ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Evaluation, Selection, and Deep Adaptation of General-Purpose Large Models for Power Industry Applications

Ke Shi[1,*], *Jing Niu*[1], *Weixiang Qiao*[1] *and Xing Zhang*[2]

[1] Power Dispatching Control Center of Guizhou Power Grid Co., Ltd., GuiZhou, China
[2] Power Dispatching Control Center of Zunyi Power Supply Bureau of Guizhou Power Grid Co., Ltd., GuiZhou, China

SUMMARY: *Building a practical research model for evaluating, selecting, and adapting general-purpose large models in the two energy application scenarios of substations' intelligent inspection and transmission corridors visualisation. The benchmark includes 18,642 retained multimodal evidence records that include: visible images; thermal frames; OCR string; equipment metadata; corridor attribute; rule clause; and historical ticket text. Anonymised six models were evaluated at set data divisions, prompts templates, inference upper limit and scoring script. Targeted power service judgment: Object localisation, risk inference, rule-based evidence, unsupervised alarm control, and robustness to field perturbations.Based on a weighted-score screen of the candidates, an adaptive selection result of the selected model included retrieval evidence, LoRA tuning, visual-grounding calibration, and safety verifier. Adapted Power-GM obtained the best comprehensive scores of 89.0%, 86.0%, 87.0%, 88.0% and 84.0%, respectively, for visual anchoring, risk judgement, rule obedience, hallucination suppression, and robustness. Eight of the selected tasks surpassed the most powerful open multimodal baseline by 9.9%.-20.3 percentage points and the closed multimodal baseline by 3.3-8.5 percentage points. The best response-surface area is LoRA ranking 48 and retrieval top-k 6, which still has a power-biz score of around 89.0% inside the latency bound. Ablation demonstrated that retrieval enhanced rule adherence, LoRA strengthened task reasoning, grounding calibration reduced object-Region misalignment, and the safety verifier decreased hallucinated risk assertions. This study is confined to the two tested scenes, with fixed model labels, task definitions, scoring scripts, test record retention policies, and only included the evidence types from the benchmark collection.8.5 percentage points. The best response-surface region was LoRA rank 48 with retrieval top-k 6, where the Power-Biz score remained near 89.0% within the latency target. Ablation showed that retrieval improved rule compliance, LoRA strengthened task reasoning, grounding calibration reduced object-region mismatch, and the safety verifier reduced hallucinated risk statements. The conclusion is limited to the two tested scenes, fixed model labels, fixed task definitions, fixed scoring scripts, retained test records, and the evidence types in the benchmark corpus.*

KEYWORDS: *Large power grid model; Empirical verification; Model choice; Deep adaptation; Substation intelligent inspection; Transmission corridor visualisation.*

1 Introduction

Representative entries for the intelligent inspection of substations and transmission corridors

*18786755904@163.com
https://doi.org/10.65102/is20261026
Published: 30 April, 2026

visualisation in power operation and maintenance applications. Both transform heterogeneous field evidence into operator-revocable inspection judgments. Substation images are enclosed but have many details; cabinets, reflecting planes, nameplates, oil traces, pointer meters, infrared pseudo-colours, as well as small faults. The corridor image covers longer corridors and includes vegetation, conductor, crane, vehicle, bird, foreign object, insulator, tower component, shadow, hazy conditions, etc. Error has an immediate operational cost: if the wrong localization leads a robot to the wrong cabinet; If the false vegetation risk causes unnecessary clearing; Or If fabricated rule evidence erodes trust in automatic tickets.

The primary Inspection methods at present are detection-based Segmentation-Based Classification-Defects Recognition and so forth. Task-specific-defective-object recognition. UAV imaging, visible detection, thermal sensing, edge processing, light-weight substation detector, visual-thermal-line recognition and YOLO-based vegetation evaluation have increased the accuracy of perception [1-5]. Mostly short-term prediction results that lack the combination of visual materials, rules, and field action meanings.

Multimodal generalised models can handle both image and textual data, be prompted with instructions or recognize OMRs/OCR; Generate structured summaries, connect perception results to inspection records. Unified image-Text interface, little-shot Vision-Language reasoning, unfrozen visual encoder-LLM connection, and visual instruction tuning all show this capability [6-9]. These abilities match the power-inspection function well; field judgement usually involves judging images, thermal information, equipment names, rule contents, historical ticket data, recommended measures, etc.

The direct transfer remains untrustworthy. Models may name the correct asset while grounding evidence on nearby breakers, terminal blocks, nameplates, conductors, cross-arms, tree crowns, or background objects. Assigning risk levels without considering the relevant local regulations, particularly for vegetation clearance, thermal anomaly, and defect severity operations. Generic benchmarks also miss OCR-heavy tickets, small-object localization, route-context reasoning, latency, memory, cost, auditability, and network-isolation constraints. Existing multimodal benchmarks and hallucination studies provide useful dimensions [10-12], but power inspection requires evidence-level judgement: asset or corridor-object identification, abnormal-region localization, task-specific risk grading, rule-consistent evidence, and structured platform output.

The prompt-engineered function is not stable. Retrieval-augmented generation provides rule and ticket evidence (labeling), relevance-aware retrieval removes irrelevant context (filtering) [13, 14], and LoRA reduces rank through freeze base weight [15]. Retrieval mainly addresses rule availability; LoRA improves field-expression mapping. visual grounding and hallucination control need to be calibrated separately, etc. Therefore, this study has integrated the entire control system for a model screening-retrieval-LoRA tuning-grounding calibration-safety verification-deployment feasibility process into one set of tests.

Two of the two-target scenes need separate empirical verification. Substation intelligent inspection models need to identify densely packed equipment objects, recognise sparse text information, determine infrared abnormalities, and produce safe narrative results without falsely reporting faults. Transmission corridor visualisation requires connecting a large scene of hazardous factors at the route level, such as vegetation obstruction, foreign objects, tower-component misalignment, and weather-caused impaired visibility. A single overall score does not reflect whether the chosen model is appropriate in both scenarios. Therefore, this experiment presents the overall selection results, scene-task results, ablation experiments, deployment efficiency, perturbation-robustness indicators, and residual-error-related explanations. The data splitting controls at the equipment-bay or route-span levels and reduces scene exposure. Anonymised candidate models are of the same type according to function, and

this work selects them empirically without branding any particular commercial prototype.

How to evaluate, select and adapt general-purpose large models to produce traceable results on the topic of substation inspections and transmission corridors through visualisation. Therefore, based on this, we build a new Power-Inspection benchmark in the form of multi-modal data from five aspects: visualisation; textual description; metadata tagging; rule restrictions; risk assessment and scoring criteria for each item are set accordingly. In addition to a traditional adaptive Design combining Retrieval, LoRA, grounding calibration and Safety Verification. Based on the retention of data, controlled comparison and assessed returns.

Power-inspection output has operational risk. A model answer can prefill a ticket, prioritize the defect queue, and initiate review, etc. Match it with the asset list, ground the visible area, assign a scene-conforming risk level, refer to relevant regulations, and not perform unauthorised shutdowns, clears or emergency operations. Therefore, the test will evaluate structured evidence fields instead of paragraph fluency.

Model selection should meet deployment prerequisites. Basic models are still based on image-OCR-metadataramourous-text combinations, but these may misinterpret risk expressions, fail to notice background distractions, or refer to incorrect rules. They cannot be reflected in the public benchmark scores. Test with the same objects, Images, Templates and Limits as in the Power Service workflow.

These two situations constitute the empirical Domain. Substation inspection tests for densely packed equipment ground connections, dense-heat/fused tickets, etc. Corridor visualization tests large-scene localization, vegetation and foreign-object risk, and route summaries. To determine the optimal model parameters that yield valid inspection results based on given Power-Service data.

This positioning falls somewhere between perception research and automatic operation Control. Evaluate a large model that integrates visual evidence, rule evidence and inspection actions. At this time, there are many error-prone deployments, as well as opportunities for control and selection to bring practical significance.

