Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(6), 1518-1529
Published Online November 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.156137

视觉语言模型在火灾与烟雾识别中的应用研究

孙　强，田洪超，沈浩强，杜尚波，赵海峰*

沈阳工业大学化工装备学院，辽宁 沈阳

收稿日期：2026年10月22日；录用日期：2026年11月9日；发布日期：2026年11月18日

摘　要

早期烟雾通常面积小、对比度低，且易与云雾、蒸汽混淆，给火灾监测中的漏检控制和误报判别带来困难。视觉语言模型通过图文对应关系，将烟火属性及场景信息用于目标识别。本文从语言信息的作用出发，将相关研究分为图文对齐判别、文本引导定位和指令驱动理解三类，梳理提示适配、语义增强检测和检测器协同方法，比较其数据来源、验证任务及评价指标。已有研究显示，类别和区域描述可辅助少样本判别与烟火特征学习，场景描述可辅助受控燃烧辨识与疑似告警复核。训练阶段采用语义监督的轻量检测器侧重持续检测，推理阶段调用视觉语言模型则侧重场景解释，两者的计算开销与适用条件有所不同。当前研究仍面临弱烟雾漏检、跨场景性能下降及生成描述缺少图像依据等问题，单幅图像的分类或检测分数也难以反映长期监测效果。后续应加强局部细节与时序信息建模，完善按独立事件和地域划分的评测，并在多源观测条件下同时考察事件召回、误报频次和告警延迟。

关键词

视觉语言模型，火灾检测，烟雾识别，多模态学习

A Study on the Application of Vision-Language Models to Fire and Smoke Recognition

Qiang Sun, Hongchao Tian, Haoqiang Shen, Shangbo Du, Haifeng Zhao*

School of Chemical Equipment, Shenyang University of Technology, Shenyang Liaoning

Received: October 22, 2026; accepted: November 9, 2026; published: November 18, 2026

Abstract

Early smoke often occupies small image regions, has low contrast, and resembles clouds, mist, or

*通讯作者。

文章引用: 孙强, 田洪超, 沈浩强, 杜尚波, 赵海峰. 视觉语言模型在火灾与烟雾识别中的应用研究[J]. 人工智能与机器人研究, 2026, 15(6): 1518-1529. DOI: 10.12677/airr.2026.156137

steam, increasing the risk of missed detections and false alarms. Vision-language models use image-text correspondences to incorporate smoke and fire attributes and scene information into recognition. This review groups existing studies into image-text discrimination, text-guided localization, and instruction-driven understanding according to the role of language. It compares prompt adaptation, semantically enhanced detection, and detector collaboration in terms of data sources, evaluation tasks, and performance measures. Existing studies indicate that category and region descriptions can support few-shot discrimination and visual feature learning, while scene descriptions can help identify controlled combustion and verify suspected alarms. Lightweight detectors trained with semantic supervision emphasize continuous detection; vision-language models used during inference emphasize scene interpretation. Their computational costs and operating requirements differ. Remaining limitations include missed faint smoke, reduced performance across scenes, and descriptions without adequate visual support. Scores on individual images also fail to characterize long-term monitoring performance. Further work should improve local detail and temporal modeling, evaluate generalization across independent events and regions, and assess event-level recall, false-alarm frequency, and alarm delay using observations from multiple sources.

Keywords

Vision-Language Models, Fire Detection, Smoke Recognition, Multimodal Learning

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

火灾与烟雾识别是林区瞭望、建筑安防和无人机巡护中的重要技术环节，其目标不仅是判断画面中是否存在烟火，还包括在目标尚不明显时及时发现异常，并控制长期监控中的误报。随着深度学习的发展，视觉监测逐步由人工设计颜色、纹理和运动特征，转向利用数据学习烟火的判别表示，相关研究已覆盖分类、目标检测和语义分割等任务[1]-[3]。然而，烟火的视觉表现具有明显的阶段性和场景依赖性。火焰通常亮度较强，烟雾则具有半透明、低对比度和不规则扩散等特点；远距离初生烟羽占据的像素较少，容易与云雾、蒸汽和扬尘混淆。仅依据局部外观识别烟火，难以充分解释目标与周围环境的关系，而相似外观在不同场景下又可能具有不同的告警含义。因此，烟火视觉识别既要发现弱小目标、排除相似干扰，也要判断场景中的烟火是否需要告警。评价方法除了考察单幅图像的识别结果，还应反映连续观测中的事件发现能力和误报情况。

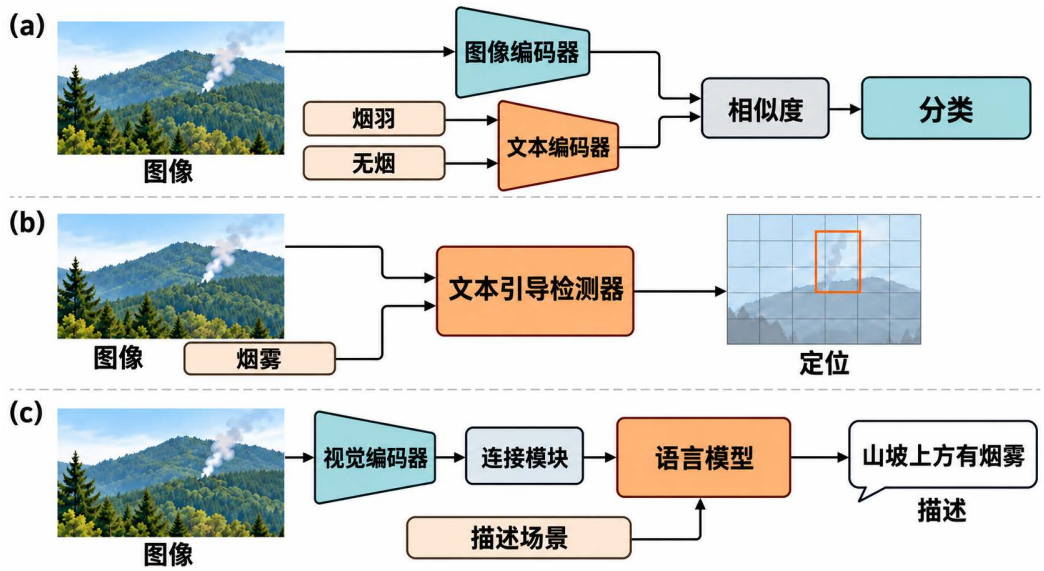
已有综述从任务类型、模型结构和应用环境等角度整理了烟火视觉识别研究。夏雪等围绕识别、检测与分割梳理视觉烟雾方法[4]，Cheng 等归纳模型、数据资源与评价指标[5]，Gragnaniello 等以应用场景组织视频火灾研究，Khan 等则关注室内外环境变化与误报问题[6] [7]。这些综述侧重视觉模型及其应用，语言与视觉的结合则带来了新的识别和交互方式。视觉语言模型(Vision-Language Models, VLMs)通过联合视觉与语言表示，将类别属性、目标位置和场景关系引入识别过程，使研究进一步涉及图文匹配、文本引导定位、火场描述与疑似告警复核。Boroujeni 等讨论了视觉语言技术在火灾影像中的应用，Raza 等将多模态理解纳入野火遥感监测的发展方向[8] [9]。比较此类研究时，应区分训练阶段的语义监督与推理阶段的语言交互，进而分析语言信息是否改善了弱烟雾检出和复杂场景中的告警结果。

