FROM CLUES TO CAUSES: TEACHER-GUIDED REASONING AND ANALYSIS THROUGH CASCADED EVIDENCE FOR WAFER DEFECT MAP DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

A complete wafer defect diagnosis must go beyond naming a pattern: it must say where the defect sits, what it looks like, which class it belongs to, why it arose, and how recurrence is prevented. Existing methods address localization and classification, leaving root cause and prevention underexplored — partly a gap of observation, since public corpora are end-of-line bin maps that cannot attribute a pattern to a step. We work from per-operation *adder* maps — the inspections before and after one operation, overlaid — which fix the responsible step and leave the mechanism to be inferred. We propose TRACE, *Teacher-guided Reasoning and Analysis through Cascaded Evidence*, a reinforcement-learning framework that verifies the five diagnostic fields in sequence and halts verification at the first failed premise, so that no downstream conclusion is rewarded on top of a wrong one. To guide the policy without disclosing the answer, a frozen oracle-holding snapshot of that same policy returns one constrained clue for the stage that broke. TRACE then takes a second, dense signal from the same clue — the centred shift the clue induces in the policy's own logits — and uses it to precondition the verifier-driven policy gradient rather than as an auxiliary teacher-matching loss. The transformation is positive definite, so it preserves the zeros of the token-level logit gradient while amplifying the update along clue-informed directions. Neither the teacher nor its clues are present at inference. On a real-world industrial wafer corpus TRACE improves complete diagnostic-chain accuracy and root-cause discrimination between wafers that share an operation, a recipe and a defect class but arise from different mechanisms.

1 INTRODUCTION

A wafer defect map is an *adder* map: the wafer is inspected before an operation and again after it, and only the defects that were not there before are kept, so what remains is attributed by construction to that one operation. Rasterised onto the die grid, it marks which dies acquired a defect at that step, and rings, edge arcs, chords and clusters are the residue of specific mechanisms. Research on wafer maps has nonetheless concentrated on naming the pattern, and the reason is in the data: WM-811K and MixedWM38 (Wu et al., 2015; Wang et al., 2020) are end-of-line bin maps accumulated over every operation a wafer ever saw, so an edge ring on one may have been written by the etcher, the coater or the polisher and nothing in the image says which. The dominant formulation is therefore closed-set classification (Yang & Sun, 2021; 2022b), recently coupled to language under rubric-shaped rewards (Xu & Lian, 2026). The limit is not ambition but observation.

An adder map moves the question but does not answer it. Knowing the operation narrows the search to one tool and one step, but not to which mechanism inside it went wrong, and mechanisms inside one step are repaired by different people. Closing an excursion requires five answers in order: where the pattern sits, what it looks like at die resolution, which class it belongs to, which mechanism produced it, and what control change prevents recurrence. The first three state evidence, the last two commit to action. Two wafers can carry the same class added by the same operation — a *Scratch* from the same copper CMP polish, recipe and equipment type — and still have different causes: a

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.

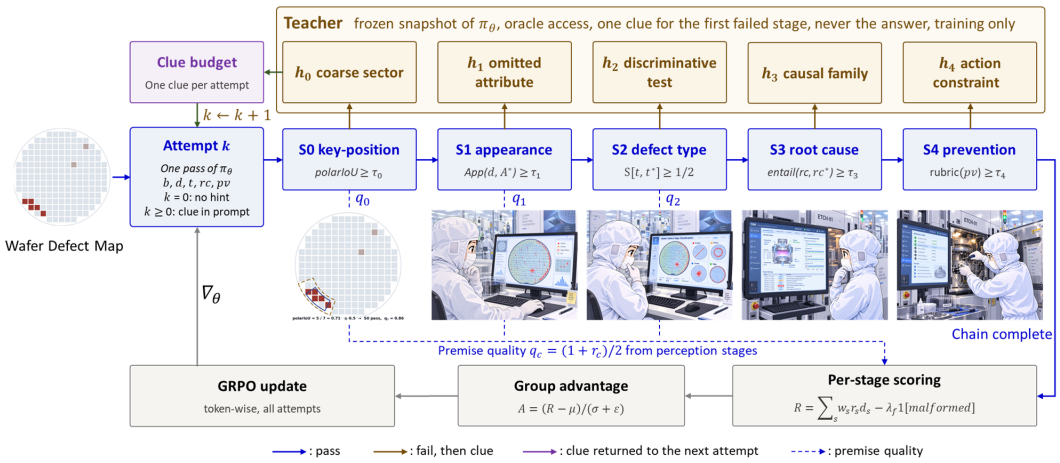


Figure 1: **TRACE at a glance.** Five fields in one pass; verification halts at the first failed gate; an oracle-holding snapshot of the policy returns one clue for that stage, never the answer.

shed conditioning-disc diamond, or an agglomerated slurry particle. Only how the damage lies on the wafer separates them, a long continuous arc against short broken traces, and the repairs separate with it, so a model that maps pattern to cause by lookup is confidently wrong on one of the two (Appendix B). Outcome-level reinforcement learning with verifiable rewards (Shao et al., 2024; DeepSeek-AI, 2025) compounds this by paying a rollout for a cause that reads plausibly whether or not the model located the defect.

We address the data and the learning side together. Our corpus pairs each adder map with the process context an engineer consults — technology, part, stage, operation, equipment type and recipe — and with five expert-annotated fields (Appendix B). Three are stored in a form the policy never emits, so the reference is independent of how an answer is written: position as a die set rather than a geometric primitive (Appendix J); appearance as morphological attributes annotated from the map alone; and root cause as one of the mechanisms the operation admits, beside the engineer's wording.

We propose TRACE, *Teacher-guided Reasoning and Analysis through Cascaded Evidence*, which scores the diagnosis stage by stage along position → appearance → type → root cause → prevention and stops at the first failure, since grading a cause on top of a wrong location produces noise rather than signal; Fig. 1 places one training step in full. A frozen snapshot of the policy, holding the oracle, returns one constrained clue for the stage that broke — a coarse annulus sector, an unasserted visual attribute, a discriminative criterion, a mechanism family, or a constraint on the action space — never the target answer. The clue carries a second signal its words leave unused: conditioning the *policy* on it shifts its own next-token logits at every prefix, densely and at no annotation cost. Matching that shift, as on-policy self-distillation does (Zhao et al., 2026), makes a privileged conditional an optimisation target and suppresses the forking deliberation depends on (Kaur et al., 2026). We instead enter it as a *direction* (Park et al., 2026), amplifying only the component of the student's own verifier-driven gradient that lies along it: the map is positive definite, so the token-level stationary points stay put, and because the direction is built from the clue rather than the oracle, one constraint bounds both channels. This separates TRACE from teacher-assisted repair, which inserts the missing evidence itself (Huang et al., 2026; Yu et al., 2026); every number we report is clue-free, since a signal obtained under hints need not improve the unhinted policy (Xia et al., 2026).