2 Methods

2.1 Empirical Scenario Corpus and Evidence Object Construction

Empirical samples were collected in the substations' intelligent inspections and transmission corridors' visualisations respectively. Deduplicate, blur filter, balance tasks, and check labels to retain 18,642 pieces of evidence information. Each record was organised as an evidence object that contained both images and thermal information; it also had OCR strings in addition to metadata, rule-based evidence, annotation details, tickets, and risk identification results.

$$\mathcal{D} = \{(x_i^v, x_i^t, g_i, y_i, s_i, c_i)\}_{i=1}^{N}, \quad N = 18{,}642. \tag{1}$$

In the expression, x_i^v is a visible-light image or its crop, while x_i^t is the corresponding text evidence; thermal cue, OCR string, g_i or any other metadata field are considered as y_i; s_i is a visual-grounding label, c_i is an object/defect label, is a risk-level label, and is a rule/ticket evidence connected with this instance. Notations are adopted here to make the evaluated objects under these two conditions identical. A substation image; may be a breaker compartment, indication pointer meters, heat spots, insulators, oil leakage areas. In general, for corridor images: may stand for conductor segments, towers, foreign objects, construction machinery, and plants close to the warning line. Field was still used even if there was no abnormality; in order to verify the accuracy of a false alarm.

Table 1 gives the retained corpus composition. A split of 70:15:15 was employed, but it was not randomly selected images but rather based on the equipment bay, inspection route and tower span. Images and their associated metadata for the same Bay or Route did not belong to either the training or test set simultaneously. This setting reduced leakage from repeated camera angles, repeated tower silhouettes, and identical background vegetation. The substation part had 10,215 records. The corridor part contained 8,427 records. The collection contained 36 types of substations and 11 objects in the corridors. Retained labels included normal, attention, warning, and critical risk categories; The distribution of these labels was even across all tasks rather than uniformly distributed across raw-image data.

Table 1: Corpus composition and split used in the empirical test.

Scenario source	Evidence type	Total records	Train	Validation	Test	Main label fields
Substation inspection	Visible equipment frames	6,820	4,774	1,023	1,023	object, defect, grounding
Substation inspection	Thermal/OCR composite records	2,764	1,935	414	415	hotspot, reading, component
Substation inspection	Ticket-rule pairs	1,631	1,141	244	246	risk phrase, rule clause
Transmission corridor	UAV and tower-camera frames	6,124	4,287	918	919	line, tower, vegetation, foreign object
Transmission corridor	Route metadata and weather records	1,303	912	195	196	span, buffer, weather, visibility
Total	Multimodal evidence records	18,642	13,049	2,794	2,799	scenario, task, risk, rule

Retained negative and boundary samples to match the field inspection environment. Records of substations include normal equipment, reflections, dust, harmless namespce artefacts, and no-defect thermal deviations. Corridor records include vehicles that have exceeded the buffer zone or warning distance range, etc.; tower-shaped backgroud structures. These sample data did not improve the model's performance by penalising excessive warning of uncertain samples.

Every record is associated with tasks of the same type in different datasets. Substation Tasks include identifying equipment; locating defects; describing infrared anomalies; reading meters and nameplates; generating safety tickets, etc. Tasks involve: line and tower locationing; foreign object removal (FUR); vegetation-risk evaluation (VE-RISK), route-context summery, multi-modality-warning-analysis. To allow for multiple tasks using a single instance without experiencing split-leakage.

Rules are transformed into retrievable entities that contain the following information: rule id; Applicable objects; Voltage/Scene conditions; Risk phrases; Action descriptions. Anonymise and clean historical tickets, etc. OCR strings were retained only when readable or metadata-aligned. Check the image-grounding labels again by two annotators, and remove samples where the Iou difference is greater than 0.25 after resolving.

Annotation performed two passes. Mark the first-pass detected scene objects, abnormalities, abnormal areas, risk phrases and visibility status. The second pass matched the rule evidence and historical ticket phrase annotations to obtain objects [16-18]. Only a record could be accepted if the objects' labels and risk levels were provided with rule-based evidence or

documented under normal circumstances.

Due to the difference in Structure and complexity of the two situations, Scores were calculated individually for each task before being averaged according to this circumstance. This prevented equipment-identification performance from masking weak vegetation-risk reasoning and kept corridor localization from dominating the final comparison.

Data Cleaning Followed the Operation Standard. Severely blurred images were excluded from selection only if an abnormal item was not clearly identified by a human reviewer. A mild blur, an exposure variance, a lack of contrast, partial obscuration and temperature colour difference are all present during regular field observation. Before evaluating, OCR strings were not manually corrected. The original unprocessed OCR string and the aligned equipment metadata were both kept; thus, when noisy in ORECOO processing, they could be used for reference. Thus, making this a more suitable inspection-type system, where the perception modules provide incomplete data for subsequent processing by the reasoning part.

Finally, the entire audited Corpus to ensure that there were sufficient positive examples (POS), negative Examples(NEG)and Boundary Cases(BOUR). Substation defect location; Retained local stains, leakage-like shadows, normal screw holes, and true defects in the test set to prevent over-reliance on an obvious anomaly texture indication. Infrared anomaly interpretation: keep the normal warm parts and real hot spots together. Corridor visualisation tests for test subsets; it retained foreground vegetation, distant plantations not within line-corridors protection ranges, constructions without buffers inside corridors, and additional debris or close-encounter conditions related to transmission towers. These boundary samples are chosen to show overwarning performance more evidently compared with clean abnormal samples.

In addition, the annotation scheme separated object presence from actionable risks. Truck appeared in the corridor of an image but was not labelled as a risk. It turned into a risk sample after determining its Location, Route Context or relationship with the corridor had triggered an alert. In the same way, for sample data of substations with high temperatures; a clear hot area alone could not be determined as critical until the equipment category and rule evidence confirmed this. Therefore, this separation enabled verification of whether the model was able to eliminate transforming all salient objects into operational events. Figure 1 presents the data organisation mechanism. It is described in detail by merging the forms of explicit Frames, thermal images, OCR strings, ground-based data such as labels, risks and rules. The other objects after the above-mentioned ones include visual anchoring, rule adherence, robustness testing, and hallucination suppression performance.

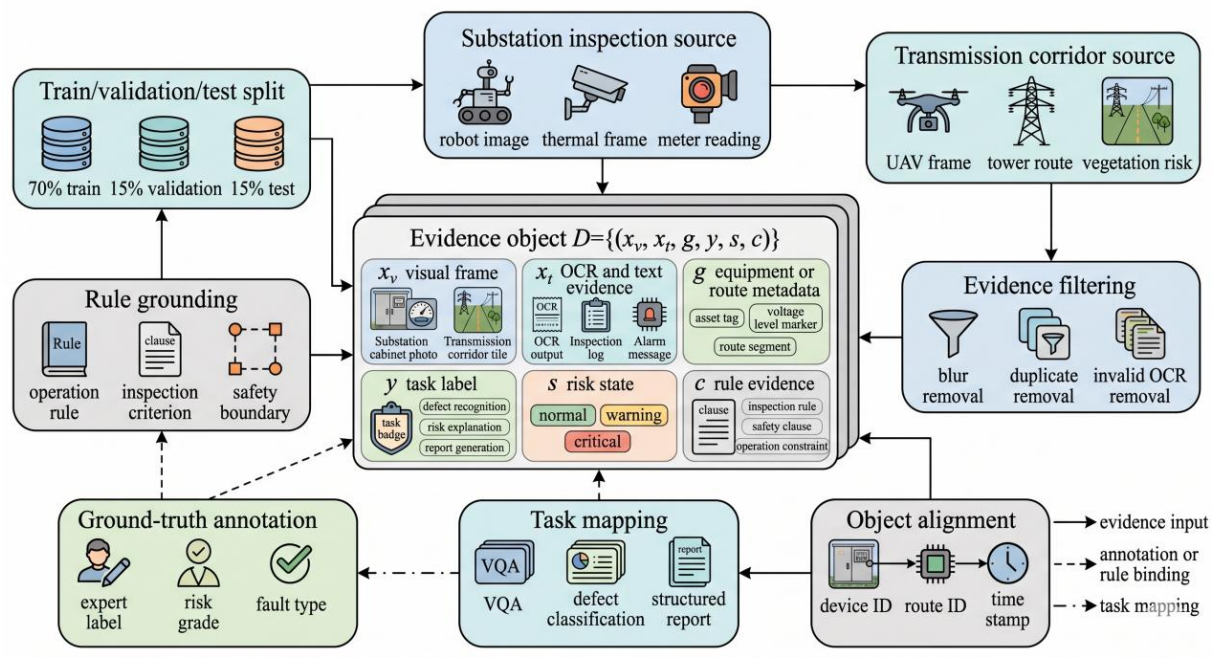

Figure 1: Mechanism of multi-source scenario data organization for power inspection large-model evaluation.

