BEYOND AGGREGATE ACCURACY: A DEPLOYMENT-RISK-AWARE EVALUATION PROTOCOL AND BENCHMARK FOR MIXED-DEFECT WAFER-MAP VQA

Anonymous authors
Paper under double-blind review

ABSTRACT

Real wafer diagnosis is compositional: a model must determine whether a wafer is abnormal, identify all co-occurring defect patterns, localize them, and produce image-grounded diagnostic language. Existing wafer-map benchmarks report aggregate task scores and therefore cannot show whether a screening model is deployable, whether its later answers are grounded in the image, or whether the benchmark itself can be solved without multimodal reasoning. We propose a *deployment-risk-aware evaluation protocol* for hierarchical industrial VQA and instantiate it as **Mixed-WM38 VQA**, a benchmark of 190,075 linked question–answer instances over 38,015 wafer maps and 38 normal, single-defect, and mixed-defect combinations. The protocol has three parts: a class-conditional risk decomposition that separates false alarms (FA) from fatal misses (FM) and requires Q1–Q3 tuple correctness; a model-independent determinism audit that detects, and in our own first release removed, later answers that are deterministic functions of earlier labels; and a mandatory *shortcut floor*, a classical classifier-to-template pipeline that any multimodal method must beat before its language metrics are read as reasoning. On the complete 7,983-wafer deployment-risk evaluation set, with every wafer included, an always-abnormal gate already attains 97.34% accuracy because only 212 wafers are normal. On the evaluated 4B backbone, SFT, replication, domain-QA, CoT, and GRPO variants remain near this score while rejecting every normal wafer. Token probing shows that the correct absence answer is represented but approximately $18\times$ less likely than the affirmative answer. Controlled interventions expose a non-trivial FA–FM frontier rather than a universal fix, and the same replicated recipe yields 7.6% or 30.6% fatal misses depending only on the training seed; a second intervention that removes replication entirely is more seed-sensitive still, spanning the entire fatal-miss range (1.25% to 100%) across three seeds, two of which collapse to the always-normal shortcut. On a 7B backbone the same margin-plus-selective-GRPO chain holds FA and FM below 0.5% at both seeds, so the frontier's severity is backbone-dependent. Because FA and FM are prevalence-free, the same operating points transfer to a fab-like 90%-normal prior, where the always-abnormal rule drops to 10.0% accuracy and the ranking inverts. Applied to our own benchmark, the shortcut floor shows that a compact CNN plus deterministic lookup exceeds the VLM on every Q1–Q5 metric, so its language scores measure structured-format competence, not open-ended reasoning. The central result is a four-way separation: aggregate accuracy, class-conditional deployment risk, internal cross-stage consistency, and end-to-end cascade correctness measure different properties and can move in different directions. Hierarchical diagnostic systems should be selected with all four audits, the shortcut floor, and run-to-run stability, not aggregate accuracy or fluent text alone. (verbatim generated outputs are in the anonymous release: https://anonymous.4open.science/r/WM38_VQA-9C2B/README.md)

1 INTRODUCTION

A wafer-screening model can score 97% accuracy and still reject every good wafer: on our 7,983-wafer deployment-risk evaluation set only 212 wafers are normal, so an always-abnormal gate is already

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

97.34% accurate. Such a gate is unusable in production, and it can initiate a complete downstream narrative (defect types, locations, morphology, and a plausible process explanation) for a wafer that is actually normal.

The issue is not imbalance alone, but how imbalance interacts with a hierarchical answer cascade. A majority shortcut at the first decision can be rewarded by aggregate accuracy, while mutually consistent downstream text makes the resulting failure appear coherent. Existing wafer-map work largely studies closed-set recognition on WM-811K (Wu et al., 2015; Nakazawa & Kulkarni, 2018) or mixed-defect classification (Kyeong & Kim, 2018; Wang et al., 2020). Recent industrial MLLMs extend inspection toward natural-language reasoning (Gu et al., 2024; Jiang et al., 2025; Kang et al., 2026), but do not characterize the class-conditional risk of the gate, its self-conditioned propagation through a linked answer chain, or how much of their benchmarks a non-multimodal shortcut can solve.

This paper asks three questions: (i) what must an evaluation protocol report before a hierarchical industrial VQA model can be called deployable; (ii) can standard training recipes escape the majority shortcut, and do targeted interventions improve both sides of deployment risk; and (iii) how much of a linked wafer-VQA benchmark is solvable without multimodal reasoning at all? We answer them with **Mixed-WM38 VQA**, a benchmark we construct and release as the vehicle for the protocol. The contribution is the protocol and what it reveals; the dataset is derived from the public MixedWM38 corpus (Wang et al., 2020) and is not claimed as new imagery. Our contributions are fourfold.

A deployment-risk-aware evaluation protocol. We report FA, FM, balanced accuracy, Q1–Q3 tuple correctness, and cross-stage support, each with its denominator, and show how the class-conditional pair (FA, FM) transfers across priors while any single-prior accuracy ranking does not. The protocol exposes an always-abnormal operating point hidden behind approximately 97.3% gate accuracy in every tested pre-intervention recipe.

Two benchmark-validity audits: determinism and a shortcut floor. A model-independent pairwise (Q_i, Q_j) determinism audit over the full corpus detected and removed a shortcut in our own first release, in which location was a deterministic relabeling of defect type. A classifier-to-template *shortcut floor* then bounds, adversarially against our own benchmark, how much of the Q1–Q5 protocol is solvable by 38-way classification plus lookup. It exceeds the VLM on every metric; we release it and argue it should be mandatory for wafer VQA.

An empirical diagnosis of hidden deployment risk. Token probabilities localize a suppressed absence answer that is reachable but $18\times$ less likely than the affirmative. Margin and selective-reward interventions show that reducing FA can shift risk toward FM, threshold recalibration cannot reproduce the trained frontier, and the identical recipe can land at opposite ends of the FA–FM plane depending only on the seed.

A released benchmark as the vehicle. Mixed-WM38 VQA provides 190,075 linked question–answer instances over 38,015 wafer maps and 38 label combinations, covering anomaly screening, set-valued defect recognition, per-image spatial grounding, morphology description, and process-cause hypothesis generation, with source mappings, fixed splits, the deployment-risk evaluation set, construction and audit scripts, the shortcut-floor pipeline, evaluation code, and a dataset card.

2 Related Work

Wafer-map recognition. WM-811K established the standard single-defect classification setting (Wu et al., 2015); subsequent CNN methods refined closed-set recognition (Nakazawa & Kulkarni, 2018). Mixed-type recognition was posed by Kyeong & Kim (2018) and scaled in MixedWM38 (Wang et al., 2020). These works predict labels rather than linked natural-language diagnoses.

Industrial VQA and post-training. Prompt-guided industrial MLLMs (Gu et al., 2024), MMAD-style capability evaluation (Jiang et al., 2025), domain-knowledge injection (Kang et al., 2026), and RL-post-trained anomaly reasoners (Chao et al., 2025) broaden inspection beyond classification. Two published lines are closest in structure. MAU-GPT (Wang et al., 2026) introduces MAU-Set, five hierarchical industrial-anomaly QA tasks spanning binary detection to open-ended reasoning, and adapts a generalist MLLM with anomaly-aware experts; it reports per-task aggregate scores on natural-image industrial objects and contains no class-conditional gate audit, determinism audit, or

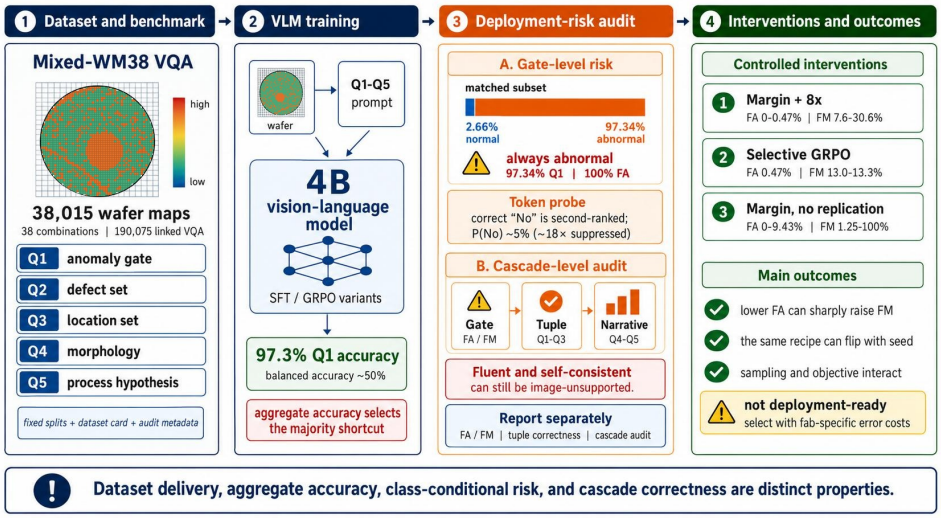


Figure 1: Overview and central result. A 4B VLM reaches approximately 97.3% Q1 accuracy on the 2.66%-normal deployment-risk set, where an always-abnormal rule scores 97.34% while producing 100% false alarms. Token probing and cascade auditing expose the hidden shortcut; controlled interventions reveal seed-sensitive FA–FM operating points (the no-replication recipe spans FM 1.25% to 100%). Aggregate accuracy, class-conditional risk, cross-stage consistency, and cascade correctness therefore require separate audits.

shortcut floor. IAD-R1 (Li et al., 2025) post-trains an anomaly reasoner with a consistency-oriented reward under GRPO (Shao et al., 2024; DeepSeek-AI, 2025); our selective reward differs in scope and purpose: it fires only on normal-gold prompts to counter a measured token-level absence bias, rather than rewarding reasoning–answer agreement, and we evaluate it by FA/FM rather than by aggregate accuracy. GRPO also motivates our reinforcement-stage ablation. Concurrent WaferSAGE (Xu et al., 2026) also derives wafer-map VQA from WM-811K and MixedWM38, but with a different goal: questions, answers, and rubrics are synthesized by proprietary teacher models to distill API-scale capability into a small VLM, scored by an LLM judge on a 31-wafer, 186-question test set of defect wafers. Our object is different: deterministic, label- or pixel-derived answers over the complete corpus with exact scoring on all 26,613 test wafers; an anomaly gate over a 2.66%-normal population with FA/FM decomposition and seed-stability reporting, which its scope (no normal-wafer task) cannot express; and two validity audits, determinism and a shortcut floor, that neither it nor the lines above report.

