# ·应用规程约束风险感知偏好强化学习的·电力设备运维方法

罗海宸，李明轩，荣命哲，杨爱军，王小华
（西安交通大学电工材料电气绝缘全国重点实验室，710049，西安）

**摘要**：针对电力设备智能运维中多模态证据融合困难、通用大模型运维规程约束不足，以及高风险任务输出偏好难以控制等问题，提出一种面向电力设备运维的规程约束风险感知偏好强化学习方法。该方法将任务问题、现场多模态证据、规程约束及设备与场景元信息组成结构化运维输入，建立状态判断正确性、证据一致性、规程一致性、风险保守性和建议可执行性五维偏好裁决标准，并将综合偏好得分、安全评分差距、任务复杂度和证据充分性纳入动态间隔偏好优化，使模型输出更保守、可复核且符合运维流程。所构建的数据集覆盖6种任务类别，包括输电线路感知、隔离开关状态监测、电力现场安全监测、变电站监测、电力器件识别和发电厂综合监测。研究结果表明：所提方法在电力设备运维验证集上的选择题准确率达到87.89%，相较初始策略模型准确率提升了13.56个百分点，相较近端策略优化模型提升了5.54个百分点，从而验证了所提方法在离线电力设备运维场景验证中的有效性。该研究为提升多模态大模型在高风险电力设备运维场景中的规程遵循能力与安全输出能力提供了一种可行方案。

**关键词**：电力设备运维；偏好强化学习；规程约束；风险感知动态间隔
**中图分类号**：TM 507 **文献标志码**：A
**DOI**：10.7652/xjtuxbXXXXXX001 **文章编号**：0253-987X(XXXX)XX-0001-11

# A Power Equipment Maintenance Method Using Rule-Constrained Risk-Aware Preference Reinforcement Learning

LUO Haichen, LI Mingxuan, RONG Mingzhe, YANG Aijun, WANG Xiaohua
(State Key Laboratory of Electrical Insulation and Power Equipment, Xi'an Jiaotong University, Xi'an 710049, China)

**Abstract**: To address difficulties in multimodal evidence fusion, inadequate enforcement of maintenance rules in general-purpose large models, and hard-to-control output preferences in high-risk power equipment maintenance, this paper proposes a rule-constrained risk-aware preference reinforcement learning method. The method organizes task queries, multimodal field evidence, maintenance rules, and equipment and scenario metadata into structured inputs. It establishes a five-dimensional preference criterion covering state-assessment correctness, evidence consistency, rule compliance, risk conservativeness, and recommendation actionability. Comprehensive preference scores, safety-score gaps, task complexity, and evidence sufficiency are incorporated into dynamic-margin preference optimization to encourage conservative, verifiable, and procedure-compliant outputs. The constructed dataset covers six task categories: transmission-line perception,

收稿日期：2026-04-23。
网络出版日期：XXXX-XX-XX 网络出版地址：

disconnector-state monitoring, on-site safety monitoring at power facilities, substation monitoring, power-component recognition, and comprehensive power-plant monitoring. Experimental results show that the proposed method achieves a multiple-choice accuracy of 87.89% on the power equipment maintenance validation set, exceeding the initial policy model and the proximal policy optimization model by 13.56 and 5.54 percentage points, respectively. These results validate the method for offline power equipment maintenance scenarios. This study provides a feasible approach to improving rule compliance and safe outputs of multimodal large models in high-risk maintenance.

**Keywords**: power equipment maintenance; preference reinforcement learning; rule constraint; risk-aware dynamic margin

电力设备是电力系统的重要组成,其安全稳定运行直接关系电网的可靠性[1]。传统运维方式存在人力投入大、效率低和成本高等问题[2-3],而智能运维能够辅助掌握设备状态、研判运行趋势并制定检修计划[4]。

现有电力设备的智能运维研究主要集中于红外故障诊断、视频图像识别、深度学习故障诊断、一次设备健康管理、绝缘子缺陷检测和隔离开关状态感知等专用任务[5-12]。这些方法在设备对象和缺陷类型固定,标注数据充分时具有较好效果,但对于跨设备、跨场景和开放式任务的处理能力较为有限。

通用多模态大模型提升了图文和视频的理解能力[13-19],但在高风险电力设备运维任务中仍需满足判断有证据、建议合规、高风险保守的要求。现有偏好优化方法通常未显式建模设备对象、运维规程、证据充分性和风险边界,因此需要进一步开展面向电力设备运维的安全偏好对齐。

除多模态理解外,电力设备运维任务还具有高风险和强规程约束特征。PPO-RLHF通过奖励模型和策略梯度更新对大模型进行偏好对齐[20],但训练流程复杂,且奖励模型与策略模型的目标解耦,容易在专业高风险场景中引入奖励偏差。直接偏好优化等方法进一步证明,可以不显式训练奖励模型而直接优化偏好目标[21],但多数方法是面向通用问答或通用对话场景,通常未显式建模设备对象、电压等级、操作规程、现场证据充分性和风险保守性等电力设备运维特征。然而,对于电力设备运维而言,一个回答是否优质并不只取决于语言流畅度或一般正确性,还取决于其是否基于现场证据、是否符合操作规程、是否在证据不足时主动提示人工复核。

综上所述,现有相关研究与高安全行业智能模型的实际需求之间仍存在三方面差距:第一,专用智能模型难以统一处理多模态、多任务和开放式运维需求;第二,通用多模态大模型和电力领域多模态基座模型虽然具备较强的多模态理解能力,但仍需要进一步面向运维规程和高风险决策边界进行偏好对齐;第三,现有偏好对齐方法通常未将设备对象、规程一致性、现场证据充分性和保守输出原则系统纳入偏好裁决与策略优化约束。