2.2 Model Evaluation, Selection Score, and Deep Adaptation Strategy

Six anonymised candidates were tested under the same division, output scheme, token restrictions and re-attempt policies: Text-GM, three open multimodal systems with scales of 7B/13B/34B; closed-MM and power-GM. Anonymisation of the labels was used to assess a repeatable selection procedure, not vendor ranking. Selection adopted the workflow-based indicator: Visual grounding, Risk Reasoning, Rule Compliance, Hallucination Control, Robustness, Latency Efficiency and Cost-Efficiency. The final screening score is obtained by weight-aggregation of penalties for latency, costs and hallucination exposure.

$$S_m = \sum_{j=1}^{J} w_j \, \phi_j(q_{m,j}) - \lambda_L \psi(L_m) - \lambda_C \psi(C_m) - \lambda_H H_m. \tag{2}$$

In this formula, S_m is the selection score for model m, $q_{m,j}$ is the measured value of indicator j, w_j is its weight, $\phi_j(\cdot)$ normalizes the indicator into a 0 to 100 scale, L_m is mean inference latency, C_m is relative inference cost, H_m is the hallucination exposure rate, and λ_L, λ_C, and λ_H are penalty coefficients. The weights were fixed before testing and were not learned from the test set. Table 2 lists the indicators' definitions and Weights. Since visual grounding and following the rules determine correct operation for users; therefore, these indicators were assigned higher scores.

Table 2: Evaluation Indicators, Scoring Definitions and Fixed Weights.

Indicator	Scoring definition	Weight	Business reason
Visual grounding	Correct object label and IoU ≥ 0.5 with expert region	0.19	prevents wrong equipment localization
Risk reasoning	macro-F1 over normal, attention, warning, critical	0.16	supports alarm prioritization
Rule compliance	applicable clause or ticket evidence used correctly	0.18	supports auditable operation records
Hallucination control	absence of unsupported object, defect, or rule claim	0.16	reduces false tickets and over-warning
Robustness	retention under perturbation pairs	0.12	tests field imaging stability
Latency efficiency	normalized inverse mean response time	0.10	supports online inspection calls
Cost efficiency	normalized inverse request cost and memory burden	0.09	supports repeated deployment

Based on filtering, the best-balanced model is composed of four parts: retrieval, LoRA, grounding calibration and safety verification. Supply filtered rule clauses, ticket phrases, object descriptions, voltage notes and maintenance actions. Tuned LORA attention and feed-forward projections, then froze the base weights. Grounding calibration forced regions to answer with detector or segmentation candidate. The safety verifier did not accept unsafe operations that were unsupervised or lacked visual verification.

$$\theta' = \theta + \Delta\theta, \quad \Delta\theta = BA, \quad A \in \mathbb{R}^{r\times d},\ B \in \mathbb{R}^{d\times r}. \tag{3}$$

According to this formula, θ where represents the frozen base weight, $\Delta\theta$ indicates a trainable low-rank update; A and B, d represent the hidden dimension. The ranked results are: 8; 16; 32; 48; and 64. Retrieval Depth Top-K is set between 1 and 10. Finally, the adapted model was trained only in the training set and validated with the hold-out data. Test set was not accessed again after that point. The low-rank Design followed the parameters-efficient principle of LoRA [16], and the Retrieval Design followed the evidence-access principle of retrieval-augmented generation [17, 18].

Table 3 shows the selected candidates. The chosen Power-GM employed a multi-modal back-bone, a domain adapter, a retrieval store, a grounding calibrator and a verifier. Only six items were allowed as the output: Object, Location; Abnormal Evidence; Risk Level; Rule Basis; Recommended Action. Organised output reduced the variety of generated prose and allowed for more accurate processing by the scoring script. Only free-form narration was retained at the end of the explanation field and did not appear among the scores.

Table 3: Candidate model and Adaptation Settings.

Model label	Input modality	Adaptation state	Retrieval	Grounding calibration	Main deployment role
Text-GM	text only	prompt-only	yes	no	ticket and rule baseline
Open-MM-7B	image + text	prompt-only	optional	detector-assisted	lightweight multimodal baseline
Open-MM-13B	image + text	prompt-only	optional	detector-assisted	mid-scale open baseline
Open-MM-34B	image + text	prompt-only	optional	detector-assisted	high-capacity open baseline
Closed-MM	image + text	prompt-only	limited	service-dependent	closed general reference
Power-GM	image + text + rule evidence	LoRA + RAG + verifier	yes	yes	adapted inspection model

Figure 2 is the deep-adaptation Design based on Table 3. It connects image evidence, OCR/metadata, retrieval evidence, LoRA, grounding calibration and safety verification of a general-purpose multimodal model. The retrieval provides rules of evidence, LoRA modifies domain representations; Grounding Calibration restricts objects-Regions References; The Verifier blocks unsupported or unsuitable Responses.

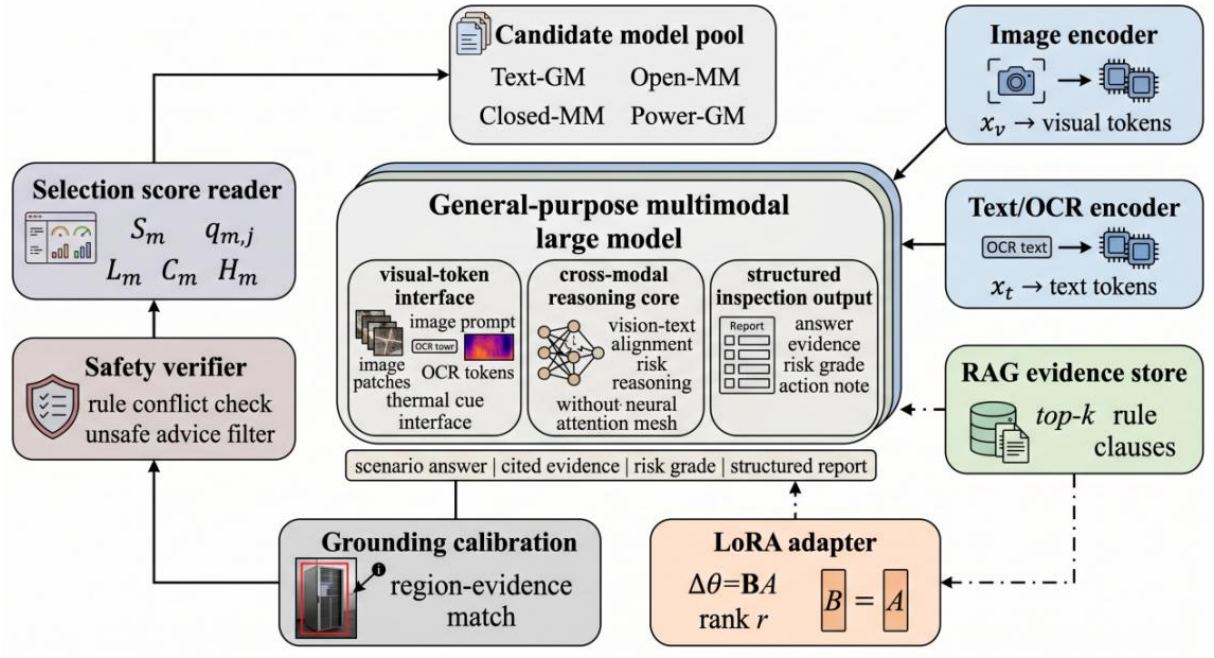

Figure 2: Based on the retrieval-based model-selection and deep-adaptation mechanisms of LoRA, ground-linguistic calibration and security verification.

The open candidates included common 7B-34B deployment ranges and instruction-following settings [19-21]. MMMU only showed in general terms of multimodal capabilities did not ensure the power-supply's reliability.

The retrieval store was index in the scenes, objects' types, risk phrases and rule IDs. A unit was accepted only when at least two keys matched the current evidence object, preventing irrelevant clauses such as vegetation rules for oil-leak samples or low-voltage rules for high-

voltage equipment. Ranked based on the degree of embedding similarity and rule-key agreement. Within 90 words for each unit to avoid the phenomenon of visual blockage.

The LORA was built with the help of a five-component-target structure containing an object, location, abnormal-evidence, risk-level, rule- foundation and recommended response. secondary free text explanation. early stop using validation power-biz score; Trained until two consecutive non-improved epochs occurred, then selected the best checkpoint.

