electronics

MDPI

Article

Trustworthy Evaluation of Vision–Language Models for UAV Power-Line Inspection: A Resolution- and Parity-Controlled Benchmark

Linghao Zhang [1,2,*], Siyu Xiang [1,2], Junwei Kuang [1,2] and Peiyu Yi [1,2]

[1] State Grid Sichuan Electric Power Research Institute, Chengdu 610041, China; xiangsiyu4024@163.com (S.X.); kuangjunwei1258@163.com (J.K.); yipeiyu1994@gmail.com (P.Y.)
[2] Power System Security and Operation Key Laboratory of Sichuan Province, Chengdu 610041, China
* Correspondence: gwscdl-dkyzlh@sc.sgcc.com.cn

Abstract

Vision–language models (VLMs) are often reported to outperform vision-only backbones for UAV power-line inspection. We examine that claim using ElecVQA-Bench, a 56,972-item resource derived from InsPLAD, and an audit of partition, evaluated items, label space, replication, input resolution, and side information. The original split has substantial source-photograph dependence and is used for paired diagnostic comparisons. Reported tower-disjoint retraining provides a limited check on new-tower performance within the same source dataset. On the original binary task, no difference is detected by a paired item-level test between the selected Swin and Qwen runs; this does not establish equivalence. Fine-grained results change with resolution, prompt information, class support, and seed. A 14-tower, three-seed comparison changes direction across seeds, and classes represented by only a few items substantially affect macro recall. The contribution is a reproducible task representation and reporting protocol for examining these dependencies, rather than a general ranking of model families or evidence of end-to-end UAV deployment.

Keywords: vision–language model; vision transformer; UAV inspection; multimodal benchmark; evaluation bias; input resolution; cost-sensitive learning; abstention; low-rank adaptation; quantization

Check for updates

Academic Editors: Muhammad Ali Farooq, Peter M. Corcoran and Alam Noor

Received: 9 October 2026
Revised: 22 October 2026
Accepted: 28 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Inspection of overhead transmission infrastructure is critical to grid reliability, and UAV imagery has become the dominant acquisition mode. Two model families compete to interpret that imagery. Purpose-trained vision backbones—convolutional networks and, increasingly, vision transformers—are the incumbent option: compact, fast, and well understood. Vision–language models are the challenger, motivated by the argument that web-scale pre-training yields better few-shot behavior and an interpretable rationale rather than a bare class label.

The study asks how reported model comparisons change when evaluation choices are made explicit. It does not isolate an intrinsic model-family effect: training resources, side information, visual encoders, and tokenization are not fully matched. The contribution is an audit of the tested configurations, with separate statements of observed differences and unresolved confounding.

The first choice is the evaluation partition and the provenance of the checkpoint being evaluated. An earlier comparison used different task files and training configurations. The revised comparison evaluates stored predictions on the same binary item set. Because both the file and the earlier training configuration differed, the old-to-new change cannot be attributed to partition choice alone.

The second choice is input resolution. InternVL uses dynamic 448 px tiles, with a recorded mean aggregate input area approximately equal to a 768 px square. The tested vision-only resolutions bracket this mean; they are not exact per-item pixel, visual-token, or compute matches. The associated score changes motivate reporting preprocessing budgets alongside recognition performance.

The third choice is replication. On the reported tower14 split, the Qwen–ResNet difference changes direction across three matched seeds. This is a descriptive observation for those models and runs, not an estimate of the variance of either entire model family.

The remaining choices concern the evaluated items, the label space, and side information. The VLM prompt supplies an asset name, definitions, and asset-specific candidates, whereas the vision-only classifier receives pixels and a fixed label set. Comparisons are therefore between the tested systems, not information-matched architectures. Explicit abstention reporting further separates an automated normal decision from an item routed for review.

Our contributions are as follows.

1. A six-axis reporting and parity-audit protocol that records partition, evaluated items, label space, replication, input budget, and side information, and identifies unmatched axes rather than treating them as fully controlled.
2. A resolution and split-provenance analysis distinguishing original-split diagnostics from reported tower-disjoint retraining. The observed Qwen–ResNet ordering is not stable across the tested splits and seeds.
3. Separate accounting for defects predicted normal, normal items predicted defective, and abstentions, with a cost model in which abstentions are routed to manual review.
4. A prevalence- and review-cost sensitivity analysis. The simple odds-ratio re-basing applies when abstention is priced at zero; a general crossover expression retains the abstention terms.
5. A rare-class decomposition and class-set sensitivity analysis showing how four- and five-item classes affect the reported seven-class macro recall. Their absolute contributions sum to 22.9 percentage points in one comparison, while their net contribution is 5.72 points.
6. ElecVQA-Bench, a 56,972-item derived benchmark built from the public InsPLAD dataset [1] by a deterministic pipeline. Construction code, split manifests, prompt templates, the evaluation harness, and adapter weights are prepared for public release upon acceptance; no independently collected photographs are added.

Section 2 reviews related work. Section 3 describes the construction and statistics of ElecVQA-Bench. Section 4 presents the cost-sensitive framework, the parity audit, and the experimental configuration. Section 5 reports results, Section 6 discusses what the controlled comparisons establish and states the limitations, and Section 7 concludes.

2. Related Work

Power-line inspection from UAV imagery. InsPLAD [1] provides 10,607 UAV images across 17 asset categories with attached defect-classification and anomaly-detection subsets, together with baselines using DifferNet [2] and CS-Flow [3], both normalizing-flow density estimators originally proposed for industrial surface inspection. Nguyen et al. [4] survey vision-based power-line inspection and identify data scarcity and class imbalance

as recurring obstacles. Tao et al. [5] detect insulator defects from aerial imagery with a cascaded convolutional architecture, and the wider literature builds predominantly on general-purpose detectors such as Faster R-CNN [6] and the YOLO family [7]. These works report detection or classification accuracy on their own partitions. We did not identify, in the literature reviewed here, a study that verifies partition and label-space equivalence before comparing model families. This review is selective rather than systematic, and motivates the audit rather than claiming exhaustive coverage.

Broader industrial monitoring and reliability assessment. Data-driven distributed monitoring frameworks optimize interconnected and manufacturing systems by integrating sensor information, subsystem couplings, fault diagnosis, and control decisions [8,9]. ElecVQA-Bench addresses a narrower perception-quality question: whether two model families classify the same ground-truth ROI crop consistently under controlled evaluation choices. It does not evaluate distributed state estimation, fault propagation, information exchange between subsystems, closed-loop control, or end-to-end system reliability. The approaches are therefore complementary: visual inspection can supply one input to such frameworks, but the present benchmark does not demonstrate a system-level reliability improvement.

Industrial defect and anomaly detection. The evaluation conventions adopted here—crop-level scoring, heavy normal-to-defect imbalance, and a small number of rare defect categories—follow the industrial anomaly-detection literature established by MVTec AD [10] and memory-bank methods such as PatchCore [11]. Recent surveys [12] document the migration of that field toward foundation-model backbones; we examine the evaluation methodology behind that migration.

Cost-sensitive evaluation. That aggregate accuracy is an unsafe summary under asymmetric misclassification costs is long established. Provost and Fawcett [13] argue for evaluating classifiers across the full range of plausible cost ratios rather than at a single operating point; Drummond and Holte [14] formalize this as cost curves; Elkan [15] gives the general reduction of cost-sensitive learning to a threshold problem. Calibration research [16] separately shows that aggregate correctness does not certify that a model's confidence, or in a discrete answer space its qualitative error profile, is trustworthy. We adapt this reasoning to the miss, hallucination, and abstention vocabulary of industrial visual question answering.

Class imbalance and long-tailed recognition. Several test classes have single-digit support, so macro recall gives a handful of items the same class weight as thousands of normal items. Specific class-balanced reweighting [17] and representation–classifier decoupling [18] are relevant mitigations, but those particular methods were not evaluated. The vision-only training does use conventional class-weighted cross-entropy; this is distinct from claiming no imbalance handling at all. The per-class analysis reports the effect of limited support rather than establishing reliable rare-class generalization [19].

Vision–language models and hallucination evaluation. General-purpose VLMs derive from contrastive pre-training [20] and instruction tuning [21,22]. InternVL [23] and Qwen2.5-VL [24] are used here as representative open families. Because the experimental focus is evaluation methodology rather than architecture, we do not summarize the technical reports of every model generation. The hallucination rate defined in Section 4.1 is the industrial-inspection analogue of the object-hallucination metrics of CHAIR [25] and POPE [26], with the difference that the present setting has a closed answer space and an explicit abstention token, so hallucination and refusal are separately measurable rather than confounded.

Parameter-efficient adaptation and statistical testing. LoRA [27] and its quantization-aware variant QLoRA [28] are the default route to adapting large models on a single accelerator. The multi-task ablation is grounded in the classical result that jointly trained

tasks can regularize one another [29], tempered by the observation that task pairing depends on task relatedness [30]. ResNet-50 [31] is used as the convolutional baseline and Swin-T [32] as the hierarchical vision-transformer baseline. McNemar's exact test [33] is applied following the analysis of paired tests for classifier comparison by Dietterich [34].

3. Materials: ElecVQA-Bench

3.1. Source Data and Construction

ElecVQA-Bench is derived from three InsPLAD subsets. The source release reports 10,607 UAV photographs and 17 asset categories. The construction audit retained 10,552 detection images and 28,824 boxes after normalization, alongside fault-classification and anomaly-detection crops. Normalized labels and derived items do not add independently collected imagery. The revision evidence package records the split-manifest digests, source-key derivation, rare-class support, selected checkpoint arguments, copied prediction files, and sign-flip reproduction checks. Construction resources are prepared for release under the access policy stated in the Data Availability Statement.

Algorithm 1 summarizes crop extraction, task construction, and the original filename-grouped split. This split is identifier-disjoint at the reported parent level; it is not certified as independent at the global source-photograph level.

Algorithm 1: ElecVQA-Bench construction

Input: raw InsPLAD subsets; expansion ratios $R = \{1.0, 1.5, 2.0, 3.0\}$

Output: item sets for T1 to T3; digested split manifest

1. Audit source data for category mapping, duplication, and box validity.
2. For each annotated instance, expand its box by every r ∈ R about its center, clip to the image, and enlarge to a 64 px minimum short side.
3. T1 (grounding): image, asset name, normalized box.
4. T2 (binary assessment, r = 2.0): crop, P3 prompt, label ∈ {normal, defect}. "Cannot judge" is an available answer but never a ground-truth label. Negatives are drawn from `good` crops of the same asset type.
5. T3 (fine-grained typing, r = 2.0): crop, P3 prompt, label ∈ closed defect set ∪ {good}. "Cannot judge" is again available but not a ground-truth label.
6. Group items by the filename-derived `parent_id`. Apply a group-level 70/10/20 split with seed 42, stratified by each group's dominant defect class. Asset–label combinations with fewer than five instances are assigned to the test partition.
7. Assert that no `parent_id` crosses a split, that no perceptual-hash collision crosses a split, and that every class present in the training split has at least five training instances. Test-only rare classes are reported separately and are not covered by the training-support assertion.
8. Write SHA-256 digests of splits, scripts, and prompts to a manifest.

Questions are phrased neutrally, and an abstention answer is available. No training target is abstention. Consequently, observed abstention is not established here as a calibrated uncertainty estimate; it may reflect the prompt, adaptation, and model disposition.

The rare-allocation rule operates on asset–label combinations rather than on the seven global classes. Combinations with fewer than five records before parent grouping are retained in test. This was an engineering retention rule, not a pre-registered zero-shot hypothesis. It deliberately changes test composition, and test-only classes are interpreted separately from ordinary training-supported classification. Group integrity takes precedence over rare allocation: all members of a source group must be assigned together. Exact

per-combination counts and the normalization audit are needed to reproduce this rule; no missing class count is inferred from the aggregate totals.

3.2. Benchmark Statistics

Table 1 gives the task-item counts. Split proportions are the realized output of the `parent_id`-grouped split in Algorithm 1. Because group sizes are uneven, realized proportions deviate from the 70/10/20 target by up to 1.5 percentage points; realized counts, not target ratios, are authoritative. The audited split manifest contains 34,815 train, 4996 validation, and 9941 test records, totaling 49,752 base crop/image records and 49,706 unique crop identifiers. The composition table reports 56,972 task items after task-specific expansion and contains 49,704 unique crop identifiers. The mapping is exact: 42,482 crop identifiers have the same row count in the base and task layers; 7222 identifiers are shared by T2 and T3 and contribute one additional task row each; and two base identifiers are absent from the three reported task files. Thus $49{,}752 + 7222 - 2 = 56{,}972$. Forty-six duplicate T1 rows are internal to both layers and do not change this net difference. The SHA-256 digest of each split file matches its manifest entry; this confirms file identity, not split validity. The filename-derived source proxy is the first `DJI_\d+` token in `image_path`; no T2 or T3 record required the `parent_id` fallback. This proxy is not a globally unique source-photograph identifier: 707/877 T2 keys and 826/964 T3 keys occur with more than one date prefix. A conservative date+DJI key, available for all records, reduces train–test overlap to 750/2021 (37.11%) in T2 and 5267/7284 (72.31%) in T3. The original DJI-only proxy gives 1975/2021 (97.72%) and 7255/7284 (99.60%), but should be read as a coarser upper bound relative to the date+DJI proxy matching rate rather than as a unique-photograph audit. The date+DJI key is also a filename-derived proxy and does not by itself confirm unique photograph identity. The identifier `parent_id` remains filename-derived and is not, by itself, a globally normalized source-photograph identifier.

Table 1. ElecVQA-Bench composition.

Task	Total	Train	Validation	Test
T1, grounding	10,605	7448	1068	2089
T2, binary assessment	9997	6949	1027	2021
T3, fine-grained typing	36,370	25,460	3626	7284
Total	56,972	39,857	5721	11,394

The counts in Table 1 are derived task items, not independent photographs. The resource adds task-specific representations, prompt and answer schemas, label crosswalks, explicit abstention handling, and item-level provenance. Table 2 separates these additions from source data; T1 is constructed but not evaluated.

The original split is a paired diagnostic, not an independent-photograph validation split. The verified proxy-key audit reports two matching rates: under the coarse DJI-only proxy, 1975/2021 (97.72%) T2 test items and 7255/7284 (99.60%) T3 test items share a key with training; under the more conservative date+DJI proxy, the rates are 750/2021 (37.11%) and 5267/7284 (72.31%). Parent-identifier disjointness and perceptual-hash checks therefore give narrower guarantees than normalized source-photograph independence. Table 3 distinguishes the reported exclusions from their intended interpretation. Related images can inflate absolute performance; using the same test items across models and clustering a significance test do not repair training leakage. Tower and tower14 rows are used only as reported new-tower checks within this dataset, not external validation.