Absence reasoning and benchmark design. VLMs systematically under-predict negation or absence answers (affirmative bias, Kang et al., 2025); multi-tier VQA benchmarks (Kou et al., 2026) and spatial-reasoning studies (Liu et al., 2025) motivate our task design.

3 Mixed-WM38 VQA Dataset and Benchmark

3.1 Dataset Scope and Five-Stage Protocol

Mixed-WM38 contains 38,015 maps over one normal class, eight single-defect classes, and 29 mixed-defect combinations with up to four co-occurring patterns. Single-defect maps originate from real production data, while mixed classes are partially simulated in the upstream MixedWM38 release (Wang et al., 2020). We retain the upstream class-stratified population, so normal wafers account for 2.63% of the data rather than being artificially balanced. This preserves the benchmark prior—not necessarily a fab's production prevalence—and makes class-conditional risk reporting essential.

For each wafer, the benchmark asks: **Q1**, whether any anomaly is present; **Q2**, the set of defect types; **Q3**, the location of each defect in an image-derived eight-sector vocabulary; **Q4**, a morphology description; and **Q5**, a process-cause hypothesis. Q1 uses exact match. Q2–Q3 use set exact match

Statistic	Value
Wafer maps	38,015
Label combinations (normal / single / mixed)	1 / 8 / 29
Maximum co-occurring defects	4
Normal-class proportion	2.63%
Questions per wafer	5
Full held-out test wafers	26,613
Complete deployment-risk evaluation set	7,983
Normal / abnormal in risk-evaluation set	212 / 7,771
Image-derived location sectors	8

Table 1: Released dataset scale and fixed evaluation denominators. All intervention comparisons evaluate 100% of the fixed 7,983-wafer deployment-risk set; FA and FM are computed over every one of its 212 normal and 7,771 abnormal wafers, respectively.

Checkpoint	Tuple	Q2→Q3	Q2→Q4	Q2→Q5
SFT	11.9%	97.0%	69.6%	92.5%
GRPO (4-reward)	12.9%	100%	72.7%	94.1%
GRPO (+spatial)	13.1%	100%	73.2%	94.7%
CoT	8.6%	100%	87.3%	99.1%
Domain-QA	5.8%	100%	76.8%	94.5%

Table 2: Consistency is not correctness. CoT raises Q2-to-Q4/Q5 support while reducing the jointly correct Q1–Q3 tuple rate: the model becomes more self-consistent about a worse diagnosis. All rows are first-release (v1) checkpoints and labels, where Q3 is a lookup of Q2 (hence Q2→Q3 ≈100%).

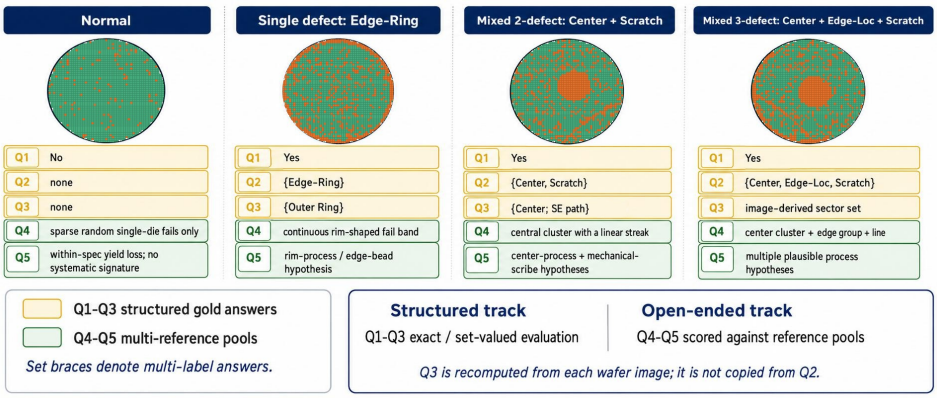


Figure 2: Representative Mixed-WM38 VQA annotation examples (real test-split wafer maps, rendered in the figure palette). Q1–Q3 form the structured anomaly, defect-set, and image-derived location track; Q4–Q5 are evaluated against morphology and process-cause reference pools. Mixed-defect wafers require set-valued answers, and the full-set determinism audit verifies that Q3 is recomputed from each image rather than copied from Q2. Location labels combine a radial term (e.g. "Inner Ring", "Outer Ring", global) with, for direction-sensitive classes (Loc, Edge-Loc, Scratch), one of the eight compass sectors counted in Table 1; the figure shows both kinds.

and per-label F1. Q4–Q5 are scored against per-class multi-reference pools using ROUGE-L, SBERT similarity, and CIDEr. We deliberately call Q5 a hypothesis: its templates are literature-grounded but are not equivalent to verified fab root-cause attribution.

How the cascade is generated and what "propagation" means. Each wafer is answered as a five-turn conversation: Q1–Q5 are issued as five successive user turns, and each turn's generation is conditioned on a dialogue history containing the model's own previously generated answers, never on gold labels. Q2–Q5 are therefore *not* independently prompted; a wrong Q1 remains in context for every subsequent turn. We do not, however, perform counterfactual interventions on Q1 (forcing the opposite gate answer and regenerating), so the downstream behavior we report is *self-conditioned correlated failure*, not an estimated causal effect of the gate; we use "cascade" in this descriptive sense throughout.

Dataset release. The release contains source-wafer mappings, five-stage annotations, canonical vocabularies, fixed splits, the deployment-risk set, construction scripts, determinism-audit reports, the shortcut-floor pipeline, evaluation code, and a dataset card. Where upstream terms prevent redistribution of raw images, identifiers and reconstruction scripts reproduce the benchmark from the authorized source.

3.2 Reference Construction and Audit

Q2 answers are canonical sets over the eight base patterns; Q3 is computed per image from the spatial support of each labeled pattern. Q4 references describe observable shape and spatial arrangement, while Q5 references express plausible process causes and checks. Normal wafers use a short, explicit no-defect chain rather than empty strings, so a correct gate can terminate the diagnostic cascade consistently.

A multi-stage benchmark can appear to test reasoning while leaking later answers through earlier labels. We therefore check all 38,015 wafers and 190,075 linked records for schema validity, label consistency, and pairwise (Q_i, Q_j) determinism. The audit caught an early Q3 design in which every defect type mapped to one fixed location phrase; after rebuilding Q3 from per-image pixel statistics, within-class Q3 entropy rises from 0 to 3.64 bits and the conditional-determinism ratio falls from 1.000 to 0.280 (Appendix B.6).

3.3 Deployment-Risk Evaluation

Let $\mathcal{N}$ and $\mathcal{A}$ denote normal and abnormal wafers, and let $\hat{y} = 1$ mean "abnormal." We report

$$\mathrm{FA} = \frac{1}{|\mathcal{N}|} \sum_{x \in \mathcal{N}} \mathbf{1}[\hat{y}(x) = 1], \qquad \mathrm{FM} = \frac{1}{|\mathcal{A}|} \sum_{x \in \mathcal{A}} \mathbf{1}[\hat{y}(x) = 0]. \tag{1}$$

If p_N is the normal prevalence, Q1 accuracy decomposes as

$$\mathrm{Acc}_{Q1} = p_N(1 - \mathrm{FA}) + (1 - p_N)(1 - \mathrm{FM}). \tag{2}$$

On the complete deployment-risk evaluation set, $p_N = 212/7{,}983 = 2.66\%$; hence an always-abnormal classifier has FA $= 100\%$, FM $= 0\%$, and Q1 accuracy $= 97.34\%$. We also report balanced accuracy,

$$\mathrm{BAcc} = 1 - \frac{1}{2}(\mathrm{FA} + \mathrm{FM}), \tag{3}$$

which removes prevalence but still weights the two errors equally. Because fab costs are application-specific, we do not collapse FA and FM into a single universal utility. Instead, the reported operating points can be evaluated under

$$\mathcal{R}(c_{\mathrm{FA}}, c_{\mathrm{FM}}) = c_{\mathrm{FA}} p_N \mathrm{FA} + c_{\mathrm{FM}}(1 - p_N)\mathrm{FM}, \tag{4}$$

for a deployment's chosen cost ratio. Because FA and FM are prevalence-free, Eq. (2) transfers every reported operating point to any prior. The benchmark prior is defect-heavy; at an illustrative fab-like prior of $p_N = 90\%$ the always-abnormal rule falls from 97.34% to 10.0% accuracy, an always-normal collapse rises from 2.66% to 90.0%, and the trained interventions reported below reach 91–99% (Appendix E.9). No single-prior accuracy ranking survives a prior change; only the class-conditional pair does, which is why the protocol reports it. Finally, Q1–Q3 tuple correctness requires the gate, type set, and location set to be correct on the same wafer; it is all-or-nothing by design, since a diagnosis with the wrong location is not actionable, and per-stage F1 is always reported beside it so partial credit remains visible. Heuristic cross-stage support checks whether each predicted Q2 type is echoed by Q3–Q5; it measures internal agreement, not factual validity. Reporting both separates a correct diagnosis from a merely coherent answer chain.