Safety verifiers carried out deterministic tests. Verifying Risk Level, Retrieve Rule Existence; Verify Object-Field Consistency; Check Action-Risk Compatibility. Violations led to one of the following: constrained regeneration or downgrade certainty. Verifier does not change the original picture prediction, but blocks unsupported structured conclusions.

All candidates were set at a lower temperature, capped output length, and using the same-six-field structure. schema failure was resolved through deterministic parsing, and unresolved cases were treated as schema failures and added to the exposure list of hallucinated risks or rules that generated unsupported situations.

Adapter training used only the training subset. Selected validation ranking, retrieved depth, verifier threshold, and stop training epoch. Finally, the prompts and schemas were locked at this time to avoid reporting variations due to multiple rounds of manual fine-tuning by the participants.

2.3 Experimental Protocol and Deployment-Oriented Validation

There were three stages in this experiment: preliminary tests on all six individuals; Validation Selection Based Retrieval-Depth-and-LoRA-Ranking; Fixed Test Selected Against Unadorned Baselines, Partial Adaptations, Ablation Variants; The same scoring scripts were applied consistently.

It contained little information. Each record provided the image or crop, available OCR/metadata, task instruction, and, when enabled, retrieved rule units in a fixed schema. The model outputted six fields: Object, Location, Abnormal Evidence, Risk Level, Rule Basis and Recommended Action. Output an "uncertain" result under uncertain circumstances to minimize false critical alarms caused by vague samples.

Combining primary business evaluation metrics with grounding, risk, rules, hallucination and robustness. Robustness using clean--perturbed pairs. The substation disturbance includes blurriness, exposure displacement, occlusion, temperature colour changes, and OCR noise. Corridor Perturbation included haze, Scale Variation, Conductor-Background Confusion, Moving Object Overlap, Vegetation-Texture Interference. Stability of each pair formed the retention rate.

$$R_m = 1 - \frac{1}{|\Omega|}\sum_{i\in\Omega}\frac{|a_m(x_i)-a_m(\tilde{x}_i)|}{\max(|a_m(x_i)|,\epsilon)}. \quad (4)$$

According to Formula (4), R_m is the robustness score for model m, Ω includes all clean-perturbed pairs; and x_i are clean samples, while ; returns a scalar task score from an evaluation script; does not result in division by zero. Predication was accepted if the objects' category, evidence area, risk degree and the rule base did not exceed a reasonable deviation in terms of change. This definition is stricter than ordinary answer similarity because a small wording change is acceptable, while a region shift or unsupported risk escalation is not.

$$P = \alpha G + \beta F + \gamma U + \eta R - \delta H. \quad (5)$$

In this formula, P is the Power-Biz score used in the result figures, G is visual grounding, F is risk reasoning macro-F1, U is rule compliance, R is robustness, H is hallucination exposure, and β, γ, η, α and δ are fixed validation-stage coefficients.

Latency is recorded at a constant resolution, output length and query depth. Warm-up requests were excluded, and repeated-batch means were reported. Recorded peak memory at inference time. Relative Cost of Token Use, Retrieval Calls and Service-Request Costs. For closed Service, memory was treated as a burden equivalent to that of local Model for comparison. Figure 3 illustrates the evaluation scheme.

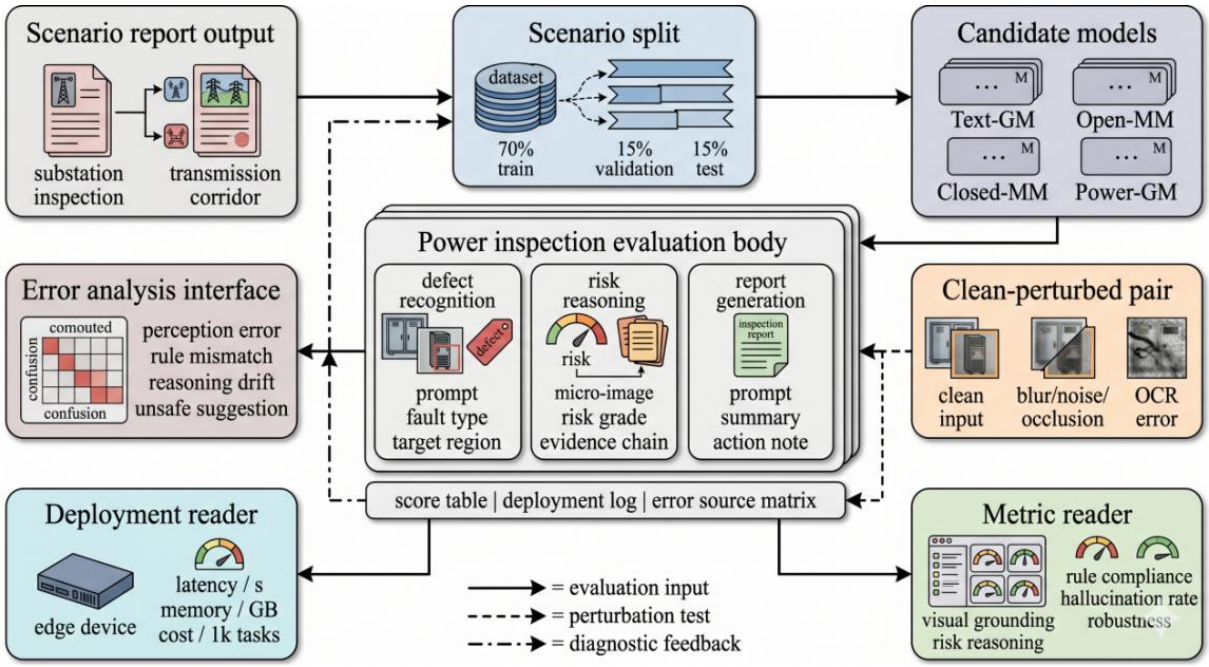

Figure 3: Evaluation Routine Connecting Scenario Split, Perturbation Tests, Deployment Indicators and Error Analysis.

Linking scenario Split, Candidate Models, Clean-Perturbed Pairs, Metric Readers, Deployment Readers, and Error Analysis. Figure out the path from evidence records to model output, and finally reach the score report; it also minimises fluctuations caused by prompts, retrievals or splits.

All comparisons used the fixed test split. The mean task score was reported, and any difference below 1% would not be distinguished. Ablation removes each of the following components in turn: retrieval, LoRA, grounding calibration or safety verifier. Using 400 stratified failures of the final-adapted model for error analysis. Each case was assigned a single error label at first; therefore, it chose the first cause of failure.

Visual grounding was correct only when the object class was correct and the referenced region reached at least 0.5 IoU with the expert region. Natural-language Locations were assigned to the closest detectors or segmentation candidates. Class error but correct instance was considered a false ground.

Based on the Rule-Basis filed in this rule. Only a Response could be generated if the cited clauses or ticket evidence referred to the same scene, Object Type and Risk Conditions. Correct risk without applicable rule evidence received risk credit but no rule-compliance credit. Hallucinations included unsupported objects, defects, serious-risk expressions and nonexistent regulations [23-25]; The larger penalty was for changes in the recommended action.

After the test-split, clean-perturbed pairs were produced and were not included in the training set. Perturbations of routine image variations. Stable judgment of objects, regions, risk levels and rule bases. Word change disregarded; risk escalation or region alteration without supplementary materials recorded as unstable factors.

Each model generated the following three files: structured-prediction result, evaluation metric, and error. Stored the raw responses along with their six-parsed attributes. Records of metrics such as grounding, risk, rules, hallucination, robustness, latency and cost. Error records store the first failure location.

3 Results and Discussion

3.1 Overall Model Selection and Adaptation Response

The first-result Section asks which candidate models to select before deep adaptation, as well as whether the chosen model still stays in an operational range after adjustment. Because the selection of models for power inspections is not equivalent to having a generalised multimodal capacity. The capacity is too large, there will be over-optimised response speed but at the expense of performance; The lighter version can meet deployment requirements and may not pass some tests for recognising specific targets. Starting with the weighted selection landscape empirically, it appears as shown in Figure 4.