针对以上不足,本文提出一种规程约束风险感知偏好强化学习方法,并围绕典型电力设备运维任务开展离线场景验证。主要包括以下贡献。

(1)将电力设备运维样本表示为任务问题—多模态现场证据—规程约束—设备场景元信息的结构化运维输入,为后续偏好强化学习提供面向运维流程的状态空间表达。

(2)提出一种电力设备运维规程约束风险感知动态间隔偏好的优化方法。该方法在广义人类反馈强化学习框架下,不训练显式奖励模型,而是构建统一的电力设备运维偏好评分标准,对候选回答的状态判断、证据使用、规程一致性、风险保守性和处置建议进行离线裁决;在此基础上,以综合偏好得分确定优劣回答,并将任务复杂度、安全评分差距和证据不足程度分别写入动态间隔与样本权重,使策略更新直接服务于判断有证据、建议合规、高风险保守的运维需求。

(3)构建电力设备运维验证集,并从选择题准确率、生成质量、策略收敛特性、偏好机制消融、偏好质量评估、类别数据移除影响等方面评估所提方法,同时明确离线验证与真实工程部署之间的边界。

## 1 电力设备运维多模态基座模型

本文使用课题组前期工作PowerLLaVA[22]作为基座模型,其能够统一处理巡检图像、红外图像、现场视频和运维文本等输入,并生成文本形式的设

备状态判断、异常描述和规程问答结果。

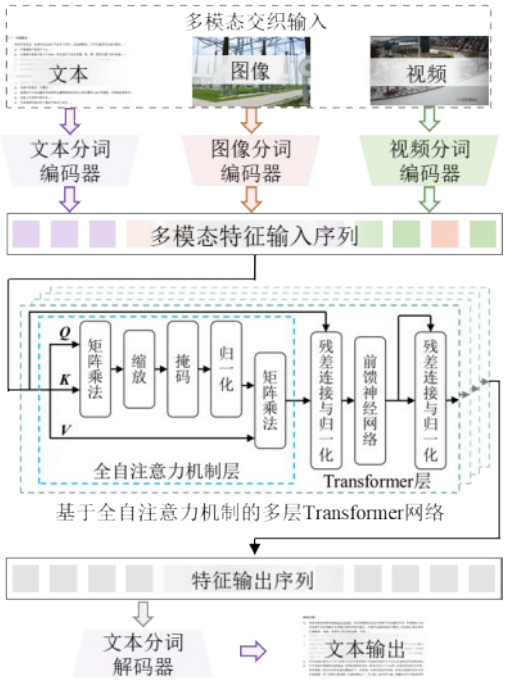



图 1　电力设备运维多模态基座模型结构

Fig. 1　Architecture of the multimodal base model for power equipment maintenance

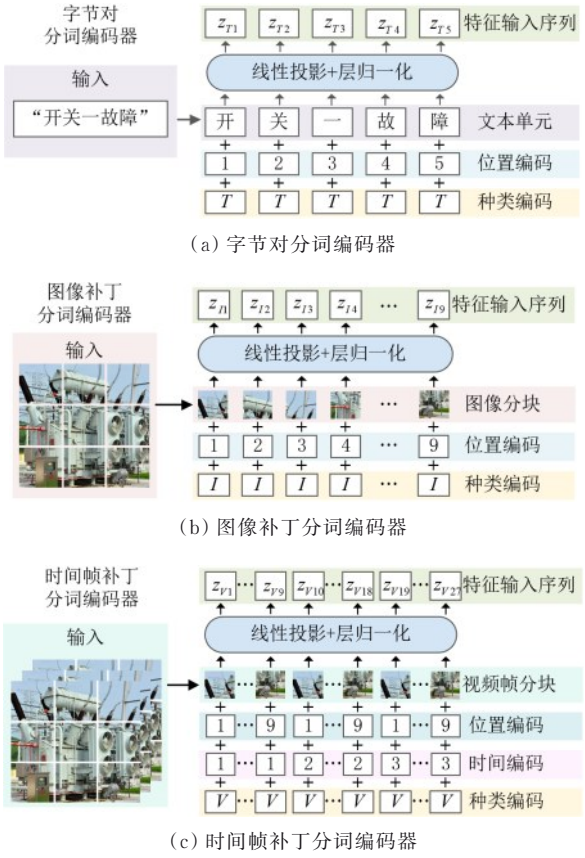
### 1.1　网络模型

电力设备运维多模态基座模型采用统一的模型架构处理不同运维任务和多模态数据，其结构如图 1 所示。模型首先利用特定模态的编码器，将来自文本、图像和视频的输入数据转换为维度一致的特征序列并进行拼接；随后，使用基于全自注意力机制的 Transformer 主干网络[23]对特征序列进行建模；最后，通过文本解码器生成设备状态判断、风险提示或处置建议。

### 1.2　分词器

分词器用于将不同模态数据与特征序列进行互相转换，并承担多模态输入的统一表示。编码器负责将自然语言、图像和视频等数据映射为特征序列；解码器则将 Transformer 主干网络输出的隐状态转换为文本单元。考虑到电力设备运维场景中的可复核性和可执行性要求，本文模型最终输出统一为文本形式。

模型分别采用字节对编码器、图像补丁编码器和时间帧补丁编码器处理文本、图像和视频的输入[24-26]，并通过模态类型编码区分数据来源。编码后的不同模态特征序列被映射至统一维度后输入 Transformer 主干网络，最终由文本解码器生成设备状态判断、异常描述或规程问答结果。




(a) 字节对分词编码器

(b) 图像补丁分词编码器

(c) 时间帧补丁分词编码器

图 2　处理多模态数据的各类编码器

Fig. 2　Encoders for processing different multimodal data

如图 2 所示，不同任务的特征输入序列 $\mathbf{z}$ 可以由文本单元序列 $\mathbf{z}_{Ti}$、图像单元序列 $\mathbf{z}_{Ii}$ 和视频单元序列 $\mathbf{z}_{Vi}$ 等模态的不同组合构成，其中 $i$ 为输入序列编号，下标 $I$、$V$ 和 $T$ 分别表示图像、视频和文本描述。

统一多模态结构能够为不同电力设备运维任务提供共享表示基础。本文后续工作并不改变其主干结构，而是通过规程约束风险感知偏好强化学习进一步约束其在高风险任务中的输出偏好。