4 Gate Diagnosis and Interventions

4.1 Diagnosis: A Reachable but Suppressed Absence Token

On all 212 genuinely normal wafers in the deployment-risk evaluation set, the correct "No" token remains the second-ranked candidate with approximately 5% probability, corresponding to an approximately $18\times$ gap from "Yes." Thus the absence token is reachable but systematically under-selected, a pattern consistent with affirmative bias (Kang et al., 2025). Replicating normal training examples $8\times$ changes zero gate decisions, showing that simple exposure rebalancing is insufficient for the tested setup. The probe is a teacher-forced log-probability measurement: it localizes where the failure lives, but it is not by itself a causal-perturbation audit of image use.

4.2 Fix 1: Gate Margin Loss

For a normal training wafer x, define $\Delta(x) = \log P(\text{Yes} \mid x) - \log P(\text{No} \mid x)$. We add

$$\mathcal{L}_{\text{gate}} = \frac{1}{|\mathcal{B}_N|} \sum_{x \in \mathcal{B}_N} \max(0, m + \Delta(x)) \tag{5}$$

to SFT, where $\mathcal{B}_N$ is the set of normal wafers in a batch; the term is omitted for batches with $|\mathcal{B}_N| = 0$, which is common under the unreplicated class distribution. The intervention directly closes the measured token gap; we evaluate it both with and without the $8\times$ replicated normal pool.

4.3 Fix 2: Selective GRPO Reward

We also test a reward term that fires only for normal-gold Q1 prompts during GRPO. Because GRPO computes group-relative advantage among rollouts of the same wafer, the term rewards "No" only where it is correct rather than applying a global label shift. Its two observed seeds yield FM 13.3% and 13.0%, whereas the two replicated-margin runs yield 30.6% and 7.6%. With only two seeds per recipe, we interpret this as lower *observed dispersion*, not proof that the reinforcement mechanism is intrinsically more predictable.

4.4 Isolation: Removing Minority Replication

The two fixes above are both trained on the $8\times$-replicated normal pool. A single-factor isolation retains the identical margin objective but removes the duplicated normal examples, isolating whether repeated minority exposure, rather than the margin term itself, drives the boundary shift reported in the Results section.

5 Experiments

Setup. We use a 4B Qwen3.5 vision-language backbone with LoRA SFT, followed where indicated by GRPO continuation. Pre-intervention variants include plain SFT, $8\times$ normal replication, domain-QA pretraining, a templated CoT stage, and three GRPO reward compositions. Every row is evaluated on all 7,983 wafers of the fixed deployment-risk set (212 normal, 7,771 abnormal), fixed before any model was trained as the first 7,983 records of the released test manifest, not selected by class or difficulty (Appendix A). Paired rows differ only in training seed; the no-replication isolation keeps the margin objective, removes duplicated normal examples, and has three seeds. Confidence intervals, Q4–Q5 metrics, zero-shot and shortcut baselines, exclusions, and full hyperparameters are in Appendix E.

6 Results and Analysis

6.1 Aggregate Accuracy Selects the Majority Shortcut

All tested pre-intervention recipes—SFT, three GRPO reward variants, CoT, domain-QA, and $8\times$ replication—land near 97.3% Q1 EM and 100% FA. Two checkpoints never emit "No" across all 26,613 wafers in the held-out test split. Equation 2 explains why: at the gate, their aggregate score is almost entirely supplied by the abnormal majority. Balanced accuracy falls to approximately 50%, revealing that Q1 EM alone cannot certify a screening function.

6.2 Threshold Calibration Alone Cannot Match the Fixed Interventions

Why retrain instead of recalibrating the baseline's $\log P(\text{Yes}) - \log P(\text{No})$ threshold? Sweeping τ on all 3,800 validation wafers (Appendix E.2) gives a steep, narrow cliff ($\tau \in [2.5, 4.25]$: FA $100\% \to 0\%$ while FM $\sim 0\% \to 100\%$) with no region where both are small: the FA-minimizing point ($\tau = 3.375$) leaves FM at **51.9%**, and the FA=FM balance point ($\tau = 3.125$: 19.0%/18.6%) is dominated on both axes by four of the seven trained runs in Table 3 (both selective-GRPO seeds, margin seed 123, and no-replication seed 42). Applying the validation-selected thresholds unchanged to the same 7,983-wafer deployment-risk set confirms the comparison on a common split: $\tau = 3.375$

Model / intervention	Q1 EM	BAcc	Q2 F1	Q3 F1	FA ↓	FM ↓
Always abnormal (derived)	0.973	0.500	–	–	100%	0%
SFT baseline	0.973‡	≈0.500	0.507	0.258	100%	≈ 0%
+ 8× normal replication	0.973	≈0.500	0.519	0.294	100%	≈ 0%
+ domain-QA	0.973	≈0.500	0.528	0.241	100%	≈ 0%
GRPO (independent)	0.973	≈0.500	0.525	0.284	100%	≈ 0%
GRPO (consistency)	0.973	≈0.500	0.526	0.291	100%	≈ 0%
GRPO (full/spatial)	0.973	≈0.500	0.526	0.286	100%	≈ 0%
Margin + 8× (seed 42)	0.702	0.847	0.425	0.285	0%	30.6%
Margin + 8× (seed 123)	0.926	**0.960**	0.570	0.443	0.47%	7.6%
Selective GRPO (seed 42)	0.870	0.931	0.581	0.401	0.47%	13.3%
Selective GRPO (seed 123)	0.873	0.933	0.587	0.404	0.47%	13.0%
Margin, no replication (seed 42)	**0.985**	0.947	**0.659**	**0.530**	9.43%	**1.25%**
Margin, no replication (seed 123)	0.027	0.500	0.035	0.032	0%	100%
Margin, no replication (seed 7)	0.027	0.500	0.188	0.130	0%	100%

Table 3: Gate accuracy reverses the apparent model ranking. Every pre-intervention recipe remains near the 97.34% always-abnormal baseline while rejecting all normal wafers. Interventions expose different FA–FM operating points: replicated margin is seed-sensitive; selective GRPO has similar observed FM across two seeds. The no-replication recipe's own three seeds are the most seed-sensitive of any variant in this table: seed 42 attains the lowest measured FM among trained runs and the best Q1–Q3 scores at 9.43% FA, while seeds 123 and 7—identical recipe, only the seed differs—collapse to the opposite degenerate shortcut (FA 0%, FM 100%, BAcc at chance). No row is deployment-ready or universally optimal without a cost ratio, and no single seed of the no-replication recipe should be read as representative. ‡: the SFT baseline row is measured on the full 26,613-wafer test split (700 normal), where FA $= 100\%$ bounds Q1 EM by 0.9737; all other rows use the 7,983-wafer deployment-risk prefix (212 normal, bound 0.9734).

gives FA 1.42% / FM 52.18% and $\tau = 3.125$ gives FA 16.51% / FM 23.51%, both still dominated on both axes by the same trained runs (Appendix E.2). The baseline's representation is not cleanly separable along this axis—the trained interventions do more than move a decision boundary on a fixed one.

6.3 Closing FA Alone Produces an Unstable Boundary Shift

The replicated-margin intervention nearly eliminates FA at both seeds, confirming that the measured token gap is controllable. Its FM cost, however, differs sharply: 30.6% at seed 42 versus 7.6% at seed 123 under the identical recipe. The selective-GRPO runs show a narrower observed range (13.3% and 13.0%). Exact Clopper–Pearson 95% intervals on the fixed evaluation set separate finite-test noise from run-to-run variation: the two margin-seed FM intervals, $[29.6, 31.7]\%$ and $[7.0, 8.2]\%$, do not overlap, whereas the two selective-GRPO intervals, $[12.6, 14.1]\%$ and $[12.3, 13.8]\%$, almost coincide. These intervals quantify test-set uncertainty only; two seeds are insufficient to infer a mechanism-level variance distribution. The defensible conclusion is that a single favorable intervention run does not characterize deployment risk.