Contributions. (i) TRACE: a cascade of five verified stages that halts at the first failed premise, a constrained clue for the stage that breaks, and a second, dense clue from the same teacher entering optimisation as a preconditioner rather than a target. (ii) A benchmark scored by complete-chain consistency and counterfactual root-cause reasoning, which measures what defect classification cannot express. (iii) A die-level position reference that privileges no output format, so boxes, masks and annular sectors are scored against one common target. (iv) A corpus of per-operation adder maps in which the responsible operation is fixed by construction and the cause is annotated at the mechanism level, its appearance layer collected in a separate pass with class and cause withheld.

2 Related Work

Wafer inspection has been posed almost exclusively as pattern classification, through self-proliferating (Yang & Sun, 2021), attention-based (Yang & Sun, 2022b) and classical-quantum hybrid (Yang & Sun, 2022a) architectures, with generative modelling extending it past recognition (Yang & Hsiao, 2025); the reference corpora (Wang et al., 2020; Wu et al., 2015) supply pattern labels and nothing about the process behind them, and WaferSAGE, the first to couple maps with language and reinforcement learning, synthesizes its morphology and root-cause supervision rather than tracing it to an engineer, putting the ceiling at the generator rather than the fab (Xu & Lian, 2026). Industrial inspection more broadly has moved toward multimodal understanding — large-scale image-text supervision (Ni et al., 2026), richer sensing (Zhu et al., 2025), zero-shot anomaly description (Xu et al., 2025), process context in the visual encoder (Li et al., 2025b), retrieval-grounded explanation (Zhang et al., 2026), reasoning benchmarks (Yao et al., 2026; Li et al., 2025a) and a parallel line on localization and recognition (Qu et al., 2026; Lee et al., 2026; Ma et al., 2025a; He et al., 2025; Ma et al., 2025b; Qu et al., 2025; Hicsonmez et al., 2026; Zhang et al., 2025; Gu et al., 2025) — yet all score whether an anomaly is present, where it lies, or whether a description is faithful; none commits to a verified cause or a control action. On the learning side, privileged teachers almost always enter through the loss: self-distillation matches a privileged conditional (Zhao et al., 2026), which suppresses self-correction at long budgets (Kaur et al., 2026); GRPO-style objectives have reached anomaly interpretation and localization (Li et al., 2026; Yao et al., 2026; Jin et al., 2026); and hint-based variants replace the answer so all-fail groups still carry gradient (Yu et al., 2026; Huang et al., 2026), with hint reliance governing transfer to the hint-free policy (Xia et al., 2026) — all making the teacher a second optimisation target. On-policy reverse distillation instead takes a teacher's policy shift only as a direction and amplifies the aligned component of the student's own gradient (Park et al., 2026), closest in form to gradient surgery for multi-task learning (Yu et al., 2020) except that nothing is removed; TRACE adopts that entry point but not its source: rather than a pre-/post-training checkpoint pair, whose difference is dominated by the chosen reference, the direction is the centred logit gap between two conditionings of the *same* weights at the same step.

3 Method

3.1 Problem Formulation and the Evidence Cascade

A wafer map $\mathbf{W}$ is an adder map on a die grid of $R \times C$ cells, each labelled *added*, *clean* or *no-die*, with the die-space centre (c_r, c_c) and the wafer radius ϱ in dies, which normalises the radial coordinate to $[0, 1]$. Process context m carries technology, part, stage, operation, equipment type and recipe — fields that are not nested, since one operation recurs at several stages and one stage can be qualified on more than one equipment type — while lot and wafer identifiers and individual equipment identities stay out of the policy's context. Given $(\mathbf{W}, m)$ and a query q the policy emits

$$y = (b,\ d,\ t,\ rc,\ pv), \qquad b = \left\{(r_j^-, r_j^+, \theta_j^-, \theta_j^+)\right\}_{j=1}^{k}, \quad 1 \le k \le 4, \tag{1}$$

where d, rc, pv are free text, t is a defect class and b is the defect region written as k annular sectors. The supervision is $y^* = (M^*, A^*, t^*, rc^*, pv^*)$, and three fields are stored in a form the policy never emits: b rasterises to a die set $M(b)$ compared against the annotated M^*; A^* is a set of morphological attribute–value pairs, not a sentence, annotated from $\mathbf{W}$ and M^* alone; and $rc^* = (\omega^*, \text{text})$ pairs a mechanism label from the few the operation admits with the engineer's wording (Appendix B).

The operation that wrote the map is given in m, so S3 asks which of that operation's mechanisms is responsible, not which step. The five fields do not carry equal evidential status: position, appearance and type jointly constrain the mechanism, which the type alone leaves underdetermined. TRACE therefore verifies them as a chain — S0 position, S1 appearance, S2 type, S3 root cause, S4 prevention — and stops at the first stage whose score falls below its threshold, as shown in Fig. 2. Writing F for that index, every $s > F$ scores $r_s = 0$, and a rollout that fails nowhere has $F = 4$. All five fields are emitted in one response: verification short-circuits, generation does not. Short-circuiting is a measurement decision, not an efficiency trick: once it is enabled, every r_s entering the reward was measured under a premise that had itself cleared its threshold, and Section 3.5 extends the argument

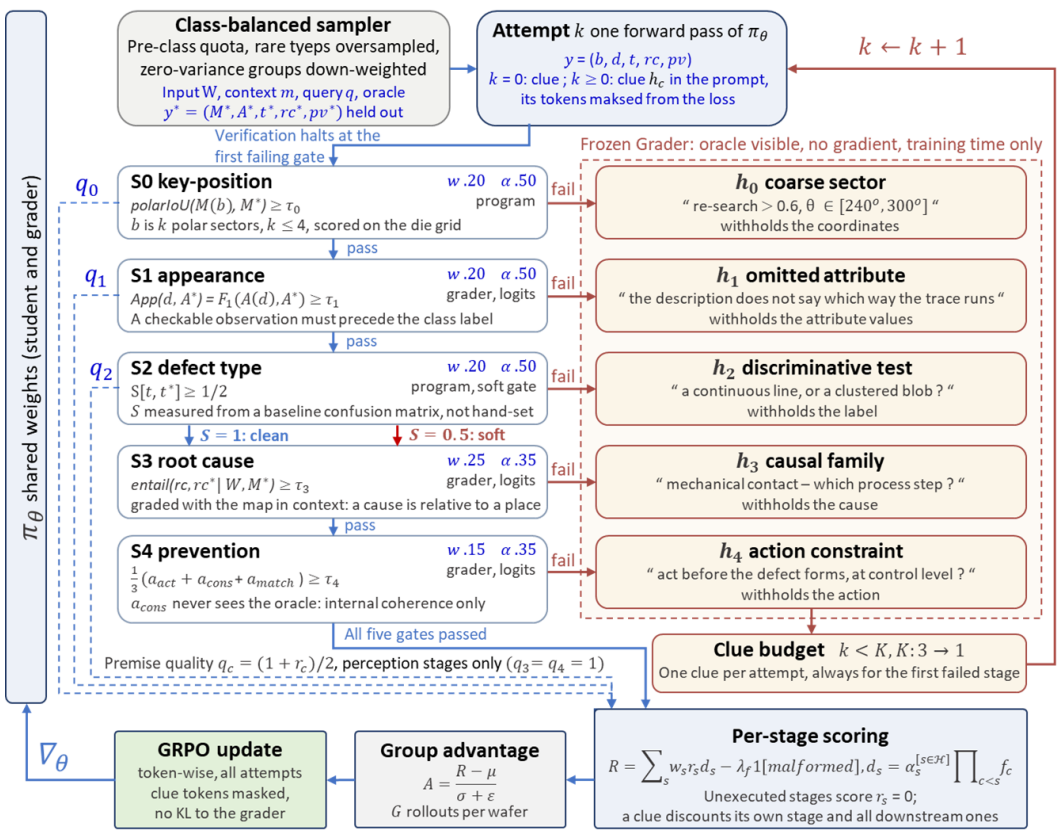


Figure 2: **The TRACE training loop in full.** Per-gate thresholds, stage weights w_s and clue discounts α_s; every stage below the break is left unverified and score $r_s = 0$ (Appendix C).