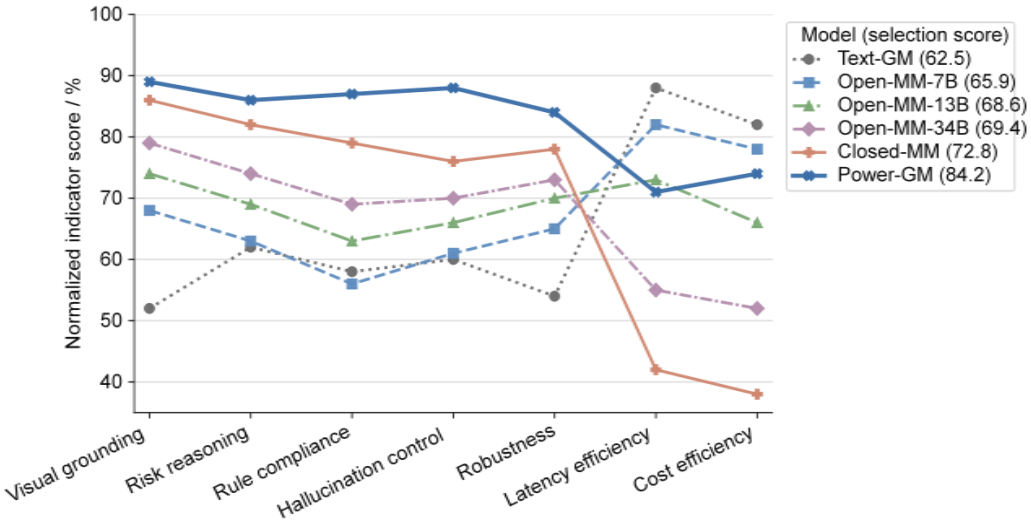

Figure 4: Selection Landscape of the Seven Deployment-Oriented Indicators.

Fig. 4 shows that adapted Power-GM achieved the highest overall selection score, 84.3%. The closed-MM has a high score in terms of visual recognition (grounding), risk identification, rules obeyed; But it is lacking in latency and cost-effectiveness by 42.0% and 38.0%. Open-MM-34B surpassed Open-MM-13B in terms of visual grounding, risk reasoning and rule adherence; its performance increased from 74.0%, 69.0% and 63.0% respectively, to 79.0%, 74.0% and 69.0%; Latency and cost-efficiency decreased accordingly: From 73.0% and 66.0% to 55.0% and 52.0%. Text-GM is effective, with a latency of 88.0 per cent and cost under US$0.8; it reached $0.52.

Selection results are more favourable to engineering balance over model size individually. Power-GM achieved 89.0% visual grounding, 86.0% risk reasoning, 87.0% rule compliance, 88.0% hallucination control and 84.0% robustness; Latency stayed under a controlled

deployment boundary of repetition inspection. There is such a balance that one part's intensity is too strong for practical significance, and there should be unevenness all over the place.

Select first and then validate through adaptation configuration searches based on the selected data. As shown in Figure 5, the response surface.

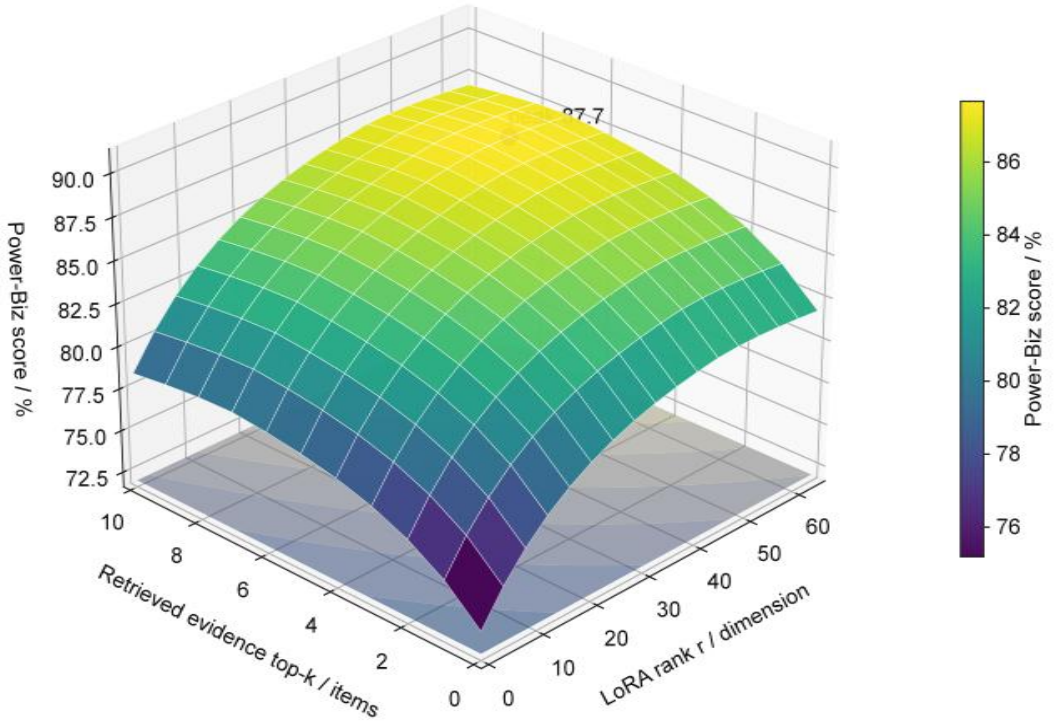

Figure 5: Three-dimensional plot of the LoRA ranking and retrieval depth.

As shown in Figure 5 is the Validation power-Biz Response Surface under LoRA Rank and Retrieval Top-K, showing the expected latency baseline plot. The score increased sharply from ranking 0-32, with the highest in first place; Then it was near ranking 48-50 and position K+6 reached 89.2 percent. Rank 64 produced no further gain, and top-k above 7 increased latency with limited score improvement. Therefore, the final configuration is: LoRA Rank 48 and Top-K Retrieval 6 across all tests.

In terms of Retrieval is not complicated or out of context. Top-K1~3 often miss the corridor-vegetation or substation-thermal rule; but Top-K6 has given us sufficient options without overcrowding it. LORA also needed to be of high degree, yet a low rank was necessary to avoid overfitting. Rank 48 entered the steady-state operating area.

Fig. 6 reports validation convergence. RAG-only adaptation raised Power-Biz from 73.5% to 80.2%, mainly through better rule basis and ticket wording. LoRA-only reached 83.6% by improving domain expression and risk mapping. The RAG+LoRA model reached an accuracy of 88.7%; After adding a verifier, the stability ranged within [89.1%, 89.2%], and non-supported high-risk statements were suppressed while uncertainty could not be avoided without evidence present.

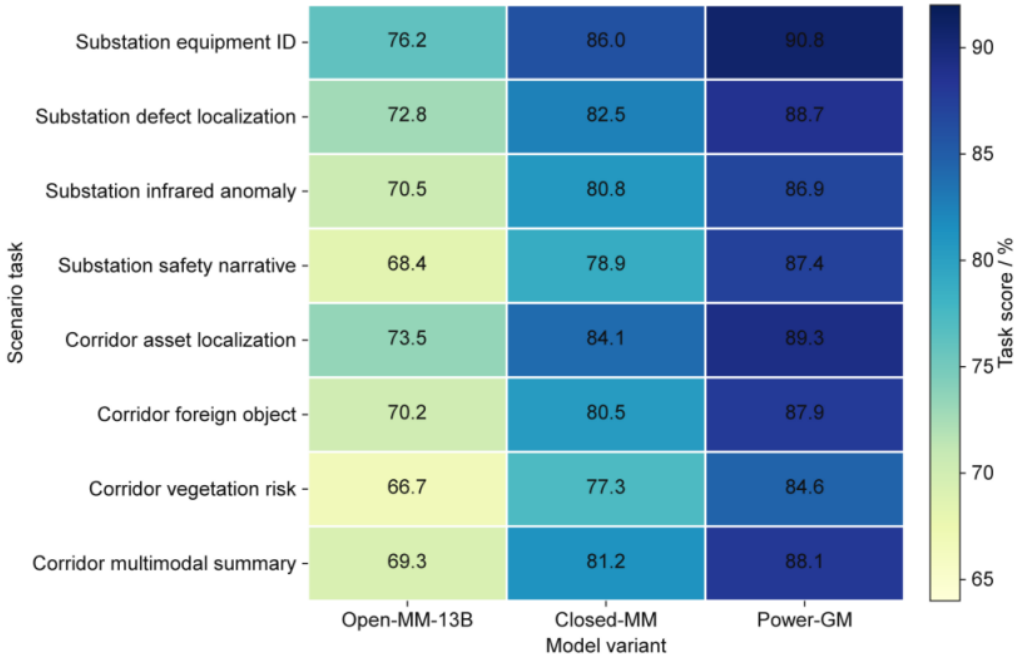

Figure 6: Adaptation Convergence Curves for Retrieval, Low-Rank Tuning, and Safety Verification.

