BUDGETMATCH: AUDITING REFERENCE ACCESS IN FROZEN-FEATURE ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero- and few-shot industrial anomaly detectors can appear in the same quantitative comparison even when their reference access differs materially: an inductive detector uses clean training-normal images, whereas an offline transductive detector may use the complete test cohort, including other anomalous queries. Setting labels such as batched zero-shot separate such rows without pricing the difference. We introduce **BudgetMatch**, an auditable evaluation contract that describes each protocol on four cards and whose certificate lists every field two protocols differ on, admitting a single-factor comparison only inside a predeclared intervention set. A registered study measures the difference in clean reference images. With a global reference-patch 1-nearest-neighbor scorer fixed, going from one to sixteen clean images raises macro image AUROC on all ten dataset–path pairs of MVTec AD and VisA, by $+5.60$ and $+6.31$ percentage points, and by $+6.13$ on Real-IAD. Read post hoc, the first two exceed 36 of the 46 gains audited papers report for their own methods (33 with the archived AnomalyDINO ladder). Under that nearest-neighbor reduction, a clean bank of the full pool's size beats the self-excluded cohort by $+3.49$ and $+7.07$ points (registered; all five paths agree in a later-registered follow-up). Post hoc, the cohort's anomalies cost $+3.61$ and $+7.23$ points, against $+0.41$ and $+0.50$ for its source. MuSc's interval average shrinks the margin to $+1.00$ (undetermined) and $+4.01$; on Real-IAD, which has no ceiling, a smaller clean pool still wins by $+3.64$. Post hoc, the worth depends on the scorer: one clean image matches or beats the cohort in 15 of 27 categories under 1-NN, but in only 4 under official MuSc, which sixteen clean images fail to reach in 8. Applied post hoc to 66 printed comparisons of 12 papers, BudgetMatch finds verifiably different access under one setting label in 54; in 30, test-cohort access is undeterminable from a listed method's own paper. We therefore read a ranking as undecidable when its margin is below the measured worth of a field its rows leave undeclared. Every value from our runs is hash-bound.

1 INTRODUCTION

Industrial anomaly detection asks a model to separate normal from anomalous products when anomalies are scarce and heterogeneous. Frozen encoders make a particularly simple family of solutions possible: extract patch features, retain a reference bank, and score each query by its distance to that bank (Roth et al., 2022; Jeong et al., 2023; Li et al., 2024b; Damm et al., 2025). The apparent simplicity hides an evaluation choice. A conventional inductive system builds its bank from clean training-normal images. An offline transductive system can instead inspect the complete unlabeled test cohort and score each query relative to the other queries (Li et al., 2024b). The two systems therefore do not receive the same information or the same reference resource.

This distinction matters because a single leaderboard number can otherwise be read as an algorithmic comparison when it is also a protocol comparison. The full test cohort is not available when queries arrive individually, contains anomalous images, grows with the benchmark, and couples one query's score to the identities of other queries. Conversely, a small clean support bank has verified normality but a deliberately limited budget. Equal backbones and nominally similar distance functions do not remove these differences.

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.

The fix is not to attribute the gap between the two systems to one favored cause. Changing from clean support to the test cohort changes all eight access and resource fields of Section 3 at once (Figure 1). We therefore treat each arm (each reference protocol) as one complete protocol and separate two research questions. *RQ1 (controlled sensitivity):* within the clean-support arm, how sensitive is a fixed scorer to a predeclared increase from one to sixteen reference images? *RQ2 (complete-protocol position):* where does the deterministic full-pool protocol lie relative to those clean-support curves on the same finite benchmarks? RQ1 changes one declared resource factor; RQ2 is descriptive. A separately registered decomposition then asks which fields make up the RQ2 gap, using diagnostic arms that, unlike the two protocols they explain, need test labels.