Table 2. Source material and derived-task increment; no new independently collected imagery is added.

Aspect	InsPLAD Source	ElecVQA-Bench Representation
Visual observations	UAV images, asset annotations and crops	Reuses source imagery; deterministic ROI/task items
Assessment target	Asset/defect annotations	T2: normal/defect; T3: closed defect labels plus good
Answer handling	Source annotation labels	Fixed answer schema and explicit abstention option
Traceability	Source records	Derived-item identifiers, prompt versions, split and prediction records
Validation scope	Source benchmark tasks	T2/T3 classification evaluated; T1 grounding not evaluated

Table 3. Reported split dependence and interpretation. Proxy-key matching rates are not global near-duplicate-image audits.

Split	Reported Dependence/Exclusion	Permitted Interpretation
Original T2/T3	DJI-only proxy: 97.72%/99.60%; date+DJI proxy: 37.11%/72.31%	Paired diagnostic with substantial proxy-key dependence; neither proxy establishes unique photograph identity
Crop retraining	488/6672 test image paths also in train; all 42 test towers seen	Crop/item diagnostic; not unseen-photo or unseen-tower validation
Image retraining	Image paths and parent groups excluded; all 41 test towers seen	Path-group diagnostic; not unseen-tower validation
Tower/tower14	Reported no crossing of image path, parent group or tower	Limited within-source new-tower check; source normalization and checkpoint provenance remain required

T3 contains 36,370 items at ROI expansion ratio 2.0: 39,145 candidate crops minus 2775 exclusions (7.1%). The construction audit and per-reason exclusions are prepared for release. They are distinct from an independent source-photograph overlap check. T1 remains a resource for future grounding experiments.

T2's test set contains 1255 normal and 766 defective items, giving $\pi_N = 0.621$ and $\pi_D = 0.379$, not the targeted 1:1 ratio. Negatives are matched 1:1 before splitting, but `parent_id` grouping distributes them unevenly and defective crops cluster more strongly by parent group. The realized 37.9% prevalence is a split property, not a designed value, and Section 4.4 shows it is the most influential quantity in the cost analysis.

The original T3 test set contains six defect types and `good`, giving seven ground-truth classes. Unless another class set is explicitly stated, original full-T3 balanced accuracy means unweighted seven-class macro recall, with abstentions counted as errors. T2 uses `normal` for the no-defect concept and T3 preserves `good`; Table A1 gives the crosswalk. Table A2 gives the available test support. The verified rare-class support is 0/0/4 train/validation/test items for `peeling-paint` (four test source keys from three towers) and 15/2/5 for `torned-up` (five test source keys from three towers). Other class- and asset-specific cross-split and independent-cluster counts are not established by these test totals, limiting rare-class inference.

3.3. Prompts and Evaluation Protocol

Four prompt specifications are prepared for release. P1 contains the question and candidates; P2 adds definitions; P3 uses an ordered decision tree; P4 adds reference images.

Main reported results use P3, with guided JSON output and format failures recorded separately. Exact templates and per-asset candidate lists are part of the planned benchmark release. Asset identity and the decision tree remain additional information relative to a pixels-only classifier.

We declare three asymmetries between the families. First, guided decoding guarantees a syntactically legal VLM label, whereas a vision-only classifier emits one by construction; the VLM therefore solves constrained generation and the classifier discrimination. Second, the VLM receives only the answer options possible for the asset, typically two to four, while the classifier performs unconstrained seven-way classification and can predict an impossible class. Stored predictions show this affects 7 of 7284 items (0.1%), and oracle correction changes balanced accuracy by at most 0.55 points. Third, the VLM prompt supplies an asset name and decision tree, while the classifier receives only image and fixed label set. Cross-family comparisons are therefore system-level comparisons, and no result claims equivalent input information.

4. Methods

4.1. Notation

For T2, let $Y \in \{N, D\}$ denote the ground truth and $\widehat{Y} \in \{N, D, A\}$ the output (normal, defective, abstain). Define $\beta = P(\widehat{Y} = N \mid Y = D)$, the rate of defects explicitly predicted normal, and $\alpha = P(\widehat{Y} = D \mid Y = N)$, the false-alarm rate. Define $\gamma_D = P(\widehat{Y} = A \mid Y = D)$ and $\gamma_N = P(\widehat{Y} = A \mid Y = N)$. Every denominator contains all items of the ground-truth class, including abstentions; none is conditioned on the model answering. Table 4 gives the full probability accounting. The defect non-detection rate for the automated stage is $\beta + \gamma_D = 1 - \text{TPR}$, whereas only the explicit-normal event β receives the missed-defect cost in the manual-review model below. Thus, $\text{TPR} = 1 - \beta - \gamma_D$ and $\text{TNR} = 1 - \alpha - \gamma_N$.

$$\text{BalAcc} = \frac{1}{2}(\text{TPR} + \text{TNR}) = 1 - \frac{1}{2}(\alpha + \beta + \gamma_D + \gamma_N). \tag{1}$$

Equation (1) is an identity rather than an approximation. Section 5.1 uses it to cross-validate the measured abstention rates.

Table 4. Class-conditional decision accounting. Each row sums to one; all rates use the complete ground-truth row as denominator.

Ground Truth	Output N	Output D	Output A
Normal N	$1 - \alpha - \gamma_N$	α	γ_N
Defective D	β	$1 - \beta - \gamma_D$	γ_D

4.2. The Balanced-Accuracy Blind Spot

Proposition 1. *For $1/2 \le b < 1$, fix $\text{BalAcc} = b$ and set $s = 2(1 - b)$. By Equation (1), the compatible error profiles form the set*

$$\Delta_b = \{(\alpha, \beta, \gamma_D, \gamma_N) \in [0,1]^4 : \alpha + \beta + \gamma_D + \gamma_N = s,\ \alpha + \gamma_N \le 1,\ \beta + \gamma_D \le 1\},$$

a three-dimensional polytope. Consequently, for any $1/2 \le b < 1$ there exist models of identical balanced accuracy whose errors are dominated respectively by misses ($\beta \to s$), by hallucinations ($\alpha \to s$), and by abstentions ($\gamma_D + \gamma_N \to s$), and no function of b alone distinguishes them.

Proof. Equation (1) imposes one linear equality on four non-negative quantities subject to two row-stochasticity constraints. The three named cases are the coordinate extrema of Δ_b and are feasible whenever $s \le 1$. □

This elementary identity explains why balanced accuracy alone cannot identify the error composition [13,14]. The observed models illustrate differing error profiles, rather than literally attaining every extreme point of the feasible set.

4.3. Cost-Sensitive Risk with Abstention

Let C_m be the cost of a defect explicitly predicted normal, $C_h > 0$ the false-alarm cost, and C_a the cost of routing an abstention to review. Dividing by C_h, write $\rho = C_m/C_h$ and $\kappa = C_a/C_h$. With class priors satisfying $0 < \pi_D < 1$, the normalized expected cost per item is

$$R_i(\rho, \kappa) = \pi_D(\rho\beta_i + \kappa\gamma_{D,i}) + \pi_N(\alpha_i + \kappa\gamma_{N,i}). \tag{2}$$

Equation (2) aggregates these terms with the class priors π_D and π_N, and the case $\kappa = 0$ makes review cost-free in the accounting model. The illustrated range is $0 \le \kappa \le 1$ with $\rho \ge 1$; this is an explicit cost scenario, not a field estimate. Abstentions are assumed to be routed for review. Human review error, queueing delay, and downstream consequences are not fully modeled. For $\beta_A \neq \beta_B$, equating the risks gives

$$\rho^*(\kappa) = \frac{\pi_N(\alpha_B - \alpha_A) + \kappa[\pi_D(\gamma_{D,B} - \gamma_{D,A}) + \pi_N(\gamma_{N,B} - \gamma_{N,A})]}{\pi_D(\beta_A - \beta_B)}. \tag{3}$$

Because the compared models are discrete classifiers with token outputs and no tunable threshold, a full cost curve in the sense of Drummond and Holte [14] is unavailable without a confidence signal. This study does not extract such a signal. The abstention term is used as an accounting device rather than as a new theory of selective classification.

4.4. Prevalence Sensitivity

When $\kappa = 0$, the crossover in Equation (3) is proportional to π_N/π_D. For nonzero review cost, additional class-conditional abstention terms remain and this simple proportionality is not general. The benchmark prevalence $\pi_0 = 0.379$ is a construction property, not a field estimate. At zero review cost only, the crossover at another assumed prevalence π is

$$\rho^*(\pi) = \rho_0^* \cdot \frac{\pi_0}{1 - \pi_0} \cdot \frac{1 - \pi}{\pi}, \quad \kappa = 0. \tag{4}$$

Equation (4) applies to the zero-cost abstention case; the general surface is given by Equation (3). Any crossover ratio derived from a constructed benchmark therefore carries an implicit prevalence assumption that must be reported with it.

For practical use, first fix the unit—crop, asset, image or tower—and estimate prevalence and class-conditional errors on that unit. Use a representative target-line sample and independent verification of both alarm and non-alarm outcomes; precision among enriched alarm samples is not prevalence. Stratified samples require population weights. False-alarm cost includes unnecessary verification/dispatch; abstention cost includes review and queueing; missed-defect cost is a scenario-dependent expected consequence. Use justified ranges when point estimates are unavailable. Prior re-basing additionally assumes the class-conditional error rates remain applicable: changing prevalence alone does not correct distribution shift, which requires local validation.

4.5. Parity Audit

An audit verdict records comparability, not training–test independence or a causal effect attributable to model family. Every cross-family comparison carries at least the side-information and resource caveats. Resolution bracketing is documented rather than described as exact per-item matching.

4.6. Models, Adaptation, and Statistical Testing

The tested systems are language-grounded generative VLMs and the vision-only ResNet-50 and Swin-T backbones [31,32]. Including Swin-T prevents a simple identification of all vision-only models with convolutional architectures, but does not isolate or exclude architecture-specific effects.

Seven models were evaluated zero-shot. The open-weight models were InternVL3.5-2B and -8B, Qwen3-VL-2B and -8B, and Qwen2.5-VL-7B; the hosted models were Qwen3.8-27B and Qwen3.5-122B-A10B-FP8 with `enable_thinking = false`. Adaptation used LoRA [27] rank 32, $\alpha = 64$, dropout 0.05, a frozen vision encoder, learning rate 10^{-4} with cosine schedule and 3% warmup, three epochs unless stated otherwise, effective batch 16, and bf16 precision; this trained about 0.25% to 0.41% of parameters. The vision-only baselines were ImageNet-initialized and trained for 10 epochs with the settings in Table A3, at 224 × 224 unless stated. Table A4 records the identifier, endpoint field, and access date of each hosted service.

Two scoring granularities are used on T3. The seven-way score is macro-averaged recall over the seven ground-truth classes after mapping each model's answer-option output back to those labels. The three-way score is macro-averaged over the coarser option-level partition. The two average over different partitions and weights, so neither is constrained to exceed the other and no monotonic relation should be inferred.

Input resolution is audited using recorded preprocessing budgets. InternVL tiles at 448 px and records a mean 2.94 patches and 752.9 visual tokens on original T3; the aggregate area is approximately a 768 px square on average. The vision-only 672 and 896 px settings bracket that mean. Qwen records 329.5 mean visual tokens at its native setting. These token counts are architecture-dependent and do not measure equal information or compute. Instrumented pixel-budget passes log input and prediction hashes; their native scores differ from the main pass by 0.31 points for Qwen and 0.56 for InternVL. Within-pass contrasts and explicitly identified earlier per-class passes must therefore remain distinct.

Quantization used weight-only NF4 and INT8 with bf16 activations. Zero observed events are reported with an approximate one-sided rule-of-three bound $3/n$, explicitly conditional on independent trials. Correlated crops do not automatically supply n independent field observations, so this bound is not an operational safety guarantee.

Paired item-level correctness is assessed using exact McNemar tests [33,34]; these test ordinary correctness, not macro recall or training-seed equivalence. The reported macro-recall tests swap paired predictions by cluster using 10,000 sign-flip draws; the cluster is tower for the tower/tower14 and original-T3 analyses, and parent group for the image analysis. A follow-up rerun of the seven retained contrasts from copied prediction files, using 10,000 randomizations and randomization seed 42, reproduced the reported macro differences and p-values to displayed precision. The seven specific contrasts, cluster keys, and prediction sources are listed in Appendix A (Table A5). This is a computational reproduction, not a new experiment or an independent statistical method. Model exchangeability and the adequacy of the selected clustering are assumptions, not consequences of the permutation procedure. Multiplicity families are stated beside each reported Holm-adjusted p-value. Small cluster counts limit inference. The shifted-null endpoints in the earlier draft are omitted because a confidence-interval construction and coverage justification were not established; they are not used to claim significance, precision or equivalence. Retained p-values describe the reported runs and do not quantify training randomness or cure source-image overlap.

The reported hardware is one NVIDIA RTX PRO 6000 Blackwell (96 GB). The itemized, non-overlapping numerical entries in Table A6 sum to 196.5 GPU-h. This is a subtotal of the listed work, rather than a comprehensive inference or wall-clock measurement. Table A3

gives the run configuration, and Table A9 records the run and comparison unit behind each core result.

5. Results

Percentages are normally shown to two decimal places and unit-fraction scores to three or four. Display rounding can affect a difference by 0.01 points; differences should be calculated from the unsimplified result. A single precision estimate based only on item count is not assumed for clustered or rare-class macro recall.

5.1. Zero-Shot Error Profiles

Zero-shot T2 balanced accuracy spans 42.0% to 74.3% across seven models. Format failure is 0.0%, so low scores reflect model behavior rather than parsing. Refusal is a legal output and is scored separately.

Error regimes invert with scale: 2B models approximate an always-defect policy (α 60–70%, β near zero), whereas 8B–27B models show the reverse (α below 5%, β 40–65%). The 7B model lies between them. Figure 1 plots the (α, β) plane.