to the gradient. Placing appearance before type forces a *checkable observation* before a class may be named, so the label cannot be reached through a global prior. S2 is the single soft gate, resolved by a class-similarity matrix S measured from a baseline confusion matrix: $S[t, t^*] = 1$ is a clean pass, 0.5 a pass that propagates reduced premise quality, 0 a break (Appendix C).

3.2 Polar Grounding and the Multi-Sector Penalty

S0 is scored at die resolution. Each die is mapped to normalised polar coordinates about the grid centre, belongs to $M(b)$ when it falls inside any predicted sector, and the overlap is a set ratio,

$$\text{polarIoU} = \frac{|M(b) \cap M^*|}{|M(b) \cup M^*|}. \tag{2}$$

No geometric integration is performed, so sector intersection, angular wrap and overlapping annuli all disappear. Chordal scratches are the one pattern a single sector cannot enclose, and allowing several solves this only if the cost multiplies the overlap rather than being deducted from it,

$$r_0 = \text{polarIoU} \cdot \big(1 - \lambda_k(k-1)\big), \qquad \lambda_k = 0.08. \tag{3}$$

An additive penalty does not survive the weighting: the gain from a better primitive enters R attenuated by w_0 while the cost does not, so the argmax collapses to $k = 1$. Scratches crossing the wafer centre are flagged from the ground truth and S0 recorded as not applicable (Appendix E).

3.3 Stage Verification and Constrained Clues

S1, S3 and S4 are graded by the policy itself: the grader $\bar{\pi}_\theta$ is a snapshot of π_θ frozen and refreshed every N optimiser steps, and its only privilege is sight of the oracle. Scores are read from the logits rather than generated, as $s = P(\texttt{Yes})/(P(\texttt{Yes}) + P(\texttt{No}))$, which is continuous and immune to

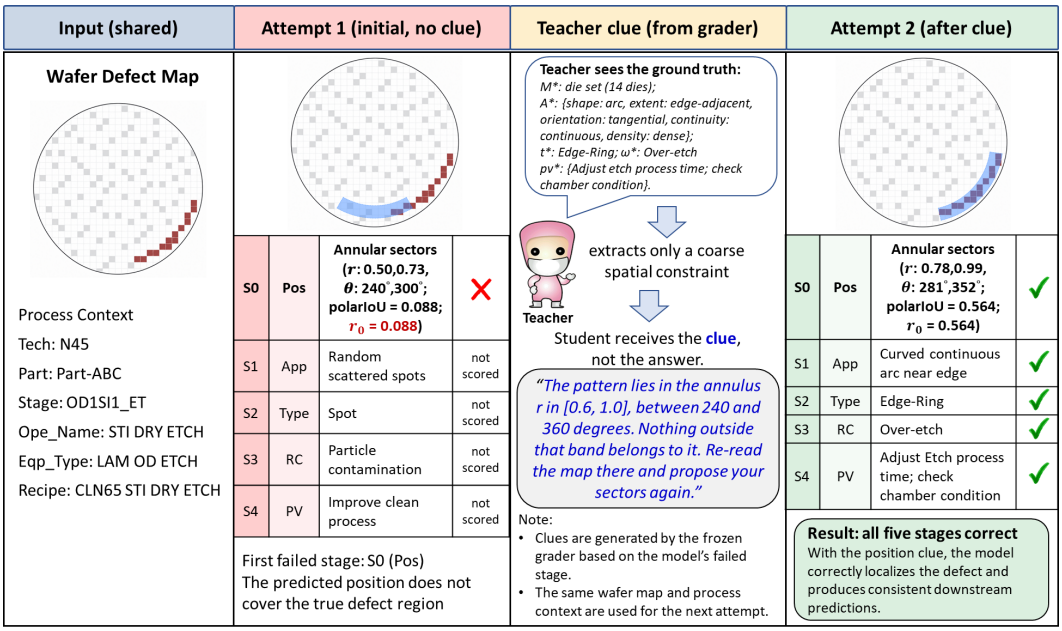


Figure 3: **One clue, one correction.** S0 fails and the stages below go unscored; the grader returns only a quantised annulus, and attempt 2 re-reads the map inside it (Appendix I).

parsing failure. S1 asks one such question per candidate (attribute, value) pair rather than whether two descriptions are equivalent, and r_1 is the F_1 of the set $A(d)$ so recovered against A^*: omitting an attribute costs recall, a wrong value costs both, and asserting one A^* marks inapplicable costs precision, so invented morphology is penalised and naming the class earns nothing. S3 is asked once per admissible mechanism whether the stated cause is that one, with $\mathbf{W}$ and M^* in context, and r_3 is the score on ω^*: grading against a label rather than the engineer's sentence keeps wording out of the reward. r_4 averages actionability, consistency against the cause *the model itself stated* — the one term never shown the oracle — and agreement with pv^* (Appendix F).

When the chain breaks at stage c, the grader returns a clue h_c that narrows that stage's search space and names no point inside it: a coarse annulus at S0, the attributes of A^* left unasserted at S1, the discriminative comparison at S2, the mechanism's family at S3, and a constraint on the action space at S4. Writing y_c^* for the sub-answer, the design constraint is

$$0 \;<\; I\big(y_c^*;\, h_c \mid \mathbf{W}, q\big) \;<\; H\big(y_c^* \mid \mathbf{W}, q\big), \tag{4}$$

the left inequality demanding that the clue be useful and the right that it leave the sub-answer unidentified, with both entropies over the annotation distribution given the map. It is a criterion we verify rather than prove, except for h_1, which ranges over attribute identities and never values, so at least two assignments survive any clue it emits; it is also the line separating a clue from repair, which inserts the missing evidence itself (Huang et al., 2026). At most one clue is issued per attempt and K per rollout, so a rollout runs at most $K+1$ attempts, and clue tokens are excluded from the loss. No divergence penalty against the grader appears in the objective, since pulling the policy toward an oracle-conditioned distribution suppresses self-correction (Kaur et al., 2026).

