第 52 卷 第 7 期：2998-3009 高电压技术 Vol.52, No.7: 2998-3009
2026 年 7 月 31 日 High Voltage Engineering July 31, 2026
DOI: 10.13336/j.1003-6520.hve.20260524

MM-Power：面向电力巡检场景的视觉大模型多模态评测基准

闫云凤[1]，邓以恒[1]，蔡芳仪[2]，陈　毅[3]，龚世超[1]，熊周智[1]，曾子墨[4]，蔡舒瑶[1]，林嘉扬[1]，齐冬莲[1]
（1. 浙江大学电气工程学院，杭州 310027；2. 西安交通大学电气工程学院，西安 710049；
3. 浙江大学海洋学院，舟山 316021；4. 同济大学汽车与能源学院，上海 201804）

摘　要：随着视觉大模型在通用多模态任务上的快速发展，其在电力巡检等专业场景中的应用潜力日益受到关注。然而，现有研究对于此类模型在工业细粒度感知、视觉推理和抗幻觉方面的真实能力边界仍缺乏系统认识。针对通用多模态基准难以覆盖电力巡检中专业目标繁多、小目标缺陷突出、成像视角复杂以及误报代价较高等问题，该文提出面向电力巡检任务的多模态评测基准 MM-Power。该基准从细粒度感知、视觉语义一致性与幻觉评估、视觉逻辑推理 3 个层面构建 12 项任务，共包含 2 708 道选择题，覆盖可见光与红外两种模态、输电/变电/配电 3 类业务场景、云台/无人机/机器人/手持摄像机/手持红外测温仪 5 类采集视角，以及 130 类设备状态、缺陷与异常。进一步地，该文选取 16 个具有代表性的开源与闭源视觉大模型，在中英文提示词条件下开展统一评测。结果表明：英文提示词下最佳模型总体准确率仅为 77.34%，中文提示词下最佳模型总体准确率为 75.54%；模型在存在性判断、负样本鉴别和空间关系推理等任务上普遍表现较弱，提示词语言还会对部分模型的工业场景表现产生明显影响。研究结果表明，当前通用视觉大模型距离可靠的电力巡检应用仍有较大差距，MM-Power 可为电力领域多模态模型的评测、优化与部署提供统一基准。

关键词：视觉大模型；电力巡检；多模态评测；幻觉评估；视觉推理；工业智能

MM-Power: A Multi-modal Benchmark for Evaluating Vision-language Models in Power Grid Inspection

YAN Yunfeng[1], DENG Yiheng[1], CAI Fangyi[2], CHEN Yi[3], GONG Shichao[1], XIONG Zhouzhi[1], ZENG Zimo[4], CAI Shuyao[1], LIN Jiayang[1], QI Donglian[1]
(1. College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China;
2. School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
3. Ocean College, Zhejiang University, Zhoushan 316021, China;
4. College of Automotive and Energy Engineering, Tongji University, Shanghai 201804, China)

Abstract：General-purpose vision-language models have shown strong performance on widely used multimodal benchmarks, yet a systematic understanding of their true capability boundaries in industrial fine-grained perception, visual reasoning, and hallucination resistance remains lacking. In power grid inspection, generic multimodal benchmarks often fail to cover the large variety of domain-specific targets, prominent small-target defects, complex imaging perspectives, and the high cost of false alarms. To address the mismatch between existing generic benchmarks and the demands of real-world power inspection, this paper presents MM-Power, a multimodal benchmark tailored to power grid inspection scenarios. MM-Power contains 2 708 multiple-choice questions organized into three tracks and 12 tasks, covering fine-grained perception, visual-semantic consistency and hallucination assessment, and visual reasoning. The benchmark spans visible-light and infrared modalities, transmission/substation/distribution scenarios, five acquisition perspectives including pan-tilt cameras, unmanned aerial vehicles, robots, handheld cameras, and handheld infrared thermometers, and 130 categories of equipment states, defects, and anomalies. Based on a unified evaluation of 16 representative proprietary

基金资助项目：国家自然科学基金(62476242)；浙江省“尖兵”“领雁”研发攻关计划项目(2025C01058).
Project supported by National Natural Science Foundation of China (62476242), Pioneer R&D Program of Zhejiang Province (2025C01058).

and open-weight vision-language models with both English and Chinese prompts, the best model achieves only 77.34% overall accuracy in English and 75.54% in Chinese. Significant weaknesses are observed in existence judgments, negative-sample identification, and spatial relation reasoning, while prompt language also has a noticeable impact on the industrial-scenario performance of some models. These findings indicate that current vision-language models are still far from reliable deployment in practical power inspection, and MM-Power provides a unified benchmark for future evaluation, optimization, and deployment of multimodal models in the power domain.

Key words：vision-language model; power grid inspection; multimodal benchmark; hallucination evaluation; visual reasoning; industrial intelligence

0 引言

近年来，随着无人机、机器人、云台和红外测温设备在输电线路、变电站和配电场景中的广泛应用，电力巡检正由人工经验主导逐步转向数据驱动。与此同时，以 GPT、Gemini、Claude、Qwen-VL 和 InternVL 为代表的视觉大模型在多模态理解、视觉问答和跨模态推理等任务上持续取得进展，并在 MMBench、MM-Vet、SEED-Bench 和 MMMU 等通用基准上表现突出[1-6]。这使得视觉大模型在电力巡检中的应用潜力受到越来越多关注。相较于传统目标检测模型，视觉大模型具备更强的跨任务泛化能力，理论上能够在统一框架下完成设备识别、缺陷判断、位置定位和交互问答等多类任务。

然而，通用基准上的高分并不等价于模型已经具备工业场景的可靠性。电力巡检任务与自然图像理解存在显著差异：一方面，巡检图像中包含大量专业设备与细粒度部件，缺陷形态往往尺度很小、边界不清、上下文复杂，模型需要在高分辨率图像中发现低占比目标；另一方面，电力巡检并非单纯的“看见”任务，很多判断依赖设备结构知识、空间关系理解以及对异常状态的逻辑推断，例如刀闸分合状态、导线与树障相对位置、正常设备是否被错误描述为故障等[7-10]。这意味着，在电力场景中，视觉大模型需要同时满足细粒度感知、视觉推理与抗幻觉 3 个要求[11-16]，而这些能力在现有通用基准中并未被系统联合考查。

围绕上述需求，现有电力巡检研究主要集中于基于卷积神经网络或目标检测框架的专用视觉方法，并在绝缘子破损、金具异常、异物悬挂和红外热异常等任务中取得了较好效果[17-22]。这类方法对单一对象或单一故障模式通常能够达到较高精度，但往往依赖明确的任务定义与专门训练，难以适应电力巡检中不断变化的对象类型和多样化问答需求。另一方面，面向视觉大模型的现有评测工作虽然已形成较为系统的基准体系，例如 MMBench 注重通用能力覆盖，MM-Vet 强调复合能力考查，POPE 和 H-POPE 则专门分析多模态幻觉问题[1-2,23-24]，但这些工作大多面向自然图像或开放知识场景，难以覆盖电力巡检中常见的小目标异常、局部定位、错误前提诱导和结构化空间推理问题。