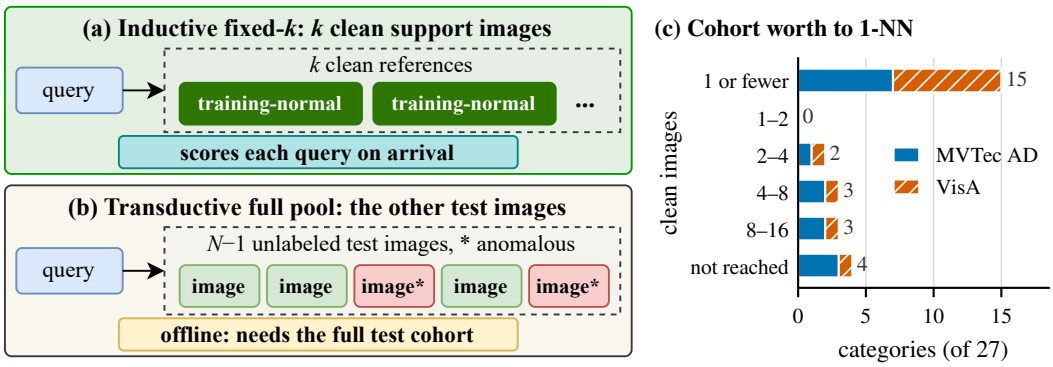


Figure 1: Two reference-access regimes for one frozen scorer, and what separates them. (a) The inductive fixed-k arm scores each arriving query against a fixed bank of k verified-normal training images. (b) The transductive, self-excluded full-pool arm scores it against the other $N-1$ images of its category's unlabeled test cohort, which contains anomalous images (marked *) and exists only offline. The two arms can share encoder, distance and scoring functional and still differ in the eight fields of Appendix Table 13. (c) For each category, the clean budget, interpolated between grid points, at which the path-averaged clean curve reaches the 1-NN full-pool arm: its shot-equivalent, a post hoc quantity (Appendix O); bars count categories by dataset, and each bin includes its upper edge.

Some benchmark harnesses already fix one access protocol by rule. The Visual Anomaly and Novelty Detection (VAND) 2025 challenge admits only the k support images and the test image (VAND Workshop Organizers, 2025), and AnomalyDINO separates a batched zero-shot setting from zero- and few-shot ones (Damm et al., 2025). Yet a setting label records one field and prices none. General reporting layers flag setup differences only once reported scores diverge (Ghosh et al., 2026) or group entries by observed protocol signature (Siravuri & Alhoori, 2026). Three of five public reviewers of Li et al. (2024b) questioned scoring against unlabeled test images (OpenReview, 2024). The authors acknowledged that this "may lead to a potentially unfair comparison" and added a row for another test-cohort method under the same zero-shot label. A repair made per table carries to no other comparison. What is missing is a record of each protocol that a program can compare field by field before two scores share a row. **BudgetMatch** supplies it. It describes each arm on four linked cards: what the method may see, what references it gets, how it scores, and what it is evaluated on. A certificate lists every field on which two arms differ before their comparison is called single-factor or complete-protocol. Our study freezes an OpenCLIP ViT-L/14-336 encoder (Radford et al., 2021; Cherti et al., 2023), a global reference-patch 1-nearest-neighbor (1-NN) scorer, five nested budgets $k \in \{1, 2, 4, 8, 16\}$, and five hash-ranked support paths for every category in MVTec AD and VisA (Bergmann et al., 2019; Zou et al., 2022).

Beyond BudgetMatch itself, released with a validator that writes its certificate, we claim three results, each with its own evidence:

- **C1. What a reference-access difference is worth can be measured in clean reference images, and the price depends on the scorer.** We apply to reference access the labeled-examples-to-match unit with which Radford et al. (2021) price zero-shot transfer. With every other field fixed, going from one clean image to sixteen improves macro image AU-

ROC (area under the receiver operating characteristic curve) on all ten dataset–path pairs of MVTec AD and VisA, by $+5.60$ and $+6.31$ percentage points. On Real-IAD, which does not saturate, the gain is $+6.13$. On that scale, and under the nearest-neighbor reduction, the self-excluded test cohort is worth less than a clean bank of its own size: a registered decomposition on one predeclared support path finds the bank ahead by $+3.49$ and $+7.07$ points. In a source $\times$ purity design, the cohort's own anomalies account for $+3.61$ and $+7.23$ points against $+0.41$ and $+0.50$ for its source (post hoc). The price moves with the scorer: read post hoc, the cohort is worth one clean image or fewer in 15 of 27 categories under the 1-NN arm (Figure 1c) but in only 4 under the official MuSc method.