本文围绕语言参与烟火视觉识别的方式组织相关研究，采用叙述性综述方法，文献更新截至 2026 年 10 月 14 日。文献选择结合期刊官网、会议论文集、相关综述及引文追溯，并通过 Crossref 辅助核对书目

信息；重点纳入语言直接参与烟火判别、定位、场景理解或告警复核的正式发表研究，基础模型、数据资源与视觉时序工作则用于说明技术来源和评价背景。在此基础上，本文将相关方法归纳为图文对齐判别、文本引导定位和指令驱动理解三类，比较少样本适配、训练期语义增强与在线模型协同的实现机制，进而结合代表性数据集讨论证据质量及适用条件。分析中分别标明图像测试、视频告警和仿真验证的范围，避免将不同任务的分数直接排列比较，并将漏检、误报及调用开销纳入同一应用背景。本文的讨论重点是不同方法在弱烟雾发现、目标定位和告警复核中的实际效果，以及其验证数据能否支持连续监测应用。

2. 视觉语言模型的技术基础与应用路径

语言信息可用于类别匹配、目标定位和场景解释，其参与训练或推理的方式也影响系统的输出和计算开销。按语义功能，VLMs 的应用路径可分为图文对齐判别、文本引导定位和指令驱动理解(图 1)，分别侧重类别匹配、空间对应与场景解释。从实施阶段看，语言既可在推理时作为类别查询或任务指令，也可仅在训练时提供语义监督，部署时保留视觉检测路径。因此，技术路线的比较需要同时说明语言承担的任务与模型运行阶段。训练时使用文本编码器，并不意味着部署时仍需运行语言模型。下文据此介绍三类技术基础，并在应用分析中进一步区分训练期语义增强与在线 VLM 推理。



注：(a) 图文相似度分类；(b) 文本引导目标定位；(c) 指令驱动场景描述。

Figure 1. Common technical routes of vision-language models for fire and smoke recognition
图 1. 视觉语言模型用于火灾与烟雾识别的常见技术路线

2.1. 图文对齐与语义判别

对比语言－图像预训练(Contrastive Language-Image Pre-training, CLIP)将图像和文本映射至共享空间，通过图文相似度匹配类别[10]。在烟火识别中，类别文字能够包含颜色、形态和环境属性，将“烟雾/非烟雾”细化为林区烟羽、水汽或工业排放等描述，为区分相似干扰提供语义参照。由于预训练数据与消防监控影像存在差异，直接使用通用提示未必能够充分刻画弱烟雾及领域负样本，因此需要结合目标场景进行适配。在少量领域标注下，CoOp 通过学习文本上下文调整类别表示，Tip-Adapter 利用支持图像建立特征缓存，分别通过调整文本表示和利用图像参照改善类别匹配[11] [12]。这类方法保留了预训练表示，并减少对大规模参数更新的依赖，但支持集的组成仍会影响判别边界。若样本主要来自明显烟火

或少数固定背景，模型对新相机中的初生烟羽及薄雾、受控燃烧等情况仍可能判断不准。因此，少样本适配应同时考虑正样本阶段与困难负样本覆盖，并在独立场景中评价提示和支持数据的贡献。

2.2. 文本引导的目标定位

开放词汇检测将语言条件与区域特征结合，根据文本描述在图像中寻找相应目标区域。Grounding DINO 通过图文融合实现开放集合检测，YOLO-World 面向实时开放词汇识别组织视觉与语言表示[13] [14]。在烟火任务中，文字查询可由类别名称扩展为“远处的灰色烟羽”等属性描述，为类别扩展和区域检索提供灵活性。然而，烟羽边界模糊、密度不均，且常与天空或建筑背景交叠，图像级语义相似并不意味着模型已经获得可靠的定位依据。全局特征有助于判断场景是否异常，保留局部高分辨率信息有助于发现小面积烟雾，二者需要通过多尺度表示、区域监督或局部裁剪建立联系。定位评价还应关注漏检区域、边界偏移及目标尺度，避免仅以是否输出了烟雾描述衡量空间能力。对于类别固定的消防应用，领域检测监督通常仍承担空间约束；对于类别和属性不断变化的检索任务，语言条件的灵活性更有价值。实际选型时，应结合类别扩展需求、标注条件和在线计算开销确定语言模块的使用方式。

