一定程度上减少解释与结论不一致的现象。对于工业安全场景而言，这种一致性增强不仅有助于提升模型结果的可理解性，也有助于提高实际应用中的人工复核效率与系统可信度。

4　结　论

本文面向工业安全场景中的小样本识别问题，提出了一种上下文思维链学习框架。该方法通过两阶段训练范式、层级化视觉编码器、语义一致性分类头以及主动感知与迭代精炼机制的协同设计，使模型不仅能够学习结构化的工业安全推理方式，还能够在少量示例条件下对训练阶段未出现的风险类别形成较好的识别与泛化能力。实验结果表明，所提方法在UH-14数据集上的F1-score、F2-score和召回率等指标上均优于对比方法，尤其在强调降低漏检的评价指标上表现出更明显优势，说明该框架在工业小样本识别任务中具有较好的应用潜力。与此同时，本文方法在输出结果的逻辑一致性与事实可靠性方面也表现出一定优势，为工业安全监测系统中的可解释识别提供了有益思路。需要指出的是，当前研究仍主要围绕静态图像场景展开，对于连续视频环境下的动态风险理解、实际部署条件下的推理效率优化以及更大规模开放类别评测仍有进一步研究空间。未来可在视频级时序建模、轻量化部署以及更系统的数据构建与标注规范方面继续深入，以进一步提升方法的工程实用价值。

参考文献(References)

Alayrac J B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. 2022. Flamingo: a visual language model for few-shot learning//

Advances in Neural Information Processing Systems. New Orleans, USA: Curran Associates, Inc.: 23716-23736

Chen Q H and Yin X F. 2025. Tailored vision-language framework for automated hazard identification and report generation in construction sites. Advanced Engineering Informatics, 66: 103478 [DOI: 10.1016/j.aei.2025.103478]

Chen Q, Long D, Wang S, Chen Q and Yuan B. 2025. Real-time detection of personal protective equipment violations for construction workers using semi-supervised learning and video clips. Journal of Construction Engineering and Management, 151(3): 04024213 [DOI: 10.1061/JCEMD4.COENG-15310]

Chen B Y, Xu Z, Kirmani S, Ichter B, Sadigh D, Guibas L and Xia F. 2024. SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 14455-14465 [DOI: 10.1109/CVPR52733.2024.01370]

Dai Y, Wang L W, Li Y, Yan Y, Han J J and Wen F S. 2018. A brief survey on applications of new generation artificial intelligence in smart grids. Electric Power Construction, 39(10): 1-11 (戴彦, 王刘旺, 李媛, 颜拥, 韩嘉佳, 文福拴. 2018. 新一代人工智能在智能电网中的应用研究综述. 电力建设, 39(10): 1-11 [DOI: 10.3969/j.issn.1000-7229.2018.10.001]

Dong Q X, Li L, Dai D M, Zheng C, Ma J Y, Li R, et al. 2024. A survey on in-context learning//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics: 1107-1128 [DOI: 10.18653/v1/2024.emnlp-main.64]

Gu Z, Zhu B, Zhu G, Chen Y, Tang M and Wang J. 2024. AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: 1932-1940 [DOI: 10.1609/aaai.v38i3.27963]

Hu Z W. 2016. Discussion on safety hazard identification and management methods of transmission line projects. China High-Tech Enterprises, (32): 131-132 (胡正文. 2016. 输电线路工程的安全危险辨识及管理方法探讨. 中国高新技术企业, (32): 131-132) [DOI: 10.13535/j.cnki.11-4406/n.2016.32.065]

Hu Z, Chan W T, Hu H and Xu F. 2023. Cognitive factors underlying unsafe behaviors of construction workers as a tool in safety management: a review. Journal of Construction Engineering and Management, 149(3): 03123001 [DOI: 10.1061/JCEMD4.COENG-11820]

Hu E J, Shen Y L, Wallis P, Allen-Zhu Z, Li Y Z, Wang S, Wang L and Chen W Z. 2022. LoRA: Low-rank adaptation of large language models//Proceedings of the International Conference on Learning Representations (ICLR). Virtual: OpenReview.net: 1-13 [DOI: 10.48550/arXiv.2106.09685]

Jia C, Yang Y F, Xia Y, Chen Y T, Parekh Z, Pham H, Le Q, Sung Y H, Li Z and Duerig T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision// Proceedings of the 38th International Conference on Machine Learning. Online: PMLR: 4904-4916 [DOI: 10.48550/arXiv.2102.05918]

Li J, Li D, Xiong C and Hoi S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation// Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 12888-12900 [DOI: 10.48550/arXiv.2201.12086]

Li J N, Li D X, Savarese S and Hoi S. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models// Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR: 19730-19742 [DOI: 10.48550/arXiv.2301.12597]

Liu W, Huang S, Ma K and Chen H. 2014. Application of video monitoring system in power system. Guangdong Electric Power, 27(4): 57-60 (刘玮, 黄曙, 马凯, 陈皓. 2014. 视频监控技术在电力系统中的应用. 广东电力, 27(4): 57-60) [DOI: 10.3969/j.issn.1007-290X.2014.04.012]

Ma L, Ming Y, Guo T, Liao S, Zou Y X and Xiong Y. 2020. Method for anti-destruction early warning system of electric equipment based on video monitoring. Information Technology, 44(4): 115-120 (马莉, 明月, 郭婷, 廖爽, 邹雨馨, 熊一. 2020. 基于视频监控的电力设备防破坏预警系统的方法. 信息技术, 44(4): 115-120) [DOI: 10.13274/j.cnki.hdzj.2020.04.025]