Figs. 4-6 Fix the empirical setup. Fig. 4 selects Power-GM as the best engineering-balanced candidate. Fig. 5 sets LoRA rank 48 and retrieval top-k 6. Fig. 6 shows stable convergence without excessive epochs. Therefore, the same-adapted model was used in subsequent scenes-level and ablation experiments.

Power-GM outperforms Open-MM-34B through structured adaptation rather than Scale. Open-MM-34B improves over smaller open models but loses efficiency. Power-GM increases both of these factors that affect ticket usability more strongly than any other. Its 17.0-point advantage over Open-MM-34B in hallucination control indicates that verifier and rule filtering changed output behavior.

Fig. 5 shows adaptation saturation. Top-k 1–3 corrects many missing-rule cases; top-k 6 helps boundary vegetation and thermal-anomaly samples. Beyond top-7, Latency increases and Score gain drops significantly. Low LORA ranks underfit the domain structure; extremely high LORA rank additions are minimal and sensitive to repeated asset expressions.

Fig. 6 confirms that the final setting is stable. At epoch 6, all configurations stabilised around ±0.3% fluctuation with respect to the performance of the validation set. The RAG-only and LoRA-only curves remain lower, which support the later-ablated result; Retrieval enhances evidence accessibility, LoRA addresses domain alignment, and the verifier strengthens output verification.

Validation-test Deviation is in check. Figures 9A-9B show that the precision for Power-GM was 89.1% and 84.3%, respectively, during training-validation; The drop shows more strict division of routes and equipment bays, invisible corridors, and surveillance cameras. The model is still higher than the 80.0 per cent threshold, with most residuals within recognisable visual areas.

Text-GM can serve as the reference. Good at matching rules but not suitable for identifying objects' regions. This indicates that the substantial advantage is associated with linking languages to pictures, Rules and verification; The performance of natural language itself has not been raised accordingly.

3.2 Scenario-Specific Performance, Interactive Behavior, and Case Analysis

The second result is whether the selected setting can still be effective under different situations, which modules contribute to each gain. Fig. 7 presents the Scene-Task Matrix; The meanings of devices dominate Substation Inspection, and Large Scence location-based Positioning is focused primarily around Plants near Corridors Obstacles and Route Background Knowledge.

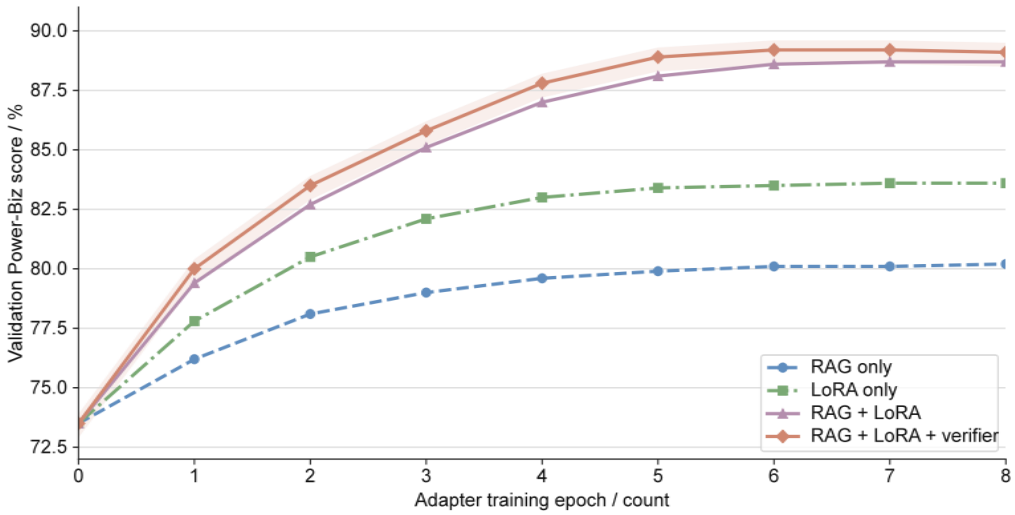

Figure 7: Substation Intelligent Inspection and Transmission Corridor Visualisation Scenario-Tasks Performance Matrix.

Fig. 7 shows that Power-GM outperformed Open-MM-13B and Closed-MM on all eight tasks. Substation Work: Power-GM obtained 90.8% in the identification of equipment; received 88.7% in defect location; Received 86.9% Infrared anomaly detection; Receiving 87.4%. Among them, the greatest increase was seen in the safety narration generation problem; From 78.9 per cent to 87.4 per cent, which required linking an objects-abnormalities-risk phrases-rule base structure ticket for this kind of problem.

Power-GM achieved 89.3%, 87.9%, 84.6%, and 88.1% accuracy for tasking of the corridors respectively: asset location; Foreign objects detection; Vegetation risk prediction, multimodal path summarization; Vegetation risk remained the weakest corridor task because visually similar vegetation may lie outside or inside the protection buffer, and a single image may not provide reliable distance evidence. Closed-MM dropped from 84.1% in asset localization to 77.3% in vegetation risk; and open-MM-13b also declined to 66.7%. Adapted model obtained the highest scores in which both rule-context interpretation and spatial-risk assessment were needed.

As shown in Figure 8b, which is an Ablation Matrix for all components.

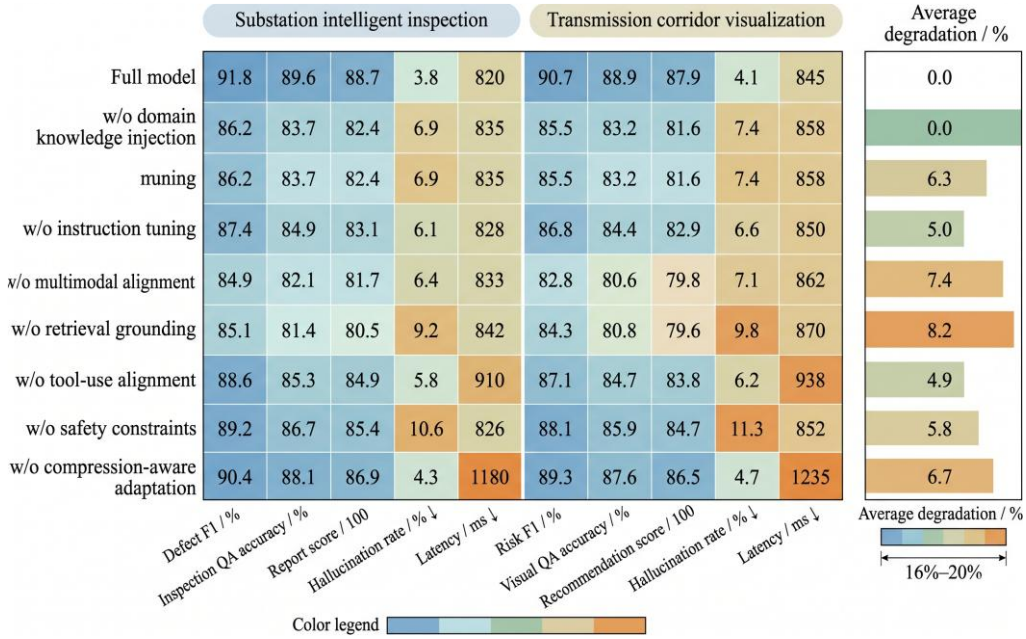

Figure 8: Module Ablation Matrix of the Adapted power-GMM structure.

In total, all aspects of the Power-GM's evaluation scores are as follows: Visual Grounding 89.0%, Risk Reasoning 86.0%; Rule Adherence 87.0%, Hallucination Avoidance 88.0%, Robustness 84.0%. Removing retrieval reduced rule compliance to 80.1% and hallucination control to 82.6%, showing more unsupported rule statements. Removing LoRA reduced visual grounding to 83.8% and risk reasoning to 81.1%, indicating weaker power-specific visual-text mapping.