The isolation run changes one factor: it removes the $8\times$ repeated normal pool while retaining the margin objective. At seed 42, FM falls to 1.25% (95% CI $[1.0, 1.5]\%$) and Q1–Q3 all improve, while FA rises to 9.43% (95% CI $[5.9, 14.2]\%$) — on its own, a distinct low-FM operating point. A second seed of the identical recipe (seed 123) instead collapses to FA 0% (95% CI $[0, 1.7]\%$), FM 100% (95% CI $[99.95, 100]\%$): the model answers "No" on every one of the 7,983 evaluation wafers, verified against raw predictions rather than an artifact of scoring. A third seed (seed 7) collapses in the same way (FA 0%, FM 100%). The no-replication seeds therefore span the entire FA–FM tradeoff range (Fig. 3, left), a wider range than either the replicated-margin or selective-GRPO recipes show across their own paired seeds. This contradicts the hypothesis the ablation was designed to test, that removing repeated minority exposure would reduce seed sensitivity: instead it makes even the qualitative direction of the failure seed-dependent. Data replication and objective reweighting are not interchangeable remedies, and neither should be trusted from a single run, including the seed-42 result that looked like the safest point in the paper until its other two seeds landed. On a second

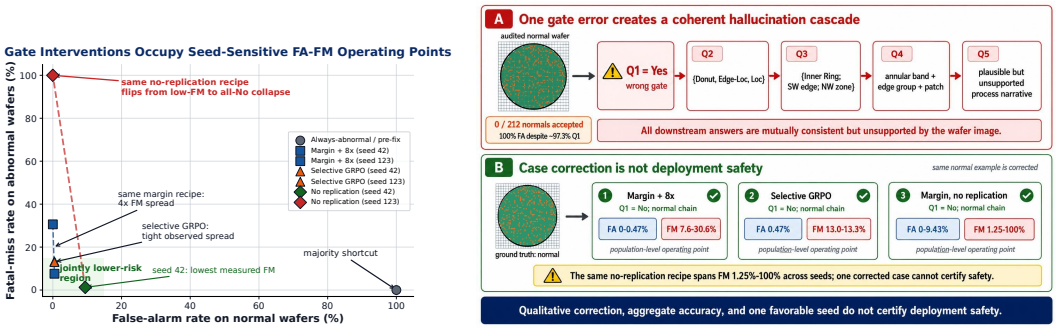


Figure 3: **Left:** FA–FM risk map for all gate interventions. The always-abnormal shortcut occupies the extreme right despite 97.34% Q1 accuracy. Paired replicated-margin and selective-GRPO seeds move left but incur different FM costs (the always-abnormal and pre-intervention rows have FM ≈ 0 by construction; "lowest" refers to trained runs). Most strikingly, the identical no-replication recipe occupies opposite extremes across seeds: seed 42 reaches FA 9.43% / FM 1.25%, the lowest FM among trained runs, whereas seeds 123 and 7 collapse to FA 0% / FM 100%. Lower-left is jointly better, but a fab-specific cost ratio and multi-seed evidence are required to select an operating point. **Right:** Hallucination-cascade and deployment-risk audit. A wrong Q1 gate on an audited normal wafer is followed, within the same self-conditioned dialogue, by a mutually consistent but image-unsupported Q2–Q5 narrative and corresponds to 0/212 normal wafers accepted. All three interventions correct the illustrated case, yet their held-out FA–FM operating points differ sharply; the no-replication recipe alone spans FM 1.25% to 100% across three seeds. Correcting one example, preserving aggregate accuracy, or reporting one favorable seed therefore does not certify deployment safety.

backbone (Qwen2.5-VL-7B; Appendix E.7) the same chain reproduces the shortcut, then holds FA and FM below 0.5% at both seeds: the frontier's severity is backbone-dependent; the audit is not.

6.4 The Gate Controls the Downstream Narrative

On normal wafers (Fig. 3, right), an erroneous "Yes" is followed in the same completion by a mutually consistent set of defect types and locations and by plausible morphology and process language; consistency makes the false narrative look trustworthy. A wrong "No" on an abnormal wafer instead collapses downstream answers into the short normal template, which can spuriously raise n-gram scores. We therefore prioritize tuple correctness over Q4–Q5 overlap metrics (example in Appendix F.2).

Across pre-intervention checkpoints, Q1–Q3 tuple correctness is only 5.8–13.1%, far below Q1 EM, and location F1 trails type F1 throughout; neither CoT nor domain-QA repairs the gate (Appendix E).

6.5 Consistency Is Not Correctness

Table 2 shows why internal agreement cannot replace tuple correctness. CoT increases Q2-to-Q4 and Q2-to-Q5 support, yet its Q1–Q3 tuple accuracy falls to 8.6%. In other words, the model becomes more consistent about a worse answer chain. Domain-QA gives the lowest tuple correctness, despite competitive type F1. The mixed-defect benchmark therefore separates three properties that are often conflated: per-stage accuracy, cross-stage agreement, and end-to-end factual correctness.

6.6 The Shortcut Floor: What the Benchmark Measures Beyond Classification

A linked-answer benchmark should be audited adversarially against itself. We therefore report four shortcut baselines on the same 7,983-wafer set (Appendix E.6): a question-only model, blank- and shuffled-image inputs, and a *classifier-to-template mapper* that predicts the 38-way class with a compact CNN and deterministically maps it to Q1–Q5 through the benchmark's own construction library. Question-only Q2 reaches 92% of the real-image VLM's Q2 F1, and the mapper exceeds the VLM on every metric (Q1 EM 1.000 with FA and FM both 0%; Q2 F1 0.977 vs. 0.507; Q3 F1 0.976

vs. 0.258; Q5 CIDEr 3.116 vs. 0.569). Two conclusions follow, and they are complementary rather than contradictory. First, the anomaly signal is fully recoverable from the image by a small classifier, so the VLM's 100% false-alarm gate is a decision-layer pathology of the language model, consistent with the token-level absence bias, not a perceptual limit of the task. Second, the current Q1–Q5 scoring protocol is highly compatible with classification plus deterministic lookup: its language metrics measure structured diagnostic-format competence, and no VLM result on this benchmark should be presented as proof of open-ended multimodal reasoning. The mapper wins precisely because the structured answers are functions of the class label, which is the same property that makes the gate the VLM's single point of failure. We release the mapper and propose it as a **mandatory floor** for wafer VQA: a multimodal method should be required to beat its own classifier-to-template control before its language scores are interpreted as reasoning. To our knowledge this control is not reported by existing industrial VQA benchmarks.

6.7 A Four-Way Separation, and Its Implications

Taken together, the experiments establish a four-way separation on a single fixed evaluation set. Aggregate Q1 accuracy is near 97.3% for every pre-intervention recipe; class-conditional risk simultaneously sits at the worst possible FA operating point; heuristic cross-stage support reaches 97–100%; and jointly correct Q1–Q3 tuples occur on only 5.8–13.1% of wafers. Each quantity can improve while another degrades—CoT raises consistency while lowering tuple correctness, and the margin fix lowers FA while raising FM—so no one of them can proxy for the others.

The failure is a model-selection failure as much as a model failure: selecting by highest Q1 EM favors the majority shortcut, and selecting by lowest FA alone favors an intervention that may clear thousands of defective wafers. A deployment-facing protocol should instead specify an acceptable cost region in the FA–FM plane, report both denominators, and select checkpoints inside that region; the same principle extends to per-label false positives and misses in multi-label Q2.

Fluent Q5 text is likewise not causal competence: a gate error yields a plausible but unsupported process narrative, and a fatal miss inflates overlap metrics by routing an abnormal wafer to the short normal template, so Q4–Q5 metrics must be read gate-conditionally.

7 Limitations

The benchmark is derived and partly simulated, lacks lot IDs and expert adjudication for Q4–Q5, and uses limited backbone/seed replication. Its protocol is largely solvable by classification plus lookup, while typed-pair scoring and unseen-combination generalization remain untested.

8 Conclusion

We proposed a deployment-risk-aware evaluation protocol for hierarchical industrial VQA and instantiated it as Mixed-WM38 VQA, released with its splits, determinism audit, shortcut floor, and evaluation code. The protocol reveals what aggregate scores conceal: a 97%-accurate gate that rejects every normal wafer and then narrates a coherent but unsupported diagnosis; an absence token whose restoration cost is seed- and backbone-dependent; and a classifier-to-template pipeline that outscores the VLM on every metric, so its language scores must not be read as reasoning. Aggregate accuracy, class-conditional risk, cross-stage consistency, and end-to-end tuple correctness measure different properties and can move in different directions; hierarchical industrial VQA should be selected with all four, the shortcut floor, and run-to-run stability, not aggregate accuracy or fluent text alone.

Reproducibility Statement

All numbers in this paper are computed over 100% of a named, fixed population (Table 1; Appendix A). The anonymous release package accompanying this submission contains the Mixed-WM38 VQA annotations, source-wafer mappings to the upstream MixedWM38 corpus, fixed train/validation/test split manifests, the deterministic construction rule and manifest of the 7,983-wafer deployment-risk evaluation set, Q4/Q5 multi-reference pools with source notes, construction and determinism-audit

scripts, the gate-margin trainer and all GRPO reward implementations, evaluation and Clopper–Pearson interval code, a datasheet, and Croissant metadata. Training hyperparameters, reward weights, and the run-inclusion rule are given in Appendix D. Where the upstream license prevents redistribution of raw wafer images, identifiers and reconstruction scripts reproduce the benchmark from the authorized source.

Ethics Statement

Mixed-WM38 VQA is a derived annotation and evaluation layer over a published, publicly available wafer-map corpus; it contains no personal data. Mixed-defect classes are partially simulated upstream, and Q5 references are literature-grounded hypotheses rather than verified fab root causes; the dataset card documents these limits and states that no reported operating point is deployment-ready. The paper's purpose is to make hidden screening risk visible; we explicitly caution against using any single favorable run or aggregate score as evidence of deployability in a safety-relevant manufacturing setting.

AI Use Statement

AI assistants (Anthropic's Claude) were used substantially in this work, under continuous author direction and review, for: (i) coding assistance — writing and debugging the dataset-construction, training, evaluation, and audit scripts released with this paper; (ii) experiment orchestration — preparing and monitoring compute jobs specified by the authors; (iii) language editing and LaTeX formatting; and (iv) drafting portions of the manuscript text and the Q4/Q5 reference-answer templates, all subsequently reviewed and edited by the authors. AI assistants were not used to generate, select, or filter any reported quantitative result (every number is the direct output of the released scripts run on the full fixed populations), to fabricate data or references, or to make research claims the authors did not verify against the underlying artifacts. The authors take full responsibility for all content of this paper.