### 1.3　基于全自注意力机制的 Transformer 网络

模型中负责跨模态语义建模的主干网络为基于全自注意力机制的 Transformer 网络，具体结构如图 1 所示。特征输入序列经过多个串行的 Transformer 层进行计算，每个 Transformer 层包括 1 个全自注意力机制层、2 个残差连接与归一化层和 1 个前馈神经网络层。具体来说，全自注意力机制层的表达式为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{n}}\right)\boldsymbol{V} \quad (1)$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别为特征输入序列经不同线性变

换得到的查询矩阵、键矩阵和值矩阵；$n$为输入特征维度；Attention(•)为全自注意力机制；Softmax(•)为归一化指数函数。

在每轮生成迭代中，最后一层Transformer输出的隐藏状态经语言模型头和Softmax函数映射为词表概率分布。模型根据该分布预测当前文本单元，并将预测结果拼接到输入序列末尾继续迭代，直至生成特殊结束符(＜eos＞)，并最终得到完整文本回答。

## 2 电力设备运维规程约束风险感知偏好强化学习方法

在电力智能运维任务中，PowerLLaVA模型虽然已经具备基础理解和问答能力，但在高风险场景下仍需要进一步对齐其输出偏好。因此，本文提出电力设备运维规程约束风险感知动态间隔偏好优化方法(OM-RDMPO)。该方法使用若干个优劣问答对作为数据进行训练，遵循策略采样—离线偏好裁决—策略更新的闭环，采用和传统强化学习相似的思想，但不训练独立奖励模型。

### 2.1 统一偏好评分标准与离线裁决

与通用问答不同，电力设备运维任务样本通常同时包含用户问题、现场图像或视频、规程条款以及设备场景信息。本文将一个训练样本表示为

$$x=(q,M,R,e);\ M=\{I,V,T\} \tag{2}$$

式中：$x$为结构化运维输入；$q$表示用户问题或运维任务指令；$R$表示操作规程、安全条款或操作票信息；$M$表示现场多模态证据集合；$e$表示设备与场景元信息，包括设备类型、电压等级、任务类别、运行状态以及是否涉及带电或操作风险等属性。该定义使后续偏好裁决能够显式利用电力设备运维中的设备对象、规程边界和现场证据。

基于以上结构化运维输入数据，本节首先构建统一评分标准用以离线裁决优劣回答。将PowerLLaVA基座模型记为初始策略$\pi_0$，对于结构化运维输入$x$，初始策略$\pi_0$在采样温度$T_s$下生成$K$个候选回答，表达式为

$$\begin{cases} \mathcal{Y}_0(x)=\{y_1,y_2,\cdots,y_N\} \\ y_j\sim\pi_0(\cdot|x,T_s),j=1,2,\cdots,N \end{cases} \tag{3}$$

式中：$\mathcal{Y}_0(x)$为候选回答集合；$y_j$为第$j$个候选回答；$N$为候选回答数量，取值为2；$\pi_0(\cdot|x,T_s)$为在$x$和采样温度$T_s$条件下，由初始策略生成回答的概率分布。本文取$T_s=0.7$，以在候选回答之间形成适度差异，同时避免过高随机性导致的明显无效回答。

为计算策略更新目标，需要比较当前策略对优劣回答的相对概率。为避免长回答因文本单元更多而天然获得更低序列概率，本文采用长度归一化对数概率作为策略评分，表达式为

$$l_\pi(y|x)=\frac{1}{|y|}\sum_{t=1}^{|y|}\log\pi(y_t|x,y_{<t}) \tag{4}$$

式中：$l_\pi(y|x)$表示当前策略$\pi$对回答$y$的长度归一化对数概率；$|y|$为回答长度；$y_t$为第$t$个文本单元；$y_{<t}$为此前已生成的上下文。

离线偏好裁决采用少量人工种子标注，以及通用领域大模型统一评分的方式完成。首先，由电力设备运维相关人员对少量样本进行人工排序和维度打分，形成种子示例；随后，将结构化运维输入$x$、候选回答、人工种子示例和统一评分规则共同输入通用领域大模型，使其对每个候选回答给出5个维度的原始分数、简要理由和最终偏好。裁决时不提供候选回答来源或模型名称，并随机打乱候选回答顺序，以减少位置偏差和模型来源偏差。

5个维度均采用0～5分制，并归一化到[0,1]区间，表达式为

$$d_k(y,x)=\frac{\widehat{d_k}(y,x)}{5},k=1,2,3,4,5 \tag{5}$$

式中：$d_k(y,x)$为根据统一评分规则给出的第$k$个维度原始分数；$d_k(y,x)$为归一化后的维度得分。5个维度含义如下：$d_1$为状态判断正确性，衡量回答对设备状态、故障类别或风险对象的判断是否正确；$d_2$为证据使用一致性，衡量回答是否基于图像、视频、文本和规程证据，是否存在无依据臆测；$d_3$为规程一致性，衡量回答是否符合运维规程、安全条款和操作票要求；$d_4$为风险保守性，衡量回答在高风险或证据不足时是否提示不确定性、人工复核和风险边界；$d_5$为处置建议可执行性，衡量回答是否给出符合现场运维流程的可执行建议。一般地，0分表示严重错误或违反安全要求，3分表示部分正确但证据或规程说明不足，5分表示判断正确、证据充分、规程合规且建议可执行。

基于5个维度，回答的综合偏好得分$S(y,x)$定

义为

$$S(\boldsymbol{y},\boldsymbol{x})=\sum_{k=1}^{5}\mu_k d_k(\boldsymbol{y},\boldsymbol{x});\ \sum_{k=1}^{5}\mu_k=1 \quad (6)$$

式中：$\mu_k$为第$k$个维度的权重。本文所有优劣回答判断均依据$S(\boldsymbol{y},\boldsymbol{x})$完成，即综合得分最高的候选回答记为更优回答$\boldsymbol{y}^+$，综合得分最低的候选回答记为更劣回答$\boldsymbol{y}^-$，表达式为

$$\boldsymbol{y}^+=\arg\max_{\boldsymbol{y}_j\in\mathcal{Y}_0(\boldsymbol{x})} S(\boldsymbol{y}_j,\boldsymbol{x}), \quad (7)$$

$$\boldsymbol{y}^-=\arg\min_{\boldsymbol{y}_j\in\mathcal{Y}_0(\boldsymbol{x})} S(\boldsymbol{y}_j,\boldsymbol{x}). \quad (8)$$