3.4 Stage-Wise Reward with Premise Propagation

A rollout is scored stage by stage, with weights $(w_0, \ldots, w_4) = (.20, .20, .20, .25, .15)$ and $\mathcal{H}$ the set of stages that received a clue:

$$R = \sum_s w_s\, r_s\, d_s - 0.15 \cdot \mathbf{1}[\text{malformed}], \qquad d_s = \alpha_s^{[s\in\mathcal{H}]} \prod_{c<s} f_c, \tag{5}$$

$$f_c = \min(\alpha_c, q_c) \text{ if clued and marginal;} \quad \alpha_c \text{ if clued;} \quad q_c \text{ if marginal;} \quad 1 \text{ otherwise,} \tag{6}$$

where $q_c = (1 + r_c)/2$ for the perception stages and 1 for S3, *marginal* means passed but not cleanly ($S[t, t^*] = 0.5$ at S2, within $\delta = 0.1$ of the threshold at S0 and S1), and the discounts start

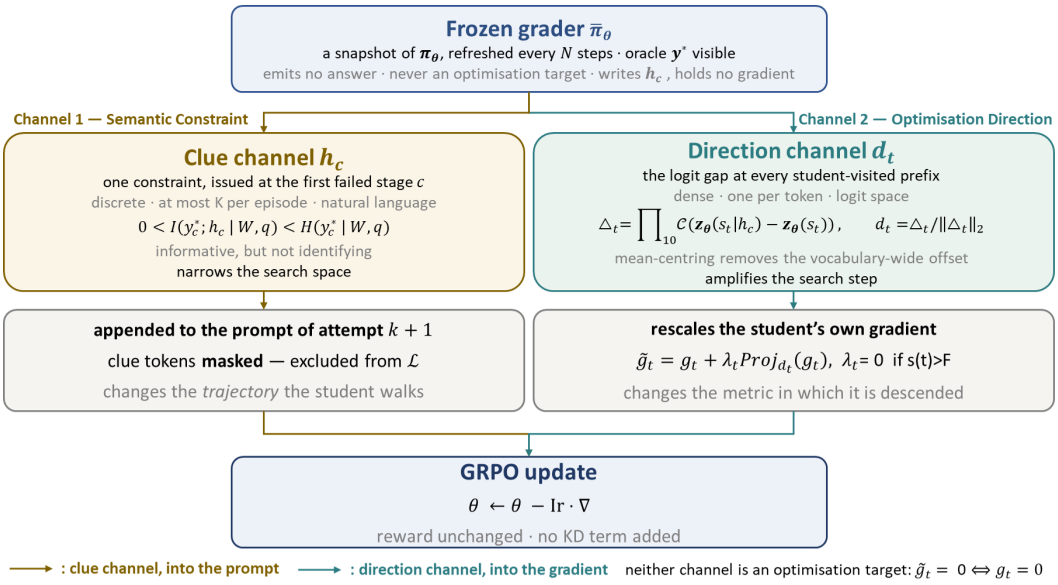


Figure 4: **Two clues, one teacher.** The same clue edits the next attempt as text and, as a conditioning of the policy's own logits, rescales its gradient; neither channel is an optimisation target.

at $\alpha = 0.50$ for perception and 0.35 for reasoning. The product runs over $c < s$ *strictly*, since stage c's own imperfection is already in r_c, and a clued and marginal stage takes the minimum rather than the product. Table 6 shows the consequence: a rollout that grounds every perception stage barely and then names a correct cause keeps about half the credit of one that grounds them cleanly, which a single global α cannot express.

3.5 The Privileged Direction Channel

The two channels answer different questions. The clue says *where to look again* and acts only through the continuation the policy then samples; the direction says *how far to move along what the clue implies* and acts on the gradient, with no teacher distribution as the thing to imitate.

The grader reaches the student through one narrow port: a short piece of text, at most K per rollout. But conditioning the policy's *own* weights on that text changes the distribution it assigns to every token, and that difference is dense, free and so far unused, as Fig. 4 shows. The grader writes h_c; both logit terms below are the policy's, so no second model is resident. Proofs are in Appendix A.

The privileged logit shift. Let $z_\theta(s_t)$ be the next-token logits at a student-visited prefix s_t and $z_\theta(s_t \mid h)$ the same logits with a privileged context h inserted. With $\mathcal{C}(v) = v - \frac{1}{|\mathcal{V}|}(\mathbf{1}^\top v)\mathbf{1}$,

$$\mathbf{\Delta}_t = \Pi_{10}\, \mathcal{C}\big(z_\theta(s_t \mid h) - z_\theta(s_t)\big), \qquad \mathbf{d}_t = \mathbf{\Delta}_t / \|\mathbf{\Delta}_t\|_2, \tag{7}$$

where h is the clue the current attempt carries and Π_{10} keeps the ten coordinates of largest student probability; on the first attempt of an episode there is no clue, and there $\lambda_t = 0$, as it is on the measure-zero event $\mathbf{\Delta}_t = \mathbf{0}$. Both terms come from the same weights at the same step and differ only in conditioning, so length preference, register and formatting cancel to first order — which distinguishes Eq. (7) from a shift between a post-trained checkpoint and a pre-training reference (Park et al., 2026), which needs two resident models and a choice of reference that the construction itself does not settle.

Proposition 1 (centred shift as a pointwise likelihood)**.** *If the model defines a joint distribution over* (s_t, y_t, h)*, then for every* v *retained by* Π_{10}*,* $\mathbf{\Delta}_t(v) = \mathcal{C}\big(\log p_\theta(h \mid s_t, y_t = v)\big)$.

Mean-centring is therefore not cosmetic. After it, $\mathbf{\Delta}_t$ is, on the coordinates Π_{10} retains, the model's own log-likelihood of the withheld information given that it emits v. Centring removes every term constant in v, so there it is the model's own pointwise mutual information $I_\theta(y_t = v; h)$, under the joint the proposition assumes. $\mathbf{d}_t$ points toward the tokens that most raise the model's belief in

the constraint the grader supplied — a step-level signal in a setting where process-level labels are prohibitive to collect.

Directional rescaling. Let $\mathbf{g}_t$ be the logit gradient of the surrogate of Section 3.6, equal to $A_i \nabla_{z_t} \log \pi_\theta(\hat{y}_t \mid s_t)$ outside the clipped region, and $u_t = \mathbf{d}_t^\top \mathbf{g}_t$. We replace it by

$$\widetilde{\mathbf{g}}_t = \big(\mathbf{I} + \lambda_t \mathbf{d}_t \mathbf{d}_t^\top\big)\mathbf{g}_t = (1+\lambda_t)\,\mathrm{Proj}_{\mathbf{d}_t}(\mathbf{g}_t) + \mathbf{g}_t^\perp, \qquad \lambda_t \geq 0, \tag{8}$$

and back-propagate $\widetilde{\mathbf{g}}_t$ in place of $\mathbf{g}_t$; both $\mathbf{d}_t$ and λ_t are constants with respect to θ.