References

Yuhao Chao, Jie Liu, Jie Tang, and Gangshan Wu. Anomalyr1: A grpo-based end-to-end mllm for industrial anomaly detection, 2025.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1932–1940, 2024.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *The Thirteenth International Conference on Learning Representations*, 2025.

Hyunju Kang, Woohyun Lee, Jaewon Kim, and Hogun Park. Judo: A juxtaposed domain-oriented multimodal reasoner for industrial anomaly qa, 2026.

Inha Kang, Youngsun Lim, Seonho Lee, Jiho Choi, Junsuk Choe, and Hyunjung Shim. What "not" to detect: Negation-aware vlms via structured reasoning and token merging, 2025.

Qian Kou, Xiaofeng Shi, Yulin Li, Xiaosong Qiu, Xinyang Wang, Hua Zhou, and Cao Dongxing. Mechvqa: Benchmarking and enhancing multimodal llms on comprehensive mechanical drawing understanding, 2026.

Kiryong Kyeong and Heeyoung Kim. Classification of mixed-type defect patterns in wafer bin maps using convolutional neural networks. *IEEE Transactions on Semiconductor Manufacturing*, 31(3): 395–402, 2018.

Yanhui Li, Yunkang Cao, Chengliang Liu, Yuan Xiong, Xinghui Dong, and Chao Huang. Iad-r1: Reinforcing consistent reasoning in industrial anomaly detection, 2025.

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. Can multimodal large language models understand spatial relations? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 620–632, 2025.

Takeshi Nakazawa and Deepak V. Kulkarni. Wafer map defect pattern classification and image retrieval using convolutional neural network. *IEEE Transactions on Semiconductor Manufacturing*, 31(2):309–314, 2018.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

Junliang Wang, Chuqiao Xu, Zhengliang Yang, Jie Zhang, and Xiaoou Li. Deformable convolutional networks for efficient mixed-type wafer defect pattern recognition. *IEEE Transactions on Semiconductor Manufacturing*, 33(4):587–596, 2020.

Zhen Wang, Zhi Fan, Shuo Tan, Yong Zhong, Yuan Yuan, Hao Li, Hao Jiang, Wei Zhang, Feng Shao, Hui Wang, and Jun Xiao. MAU-GPT: Enhancing multi-type industrial anomaly understanding via anomaly-aware and generalist experts adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26787–26795, 2026.

Ming-Ju Wu, Jyh-Shing R. Jang, and Jui-Long Chen. Wafer map failure pattern recognition and similarity ranking for large-scale data sets. *IEEE Transactions on Semiconductor Manufacturing*, 28(1):1–12, 2015.

Ke Xu et al. Wafersage: Large language model-powered wafer defect analysis via synthetic data generation and rubric-guided reinforcement learning, 2026.

A Appendix Scope and Evaluation Populations

This appendix provides dataset construction, complete evaluation tables, statistical controls, seed checks, shortcut baselines, failure analyses, and release documentation for the main text. All values are copied from completed experiment records; no missing result is estimated. Each headline metric uses 100% of its named split or fixed evaluation set. The teacher-forced token probe is a separate predefined diagnostic cohort and is not extrapolated to the full test population.

Named population	Records
Dataset / linked Q1–Q5 records	38,015 / 190,075
Training / validation / test	7,602 / 3,800 / 26,613
Deployment-risk evaluation set	7,983
Normal / abnormal risk records	212 / 7,771
Threshold-selection validation set	3,800

Table 4: Complete named populations used throughout the paper.

For rapid verification, each headline claim maps to its complete-population evidence as follows: the majority-shortcut result to Table 10 (rows 1–6) and Sec. E.2; the seed-sensitivity result to Table 10 (intervention rows) with Clopper–Pearson intervals in Sec. 6.3; consistency-versus-correctness to Table 12; the absence-token diagnosis to Sec. F.1; and the shortcut-compatibility bound to Table 15. Every entry evaluates 100% of its named population.

B The Mixed-WM38 VQA Dataset and Benchmark

B.1 Provenance and Contribution Boundary

Mixed-WM38 VQA is a new multimodal annotation and evaluation layer derived from the published MixedWM38 wafer-map corpus (Wang et al., 2020). It is not a newly collected fab-image corpus. The upstream source contains single-defect wafer maps drawn from production data and 29 mixed-combination classes supplemented with simulated or generative examples because the original WM-811K population does not contain mixed-pattern labels (Wu et al., 2015; Wang et al., 2020). Our contribution is the linked five-stage VQA annotation, image-derived location reconstruction, reference libraries, fixed splits, release metadata, construction audits, and deployment-aware evaluation package.

Raw-image redistribution remains subject to the upstream license. Where direct redistribution is not allowed, the release provides source identifiers, mappings, and reconstruction scripts. The anonymous package therefore distinguishes clearly between upstream image ownership and the new VQA annotations and evaluation resources delivered by this work.

B.2 Composition and Fixed Splits

The benchmark contains 38,015 wafer maps spanning one normal class, eight single-defect base classes, and 29 mixed combinations with up to four simultaneous patterns. Each wafer contributes five linked questions, producing 190,075 question–answer instances. The base defect vocabulary is {Center, Donut, Edge-Loc, Edge-Ring, Loc, Near-Full, Random, Scratch}. The original 38-way combination taxonomy is retained rather than rebalanced; normal wafers constitute 2.63% of the population.

On the full test split, the defect-arity distribution used for the arity audit is: 700 normal wafers, 4,913 single-defect wafers, 9,100 two-defect wafers, 9,100 three-defect wafers, and 2,800 four-defect wafers. All 38 combinations present in test also occur in training; the current benchmark therefore measures performance on a closed combination vocabulary rather than unseen-combination generalization.

B.3 Five-Stage Diagnostic Protocol

Every wafer is represented as one linked diagnostic cascade:

Dataset statistic	Value
Wafer maps	38,015
Linked Q1–Q5 records	190,075
Normal / single / mixed combinations	1 / 8 / 29
Maximum defect arity	4
Normal-class proportion	2.63%
Train split	7,602 (20%)
Validation split	3,800 (10%)
Test split	26,613 (70%)
Deployment-risk evaluation set	7,983
Normal / abnormal in risk set	212 / 7,771

Table 5: Dataset composition and fixed denominators. Splits are stratified by the 38-way combination label and are disjoint by sample identifier.

Stage	Capability	Gold representation	Primary evaluation
Q1	Anomaly gate	Yes/No defect presence	exact match, FA, FM, balanced accuracy
Q2	Defect-set recognition	set over eight base patterns	set F1, exact-set match
Q3	Spatial grounding	per-defect radial label plus an 8-sector direction where applicable	set F1, tuple correctness, determinism audit
Q4	Morphology description	class-keyed multi-reference text	ROUGE-L, SBERT, CIDEr, cross-stage support
Q5	Process-cause hypothesis	literature-grounded multi-reference text	ROUGE-L, SBERT, CIDEr, cross-stage support

Table 6: The linked five-stage protocol. Q5 is intentionally named hypothesis generation, not verified root-cause diagnosis.

The release record schema contains `sample_id`, source image mapping, 38-way `class_id`, defect arity, split assignment, Q1–Q5 gold fields, and evaluation metadata. The Croissant metadata file additionally records the image path, base defect set, and task-level field descriptions.

B.4 Construction Pipeline

Q1 and Q2. Q1 is derived from whether the upstream class is normal. Q2 is the canonical set of active base defects encoded by the upstream MixedWM38 combination label.

Q3. The initial dataset version used a static mapping from Q2 type to Q3 location, which made Q3 a deterministic relabeling of Q2. Dataset v2 replaces this shortcut with per-image pixel analysis. Undilated connected components are extracted from the failing-die map. Center, Donut, Edge-Ring, Near-Full, Random, and Normal retain physically meaningful radial or global labels. Direction-sensitive classes (Loc, Edge-Loc, and Scratch) receive an image-derived compass sector from an eight-sector angular partition. When several direction-sensitive patterns co-occur, blobs are assigned using class-specific geometry: elongation for Scratch, proximity to the wafer boundary for Edge-Loc, and the largest remaining local component for Loc.

Q4. Morphology references are assembled from a curated class-keyed library with three length/style variants per combination. The text describes observable geometry rather than process causation. The base templates were checked against the canonical wafer-map pattern definitions; mixed references compose the corresponding base descriptions.

Q5. Process-cause references are assembled from a separate class-keyed library with three variants. Each single-defect class has source notes pointing to semiconductor-process literature; mixed-class entries explicitly combine the mechanisms of their component classes. One Donut mechanism was corrected during the literature audit from a thermal/CMP explanation to photoresist redeposition during develop/rinse. Q5 remains a plausible process hypothesis because no per-wafer fab history or engineer adjudication is available.

B.5 Full-Population Integrity and Quality Audits

All structural checks below were run over the complete designated populations, not estimated from partial evaluations.