- **C2. That worth is as large as most margins published tables rank by, so a setting label cannot stand in for it.** The one-to-sixteen effect exceeds 36 of the 46 gains the audited papers report for their own methods, and 33 when measured with the archived AnomalyDINO ladder instead ($+1.74$ and $+8.93$; post hoc; Appendix U). We therefore read a ranking as undecidable when its margin is smaller than the measured worth of a field its two rows leave undeclared. Separating transductive rows into their own block, as some tables already do (Damm et al., 2025), removes the label ambiguity but not this question.
- **C3. Whether two scores may share a row can be checked field by field, and, in published tables, rows printed under one setting label can differ in access.** Four cards give the fields, the difference set gives what is unequal, and one admissibility rule turns them into a verdict a program can return (Section 3). Applied post hoc to all 66 printed comparisons of 12 papers, it finds verifiably different access under one label in 54: 48 differ in the source or purity of the data a method trains on or draws references from and 6 in test-time access. Test-time access is undeterminable in 30, where a method could not be coded from its own paper (Appendix G).

Authors, reviewers and maintainers can apply the C2 and C3 checks to any table (Section 5).

2 Related work

Reference-based anomaly detection. Patch-level methods score queries against normal training data, kept as a feature bank (Cohen & Hoshen, 2020; Roth et al., 2022) or as Gaussian statistics (Defard et al., 2021). Frozen vision-language and self-supervised encoders extend this to zero- and few-shot settings via text prompts or a few normal images (Chen et al., 2023; Jeong et al., 2023; Zhou et al., 2024; Damm et al., 2025; Lendering et al., 2026). MuSc instead performs offline mutual scoring over the test cohort (Li et al., 2024b), and MuSc-V2 tests it under smaller subsets, very small batches and varying normal-sample ratios, so sensitivity to the available cohort is not itself new (Li et al., 2026). Online methods may accumulate a memory from earlier test observations (He et al., 2025), and FiSeCLIP uses the other images of the current test batch as mutual references (Bai et al., 2025). Other work widens what a detector may hold, from retrieval memories and expert routing to learned visual tokens and agents (Xu et al., 2026; Park et al., 2026; Ali et al., 2026; Hou et al., 2026; Zhang et al., 2026a). None records, in machine-readable form, which references a method may use, whether it sees the other queries, whether those references are clean, and whether it keeps them.

Transduction and reference access. A transductive learner predicts the given test points rather than a rule for future inputs (Gammerman et al., 1998), so the difference from an inductive system is not label leakage under another name but a difference in what each is allowed to know. Few-shot learning met the same problem: transductive learners "have more information available" at prediction time, "which may lead to unfair comparisons" (Bronskill et al., 2020), and their advantage depends on the class composition benchmarks assume for the test set (Veilleux et al., 2021). Batch-based zero-shot anomaly detection in 3D medical imaging loses accuracy as its unlabeled test cohort is split into smaller chunks (Le-Gia, 2026); Appendix D lists further instances.

Evaluation documentation and benchmark discipline. Evaluation Cards come closest: their comparability flag is raised when reports of one model on one benchmark diverge by more than 5% of the metric range, and it then surfaces the setup fields that differ (Ghosh et al., 2026); a recent dataset groups leaderboard entries by observed protocol signature and leaves entries of unknown comparability unranked (Siravuri & Alhoori, 2026). Concurrent harness and protocol cards for

LLM agents record how each score was produced (Zhang et al., 2026b; Zhang & Lee, 2026), but concern agent execution rather than reference data. None of these turns its card into a field-level rule for when two scores may share a row. BudgetMatch supplies that rule: it encodes what a detector may see and keep, compares the rows of different methods before any score is read, and returns a verdict, admissible or not against a predeclared intervention set, checked against hashed identities. Benchmark audits have shown that rankings depend on protocol choices, aggregation and the tuning budget (Dodge et al., 2019; Musgrave et al., 2020; Han et al., 2022; Zhao et al., 2023; Xie et al., 2024; Goldblum et al., 2023); we extend the documentation layers of Appendix D to the reference data itself.

Access in anomaly-detection practice. Practice is uneven rather than careless. Li et al. (2023) state batch-level access as a numbered assumption, Damm et al. (2025) separate a batched zero-shot setting, in which they place MuSc and ACR, from zero- and few-shot ones, and an online zero-shot detector lists MuSc "as a reference under different test-time information assumptions" (Zhao et al., 2026). Yet in a bounded audit of 12 cited papers, eight put methods with different access in one table (Appendix E). Coding every row of the main comparison tables of a second set of 12 papers finds verifiably different access under one setting label in 54 of 66 printed comparisons (Appendix G). Both sets were chosen, not sampled, so the counts say the practice exists, not how common it is. Concurrent work varies which candidates fill a fixed-size memory (Chae et al., 2026b) or holds a patch budget fixed and varies purity (Kalantary & Khoramdel, 2026; Chae et al., 2026a); we take the complementary axis, purity fixed and budget varied, and treat test-cohort access as one complete protocol.