如果2个候选回答的综合得分差小于阈值$\varepsilon_S=0.1$，则该样本被标记为低置信偏好样本，将进入人工复核或从偏好训练集中剔除。

本文最终构建的偏好学习数据集包含4.16万条候选回答偏好数据，其中人工偏好数据约3 000条，图像约1.2万张，视频约2 000个。本文将数据集的85%作为训练集，15%作为验证集，为避免训练集与验证集之间的数据泄漏，本文在数据划分前采用样本组级去重方式。具体而言，同一原始运维样本及其派生的多个候选回答偏好对划入同一数据组。进一步地，对问题文本、规程依据和标准答案进行规范化处理，去除选项编号、标点和模板化表述差异后，将语义相同或近似重复的样本归入同一组，并按组进行训练集和验证集的划分。偏好数据覆盖输电线路感知、隔离开关状态监测、施工现场安全监测、变电站检测、电力器件识别、发电厂综合检测及其他任务，类别分布如图3所示。

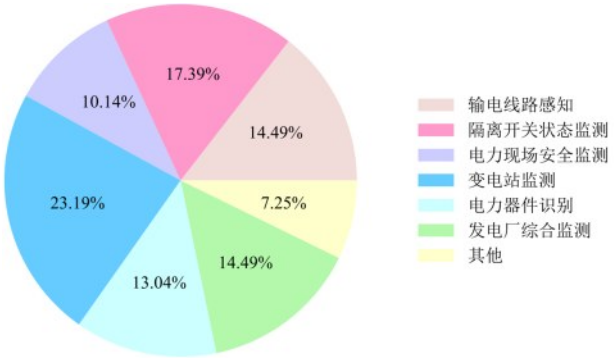



图3　电力设备运维偏好数据任务类别分布

Fig. 3　Distribution of task categories in the preference dataset for power equipment maintenance

## 2.2　统一评分驱动的动态间隔构造

间隔用于规定更优回答与更劣回答在策略概率上应被拉开的最小幅度。对于普通、低风险且优劣差异不明显的样本，过大的间隔可能导致模型过度更新；对于高风险、复杂且优劣回答在安全边界上差异明显的样本，则需要更大的间隔强化模型对安全偏好的学习。因此，本文不采用固定间隔，而是根据统一偏好评分标准派生任务复杂度和安全评分差距，并据此自适应构造动态间隔。

首先，本文用优劣回答的平均长度近似刻画任务复杂度。长回答通常对应多步判断、原因分析或操作建议，因而需要更强的偏好约束。任务复杂度因子$c_{\text{task}}(x)$定义为

$$c_{\text{task}}(\boldsymbol{x})=\min\left(\frac{|\boldsymbol{y}^+|+|\boldsymbol{y}^-|}{2L_0},c_{\max}\right) \quad (9)$$

式中：$|\boldsymbol{y}^+|$、$|\boldsymbol{y}^-|$分别为更优回答和更劣回答的长度；$L_0$为长度归一化常数，其取值为256，接近候选回答长度的经验中位量级，使一般状态判断类短回答与包含原因分析、规程解释和处置建议的长回答在任务复杂度上形成区分；$c_{\max}$为复杂度上限，取值为2，用于避免极长回答所导致的间隔过大。

其次，本文从统一偏好评分中的规程一致性和风险保守性2个维度派生安全评分。由于这2个维度最直接反映电力设备运维安全边界，安全评分$S_{\text{safe}}(\boldsymbol{y},\boldsymbol{x})$定义为

$$S_{\text{safe}}(\boldsymbol{y},\boldsymbol{x})=v_3 d_3(\boldsymbol{y},\boldsymbol{x})+v_4 d_4(\boldsymbol{y},\boldsymbol{x}), \quad (10)$$

式中：$d_3(\boldsymbol{y},\boldsymbol{x})$为规程一致性得分；$d_4(\boldsymbol{y},\boldsymbol{x})$为风险保守性得分；$v_3$、$v_4$分别为安全评分内部权重，其中，$v_3=0.6$，该设置使规程一致性在安全评分中占略高权重，以体现电力设备运维任务中操作规程和安全条款的硬约束属性；$v_4=0.4$，该设置使风险保守性在高风险或证据不足场景下约束模型，以避免产生过度确定判断。因此，安全评分并不是额外引入的新标准，而是由2.2.1节的统一评分标准直接得到。

基于安全评分，优劣回答之间的安全评分差距$\Delta S_{\text{safe}}(\boldsymbol{x})$定义为

$$\Delta S_{\text{safe}}(\boldsymbol{x})=\max\left(0,S_{\text{safe}}(\boldsymbol{y}^+,\boldsymbol{x})-S_{\text{safe}}(\boldsymbol{y}^-,\boldsymbol{x})\right) \quad (11)$$

当该值较大时，说明2个回答在电力设备运维安全边界上的差异明显，训练时应更强地拉开二者的策略概率。综合任务复杂度和安全评分差距，本文构造动态间隔$m_{\text{om}}(\boldsymbol{x})$，表达式为

$$m_{\text{om}}(\boldsymbol{x})=\text{clip}(m_0+\alpha c_{\text{task}}(\boldsymbol{x})+\beta\Delta S_{\text{safe}}(\boldsymbol{x}),m_{\min},m_{\max}) \quad (12)$$

式中：$m_0$为基础间隔，取值为0.2；clip(·)为阶段函

数；$\alpha$、$\beta$分别为任务复杂度和安全评分差距的权重，取值分别为0.1、0.4；$m_{\min}$和$m_{\max}$分别为间隔下限和上限，取值分别为0.05、0.6。式(12)表明，对于任务更复杂、优劣回答安全边界差异更明显的样本，模型需要在训练中更充分地区分优劣回答。