Audit	Population	Result
Split identity overlap	all 38,015 records	zero train/validation/test `sample_id` overlap; union equals the full benchmark
Image-path integrity	all 26,613 test records	every referenced image exists
Q4/Q5 reference completeness	all 26,613 test records	three references per task; no empty strings
Reference-key resolution	all 38 combinations	every class key resolves to Q4/Q5 library entries
Q3 patch integrity	all affected dataset files	intended Q3 fields changed; zero unexpected diffs
Q5 citation registry	all 8 base classes and 29 mixed keys	base classes have specific source notes; mixed keys point to component sources
Combination coverage	all train/test labels	38/38 test combinations occur in training

Table 7: Full-population dataset and release audits.

B.6 Quantified Q2-to-Q3 Determinism Audit

We group all 26,613 test wafers by their Q2 gold combination and measure the distribution of Q3 answers within each group. Before rebuilding Q3, every group had a single location answer. After rebuilding, dynamic direction labels create substantial within-class spatial variation.

Metric (record-weighted over 38 groups)	Before	After
Mean unique Q3 answers per Q2 group	1.0	47.6
Q3 entropy within a Q2 group (bits)	0.0	3.64
Conditional determinism ratio	1.000	0.280
Fully deterministic groups	38/38	8/38

Table 8: The Q3 rebuild removes the label-lookup shortcut for the direction-sensitive majority of the dataset. The remaining deterministic groups correspond to classes with fixed radial/global semantics.

The Q3 scoring and reward audit also exposed implementation errors that are retained in the experiment registry for reproducibility. Linear Track originally inspected only the largest connected component, which was often a co-occurring round defect. Using the top-three component aspect ratios raised its full-test consistency score from 0.183 to 0.817. Continuous Edge Ring and Local Edge checks were corrected to avoid contamination from a co-occurring Center pattern. A remaining limitation is that the current marginal radial/angular signature does not reliably distinguish Local Edge from Local Zone in all mixed cases; a joint two-dimensional radial–angular representation is needed for a stronger geometric audit.

B.7 Dataset Release Package and Intended Use

The release package contains the VQA annotations, source mappings, fixed split files, Q4/Q5 reference pools, Q5 source notes, construction and patch scripts, determinism-audit outputs, dataset card, Croissant metadata, reward implementations, gate-margin trainer, and evaluation code. Intended uses include mixed-defect VQA, class-conditional deployment-risk analysis, cascade evaluation, and research on negation and absence behavior. The data should not be used as evidence that a model is deployment-ready, as verified root-cause ground truth, or as newly collected production imagery. Lot identifiers are absent upstream, so a lot-disjoint split cannot be produced.

C Evaluation Protocol

C.1 Metrics

For Q1, let normal wafers be the negative class and abnormal wafers the positive class. We report

$$\mathrm{FA} = \frac{\#\{\text{normal predicted abnormal}\}}{N_{\text{normal}}},$$
$$\mathrm{FM} = \frac{\#\{\text{abnormal predicted normal}\}}{N_{\text{abnormal}}}.$$

Balanced accuracy is $1 - (\mathrm{FA} + \mathrm{FM})/2$. Q2 and Q3 use set F1 and exact-set match. Q1–Q3 tuple correctness requires all three stages to be correct jointly. Q4 and Q5 use ROUGE-L, SBERT similarity, and CIDEr against the three-reference pools. Cross-stage support checks whether claimed Q2 labels are represented consistently in Q3–Q5; it is a diagnostic of internal agreement, not factual correctness.

Aggregate gate accuracy decomposes as

$$\mathrm{Acc}_{Q1} = p_N(1 - \mathrm{FA}) + (1 - p_N)(1 - \mathrm{FM}),$$

where $p_N = 212/7{,}983 = 2.66\%$ on the deployment-risk set. Therefore an always-abnormal rule has $\mathrm{FA} = 100\%$, $\mathrm{FM} = 0\%$, and $\mathrm{Acc}_{Q1} = 97.34\%$.

C.2 Risk Uncertainty

Exact Clopper–Pearson 95% intervals are computed from the complete 212-normal and 7,771-abnormal denominators. Baseline FA is $[98.28\%, 100\%]$. Margin seed 42 has FA $[0, 1.72\%]$ and FM $[29.6\%, 31.7\%]$; margin seed 123 has FA $[0.01\%, 2.60\%]$ and FM $[7.0\%, 8.2\%]$. Selective-GRPO seeds 42 and 123 have FM intervals $[12.6\%, 14.1\%]$ and $[12.3\%, 13.8\%]$, respectively. No-replication seed 42 has FA $[5.9\%, 14.2\%]$ and FM $[1.0\%, 1.5\%]$; seeds 123 and 7 both have FM $[99.95\%, 100\%]$. These intervals quantify finite evaluation-set uncertainty and do not substitute for additional training seeds.

D Training and Reproducibility Details

D.1 SFT and Margin-Loss Training

The main experiments use Qwen3.5-4B with LoRA rank 32, $\alpha = 64$, and dropout 0.05. Standard SFT uses learning rate 8×10^{-5}, three epochs, per-device batch size 1, and gradient accumulation 16. The gate-margin intervention adds, only for normal-gold Q1 examples,

$$\mathcal{L}_{\text{gate}} = \tfrac{1}{|\mathcal{B}_N|} \textstyle\sum_{x \in \mathcal{B}_N} \max\big(0,\, m + \log P(\text{Yes} \,|\, x) - \log P(\text{No} \,|\, x)\big),$$

with margin $m = 1.0$ and weight $\lambda = 0.5$. The replicated version duplicates the 200 normal training records eightfold, raising the normal share from 2.63% to 17.77%. The no-replication isolation uses the same margin objective without duplication.

D.2 GRPO Reward Variants

For completion y, the reward family is

$$R(y) = w_f R_{\text{format}} + w_a R_{\text{acc}} + w_c R_{\text{cons}} + w_d R_{\text{dom}} + w_s R_{\text{spat}} + w_q R_{\text{q1}}.$$

R_{format} checks the five-section output schema; R_{acc} combines Q1 exact match, Q2/Q3 F1, and Q4/Q5 semantic similarity; R_{cons} penalizes Q1-to-Q2/Q3 contradictions; R_{dom} checks lexical compatibility; R_{spat} compares Q3 against the pixel-derived spatial signature; and R_{q1} fires only on normal-gold prompts.

All GRPO variants use learning rate 10^{-6}, one epoch, per-device batch size 4, gradient accumulation 4, four generations per prompt, maximum completion length 512, and KL coefficient $\beta = 0.04$. Paired-seed comparisons change only the seed.

Variant	w_f	w_a	w_c	w_d	w_s	w_q
Independent	0.5	2.0	–	–	–	–
Consistency	0.5	2.0	0.5	0.3	–	–
Full/spatial	0.5	2.0	0.5	0.3	0.5	–
Selective Q1 gate	0.5	2.0	–	–	–	1.5

Table 9: Fixed GRPO reward weights. They are controlled first-pass settings, not test-selected hyperparameters.

Checkpoint	Q1 EM	Q2 F1	Q3 F1	Q4 CIDEr	Q5 CIDEr	FA	FM
SFT baseline (v2)	0.974	0.507	0.258	0.351	0.569	100%	$\approx 0\%$
$+8\times$ replication	0.973	0.519	0.294	0.370	0.525	100%	$\approx 0\%$
Domain-QA	0.973	0.528	0.241	0.369	0.511	100%	$\approx 0\%$
GRPO independent	0.973	0.525	0.284	0.357	0.580	100%	$\approx 0\%$
GRPO consistency	0.973	0.526	0.291	0.365	0.591	100%	$\approx 0\%$
GRPO full/spatial	0.973	0.526	0.286	0.358	0.600	100%	$\approx 0\%$
Margin + $8\times$ (seed 42)	0.702	0.425	0.285	0.480	0.626	0%	30.6%
Margin + $8\times$ (seed 123)	0.926	0.570	0.443	0.658	0.892	0.47%	7.6%
Selective GRPO (seed 42)	0.870	0.581	0.401	0.748	0.926	0.47%	13.3%
Selective GRPO (seed 123)	0.873	0.587	0.404	0.730	0.939	0.47%	13.0%
Margin, no replication (seed 42)	0.985	0.659	0.530	0.799	1.078	9.43%	1.25%
Margin, no replication (seed 123)	0.027	0.035	0.032	0.085	0.112	0%	100%
Margin, no replication (seed 7)	0.027	0.188	0.130	0.163	0.242	0%	100%

Table 10: Complete results on the fixed 7,983-wafer deployment-risk set. Q4/Q5 CIDEr must be interpreted conditionally on Q1 because a fatal miss routes an abnormal wafer to the short normal template. Two of three no-replication seeds collapse to predicting normal for every wafer.

D.3 Run Inclusion Rule

Only checkpoints with a valid closed reasoning block and parsable Q1–Q5 output schema are included. Early runs affected by an open reasoning-block/chat-template bug exhausted the completion budget and are excluded based on implementation validity before metric inspection. Tool-use attempts that never learned a persistent tool call are recorded in the experiment registry but are not presented as successful model variants. The deployment-risk evaluation set is the fixed 7,983-record prefix of the released test manifest and is reused with identical sample identifiers across every comparison; the prespecified per-run checkpoint-selection rule is documented in the release scripts and is not restated here.

E Complete Experimental Results

E.1 Main Comparison

Replication alone does not alter the 100% FA shortcut. Both trained fixes move the gate away from the always-abnormal extreme, but neither yields a seed-independent deployment point. The selective-GRPO pair is close in observed FM, whereas the replicated-margin pair differs by 23 points. The no-replication recipe is the least reliable: two of three completed seeds collapse to 100% FM. The third seed therefore strengthens, rather than softens, the seed-sensitivity conclusion of Sec. 6.3—the favorable seed-42 operating point is a one-in-three outcome under the identical recipe, and no single run of this recipe should be treated as representative.