BudgetMatch is not a new detector: our study freezes the representation and the scorer, and the contract makes the permitted information flow something a program can check. Nor is it a re-benchmarking study, since the registered study (v4 in the artifact) added no method, backbone, dataset or hyperparameter after seeing its own results. Each later protocol carries its own identity, and Appendix X states for each what evidence of registration survives. What is meant to be reused is the specification, the admissibility rule and the certificate—not a leaderboard.

3 Protocol and Estimands

3.1 Reference-access protocol specification and comparison admissibility

Write d for a dataset, c for a category, Q_{dc} for that category's test cohort of $N = |Q_{dc}|$ images, and $x_{dci} \in Q_{dc}$ for its ith query. We describe a protocol $\mathcal{P}$ by four cards, its signature $\sigma(\mathcal{P}) = (A, R, F, E)$. Each card answers one question about the method, and Appendix Table 13 fills them in for our two arms:

1. the *access card*—what may the method look at? *Source*, the data it may draw on; *purity*, whether it knows those images are clean; *query visibility*, whether it sees the other queries; *self-exclusion*, whether it excludes the query itself; and *cohort coupling*, whether one query's score can depend on which other queries happen to be present;
2. the *resource card*—what does it get to keep? The *reference budget*, the *support-selection rule* that fills it, and the *reuse/update* policy. Patch count and storage follow from these and are reported, but they are not separate choices;
3. the *functional card*—how does it turn that into a score? Encoder, preprocessing, feature layers, distance, and the aggregation that reduces them;
4. the *evaluation card*—what is it scored on? Query cohort and order, category aggregation, the primary and secondary metrics, and the *uncertainty unit*, meaning what a reported range spreads over.

When we code a published row, whose functional card is the method itself, we compare only the access and resource cards, eight fields, and call that part of σ the row's signature. Two protocols differ on a set of fields,

$$D(\mathcal{P}, \mathcal{P}') = \{f : \sigma_f(\mathcal{P}) \neq \sigma_f(\mathcal{P}')\}. \tag{1}$$

Declare in advance the set I of fields meant to change. We call a contrast *admissible for* I when $\varnothing \neq D(\mathcal{P}, \mathcal{P}') \subseteq I$ and the same units are paired on both sides: the only fields that moved are

Table 1: A published comparison, coded: the MVTec AD blocks of Table 1 in Li et al. (2024b), each row from its own paper (Appendix F). The last column counts the access and resource fields, of eight, on which a row differs from MuSc. Ten source rows print as eight lines: MuSc appears, labeled 0-shot, in both blocks, and PatchCore and WinCLIP share one coding. †Coded with a field the paper does not state.

Label	Published rows	Test-time references	Extra training data	≠ MuSc
0-shot	WinCLIP	none	none	8
	APRIL-GAN	none	labeled auxiliary set	8
	ACR	test cohort, batch statistics	other categories	6
	MuSc	test cohort, self-excluded	none	reference
4-shot	RegAD†	4 clean training images	other categories	8
	PatchCore, WinCLIP	4 clean training images	none	8
	APRIL-GAN	4 clean training images	labeled auxiliary set	8
	GraphCore†	4 clean training images	not stated	8

the ones declared in I; with $|I| = 1$ the contrast is *single-factor*. If anything outside I differs, we label the comparison *complete-protocol*—a comparison between two whole protocols—rather than picking whichever explanation we prefer. The rule is necessary, not sufficient: passing it does not prove that no unrecorded difference remains, which is the point Appendix V returns to. Nor is a coarse label that records only source and budget enough, because two arms can share both and still differ in scorer, selection rule, purity guarantee or evaluation cohort. On our two contrasts, the released validator, which writes the certificate, returns one field for RQ1 (the clean reference budget) and eight for RQ2—exactly the rows Appendix Table 13 marks unequal.

Applied to published tables (C3). The rule ranks no access mode above another; it asks only that a difference be declared or matched. On a published comparison it separates what a shared label merges: in Table 1 the five rows labeled zero-shot carry four signatures, as do the five labeled four-shot. Of the 20 row pairs that share a label, only one compares methods alone, while four more turn on a field a paper does not state.