2.3. 视觉指令理解与领域适配

生成式 VLMs 可依据指令输出描述、问答或结构化判断。BLIP-2 以查询模块连接预训练视觉与语言模型，LLaVA 通过视觉指令调优学习图像问答与描述[15] [16]。用于烟火任务时，模型可围绕火源位置、周围物体和活动状态组织信息，使识别结果便于人员核对。比较领域适配方法时，需要说明训练和推理分别使用了哪些目标任务样本。零样本评价通常不针对目标测试任务进行专门微调，也不在输入中提供标注示例；少样本设置则允许利用少量目标任务样本。提示学习可采用不同监督规模，不能仅凭提示的形式判断其是否属于少样本方法。本文重点讨论少样本提示与缓存适配，并将其与领域微调区分，FireCLIP 和 ForestFireVLM 分别体现提示调优与火情属性学习[17] [18]。具体而言，ForestFireVLM 以 Qwen2.5-VL 为基础，采用低秩适配(Low-Rank Adaptation, LoRA)更新少量适配参数，使火场图像与结构化属性标注建立对应。推理时，系统以 JSON 字段及候选答案约束生成，分别输出烟雾、火焰、火势和环境属性，并允许在证据不足时回答“无法判断”[18]。领域微调调整属性判别与表达方式，结构约束提高输出的可核对性，但二者均不能替代视觉证据核验。在比较方法时，还需区分推理阶段的自然语言指令与训练阶段的可学习提示，前者规定当前任务要求，后者通过训练调整提示参数。指令理解的优势主要体现在环境关系和可读输出，实际应用仍需限定问题范围，对图像中可见的目标与位置作出判断，并将缺少直接证据的火因或危险程度作为不确定推断处理。

3. 视觉语言模型在火灾与烟雾识别中的应用进展

3.1. 零样本与少样本烟火判别

图文判别是 VLMs 的基础应用。Wei 和 Kulkarni 将其用于野火烟雾二元分类[19]。云雾、水汽和工业排放等相似对象容易引起误报，提示描述因而需要体现烟火与这些干扰之间的差别。FireCLIP 针对伪烟混淆与区域数据不均衡，利用多模态提示学习适配 CLIP，其数据来自佛山五台摄像机一年内的告警图像[17]。研究将野火、工业排放与水汽纳入判别，以少量可学习参数调整图文先验，适于领域标注有限的监控任务。Saidi 等基于 4B 模型，在 FIgLib-Test 上比较基础、召回增强与平衡提示，以推理指令调节漏检和误报[20]。这类策略无需训练提示参数，但受模型原有视觉能力与表达稳定性影响。比较此类结果时，需要说明模型是否更新参数、使用了哪些支持样本，以及提示是否在测试前确定，以免混淆提示设计与领域训练的作用。ForestFireVLM-7B 在场景字段评测中的平均准确率为 76.6%，在 FIgLib-Test 上的精确

率、召回率和 F1 分别为 98.6%、57.2%和 72.4% [18]。描述能力与弱烟雾发现具有不同难度，高精确率仍可能伴随较多漏检，因而需要分别评价场景理解与早期烟羽召回。

3.2. 文本引导与多粒度语义增强检测

SmokeBench 分别评价分类、分块、网格与边界框定位，揭示烟雾存在性判断与空间感知的差距[21]。小烟羽易在缩放和编码中被背景淹没，语言描述难以补回丢失的视觉信息；保留局部细节与建立区域语义对应，是有效定位的前提。语言既可在推理时作为目标查询，也可在训练时作为特征监督。MVLFireNet 采用后者，以全局场景与局部目标描述构建多粒度语义增强分支，并结合多尺度空间注意和跨调制融合[22]。全局描述为场景特征学习提供监督，局部描述强调烟火目标属性，两级监督用于改善目标与背景的区分。其内部机制是将全局和局部视觉表示分别与对应文本表示进行对比学习，并将两级语义对齐损失与分类、边界框回归损失联合优化，使语言监督参与视觉特征学习。多尺度空间注意通过不同感受野的卷积分支补充空间细节，跨调制融合则以深层语义引导浅层特征去噪，并以浅层细节补偿深层特征的位置偏差[22]。因此，文本语义、空间细节和跨层融合分别约束不同的识别环节，其收益需结合各模块的消融结果解释。以平均精度均值(mean Average Precision, mAP)衡量，MVLFireNet 在 FSDataset-VL 上的 mAP@0.5 与 mAP@0.5:0.95 为 86.3%和 54.8%，检测网络含 2.41 M (百万)参数，计算量为 6.8 GFLOPs (十亿次浮点运算) [22]。这类方法在训练中利用文本监督，推理时保留视觉检测网络，可减少在线语言处理的开销；需要临时增加检索类别时，则需采用支持文本查询的检测方法。

3.3. 火情语义理解与多源信息融合

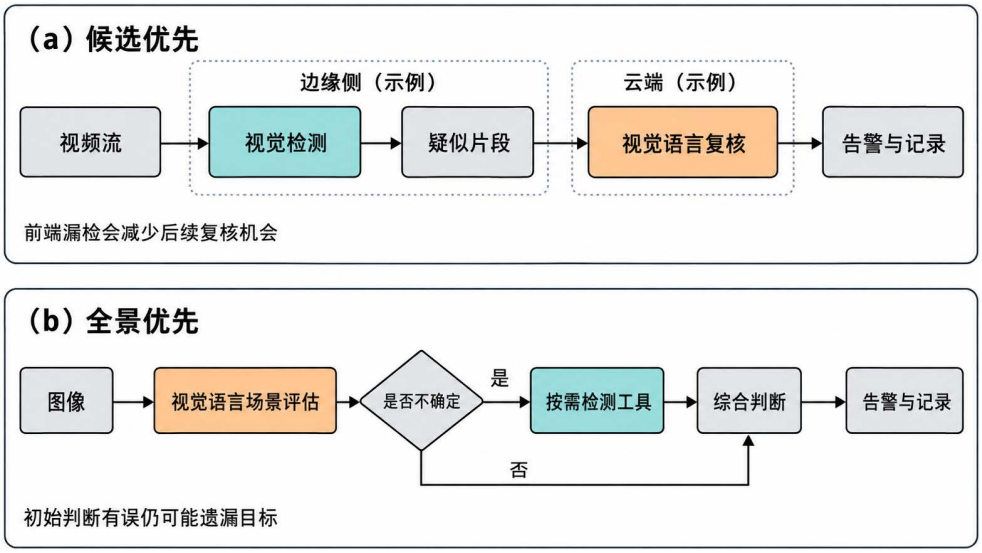
火情理解关注烟火、环境与活动的关系。炉灶或壁炉附近的受控燃烧，与家具、线路失控着火具有不同告警含义。ConFire 利用视觉描述及后续语义分类，区分无火、受控火和危险火[23]，将火焰存在性与场景异常性联系起来。ForestFireVLM 与 Qwen2-Wildfire 通过领域图文适配支持火场描述和场景识别[18] [24]。结构化输出可分别呈现目标、位置与环境关系，便于复核；但风速、起火原因和燃烧可控性可能依赖图像之外的信息，需要区分可见属性与情境推断。Sandeep 等通过多智能体与检索增强生成组织影像、传感器及气象资料，构建野火问答流程[25]。图像用于识别烟火及周围物体，传感器和气象资料补充环境信息，检索则为回答提供可追溯的资料来源。不同数据须在采集时间和位置上相互对应；问答正确率也应与烟火检出性能分开报告。