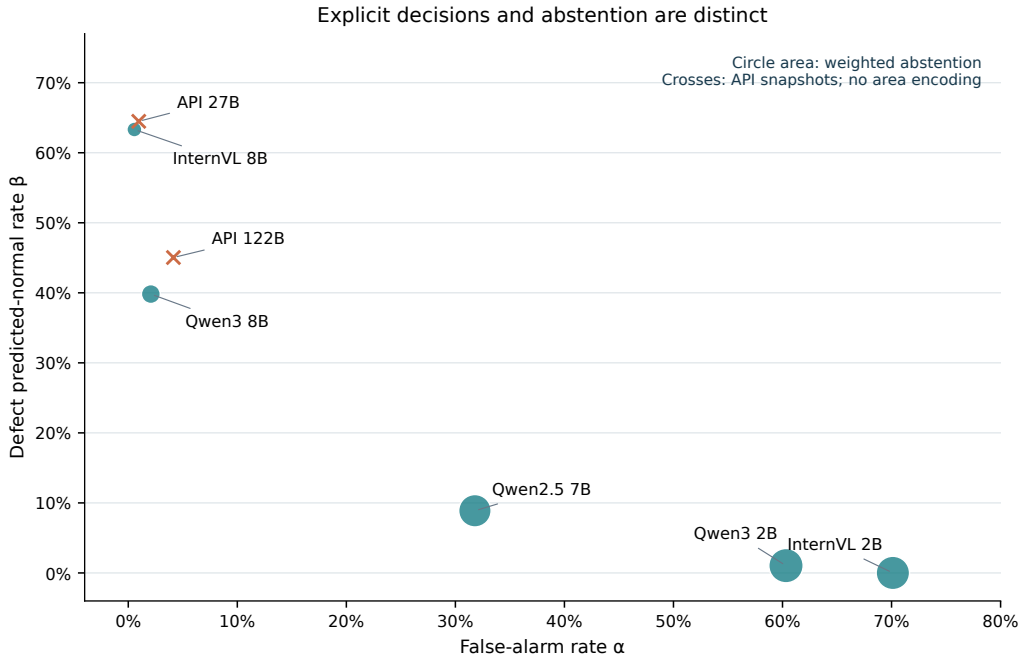


Figure 1. Original T2 explicit error rates. Marker area encodes prevalence-weighted abstention for the five local models; API snapshots use crosses because class-conditional abstention cannot be recovered separately. All points are reported diagnostics, not independent field validation.

The abstention columns of Table 5 independently check Equation (1). For every locally run model, measured $\gamma_D + \gamma_N$ reproduces the value implied by BalAcc, α, and β to within rounding: 0.4592, 0.4928, 0.4368, 0.0303, and 0.0961 for InternVL3.5-2B, Qwen3-VL-2B, Qwen2.5-VL-7B, InternVL3.5-8B, and Qwen3-VL-8B, respectively.

InternVL3.5-2B has no observed explicit-normal predictions among defective items, but abstains on 20.1% of them. Its automated defect non-detection rate is therefore 20.1%, not zero. This is an example of why β, γ_D, and their sum must be distinguished. The reported adapted runs show negligible abstention; this does not establish calibrated uncertainty.

Table 5. Zero-shot T2 performance (n = 2021; 1255 normal, 766 defective). Abstention rates are measured from stored predictions for the five locally run models and derived from Equation (1) for the two API-served models.

Model	Parameters	BalAcc	β (Miss)	α (Hallucination)	γ_D	γ_N
InternVL3.5-2B	2B	0.420	0.00% ‡	70.12%	20.10%	25.82%
Qwen3-VL-2B	2B	0.447	1.04%	60.32%	22.19%	27.09%
Qwen2.5-VL-7B	7B	0.578	8.88%	31.79%	20.89%	22.79%
Qwen3.8-27B	27B (API)	0.665	64.49%	0.96%	≤1.55% †	≤1.55% †
InternVL3.5-8B	8B	0.666	63.32%	0.56%	1.83%	1.20%
Qwen3.5-122B	122B (API)	0.710	45.04%	4.14%	≤8.76% †	≤8.76% †
Qwen3-VL-8B	8B	0.743	39.82%	2.07%	4.83%	4.78%

‡ Zero explicit-normal decisions among 766 defective items; the approximate one-sided 95% upper bound under independent trials is 0.39%. This is not a cluster-adjusted field bound. † Derived bound: the API result records do not store per-class predictions, so Equation (1) identifies only the sum $\gamma_D + \gamma_N$.

5.2. Cost-Sensitive Ranking and Its Assumptions

Applying Algorithm 2 to InternVL3.5-2B against InternVL3.5-8B yields the crossover surface of Table 6.

Algorithm 2: Abstention-aware cost crossover

1. Fix the evaluation unit, target prevalence, and plausible (ρ, κ) domain.
2. If $\beta_A = \beta_B$, compare the remaining risk terms; if they also agree, report a tie over the domain.
3. Otherwise compute $\rho^*(\kappa)$ from Equation (3).
4. If the crossover is outside the declared domain, evaluate both risks at an interior point and report the preferred model on that domain; do not present an out-of-domain threshold as an operational recommendation.
5. If it is inside, evaluate the risks on both sides, then repeat over the stated prevalence and review-cost scenarios.

Table 6. Crossover cost ratio $\rho^*(\kappa)$ over defect prevalence π_D and abstention cost κ, computed from the measured abstention rates of Table 5.

π_D	$\kappa = 0$	$\kappa = 0.25$	$\kappa = 0.5$	$\kappa = 1$
0.379 (this benchmark)	1.8	2.0	2.3	2.7
0.10	9.9	10.8	11.8	13.7
0.05	20.9	22.8	24.7	28.5
0.01 (low-prevalence sensitivity point)	108.8	118.5	128.1	147.5

In the displayed scenarios, pricing review at the false-alarm cost moves the crossover from approximately 1.8 to 2.7; it does not eliminate it. At the illustrative 1% prevalence, the corresponding ratios are approximately 109 to 148. Figure 2 plots this crossover surface over the illustrated prevalence and review-cost scenarios. These are sensitivity calculations, not measured field prevalence or universal operational cost thresholds.

EAR = $(\beta + \epsilon)/(\alpha + \epsilon)$ summarizes the two explicit error rates but excludes abstention. Its magnitude is sensitive to the smoothing constant: InternVL8B changes from 6.94 to 96.09 across the displayed settings, about fourteenfold. The tabulated model ranking, including 27B versus 122B, does not reverse in these settings (Table A7). EAR is therefore descriptive and is not a basis for a regression or deployment recommendation.

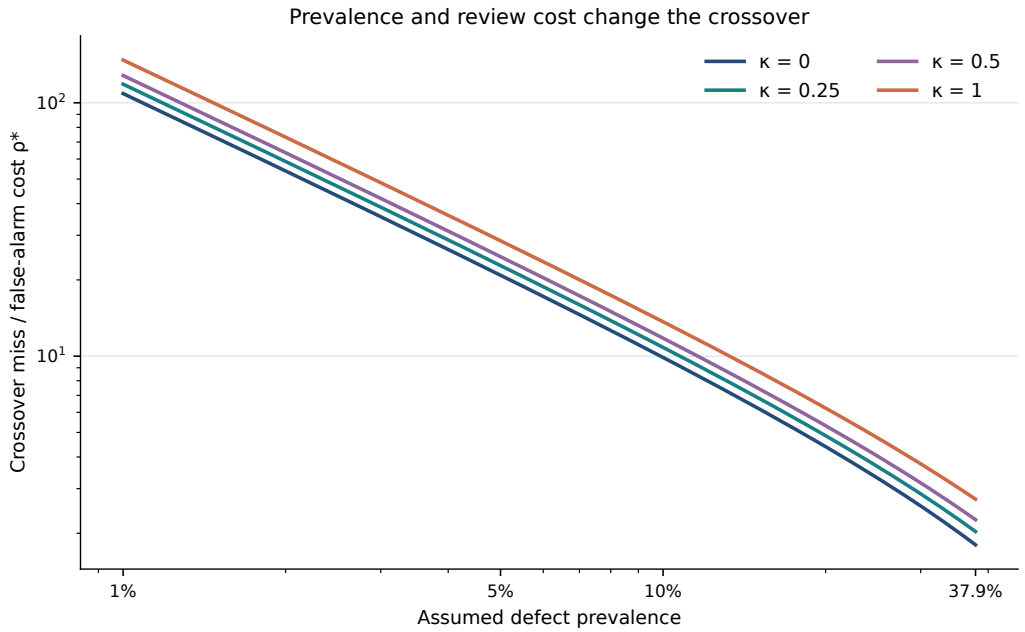


Figure 2. Cost crossover calculated from the reported InternVL3.5-2B/InternVL3.5-8B class-conditional rates across illustrative prevalence and review-cost scenarios. Neither axis represents measured target-line cost or prevalence. The asterisk in the vertical-axis label denotes the crossover ratio ρ^*, that is, the miss/false-alarm cost ratio at which the two models have equal expected cost (Equation (3)); it is not a significance marker.

5.3. Adaptation

Seven-way scores map each model's answer-option output back to the seven ground-truth labels. Because the three-way scores of Section 5.6 average over a different partition, they are not constrained to exceed these values; the 0.881 three-way and 0.891 seven-way scores for Qwen3-VL-8B are therefore not inconsistent. Macro recall and macro-F1 can also rank models differently because macro-F1 includes precision. Table 7 gives the adapted-model scores on the full T2 and T3 test files.

Table 7. LoRA-adapted models on the full T2 ($n = 2021$) and T3 ($n = 7284$) test sets, seed 42.

Model	T2 BalAcc	T2 β	T2 α	T2 Abstention	T3 BalAcc (K = 7)	T3 Macro-F1
InternVL3.5-2B + LoRA	0.949	9.79%	0.40%	0.000	0.742	0.772
InternVL3.5-8B + LoRA	0.959	7.57%	0.72%	0.000	0.779	0.795
Qwen3-VL-8B + LoRA	0.971	5.09%	0.64%	0.000	0.891	0.919

The adapted-model table alone has the same three-model ordering on the binary and seven-class tasks. Table 8 places the adapted models against the vision-only backbones on the identical T2 test file. The cross-family label-space comparison in Table 9 motivates the resolution and prompt diagnostics because that broader ordering differs across tasks.

Holm adjustment leaves the displayed McNemar verdicts in Table A11 unchanged. These paired ordinary-accuracy comparisons are not equivalence tests for balanced accuracy. Adaptation changes the error composition: the two 8B models reduce their high explicit-normal defect rates, whereas the 2B InternVL run exchanges its near-zero explicit-normal rate and heavy abstention/false alarms for a different adapted error profile.

The seed-41 Qwen evaluation has no format failures. Its option-level T3 balanced accuracy is 0.9379, and its seven-label macro recall is 0.9010. Replication changes Qwen T2 by 0.86 points and seven-way macro recall by 1.02 points, so the global result is not a single-seed

estimate. Two seeds remain too few for a precise variance estimate, and the InternVL seven-way seed difference of 3.50 points is more than three times the Qwen difference. Figure 3 shows both measured seeds for these two 8B adaptations, and Table A10 tabulates them.

Table 8. Cross-family comparison at full training data on the identical T2 test file ($n = 2021$). Vision-only scores are three-seed means with SD; the paired McNemar test uses the stored seed-42 vision-only prediction file and the seed-42 Qwen prediction file, not the three-seed mean. Algorithm 3: PASS at lines 1 to 4, PASS with caveat at lines 5 and 6.

Model	Family	BalAcc	Seeds	Discordant Pairs vs. Qwen3-VL-8B	p (Exact)
Swin-T	vision-only (transformer)	97.17 ± 0.23	3	24, 24	1.000
Qwen3-VL-8B + LoRA	VLM	97.14	1	—	—
ResNet-50	vision-only (convolutional)	96.54 ± 0.10	3	22, 30	0.332
InternVL3.5-8B + LoRA	VLM	95.86	1	—	—
InternVL3.5-2B + LoRA	VLM	94.91	1	—	—

Algorithm 3: Six-axis reporting and parity audit

Input: result records with checkpoint/training provenance, test file, evaluated IDs, ground-truth labels, seeds, preprocessing/input-budget record, and side information.

1. Check the task and test-file identity; different tasks/files cannot establish a paired contrast.
2. Check item counts and evaluated-ID sets; equal counts alone do not establish item parity.
3. Check the ground-truth class set, label mapping, and metric definition.
4. Record every training seed and checkpoint source. Unequal or limited replication is a caveat; descriptive point differences remain reportable but do not establish training stability.
5. Compare resolution, preprocessing, source-pixel and visual-token records. Distinguish exact matches from bracketing settings and unmatched compute; an area match is not an information match.
6. Record asset identity, definitions, candidate restrictions and output constraints. Unequal side information permits a system-level comparison only.
7. Return FAIL for a task/item/metric mismatch; otherwise return PASS with the accumulated caveats, or PASS only if no mismatch was found on the declared axes.

Table 9. Balanced accuracy (%) across label-space granularity. All rows are scored on the stated file with the stated ground-truth class set.

K	Test File, n	ResNet-50 (224 px)	Swin-T (224 px)	InternVL3.5-2B	InternVL3.5-8B	Qwen3-VL-8B
2	T2, 2021	96.54	97.17	94.91	95.86	97.14
3	T3, 7284	92.26	93.09	85.77	89.17	88.10
7	T3, 7284	57.37	54.23	74.16	77.90	89.08

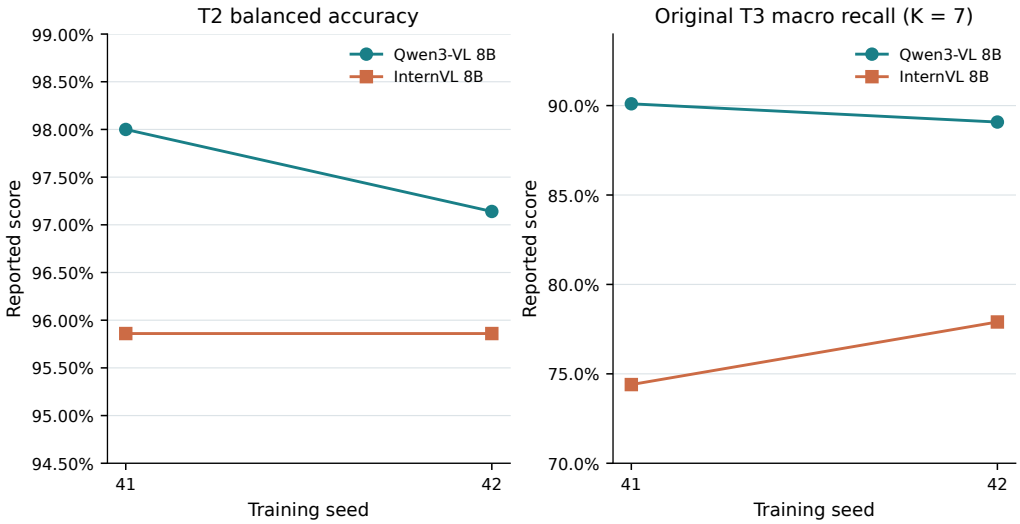


Figure 3. Individual global seeds 41 and 42 for the two 8B adaptations: T2 balanced accuracy and original T3 seven-class macro recall. Points are training runs, not independent datasets.