对于电力领域应用而言，评测基准的作用并不在于复现完整的巡检处置流程，而在于把复杂应用拆解为可稳定测量的基础能力单元。开放生成式评测虽然形式灵活，但在工业场景中往往面临答案边界不清、专业术语表达差异大、模型间结果难以公平比较等问题。相比之下，选择题形式能够把视觉证据与候选判断明确对应起来，使评分标准更清晰、测试结果更稳定，也更适合电力巡检这类术语严格、结论边界明确的场景。同时，复杂巡检决策必须建立在感知、定位、证据一致性和基本推理等底层能力可靠的前提上；若这些基础能力尚不稳定而直接评估复杂应用流程，不利于识别模型真正的薄弱环节。因此，从基础能力入手构建评测基准，更有助于揭示当前视觉大模型在电力巡检中的能力边界。

基于上述认识，本文提出面向电力巡检场景的多模态评测基准 MM-Power。该基准围绕细粒度感知、视觉语义一致性与幻觉评估、视觉逻辑推理 3 个层面组织任务。本文的工作重点不在于提出新的模型结构，而在于构建一套能够稳定揭示模型薄弱环节的评测体系，并通过对 16 个代表性视觉大模型开展统一双语测试，分析其在真实电力巡检场景中的表现差异。MM-Power 共包含 2 708 道题目、12 项任务和 130 类设备状态、缺陷与异常，覆盖可见光与红外两种模态、输电/变电/配电 3 类场景以及 5 类采集视角，可为电力领域视觉大模型的评测、优化与部署提供统一依据。与 MMBench、MM-Vet、POPE 等通用多模态基准相比，MM-Power 不以自然图像或开放知识问答为主要对象，而是面向电力巡检中的专业设备识别、小目标异常和空间关系推

理等典型问题，更适合评价视觉大模型在电力巡检场景中的基础可靠性。

1 MM-Power 基准构建

1.1 基准属性、设计目标与总体框架

作为面向电力巡检场景的评测基准，MM-Power 具有高质量基准的核心属性：代表性、覆盖性、客观性和可复现性。代表性体现在图像来源于真实电力巡检业务场景；覆盖性体现在任务同时涵盖细粒度感知、幻觉评估和视觉推理；客观性体现在所有题目均采用选择题形式，并具有明确视觉证据和唯一标准答案；可复现性体现在统一提示模板、统一评分指标和固定选项顺序。

在电力运维场景下，不同模型往往呈现显著的能力失衡现象。例如，部分模型具备识别设备属性的能力，但存在误报、漏报等问题；部分模型误报漏报率低，但在更复杂的推理场景下表现欠佳。只有在统一框架下对模型的基础能力进行系统考察，才能较完整地呈现模型在实际部署中的能力结构。

因此，通过围绕实际应用中的 3 类核心能力缺口，本文构建 MM-Power 的设计如下：1）细粒度感知，即模型是否能够识别具体设备、部件属性及局部异常；2）视觉语义一致性与幻觉评估，即模型在面对正常设备或错误前提时是否仍然坚持基于图像作答；3）视觉逻辑推理，即模型能否基于局部视觉证据完成计数、状态判断和空间关系推断。围绕这 3 类目标，本文把整个 benchmark 划分为 3 个 Track、12 项任务，覆盖 2 708 道题目、12 项任务以及多模态多场景数据来源，形成图 1 所示总体结构，其中每个任务下方数值表示该任务的题目数量。

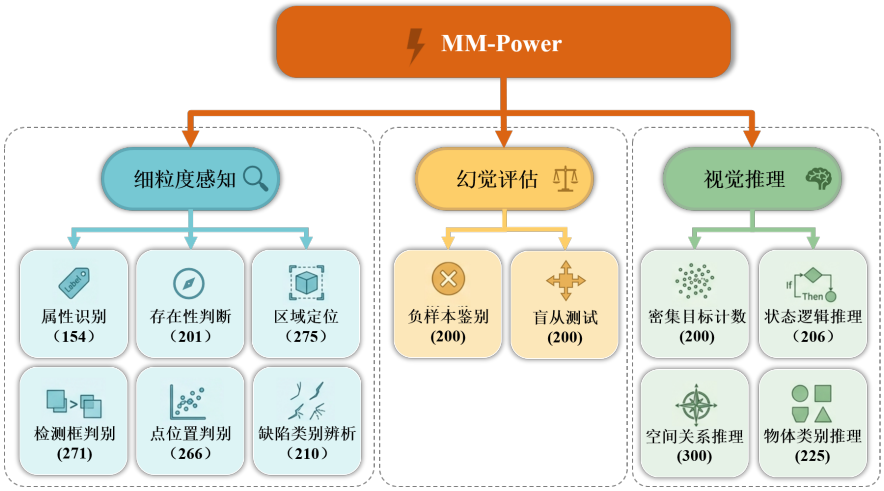


图 1 MM-Power 基准总体框架

Fig.1 Overall framework of the MM-Power benchmark

通过上述组织结构，将工业巡检中相互交织的识别、判断和推理过程拆解为可重复测量的能力单元，从而避免仅凭单一任务对模型作出过于粗略的评价。

1.2 数据来源与覆盖范围

本文构建 MM-Power 所使用的图像均来自真实电力巡检业务场景，覆盖可见光与红外两种模态，贯穿输电、变电、配电 3 类典型作业环境，并包含云台、无人机、机器人、手持摄像机和手持红外测温仪 5 类采集视角。这种数据组织方式兼顾了远距离线路巡检、近距离设备检查以及热异常识别等不同任务形态，有助于减少模型因适应单一成像模式而获得偏高评价的情况。

MM-Power 中的 130 类设备状态、缺陷与异常并非与 12 项任务形成简单的一一对应关系，而是根据视觉证据类型、业务风险形态和题目可判别性映射到不同任务中。具体而言，属性识别任务主要覆盖设备类别、部件材质、颜色状态等外观属性；存在性判断任务主要覆盖局部异物、细小缺陷、人员作业状态等目标占比较低但业务风险较高的对象；区域定位、检测框判别和点位置判别任务主要选取具有明确空间位置标注依据的设备和缺陷；缺陷类型辨析任务侧重于同类设备下外观相近但处置含义不同的缺陷模式。幻觉评估任务则从正常设备图像和视觉事实明确的图像中构造负样本鉴别与错误前提诱导问题，用于考查模型是否会在缺乏视觉证据