### 2.3 证据加权偏好损失与策略锚定优化目标

在得到更优回答$\boldsymbol{y}^+$、更劣回答$\boldsymbol{y}^-$以及动态间隔$m_{\text{om}}(\boldsymbol{x})$后，本文进一步构造策略优化目标。首先，定义当前策略对优劣回答的归一化对数概率差$g_\pi(\boldsymbol{x})$，表达式为

$$g_\pi(\boldsymbol{x})=l_\pi(\boldsymbol{y}^+|\boldsymbol{x})-l_\pi(\boldsymbol{y}^-|\boldsymbol{x}) \tag{13}$$

若该值小于动态间隔$m_{\text{om}}(\boldsymbol{x})$，说明模型尚未充分区分优劣回答，需要继续提高更优回答的相对概率。除动态间隔外，证据不足样本还应在损失函数中获得更高权重。将现场证据充分性记为$C_{\text{evi}}(\boldsymbol{x})\in[0,1]$，其由离线偏好裁决模型在统一评分过程中同步给出，并由人工种子示例进行校准，用于衡量图像、视频、文本和规程信息是否足以支持确定性判断，打分时采用5分制并归一化到$[0,1]$区间。证据不足程度$U_{\text{evi}}(\boldsymbol{x})$可定义为

$$U_{\text{evi}}(\boldsymbol{x})=1-C_{\text{evi}}(\boldsymbol{x}) \tag{14}$$

$C_{\text{evi}}(\boldsymbol{x})$越低，说明现场证据越不足，模型越应避免给出过度确定的判断。基于证据不足程度，样本权重$\omega_{\text{om}}(\boldsymbol{x})$定义为

$$\omega_{\text{om}}(\boldsymbol{x})=1+\eta U_{\text{evi}}(\boldsymbol{x}) \tag{15}$$

式中：$\eta$为缩放系数，取值为1。该权重使证据不足的样本在训练中获得更强约束，促使模型优先学习安全、保守和可复核的输出偏好。

最终，规程约束风险感知动态间隔偏好损失$\mathcal{L}_{\text{OM-RDMPO}}(\pi)$可定义为

$$\mathcal{L}_{\text{OM-RDMPO}}(\pi)=-E_{(\boldsymbol{x},\boldsymbol{y}^+,\boldsymbol{y}^-)}\left[\omega_{\text{om}}(\boldsymbol{x})\log\sigma\left(\frac{g_\pi(\boldsymbol{x})-m_{\text{om}}(\boldsymbol{x})}{\tau}\right)\right] \tag{16}$$

式中：$\sigma(\cdot)$为Sigmoid函数；$E_{(\boldsymbol{x},\boldsymbol{y}^+,\boldsymbol{y}^-)}$表示对偏好训练样本$(\boldsymbol{x},\boldsymbol{y}^+,\boldsymbol{y}^-)$求期望；$\tau$为偏好温度系数，取值为0.1，用于调节优劣回答评分差对偏好损失的影响。

当优质回答与劣质回答之间的策略评分差尚未达到动态间隔要求时，该损失推动模型提高优质回答的相对概率；当策略评分差已经超过动态间隔时，该损失逐渐减小。

为避免偏好强化学习导致模型过度偏离初始策略原有的多模态理解能力，本文以初始策略$\pi_0$作为锚点，引入KL策略约束$\mathcal{L}_{\text{anchor}}(\pi)$，表达式为

$$\mathcal{L}_{\text{anchor}}(\pi)=E_x\left[D_{\text{KL}}\left(\pi(\cdot|\boldsymbol{x})\|\pi_0(\cdot|\boldsymbol{x})\right)\right] \tag{17}$$

式中：$D_{\text{KL}}$为KL散度；$E_x$为对训练集中的结构化运维输入$\boldsymbol{x}$求期望。式(17)用于限制策略更新幅度，避免模型为满足局部偏好样本而牺牲原有语言能力和多模态理解能力。最终训练目标表达式为

$$\mathcal{L}_{\text{total}}(\pi)=\mathcal{L}_{\text{OM-RDMPO}}(\pi)+\lambda_{\text{KL}}\mathcal{L}_{\text{anchor}}(\pi) \tag{18}$$

式中：$\lambda_{\text{KL}}$为策略锚定权重，取值为0.05。具体训练流程如图4所示。

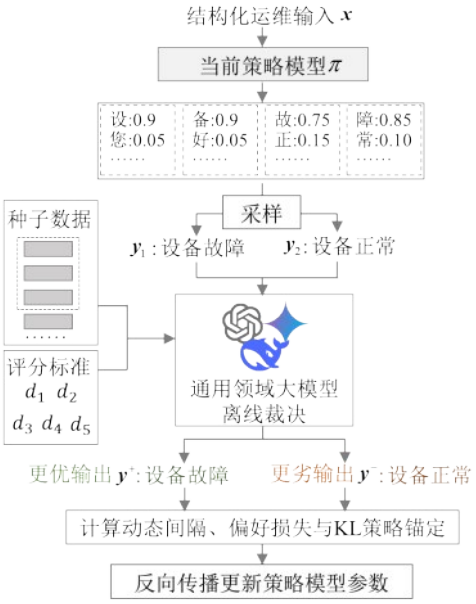



图4 OM-RDMPO训练流程

Fig. 4 Training workflow of OM-RDMPO

## 3 模型验证

### 3.1 实验设置

本文模型在8张NVIDIA A800显卡上进行偏好强化学习训练，训练精度为16位浮点数。模型的Transformer主干包含28层Transformer层，总参数数量为130亿。偏好优化训练学习率为$5\times10^{-7}$，训练轮次数为3，最大序列长度为4 096。所有对比模型在测试时采用统一提示词、统一输出格式和尽可能一致的解码参数。

OM-RDMPO中动态间隔和证据权重相关参数主要依据归一化偏好评分范围、候选回答长度分布和训练稳定性确定；参数设置原则是在普通样本上保持较小更新幅度，在安全边界差异明显或证据

不足样本上增强偏好约束，并通过KL锚定限制策略模型相对初始模型的偏离。

### 3.2　评价指标

本文采用BLEU-4（本位定义为$B_4$）[27]和选择题准确率评价模型。其中，$B_4$是用于衡量生成回答与标准答案的字词匹配度；为减小等义表达造成的偏差，每条数据设置3个不同表述的标准答案，并取最高分，计算公式为

$$B_4 = B_{\mathrm{P}} \exp\left( \sum_{n=1}^{4} \omega_n \log p_n \right) \tag{19}$$