3.2 Fixed scoring functional

For feature layer $\lambda \in \Lambda$ and aggregation radius $\rho \in \Gamma$ (Section 4), let $z_{jm}^{\lambda\rho}$ be patch m of reference image j and $u_{ip}^{\lambda\rho}$ query patch p. The per-reference distance is

$$a_{ipj}^{\lambda\rho} = \min_m \|u_{ip}^{\lambda\rho} - z_{jm}^{\lambda\rho}\|_2. \tag{2}$$

The formal experiment fixes the absolute q=1 functional, where q counts the lowest per-reference distances averaged (a fixed count, not a fraction of the bank):

$$s_{ip}(J) = \frac{1}{|\Lambda||\Gamma|} \sum_{\lambda,\rho} \min_{j \in J} a_{ipj}^{\lambda\rho}, \qquad S_i(J) = \max \mathrm{Up}\big(s_{i\cdot}(J)\big), \tag{3}$$

where Up bilinearly upsamples the patch-score map to the input resolution. This is a *global reference-patch 1-NN scorer*, not the complete PatchCore algorithm (Roth et al., 2022). Enlarging J cannot increase a raw patch score, but AUROC need not be monotone, so we report the whole frozen curve without a monotonicity claim.

3.3 Arms and semantic isolation

In the fixed-k arm, J_{dcrk} contains the first k clean training-normal images in predeclared support path $r \in \{0, \ldots, 4\}$, labeled p0–p4 in tables and figures. Paths are hash-ranked and nested, so $J_{dcr,1} \subset J_{dcr,2} \subset \cdots \subset J_{dcr,16}$. Only the current query and that prefix are semantically visible; numerically, the inference batch can still move scores (Appendix V). Feature caching is allowed only after an elementwise cache/no-cache equality test; unselected support and other queries must have no influence.

In the full-pool arm, the reference bank for query i is the set of all other images in the same category cohort. Self-exclusion is mandatory. The arm is computed once per category and its cohort identities and order are hashed. Because this arm changes multiple protocol-card fields relative to fixed-k, the contrast is a complete-protocol comparison; it does not separately identify source, purity, or budget effects. The diagnostic decomposition of Appendix R separates them on all five support paths, using test labels.

3.4 Frozen questions and reporting units

Let C_d be the categories of dataset d and Y_{dcrk} the category-level image AUROC for fixed-k. The primary paired difference on fixed benchmark strata for dataset d and support path r is

$$\Delta_{dr} = \frac{1}{|C_d|} \sum_{c \in C_d} \left(Y_{dcr,16} - Y_{dcr,1} \right). \tag{4}$$

We report all five Δ_{dr} values plus their mean, sample standard deviation, minimum, and maximum as planned-support sensitivity, not as an i.i.d. population confidence interval or confirmatory p-value. The registered protocol, v4 in the artifact, applies this estimand to MVTec AD and VisA and freezes the complete five-budget curves below.

The complete $k \in \{1, 2, 4, 8, 16\}$ curve is secondary. If T_{dc} is the deterministic full-pool category AUROC, then its dataset macro value is $T_d = |C_d|^{-1} \sum_c T_{dc}$. For each path we may describe

$$k^*_{dr} = \min\{k : Y_{drk} \geq T_d\}, \tag{5}$$

or "none," where Y_{drk} is the dataset macro of Y_{dcrk}. This is an empirical crossing between complete protocols, not a causal threshold. Image average precision (AP), pixel AUROC/AP, the area under the per-region-overlap curve (AUPRO), and F1-max are secondary recomputations from the same raw artifacts; F1-max tunes its threshold on the test data.

A cell is one scored unit, a (dataset, category, support path, budget) combination or one category's full-pool arm. It binds the protocol, data, ordered query and support, encoder, configuration, code, environment and metric identities and is committed with its completion marker written last; every attempt goes to an append-only ledger (Appendix I). Throughout, *registered* means that a protocol's document, including its decision rules, was frozen before any value of that protocol existed. The commitments are internal and carry no third-party timestamp; Appendix X states for every protocol what evidence of timing survives and lists the post hoc analyses.