Proposition 2 (stationarity and first-order gain)**.** *Let J collect the per-token objective of one response, $\mathbf{g}_t = \nabla_{z_t} J$, $\|\mathbf{d}_t\|_2 = 1$ and $\lambda_t \geq 0$. Then $\widetilde{\mathbf{g}}_t = \mathbf{0}$ for all t iff $\nabla_z J = \mathbf{0}$, and, with L the smoothness constant of J in z and $\bar{\lambda} = \max_t \lambda_t$, for any η with $\eta L(1+\bar{\lambda}) \leq 1$,*

$$J(z_{k+1}) - J(z_k) \;\geq\; \tfrac{\eta}{2}\|\nabla_z J(z_k)\|_2^2 + \tfrac{\eta}{2}\textstyle\sum_t \lambda_t u_t^2. \tag{9}$$

Two consequences matter. The token-level stationary points of Eq. (11) are unchanged, so the dense channel cannot move that set onto the oracle-conditioned distribution — the objection to privileged distillation (Kaur et al., 2026; Zhao et al., 2026) — and the no-KL principle holds by construction. The claim is about the zeros of the token-level logit gradient, not about parameter-space stationary points or the trajectory: $\mathbf{d}_t$ moves with the policy. And the added term depends on u_t only through its square, so past the warm-up a token whose verifier gradient *opposes* the privileged direction gains as much as one that agrees: amplification is not agreement.

Gating by the cascade. The advantage A_i is episode-level and multiplies every trainable token, including those of stages $s > F$ the cascade has declared unevaluated. They carry credit the reward has already ruled meaningless, and early in training most rollouts break at S0 or S1, so amplifying them would magnify exactly that. We therefore set

$$\lambda_t = \lambda\, \nu_{s(t)}\, \psi_t \;\text{ if } s(t) \leq F,\;\; 0 \text{ otherwise}; \qquad \psi_t = 1 \text{ if } u_t \geq 0,\;\; \min\!\big(\tfrac{n}{N_\mathrm{w}}, 1\big) \text{ otherwise}, \tag{10}$$

with $\nu_s \in [0,1]$ a per-stage strength, n the optimiser step and N_w a warm-up horizon. The first case carries the conditional-validity argument of Section 3.1 from the reward into the gradient: a stage that was not evaluated neither scores nor steers. The second damps the negative branch early, since $\mathbf{\Delta}_t$ also carries reward-irrelevant structure that amplifying a single sign of u_t would let accumulate. Because $\lambda_t \geq 0$ throughout, Proposition 2 holds.

Proposition 3 (the direction inherits the clue's bound)**.** *Let $\mathcal{E}$ collect everything $\mathbf{d}_t$ is computed from besides the clue. Then wherever the channel is active, $I(y_c^*; \mathbf{d}_t \mid \mathcal{E}) \leq I(y_c^*; h_c \mid \mathcal{E})$.*

Proposition 3 is why the direction is built from h_c and not the oracle; the oracle-conditioned variant is an ablation. The bound is on information transmitted, not acceleration obtained: the gain is quadratic in the alignment with a gradient the verifier already produced, so the channel supplies none of its own. The cost is one extra forward pass with h in context and no second resident model.

3.6 Optimisation and Schedule

Rollouts are grouped by wafer and optimised with GRPO. For G rollouts the advantage is $A_i = (R_i - \mu)/(\sigma + \varepsilon)$ and the objective is the clipped surrogate over the policy tokens of every attempt,

$$\mathcal{L} = -\frac{1}{\sum_i |\tau_i \setminus \mathrm{clue}|} \sum_i \sum_{u \in \tau_i \setminus \mathrm{clue}} \min\!\big(\rho_u A_i,\; \mathrm{clip}(\rho_u, 1\pm\epsilon) A_i\big) - \eta\, \mathbb{H}[\pi_\theta], \tag{11}$$

with $\rho_u = \pi_\theta(u)/\pi_{\theta_\mathrm{old}}(u)$, clue tokens masked and no divergence term against the grader. The $\mathbf{g}_t$ of Eq. (8) is the logit gradient of Eq. (11); the loss itself is not modified. A five-stage chain collapses at the start of training, when nearly every rollout breaks at S0 and every reward in a group is equal, so a four-phase schedule (Table 7) warm-starts with supervised fine-tuning and disables short-circuiting in Phase 1, which is safe because premise propagation is already the soft form of the same mechanism. The transition is measured rather than scheduled, on the shadow reward short-circuiting *would* have produced (Appendix H).

4 EXPERIMENTS

Every number is measured *clue-free* with a *single* attempt: no clue and no oracle enter the policy's context and neither channel is active, so the frozen snapshot appears only as the evaluator. Subscripts are standard deviations over three training seeds on `test_time`; the prompting rows are deterministic and run once, subscripted 0.0^*; n/a marks a metric undefined for that method. Splits run over lots along the time axis with a two-week embargo, since a wafer-level split leaks equipment and recipe. **Chain@5**, the fraction of wafers clearing all five gates, is the primary metric; **CF-RC**, mechanism accuracy on wafers matched on the whole process context but differing in mechanism, is the primary evidence that the model reads the map and not the operation's prior (Appendix I). Decoding is greedy, every wafer is answered once, and every trained row shares one backbone, corpus, prompt and update budget.

Positioning, paradigms and components (Tables 1 and 2). The upper block of Table 1 establishes the gap: those formulations classify, at best localise, or look a cause up from a class they are handed, so Chain@5 is undefined for them, and that column of n/a is the contribution rather than a hole. TRACE reaches 48.6 Chain@5 and 69.3 CF-RC, 13.9 and 13.7 points above GRPO (Shao et al., 2024) and 7.6 and 9.3 above `OPSD-KL` (Zhao et al., 2026), which makes the identical signal a matching target. It clears OPRD (Park et al., 2026), which takes a checkpoint shift as a direction, by 4.3 and 4.2, and it clears TRACE $-$ dir., which keeps the clue and drops the direction, by 4.6 and 4.7; the two controls fall together, at 44.3 and 44.0 Chain@5, so neither channel accounts for the result alone. In Table 2b the largest fall is A2, which keeps the clue but blinds its writer to the oracle: Chain@5 drops 13.2 points, further than A1, which replaces every clue with the missing sub-answer. What the clue channel spends is the teacher's privilege, not the act of hinting. The lookup control is the row worth pausing on: handed the class, it reaches RC 72.4, three points below ours, then falls to CF-RC 47.8, a drop of 24.6 against our 21.5 that measures how much of a root-cause score the operation's prior supplies free.