时输出故障结论。视觉推理任务进一步覆盖密集重复结构、设备开合状态、树障与导线相对位置、跨图物体类别匹配等场景，以评估模型对局部视觉证据的组合推理能力。

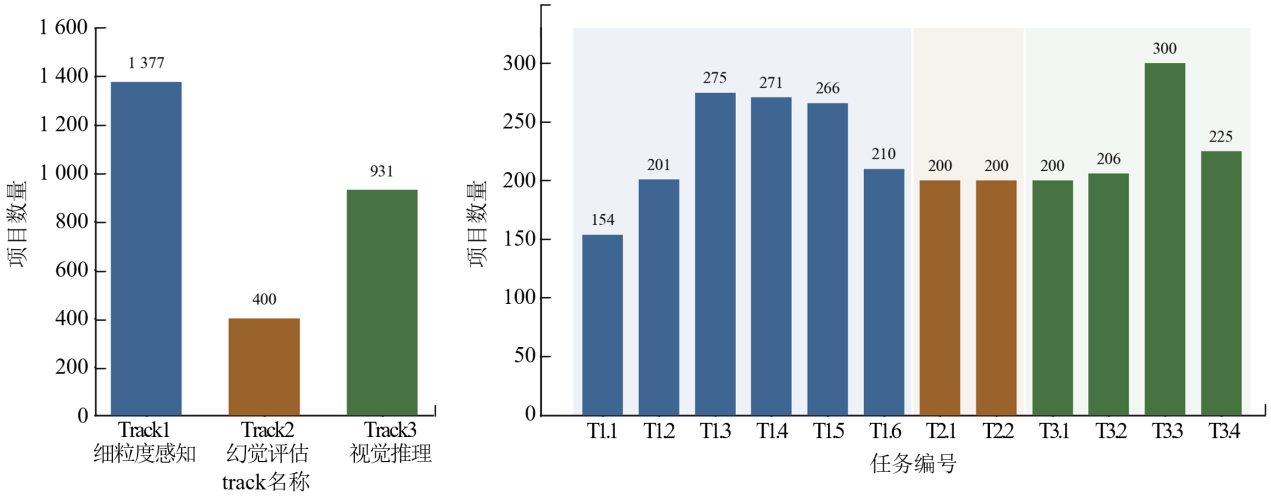
在业务场景分布上，输电、变电和配电场景并未被强制设置为完全等量，而是依据真实巡检任务中各类问题的出现形式和可判别性进行组织。例如，远距离输电线路图像更常用于小目标存在性判断、空间关系推理和密集目标计数；变电站和配电场景更常用于设备状态、部件属性和局部定位判断；红外图像主要用于热异常和状态类问题。该设计旨在使每类任务对应其在实际巡检中最典型的视觉困难，而不是追求形式上的均匀抽样。

从规模上看，MM-Power 各任务的具体分布见表 1 和图 2。Track 1 题目数量 1 377 题，占比为 50.85%，说明巡检业务仍首先依赖扎实的视觉感知基础；Track 3 题目数量 931 题，占比为 34.38%，表明本文并未将巡检问题简化为部件识别，而是保留了较强的推理要求；Track 2 题目数量 400 题，对应的是部署过程中更为敏感的可靠性问题。

表 1 MM-Power 各 Track 与任务规模统计

Table 1 Statistics of track and task scales in MM-Power

Track	任务编号	任务名称	题目数
细粒度感知	T1.1	属性识别	154
细粒度感知	T1.2	存在性判断	201
细粒度感知	T1.3	区域定位	275
细粒度感知	T1.4	检测框判别	271
细粒度感知	T1.5	点位置判别	266
细粒度感知	T1.6	缺陷类型辨析	210
幻觉评估	T2.1	负样本鉴别	200
幻觉评估	T2.2	盲从测试	200
视觉推理	T3.1	密集目标计数	200
视觉推理	T3.2	状态逻辑推理	206
视觉推理	T3.3	空间关系推理	300
视觉推理	T3.4	物体类别推理	225



(a) 3 个 Track 的题量统计

(b) 12 项具体任务的题量统计

图 2 MM-Power 的题量分布

Fig.2 Question distribution in MM-Power

1.3 任务设计与样例说明

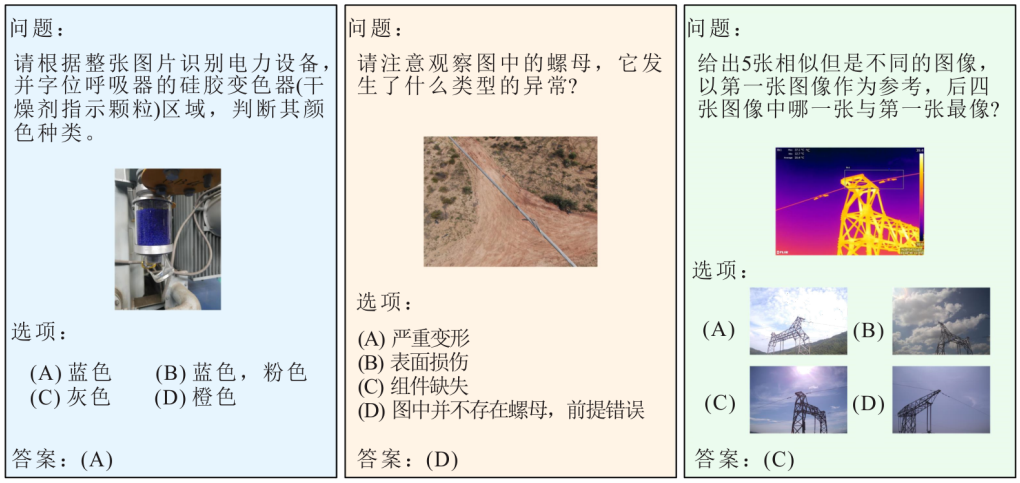
从任务设计看，Track 1 主要面向专业视觉对象的细粒度解析。属性识别任务要求模型区分设备子类或部件外观状态，例如不同材质绝缘子、不同颜色状态指示器等；存在性判断任务进一步强调极小目标和局部异常识别，是检验模型细节感知能力的关键；区域定位、检测框判别和点位置判别将识别要求进一步延伸至空间感知与局部定位；缺陷类型辨析任务则重点考查模型在同类设备内部区分相近缺陷模式的能力。上述任务共同构成了对“是否看得准、找得对”的基础考查。

Track 2 的重点则不在于发现更多异常，而在于识别模型是否会产生与视觉证据不一致的回答。负样本鉴别使用正常设备图像，但问题设置会诱导模型在多个故障选项中作答；盲从测试进一步构造与图像事实相矛盾的问题前提，用以考查模型在语言诱导与视觉证据发生冲突时的响应方式。这两类任务分别对应工业应用中最常见的两类风险，即无中生有的误报以及对错误前提的盲从。

Track 3 聚焦视觉逻辑推理。密集目标计数要求模型在重复结构和局部遮挡条件下完成准确计数；状态逻辑推理要求模型依据刀闸、箱门、表计等视

觉线索判断设备状态；空间关系推理关注树障、导线和设备之间的相对位置关系；物体类别推理则通过跨图匹配方式考查模型的类内抽象与对比判断能力。与 Track 1 相比，这部分任务不再局限于“识别对象”，而是进一步考查模型能否将局部视觉证据组织为有效判断。图 3 给出了 3 类代表性样例。

在题目设计过程中，本文遵循以下原则：第一，题干必须能够由图像中的可见证据直接支持，避免依赖图像外不可验证的运行记录或处置经验；第二，正确选项与干扰选项应在语义上具有可比性，干扰项优先选取同一业务场景下容易混淆的设备、部件或缺陷类别，以避免模型仅凭常识或排除法作答；第三，对于负样本和盲从测试，题目有意引入与图像事实不完全一致的语言前提，以模拟实际问答系统中用户描述错误或模型受到先验诱导的情况；第四，中英文题目保持语义等价，由具备电力巡检背景的研究人员复核专业术语翻译，尽量减少因翻译不一致造成的额外偏差。