3.4. 协同告警与连续监测

检测器与 VLM 的协同使连续感知和情境分析能够在同一告警流程中分工。FIRE-TASTIC 利用检测器提供的时空信息，在疑似帧上执行描述或问答；FOCUS 结合检测、运动过滤与语义确认，根据场景配置识别模块[26] [27]。这类候选优先流程能够减少大模型调用，但后续分析依赖前端对疑似烟火的覆盖，初筛遗漏的目标通常难以进入复核。另一类协同方式利用语义判断调整检测策略。Kim 等通过 Florence2 判断整帧烟火存在性：判断为阳性时，将 YOLOv8s 阈值从 0.5 降至 0.3，否则保留 0.5。以原文的逐图检出结果统计，综合 F1 由 0.863 升至 0.8895；烟雾类别的召回率由 0.833 升至 0.885，精确率由 0.944 降至 0.931 [28]。该结果来自包含合成资料的图像测试，性能变化同时受到语义判断与检测阈值调整的影响，尚不能单独反映告警复核的作用。比较协同方法时，应明确语义模块处于初筛之前还是之后，并结合阈值变化分析误报与漏检的权衡。

除候选优先流程外，全景优先方法尝试先理解环境，再根据不确定性调用检测工具。ContextFireAgent 通过全景分析、多角色判断和一致性保留规则完成复核，在 410 张 RGB (红绿蓝三通道)图像上的准确率与 F1 为 0.983 和 0.978，YOLO 调用比例为 10.0% [29]。相较原始 VLM，其误报由 5 个减至 2 个，漏检

由 3 个增至 5 个，说明误报减少的同时仍可能增加漏检；这一结果来自图像测试，尚不能据此推断连续视频中的告警效果。FireAgent 采用边缘检测初筛与云端图文模型复核，专项测试包含 90 张正样本和 10 张干扰图像，误报由 3 张降至 0 张[30]，但这一规模的测试尚不足以说明长期运行中的误报负担。协同方法的评价因而需要分别报告复核前后的误报和漏检，尤其应检查被语义模块否决的真实烟火样本。图 2 展示候选优先与全景优先两类常见流程，动态阈值调节属于其他协同策略；云边分工则描述计算模块的部署位置，可与不同调用流程组合。



注：(a) 候选优先，云边分工为部署示例；(b) 全景优先，按不确定性调用检测工具。动态阈值调节等策略见正文。

Figure 2. Two common invocation orders for detector and vision-language model collaboration
图 2. 检测器与视觉语言模型协同的两类常见调用顺序

协同识别进一步面向连续监测时，还需利用时间证据区分单帧干扰与持续发展的烟火事件。FIgLib 与 SmokeyNet 结合固定相机观测及时间关联，STV-FSANet 通过检测、跟踪与轨迹验证约束告警[31] [32]。时序方法与 VLM 可提供不同信息：前者利用烟羽的运动、扩散和持续性判断目标是否真实存在，后者分析烟火与周围环境的关系。无人机监测还涉及视角、目标尺度和观察路径变化，Danish 等据此梳理数据、模型及平台约束[33]。FIRE-VLM 结合视觉语言感知与强化学习，在火灾数字孪生中开展无人机跟踪[34]，将语义识别延伸至观察策略。由于仿真环境能够控制火场演变和平台运动，其结果更适于分析感知与行动的耦合机制，对真实烟雾外观、飞行扰动和通信条件的适应性仍需实测。因此，图像复核、视频事件判别和无人机主动观测的结果应分别讨论，不能将某一层面的实验效果直接推广到整个监测流程。

Table 1. Representative applications of vision-language models in fire and smoke recognition
表 1. 视觉语言模型在火灾与烟雾识别中的代表性应用

研究	语义作用与推理方式	验证数据与层级	结果及适用限制
FireCLIP (2025) [17]	提示调优；在线图文匹配	相机告警图像、FASDD；图像判别	关注伪烟混淆；需领域支持集
ForestFireVLM (2025) [18]	领域微调；在线判断与描述	FIgLib-Test；图像判别	精确率 98.6%、召回 57.2%；描述与发现分别评价

续表

MVLFireNet (2026) [22]	训练期图文监督；在线视觉检测	FSDataset-VL；检测框评价	mAP@0.5 为 86.3%；检测网络 2.41 M
ConFire (2026) [23]	在线描述及语义分类	重标注公共集、ConFire；图像情境分类	区分无火、受控火与危险火；依赖情境标签
FIRE-TASTIC、FOCUS (2025) [26] [27]	检测或运动筛选；按需语义确认	各研究的视频测试集；视频识别	按片段与场景调用；受候选召回制约
YOLO-Florence2 (2025) [28]	整帧语义判断；动态检测阈值	含合成资料的图像集；每图检出	综合 F1 为 0.8895；需结合阈值解释
ContextFireAgent (2026) [29]	全景判断；按需调用检测工具	Wildfire Dataset，410 张；图像二分类	F1 为 0.978；误报减少伴随漏检增加
FireAgent (2026) [30]	边缘检测；云端语义复核	专项 100 张；图像复核	误报 3 张降至 0 张；需持续视频验证

从表 1 所列研究看，方法选型需要同时考虑标注条件、输出目标与运行资源。领域样本有限且主要输出类别时，可优先比较提示学习与少样本适配；需要持续输出检测框时，训练期语义增强能够在视觉路径中利用语言监督；需要解释环境关系或复核疑似告警时，则需考察在线 VLM 的实际贡献。候选优先流程应重点检查初筛召回，全景优先流程应重点控制语义分析的调用成本，时序确认还需权衡误报过滤与发现延迟。不同研究的数据分布、监督规模和输出粒度并不一致，较高的图像分数不能直接说明更适合连续监控。表中同时列出各方法的语义作用、推理方式、验证数据及评价限制，所列结果用于说明方法特征和适用条件，不宜据此对模型性能作直接排序。