Peng M Z, Xu Y, Hu Y B, Wu Y H and Yuan H D. 2023. Intelligent inspection technology of substation secondary equipment based on artificial intelligence. High Voltage Engineering, 49(S1): 90-96 (彭明智, 许尧, 胡永波, 吴永恒, 袁洪德. 2023. 基于人工智能技术的变电站二次设备智能巡检技术. 高电压技术, 49(增刊1): 90-96)

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Online: PMLR: 8748-8763 [DOI: 10.48550/arXiv.2103.00020]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2024-05-18]. https://arxiv.org/pdf/1804.02767.pdf

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Wang B X. 2019. Research on Power Vision Terminal Target Detection System Based on Deep Learning[D]. Beijing: North China Electric Power University (Beijing): 20-37 (王碧霄. 2019. 基于深度学习的电力视觉终端目标检测系统研究[D]. 北京: 华北电力大学(北京)): 20-37

Wang P, Bai S, Tan S N, Wang S J, Fan Z H, Bai J Z, et al. 2024.

Qwen2-VL: enhancing vision-language models′ perception of the world at any resolution[EB/OL].[2026-03-18]. https://arxiv.org/abs/2409.12191

Xiang Q T, Ye G, Liu Y, Goh Y M, Wang D and He T T. 2023. Cognitive mechanism of construction workers' unsafe behavior: A systematic review. Safety Science, 159: 106037 [DOI: 10.1016/j.ssci.2022.106037]

Xiao D, Dianati M, Jennings P and Woodman R. 2025. HazardVLM: A Video Language Model for Real-Time Hazard Description in Automated Driving Systems. IEEE Transactions on Intelligent Vehicles, 10(5): 3331-3343 [DOI: 10.1109/TIV.2024.3451350]

Xiao B, Zhang Y, Chen Y and Yin X. 2021. A semi-supervised learning detection method for vision-based monitoring of construction sites by integrating teacher-student networks and data augmentation. Advanced Engineering Informatics, 50: 101372 [DOI: 10.1016/j.aei.2021.101372]

Yan Y F, Chen X, Jin H Y, Qi D L, Chu H D and Wang J W. 2024. Research status and prospect of power worker behavior analysis based on computer vision. High Voltage Engineering, 50(5): 1842-1854 (闫云凤, 陈汐, 金浩远, 齐冬莲, 储海东, 汪金维. 2024. 基于计算机视觉的电力作业人员行为分析研究现状与展望. 高电压技术, 50(5): 1842-1854 [DOI: 10.13336/j.1003-6520.hve.2024.05.008]

Zeng X W and Ran Z W. 2015. Electric power tower stress comprehensive monitoring and risk pre-warning system. Guizhou Electric Power Technology, 18(8): 70-72 (曾宪武, 冉祯伟. 2015. 电力杆塔应力综合监测危险预警系统. 贵州电力技术, 18(8): 70-72) [DOI: 10.19317/j.cnki.1008-083x.2015.08.023]

Zhang J Y, Huang J X, Jin S and Lu S J. 2024. Vision-language models for vision tasks: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8): 5625-5644 [DOI: 10.1109/TPAMI.2024.3369699]

Xu H, Dong S H, Zhang J W and Zheng Y H. 2025. Context-aware attention fused Transformer tracking. Journal of Image and Graphics, 30(01): 0212-0224 (徐晗, 董仕豪, 张家伟, 郑钰辉. 2025. 融合上下文感知注意力的Transformer目标跟踪方法. 中国图象图形学报, 30(01): 0212-0224) [DOI: 10.11834/jig.240084]

Zhao L J, Liu X T, Wu Q B and Meng F M. 2026. From association to refinement: scalable-chain-of-thought-guided few-shot continual teaching behavior recognition. Journal of Image and Graphics, 31(04): 1156-1171 (赵珞君, 刘小同, 吴庆波, 孟凡满. 2026. 从联想到凝练: 可伸缩思维链引导的少样本连续教学行为识别. 中国图象图形学报, 31(04): 1156-1171) [DOI: 10.11834/jig.250327]

作者简介

冯仁帅，1995年生，第一作者，男，工程师，研究方向为AI视频目标检测自动追踪技术在配网作业现场违章行为识别中的应用。E-mail：1414945631@qq.com

赵才荣，1981年生，通信作者，男，教授，博士生导师，主要研究方向为计算机视觉，重点研究高效可信行人再识别、多模态数据驱动的自动驾驶以及垂直领域模型的知识表示与推理问题。E-mail：zhaocairong@tongji.edu.cn

刘希林，1986年生，女，高级工程师，研究方向为计算机视觉与图像识别技术在电力企业安全领域的研究与应用。E-mail：382905980@qq.com

薛宇浩，2001年生，男，硕士研究生，研究方向为计算机视觉模型安全与图像生成。E-mail：2432200@tongji.edu.cn

李鑫婧，1984年生，女，高级工程师，研究方向为计算机视觉与图像识别技术在电力企业安全领域的研究与应用。E-mail：253707773@qq.com

谭玉林，1984年生，男，工程师，研究方向为智能防污闪技术与状态监测。E-mail：541513595@qq.com