式中：$B_{\mathrm{P}}$为惩罚因子，用来降低过于简洁回答的分数；$\omega_n$为权重，总和为1，在本文实验中设置$\omega_n$为1/4；$p_n$为$n$－gram精度，它衡量模型生成的答案与标准答案之间基于长度为$n$的字词序列的精确度。设$c$为模型生成答案的长度，$r$为标准答案的长度，$B_{\mathrm{P}}$的计算公式为

$$B_{\mathrm{P}} = \begin{cases} 1, & c > r \\ \mathrm{e}^{(1-r/c)}, & c < r \end{cases} \tag{20}$$

由于$B_4$难以充分反映语义等价，本文进一步为验证集中，每个问题构建3个同类别强干扰错误答案，将每一条验证数据编写成选择题形式，并在测试时采用统一提示词和输出格式。最终，将模型完成选择题的准确率作为第二个主要评价指标。

需要指出的是，该指标仅反映离线场景适配能力，不等同于工程安全认证。此外，测试采用选项随机化、同类强干扰和输出依据约束。

### 3.3　主要结果

本文评估了GPT-4-turbo、通义千问VL-Max、Gemini-2.5-Pro、DeepSeek-V3.1、GLM-4-9B-Chat等主流高性能大模型在电力设备运维验证集上的表现。各模型采用统一提示词和输出格式；提示词要求依据给定文本、图像或视频客观判断，信息不足时说明不确定性，不得编造设备状态、测量数据或规程条款。对于支持参数设置的模型，设置温度参数为0，核采样参数为1.0，最大输出长度为512，其他闭源模型采用平台允许的最低随机性。对于仅支持文本输入的模型，仅在文本子集上进行评估，并在结果表中以*标注。为降低数据泄漏和题面线索带来的偏差，训练集与验证集按样本语义和场景来源进行去重，选项随机打乱，模型需输出选项及简要依据。

表1　电力设备运维验证集上各模型性能对比

Table 1　Performance comparison of different models on the power equipment operation and maintenance validation set

| 模型 | 选择题准确率/% | | | | | | | $B_4$/% |
|---|---|---|---|---|---|---|---|---|
| | 输电线路 | 隔离开关 | 施工现场 | 变电站 | 电力器件 | 发电厂 | 平均分 | |
| GPT－4－turbo | 69.87 | 70.15 | 69.33 | 71.90 | 71.23 | 72.14 | 71.03 | 41.50 |
| 通义千问VL－Max | 72.92 | 73.50 | 70.47 | 73.55 | 71.67 | 72.68 | 72.88 | 42.89 |
| Gemini－2.5－Pro | 72.29 | 71.81 | 70.59 | 69.72 | 71.78 | 72.10 | 71.21 | 42.48 |
| DeepSeek－V3.1* | 74.06 | 73.33 | 74.25 | 72.05 | 72.96 | 73.42 | 73.08 | 43.27 |
| GLM－4－9B－Chat* | 65.84 | 64.61 | 63.99 | 64.18 | 66.04 | 65.22 | 64.92 | 39.41 |
| 本文完整模型 | 89.51 | 90.34 | 87.28 | 85.57 | 89.16 | 87.78 | 87.89 | 51.75 |

注：*表示仅支持文本输入的模型，本文仅在文本子集上进行评估。

各模型的具体性能对比如表1所示，本文完整模型在选择题准确率总分上达到87.89%，高出完成完整验证集的通义千问VL-Max模型15.01个百分点；在仅文本任务上，本文完整模型也高出DeepSeek-V3.1模型14.81个百分点。从不同模型的回答可以看出，主流大模型虽具有较好的语言流畅性，但在设备细节、规程边界和风险提示方面易出现事实缺失或泛化回答；本文完整模型在事实一致性、规程边界意识和风险保守性方面表现更好。

### 3.4　消融实验

#### 3.4.1　偏好优化训练

为了验证规程约束风险感知偏好强化学习的重要性，本文将完整模型与初始策略模型以及近端策略优化（PPO-RLHF）模型进行对比，结果如表2所示。其中，初始策略模型表示未进行本文偏好强化学习的模型，PPO-RLHF表示在相同初始策略上采用近端策略优化[20]训练后的模型。

对比性能可以得到，通过本文方法训练后，模型的选择题准确率提升了13.56个百分点，而且对比近端策略优化，本文方法训练的模型性能提升了5.54个百分点。由此可以看出，本文提出的规程约束风险感知偏好强化学习能够显著提升模型性能。

表2 基于偏好优化训练对模型性能的影响

Table 2 Effect of preference optimization training on model performance

| 模型 | 选择题准确率/% | 选择题准确率下降百分点 | $B_4$/% |
|---|---|---|---|
| 初始策略模型 | 74.33 | 13.56 | 45.36 |
| PPO－RLHF | 82.35 | 5.54 | 48.62 |
| 本文完整模型 | 87.89 | | 51.75 |

### 3.4.2 动态间隔机制的收敛分析

为进一步分析动态间隔机制对偏好优化训练的作用，本文比较了本文完整模型、去除任务复杂度约束和去除安全评分约束的变体模型在训练后的表现，结果如图5所示。图5给出了不同模型的验证集选择题准确率随训练步数变化的收敛曲线。

本文完整模型的选择题准确率和BLEU-4分别为87.89%和51.75%，均优于去除任务复杂度约束的变体模型(86.83%、51.04%)和去除安全评分约束的变体模型(86.32%、50.86%)。完整模型约在1 200步基本收敛，另外2种变体分别约在1 450步和1 550步趋稳，且本文完整模型收敛阶段的准确率波动控制在±0.24%以内，这说明任务复杂度与安全评分差距共同提升了训练性能和稳定性。