4. 数据资源与评价体系

4.1. 从视觉标注到多粒度图文标注

视觉数据主要提供类别、检测框或掩码。FASDD 整合自然图像、无人机与遥感资料；FLAME 和 FLAME2 面向无人机火场观测，后者提供配对 RGB/红外影像[35]-[37]。不同平台的尺度与背景差异，为跨场景评价提供了条件。FSDataset-VL 在 19,866 张图像、41,983 个检测框上增加全景与局部描述[22]。DetectiumFire 包含约 2.25 万张图像和 2500 段视频，图像包括约 1.45 万张真实资料与 8000 张合成资料，并结合视觉标注及场景文本[38]。采用这些图文数据时，需要检查描述中的对象、属性与标注区域是否一致。SmokeBench 考查多粒度烟雾感知；Fire360 基于 228 段消防训练全景视频组织问答、动作描述、定位与退化对象检索，覆盖低可见度和时间记忆任务[21] [39]。相关资源见表 2。部分基准增加了描述、问答和时间关联任务，但消防训练场景仍不能完全代表真实灾害条件。

Table 2. Representative datasets and evaluation benchmarks for fire and smoke recognition
表 2. 火灾与烟雾识别的代表性数据集与评测基准

数据资源	观测与标注特点	主要任务	使用要点
FASDD [35]	自然图像、无人机及遥感影像，烟火框	多源烟火检测	区分平台、空间尺度与数据来源
FLAME [36]	无人机火场影像及类别/分割标注	火场分类、分割	关注受控燃烧与真实场景差异
FLAME2 [37]	配对 RGB/红外视频与帧级标签	多模态火场判别	核对烟雾标签定义及同事件相关性

续表

FIgLib [31]	固定相机烟雾图像及时间关联	林火烟雾与早期识别	按事件划分，保留初生烟羽与负样本
FSDataset-VL [22]	全景/局部描述与检测框对应	多粒度语义增强检测	区分图像、事件与文字条目数量
SmokeBench [21]	分类、分块、网格与边界框任务	多粒度烟雾感知	分开评价存在性与空间定位
DetectiumFire [38]	真实/合成图像、视频及场景文本	检测、生成与火情理解	分开报告真实与合成数据的作用
Fire360 [39]	消防训练全景视频与多任务标注	退化感知、问答与记忆	注意低可见度和训练场景分布

4.2. 数据质量与分层评价

烟火数据的评价需要同时说明统计单位、划分方式与标签含义。连续视频帧能够扩大图像数量，新增描述和问答能够增加标注条目，但二者均不等同于新增独立火灾事件。若训练与测试包含同一事件的相邻画面，模型可能利用相近背景获得较高分数，而对新相机和新地域的适应能力仍不明确。因此，数据集应记录事件、相机和采集区域，分别设置同域测试与跨场景测试，并公开领域适配所用样本的范围。阳性定义也需符合应用目标：FLAME2 依据专家估计的烟雾画面覆盖标注，采用约 50%的 RGB 覆盖阈值[37]，与固定相机中首次出现弱烟羽的识别任务不同。若忽略这一差异，较成熟烟团上的分类效果可能被误读为早期发现能力。对于图文数据，还应核查描述与目标区域是否对应，区分可见属性与情境推断，并检查自动生成文字中的目标遗漏、对象混淆和缺少依据的风险解释。事件之间是否独立、是否覆盖容易混淆的负样本，以及图像和文字标注是否一致，均应作为数据质量检查的内容。

在保证事件独立性和标注一致性的基础上，评价指标还应与任务输出粒度对应。分类常用精确率、召回率和 F1，F1 为前两者的调和平均；检测常用 mAP，以交并比(Intersection over Union, IoU)规定定位要求，mAP@0.5 与 mAP@0.5:0.95 的严格程度不同。属性或问答准确率评价语义理解，不能与检测分数合并形成统一排名。对于持续监控，场景配置、候选筛选和时间确认共同影响告警，图像分数只能说明其中部分能力[6] [40]。系统还需统计独立事件的发现情况、单位相机时间的误报频次，以及首次可见烟火至有效告警的延迟，并说明时间窗口和重复告警的合并规则。对模型生成的描述和回答，还应逐项核对对象、位置与属性是否具有视觉依据，可结合消防领域问答和对象幻觉评价方法核查事实一致性[39] [41]。运行层面则需区分模型推理耗时与端到端响应，考虑传输、排队和失败重试等环节。表 3 按图像判别、定位、语义理解、视频告警与部署列出评价重点，便于区分模型识别错误、语义判断错误与系统运行问题。

Table 3. Evaluation priorities for different fire and smoke recognition tasks
表 3. 不同烟火识别任务的评价重点

任务层级	典型输出	主要评价内容	应控制的条件
图像判别	烟火类别或存在性	精确率、召回率、F1 及困难负样本错误	类别定义、支持样本与领域训练
区域定位	检测框、网格或目标区域	mAP、IoU 及小目标定位表现	输入尺度、标注范围与定位阈值
语义理解	描述、属性或视觉问答	字段正确性、事实一致性及证据对应	提示版本、问题范围与外部知识
视频告警	独立事件及告警时间	事件召回、误报频次与发现延迟	事件划分、窗口长度与告警合并
系统部署	持续推理与触发式复核	端到端响应、资源占用及运行失败	硬件、网络、并发与调用策略