(a) 属性识别 (b) 幻觉评估 (c) 视觉推理

图 3 MM-Power 代表性任务样例(对应的真实业务图像)

Fig.3 Representative task examples from MM-Power (presents real-world business images corresponding to typical tasks)

1.4 标注流程与评测协议

为保证基准的专业性与可复现性，MM-Power 的问题设计与答案校验由电力巡检领域专家和计算机视觉研究人员协同完成。每一道题目均采用 2~4 个候选项的选择题形式。标注流程包括初始题目设计、交叉复核和资深专家仲裁 3 个环节，仅保留结论一致且视觉证据明确的样本。对于容易引发歧义的题目，优先通过修改题干与选项消除解释空间，而不保留边界模糊的样本，以保证评测边界清晰。在模型评测协议方面，采用统一的电力巡检专家式提示模板，要求模型仅输出一个选项，不给出解释。其中英文模板如下：

“You are a senior power inspection expert with 20 years of experience. You possess keen visual recognition capabilities and are well-versed in the various components of transmission and distribution lines and substation equipment (such as insulators, towers, fittings, transformers, etc.) as well as their common defects (such as breakage, rust, foreign objects, bird nests, vegetation encroachment, etc.). Please carefully analyze the uploaded power inspection image and answer the following multiple-choice question. You must output only one of the options (A), (B), (C), (D) without providing any explanation. For example, output: (A).

Question:{}

Options:{}”

中文模板如下：

“你是一位拥有 20 年经验的资深电力巡检专家。你具备极敏锐的视觉识别能力，熟悉输配电线路、变电站设备的各种组件(如绝缘子、杆塔、金具、变压器等)及其常见缺陷(如破损、锈蚀、异物悬挂、鸟巢、植被入侵等)。

请仔细分析上传的电力巡检图像，并回答随后提出的单项选择题。你只用输出(A),(B),(C),(D)选项中的一个，不用对结果做任何解释。比如，输出：(A).

问题：{}

选项：{}”

为考查语言因素对工业场景理解的影响，分别使用英文与中文提示词进行独立测试。每个任务的准确率按答对题数与总题数之比计算；Track 得分取同一 Track 内任务准确率的非加权平均；总体分数取 12 项任务准确率的非加权平均。上述设定既保证了不同任务之间的可比性，也避免了题量差异对最终结论的过度影响。

2 实验设置

2.1 评测模型与选择依据

为覆盖测试时点具有代表性的视觉语言模型，本文参考截至 2026 年 1 月 1 日的公开视觉模型排行榜、模型可获得性及开闭源属性，选取 16 个模型进行评测，其中既包括 Google、OpenAI、Anthropic、腾讯等提供的闭源商业模型，也包括 Qwen、GLM、InternVL 等开源权重模型，具体见表 2。模型选择综合考虑公开评测表现、模型家族覆盖、开闭源属性、实际可访问性以及是否提供推理增强模式等因素。对于支持推理增强模式(reasoning/extended thinking)的模型，评测时保持对应模式开启，以贴近实际使用条件。

表 2 参与评测的视觉大模型及版本说明

Table 2 Vision-language models participating in the evaluation and their version details

简称	评测版本	提供方	开/闭源	思维模式
Gemini-3-Pro	gemini-3-pro-preview	Google	闭源	否
Gemini-3-Flash	gemini-3-flash-preview	Google	闭源	否
Gemini-2.5-Pro	gemini-2.5-pro	Google	闭源	否
GPT-5.1	gpt-5.1	OpenAI	闭源	否
Claude-Opus-4	claude-opus-4-20250514-thinking	Anthropic	闭源	是
ChatGPT-4o	chatgpt-4o-latest	OpenAI	闭源	否
Qwen3-VL-Think	qwen3-vl-235b-a22b-thinking	阿里	开源	是
GLM-4.5V	glm-4.5v	智谱	开源	否
Qwen3-VL-Inst	qwen3-vl-235b-a22b-instruct	阿里	开源	否
Qwen-VL-Max	qwen-vl-max-2025-08-13	阿里云	闭源	否
Qwen2.5-VL-72B	qwen2.5-vl-72b-instruct	阿里	开源	否
HunYuan-T1	hunyuan-t1-vision-20250916	腾讯	闭源	是
HunYuan-Large	hunyuan-large-vision	腾讯	闭源	否
Step-3	step_3	阶跃星辰	闭源	否
Claude-Sonnet-4	claude-sonnet-4-20250514-thinking	Anthropic	闭源	是
InternVL3-78B	internvl3-78b	OpenGVLab	开源	否

2.2 评价指标与统计方式

本文采用准确率作为统一评价指标。考虑到 MM-Power 的目标并非简单反映模型在某一高频任务上的总体命中率，而是希望不同能力维度在最终分数中得到相对均衡的体现，因此不采用加权平均或基于题量的总体正确率。后续结果分析主要围绕总体排名、3 个 Track 的能力差异、双语提示词差异以及典型薄弱任务的失败模式展开。

2.3 实验设置

为保证不同模型之间的可比性，本文在模型调用过程中尽量保持统一输入与解码设置。所有题目均以单图像、多选项问答形式输入模型，模型每次仅处理一道题目。除各模型接口自身不可见的内部图像编码、缩放或压缩机制外，本文在外部输入阶段不对图像进行人工裁剪、增强或局部放大；图像按照原始长宽比输入模型。对于因模型接口限制而需要调整尺寸的情况，仅采用等比例缩放，不改变图像内容和候选区域关系。

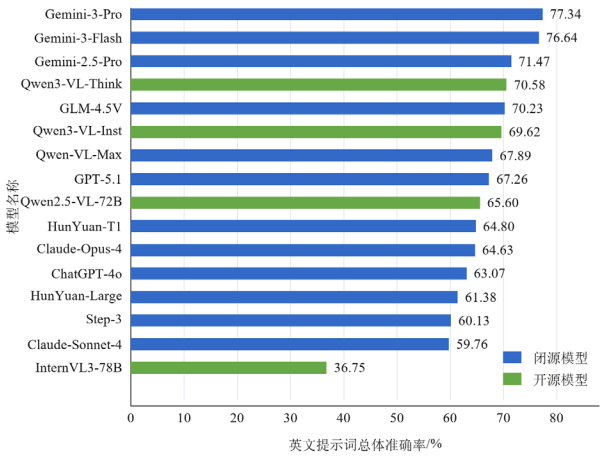
在文本解码参数方面，对于支持参数设置的模型，统一采用确定性或近似确定性设置。具体来说，将控制模型输出多样性的温度系数设置为 0、核采样数设为 1，最大输出长度限制在足以输出单个选项字母的范围内。对于不开放全部参数的商业接口，采用官方默认设置，并在模型版本表中记录实际调用版本。每道题目在同一模型、同一提示语言下调用一次，并以首次有效选项输出作为最终答案。

选项顺序在题库构建阶段完成固定。为降低选项位置对结果的影响，候选项在题目定稿时进行随机排列；同一道题在所有模型和中英文提示词条件下保持相同选项顺序，以保证模型间比较的一致性。对于输出非标准格式的结果，仅提取其中首次出现的合法选项字母；若无法提取合法选项，则判错。