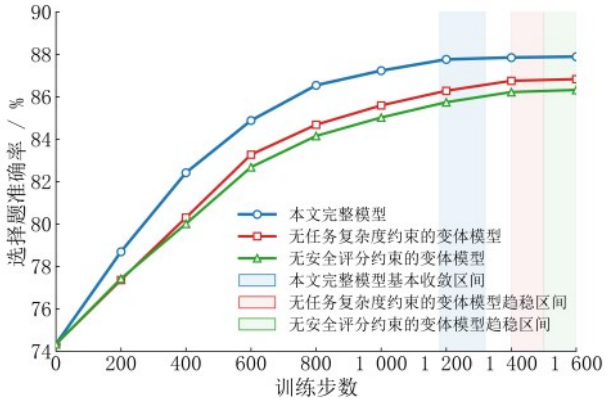



图5 动态间隔机制的收敛分析

Fig. 5 Convergence analysis of the dynamic margin mechanism

### 3.4.3 偏好裁决质量评估

由于本文偏好数据主要由人工种子样例引导的通用领域大模型离线裁决生成，为验证统一偏好评分标准的可靠性，本文从偏好数据集中随机抽取500条样本进行人工复核，并开展维度评分一致性分析。人工复核由2名种子数据标注人员完成，当2名复核者意见不一致时，由第三名复核者进行仲裁。评价结果如表3所示。

表3 偏好裁决质量评估结果

Table 3 Preference adjudication quality evaluation results

| 指标 | 评价方式 | 结果 |
|---|---|---|
| 偏好一致率 | 通用领域大模型裁决结果与人工复核偏好一致的比例 | 94.4% |
| 统一五维评分相关性 | 通用领域大模型五维评分与人工评分的平均Spearman相关系数 | 0.86 |
| 低置信样本比例 | 因得分差过小或理由不一致而进入复核或剔除的比例 | 7.2% |
| 复核后有效样本比例 | 经质量筛选后可用于偏好训练的样本比例 | 96.8% |

由表3可见，通用领域大模型离线裁决结果与人工复核偏好的一致率达到94.4%，统一五维评分与人工评分的平均Spearman相关系数为0.86，这表明外部裁决模型能够较好执行本文定义的电力设备运维偏好评分标准。同时，7.2%的低置信样本被送入人工复核或从训练集中剔除，降低了错误偏好标签对后续策略更新的影响。该实验说明，本文偏好数据并非完全依赖未校验的外部模型判断，而是通过人工种子校准、统一评分标准和低置信样本过滤共同保证偏好裁决质量。

### 3.4.4 电力设备运维数据

图6展示了移除不同类别训练数据对各类别验证性能的影响。图中，矩阵元素表示完整模型与相应数据移除变体模型在对应类别验证集上的选择题准确率下降百分点；数值越大、颜色越深，说明移除该类训练数据造成的性能下降越明显。以输电线路类数据为例，移除该类训练数据后，模型在输电线路验证集上的选择题准确率相较本文完整模型下降72.12个百分点，而在其他类别验证集上准确率的下降均不超过5个百分点。类似地，移除其他类别训练数据后，模型在相应类别验证集上的准

确率均出现显著下降，而对非对应类别的影响相对较小。该结果表明：不同运维类别包含具有差异性的设备对象、现场证据和风险模式，各类训练数据均对模型形成相应类别的证据使用、规程边界和风险保守偏好具有重要作用。

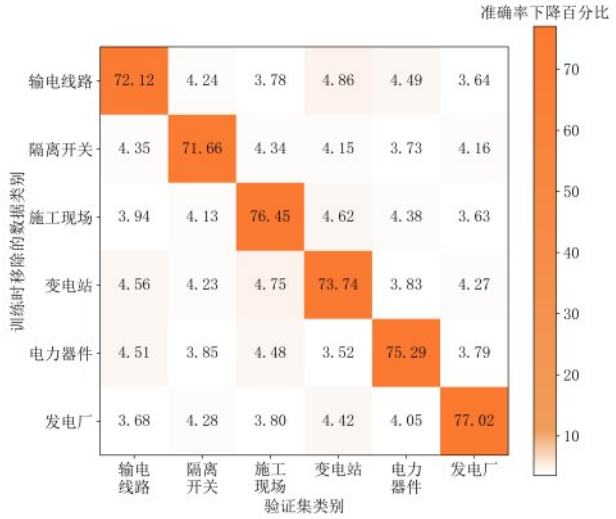



图 6　跨类别训练数据移除对选择题准确率的影响

Fig. 6 Cross-category effect of training-data removal on multiple-choice accuracy

## 4　结　论

（1）本文提出了一种面向电力设备运维的规程约束风险感知偏好强化学习方法，研究高风险运维任务中的安全偏好对齐，构建了由任务问题、现场多模态证据、运维规程和设备场景元信息组成的结构化运维输入表示。

（2）本文提出 OM-RDMPO 方法，将广义人类反馈强化学习中的策略采样、偏好裁决和策略更新闭环与电力设备运维专属约束相结合。该方法通过统一的电力设备运维偏好评分标准对候选回答进行离线裁决，并将综合偏好得分、安全评分差距、证据充分性和 KL 策略锚定纳入可微训练目标。实验结果表明，本文模型相较初始策略模型准确率提升 13.56 个百分点，相较 PPO-RLHF 模型提升 5.54 个百分点。

（3）经过偏好强化学习训练后，本文模型在电力设备运维验证集上的选择题准确率达到 87.89%，BLEU-4 达到 51.75%，高出完成通义千问 VL-Max 模型 15.01 个百分点，表明所提方法在离线电力设备运维场景验证中具有较好的有效性。与此同时，本文结果仍主要来自离线验证集，尚不能等同于真实工程现场闭环安全认证；后续将进一步扩展现场视频数据，补充开放式专家盲评、风险漏报率和工程试运行日志，以形成更完整的高安全场景大模型评价体系。

**参考文献：**

[1] 唐文虎，牛哲文，赵柏宁，等. 数据驱动的人工智能技术在电力设备状态分析中的研究与应用[J]. 高电压技术，2020，46(9)：2985-2999.
TANG W H，NIU Z W，ZHAO B N，et al. Research and application of data-driven artificial intelligence technology in condition analysis of power equipment [J]. High Voltage Engineering，2020，46(9)：2985-2999.