5. 关键挑战与发展趋势

5.1. 细粒度时空语义对齐

早期弱烟雾感知的困难在于局部视觉证据与全局语义表示之间存在尺度差异。通用视觉编码倾向于提取场景整体特征，细小烟羽在缩放、特征压缩和背景混合过程中可能失去可辨识性，此时增加文字描述难以补回已经丢失的图像信息。SmokeBench 中不同空间任务的表现差异，以及 ForestFireVLM 在 FIgLib-Test 上的有限召回，说明语义理解与弱目标发现需要分别改进[18] [21]。MVLFireNet 将全景和局部描述用于特征学习，为多粒度语义约束提供了参考[22]；视频时序研究则利用目标运动与轨迹持续性补充静态外观[31] [32]。后续可将烟羽的位置、形态和扩散过程与图像区域及文本描述对应起来，分别利用全景图像保留环境信息、局部高分辨率图像呈现弱目标、相邻帧验证目标的持续性。具体可探索将连续视频帧中的烟雾扩散过程与浮力上升、风场输运等流体力学先验相结合的 VLM。实现上，可从相邻帧提取烟羽区域、运动方向及形态变化，经时序编码与扩散过程描述对齐，并以物理一致性作为辅助训练约束。FIRE-VLM 的物理数字孪生提供了结合火场演变与语义感知的研究背景[34]，但连续视频中的弱烟雾识别仍需单独验证。考虑湍流、遮挡及相机运动的影响，物理先验宜作为容错的软约束，并通过消融比较其对召回、误报和延迟的贡献。局部裁剪可能损失上下文，较长时间窗口可能延后告警，二者均需与召回收益一并评价。研究可针对首次可见烟羽、低对比度目标及相似干扰设置子任务，比较全景、局部和时间证据的独立贡献，而不以场景描述的完整性替代早期发现能力。

5.2. 跨场景泛化与领域数据建设

跨场景泛化受到采集平台、背景结构和阳性定义的共同影响。固定相机、无人机与遥感影像具有不同的视角、尺度及成像条件，同一平台也会受到季节、天气和工业活动变化的影响，室内外综述与无人机监测研究对此已有讨论[7] [33]。预训练图文表示能够提供可迁移的语义先验，但目标场景中的烟雾外观及相似负样本仍需适配；FireCLIP 的提示调优与 FASDD 的多源数据组织，分别从模型和数据层面提供了应对思路[17] [35]。后续数据建设应增加独立事件和困难负样本，避免仅靠相邻视频帧扩大规模，并分别设置跨事件、跨相机和跨地域测试。对合成资料，应区分其与真实观测的外观及背景差异，单独评价加入合成数据前后的效果，不能将生成数量直接作为场景覆盖的证据。领域适配实验还需固定支持集、提示配置和测试范围，分析收益来自烟火属性学习还是特定背景匹配。公开数据来源、标签定义与划分规则，有助于形成可比较的迁移结论，并为新地域监测中的模型更新提供依据。

5.3. 可信语义输出与多源证据融合

生成式模型能够依据图像给出火情描述，但语言流畅并不能保证描述正确。模型可能遗漏弱烟雾、误认场景对象，或仅凭局部视觉线索就对起火原因或危险程度作出确定判断。ConFire、领域场景模型及 Fire360 拓展了受控燃烧区分、属性描述和消防问答，也提示语义标签及问题范围需要与可观察信息对应[18] [23] [39]。可信输出应区分目标存在、空间关系和风险解释，并允许在信息不足时保留不确定判断；若声称某种危险，应说明所依据的对象、区域或外部观测。可进一步研究同时输出不确定性量化结果和可溯源视觉证据的 VLM 告警系统。实现上，可对烟火存在性及关键属性设置置信度估计模块，利用独立验证集检验并校准其与实际错误率的对应关系，避免直接将模型自报置信度视为可靠概率；同时将各项判断关联至原始帧时间戳、目标框或分割区域及对应裁剪图像。对低置信度或语义判断与区域证据不一致的样本，可触发补充观测或人工复核，并结合对象幻觉评价[41]，检验字段正确性、证据对应关系及置信度校准误差。热红外可补充 RGB 成像不足，气象和现场传感资料可提供环境背景，检索增强机制则能够关联外部信息，但融合的前提是来源、采集时间与位置对应明确[9] [25] [37]。多角色复核若依赖同一

模型和同一幅图像，仍可能共享判断偏差；意见一致本身不能代替独立观测或人员核实。未来应结合区域证据、字段约束和来源标记组织输出，分别评价感知错误、生成错误及外部资料错误，并在弱烟雾、受控火源和视觉干扰等场景检验复核收益。评价时应同时报告描述是否易于理解、关键对象和属性是否正确，以及复核是否改变了最终告警结果。

5.4. 轻量协同部署与连续运行验证

连续检测与火情解释对计算资源的要求不同，部署时可据此安排检测器和语言模型的分工。训练阶段引入语言监督、推理阶段保留视觉路径的轻量检测网络，推理时无需再次运行文本模块；在线调用的生成式 VLM 主要用于对触发事件进行情境分析和告警复核。FireAgent 在单张 RTX 3090 上的前端耗时小于 15 ms/帧，而 Qwen2.5-VL-7B 图文响应约 2~5 s [30]，说明两类模块的处理时间存在明显差异，模块耗时也不能直接代表包含传输、排队和重试的端到端响应。候选优先流程能够减少语言模型调用，但前端漏检的目标通常也不会进入后续复核；全景优先流程保留了环境关系，却需承担更频繁的语义分析成本。ContextFireAgent 的误报减少伴随漏检增加，进一步说明协同优化需要同时检查两类错误[29]。后续应在同一测试集控制阈值、输入尺度与候选集合，比较复核前后的错误变化，再评价并发事件下的响应、资源占用和调用失败。量化、压缩或调度策略改变后，也需重新检验弱烟羽和困难负样本。随着静态验证转向持续视频，统一事件定义、重复告警合并及失败处理规则，有助于解释图像测试成绩与实际监测效果之间的差异。

6. 结论

视觉语言模型在火灾与烟雾识别中的作用主要体现在图文类别匹配、文本引导定位、训练期语义监督和火情描述等方面。已有研究说明，语言信息可帮助模型利用目标属性和场景关系，但其效果与训练样本、输出任务及模型调用方式密切相关。训练期语义增强有利于保留较轻的在线检测流程，在线视觉语言推理则可补充场景解释和告警复核，两者不能仅凭单项测试分数比较。尤其需要注意，描述更完整或误报减少，并不能保证早期弱烟雾检出和连续监测性能同步提高。后续研究应加强局部细节与时间关联建模，采用独立事件、相机和地域划分验证集，并核对生成描述与图像及外部观测是否一致。系统评价还应覆盖事件召回、误报频次、告警延迟及运行失败等问题。只有在持续视频和独立事件上分别验证检测、复核与告警输出，才能判断语言模块在实际火灾监测中带来的改善及其计算代价。

参考文献

[1] 张冰, 李士心. 森林火灾检测算法研究综述[J]. 传感器技术与应用, 2025, 13(6): 838-847.