E.2 Threshold Calibration

For every validation wafer, we compute $\Delta = \log P(\text{Yes}) - \log P(\text{No})$ and classify as abnormal when $\Delta > \tau$. The sweep uses the complete 3,800-wafer validation split.

Applying the validation-selected thresholds unchanged to the complete 7,983-wafer test set yields: $\tau = 3.375$, FA 1.42% (3/212) and FM 52.18% (4,055/7,771); $\tau = 3.125$, FA 16.51% (35/212) and FM 23.51% (1,827/7,771). The thresholds transfer without pathological drift, yet every transferred operating point remains dominated by the stronger trained interventions on both axes. Calibration

τ	Validation FA (%)	Validation FM (%)
≤ 2.0	100.0	0.0
2.875	78.0	4.08
3.000	39.0	14.41
3.125 (balance)	19.0	18.57
3.250	8.0	29.27
3.375 (FA-min)	0.0	51.89
3.500	0.0	69.97
3.625	0.0	81.78
3.875	0.0	96.70
≥ 5.0	0.0	100.0

Table 11: Representative points from the complete validation threshold sweep.

generalizes but cannot reproduce the trained frontier—retraining is doing real representational work, not merely moving a decision boundary.

E.3 Cross-Stage Consistency

Checkpoint	Q1–Q3 tuple	Q2→Q3	Q2→Q4	Q2→Q5
SFT	11.9%	97.0%	69.6%	92.5%
GRPO (4-reward)	12.9%	100%	72.7%	94.1%
GRPO (+spatial)	13.1%	100%	73.2%	94.7%
CoT	8.6%	100%	87.3%	99.1%
Domain-QA	5.8%	100%	76.8%	94.5%

Table 12: Internal support is not factual correctness. CoT increases textual agreement while reducing the jointly correct tuple rate. Same rows as Table 2. All rows are first-release checkpoints scored on the first-release (v1) labels, where Q3 is a lookup of Q2, so Q2→Q3 agreement is near 100% by construction; the within-table comparison is unaffected (v2 SFT tuple rate: 0.44%).

E.4 Multi-Label Arity and Set-Size Analysis

Gold arity	n	Q2 F1	Exact-set match	Q1 EM
Normal (0)	700	0.000	0.00%	0.00%
1	4,913	0.317	11.48%	99.96%
2	9,100	0.467	8.25%	100%
3	9,100	0.592	19.80%	100%
4	2,800	0.660	2.25%	100%

Table 13: Full-test Q2 results by gold label cardinality. F1 increases mechanically with larger sets, whereas exact-set match exposes the difficulty of four-way compositions.

The baseline over-predicts the Q2 set size on 32.30% of test wafers, under-predicts it on 25.93%, and matches the correct cardinality on 41.77%; the mean predicted-minus-gold size is $+0.12$. Thus fewer than half of predictions recover even the number of active defects.

E.5 Zero-Shot Baselines

E.6 Shortcut Baselines

Question-only Q2 reaches 92% of the real-image VLM's Q2 F1, demonstrating substantial text/prior reliance. The mapper's Q1 EM of 1.000 (FA and FM both 0% on the same 7,983 wafers) additionally shows that the anomaly signal is fully recoverable from the image by a compact CNN; the VLM gate failure is therefore a decision-layer pathology of the language model, consistent with the affirmative-bias diagnosis, rather than a perceptual limit of the task. More importantly, the classical classifier-to-template mapper exceeds the VLM on every reported metric. We report this negative-control audit against our own benchmark deliberately: it establishes that the current Q1–Q5 scoring

Model	Q1 EM	Q2 F1	Q3 F1
Qwen2.5-VL-7B-Instruct	0.955	0.314	0.243
Qwen3-VL-4B-Instruct	0.972	0.415	0.238
Fine-tuned 4B baseline	0.974	0.507	0.258

Table 14: Zero-shot results on the complete 7,983-wafer deployment-risk set. FA/FM was not separately computed for these zero-shot runs and is therefore not inferred.

Metric	Question-only	Blank image	Shuffled image	Classical mapper	Real-image VLM
Q1 EM	0.973	0.973	0.973	**1.000**	0.974
Q2 F1	0.468	0.435	0.427	**0.977**	0.507
Q3 F1	0.159	0.168	0.174	**0.976**	0.258
Q4 ROUGE-L	0.387	0.373	0.357	**0.954**	0.414
Q4 SBERT	0.725	0.678	0.663	**0.983**	0.714
Q4 CIDEr	0.240	0.235	0.245	**3.081**	0.351
Q5 ROUGE-L	0.380	0.383	0.388	**0.962**	0.442
Q5 SBERT	0.596	0.571	0.586	**0.982**	0.644
Q5 CIDEr	0.304	0.366	0.373	**3.116**	0.569

Table 15: Shortcut baselines on the complete 7,983-wafer set. The classical baseline predicts the 38-way class with a compact CNN and deterministically maps that class to Q1–Q5 using the benchmark construction library.

protocol is highly compatible with classification plus deterministic lookup and should not be presented as proof of open-ended multimodal reasoning. Read together with the deployment-risk results, the two findings are complementary rather than contradictory—the classical pipeline wins precisely because the benchmark's structured answers are deterministic functions of the class label, which is the same property that makes the gate the single point of failure for a VLM. This audit is, to our knowledge, rarely reported for industrial VQA benchmarks, and we release the mapper so future work can use it as a mandatory floor: a proposed multimodal method should be required to beat its own classifier-to-template control before its language metrics are interpreted as reasoning.

E.7 Second-Backbone Generalization Check

To test whether the FA–FM frontier of Section 6.3 is a property of the evaluation protocol or of the particular 4B backbone, we repeated the full intervention chain (SFT $\rightarrow$ gate-margin loss $\rightarrow$ selective GRPO) on Qwen2.5-VL-7B-Instruct, changing only the backbone. No reward weight, loss coefficient, data split, or training hyperparameter was retuned.

Stage (Qwen2.5-VL-7B)	Q1 EM	Q2 F1	Q3 F1	FA	FM
SFT	0.9734	0.172	0.132	100%	$\approx 0\%$
+ gate-margin loss	0.9914	0.825	0.613	**0%**	0.89%
+ selective GRPO (seed 42)	**0.9955**	**0.856**	**0.629**	0.47%	**0.45%**
+ selective GRPO (seed 123)	0.9950	0.854	0.627	0.47%	0.50%

Table 16: Second-backbone check on the same complete 7,983-wafer deployment-risk set (212 normal, 7,771 abnormal) used throughout the paper. The SFT stage reproduces the majority shortcut exactly: Q1 EM equals the 0.9734 always-abnormal bound with 100% FA. Both selective-GRPO seeds then place FA and FM simultaneously below 0.5%, which no 4B configuration in Table 3 attains.

Three observations follow. First, the shortcut itself transfers: the 7B SFT stage lands exactly on the always-abnormal bound (Q1 EM $= 0.9734$, FA $= 100\%$), so the diagnosis the protocol is built to expose is not backbone-specific. Second, the gain is localized to the margin stage rather than to reinforcement learning: the margin objective alone drives FA to 0% (0 of 212 normal wafers, 95% CI $[0, 1.7]\%$) at FM 0.89% (95% CI $[0.69, 1.12]\%$), and the subsequent GRPO stage trades one false alarm for roughly half the fatal misses (FM 0.45%, 95% CI $[0.31, 0.63]\%$). Third, the two GRPO

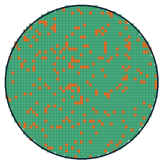
seeds are tightly clustered — identical FA (1 of 212 in both runs) and FM intervals that overlap substantially ($[0.31, 0.63]\%$ and $[0.36, 0.69]\%$) — in contrast to the 4B seed pairs of Table 3, whose FM values differ by a factor of four under the replicated-margin recipe. At matched FA (0.47%), fatal misses fall from $13.3\%/13.0\%$ on the 4B backbone to $0.45\%/0.50\%$ here, a reduction of roughly $29\times$.

All three post-SFT rows are verified non-degenerate against raw predictions rather than inferred from aggregate scores: each produces both answer classes on the gate (Yes/No counts $7{,}702/281$, $7{,}737/246$, and $7{,}733/250$ respectively), and each exceeds the 0.9734 always-abnormal bound on Q1 EM, which the always-normal and always-abnormal shortcuts cannot do.

Two limits of this check should be read with it. The two backbones differ in both parameter count and model family (Qwen3.5-4B versus Qwen2.5-VL-7B), so the improvement cannot be attributed to scale alone; a clean scaling ladder within one family is required for that claim and we do not make it. And the shortcut-floor conclusion of Section 6.6 is unchanged: at Q2 F1 0.856, the 7B chain still does not reach the classifier-to-template floor (0.977), so its language metrics remain bounded by a non-multimodal baseline. What the check does establish is that the severity of the FA–FM frontier is backbone-dependent, and that the protocol — not any single reported operating point — is what makes that distinction visible: without the class-conditional decomposition and the non-degeneracy checks, all four rows above would be indistinguishable at 97–99% aggregate Q1 accuracy.