[2] 周俊煌，黄廷城，谢小瑜，等. 视频图像智能识别技术在输变电系统中的应用研究综述[J]. 中国电力，2021，54(01)：124-134＋166.
ZHOU J H，HUANG T C，XIE X Y，et al. Review on the application of video image intelligent recognition technology in power transmission and transformation systems [J]. Electric Power，2021，54 (1)：124-134，166.

[3] 段俊峰，李晨坤，姚文轩，等. 基于人工智能赋能的新型电力系统关键技术综述[J]. 湖南电力，2024，44(01)：1-10.
DUAN J F，LI C K，YAO W X，et al. Review of key technologies for new power systems empowered by artificial intelligence [J]. Hunan Electric Power，2024，44(1)：1-10.

[4] 盛戈皞，钱　勇，罗林根，等. 面向新型电力系统的电力设备运行维护关键技术及其应用展望[J]. 高电压技术，2021，47(09)：3072-3084.
SHENG G H，QIAN Y，LUO L G，et al. Key technologies and application prospects of operation and maintenance of power equipment for new power systems [J]. High Voltage Engineering，2021，47(9)：3072-3084.

[5] 刘传洋，吴一全. 基于红外图像的电力设备识别及发热故障诊断方法研究进展[J]. 中国电机工程学报，2025，45(06)：2171-2196.
LIU C Y，WU Y Q. Research progress of power equipment identification and overheating fault diagnosis based on infrared images [J]. Proceedings of the Chinese Society for Electrical Engineering，2025，45(6)：2171-2196.

[6] 贺明强，靳　君，关新宇，等. 基于计算机视觉的电力设备状态监测与故障诊断策略分析[J]. 集成电路应用，2024，41(02)：224-225.

HE M Q, JIN J, GUAN X Y, et al. Analysis of computer vision-based strategies for condition monitoring and fault diagnosis of power equipment [J]. Integrated Circuit Applications, 2024, 41 (2): 224-225.

[7] 谢 庆, 张煊宇, 王春鑫, 等. 新一代人工智能技术在输变电设备状态评估中的应用现状及展望[J]. 高压电器, 2022, 58(11): 1-16.
XIE Q, ZHANG X Y, WANG C X, et al. Application status and prospects of new-generation artificial intelligence technology in condition assessment of power transmission and transformation equipment [J]. High Voltage Apparatus, 2022, 58(11): 1-16.

[8] 蒲天骄, 乔 骥, 韩 笑, 等. 人工智能技术在电力设备运维检修中的研究及应用[J]. 高电压技术, 2020, 46(02): 369-383.
PU T J, QIAO J, HAN X, et al. Research and application of artificial intelligence technology in operation, maintenance and overhaul of power equipment [J]. High Voltage Engineering, 2020, 46 (2): 369-383.

[9] 陈浈斐, 章黄勇, 马宏忠, 等. 基于深度学习的电力设备故障诊断方法研究综述[J]. 电气自动化, 2022, 44(01): 1-2+6.
CHEN Z F, ZHANG H Y, MA H Z, et al. Review of deep learning-based fault diagnosis methods for power equipment [J]. Electrical Automation, 2022, 44(1): 1-2, 6.

[10] 谢庆, 王春鑫, 李帆, 等. 知识及数据驱动的电力一次设备健康管理方法综述[J]. 高电压技术, 2024, 50(02): 605-620.
XIE Q, WANG C X, LI F, et al. Review of knowledge- and data-driven health management methods for primary power equipment [J]. High Voltage Engineering, 2024, 50(2): 605-620.

[11] 刘开培, 李博强, 秦 亮, 等. 深度学习目标检测算法在架空输电线路绝缘子缺陷检测中的应用研究综述[J/OL]. 高电压技术: 1-12[2023-03-09]. DOI: 10.13336/j.1003-6520.hve.20220273.
LIU K P, LI B Q, QIN L, et al. Review on the application of deep learning-based object detection algorithms in defect detection of insulators on overhead transmission lines [J/OL]. High Voltage Engineering: 1-12 [2023-03-09]. DOI: 10.13336/j.1003-6520.hve.20220273.

[12] 陈富国, 杨爱军, 马慧珍, 等. 高压隔离开关状态智能感知系统设计与实现[J]. 自动化技术与应用, 2021, 40(08): 131-135+152.
CHEN F G, YANG A J, MA H Z, et al. Design and implementation of an intelligent perception system for the state of high-voltage disconnectors [J]. Techniques of Automation and Applications, 2021, 40(8): 131-135, 152.

[13] Lu J, Clark C, Zellers R, et al. Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks [J]. arXiv preprint arXiv:2206.08916, 2022.

[14] Zhu X, Zhu J, Li H, et al. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16804-16815.

[15] Alayrac J B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning[J]. arXiv preprint arXiv:2204.14198, 2022.

[16] Wang P, Yang A, Men R, et al. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework [J]. arXiv preprint arXiv:2202.03052, 2022.

[17] Liu H, Li C, Wu Q, et al. Visual instruction tuning [C]//Advances in Neural Information Processing Systems. 2023.

[18] Bai J, Bai S, Yang S, et al. Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond [J]. arXiv preprint arXiv: 2308.12966, 2023.

[19] Lin B, Ye Y, Zhu B, et al. Video-LLaVA: learning united visual representation by alignment before projection[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

[20] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv: 1707.06347, 2017.

[21] Rafailov R, Sharma A, Mitchell E, et al. Direct preference optimization: your language model is secretly a reward model [C]//Advances in Neural Information Processing Systems. 2023.

[22] Wang J, Li M, Luo H, et al. Power-LLaVA: Large Language and Vision Assistant for Power Transmission Line Inspection [C]//2024 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2024: 963-969. DOI: 10.1109/ICIP51287.2024.10648271.

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[24] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[J]. arXiv preprint arXiv:1508.07909, 2015.

[25] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv: 2010.11929, 2020.

[26] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding?[C]// ICML. 2021, 2(3): 4.

[27] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.

（编辑　刘懿莹）