4 Registered experiment and validation

4.1 Datasets, encoder, and frozen matrix

We use the official category structures of MVTec AD (Bergmann et al., 2019) and VisA (Zou et al., 2022): 15 and 12 categories, respectively. Clean supports come only from each dataset's training-normal split; evaluation uses the corresponding test cohort. The encoder is OpenCLIP ViT-L/14-336 with one fixed local weight artifact (Radford et al., 2021; Cherti et al., 2023). Preprocessing, feature layers, radii, distance, q=1, and metric code are identical across arms. The exact setup uses 518×518 inputs, layers $\{5, 11, 17, 23\}$, radii $\{1, 3, 5\}$ for the local neighborhood aggregation with multiple degrees (LNAMD) of Li et al. (2024b), and bilinear interpolation with aligned corners. Scoring uses only the distance stage of their mutual scoring mechanism (MSM), applied to each arm's reference bank. Encoder, code, and environment hashes are given in Appendix A.

The registry contains 675 fixed-k scoring cells (27 categories $\times$ five budgets $\times$ five support paths) and 27 deterministic full-pool category cells (Appendix Table 4). Each cell sees only the references its registry row names, and nothing is trained: the 702 cells are not 702 tuned configurations. Before the formal run the implementation passed its released test suite, from a hand-computable NumPy oracle to a scheduler drill that interrupts a category mid-run (Appendix I). An engineering pilot, whose values are not evidence, showed that integer casting of resized masks discards boundary pixels, so masks are binarized at $m > 0.5$ (Appendix J).

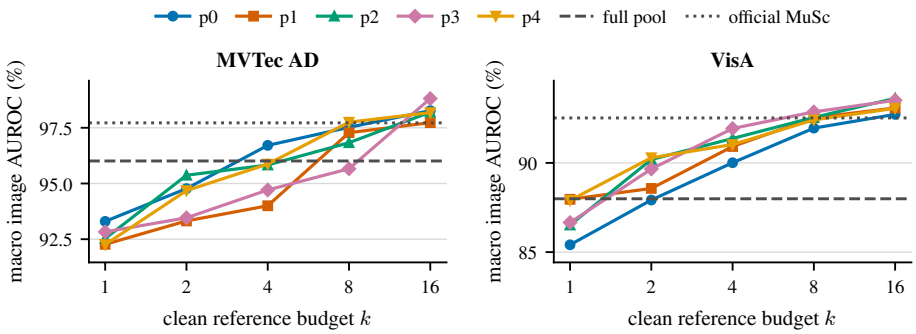


Figure 2: Dataset-macro image AUROC across the five frozen clean reference budgets k. Each colored line is one predeclared deterministic support path; the dashed line is the self-excluded full-pool protocol under this paper's 1-NN functional, not a matched-access baseline; the dotted line is the official MuSc method (97.72 and 92.52), placed post hoc from archived runs of its unmodified program, which change the functional as well as the access (Appendix O). Points are fixed-benchmark values, not random-seed estimates; Appendix Table 5 lists them.

4.2 Formal results

Unless another protocol or an archived run is named, every value here comes from the 702 registered cells, used only after each carried a hash-valid completion marker and an independent artifact audit reproduced its metrics, with no outcome-dependent filtering. The five support paths are predeclared deterministic selections, so every mean, standard deviation, and range across them describes fixed-benchmark path sensitivity, not sampling uncertainty.

Table 2: Primary contrast Δ_{dr}: k=16 minus k=1 macro image AUROC (percentage points), from unrounded values. Paths are predeclared support selections; SD is descriptive. Every cell maps to a row in `paper/audit/provenance/table-cell-provenance.csv`.

Dataset	p0	p1	p2	p3	p4	mean ± SD	range
MVTec AD	+4.97	+5.45	+5.67	+5.98	+5.93	+5.60 ± 0.41	[+4.97, +5.98]
VisA	+7.32	+5.12	+7.10	+6.86	+5.16	+6.31 ± 1.08	[+5.12, +7.32]

Sixteen clean images beat one on every path (RQ1, RQ2; C1). The primary contrast in Table 2 is positive for all ten dataset–path pairs (Figure 2). Its descriptive mean is $+5.60$ percentage points on MVTec AD (sample standard deviation, SD, 0.41, range $+4.97$ to $+5.98$) and $+6.31$ on VisA (SD 1.08, range $+5.12$ to $+7.32$). Thus the direction is stable across the planned support selections, while its magnitude remains path- and dataset-dependent. The protocol is a locked replication of a direction explored before its supports were frozen, and the magnitude is descriptive. With categories rather than paths as the unit, a post hoc category-stratified bootstrap of the path-mean contrast gives $[+1.63, +10.78]$ on MVTec AD and $[+3.54, +9.72]$ on VisA (Appendix U). Clean support first meets or exceeds the deterministic full-pool macros (96.01% and 87.99%) at budgets $(4, 8, 8, 16, 8)$ across the five MVTec AD paths and $(4, 2, 2, 2, 2)$ across the VisA paths. Post hoc, the VisA cohort is worth less than four verified-normal images under MuSc's own reduction (matched in its declared functional but not its aggregation width); on MVTec AD it is bounded only from below (Appendix N).