[2] Elhanashi, A., Essahraui, S., Dini, P. and Saponara, S. (2025) Early Fire and Smoke Detection Using Deep Learning: A Comprehensive Review of Models, Datasets, and Challenges. *Applied Sciences*, **15**, Article No. 10255. https://doi.org/10.3390/app151810255

[3] 陈漓, 崔燕, 邓乾涛. 基于空-频域联合表征的 YOLOv11 改进模型及初期火灾烟雾检测应用[J]. 安防技术, 2026, 14(2): 57-72.

[4] 夏雪, 袁非牛, 章琳, 杨龙箴, 史劲亭. 从传统到深度: 视觉烟雾识别、检测与分割[J]. 中国图象图形学报, 2019, 24(10): 1627-1647.

[5] Cheng, G., Chen, X., Wang, C., Li, X., Xian, B. and Yu, H. (2024) Visual Fire Detection Using Deep Learning: A Survey. *Neurocomputing*, **596**, Article No. 127975. https://doi.org/10.1016/j.neucom.2024.127975

[6] Gragnaniello, D., Greco, A., Sansone, C. and Vento, B. (2024) Fire and Smoke Detection from Videos: A Literature Review under a Novel Taxonomy. *Expert Systems with Applications*, **255**, Article 124783. https://doi.org/10.1016/j.eswa.2024.124783

[7] Khan, R.A., Bajwa, U.I., Raza, R.H. and Anwar, M.W. (2025) Beyond Boundaries: Advancements in Fire and Smoke

Detection for Indoor and Outdoor Surveillance Feeds. *Engineering Applications of Artificial Intelligence*, **142**, Article 109855. https://doi.org/10.1016/j.engappai.2024.109855

[8] Boroujeni, S.P.H., Mehrabi, N., Afghah, F., McGrath, C.P., Bhatkar, D., Biradar, M.A., *et al*. (2025) Toward AI-Driven Fire Imagery: Attributes, Challenges, Comparisons, and the Promise of VLMs and LLMs. *Machine Learning with Applications*, **22**, Article 100763. https://doi.org/10.1016/j.mlwa.2025.100763

[9] Raza, M., Yin, X., Abbas, T., Ahmad, I., Aslam, R.W., Tanveer, M.F., *et al*. (2026) Remote Sensing Techniques Shaping the Future of Wildfire Early Detection. *WIREs Data Mining and Knowledge Discovery*, **16**, e70094. https://doi.org/10.1002/widm.70094

[10] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. (2021) Learning Transferable Visual Models from Natural Language Supervision. *Proceedings of the* 38*th International Conference on Machine Learning*, **139**, 8748-8763.

[11] Zhou, K., Yang, J., Loy, C.C. and Liu, Z. (2022) Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, **130**, 2337-2348. https://doi.org/10.1007/s11263-022-01653-1

[12] Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., *et al*. (2022) Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M. and Hassner, T., Eds., *Lecture Notes in Computer Science*, Springer, 493-510. https://doi.org/10.1007/978-3-031-19833-5_29

[13] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., *et al*. (2024) Grounding DINO: Marrying DINO With Grounded Pre-Training for Open-Set Object Detection. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T. and Varol, G., Eds., *Lecture Notes in Computer Science*, Springer, 38-55. https://doi.org/10.1007/978-3-031-72970-6_3

[14] Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X. and Shan, Y. (2024) YOLO-World: Real-Time Open-Vocabulary Object Detection. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 16901-16911. https://doi.org/10.1109/cvpr52733.2024.01599

[15] Li, J., Li, D., Savarese, S. and Hoi, S. (2023) BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. *Proceedings of the* 40*th International Conference on Machine Learning*, **202**, 19730-19742.

[16] Liu, H., Li, C., Wu, Q. and Lee, Y.J. (2023) Visual Instruction Tuning. *Advances in Neural Information Processing Systems* 36, New Orleans, 10-16 December 2023, 34892-34916. https://doi.org/10.52202/075280-1516

[17] Wu, S., Qiao, Y., He, S., Zhou, J., Wang, Z., Li, X., *et al*. (2025) Fireclip: Enhancing Forest Fire Detection with Multimodal Prompt Tuning and Vision-Language Understanding. *Fire*, **8**, Article 237. https://doi.org/10.3390/fire8060237

[18] Seidel, L., Gehringer, S., Raczok, T., Ivens, S., Eckardt, B. and Maerz, M. (2025) Advancing Early Wildfire Detection: Integration of Vision Language Models with Unmanned Aerial Vehicle Remote Sensing for Enhanced Situational Awareness. *Drones*, **9**, Article 347. https://doi.org/10.3390/drones9050347

[19] Wei, T. and Kulkarni, P. (2024) Enhancing the Binary Classification of Wildfire Smoke through Vision-Language Models. 2024 *Conference on AI, Science, Engineering, and Technology* (*AIxSET*), Laguna Hills, 30 September-2 October 2024, 115-118. https://doi.org/10.1109/aixset62544.2024.00021

[20] Saidi, M.A., Fourati, R. and Kechaou, Z. (2026) Safety-Oriented Prompt Design for Small Vision-Language Models in Early Wildfire Smoke Detection. 2026 *IEEE International Conference on Complex Systems and Intelligent Computing* (*ICCSIC*), Tabarka, 26-29 March 2026, 45-51. https://doi.org/10.1109/iccsic69301.2026.11521386

[21] Qi, T., Li, W. and Barnes, N. (2026) Smokebench: Evaluating Multimodal Large Language Models for Wildfire Smoke Detection. 2026 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Tucson, 6-10 March 2026, 1043-1053. https://doi.org/10.1109/wacv61042.2026.00108

[22] Ma, Y., Shan, W., Sui, Y. and Wang, M. (2026) A Lightweight Forest Fire Detection Model with Multi-Granularity Vision-Language Enhancement. *Fire*, **9**, Article 409. https://doi.org/10.3390/fire9090409