The grounding correction obtained the best-grounded results. Thus, in this context of "no Visual Grounding", the robustness dropped to 77.8 per cent. The error log shift to adjacent substation's components or corridor background objects occurred. After removal of the safety verifier, ground retention was approximately 87.9% and hallucination control was around 81.5 per cent. Mainly filtered out unsupported rules, overly serious risk claims and unsafe-ticket behavior.

The ablation results validate the Figure 2 process. Retrieval supply rules and ticket evidence; LoRa adapts to the domain behaviour, ground calibrates object-Region References, and verifies unsupported outputs. Effects are as follows: Retrieval cannot eliminate erroneous areas; Grounding correction is unavailable for missing conditions. Therefore, the final setting was better than its parts.

Need for task-level report. The aggregate score conceals the weak vegetation-risk result; The ground only score misses the retrieval-driven rule-compliance benefit. Therefore, the Scene-TASK and Abolation Matrices can simultaneously show deployable strengths and remaining evidence weaknesses, particularly route geometry and multiple-frame context.

As in the latest power-asset detection studies, constraint-based detectors can achieve relatively good performance under a limited number of defects or lines [22, 24]. Power-GM adds the perception layer with rule evidence, structure risk judgment and hallucination filtering. Lightweight inspection algorithms are still applicable to edge devices, but they cannot achieve high-performant ticket-level reasoning on their own.

Closed-<mM> was still popular overall recognitions; Substation equipment Recognition rate: 86.0%, Corridor Asset localisation rate of: 84.1%. Its deficiency was rule-based task-oriented work. Power-GM outperformed it by 8.5 percentage points in the substation safety narrative and 7.3 percentage points in corridor-vegetation-risk, showing filtered retrieval and

trained result structure characteristics.

Open-MM-13B served as the lower-bound multimodal reference. Its localisation capability met requirements, but rule-based work was unsatisfactory. Power-GM exceeded it by 20.3 percentage points in corridor vegetation risk and 19.4 percentage points in substation safety narrative. Tasks that need to be reasoned jointly across objects, scenes, risks and rules.

3.3 Efficiency, Error Diagnosis, and Deployment Interpretation

Finally, connect the accurate results with the deployment limits and remaining failure domains in this way. In power inspection, a model with high offline accuracy may still be unsuitable if it cannot respond within the platform tolerance, fit into the available memory budget, or keep false ticket rates under control. Fig. 9 shows the deployment envelope for comparison of scores, latencies, memories and costs under a single experiment view.

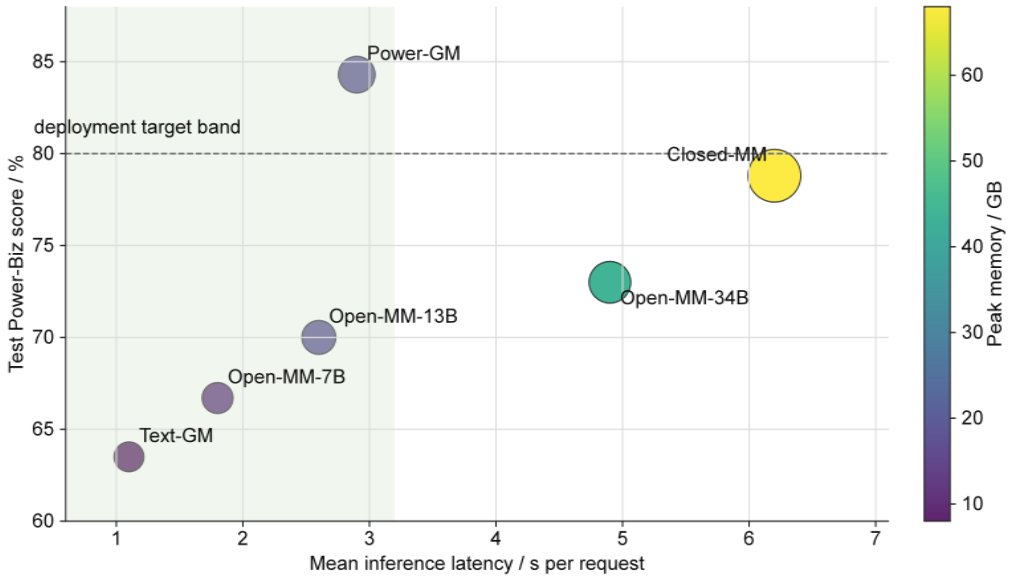

Figure 9: Deployment envelope of candidate models based on scores, latencies, memories and costs constraints.

Fig. 9 shows the deployment range. Text-GM was fastest and cheapest, with 1.1 s latency and 8 GB memory, but its Power-Biz score was only 63.5%. Open-MM-7B, Open-MM-13B, and Open-MM-34B achieved up to 66.7% and beyond; Latency increased from 1.8s to over 4.9s and Memory went from 12GB to 44GB+. Closed-MM scored 78.8%, but latency reached 6.2 s and relative cost was the highest. Power-GM achieved 84.3% with 2.9 s latency and 18 GB local memory.

Only Power-GM exceeded 80.0% and had a latency of less than 3.2 seconds. The position for repeatedly inspected call items is mostly normal or of low risk. Closed-MM is accurate but costly and slow; small open models are deployable but less reliable. Power-GM can reasonably achieve a high degree of accuracy, timely response and low deployment consumption simultaneously.

Residual errors are analysed using the last configuration selected. Figure 10 shows the error-source distribution based on 400 failed samples.

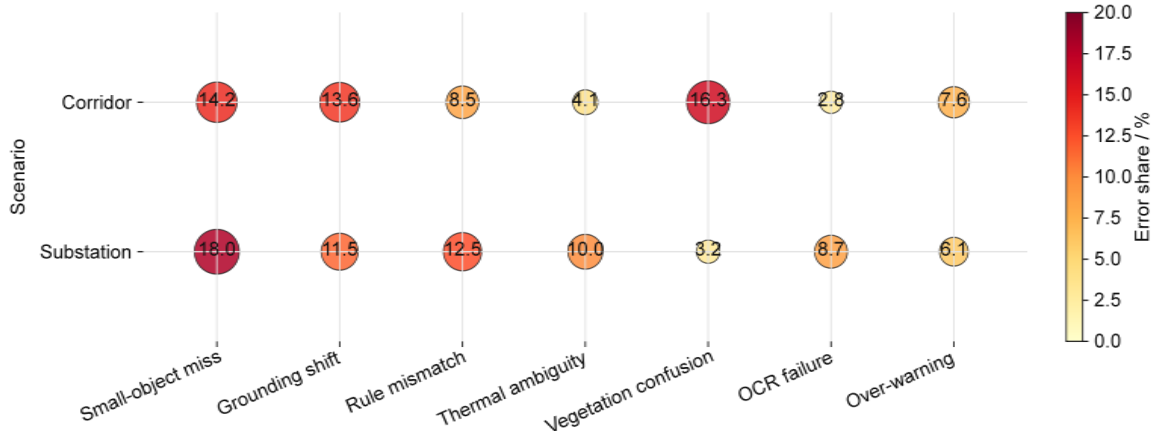

Figure 10: residual error diagnosis after deep adaptation.

Errors varied by Scene. The primary faults in the substations include: small object omission (18.0%), rule misassignment (12.5%) and grounding displacement (11.5%); The problems of temperature ambiguity (10.0%), Ocr failure (8.7%), over-warning (6.1%) and vegetation confusion (3.2%) also occurred frequently. In corridor samples, vegetation confusion led at 16.3%, followed by small-object miss at 14.2%, grounding shift at 13.6%, rule mismatch at 8.5%, over-warning at 7.6%, thermal ambiguity at 4.1%, and OCR failure at 2.8%.

This error explains the task result(s). Substation weaknesses concentrate on small targets, thermal ambiguity, and OCR-dependent evidence, affecting defect localization, infrared interpretation, and ticket generation. Corridor weaknesses concentrate on vegetation distance and region shift, affecting vegetation-risk and foreign-object tasks. Therefore, the residuals reflect not random outputs but scenes with good observations.

The verifier decreased unapproved high-risk claims. Some baselines identified the correct conductor or cabinet but attached severe-risk phrases without rule support. Finally, either referred to a suitable rule or downgraded the result as uncertain. Cannot restore the missing visual information such as corridor vegetation distance without route geometry, stereo cues and so on.