3 结果分析与讨论

3.1 总体结果与第一梯队模型格局

图 4 为通用视觉榜单与 MM-Power 电力巡检基准的模型排名差异。结果表明，通用榜单排名与电力巡检任务表现并不完全一致。图 4(a)给出了各模型在英文提示词条件下的总体得分排序，表 3 进一步列出了中英文提示词下的总体得分及其差值。总体来看，Gemini 系列模型在 MM-Power 上取得相对领先的表现。其中，Gemini-3-Pro 的总体得分 77.34%，位居首位，Gemini-3-Flash 以 76.64%紧随其后，Gemini-2.5-Pro 的总体得分为 71.47%。在开源权重模型中，Qwen3-VL-Think 和 GLM-4.5V 分别达到 70.58%和 70.23%，与部分闭源接口模型处于相近水平。



(a) 各模型在 MM-Power 上的综合平均准确率

排名	模型名称	评分	置信区间	开发者
1	gemini-3-pro	1 324	±13	Google
2	gemini-2.5-pro	1 249	±7	Google
3	chatgpt-4o-latest-20250326	1 237	±7	OpenAI
4	gpt-4.5-preview-2025-02-27	1 226	±11	OpenAI
5	gemini-2.5-flash-preview-09-2025	1 225	±11	Google
6	gpt-5-chat	1 222	±8	OpenAI
7	o3-2025-04-16	1 219	±7	OpenAI
8	gpt-4.1-2025-04-14	1 216	±7	OpenAI
9	gemini-2.5-flash	1 214	±7	Google
10	claude-opus-4-20250514-thinking-16k	1 210	±15	Anthropic

(b) 相同或相近模型在通用视觉大模型榜单中的排名

图 4 通用视觉榜单与 MM-Power 电力巡检基准的模型排名差异

Fig.4 Ranking differences between a general vision-language leaderboard and the MM-Power power inspection benchmark

为进一步分析通用多模态基准与电力巡检专用基准之间的差异，本文将 MM-Power 上的模型排序与测试时点公开通用视觉语言模型榜单中的相对排序进行对照，结果见图 4(b)。可以看出，部分模型虽然在通用榜单中排名较高，但其在 MM-Power 上的相对表现并不完全一致。这一结果表明，通用视觉语言任务中的综合表现并不能完全代表模型在电力巡检场景中的细粒度感知、负样本鉴别和空间关系推理能力。因此，电力巡检场景仍需要专门的领域评测基准进行能力诊断；MM-Power 可作为通用多模态评测的补充，用于揭示通用榜单难以反映的行业适配性和可靠性差异。

结果表明，电力巡检场景对通用视觉大模型仍具有较高挑战。即使最佳模型的错误率仍接近 1/4，距离高可靠巡检辅助应用仍有明显差距。英文提示词下，最高与最低模型的总体准确率相差 40.59%，说明 MM-Power 具有较好的区分度，能揭示模型在工业视觉任务中的能力差异。此外，模型规模并非决定性能的唯一因素：Gemini-3-Flash 的表现接近 Gemini-3-Pro，而 InternVL3-78B 明显低于多数同量级或更小规模模型，表明训练数据、视觉编码策略、指令对齐和推理机制等因素可能共同影响工业场景表现。图 5 进一步展示了部分代表性定性结果。

表 3 各模型在中英文提示词下的总体准确率比较

Table 3 Comparison of overall accuracy of each model under Chinese and English prompts

模型	提供方	英文准确率/%	中文准确率/%	差值(中文−英文)/%
Gemini-3-Pro	Google	77.34	75.54	−1.81
Gemini-3-Flash	Google	76.64	75.44	−1.19
Gemini-2.5-Pro	Google	71.47	69.91	−1.55
Qwen3-VL-Think	阿里	70.58	69.64	−0.95
GLM-4.5V	智谱	70.23	70.76	+0.53
Qwen3-VL-Inst	阿里	69.62	69.32	−0.30
Qwen-VL-Max	阿里云	67.89	68.08	+0.19
GPT-5.1	OpenAI	67.26	67.13	−0.13
Qwen2.5-VL-72B	阿里	65.60	65.65	+0.05
HunYuan-T1	腾讯	64.80	68.16	+3.36
Claude-Opus-4	Anthropic	64.63	64.73	+0.10
ChatGPT-4o	OpenAI	63.07	60.97	−2.10
HunYuan-Large	腾讯	61.38	61.56	+0.18
Step-3	阶跃星辰	60.13	61.95	+1.82
Claude-Sonnet-4	Anthropic	59.76	57.68	−2.08
InternVL3-78B	OpenGVLab	36.75	41.21	+4.46

3.2 双语提示词影响与模型家族差异

从双语结果看，提示词语言对不同模型的影响并不一致。图 6 显示，Gemini-3-Pro、Gemini-3-Flash、Gemini-2.5-Pro 和 GPT-5.1 在中英文条件下的总体分数变化相对有限，说明这类模型在多语言工业视觉任务中的稳定性较好；与之相比，HunYuan-T1、Step-3 和 InternVL3-78B 在中文提示词下分别提升 3.35%、1.82% 和 4.46%，而 ChatGPT-4o 与 Claude-Sonnet-4 则在中文提示词下出现超过 2%的下降。

以上差异表明，提示词语言不仅会影响模型的语言解析过程，还可能改变其在视觉证据与领域术语之间的对齐方式，造成双语提示差异的原因可能来自 3 个方面。首先，电力巡检中存在大量专业设备名词和缺陷描述，不同语言下术语与模型预训练语料中的视觉概念对齐程度不同。例如，同一类缺



图 5　不同模型在 12 项任务示例题目上的表现

Fig.5　Performance of different models on example questions from 12 tasks

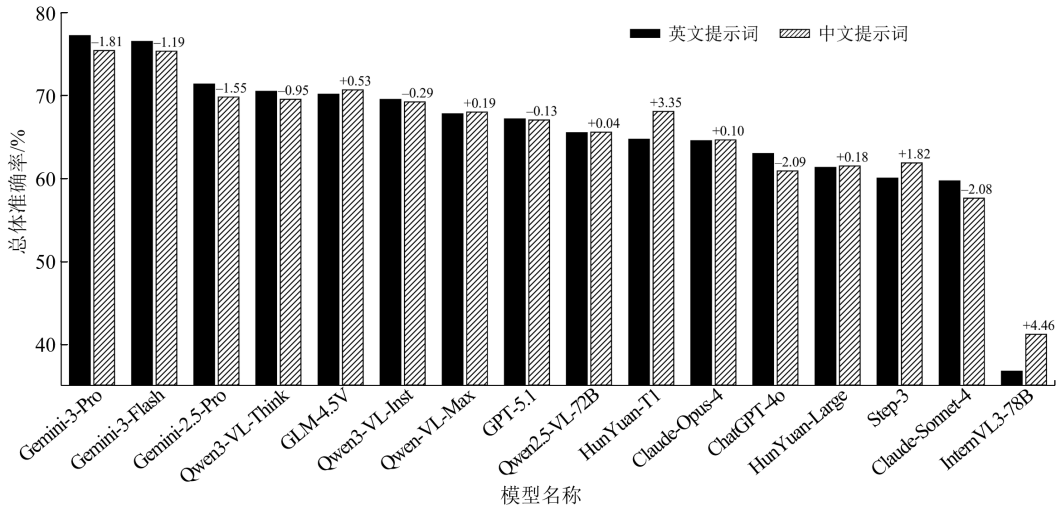


图 6　中英文提示词下的总体准确率对比(柱状图顶部标出了差值绝对值较大的模型，用于刻画语言因素对工业场景评测的影响)