Method family	Ref.	Fields emitted					Pos	Type	RC	Chain@5	CF-RC
		b	*d*	*t*	*rc*	*pv*					
CNN pattern classifier	Wu et al. (2015)	–	–	✓	–	–	n/a	$86.2_{\pm 0.8}$	n/a	n/a	n/a
ResNet / ViT fine-tune	He et al. (2016); Dosovitskiy et al. (2021)	–	–	✓	–	–	n/a	$89.4_{\pm 0.6}$	n/a	n/a	n/a
Attention and mixed-type CNNs	Yang & Sun (2022b); Wang et al. (2020)	–	–	✓	–	–	n/a	$\mathbf{90.1}_{\pm 0.5}$	n/a	n/a	n/a
Self-supervised pretrain + probe	Oquab et al. (2023)	–	–	✓	–	–	n/a	$87.3_{\pm 0.9}$	n/a	n/a	n/a
Joint segmentation + classification	Ronneberger et al. (2015)	✓	–	✓	–	–	$\mathbf{86.5}_{\pm 0.7}$	$88.2_{\pm 0.6}$	n/a	n/a	n/a
Detector (box) + classifier	Carion et al. (2020); Zhu et al. (2021)	✓	–	✓	–	–	$77.8_{\pm 1.0}$	$86.4_{\pm 0.8}$	n/a	n/a	n/a
Rule system, oracle type	—	–	–	–	✓	✓	n/a	n/a	$72.4_{\pm 0.0^*}$	n/a	$47.8_{\pm 0.0^*}$
General VLM, zero-shot	Bai et al. (2025)	✓	✓	✓	✓	✓	$44.3_{\pm 0.0^*}$	$66.1_{\pm 0.0^*}$	$32.4_{\pm 0.0^*}$	$5.2_{\pm 0.0^*}$	$28.6_{\pm 0.0^*}$
General VLM + few-shot CoT	Bai et al. (2025)	✓	✓	✓	✓	✓	$51.6_{\pm 0.0^*}$	$71.3_{\pm 0.0^*}$	$38.7_{\pm 0.0^*}$	$9.8_{\pm 0.0^*}$	$32.1_{\pm 0.0^*}$
WaferSAGE-style rubric RL	Xu & Lian (2026)	✓	✓	✓	✓	✓	$58.2_{\pm 1.3}$	$77.5_{\pm 1.0}$	$48.6_{\pm 1.5}$	$17.4_{\pm 1.2}$	$40.2_{\pm 2.5}$
Domain VLM, SFT on our corpus	—	✓	✓	✓	✓	✓	$69.4_{\pm 1.1}$	$86.3_{\pm 0.7}$	$63.8_{\pm 1.3}$	$31.5_{\pm 1.4}$	$53.1_{\pm 2.0}$
TRACE (ours)	—	✓	✓	✓	✓	✓	$75.2_{\pm 0.9}$	$88.1_{\pm 0.5}$	$\mathbf{75.4}_{\pm 1.0}$	$\mathbf{48.6}_{\pm 1.2}$	$\mathbf{69.3}_{\pm 1.9}$

Table 1: **Comparison with prior methods.** ✓ marks a field emitted natively, n/a a metric the output leaves undefined, 0.0^* a row with no seed to vary; every row is trained on our adder corpus (Appendix I).

Method	Pos	App	Type	RC	PV	Ch@5	CF
Zero-shot VLM	$44.3_{\pm 0.0^*}$	$47.5_{\pm 0.0^*}$	$66.1_{\pm 0.0^*}$	$32.4_{\pm 0.0^*}$	$31.2_{\pm 0.0^*}$	$5.2_{\pm 0.0^*}$	$28.6_{\pm 0.0^*}$
+ SFT	$69.4_{\pm 1.1}$	$75.1_{\pm 1.2}$	$86.3_{\pm 0.7}$	$63.8_{\pm 1.3}$	$66.2_{\pm 1.5}$	$31.5_{\pm 1.4}$	$53.1_{\pm 2.0}$
+ GRPO	$70.3_{\pm 1.2}$	$76.4_{\pm 1.3}$	$86.9_{\pm 0.8}$	$66.1_{\pm 1.5}$	$68.1_{\pm 1.6}$	$34.7_{\pm 1.6}$	$55.6_{\pm 2.3}$
+ OPD, full answer	$71.1_{\pm 1.0}$	$77.9_{\pm 1.2}$	$87.5_{\pm 0.7}$	$67.8_{\pm 1.6}$	$69.3_{\pm 1.4}$	$36.2_{\pm 1.7}$	$55.2_{\pm 2.4}$
+ FFR-style	$71.8_{\pm 1.1}$	$78.0_{\pm 1.4}$	$87.6_{\pm 0.8}$	$69.3_{\pm 1.4}$	$71.4_{\pm 1.5}$	$39.8_{\pm 1.5}$	$58.7_{\pm 2.1}$
+ OPSD-KL	$72.2_{\pm 1.0}$	$79.1_{\pm 1.2}$	$87.7_{\pm 0.6}$	$70.2_{\pm 1.5}$	$72.2_{\pm 1.4}$	$41.0_{\pm 1.4}$	$60.0_{\pm 2.0}$
+ OPRD	$73.3_{\pm 1.1}$	$80.4_{\pm 1.0}$	$87.9_{\pm 0.7}$	$72.8_{\pm 1.2}$	$73.5_{\pm 1.3}$	$44.3_{\pm 1.5}$	$65.1_{\pm 1.8}$
TRACE $-$ dir.	$74.1_{\pm 1.0}$	$80.6_{\pm 1.1}$	$87.8_{\pm 0.6}$	$72.3_{\pm 1.4}$	$73.9_{\pm 1.5}$	$44.0_{\pm 1.4}$	$64.6_{\pm 1.9}$
TRACE	$\mathbf{75.2}_{\pm 0.9}$	$\mathbf{82.3}_{\pm 1.2}$	$\mathbf{88.1}_{\pm 0.5}$	$\mathbf{75.4}_{\pm 1.0}$	$\mathbf{76.8}_{\pm 1.2}$	$\mathbf{48.6}_{\pm 1.2}$	$\mathbf{69.3}_{\pm 1.9}$

(a) Main comparison. OPD distils the answer, FFR repairs context, OPSD-KL matches, OPRD directs.

Variant	Pos	Type	RC	PV	Ch@5	HRR
Full model	$75.2_{\pm 0.9}$	$88.1_{\pm 0.5}$	$75.4_{\pm 1.0}$	$76.8_{\pm 1.2}$	$48.6_{\pm 1.2}$	$45.2_{\pm 1.4}$
A1 clue $\rightarrow$ full answer	$74.3_{\pm 1.1}$	$87.5_{\pm 0.7}$	$70.5_{\pm 1.5}$	$72.4_{\pm 1.7}$	$38.0_{\pm 1.8}$	$54.0_{\pm 1.9}$
A2 oracle-blind clue writer	$74.2_{\pm 1.2}$	$87.1_{\pm 0.8}$	$68.0_{\pm 1.8}$	$70.8_{\pm 1.8}$	$35.4_{\pm 2.0}$	$57.5_{\pm 2.1}$
A3 cascade $\rightarrow$ parallel	$74.7_{\pm 1.0}$	$88.0_{\pm 0.6}$	$70.7_{\pm 1.6}$	$72.1_{\pm 1.7}$	$40.7_{\pm 1.8}$	$50.5_{\pm 1.8}$
A4 per-stage $\rightarrow$ global α	$73.7_{\pm 1.1}$	$88.0_{\pm 0.6}$	$72.7_{\pm 1.4}$	$74.0_{\pm 1.5}$	$44.6_{\pm 1.6}$	$52.8_{\pm 1.6}$
A5 merge S4 into S3	$75.0_{\pm 0.9}$	$88.1_{\pm 0.5}$	$74.1_{\pm 1.2}$	$71.3_{\pm 1.7}$	$45.1_{\pm 1.5}$	$50.5_{\pm 1.5}$
A6 no direction, $\lambda = 0$	$74.1_{\pm 1.0}$	$87.8_{\pm 0.6}$	$72.3_{\pm 1.4}$	$73.9_{\pm 1.5}$	$44.0_{\pm 1.4}$	$47.5_{\pm 1.6}$
A7 direction ungated	$74.0_{\pm 1.1}$	$88.0_{\pm 0.6}$	$73.4_{\pm 1.3}$	$74.9_{\pm 1.4}$	$45.2_{\pm 1.6}$	$51.5_{\pm 1.5}$
A8 symmetric branches	$75.1_{\pm 0.9}$	$88.0_{\pm 0.5}$	$74.0_{\pm 1.2}$	$75.6_{\pm 1.3}$	$46.4_{\pm 1.4}$	$49.0_{\pm 1.5}$

(b) Component ablation. A7 keeps the cascade in the reward and drops it from the gradient.