5.4. Multi-Task Mixing and Epoch Budget

The mixed corpus outperforms both specialists on T2, by 2.4 points over the T2-only specialist and 5.8 points over the T3-only specialist. This is consistent with cross-task regularization [29,30], although a training-volume contribution cannot be separated without a volume-matched control.

Extending training from three to ten epochs leaves T2 essentially unchanged while evaluation loss rises from 0.035 to 0.099, and the five-class T3 diagnostic falls by 10.1 points (Table A8). The pattern indicates over-fitting in the thinner, more class-diverse task. We therefore use three epochs and recommend governing early stopping by the hardest task in a multi-task mixture rather than the headline task.

Direct measurement of S200 gives 126 normal and 74 defective items at T2, and 177 `good`, 12 `rust`, 7 `missing-cap`, 3 `corrosão`, and 1 `nest` item at T3, with `peeling-paint` and `torned-up` absent. The S200 T3 results are therefore a five-class diagnostic and are not comparable to any seven-class value elsewhere in this paper.

5.5. Partition Parity at Binary Screening

The earlier comparison mixed a 7851-item seven-class file with the 2021-item binary T2 file and used a superseded vision-only training configuration. It is retained as an audit failure, not a clean intervention on partition alone. Equality of class distributions cannot be asserted across these different tasks without an explicit binary mapping. The revised T2 comparison below uses the identical item file, with training provenance and replication caveats stated separately.

The vision-only budget sweep in Table A13 is supplementary. It reports three-seed means on T2 but does not establish a cross-family few-shot ordering: the VLM budget runs used S200 and had different replication.

The displayed Swin-T three-seed mean differs from the Qwen seed-42 score by 0.03 percentage points; this mixes a mean with an individual run and is only descriptive. The separate seed-42 paired predictions have 24 versus 24 discordant items, giving exact McNemar $p = 1.000$. That test concerns ordinary item correctness, not equivalence of balanced accuracy, training variability, or the two model families.

5.6. Label-Space Parity: Seven-Way Defect Typing

At 224 px, the seven-way row reverses the binary row. The 224 px gaps are 20.53 points for ResNet-50 and 23.67 points for Swin-T against InternVL3.5-8B, and 31.71 and 34.85 points against Qwen3-VL-8B. Section 5.7 shows that pixel budget accounts for most of the InternVL gap but not the difference between the two VLMs. The three-way row averages over the option-level partition and is not directly comparable to the seven-way row.

The low-resolution error pattern motivates examining spatial detail and prompt information separately. It does not identify either as a sufficient cause of the observed cross-family difference.

5.7. Separating Pixel Budget, Language Grounding, Scale, and Model Identity

5.7.1. Resolution and Pixel-Budget Control

Raising the vision-only inputs from 224 to 448 px increases their reported seven-class macro recall. At 672 px, ResNet-50 scores 78.47% and Swin-T 73.20%, compared with the main InternVL score of 77.90%. These are bracketing input-area diagnostics, not exact budget matches, and the 448 px setting was selected from a test-set sweep.

The highest-scoring 448 px test-sweep setting has a selection advantage. Results at 672 and 896 px are also reported to bracket the mean InternVL input area; no interpolation is presented as a measured matched-budget score.

Table A14 reports a paired significance test for the Qwen comparison. It illustrates the divergence between raw accuracy and macro recall rather than establishing the gap.

The pattern is not monotonic in the macro-recall gap. The 672 px row has a large macro-recall gap but the weakest raw-accuracy signal ($p = 0.923$). Because good constitutes 89.5% of the T3 test set, item-level pairing is dominated by a near-ceiling class and can miss rare-class macro-recall differences. A tower-clustered macro-recall test therefore replaces these raw-accuracy tests as the inferential summary.

The two-decimal ResNet-50 896 px and Swin-T 672 px entries in Table 10 follow the earlier V9.2 resolution-summary reporting pass. By contrast, all macro gaps in Table 11 are computed from the unsimplified metrics stored in the current round2 statistical-analysis pass. The exact current-pass metrics are 0.7614141 and 0.7317078, corresponding to 76.14% and 73.17%, respectively; they do not correspond to the tabulated 76.10% and 73.20%. For example, Qwen minus ResNet-50 at 896 px is $0.8907536 - 0.7614141 = 0.1293395$, reported as +12.93 points, whereas the earlier-pass two-decimal scores subtract to 12.98. Qwen minus Swin-T at 672 px is $0.8907536 - 0.7317078 = 0.1590458$, reported as +15.90 points, whereas the earlier-pass two-decimal scores subtract to 15.88. Because the exact metrics and the tabulated two-decimal scores come from different evaluation passes, these differences are not rounding artifacts.

Table 10. Vision-only balanced accuracy (%) at K = 7 across input resolution (T3 test, $n = 7284$), with paired McNemar tests on raw per-item correctness against InternVL3.5-8B at native tiling (77.90%).

Vision-Only Configuration	BalAcc	Discordant Pairs (b, c)	p (Exact)	Verdict on Raw Accuracy
ResNet-50, 224 px	57.37	—	—	—
ResNet-50, 448 px	81.08	75, 37	0.0004	vision-only ahead
ResNet-50, 672 px	78.47	70, 54	0.178	not detected
ResNet-50, 896 px	76.10	74, 98	0.079	not detected
Swin-T, 224 px	54.23	—	—	—
Swin-T, 448 px	75.11	66, 29	0.0002	vision-only ahead
Swin-T, 672 px	73.20	67, 53	0.235	not detected

Table 11. Reported tower-clustered macro-recall contrasts on original T3. Holm family: the five listed comparisons. No confidence interval is claimed; original-split source dependence remains.

Comparison	Macro Gap (pt)	Raw p	Holm p	Verdict
Qwen–ResNet 448	+7.99	0.287	0.834	Not detected
Qwen–ResNet 672	+10.61	0.352	0.834	Not detected
Qwen–ResNet 896	+12.93	0.278	0.834	Not detected
Qwen–Swin 448	+13.95	0.101	0.405	Not detected
Qwen–Swin 672	+15.90	0.0079	0.0395	Detected in this family

The Qwen–ResNet-50 448 px effect is +7.99 points but is not statistically established after tower clustering. Only the contrast against Swin-T at 672 px survives Holm correction. Table 12 gives the per-class breakdown.

Table 12. Qwen3-VL-8B per-class recall on the original T3 test set, and contribution to the macro-recall gap against ResNet-50 at 448 px.

Class	Test Support	Recall	Contribution vs. ResNet-50 448 px
corrosão	184	95.11%	+0.47 pt
good	6518	99.62%	−0.03 pt
missing-cap	136	91.91%	+0.32 pt
nest	103	98.06%	+0.28 pt
peeling-paint	4	75.00%	+10.71 pt
rust	334	83.83%	−0.90 pt
torned-up	5	80.00%	−2.86 pt
Macro-average	7284	89.08%	+7.99 pt

The tabulated Qwen class recalls reproduce 89.08% and the contribution column sums to +7.99 points. The four-item `peeling-paint` and five-item `torned-up` classes contribute +7.85 points net. These are arithmetic checks of the reported decomposition, not independent validation of the stored prediction source or rare-class generalization.

To measure input budget directly, the processor `max_pixels` value was constrained, and the actual grid, visual-token count, input-identifier hash, pixel-tensor hash, prediction hash, and elapsed time were logged for every item. Table A15 reports the per-budget scores and token statistics. All six prediction files differ by SHA-256. Relative to native resolution, the lower budgets change 4472, 3727, 2839, 2025, and 1063 of 7284 items.

Reducing the budget from native to 65,536 pixels lowers Qwen3-VL-8B macro recall by 10.78 points. The two largest increments occur from 65,536 to 131,072 pixels, worth 4.01 points, and from 200,704 to 401,408 pixels, worth 5.66 points. Above 401,408 pixels the change is 0.31 points despite an increase in mean visual tokens from 189.3 to 329.5; the median remains 144 because many crops already fall below the higher budgets.

InternVL3.5-8B was evaluated under the same source-pixel budgets to test whether pixel budget explains the VLM difference. This is a source-pixel control rather than a visual-token match, because InternVL's dynamic tiling is aspect-ratio driven and its patch and token counts do not track source area in the same way. The control is effective: the 65,536-pixel setting changes 4699 InternVL pixel tensors relative to native resolution. Table A16 reports the paired scores at each cap.

Across the reported source-pixel caps, Qwen macro recall spans 10.78 points and InternVL 6.16 points. Qwen remains ahead by the tabulated 7.42–13.61 points. At native settings Qwen uses fewer counted visual tokens, so a simple explanation based on more counted Qwen tokens is unsupported. This does not exclude tokenization, encoder, pre-

training, aspect-ratio policy, or compute effects. Figure 4 keeps the source-pixel and token axes distinct.

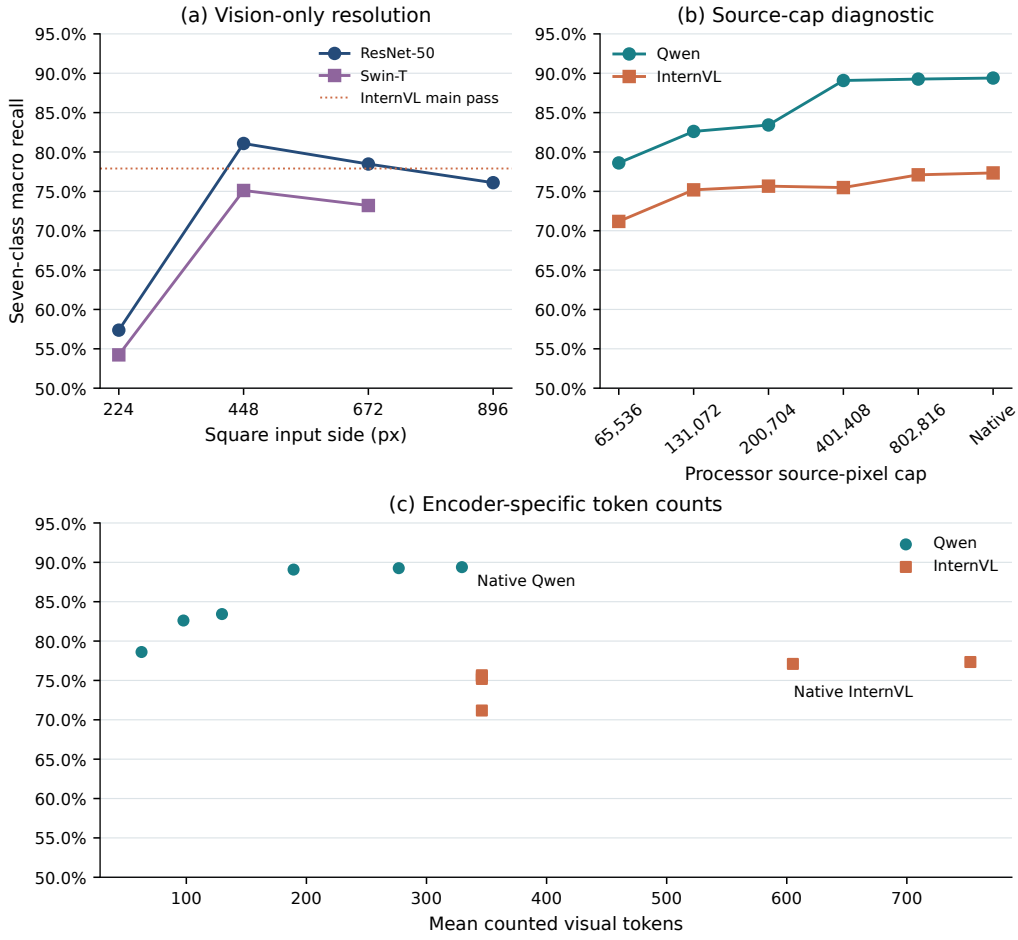


Figure 4. Budget diagnostics with distinct axes: vision-only resolution; processor source-pixel cap for both VLMs; and counted native/intermediate visual tokens. Native denotes the processor setting, not a measured common pixel area. Token counts are not comparable information or compute units across encoders.

A post-hoc token-count proxy diagnostic was added to distinguish approximate token-count similarity from a strict compute-controlled experiment (Table 13). The cross-family comparison uses Qwen at the 401,408-pixel cap, with a measured mean of 189.3 visual tokens, and Swin-T at 448 px, whose final-stage representation contains $14 \times 14 = 196$ tokens. The 6.7-token difference is a proxy, not an information or compute match. Qwen reaches 89.08% macro recall and Swin-T 75.11%, a direct gap of +13.97 points; this gap is computed from the two stored prediction summaries and is not the +13.95-point contrast in Table 11, which uses a different Qwen prediction file. No inferential p-value is claimed for this specific Qwen 401,408 comparison. For the two VLMs, Qwen at native resolution (329.5 mean tokens) most closely matches InternVL at the 65,536-pixel cap (346.1 mean tokens; 4.8% difference), where Qwen leads by 18.21 macro-recall points. The paired item-correctness counts are $b = 97$, $c = 36$, exact McNemar $p = 1.19 \times 10^{-7}$, but that test concerns raw correctness rather than the macro-recall difference. Qwen rows use `qwen_pixel_budget_v9_pred_401408.jsonl` and the native prediction file; InternVL rows use `internvl_pixel_budget_v9_pred_65536.jsonl`; the exact and ≤256 subsets use the 401,408 Qwen and InternVL prediction files. The exact 256-token subset has $K = 3$: `good` $= 132$, `rust` $= 28$, and `corrosão` $= 18$. The shared ≤256-token subset

has $K = 6$: `good` $= 4467$, `rust` $= 134$, `missing-cap` $= 11$, `nest` $= 12$, `corrosão` $= 29$, and `torned-up` $= 1$. These diagnostics remain post hoc and do not equate architecture, information, or compute.

The main-pass ResNet-50 448 px score exceeds the reported InternVL source-budget scores but is approximately eight points below the main Qwen score. The Qwen–ResNet macro-recall contrast has reported Holm $p = 0.834$ in its stated family and is strongly concentrated in rare classes; it is not treated as an established or general advantage.

A single-tile InternVL control reports 74.70% macro recall. For its listed discordant counts $b = 27, c = 18$, the two-sided exact McNemar value is $p = 0.232693$, reported as 0.233. This test is on item correctness, not on the macro-recall difference. A tile restriction and a source-pixel cap intervene at different preprocessing stages and are not equivalent controls.

Table 13. Post-hoc token-count proxy diagnostics. These are descriptive reanalyses of stored predictions, not new training runs, information matches, or strict compute-matched experiments.