Fig.6　Comparison of overall accuracy under Chinese and English prompts(models with larger absolute differences are labeled at the top of the bars to highlight the impact of language factors on evaluation in industrial scenarios)

陷在中文巡检语境中具有较固定的行业表达，而英文翻译则更接近通用视觉描述，二者会影响模型对候选项语义边界的理解。其次，不同模型的中文指令遵循能力存在差异，即使图像编码能力相近，模型也可能因对中文题干、否定表达或限制性输出要求的理解不同而产生不同答案。再次，中英文选项在文本长度、修饰结构和术语粒度上难以完全一致，改变模型对候选项的匹配倾向。因此，本文中的双语结果不应被理解为单纯的语言能力比较，而应视为视觉证据、领域术语表达和指令遵循共同作用下的工业场景评测结果。对于中文背景较强或中文语料覆盖较多的模型而言，中文提示词更接近其领域表述习惯，因此任务理解相对稳定；而对于以英文分布为主的模型，中文提示词可能引入额外的不确

定性，尤其是在专业设备名词和异常描述存在多种表达方式的情况下。这也提示，提示模板本身应纳入系统设计与评测考量，而不能简单视作与模型能力无关的外部因素。

3.3 幻觉评估、小目标感知与空间推理薄弱环节

MM-Power 的一个重要作用，在于能够将模型的失败模式具体暴露到任务层面。图 7 展示了英文提示词下 2 个幻觉评估任务的结果。相较而言，Task 2.2 盲从测试整体难度较低，多数模型准确率超过 80%；Task 2.1 负样本鉴别则明显更难，多个模型在“图中并无显著缺陷”的场景下仍倾向于给出具体故障名称。以 ChatGPT-4o 为例，其在该任务上的准确率仅为 28.9%，反映出模型在工业负样本场景中仍存在较强误报倾向。

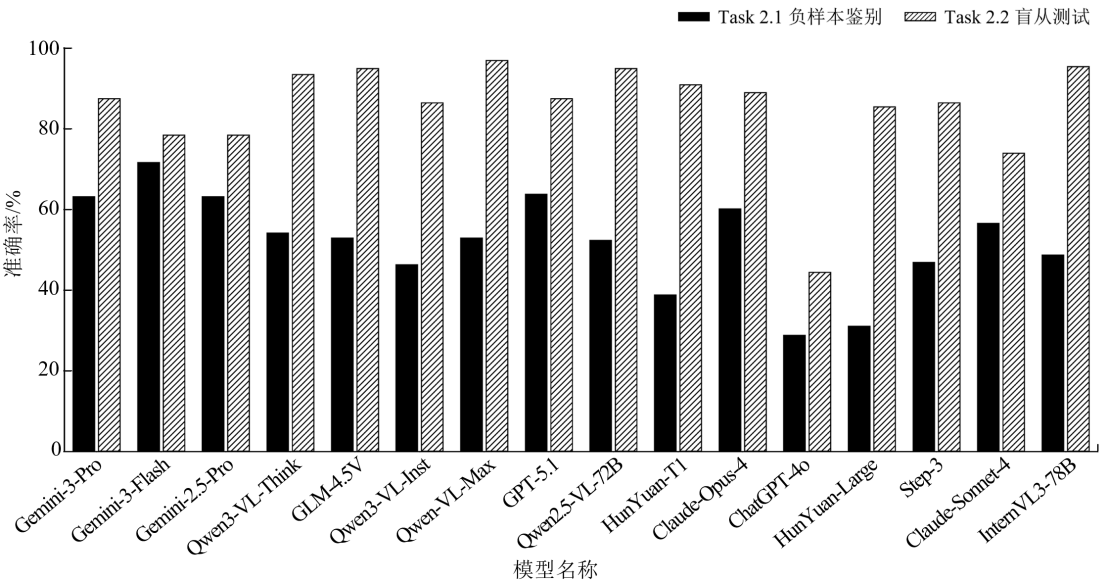


图 7 幻觉评估任务的模型表现差异

Fig.7 Performance differences of models on hallucination evaluation tasks

若将观察范围进一步扩展到全部 12 项任务，图 8 热力图所反映的结果更为清晰。图 8 显示，存在性判断(T1.2)、负样本鉴别(T2.1)以及若干空间推理相关任务在多数模型上均形成明显低谷，说明当前视觉大模型在工业场景中最欠缺的并不是对象类别的粗粒度识别，而是对局部细节、异常存在性以及结构化空间关系的稳定把握。结合热力图与原始结果可发现，从任务机制看，存在性判断、负样本鉴别和空间关系推理之所以形成低谷，主要与电力巡检图像的 3 类视觉困难有关。第一，许多缺陷或异常目标在整幅图像中所占像素比例较低，且常位于杆塔、导线、金具或背景纹理密集区域，模型在全图编码过程中容易丢失局部细节。第二，局部异常与正常结构之间的外观差异往往较小，例如轻微变形、局部遮挡、异物附着或规范着装差异，这类判断不仅需要识别对象类别，还需要对局部状态进行细粒度比较。第三，空间关系和状态逻辑问题要求模型同时完成目标识别、相对位置判断和业务语义映射，例如判断树障与导线的相对关系、设备开合状态或多个相似目标的计数，相较单纯分类更依赖结构化视觉推理。无论是杆塔局部异物、均压环缺陷，还是人员规范着装等细节判断，模型都可能因目标占比过小或背景干扰过强而产生遗漏。同时，视觉大模型在空间位置和状态关系理解上的稳定性也仍显不足。区域定位、检测框判别、点位判别、空间关系推理和状态逻辑推理等任务要求模型不仅判断“是什么”，还要进一步判断“在哪里”以及“意味着什么”，这仍是当前模型的突出难点。此外，抗幻觉能力与基础感知能力并不总是同步提升，某些模型在细粒度感知任务上得分并不低，但一旦进入负样本和错误前提场景便容易失稳，因此仅凭通用视觉得分判断工业部署可行性并不充分。

3.4 Track 级表现与工程启示

仅从总体准确率出发，并不能完整反映模型的能力结构。进一步按 Track 统计可见，不同模型在细粒度感知、幻觉评估和视觉推理 3 个维度上的表现差异较大。表 4 给出了若干代表性模型在中英文提示词下的 Track 平均准确率。Gemini-3-Pro 在 3