Table 2: **Training paradigms and component ablation**, clue-free on `test_time` with one backbone, corpus and budget. Ch@5 is Chain@5, CF is CF-RC, HRR a training-time diagnostic that A1 and A2 move (Appendix I).

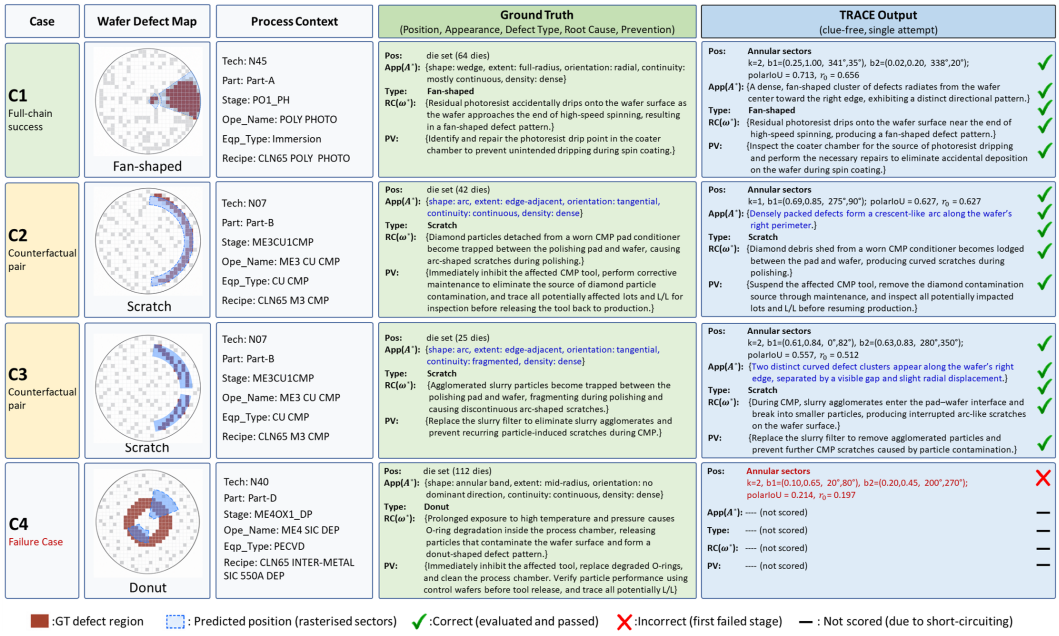


Figure 5: **Four clue-free diagnoses.** C2 and C3 are a counterfactual pair whose contexts and classes agree and whose continuity value does not, and the repairs separate with that one value; C4 breaks at S0, so the four stages below it are recorded but not scored (Appendix I).

Variant	Source of $\mathbf{d}_t$	Extra models	Chain@5	CF-RC
D0 none, $\lambda = 0$	—	0	$44.0_{\pm 1.4}$	$64.6_{\pm 1.9}$
D1 clue-cond. (TRACE)	$\mathcal{C}(z(\cdot \mid h_c) - z(\cdot))$	0	$48.6_{\pm 1.2}$	$69.3_{\pm 1.9}$
D2 oracle-cond.	$\mathcal{C}(z(\cdot \mid y^*) - z(\cdot))$	0	$50.1_{\pm 1.4}$	$70.4_{\pm 1.9}$
D3 pair, ref. mid P1	$\mathcal{C}(z_{\text{post}} - z_{\text{ref}})$	2	$44.6_{\pm 1.5}$	$62.9_{\pm 2.2}$
D4 pair, ref. SFT	$\mathcal{C}(z_{\text{post}} - z_{\text{sft}})$	2	$42.9_{\pm 1.7}$	$63.8_{\pm 2.3}$
D5 teacher-matching grad.	$\nabla \mathrm{KL}(\pi_T \| \pi_S)$	0	$43.3_{\pm 1.6}$	$61.9_{\pm 2.4}$
D6 random direction	unit vector, same norm	0	$44.2_{\pm 1.5}$	$64.1_{\pm 2.1}$
D7 no mean-centring	$z(\cdot \mid h_c) - z(\cdot)$	0	$46.5_{\pm 1.3}$	$66.8_{\pm 2.0}$
D8 templated h_1	h_1 as a bare attribute list	0	$45.8_{\pm 1.4}$	$67.4_{\pm 1.9}$

Table 3: **Direction channel.** Only $\mathbf{d}_t$ varies, except D8, which varies the wording of h_1 at fixed clue information; "Extra models" counts frozen copies beyond the policy and its grader snapshot, which every row carries, and D1 is TRACE itself (Appendix I).

Which direction carries the benefit (Table 3). A random unit direction of the same norm recovers almost nothing: D6 reaches 44.2, within noise of the 44.0 of D0, so the gain is not an enlarged gradient. Matching rather than directing is worse than doing neither, D5 at 43.3 Chain@5 and 61.9 CF-RC against D0's 44.0 and 64.6. Both checkpoint-pair references land near D0, D3 at 44.6 and D4 at 42.9, two points apart on nothing but the choice of reference — the sensitivity that motivates using two conditionings of one set of weights. Mean-centring is worth 2.1 points (D7) and wording at fixed clue information 2.8 (D8). D2 scores above D1 but is an upper reference, not a competing method: its direction is built from y^*, a privilege Eq. (4) does not bound, and at three seeds its margin is inside the noise. The remaining rows are read in Appendices I and J.

5 Conclusion

We posed wafer defect diagnosis as an evidence-grounded reasoning problem, and supplied what that needs: per-operation adder maps whose operation is fixed by construction and whose evidence layer is recorded apart from its interpretation, and an objective that will not score a stage on a false premise. TRACE spends one privileged teacher through two channels, neither a target: a clue bounded by Eq. (4), and the centred logit shift it induces (Proposition 1), entering as a preconditioner that narrows the search without moving where the token-level gradient vanishes. None of it reaches inference.

REPRODUCIBILITY STATEMENT

The diagnostic chain, the polar grounding operator, the stage-wise reward and the two clue channels are specified in Section 3 at the level of the equations that implement them: Eqs. (2) and (3) for the positional score, Eqs. (5) and (6) for the reward, Eq. (4) for the constraint that defines an admissible clue, and Eqs. (7), (8) and (10) for the direction channel, whose three claims are proved in Appendix A. The training schedule, its annealing clocks and its transition criteria are in Table 7 and Section 3.6, with the grader thresholds τ_1, τ_3, τ_4 fitted as Appendix F describes rather than set; the evaluation protocol, the split construction and the rare-class admission rule are given in Appendix I. The appearance annotation protocol — what the interface exposes, the presentation order, the annotator-disjointness condition, the complete value vocabulary and the two leakage screens — is specified in Appendix B and Tables 4 and 5. Appendix I lists the remaining implementation details, and the prompts used for the grader and for each clue operator are reproduced in Appendix K. Source code will be released as anonymous supplementary material. The fab record contains proprietary manufacturing data; Appendix B states what can be released and under which conditions.