Read post hoc and per category, the same crossing gives a unit that is not a constant: how many clean images buy what the whole cohort delivers. One clean image already matches or beats the full-pool arm in 15 of 27 categories. In 4 categories, sixteen do not reach it. The unit depends on the scorer that uses the cohort, not only on the cohort. The official MuSc method, placed post hoc from archived runs of its unmodified program on the same curves, needs more than one clean image

Table 3: The full-pool gap by field, in percentage points of the named metric (the registered decompositions of Appendix R; *Minimum*: the nearest-neighbor reduction). *Budget*: a clean training bank of the full pool's size minus the full pool; *purity*: the cohort subsample with its anomalies removed minus with them; *source*: normal references from the test cohort minus from training with the anomalies fixed, which also changes their selection. Registered verdicts: °undetermined, =equivalent to zero, otherwise present (purity is judged by the count H instead, supported at ten or more). The pixel row uses the registered primary pixel metric; the secondary AUPRO is in Appendix Table 29. Real-IAD, with no category at the ceiling: Appendix Table 33.

	Budget		Purity		Source		
Reduction, encoder, metric	MVTec	VisA	MVTec	VisA	MVTec	VisA	H
Minimum, CLIP, image AUROC, p0	$+3.49$	$+7.07$	$+3.58$	$+7.45$	$+1.50$	$+0.60^{\circ}$	$14/15$
Minimum, CLIP, image AUROC, p0–p4	$+3.43$	$+7.12$	$+3.58$	$+7.45$	$+1.43^{\circ}$	$+0.62^{\circ}$	$14/15$
Minimum, DINOv2, image AUROC, p0	$+3.92$	$+9.22$	$+3.66$	$+9.05$	$+1.02^{\circ}$	$+0.25^{\circ}$	$15/15$
Minimum, CLIP, pixel AUROC, p0	$+4.27$	$+0.74^{\circ}$	$+3.28$	$+0.56$	$+0.14^{=}$	$+0.00^{=}$	$4/15$
Interval average, CLIP, image AUROC, p0	$+1.00^{\circ}$	$+4.01$	$+2.35$	$+4.19$	$+0.56^{\circ}$	-0.26°	$12/15$

in all but 4 categories and more than sixteen in 8 (Appendix O). Swapping the frozen encoder for DINOv2 (Oquab et al., 2024), with or without registers, moves the 1-NN counts by at most two categories, under the registered threshold of four. Yet the registered rank correlation falls below 0.5, a verdict that stands although ties set the statistic here (Appendix P). A frozen extension to k=32 for 7 categories (42 overlapping cells reproduced bitwise) makes 1 of the 4 unreached cross.

What the full-pool gap is made of (C1). The full-pool placement changes eight fields at once. A registered decomposition (Appendix R) moves three of them one at a time inside one distance computation per category, on the predeclared support path p0 and, in a registered follow-up, on all five. Its arms need test labels and serve only as diagnostics. Under the nearest-neighbor reduction, budget is not what the cohort supplies. A clean training bank of the full pool's own size beats the full pool by $+3.49$ points on MVTec AD and $+7.07$ on VisA on p0, and by $+3.43$ and $+7.12$ across the five paths, path-robust present on both. Purity is what it costs. Swapping the cohort's own anomalies into that clean bank at equal size changes image AUROC by -4.99 and -7.67 points. Removing them from a cohort subsample of equal size changes it by $+3.58$ and $+7.45$. The registered purity count holds in 14 of the 15 categories where one clean image suffices. Source matters least: present on p0 for MVTec AD ($+1.50$), it is undetermined across the five paths. The follow-up's four arms at one bank size form a source $\times$ purity design whose post hoc main effects are $+3.61$ and $+7.23$ points for purity against $+0.41$ and $+0.50$ for source, with interactions of -0.06 and $+0.44$, so which term is largest does not depend on which arms define the terms. The pattern holds under the second encoder (budget $+3.92$ and $+9.22$, present; purity count 15 of 15). It holds only in part at the pixel level: the registered pixel AUROC gives budget $+4.27$ and $+0.74$ (undetermined on VisA) and a purity count of 4 of 15, not supported (the descriptive AUPRO reads like image AUROC; Table 3, AUPRO in Appendix Table 29). It changes under MuSc's own interval average, which averages the closest 30% of references so that a few similar ones cannot decide a score (Li et al., 2024b). Purity still costs, but less ($+2.35$ and $+4.19$ on the cohort subsample; count 12 of 15). The clean bank's margin is $+1.00$ on MVTec AD, undetermined, and $+4.01$ on VisA, present. On Real-IAD, with no category at the ceiling, a clean pool smaller than the cohort wins under both reductions ($+8.66$ and $+3.64$ points, both present, registered). Its interval-average margin rises with headroom (Spearman 0.78, descriptive), compatible with a ceiling effect in the MVTec AD verdict, though it does not prove one (Appendix R).