Comparison	Matching Rule	*n*	Token Budget	Macro Recall/Gap
Qwen 401,408 vs. Swin-T 448	cross-family token-count proxy	7284	189.3 vs. 196	89.08% vs. 75.11%; +13.97 pt
Qwen native vs. InternVL 65,536	closest VLM mean-token proxy	7284	329.5 vs. 346.1	89.39% vs. 71.18%; +18.21 pt
Qwen vs. InternVL, exact 256	per-item exact token equality	178	256 vs. 256	$K = 3$: 98.81% vs. 90.74%; +8.07 pt
Qwen vs. InternVL, ≤256	shared upper token bound	4654	mean 104.0 vs. 256	$K = 6$: 92.56% vs. 72.96%; +19.60 pt

5.7.2. Per-Class Decomposition

The InternVL–Swin per-class values in Table 14 reproduce the +2.01-point macro difference. Every class has weight 1/7, irrespective of support. Removing `peeling-paint` changes the renormalized six-class difference to −14.32 points. The remaining classes contribute approximately −12.27 points only when kept at the original seven-class weights; that contribution sum is not itself a six-class macro difference.

Table 14. Per-class recall (%) at K = 7 on the T3 test set: InternVL3.5-8B + LoRA at native tiling against Swin-T at 448 px. The 75.89% Swin-T value comes from an earlier Unicode-normalized per-class pass and is not the 75.11% standard pass in Table 10; it is retained only for this separately sourced decomposition; no standard-pass significance result is transferred to it. The 448 px setting is the highest-scoring point of the vision-only sweep rather than the mean-area bracketing budget of approximately 768 px.

Class	Test Support	VLM Recall	Vision-Only Recall	Contribution to Macro Gap
corrosão	184	90.76	93.48	−0.39
good	6518	99.82	99.83	−0.00
missing-cap	136	69.85	88.24	−2.63
nest	103	98.06	98.06	0.00
peeling-paint	4	100.00	0.00	+14.29
rust	334	86.83	91.62	−0.68
torned-up	5	0.00	60.00	−8.57
Macro-average	7284	77.90	75.89	+2.01

This InternVL–Swin decomposition uses the explicitly identified earlier Swin per-class pass, not the standard-pass score of 75.11%. It is therefore kept separate from the standard resolution sweep and its tests. The different score is not reconciled by assuming all passes are interchangeable. The ResNet comparison and Qwen decomposition use their separately stated sources.

What carries over is the mechanism, not the exact residual. In this comparison, the absolute contributions of `peeling-paint` and `torned-up` sum to 22.9 percentage points, but their net contribution is +5.72 points because they act in opposite directions. Excluding

those classes changes the sign and magnitude of the model difference. The magnitude is configuration-dependent, so this is not a universal 22.9-point effect; it is evidence that classes with four and five test items can dominate an unweighted seven-class average. Figure 5 shows the split-level version of the same decomposition.

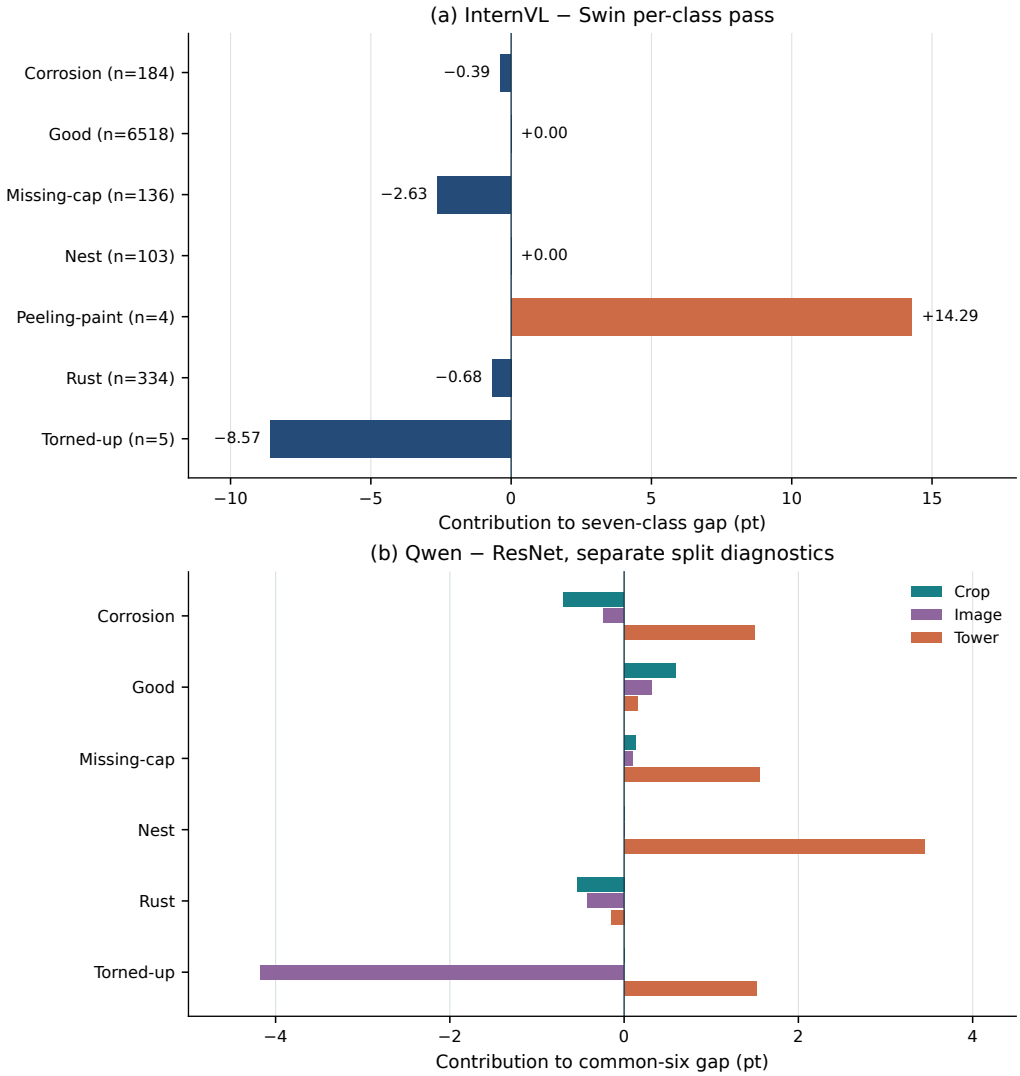


Figure 5. Point contributions to macro-recall differences. (**a**): InternVL minus the separately sourced Swin per-class pass, with original class support. (**b**): Qwen minus ResNet after split-specific retraining over six common classes. No shifted-null confidence intervals are shown.

5.7.3. Prompt Ablation

Removing definitional side information reduces reported macro recall to approximately 14.3%. The controlled P1-trained arm also has a low score, so an inference-only prompt mismatch is not a sufficient explanation in these configurations. An all-good strategy has seven-class macro recall $100/7 = 14.29\%$, but ordinary accuracy $6518/7284 = 89.48\%$. The observed rounded baseline is close to that strategy; identical individual predictions are not inferred from one aggregate score. Restoring P3 at inference gives the measured score contrasts shown in Tables A17 and A18; they are not independent causal channel estimates.

Both tested VLM adaptations depend strongly on the supplied definitions in these prompting configurations. This does not prove that language grounding is universally necessary or identify the mechanism of the remaining inter-model difference. The prompt is part of the system being evaluated.

The controlled P1 baselines are 14.29% for both reported adaptations. Restoring P3 at inference yields contrasts of 54.97 points for InternVL and 70.07 for Qwen, measured from those baselines. This describes the tested prompt transitions, rather than proving a model-intrinsic language-grounding advantage. Figure 6 plots only combinations that were reported; the P3-to-P1 arm is absent for Qwen.

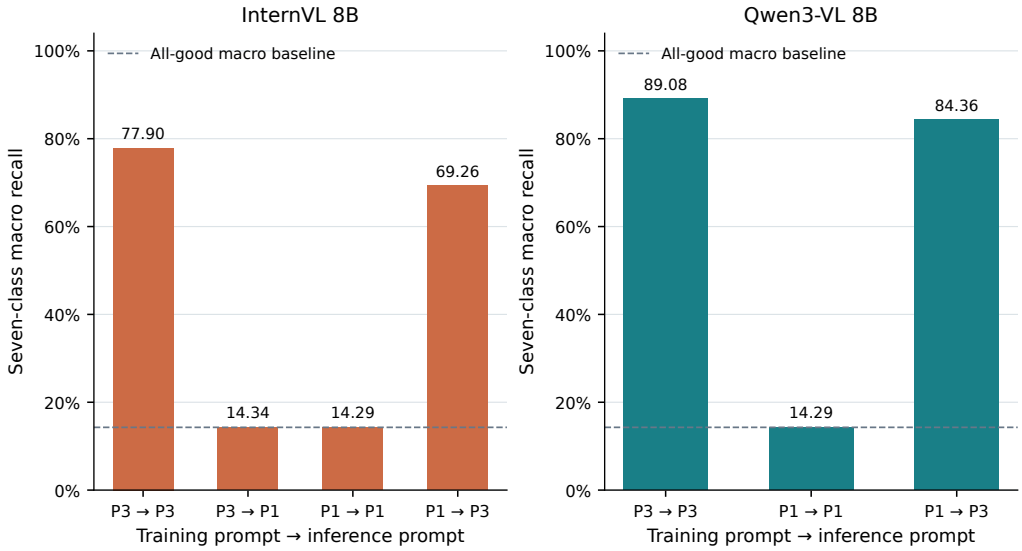


Figure 6. Measured prompt combinations only. The all-good seven-class macro-recall baseline is 14.29%; its ordinary sample accuracy would be 89.48%. The P3-to-P1 arm was reported only for InternVL, so no corresponding Qwen bar is created.

5.7.4. Summary of the Controlled Factors

The factor summary in Table A19 records associations and unresolved axes, not a complete causal decomposition. Identical nominal prompts, training hyperparameters or source-pixel caps do not imply identical information or compute. Token-matched and resource-matched experiments would add evidence, but were not performed here.

5.8. Quantization and Deployment Envelope

No precision differs from bf16 by more than 0.16 points on either test set, all p-values exceed 0.6, and the three to five discordant pairs per comparison are consistent with sampling noise. These results do not support a material accuracy difference, but no formal equivalence margin was pre-specified, so the statement is descriptive rather than a powered equivalence claim. The near-identical T3 scores do not imply identical per-item predictions. FP8 quantization failed with a framework-specific meta-tensor error and is treated as a tooling limitation, not a property of FP8. These results cover InternVL3.5-8B only; Table A20 gives the per-precision values.

The two low-precision modes have different deployment consequences. NF4 reduces memory by 61% at a 3% throughput cost; INT8 reduces memory by 44% at a 24% throughput cost from dequantization overhead. On this single-GPU, 50-item measurement, NF4 is preferable on both measured axes, but this is not a general deployment recommendation. All five candidate backbones load and accept LoRA injection on one 96 GB card, with bf16 inference footprints of 4.0–4.4 GB at the 2B tier and 15.5–16.3 GB at the 7B–8B tier.

A separate vision-only benchmark measures trained ResNet-50 and Swin-T checkpoints at 448 and 672 px on 512 test images, with five repeats per setting, 200 single-image latency samples, and batch sizes from 1 to 64. Latency includes host-to-device copy and forward pass; throughput also includes host-to-device transfer. Batch size 32 gives the best mean

throughput for every configuration. Table A21 reports the measured peak memory, latency, and throughput for each setting.

At 448 px, ResNet-50 has the highest throughput and Swin-T the lowest memory. At 672 px, ResNet-50 peak memory rises to 22.03 GB, an activation-memory effect at large batch sizes, while Swin-T remains below 10 GB. These are workstation measurements with preloaded tensors; they exclude JPEG decoding, disk input, energy, thermal throttling, and embedded accelerators. The VLM and vision-only throughput figures use different harnesses and batch policies and are reported separately. Figure 7 shows the three deployment axes.

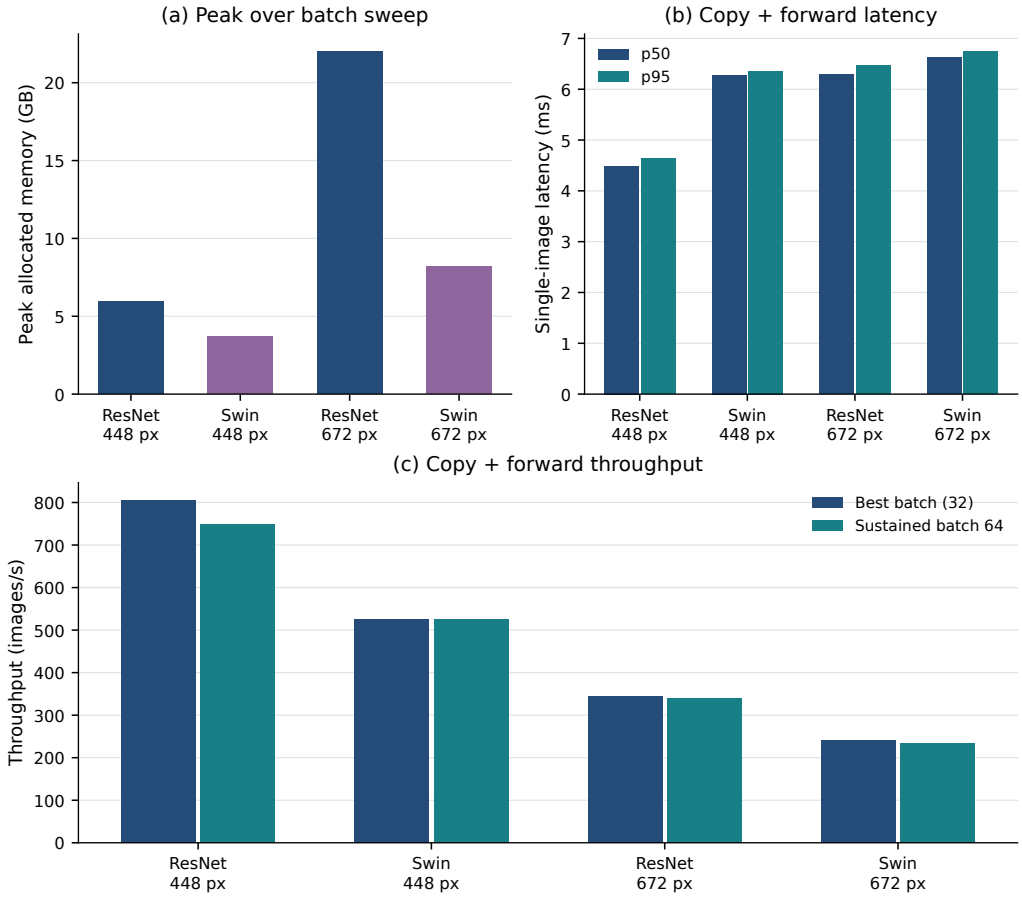


Figure 7. Workstation deployment measurements for vision-only checkpoints. Panels separate peak memory across the batch sweep, single-image p50/p95 latency, and batch throughput. Inputs were preloaded; decoding, detection, energy and embedded operation were not measured.