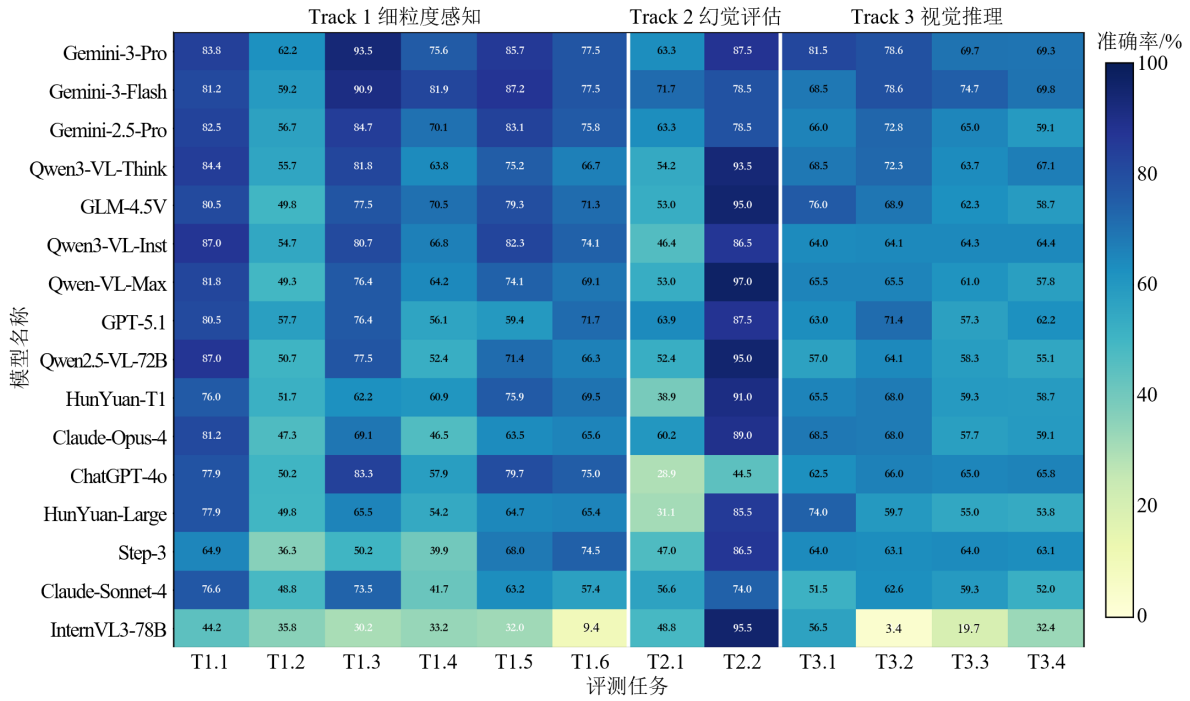


图 8 各模型在 12 项任务上的表现热力图

Fig.8 Heatmap of model performance across the 12 tasks

表 4 代表性模型在 3 个 Track 上的平均准确率

Table 4 Average accuracy of representative models across the three tracks

模型	平均准确率/%							
	Track 1-EN	Track 1-ZH	Track 2-EN	Track 2-ZH	Track 3-EN	Track 3-ZH	Overall-EN	Overall-ZH
Gemini-3-Pro	79.70	78.20	75.38	74.84	74.79	71.90	77.34	75.54
Gemini-3-Flash	79.65	80.46	75.09	65.82	72.90	72.74	76.64	75.44
Gemini-2.5-Pro	75.49	75.43	70.88	63.56	65.73	64.81	71.47	69.91
Qwen3-VL-Think	71.27	69.37	73.86	75.91	67.90	66.90	70.58	69.64
GLM-4.5V	71.47	72.86	74.01	75.16	66.48	65.42	70.23	70.76
GPT-5.1	66.97	68.03	75.68	78.28	63.48	60.20	67.26	67.13
ChatGPT-4o	70.68	69.28	36.71	39.30	64.82	59.35	63.07	60.97
HunYuan-T1	66.04	70.41	64.96	72.70	62.87	62.52	64.80	68.16

个维度上的得分相对均衡，英文条件下分别达到 79.70%、75.38%和 74.79%；Gemini-3-Flash 在细粒度感知方面略优于 Gemini-3-Pro，但中文条件下幻觉评估得分下降较明显；GLM-4.5V 与 Qwen3-VL-Think 在 3 个维度上保持了较稳定的中上水平，说明开源模型在工业基准上已具备较强竞争力；ChatGPT-4o 则呈现出“感知尚可、幻觉偏重”的失衡特征，其 Track 2 英文得分仅为 36.71%。

综合来看，视觉大模型在电力巡检中的薄弱环节并不完全一致。部分模型在感知任务上得分较高，但在负样本或错误前提场景中容易失稳；另一些模型在双语条件下稳定性较好，却在空间推理和状态判断任务中仍显不足。这表明电力巡检中的可靠应用不宜仅依据总体分数作出判断，还应同时关注模型在 3 类基础能力上的均衡性。对工程部署而言，视觉大模型更适合作为巡检辅助分析工具，而不宜直接承担自治故障判定职责；对模型研究而言，后续优化重点也不应停留在通用基准分数提升，而应更多面向高分辨率细节建模、局部区域对齐、反幻觉训练和结构化视觉推理能力的增强。

4 结论

1）本文围绕电力巡检场景，构建了视觉大模型多模态评测基准 MM-Power，并在真实业务图像

基础上形成覆盖 3 个 Track、12 项任务、2 708 道题目的完整测试体系。

2）与现有通用多模态基准相比，MM-Power 的主要特点在于面向电力巡检可靠性进行任务设计，从细粒度感知、视觉语义一致性与幻觉评估、视觉逻辑推理 3 个层面刻画模型能力，能够更有针对性地评估模型在小目标异常、负样本鉴别、错误前提诱导和空间关系推理等场景中的表现。

3）基于 16 个代表性视觉大模型的统一测试表明，当前最优模型在英文提示词下的总体准确率仅为 77.34%，在中文提示词下为 75.54%，距离高可靠巡检应用仍有明显差距。模型的共性短板主要集中在存在性判断、负样本鉴别和空间关系推理等任务上，而提示词语言又会对部分模型产生显著影响，说明工业场景中的模型能力不仅受架构和训练数据影响，也受到指令表达与领域术语对齐方式的制约。当前模型距离高可靠电力巡检辅助应用仍有明显差距。MM-Power 可为后续电力多模态模型的能力诊断、训练优化和安全部署提供参考。

参考文献 References

[1] LIU Y, DUAN H D, ZHANG Y H, et al. MMBench: is your multi-modal model an all-around player?[C]//Computer Vision-ECCV 2024: 18th European Conference. Milan, Italy: [s.n.], 2024: 216-233.

[2] YU W H, YANG Z Y, LUO L J, et al. MM-vet: evaluating large multimodal models for integrated capabilities[C]//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: ICML, 2024, 235: 57730-57754.

[3] LI B H, GE Y Y, GE Y X, et al. SEED-bench: benchmarking multimodal large language models[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024: 13299-13308.

[4] YUE X, NI Y S, ZHANG K, et al. MMMU: a massive multi-discipline multimodal understanding and reasoning benchmark for expert agi[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024: 9556-9567.

[5] LI B H, GE Y Y, CHEN Y, et al. SEED-Bench-2-Plus: benchmarking multimodal large language models with text-rich visual comprehension[J/OL]. arXiv preprint arXiv: 2404.16790, 2024.