Within the limits of deployment. Adapted Model Is Appropriate For Evidence-Based Screening, Preprocessing Of Tickets, Workload Reduction And Trace-Alarm Generation. It cannot be regarded as an independent protection-decision system. Cases of high-risk corridors remain in need of route geometric analysis combined with multiple frames of evidence from unmanned aerial vehicles (UAVs), or human examination.

Efficiency can include the indicator of deployment when selecting a model. Accuracy only favours Closed-MM on some occasions; However, the delay and expenditure rates make closed-calls beyond continuous-call consideration. Open-MM-7B is deployable but weak in reasoning. GM power occupies the viable range as retrieval and adaptation enhance scores without exceeding the closed-model limit.

Case review showed three correction patterns: grounding calibration corrected thermal anomalies assigned to the wrong cabinet component; retrieval and verification downgraded generic severe vegetation warnings when distance evidence was insufficient; and reflective normal substation images were changed from false tickets to attention-level observations with insufficient-defect evidence.

Remaining errors point to sensing and data limits. Small-object miss requires higher-resolution crops or detector proposals. Thermal ambiguity has a calibrated threshold and state. Vegetation confusion needs a Distance/Route-Buffer evidence item. OCRL error requires stronger recognition or meta-data alignment. Therefore, in the future data expansion direction,

high-resolution small-object crops, calibrated thermal labels, route geometric features, and paired OCR-metadata should be prioritised over general inspection pictures.

4 Conclusion

Aiming to develop an empirical model for assessing, screening and applying generalised-large-scale software in the field of substations' inspection and transmission corridors'. Using a retained multimodal corpus, fixed splits, unified prompts, controlled adaptation, and deployment indicators, the adapted Power-GM achieved the best engineering balance, with an 84.3% test Power-Biz score and 2.9 s mean latency.

The study set up the inspection evidence object as its unit of measurement, linked images with thermal signs, OCR data, metadata information, corridor properties, rule contents, ticket details in order. This Structure could support visual Grounding, risk Reasoning, Rule Compliance, hallucination Control, Robustness, and Deployment Tests.

(2) Selection and Adaptation need to be judged simultaneously in this study. Retrieval improved rule adherence, LoRA reinforced domain understanding, grounding calibration reduced object-Region inconsistencies, and the safety verifier curbed unsupervised high-risk outcomes. The best operating region is LOPA Rank48, with Retrieval Top- K6, and Accuracy gain remains below the latency goal.

(3) The study found the remaining boundaries. Small-object miss, grounding shift, thermal/Ocr ambiguity, and vegetation-background misidentification have still left some residuals. Future work should connect the adapted model with multi-frame UAV evidence, route geometry, LiDAR or stereo distance, calibrated thermal thresholds, and active field feedback before high-risk autonomous operation is considered.

Empirical data also show that the manuscript can be considered a bounded model selection experiment. The reported gains are influenced by the retained corpora, the two selected scenes, the fixed scoring definition and tested deployment environment. After re-calibrating the final threshold values and model selection based on this utility's equipment inventory, route plan, rules library, image acquisition device parameters, as well as the alarm activation criteria for that particular utility Environment.

Within this boundary, there is a specific path for research: Scenario Corpus Construction, Model Selection and Adaptation, Result Plotting, Deployment Interpretation. The following will be replaced or expanded with utility-specific fields in the previous benchmark, and then repeat the above test procedure prior to commercial application.

Funding

This work was supported by Research on Key Technologies for Non-Power-Outage Transformation and Upgrading of Integrated Automation Systems.

About the Author

Ke Shi, female, Ethnic Hui, born in September 1993, graduated from Newcastle University with a master's degree. Currently serves as a relay protection specialist in the Power Dispatching Control Center of Guizhou Power Grid Co., Ltd., main research direction being power system and automation; Email:18786755904@163.com

Jing Niu, female, Ethnic Han, born in October 1984, graduated from Guizhou University with a postgraduate education and a master's degree. She currently serves as the Relaying

Protection Specialist at the Power Dispatching and Control Center of Guizhou Power Grid Co., Ltd., main research direction is power system and its automation. Email: 15286085692@163.com

Weixiang Qiao, male, Han ethnicity, born in January 1993. Graduated from Guizhou University with a bachelor's degree and a bachelor's degree in engineering. Currently serves as a relay protection specialist in the Power Dispatching Control Center of Guizhou Power Grid Co., Ltd., main research direction being power system and automation; Email: 18286126843@163.com

Xing Zhang, male, Ethnic Han, born in November 1997, graduated from Sichuan University with a bachelor's degree in engineering. He currently serves as the Maintenance Specialist at the First Substation Management Institute of Zunyi Power Supply Bureau, Guizhou Power Grid Co., Ltd. His main research direction is power system and its automation. Email:15597752396@163.com

References

[1] Faisal, M. A. A., Mecheter, I., Qiblawey, Y., et al. (2025). Deep learning in automated power line inspection: A review. Applied Energy, 385, 125507.

[2] Ruszczak, B., Michalski, P., & Tomaszewski, M. (2023). Overview of image datasets for deep learning applications in diagnostics of power infrastructure. Sensors, 23(16), 7171.

[3] Wang, Q., Yang, L., Zhou, B., et al. (2023). YOLO-SS-Large: A lightweight and high-performance model for defect detection in substations. Sensors, 23(19), 8080.

[4] Santos, T., Cunha, T., Dias, A., et al. (2024). UAV visual and thermographic power line detection using deep learning. Sensors, 24(17), 5678.

[5] Rong, S., He, L., Atici, S. F., et al. (2025). Advanced YOLO-based real-time power line detection for vegetation management. IEEE Transactions on Power Delivery, 40(4), 2142-2153.

[6] Achiam, J., Adler, S., Agarwal, S., et al. (2023). GPT-4 technical report. arXiv, 2303.08774.

[7] Alayrac, J. B., Donahue, J., Luc, P., et al. (2022). Flamingo: A visual language model for few-shot learning. In Advances in Neural Information Processing Systems (Vol. 35, pp. 23716-23736).

[8] Li, J., Li, D., Savarese, S., et al. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 19730-19742).

[9] Liu, H., Li, C., Wu, Q., et al. (2023). Visual instruction tuning. In Advances in Neural Information Processing Systems (Vol. 36, pp. 34892-34916).

[10] Fu, C., Chen, P., Shen, Y., et al. (2023). MME: A comprehensive evaluation benchmark for multimodal large language models. arXiv, 2306.13394.

[11] Liu, Y., Duan, H., Zhang, Y., et al. (2024). MMBench: Is your multi-modal model an all-

around player? In Computer Vision – ECCV 2024 (pp. 216-233).

[12] Li, Y., Du, Y., Zhou, K., et al. (2023). Evaluating object hallucination in large vision-language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 292-305).

[13] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459-9474).

[14] Asai, A., Wu, Z., Wang, Y., et al. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In Proceedings of the International Conference on Learning Representations.

[15] Hu, E. J., Shen, Y., Wallis, P., et al. (2022). LoRA: Low-rank adaptation of large language models. In Proceedings of the International Conference on Learning Representations.

[16] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 8748-8763).

[17] Kirillov, A., Mintun, E., Ravi, N., et al. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015-4026).

[18] Abdelfattah, R., Wang, X., & Wang, S. (2020). TTPLA: An aerial-image dataset for detection and segmentation of transmission towers and power lines. In Computer Vision – ACCV 2020 Workshops (pp. 601-618).

[19] Touvron, H., Martin, L., Stone, K., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv, 2307.09288.

[20] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems (Vol. 35, pp. 27730-27744).

[21] Yue, X., Ni, Y., Zhang, K., et al. (2024). MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9556-9567).

[22] Sun, Y., Sun, X., Lin, Y., et al. (2025). Substation equipment defect detection based on improved YOLOv8. Sensors, 25(8), 2607.

[23] Lu, L., Chen, Z., Wang, R., et al. (2023). Yolo-inspection: Defect detection method for power transmission lines based on enhanced YOLOv5s. Journal of Real-Time Image Processing, 20(5), 104.

[24] Liu, C., Wei, S., Zhong, S., et al. (2024). YOLO-PowerLite: A lightweight YOLO model for transmission line abnormal target detection. IEEE Access, 12, 105004-105015.

[25] Li, H., Wu, Z., Sun, Y., et al. (2025). A lightweight RKM-YOLO algorithm for transmission line fault inspection. Energy Reports.