5.9. Split-Level Retraining and Tower Replication

Crop, image, and tower splits answer different questions and are not a monotonic generalization ladder. Crop-level retraining removes exact test-item reuse but can retain related crops from the same source image; image-level retraining removes same-image reuse but can retain the same tower; tower-level retraining removes same-tower reuse but does not test a new line, region, or acquisition campaign. The rows below are therefore separate diagnostics with different estimands. A checkpoint-provenance audit showed that the original crop-level and image-level rows reused checkpoints that had previously seen about 70% of the new test items. Qwen3-VL-8B and ResNet-50 at 448 px were therefore retrained on the crop-level and image-level splits; the tower-level models were already trained on the tower-disjoint split. A follow-up audit of the seven listed split-level Qwen runs confirmed the Qwen3-VL-8B base model, the stated split-specific training file and seed,

and `resume_from_checkpoint = None`; it does not add seeds or establish stability beyond those runs. The three regimes differ in residual dependence:

1. Crop-level split. Retraining removes exact test-item leakage, but same-image and same-tower overlap remains by design: 488 of 6672 test image paths also occur in the training file, and all 42 test towers occur in training.
2. Image-level split. No image path or `parent_id` crosses train and test, but all 41 test towers occur in training.
3. Tower-disjoint split. No image path, parent group, or tower crosses train and test.

All primary comparisons use the six classes common to the three test files: `corrosão`, `good`, `missing-cap`, `nest`, `rust`, and `torned-up`. Randomization is cluster-aware, with `_tower` as the cluster unit for the crop and tower splits and `parent_id` for the image split, and Holm correction is applied across the three split contrasts.

The full-label-space image result is also negative: Qwen reaches 0.9337 and ResNet-50 0.9715, a difference of −3.77 points. The single `peeling-paint` test item is predicted correctly by both models and does not create the rare-class artefact seen before retraining.

Per-class recall differences divided by six reproduce the split-level macro differences up to display rounding (Table A22). Figure 5 shows these point contributions without the earlier shifted-null endpoints. Different test populations and class supports make these separate split diagnostics rather than a repeated-measures generalization ladder.

After reported split-specific retraining, the crop and image point differences favor ResNet-50, while the nine-tower difference favors Qwen. In the three-split comparison family of Table 15, no Qwen advantage survives Holm correction. This is distinct from the supplementary nine-tower seed family, where particular contrasts are significant; those seed-dependent results do not establish a stable ordering and do not show equivalence.

Table 15. Reported common-six macro recall after split-specific retraining. Holm family: the three split contrasts; this family is separate from the seed sensitivity table. Shifted-null endpoints are omitted.

Split	Qwen	ResNet 448	Gap (pt)	Raw p	Holm p
Crop	0.9413	0.9464	−0.51	0.690	0.804
Image	0.9227	0.9667	−4.40	0.402	0.804
Tower	0.7647	0.6845	+8.02	0.0239	0.0717

Seed sensitivity on the nine-tower split. Adding Qwen seed 43 and ResNet-50 seeds 43 and 44 leaves Qwen ahead in all three contrasts, but the same-seed effect falls from +8.02 to +3.29 points and the seed-43 raw test does not detect a difference.

The third row pairs Qwen seed 43 with ResNet seed 44 and is a sensitivity contrast rather than a matched-seed replication. Across available runs, Qwen averages 0.7604 with a sample standard deviation of 0.0061 over two seeds, and ResNet-50 averages 0.7044 with 0.0194 over three seeds. The mean seed-matched difference is +5.66 points with a range of 4.73 points, so direction is consistent but magnitude and inferential status are seed-sensitive.

A larger tower replication. Because the nine-tower analysis used few test towers and an unbalanced seed design, a larger replication used 18 training towers, 10 validation towers, and 14 test towers. Its T3-only test set contains 5536 items, and both families were run with seeds 42, 43, and 44. This addresses the two principal limitations of Table A23, although it remains a new split rather than a pre-registered confirmatory trial.

Across the three seeds, Qwen averages 0.9085 common-six macro recall with a sample standard deviation of 0.0624, and ResNet-50 averages 0.9509 with 0.0053. The mean difference is −4.24 points with a range of 12.04 points. ResNet-50 leads in two of three seeds, and the Qwen seed-44 run falls to 0.8370. Under the full seven-class label space, Qwen averages

0.8740 against 0.8151 for ResNet-50, but the seed-44 contrast again reverses and all three Holm-adjusted p-values are 1.0000. The observed seed SDs are 6.2 points for Qwen and 0.5 points for ResNet-50 on this split; three runs do not estimate a family-wide variance. Table 16 lists the three matched seeds.

Table 16. Tower14 common-six macro recall: three matched seeds per tested model. Source T3-only test file contains 5536 items; effective support depends on the retained class set. Holm family: these three seed contrasts. No shifted-null confidence intervals are reported.

Seed	Qwen	ResNet 448	Gap (pt)	Raw p	Holm p
42	0.9369	0.9561	−1.91	0.0904	0.2712
43	0.9517	0.9455	+0.62	0.9189	1.0000
44	0.8370	0.9512	−11.42	0.6482	1.0000

The single-seed Swin control in Table A24 shows label-space dependence: Qwen is below Swin on the common-six score and above it on the full-seven score. These are point comparisons for one seed. The reported raw tests are not a three-seed family-level result. Figure 8 distinguishes the three-seed common-six results from this seed-42 label-space control.

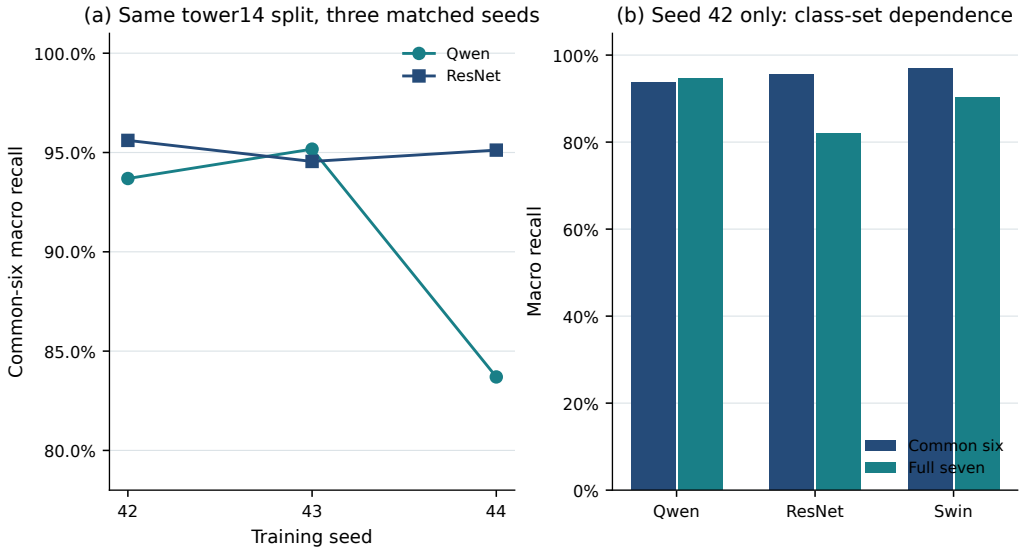


Figure 8. Tower14 results. (**a**): common-six macro recall for three matched Qwen/ResNet seeds. (**b**): seed-42 common-six and full-seven scores for Qwen, ResNet and single-seed Swin. No unreported per-seed full-seven values or statistical error bars are inferred.

Diagnosis of the seed-44 result. The drop is not a training or inference failure. Final training loss is 0.01711 and token accuracy approximately 0.998, close to the other seeds, and a rerun on a healthy accelerator reproduces the same two `torned-up` errors while changing common-six macro recall by 0.08 points. The binary prompt still classifies both images as defective; the error occurs only when choosing between `good` and `torned-up` in the multiclass prompt. Excluding the three-item `torned-up` class leaves seed 44 at 0.9377 five-class macro recall against 0.9243 and 0.9420 for seeds 42 and 43. The result is rare-class seed sensitivity amplified by macro averaging, not general model degradation. Table 17 applies this class-set arithmetic to both cross-family gaps.

Changing the retained true-label set changes the estimand and class weights. The re-averaged scores are not corrected versions of the original seven-class score. Removing only `peeling-paint` reverses the Qwen–ResNet sign; removing both rare classes leaves a small

positive Qwen difference. The InternVL–Swin sign reverses in both reduced sets. Table A25 additionally illustrates how wide four- and five-item binomial intervals are, conditional on independence; it is not a cluster-adjusted population inference. Table 18 summarizes the corresponding accuracy and deployment figures for the vision-only checkpoints.

Table 17. Arithmetic sensitivity computed from the reported seven-class per-class decomposition; not a new training experiment or inferential test. All items with retained ground-truth labels remain evaluated; predictions into excluded labels remain errors.

Class Set Retained	Qwen − ResNet-50 (448 px)	InternVL − Swin-T (448 px)
All seven classes	+7.99 pt	+2.01 pt
Exclude `peeling-paint`	−3.18 pt	−14.32 pt
Exclude `peeling-paint` and `torned-up`	+0.19 pt	−5.18 pt

Table 18. Recognition–cost summary for the vision-only checkpoints. Accuracy is seven-class macro recall on T3; memory, p50 latency, and best-batch throughput are the workstation measurements of Table A21. The accuracy and deployment measurements are not a jointly optimized deployment trial.

Configuration	K = 7 Recall	Peak Memory	p50 Latency	Best Throughput
ResNet-50, 448 px	81.08%	5.97 GB	4.49 ms	806.4 img/s
ResNet-50, 672 px	78.47%	22.03 GB	6.29 ms	344.8 img/s
Swin-T, 448 px	75.11%	3.70 GB	6.28 ms	525.7 img/s
Swin-T, 672 px	73.20%	8.19 GB	6.63 ms	241.6 img/s

The 672 px setting is closer to the mean InternVL input area than 448 px, but has lower reported macro recall and higher cost for these two backbones. Recognition-optimal, area-bracketing and deployment-efficient settings need not coincide. Pixel area rises by 2.25× from 448 to 672 px; this is not a measured FLOPs, energy or latency multiplier. VLM throughput is excluded from the joint table because its measurement program and batch policy differ.

6. Discussion

6.1. What the Controlled Comparisons Establish

The main result is sensitivity of the reported comparisons to evaluation choices. The original split is useful for paired diagnostics but has substantial source-photograph dependence. The reported tower checks reduce known tower reuse within this dataset; they do not supply new-line or external-dataset validation.

The binary score table combines a three-seed Swin mean with a Qwen individual run, whereas its McNemar result uses two seed-42 prediction vectors. These are different summaries. Neither a small point difference nor a non-significant item test establishes equivalence or a stable model-family ordering.

Input-resolution and prompt changes alter the observed fine-grained scores, but the interventions do not isolate architecture, pre-training, information or total compute. The mean-area bracketing and source-pixel-cap controls should therefore be read as budget diagnostics.

6.2. Rare Classes and Unresolved Model Differences

The class-set analysis shows why support must accompany macro recall: removing one test-only four-item class can change a model ranking, while removing both rare classes does not reverse every contrast. In the InternVL–Swin decomposition, the two absolute

contributions total 22.9 points and their net contribution is 5.72 points. The limited support prevents a precise statement of rare-class population performance.

Qwen remains above InternVL in the reported source-pixel-cap passes while using fewer counted tokens at native settings. This observation does not exclude tokenization or architecture effects. Strict information, token and compute matching, and external replication remain unresolved.

6.3. Practical Implications

Model selection should report the relevant class set, training support, input budget, side information and repeatability, rather than use a single aggregate score. The observed tower14 seed dispersion is a warning for these tested configurations, not a general rule that VLMs are unstable.

Cost rankings depend on assumed prevalence and review costs. The displayed low-prevalence values are scenarios; they are not target-line measurements. Prior adjustment assumes transferable class-conditional error rates, and a new operating environment requires local validation. The cost model also assumes a review route for abstentions and does not fully model delayed or erroneous review.

Deployment results characterize one workstation and fixed-crop classifiers. Pixel-area, visual-token and batch-throughput values are not interchangeable complexity measures. Actual UAV deployment requires detector/crop costs, input/output processing, energy and thermal measurements under a common program.

6.4. Limitations

Source validity is the principal constraint. Filename-derived parent groups and image paths are not, by themselves, globally normalized source-photograph identifiers. The revision used the first `DJI` token as a coarse source proxy; it required no fallback records, but 707/877 T2 keys and 826/964 T3 keys recur across date prefixes. The conservative date+DJI proxy reduces overlap to 37.11% in T2 and 72.31% in T3, showing that the DJI-only 97.72% and 99.60% values are upper bounds rather than unique-photograph estimates. Split-file SHA-256 digests confirm file identity, not split validity. This does not replace a global near-duplicate-image audit, and complete all-class cross-split and independent-cluster support remains necessary for stronger independence or rare-class claims. The retained comparisons do not eliminate these limitations.

Statistical coverage is also limited. Global adapted 8B results have two seeds, the 2B arm one, and only Qwen and ResNet have three matched tower14 seeds. Swin tower14 is single-seed. Some nine-tower sensitivity contrasts use unmatched seeds. The reported cluster tests require exchangeability assumptions; the earlier shifted-null endpoints are not treated as confidence intervals and are omitted. The 448 px vision setting was selected on test data, and the instrumented pixel pass differs from the main pass. The post-hoc token-count proxy diagnostics in Section 5.7.1 narrow but do not eliminate resource non-equivalence: the Swin comparison uses final-stage token count, the exact 256-token subset covers only 2.4% of items and three classes, and the shared $\leq$256-token subset is 96.0% `good`. Neither seed repetition nor resampling transforms one source dataset into external validation.

The earlier Swin Unicode-normalized per-class pass is distinct from its standard pass. This distinction is documented rather than reconciled by combining their scores. Hosted models lack pinned weight revisions. S200 covers only five T3 classes, T1 is not evaluated, and all reported recognition uses ground-truth crops.

6.5. Future Work

The most direct extensions are a near-duplicate-image audit, complete all-class train/validation/test support records, independent prediction/code verification, and stronger tower- and line-level replication. A pre-registered, per-item matched-token and matched-compute experiment would address the remaining resource effects beyond the post-hoc diagnostics. Deployment evidence requires a common end-to-end timing program and representative hardware; none of these extensions is claimed as completed.