ETHICS STATEMENT

The corpus is derived from production data of a semiconductor manufacturing line. It contains no human subjects and no personal data. The fab record retains lot and wafer identifiers and individual equipment identities, which are used for partitioning and traceability and are pseudonymised in the released form; none of them is part of the process context m that reaches the policy, so they are a release-control question rather than a model input. Process parameters that would disclose a specific manufacturer's recipe are withheld. The diagnostic annotations were written by process engineers as part of their normal work; the appearance layer was added in a separate annotation pass under the controlled conditions of Appendix B. Both are contributed with the consent of the participating organisation. The intended use of the resulting model is decision support for yield engineers, and we note one foreseeable misuse: a prevention action emitted by the model is a recommendation about equipment, and acting on it without engineering review could affect a production line. The rubric of Section 3.3 scores whether an action is specific and controllable, not whether it is safe to apply, and we make no claim that the model's output is safe to execute unsupervised.

AI USE STATEMENT

Generative AI tools were used solely for proofreading and grammatical refinement of the manuscript, as well as debugging Python code. Generative AI tools were *not* used to generate the research idea, dataset, annotations, or experimental results. All AI-assisted revisions and code debugging were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, code, and artifacts.

REFERENCES

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 2020.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. UniVAD: A training-free unified model for few-shot visual anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Jianfang He, Min Cao, Silong Peng, and Qiong Xie. RareCLIP: Rarity-aware online zero-shot industrial anomaly detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Samet Hicsonmez, Abd El Rahman Shabayek, and Djamila Aouada. VLMDiff: Leveraging vision-language models for multi-class anomaly detection with diffusion. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

Haojian Huang, Chuanyu Qin, Yinchuan Li, and Yingcong Chen. Find, fix, reason: Context repair for video reasoning. *arXiv preprint arXiv:2604.16243*, 2026.

Yizhou Jin, Yuezhu Feng, Jinjin Zhang, Peng Wang, Qingjie Liu, and Yunhong Wang. Reasoning-driven anomaly detection and localization with image-level supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Yujin Lee, Sewon Kim, Daeun Moon, Seoyoon Jang, and Hyunsoo Yoon. Bidirectional multi-modal prompt learning with scale-aware training for few-shot multi-class anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Wenqiao Li, Yao Gu, Xintao Chen, Xiaohao Xu, Ming Hu, Xiaonan Huang, and Yingna Wu. Towards visual discrimination and reasoning of real-world physical dynamics: Physics-grounded anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Yanhui Li, Yunkang Cao, Chengliang Liu, Yuan Xiong, Xinghui Dong, and Chao Huang. IAD-R1: Reinforcing consistent reasoning in industrial anomaly detection. In *AAAI Conference on Artificial Intelligence*, 2026.

Yuanze Li, Shihao Yuan, Haolin Wang, Qizhang Li, Ming Liu, Chen Xu, Guangming Shi, and Wangmeng Zuo. Triad: Empowering LMM-based anomaly detection with expert-guided region-of-interest tokenizer and manufacturing process. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b.

Hongchi Ma, Guanglei Yang, Debin Zhao, Yanli Ji, and Wangmeng Zuo. ReMP-AD: Retrieval-enhanced multi-modal prompt fusion for few-shot industrial visual anomaly detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S. Kevin Zhou. AA-CLIP: Enhancing zero-shot anomaly detection via anomaly-aware CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Tsai-Ching Ni, Cheng-Chi Chen, and Yuan-Fu Yang. Towards open-vocabulary industrial defect understanding with a large-scale multimodal dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Youngrok Park, Sangmin Bae, Hojung Jung, Jongwoo Ko, Yunseon Choi, Young Jin Kim, Pashmina Cameron, Aaron Courville, and Se-Young Yun. Eliciting weak-to-strong generalization with on-policy reverse distillation. *arXiv preprint arXiv:2609.08798*, 2026.

Zhen Qu, Xian Tao, Xinyi Gong, ShiChen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zhen Qu, Xian Tao, Xiaoyi Bao, Dingrong Wang, ShiChen Qu, Zhengtao Zhang, and Xingang Wang. AG-VAS: Anchor-guided zero-shot visual anomaly segmentation with large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

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. *arXiv preprint arXiv:2402.03300*, 2024.

Junliang Wang, Chuqiao Xu, Zhengliang Yang, Jie Zhang, and Xiaolei Li. Deformable convolutional networks for efficient mixed-type wafer defect pattern recognition. *IEEE Transactions on Semiconductor Manufacturing*, 2020.

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.

Yu Xia, Canwen Xu, Zhewei Yao, Julian McAuley, and Yuxiong He. Learning to hint for reinforcement learning. *arXiv preprint arXiv:2604.00698*, 2026.

Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M. Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ke Xu and Zhongyuan Lian. WaferSAGE: Large language model-powered wafer defect analysis via synthetic data generation and rubric-guided reinforcement learning. *arXiv preprint arXiv:2604.27629*, 2026.

Yuan-Fu Yang and Hsiu-Hui Hsiao. Photolithography overlay map generation with implicit knowledge distillation diffusion transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Yuan-Fu Yang and Min Sun. A novel deep learning architecture for global defect classification: Self-proliferating neural network (SPNet). In *SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2021.

Yuan-Fu Yang and Min Sun. Semiconductor defect detection by hybrid classical-quantum deep learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Yuan-Fu Yang and Min Sun. Semiconductor defect pattern classification by self-proliferation-and-attention neural network. *IEEE Transactions on Semiconductor Manufacturing*, 2022b.

Xincheng Yao, Zefeng Qian, Chao Shi, Jiayang Song, and Chongyang Zhang. MMR-AD: A large-scale multimodal dataset for benchmarking general anomaly detection with multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Linhao Yu, Tianmeng Yang, Siyu Ding, Renren Jin, Naibin Gu, Xiangzhao Hao, Shuaiyi Nie, Deyi Xiong, Weichong Yin, Yu Sun, and Hua Wu. KnowRL: Boosting LLM reasoning via reinforcement learning with minimal-sufficient knowledge guidance. *arXiv preprint arXiv:2604.12627*, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Kai Zhang, Zekai Zhang, Xihe Sun, Anpeng Wang, Jingmeng Nie, Qinghui Chen, Han Hao, Jianyuan Guo, and Jinglin Zhang. ADSeeker: A knowledge-grounded reasoning framework for industry anomaly detection and reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint*, 2026.

Wenbing Zhu, Lidong Wang, Ziqing Zhou, Chengjie Wang, Yurui Pan, Ruoyi Zhang, Zhuhao Chen, Linjie Cheng, Bin-Bin Gao, Jiangning Zhang, Zhenye Gan, Yuxie Wang, Yulong Chen, Shuguang Qian, Mingmin Chi, Bo Peng, and Lizhuang Ma. Real-IAD D3: A real-world 2D/pseudo-3D/3D dataset for industrial anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021.