[23] Giang, A.T., Doan, N.Q., Tran, M.D., Tran, K.D., Nguyen, H.H. and Busson, A. (2026) Towards Efficient Context-Aware Classification with Compact VLM Architectures: Indoor Fire Case Study. *Scientific Reports*, **16**, Article No. 17467. https://doi.org/10.1038/s41598-026-48743-5

[24] Du, S., Li, J. and Noto, M. (2025) Wildfire Scene Recognition Based on Qwen2-Wildfire. *Proceedings of the* 2025 8*th International Conference on Software Engineering and Information Management*, Singapore, 10-12 January 2025, 254-262. https://doi.org/10.1145/3725899.3725938

[25] Sandeep, A., Jayarathna, S., Sandaruwan, S., Samarappuli, V., Meedeniya, D. and Perera, C. (2026) Context-Aware Multi-Agent Architecture for Wildfire Insights. *Sensors*, **26**, Article 1070. https://doi.org/10.3390/s26031070

[26] Gragnaniello, D., Greco, A., Sansone, C. and Vento, B. (2025) Video Fire Recognition Using Zero-Shot Vision-Language Models Guided by a Task-Aware Object Detector. *ACM Transactions on Multimedia Computing, Communications, and Applications*, **21**, 1-24. https://doi.org/10.1145/3721291

[27] Gragnaniello, D., Greco, A., Sansone, C. and Vento, B. (2025) FOCUS: Improving Fire Detection on Videos by Scenario

Adaptation. *Image and Vision Computing*, **162**, Article 105679. https://doi.org/10.1016/j.imavis.2025.105679

[28] Kim, J., Lee, Y., Yoon, D., Jung, C. and Lee, G. (2025) An Integrated YOLO and VLM System for Fire Detection in Enclosed Environments. *Proceedings on "I Can't Believe It's Not Better*: *Challenges in Applied Deep Learning" at ICLR* 2025 *Workshops*, *Proceedings of Machine Learning Research*, **296**, 151-162.

[29] Shi, B., Zhu, Y., Li, J., Wu, X., Wang, J., Li, Z., *et al*. (2026) Contextfireagent: A Multi-Role Consensus Agent for RGB Wildfire Monitoring. *Remote Sensing*, **18**, Article 3066. https://doi.org/10.3390/rs18183066

[30] 高深远，黄蕙兰，曹昊轩，王宇森，高学鸿，黄国忠. 基于云边协同的室内火灾监测大模型智能体研究[J/OL]. 工程科学学报: 1-16. https://doi.org/10.13374/j.issn2095-9389.2026.04.16.006, 2026-10-13.

[31] Dewangan, A., Pande, Y., Braun, H., Vernon, F., Perez, I., Altintas, I., *et al*. (2022) FIgLib & SmokeyNet: Dataset and Deep Learning Model for Real-Time Wildland Fire Smoke Detection. *Remote Sensing*, **14**, Article 1007. https://doi.org/10.3390/rs14041007

[32] Ghimire, D., Kim, D., Jo, Y., Lee, E., Jeong, S. and Kim, B. (2026) STV-FSANet: Track-Level Spatio-Temporal Verification for Fire and Smoke Alarm Validation in Video Surveillance. *Sensors*, **26**, Article 4970. https://doi.org/10.3390/s26154970

[33] Danish, S., Piran, M.J., Khan, S.U., Khan, M.A., Dang, L.M., Zweiri, Y., *et al*. (2025) Vision-Based Fire Management System Using Autonomous Unmanned Aerial Vehicles: A Comprehensive Survey. *Artificial Intelligence Review*, **59**, Article No. 16. https://doi.org/10.1007/s10462-025-11415-3

[34] Webb, C., Habibpour, M., Raha, M.H., Tavakkoli, A.R., Coen, J. and Afghah, F. (2026) FIRE-VLM: A Vision-Language-Driven Reinforcement Learning Framework for UAV Wildfire Tracking in a Physics-Grounded Fire Digital Twin. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops* (*WACVW*), Tucson, 6-10 March 2026, 1493-1502. https://doi.org/10.1109/WACVW68408.2026.00161

[35] Wang, M., Yue, P., Jiang, L., Yu, D., Tuo, T. and Li, J. (2025) An Open Flame and Smoke Detection Dataset for Deep Learning in Remote Sensing Based Fire Detection. *Geo-Spatial Information Science*, **28**, 511-526. https://doi.org/10.1080/10095020.2024.2347922

[36] Shamsoshoara, A., Afghah, F., Razi, A., Zheng, L., Fulé, P.Z. and Blasch, E. (2021) Aerial Imagery Pile Burn Detection Using Deep Learning: The FLAME Dataset. *Computer Networks*, **193**, Article 108001. https://doi.org/10.1016/j.comnet.2021.108001

[37] Chen, X., Hopkins, B., Wang, H., O'Neill, L., Afghah, F., Razi, A., *et al*. (2022) Wildland Fire Detection and Monitoring Using a Drone-Collected RGB/IR Image Dataset. *IEEE Access*, **10**, 121301-121317. https://doi.org/10.1109/access.2022.3222805

[38] Liu, Z., H. Khajavi, S. and Jiang, G. (2025) Detectiumfire: A Comprehensive Multi-Modal Dataset Bridging Vision and Language for Fire Understanding. *Advances in Neural Information Processing Systems* 38, San Diego, 2-7 December 2025, 157286-157325. https://doi.org/10.52202/085713-4742

[39] Tiwari, A., Masoud, F., Nguyen, D., Kraft, J., Ji, H. and Nahrstedt, K. (2025) Fire360: A Benchmark for Robust Perception and Episodic Memory in Degraded 360° Firefighting Video. *Advances in Neural Information Processing Systems* 38, San Diego, 2-7 December 2025, 33390-33420. https://doi.org/10.52202/085713-0990

[40] Carletti, V., Greco, A., Saggese, A. and Vento, B. (2026) Real-Time Fire Detection from Video Surveillance Cameras: Where Are We Now? *IEEE Access*, **14**, 25994-26007. https://doi.org/10.1109/access.2026.3664195

[41] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X. and Wen, J. (2023) Evaluating Object Hallucination in Large Vision-Language Models. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 292-305. https://doi.org/10.18653/v1/2023.emnlp-main.20