7. Conclusions

On this single-source InsPLAD-derived resource, reported model comparisons depend on the evaluated items, labels, resolution, prompt information and seed. The task representation and audit protocol help expose those dependencies. A matched binary item test does not establish equivalence, and the available results do not establish a stable ordering of entire model families.

Three observations are directly supported by the tabulated diagnostics: single-digit-support classes can dominate macro recall; the Qwen–ResNet direction changes across the reported tower14 seeds; and residual Qwen–InternVL differences persist at the tested source-pixel caps and in post-hoc near-token-matched diagnostics. The cross-family token-count proxy comparison yields a descriptive +13.97-point macro-recall difference; no inferential test is reported for this specific prediction pair. These observations are not token-, architecture- or compute-isolated explanations. Rare-class generalization, externally reproducible training stability, and the cause of the residual model difference remain unresolved.

The original split has source-photograph dependence; reported tower-disjoint retraining provides only a limited within-source check. The benchmark adds no independent imagery, T1 is not evaluated, and the classifiers operate on ground-truth crops. Workstation measurements do not demonstrate embedded or end-to-end UAV deployment. These boundaries are part of the result, rather than evidence that VLMs are generally superior or inferior.

Author Contributions: Conceptualization, L.Z. and S.X.; methodology, L.Z., S.X. and P.Y.; software, L.Z. and J.K.; validation, J.K. and P.Y.; formal analysis, L.Z. and P.Y.; investigation, L.Z., J.K. and P.Y.; resources, S.X.; data curation, J.K. and P.Y.; writing—original draft preparation, L.Z.; writing—review and editing, L.Z., S.X., J.K. and P.Y.; visualization, J.K.; supervision, S.X. and P.Y.; project administration, S.X.; funding acquisition, S.X. The division of work follows the structure of the audit. L.Z. designed the six-axis parity protocol and the cost-sensitive risk formulation with abstention, implemented the evaluation harness and the adaptation pipeline for both the VLM and the vision-only arms, and ran the resolution control, the instrumented pixel-budget pass, the prompt ablation, and the split-level retraining experiments. S.X. conceived and coordinated the study, provided the data and computational resources, and supervised the audit protocol, the statistical design, and the multiple-comparison correction. J.K. built the ElecVQA-Bench construction pipeline from the InsPLAD release, curated the split manifests with their SHA-256 digests and the per-class support tables, instrumented the grid, token-count and input-hash logging, and prepared the tables and figures. P.Y. developed the prevalence-sensitivity analysis and the cluster-aware randomization inference, performed the quantization and deployment-envelope measurements, and independently reproduced the tower-level three-seed replication. All authors contributed to the interpretation of the results and to revising the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Project of State Grid Sichuan Electric Power Company, grant number 521997240003.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The project repository is available at https://github.com/GhostDino/elecvqa-bench (accessed on 21 October 2026). Construction scripts, split manifests with SHA-256 digests, prompt templates, the evaluation harness, per-class support tables, the construction audit log, the revision-evidence package, prediction files with reported hashes, and LoRA adapter weights are prepared for release through this repository and archival with a DOI upon acceptance. Source imagery is the public InsPLAD dataset [1], is used under its public release terms, and is not redistributed. Derived metadata, code, and adapter weights will be released under licenses compatible with those terms.

Acknowledgments: The authors thank the maintainers of the InsPLAD dataset for releasing the source imagery and annotations.

Conflicts of Interest: The authors declare that this study received funding from the Science and Technology Project of State Grid Sichuan Electric Power Company. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Abbreviations

BalAcc	Balanced accuracy
bf16	Brain floating-point, 16-bit
CI	Confidence interval
CNN	Convolutional neural network
EAR	Error-asymmetry ratio
FP8	Floating-point, 8-bit
INT8	Integer, 8-bit
LoRA	Low-rank adaptation
NF4	4-bit NormalFloat
QLoRA	Quantized low-rank adaptation
ROI	Region of interest
S200	Shared 200-item evaluation subsample
SD	Standard deviation
TNR	True negative rate
TPR	True positive rate
UAV	Unmanned aerial vehicle
ViT	Vision transformer
VLM	Vision–language model
VQA	Visual question answering

Appendix A

S200 is a fixed 200-item subsample drawn once with seed 42 and used only in Tables A8 and A12. The value 0.949 for InternVL-2B in Table 7 and 0.932 in Table A12 differ because the first is the full 2021-item test set and the second is S200. The crop, image, tower, and tower14 test sets are separate split diagnostics and are not comparable to one another or to the original T3 test set.

The reported hosted services do not expose a stable weight-revision hash, so the corresponding zero-shot results are provenance snapshots rather than version-pinned reproductions.

Appendix A.1. Prediction-Source Index

The seven contrasts use the common-six class set, with macro difference defined as Qwen minus ResNet and n equal to the total row count of the paired prediction files. The prediction-source index is:

- `crop`: `crop_qwen.jsonl` and `crop_resnet.json`
- `image`: `image_qwen.jsonl` and `image_resnet.json`
- `tower_s42`: `tower_qwen_seed42.jsonl` and `tower_resnet_seed42.json`
- `tower_s43`: `tower_qwen_seed43.jsonl` and `tower_resnet_seed43.json`

- `tower14_s42`, `tower14_s43`, and `tower14_s44`: the corresponding Qwen and ResNet prediction files in the evidence package

SHA-256 prefixes for the Qwen/ResNet files are, respectively: `397d47a5/a1f9a869`, `9b46525d/80d6869d`, `2cc4acd0/d2c99209`, `0f365ed3/a7a96d70`, `a603fa5a/b08a2827`, `b7aae39b/d53fb4bd`, and `4efc3456/1fe1c584`.

InternVL3.5-8B varies by about fourteenfold across this smoothing range. The displayed rankings, including the 27B/122B ordering, remain unchanged.

The mixed corpus exceeds the T3-only specialist by 8.6 points and the T2-only specialist by 12.5 points on this diagnostic. Extending training from three to ten epochs reduces it by 10.1 points while evaluation loss rises from 0.035 to 0.099, which is the over-fitting signal discussed in Section 5.4.

Appendix A.2. Supplementary Experimental Detail

The following tables retain secondary experimental detail from the main analyses. Moving them does not create additional independent observations.

Table A1. Label crosswalk between tasks.

InsPLAD Source Label	T2 Label (K = 2)	T3 Label (K = 7)	Definition
good	normal	good	No visible defect
rust	defect	rust	Localized oxidation on metal fittings
corrosão	defect	corrosão	Generalized corrosion of metallic surfaces
peeling-paint	defect	peeling-paint	Loss of protective coating
missing-cap	defect	missing-cap	Missing or displaced component cap
torned-up	defect	torned-up	Torn or severed element
nest	defect	nest	Foreign object, bird nest, on the asset

Table A2. Available per-class test support for the original partitions. The verified rare-class support is 0/0/4 train/validation/test for peeling-paint and 15/2/5 for torned-up; other class-wise counts and independent photograph/tower counts are not supplied by these totals.

Partition	Class	*n*
T2 (2021)	normal	1255
T2 (2021)	defect	766
T3 (7284)	good	6518
T3 (7284)	rust	334
T3 (7284)	corrosão	184
T3 (7284)	missing-cap	136
T3 (7284)	nest	103
T3 (7284)	torned-up	5
T3 (7284)	peeling-paint	4

Table A3. Software and run configuration.

Component	Setting
Software	PyTorch 2.11.0+cu128; Transformers 5.14.1; PEFT 0.20.0; Swift 4.5.2; bitsandbytes weight-only quantization
Accelerator	1 × NVIDIA RTX PRO 6000 Blackwell (96 GB)
Training precision	bf16
LoRA	rank 32, $\alpha = 64$, dropout 0.05; target modules q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj; vision encoder frozen
Optimizer	AdamW; VLM learning rate 1×10^{-4}, cosine schedule, 3% warmup, effective batch 16, weight decay 0.01
Checkpoint selection	final checkpoint after the fixed epoch budget; no early stopping
Loss	VLM autoregressive token loss; vision-only class-weighted cross-entropy
Epochs	3 for VLMs; 10 for vision-only baselines
Vision-only input	224/448/672/896 px. Training: resize, random horizontal flip, random rotation 10°, color jitter 0.2/0.2/0.2, ImageNet normalization. Evaluation: resize and ImageNet normalization.
InternVL3.5-8B input	448 px tiles, dynamic tiling to 12 tiles, 256 tokens per tile; audited native mean 2.94 patches and 752.9 visual tokens.
Qwen3-VL-8B input	processor-controlled dynamic resolution; audited native mean 329.5 visual tokens and median 144.
Hosted API models	see Table A4; endpoint, model identifier, and access date recorded; hosted weights not version-pinned
Quantization	bitsandbytes NF4 and INT8 weight-only, bf16 activations
Seeds	global VLM 41 and 42; vision-only training-budget sweep 3 seeds; tower14 42, 43, 44

Table A4. Hosted API provenance.

Model Identifier	Endpoint	Access Date	Reproducibility Status
Qwen3.8-27B	`<internal-vllm-endpoint-redacted>`	24 August 2026	service-reported model root `/root/data-tmp/Qwen3.8-27B`; no weight-revision hash
Qwen3.5-122B-A10B-FP8	`<internal-vllm-endpoint-redacted>`	24 August 2026	service identifier only; no weight-revision hash

Table A5. Seven reproduced cluster sign-flip contrasts. All reruns use 10,000 randomizations and randomization seed 42 from copied remote prediction files. This computational reproduction does not add new model runs or verify every manuscript contrast.

Contrast	Cluster Key	*n*	Clusters	Macro Difference	*p*-Value
Crop	`_tower`	6672	42	−0.005105	0.690131
Image	`_parent_id`	6620	4319	−0.044033	0.401760
Tower seed 42	`_tower`	13,574	9	+0.080213	0.023898
Tower seed 43	`_tower`	13,574	9	+0.032904	0.449855
Tower14 seed 42	`_tower`	5536	14	−0.019149	0.090391
Tower14 seed 43	`_tower`	5536	14	+0.006151	0.918908
Tower14 seed 44	`_tower`	5536	14	−0.114201	0.648235

Table A6. Experiment ledger. Rows marked as included are accounted for within another row's budget.

Group	Experiment	GPU-h	Evaluated Set	Seeds
Construction	Audit, split, ROI and VQA construction	0	—	—
Zero-shot	7-model zero-shot	0	T2 2021; T3 7284	1
Adaptation	LoRA adaptation, 8B, 2B, Qwen	41	T2 2021; T3 7284; S200	1
Baseline	Vision-only baseline, superseded	2	7851	1
Budget sweep	Vision-only training budget, re-scored	6	T2 2021	3
T3 baseline	Vision-only T3 baseline	4	T3 7284	1
Ablation	Corpus and epoch ablations	38	S200	1
Quantization	INT8 and NF4, full test	4	T2 2021; T3 7284	1
Resolution	Resolution sweep, 224/448/672/896 px	2	T3 7284	1
Tile control	InternVL single-tile control	1	T3 7284	1
Pixel budget	Qwen pixel-budget audit	1.8	T3 7284	1
Pixel budget	InternVL source-pixel control	2.0	T3 7284	1
Prompt	Prompt ablation, InternVL, three arms	10	T3 7284	1
Prompt	Prompt ablation, Qwen replication	1	T3 7284	1
Inference	McNemar and cluster-aware randomization	0 (CPU)	existing predictions	—
Verification	Independent-path verification and leakage audit	0	T2 200/full	—
Split	Crop and image split-specific retraining	included	crop and image T3 tests	1
Split	Tower split construction and retraining	7	tower T3 13,574	Qwen 2; ResNet 3
Replication	Tower14 three-seed replication	65	tower14 T3-only 5536	3/3
Replication	Tower14 Swin-T control	0.5	tower14 T3-only 5536	1
Deployment	Vision-only deployment benchmark	0.2	512 images at 448/672 px	1
Replication	Qwen global seed-41 replication	11	T2 2021; T3 7284	2
Listed subtotal		196.5		

Table A7. Sensitivity of the error-asymmetry ratio to the smoothing constant.

Model	$\epsilon = 0.1$	$\epsilon = 0.01$	$\epsilon = 0.001$
InternVL3.5-2B	0.12	0.01	0.00
Qwen3-VL-2B	0.16	0.03	0.02
Qwen2.5-VL-7B	0.45	0.30	0.28
Qwen3.8-27B	6.80	33.41	60.93
InternVL3.5-8B	6.94	41.23	96.09
Qwen3.5-122B	3.89	8.96	10.65
Qwen3-VL-8B	4.13	13.30	18.40

Table A8. S200 five-class T3 diagnostic, InternVL3.5-2B + LoRA. S200 contains 177 `good`, 12 `rust`, 7 `missing-cap`, 3 `corrosão`, and 1 `nest` item, and no instances of `peeling-paint` or `torned-up`. These values are therefore a five-class diagnostic and are not comparable to any seven-class result in this paper.

Configuration	Train Items	Epochs	T3 BalAcc (Five Classes)	Eval Loss
T2-only	6949	3	0.757	—
T3-only	25,460	3	0.796	—
Mixed	32,409	3	0.882	0.035
Mixed	32,409	10	0.781	0.099

Table A9. Core-result provenance and comparison units.

Result	Run and Unit Provenance
Binary matched comparison	Qwen seed 42 and Swin-T seed-42 prediction file on the identical 2021-item T2 test set; the tabulated Swin-T mean and SD describe three training seeds and are not the prediction vector used by McNemar's test.
Original T3 Qwen–ResNet decomposition	Qwen seed-42 seven-class prediction file and ResNet-50 448 px standard T3 pass; The reported Table 12 recall values, macro recall and contributions are arithmetically consistent; validating the underlying files remains distinct from that check.
InternVL–Swin per-class decomposition	InternVL native-tiling seed-42 pass and an earlier Unicode-normalized Swin-T per-class pass (75.89%), deliberately distinguished from the 75.11% standard pass in Table 10.
Tower14 replication	Qwen and ResNet-50 seeds 42–44 on the tower14 T3-only test set; Swin-T is a single seed-42 control.
Deployment benchmark	Vision-only checkpoints at 448 and 672 px, 512 images, five repeats, one workstation GPU, preloaded tensors; not directly comparable with the 50-item VLM throughput measurement.

Table A10. Global two-seed replication of the two 8B adapted VLMs. The T3 column is seven-class macro recall obtained by mapping option outputs back to labels.

Model	Seed	T2 BalAcc	T3 Macro Recall (K = 7)
Qwen3-VL-8B + LoRA	42	0.9714	0.8908
Qwen3-VL-8B + LoRA	41	0.9800	0.9010
Qwen3-VL-8B + LoRA	mean ± SD	0.9757 ± 0.0061	0.8959 ± 0.0072
InternVL3.5-8B + LoRA	42	0.9586	0.7790
InternVL3.5-8B + LoRA	41	0.9586	0.7440

Table A11. McNemar's exact test on paired T2 predictions ($n = 2021$). Holm family: the three comparisons in this table. The displayed ordinary-accuracy difference intervals use an item-independence approximation and are diagnostic only; they are not cluster- or seed-level equivalence intervals.