E.8 Qualitative Examples: Native versus Trained Backbone

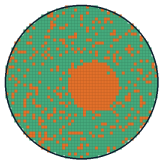
Tables 17–20 compare, on the same wafers, the native Qwen2.5-VL-7B-Instruct (zero-shot, identical five-turn protocol) with the same backbone after the full chain of Appendix E.7 (SFT → gate-margin → selective GRPO, seed 42). On the complete 7,983-wafer deployment-risk set, the strict Q1–Q3 tuple (all three stages exactly correct, sets compared as sets, directions included) is correct for 0.68% of wafers zero-shot and 30.30% after training; 2,376 wafers flip from wrong to right and 11 from right to wrong. The examples are *not* hand-picked: for each class type we show the first element of a seeded random draw (seed 0) from that type's wrong-to-right set. Correct answers are marked ✓ and incorrect ones × by the paper's own scorer; Q4 shows the first sentence of each free-text answer (unscored here). Note what the native model gets wrong: it flags clean wafers, lists defect types absent from the image, and answers Q4 with generic report templates rather than image-specific morphology.



normal_34297

	Qwen2.5-VL-7B (native, zero-shot)	Ours-7B (SFT→margin→GRPO)	Gold
Q1	Yes ×	No ✓	No
Q2	Random, Edge-Ring ×	Normal ✓	Normal
Q3	Randomly Scattered ×	None ✓	None
Q4	The wafer map displays a random scattering of red dots across the entire surface, indicating the presence of …	Observing the global map, we first notice a generally clean wafer field. …	*ref. pool*

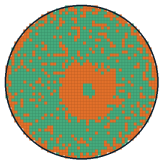
Table 17: Qualitative example 1/4 — Normal wafer: Normal. Test-split wafer map (real die layout), rendered as in Figure 2: green = pass die, red = fail die.



center_12399

	Qwen2.5-VL-7B (native, zero-shot)	Ours-7B (SFT→margin→GRPO)	Gold
Q1	Yes ✓	Yes ✓	Yes
Q2	Donut, Center, Edge-Ring ×	Center ✓	Center
Q3	Continuous Edge Ring, Inner Ring, Center ×	Center ✓	Center
Q4	The wafer map displays a pattern where red dots represent defects. …	Observing the global map, the most prominent feature is located at the wafer center. …	*ref. pool*

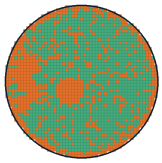
Table 18: Qualitative example 2/4 — Single defect: Center. Test-split wafer map (real die layout), rendered as in Figure 2: green = pass die, red = fail die.



donut_edge_ring_16286

	Qwen2.5-VL-7B (native, zero-shot)	Ours-7B (SFT→margin→GRPO)	Gold
Q1	Yes ✓	Yes ✓	Yes
Q2	Donut, Scratch, Edge-Ring ×	Donut, Edge-Ring ✓	Donut, Edge-Ring
Q3	Inner Ring ×	Inner Ring, Continuous Edge Ring ✓	Inner Ring, Continuous Edge Ring
Q4	The wafer map displays a central area with a distinct red coloration, which indicates a defect. …	Observing the global map, one band of defects forms a ring around the center. …	*ref. pool*

Table 19: Qualitative example 3/4 — Mixed 2-defect: Donut + Edge-Ring. Test-split wafer map (real die layout), rendered as in Figure 2: green = pass die, red = fail die.



center_edge_ring_loc_08749

	Qwen2.5-VL-7B (native, zero-shot)	Ours-7B (SFT→margin→GRPO)	Gold
Q1	Yes ✓	Yes ✓	Yes
Q2	Edge-Loc, Edge-Ring, Center, Random, Near-full ×	Center, Edge-Ring, Loc ✓	Center, Edge-Ring, Loc
Q3	Inner Ring, Local Edge, Center ×	Center, Continuous Edge Ring, Local Zone (W) ✓	Center, Continuous Edge Ring, Local Zone (W)
Q4	The wafer map displays a predominantly random distribution of red pixels, which represent defective areas. …	Observing the global map, defective dies are densely grouped in the center. …	*ref. pool*

Table 20: Qualitative example 4/4 — Mixed 3-defect: Center + Edge-Ring + Loc. Test-split wafer map (real die layout), rendered as in Figure 2: green = pass die, red = fail die.

E.9 Cost-Ratio Dominance and Prevalence Transfer (Derived)

The following analysis is derived purely from Table 10 via Eqs. (2) and (4); it introduces no new measurements. Writing $r = c_{\mathrm{FM}}/c_{\mathrm{FA}}$ and normalizing $c_{\mathrm{FA}}{=}1$, the expected per-wafer cost on the deployment-risk prior ($p_N{=}2.66\%$) is $\mathcal{R}(r) = p_N\,\mathrm{FA} + r\,(1{-}p_N)\,\mathrm{FM}$. The lower envelope over all measured operating points has exactly three regimes: the replicated-margin seed-123 point is cost-optimal for $r \lesssim 0.039$, the no-replication seed-42 point for $0.039 \lesssim r \lesssim 1.98$, and—strikingly—the always-abnormal shortcut itself for $r \gtrsim 1.98$. Under this defect-heavy evaluation prior, whenever a fatal miss costs at least about twice a false alarm, the majority shortcut is not merely accuracy-optimal but cost-optimal, which is precisely why unconstrained training gravitates to it and why the benchmark prior must not be conflated with a fab prior.

Because FA and FM are prevalence-free, the same operating points transfer to any prior by Eq. (2). At an illustrative fab-like prior of $p_N{=}90\%$ normal wafers, the ranking inverts: the always-abnormal rule falls from 97.34% to 10.0% accuracy, the always-normal collapse (no-replication seeds 123/7) rises from 2.66% to 90.0%—the mirror-image shortcut now scores the prevalence—while margin seed 123 attains 98.8%, selective GRPO 98.2–98.3%, and no-replication seed 42 91.4%. No fixed accuracy ranking survives a prior change; only the class-conditional pair (FA, FM) transfers. This derivation operationalizes the paper's central prescription: report FA and FM with their denominators, and select operating points inside a fab-specific cost region rather than by any single-prior accuracy.

F Gate Diagnostics and Failure Cases

F.1 Token-Probability Diagnostic

On a dedicated normal-wafer diagnostic cohort, the correct Q1 answer "No" has mean log probability approximately -2.97 (about 5%) versus -0.05 (about 95%) for "Yes", and is consistently the second-ranked candidate. The original result was reproduced on a larger $n = 200$ diagnostic cohort with essentially unchanged values. Sec. 4.1 states this probe over the 212 normal wafers of the deployment-risk evaluation set; the suppression statistics reported there (second-ranked "No", approximately 5% vs. 95%, $\approx$18$\times$ gap) match the cohort values reported here. This probe is teacher-forced and diagnostic; it is not a complete test-set metric or a causal image-dependence audit.

An answer-order swap leaves the suppression unchanged (4.81% correct-answer probability; 20.8$\times$ suppression), ruling out a position artifact. Inverting the question polarity so that normal requires

"Yes" raises the normal-only teacher-forced probability to 74.04%, but real generation on the complete 7,983-wafer population produces FA 0% and FM 93.96% (7,302/7,771). These controls sharpen the diagnosis of Sec. 4.1: the failure is not an artifact of answer ordering or wording, and rephrasing relocates the affirmative shortcut onto the abnormal majority rather than removing it. A prompt-level remedy that looks curative under teacher forcing can therefore be catastrophic under generation—the same aggregate-versus-deployment gap the paper documents at the gate.

F.2 Qualitative Hallucination Cascade

For normal wafer `normal_34788`, the baseline predicts Donut, Edge-Loc, and Loc; assigns Inner Ring, southwest edge, and northwest local-zone locations; describes an annular band, edge group, and internal patch; and proposes photoresist redeposition, edge-cleaning or handling effects, and localized contamination. The chain is internally coherent but unsupported because the wafer is normal. This example illustrates the mechanism; all population-level conclusions use the complete fixed evaluation sets above.

G Interpretation Boundaries

The completed evidence supports four claims. First, aggregate Q1 accuracy can select the always-abnormal shortcut. Second, cross-stage consistency does not certify factual correctness. Third, trained interventions expose a seed-sensitive FA–FM trade-off rather than a universal fix, and the derived cost analysis (Sec. E.9) shows that no single operating point—including the shortcut itself—is optimal independently of the cost ratio and prior. Fourth, the current benchmark can be solved strongly by a classifier-to-template pipeline, so its language metrics should be interpreted as structured diagnostic-format competence rather than proof of free-form causal reasoning.

The evidence does not establish fab deployment readiness, verified physical root cause, unseen-combination generalization, lot-disjoint robustness, or causal visual dependence for Q4/Q5. Genuinely incomplete analyses remain outside the claims of this appendix.

H Systems, Release, and Limitations

All reported checkpoints postdate the chat-template and completion-budget fixes described in the run-inclusion rule. The anonymous release contains annotations, fixed split manifests, source mappings, Q4/Q5 references and source notes, construction and audit scripts, reward implementations, evaluation and confidence-interval code, a datasheet, and Croissant metadata. Raw image distribution follows the upstream license.

The evidence is bounded in four ways. First, intervention training uses one 4B model family and limited seeds; the no-replication recipe collapses in two of three runs. Second, Q4/Q5 are class-conditioned references without process-engineer adjudication. Third, the classical mapper shows that current Q1–Q5 scoring is highly compatible with class prediction plus deterministic lookup. Fourth, the closed 38-combination taxonomy, missing lot identifiers, and teacher-forced token probe do not establish unseen-combination, lot-disjoint, or causal visual generalization. No reported operating point is deployment-ready.