[6] GOYAL Y, KHOT T, SUMMERS-STAY D, et al. Making the V in VQA matter: elevating the role of image understanding in visual question answering[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017: 6325-6334.

[7] 金浩远，闫云凤，张春林，等. 基于多目标序列捕捉与时空特征提取的电力人员行为识别算法[J]. 高电压技术，2025，51(12)：5788-5800.
JIN Haoyuan, YAN Yunfeng, ZHANG Chunlin, et al. Electric power personnel behavior recognition algorithm based on multi-object sequence capture and spatiotemporal feature extraction[J]. High Voltage Engineering, 2025, 51(12): 5788-5800.

[8] MENDU B, MBULI N. State-of-the-art review on the application of unmanned aerial vehicles (UAVs) in power line inspections: current innovations, trends, and future prospects[J]. Drones, 2025, 9(4): 265.

[9] VIEIRA E SILVA A L B, FELIX H D C, SIMÕES F P M, et al. InsPLAD: a dataset and benchmark for power line asset inspection in UAV images[J]. International Journal of Remote Sensing, 2023, 44(23): 7294-7320.

[10] FAISAL M A A, MECHETER I, QIBLAWEY Y, et al. Deep learning in automated power line inspection: a review[J]. Applied Energy, 2025, 385: 125507.

[11] 闫玮丹，齐冬莲，闫云凤，等. 面向电力领域的知识图谱与大模型融合关键技术及其典型应用[J]. 高电压技术，2025，51(4)：1747-1762.
YAN Weidan, QI Donglian, YAN Yunfeng, et al. Key technologies and typical applications of knowledge graph and large language model fusion in the power sector[J]. High Voltage Engineering, 2025, 51(4): 1747-1762.

[12] 范澜珊，刘云鹏，刘一瑾，等. 面向电力设备红外-可见光图像配准的自适应监督重训算法[J]. 高电压技术，2025，51(4)：1785-1800.
FAN Lanshan, LIU Yunpeng, LIU Yijin, et al. Adaptive algorithm with supervision and retraining for power equipment infrared and visible images registration[J]. High Voltage Engineering, 2025, 51(4): 1785-1800.

[13] 马超，李武峰，陈羽飞，等. ChatGPT 类大语言模型赋能电力标准数字化转型的核心技术、技术特征及应用展望[J]. 高电压技术，2025，51(4)：1727-1746.
MA Chao, LI Wufeng, CHEN Yufei, et al. Core technologies, technical characteristics, and application prospects of ChatGPT class big language models on the digital transformation of standards in electric power field[J]. High Voltage Engineering, 2025, 51(4): 1727-1746.

[14] 黄泽泽，董晓杰，李少龄，等. 基于 YOLOv7 的均压环缺陷及倾斜角度检测方法[J]. 电力信息与通信技术，2025，23(4)：25-34.
HUANG Zeze, DONG Xiaojie, LI Shaoling, et al. Detection method for defects and tilt angle of grading ring based on YOLOv7[J]. Electric Power Information and Communication Technology, 2025, 23(4): 25-34.

[15] 刘威，卢妍洁，高焜，等. 基于 YOLO-Efficient 的高空作业安全带检测[J]. 电力信息与通信技术，2025，23(7)：30-37.
LIU Wei, LU Yanjie, GAO Kun, et al. Aerial work safety belt detection based on YOLO-efficient[J]. Electric Power Information and Communication Technology, 2025, 23(7): 30-37.

[16] 赵勇彪，张雨柔，邵晓琪，等. 基于 SCAS-YOLO 的变压器电力设备渗漏油检测研究[J]. 电力信息与通信技术，2025，23(8)：41-52.
ZHAO Yongbiao, ZHANG Yurou, SHAO Xiaoqi, et al. Study on oil leakage detection of transformer power equipment based on SCAS-YOLO[J]. Electric Power Information and Communication Technology, 2025, 23(8): 41-52.

[17] 卞建鹏，李凡，郝培旭，等. 复杂环境下输电线路绝缘子的破损识别与定位[J]. 高电压技术，2022，48(2)：681-688.
BIAN Jianpeng, LI Fan, HAO Peixu, et al. Damage identification and location of transmission line insulator in complex environment[J]. High Voltage Engineering, 2022, 48(2): 681-688.

[18] 吕中亮，夏可文，卢震宇，等. 密集特征加权融合网络用于复杂天气下的绝缘子表面缺陷检测[J]. 高电压技术，2025，51(3)：1114-1125.
LYU Zhongliang, XIA Kewen, LU Zhenyu, et al. Weather-robust insulator surface defect detection with dense feature weighted fusion network[J]. High Voltage Engineering, 2025, 51(3): 1114-1125.

[19] 葛召，李洪文，刘海峰，等. 基于 YOLO-GSS 的输电线路边缘端

实时缺陷检测方法[J]. 高电压技术，2025，51(2)：669-677.
GE Zhao, LI Hongwen, LIU Haifeng, et al. Real-time defect detection method for edge-end of transmission line based on YOLO-GSS[J]. High Voltage Engineering, 2025, 51(2): 669-677.

[20] 刘宏志，黄振龙，邱彬，等. 基于改进 YOLOv11n 的输电线路主要缺陷检测方法[J/OL]. 高电压技术，2025: 1-12[2025-02-05]. https://doi.org/10.13336/j.1003-6520.hve.20241908.
LIU Hongzhi, HUANG Zhenlong, QIU Bin, et al. Major defect detection method for transmission lines based on improved YOLOv11n[J/OL]. High Voltage Engineering, 2025: 1-12 [2025-02-05]. https://doi.org/10.13336/j.1003-6520.hve.20241908.

[21] 邬开俊，徐泽浩，单宏全. 基于 FasterNet 和 YOLOv5 改进的玻璃绝缘子自爆缺陷快速检测方法[J]. 高电压技术，2024，50(5)：1865-1876.
WU Kaijun, XU Zehao, SHAN Hongquan. Rapid detection method for self-exploding defects in glass insulators based on improved FasterNet and YOLOv5[J]. High Voltage Engineering, 2024, 50(5): 1865-1876.

[22] GU Z P, ZHU B K, ZHU G B, et al. AnomalyGPT: detecting industrial anomalies using large vision-language models[C]//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024: 1932-1940.

[23] LI Y F, DU Y F, ZHOU K, et al. Evaluating object hallucination in large vision-language models[C]//Proceedings of the 2023 conference on empirical methods in natural language processing. Singapore, Singapore: Association For Computational Linguistics, 2023: 292-305.

[24] PHAM N, SCHOTT M. H-POPE: Hierarchical polling-based probing evaluation of hallucinations in large vision-language models[C]//Workshop on Statistical Foundations of LLMs and Foundation Models. 2024.



闫云凤(通信作者)
1988—，女，博士，副教授，博导
主要从事计算机视觉、运行态势感知、分布式发电设备异常监测方面的研究工作
E-mail: 21210004@zju.edu.cn

YAN Yunfeng
Ph.D.
Associate professor
Corresponding author

收稿日期 2026-04-18 修回日期 2026-06-08 编辑 卫李静