Comparison	Discordant Pairs (b, c)	Δ Accuracy	95% CI	p (Exact)	p (Holm)	Verdict
InternVL-2B vs. Qwen3-VL-8B	21, 54	−1.63 pt	[−2.47, −0.80]	0.0002	0.0005	significant
InternVL-8B vs. Qwen3-VL-8B	20, 40	−0.99 pt	[−1.74, −0.24]	0.0135	0.0270	significant
InternVL-8B vs. InternVL-2B	26, 13	+0.64 pt	[−0.01, +1.29]	0.0533	0.0533	not detected

Table A12. Training corpus and epoch budget, InternVL3.5-2B + LoRA, evaluated on the shared 200-item subsample S200. The T3 diagnostic for these configurations covers only five of the seven classes and is reported separately in Table A8.

Configuration	Train Items	Epochs	T2 BalAcc	T2 β	Eval Loss
T2-only	6949	3	0.908	17.57%	—
T3-only	25,460	3	0.874	24.32%	—
Mixed	32,409	3	0.932	13.51%	0.035
Mixed	32,409	10	0.935	12.16%	0.099

Table A13. Vision-only backbones across training budget on the T2 test partition ($n = 2021$), three seeds, mean ± SD.

Examples per Class	ResNet-50	Swin-T
4	61.48 ± 4.15	71.66 ± 8.92
8	67.14 ± 0.29	72.65 ± 8.30
16	72.70 ± 1.07	81.45 ± 6.36
32	81.59 ± 0.67	89.30 ± 1.51
64	89.56 ± 1.13	90.80 ± 1.25
128	92.24 ± 0.26	93.88 ± 0.73
all	96.54 ± 0.10	97.17 ± 0.23

Table A14. McNemar's exact test on raw per-item correctness, Qwen3-VL-8B against vision-only configurations, T3 test ($n = 7284$).

Comparison	Macro-Recall Gap	p (Exact)	Verdict on Raw Accuracy
Qwen3-VL-8B vs. ResNet-50, 448 px	8.0 pt	0.059	not detected
Qwen3-VL-8B vs. ResNet-50, 672 px	10.6 pt	0.923	not detected
Qwen3-VL-8B vs. ResNet-50, 896 px	12.9 pt	0.002	detected
Qwen3-VL-8B vs. Swin-T, 448 px	14.0 pt	0.078	not detected

Table A15. Qwen3-VL-8B pixel-budget audit on the original T3 test set ($n = 7284$). Macro recall uses the seven-class label space.

`max_pixels`	Inputs Changed vs. Native	Mean Visual Tokens	Median Visual Tokens	Macro Recall	Prediction SHA-256 (Prefix)
65,536	4472	62.7	64	0.7861	`e14507e15251`
131,072	3727	97.6	121	0.8261	`a163d5c9a9e6`
200,704	2839	129.7	144	0.8343	`033dd1a4bf96`
401,408	2025	189.3	144	0.8908	`b22525e9c05b`
802,816	1063	276.9	144	0.8926	`ff9cc972daa2`
native (100,000,000)	0	329.5	144	0.8939	`03380898f7c5`

Table A16. Source-pixel-matched comparison of Qwen3-VL-8B and InternVL3.5-8B on the original T3 test set ($n = 7284$). Inputs changed and patch and token statistics refer to the InternVL arm; Qwen token statistics are in Table A15.

`max_pixels`	Qwen Macro Recall	InternVL Macro Recall	Qwen–InternVL Gap	InternVL Inputs Changed	InternVL Mean Patches	InternVL Mean Visual Tokens
65,536	0.7861	0.7118	+7.43 pt	4699	1.35	346.1
131,072	0.8261	0.7519	+7.42 pt	3753	1.35	346.1
200,704	0.8343	0.7566	+7.77 pt	2976	1.35	346.1
401,408	0.8908	0.7548	+13.61 pt	1982	1.35	346.1
802,816	0.8926	0.7710	+12.16 pt	1092	2.36	605.1
native (100,000,000)	0.8939	0.7734	+12.05 pt	0	2.94	752.9

Table A17. Prompt ablation, InternVL3.5-8B + LoRA, T3 test ($n = 7284$). Values are seven-class macro recall, not ordinary sample accuracy. P3 supplies the asset description, decision tree, and candidate options; P1 supplies the question stem, asset name, and candidate options only.

Training Prompt	Inference Prompt	BalAcc	Δ from Baseline
P3	P3 (baseline)	77.90%	—
P3	P1	14.34%	−63.56
P1	P1 (controlled)	14.29%	−63.62
P1	P3 (cross)	69.26%	−8.64

Table A18. Prompt ablation replicated on Qwen3-VL-8B, T3 test ($n = 7284$). Values are seven-class macro recall, not ordinary sample accuracy.

Configuration	InternVL3.5-8B + LoRA	Qwen3-VL-8B + LoRA
P3 baseline	77.90%	89.08%
P1 + P1 (controlled)	14.29%	14.29%
P1 + P3 (cross)	69.26%	84.36%
Inference prompt switch, P1 → P3	54.97 pt	70.07 pt
Training prompt switch, P1-trained → P3-trained	8.64 pt	4.72 pt

Table A19. Interpretation of the tested factors; this is an audit summary, not a complete causal decomposition.

Factor	Observed Diagnostic	Inference Boundary
Resolution	Scores change across tested resolutions/caps	Bracketing/cap matches are not information or compute matches
Visual tokens	Native Qwen count lower than InternVL	Tokenization and encoder effects not isolated
Rare classes	Class-set changes alter gaps and sometimes ranks	Single-digit support does not establish population recall
Prompt definitions	Large drops in tested P1 configurations	No universal necessity or mechanism established
Scale	InternVL raw T2 accuracy gap +0.64 pt; T2 BalAcc gap +0.95 pt; T3 macro gap +3.74 pt	Metrics differ; no general scaling law
Architecture	Both CNN and transformer vision baselines included	Architecture-specific effects remain unresolved
Split and seed	Different directions across reported settings	No stable ordering of model families established

Table A20. Inference precision, InternVL3.5-8B + LoRA, full test sets (T2 $n = 2021$; T3 $n = 7284$), with paired McNemar tests against bf16 and throughput measured over 50 items.

Precision	T2 BalAcc	T2 β	T3 BalAcc	T3 Macro-F1	Memory	Throughput	T2 p	T3 p
bf16	0.9586	7.57%	0.7790	0.7954	≈16 GB	1.54 img/s	—	—
INT8	0.9592	7.44%	0.7790	0.7950	≈9 GB	1.17 img/s	1.000	0.625
NF4	0.9592	7.44%	0.7806	0.7964	6.2 GB	1.49 img/s	1.000	1.000

Table A21. Vision-only deployment benchmark on one NVIDIA RTX PRO 6000 Blackwell (PyTorch 2.11.0+cu128; 512 images; five repeats; latency $n = 200$). Peak memory is measured over the full batch sweep rather than model weights alone.

Input	Backbone	Peak Memory	p50 Latency	p95 Latency	Best Batch Throughput	Sustained Throughput (Batch 64)
448 px	ResNet-50	5.97 GB	4.49 ms	4.63 ms	806.4 img/s (batch 32)	747.8 img/s
448 px	Swin-T	3.70 GB	6.28 ms	6.36 ms	525.7 img/s (batch 32)	524.7 img/s
672 px	ResNet-50	22.03 GB	6.29 ms	6.46 ms	344.8 img/s (batch 32)	340.8 img/s
672 px	Swin-T	8.19 GB	6.63 ms	6.75 ms	241.6 img/s (batch 32)	234.0 img/s

Table A22. Per-class recall after split-specific retraining, common six classes.

Class	Crop Qwen	Crop ResNet	Image Qwen	Image ResNet	Tower Qwen	Tower ResNet
corrosão	0.9076	0.9496	0.9353	0.9496	0.9005	0.8104
good	0.9983	0.9627	0.9975	0.9784	0.9979	0.9886
missing-cap	0.9921	0.9843	0.9759	0.9699	0.9380	0.8450
nest	1.0000	1.0000	0.9710	0.9710	0.8238	0.6166
rust	0.9164	0.9486	0.9065	0.9315	0.8370	0.8463
torned-up	0.8333	0.8333	0.7500	1.0000	0.0909	0.0000

Table A23. Nine-tower seed sensitivity, common six classes. Holm family: these three contrasts only. The last contrast uses unmatched seeds and is not a matched-seed replication.

Seeds	Qwen	ResNet	Gap (pt)	Raw p	Holm p
42/42	0.7647	0.6845	+8.02	0.0239	0.0478
43/43	0.7560	0.7231	+3.29	0.4499	0.4499
43/44	0.7560	0.7057	+5.03	0.0091	0.0273

Table A24. Single-seed tower14 Swin control (seed 42). Difference is Qwen minus Swin. The source test file has 5536 items; the six- and seven-class metrics have different estimands.

Class Set	Qwen	ResNet	Swin	Gap (pt)	Raw p
Common six	0.9369	0.9561	0.9695	−3.26	0.0191
Full seven	0.9459	0.8195	0.9024	+4.35	0.9828

Table A25. Conditional exact-binomial illustration for reported rare-class counts (two-sided 95% Clopper–Pearson). These intervals assume independent Bernoulli trials; correlated crops or few towers invalidate a naive field interpretation. They are not replacements for cluster-aware evidence.

Reported Class/Run	Correct/Test	Recall	Conditional Interval
Peeling/InternVL	4/4	100%	[39.76%, 100%]
Peeling/Qwen	3/4	75%	[19.41%, 99.37%]
Peeling/Swin per-class pass	0/4	0%	[0%, 60.24%]
Torned-up/InternVL	0/5	0%	[0%, 52.18%]

References

1. Vieira e Silva, A.L.B.; de Castro Felix, H.; Simões, F.P.M.; Teichrieb, V.; dos Santos, M.M.; Santiago, H.; Sgotti, V.A.C.; Lott Neto, H.B.D.T. InsPLAD: A Dataset and Benchmark for Power Line Asset Inspection in UAV Images. *Int. J. Remote Sens.* **2023**, *44*, 7294–7320. [CrossRef]
2. Rudolph, M.; Wandt, B.; Rosenhahn, B. Same Same But DifferNet: Semi-Supervised Defect Detection with Normalizing Flows. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 1907–1916. [CrossRef]
3. Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; Wandt, B. Fully Convolutional Cross-Scale-Flows for Image-Based Defect Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2022; pp. 1829–1838. [CrossRef]
4. Nguyen, V.N.; Jenssen, R.; Roverso, D. Automatic Autonomous Vision-Based Power Line Inspection: A Review of Current Status and the Potential Role of Deep Learning. *Int. J. Electr. Power Energy Syst.* **2018**, *99*, 107–120. [CrossRef]
5. Tao, X.; Zhang, D.; Wang, Z.; Liu, X.; Zhang, H.; Xu, D. Detection of Power Line Insulator Defects Using Aerial Images Analyzed with Convolutional Neural Networks. *IEEE Trans. Syst. Man. Cybern. Syst.* **2020**, *50*, 1486–1498. [CrossRef]
6. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NeurIPS), Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.

7. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
8. Wang, H.; Luo, H.; Qiao, X.; Huo, M.; Xu, X. Data-Driven Distributed Robust Monitoring and Control Optimization for Interconnected Systems. *IEEE Trans. Ind. Inform.* **2025**, *21*, 1399–1408. [CrossRef]
9. Wang, H.; Luo, H.; Ren, L.; Huo, M.; Jiang, Y.; Kaynak, O. Data-Driven Design of Distributed Monitoring and Optimization System for Manufacturing Systems. *IEEE Trans. Ind. Inform.* **2024**, *20*, 9455–9464. [CrossRef]
10. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. MVTec AD: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9584–9592. [CrossRef]
11. Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; Gehler, P. Towards Total Recall in Industrial Anomaly Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14298–14308. [CrossRef]
12. Liu, J.; Xie, G.; Wang, J.; Li, S.; Wang, C.; Zheng, F.; Jin, Y. Deep Industrial Image Anomaly Detection: A Survey. *Mach. Intell. Res.* **2024**, *21*, 104–135. [CrossRef]
13. Provost, F.; Fawcett, T. Robust Classification for Imprecise Environments. *Mach. Learn.* **2001**, *42*, 203–231. [CrossRef]
14. Drummond, C.; Holte, R.C. Cost Curves: An Improved Method for Visualizing Classifier Performance. *Mach. Learn.* **2006**, *65*, 95–130. [CrossRef]
15. Elkan, C. The Foundations of Cost-Sensitive Learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI), Seattle, WA, USA, 4–10 August 2001; pp. 973–978.
16. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
17. Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; Belongie, S. Class-Balanced Loss Based on Effective Number of Samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9260–9269. [CrossRef]
18. Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; Kalantidis, Y. Decoupling Representation and Classifier for Long-Tailed Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 26–30 April 2020.
19. Buda, M.; Maki, A.; Mazurowski, M.A. A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef] [PubMed]
20. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 8748–8763.
21. Li, J.; Li, D.; Savarese, S.; Hoi, S. 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 (ICML), Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.
22. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS), New Orleans, LA, USA, 10–16 December 2023.
23. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. InternVL: Scaling Up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 24185–24198. [CrossRef]
24. Qwen Team. Qwen2.5-VL-7B-Instruct Model Card. 2025. Available online: https://modelscope.cn/models/Qwen/Qwen2.5-VL-7B-Instruct (accessed on 21 October 2026).
25. Rohrbach, A.; Hendricks, L.A.; Burns, K.; Darrell, T.; Saenko, K. Object Hallucination in Image Captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, 31 October–4 November 2018; pp. 4035–4045.. [CrossRef]
26. Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W.X.; Wen, J.-R. Evaluating Object Hallucination in Large Vision-Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, 6–10 December 2023; pp. 292–305. [CrossRef]
27. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
28. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS), New Orleans, LA, USA, 10–16 December 2023.
29. Caruana, R. Multitask Learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
30. Standley, T.; Zamir, A.; Chen, D.; Guibas, L.; Malik, J.; Savarese, S. Which Tasks Should Be Learned Together in Multi-Task Learning? In Proceedings of the 37th International Conference on Machine Learning (ICML), Virtual, 13–18 July 2020; pp. 9120–9132.

31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
32. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 10012–10022. [CrossRef]
33. McNemar, Q. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef] [PubMed]
34. Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput*. **1998**, *10*, 1895–1923. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.