The budget effect beyond one scorer and saturated benchmarks (C1). The budget effect is not specific to the nearest-neighbor scorer. With a principal-subspace residual scorer on the same features, k=16 beats k=1 on 10 of 10 dataset–path pairs (registered threshold eight) and in 30 of the 30 pairs of three further variants (Appendix S). Nor is it a product of saturated benchmarks: on the 30 categories of Real-IAD, none of which reaches the ceiling at any budget, the registered endpoint is positive on all five paths, by $+6.13$ points on average (Appendix T). A prospective endpoint stress test on BTAD (Mishra et al., 2021), frozen before any BTAD value was inspected, is positive on all

five paths (Appendix B). The endpoint comparison is positive on every path for six secondary image and pixel metrics (Appendix Table 6); category-level signs are in Appendix Table 7.

The price against published margins (C2). Read post hoc, the worth of one undeclared field is as large as most of the margins published tables rank by. As a descriptive comparison across backbones, the one-to-sixteen effect exceeds 36 of the 46 gains the audited papers report for their own methods, and 33 with the archived AnomalyDINO ladder (Appendix U).

5 Discussion

BudgetMatch buys one thing: a comparison a reader can interpret. Inside the fixed-k arm, nested clean supports change only the declared budget field and the images that fill it, so the curve means what it appears to mean. The full-pool arm answers a different question, and reading the two as one would invite a cause the design cannot supply.

The evidence reaches a support-sensitivity audit on fixed benchmarks and a split of the full-pool gap into budget, purity and source terms under named reductions. The arms behind that split need test labels, so they diagnose these benchmark cohorts; the two protocols they explain need none. The budget effect holds under a second scorer family, the clean bank's lead under a second encoder, and both on Real-IAD, where no category reaches the ceiling (Appendices T and R). Held-out benchmarks such as MVTec AD 2 (Heckler-Kram et al., 2026) are the natural next test.

Scoring against the whole test cohort uses no labels, though it does use test-set information (Kapoor & Narayanan, 2023). When the cohort genuinely exists, offline transduction is a legitimate way to use it. It is not interchangeable with an inductive reference set, and the difference must be visible before the two share a row. So we recommend stratifying leaderboard rows by what each method may access, merging rows only when the declared difference set is admissible, and reading a margin as undecidable when it is smaller than the measured worth of a field its rows leave undeclared. A row that leaves any of the four cards undeclared cannot support a claim about the method alone. Selection, query visibility, cohort coupling and reuse, which a table alone does not reveal, are the minimum a row should declare.

The specification also audits itself. In our own study, varying one field while pinning the rest exposed a numerical setting that the recorded identity does not bind: the inference batch size moves almost every patch score while every hash agrees. No reported value changes, because every reported run of our scorer used one batch size (Appendix V). The evaluation card should therefore record the batch size and numerical precision as well.

6 Conclusion

A setting label can say that two rows differ in reference access; it cannot say by how much. In frozen-feature anomaly detection that difference has a price in clean reference images. The price changes with the scorer that uses the references. A method may legitimately score against the test cohort, but that cohort is not a stand-in for a clean reference set it never had. Under the nearest-neighbor reduction, on every support path and under a second encoder, a clean bank of the cohort's own size still wins. Descriptively, putting the cohort's own anomalies in place of clean references costs more than that bank's whole lead over the cohort. A reduction that averages over many references, such as MuSc's own, recovers part of that cost, which is the scorer dependence again. On a benchmark without a ceiling, a smaller clean pool still wins under both reductions. Read post hoc, that price is of the same size as the margins published tables rank methods by.

BudgetMatch turns this into a check rather than an argument. Applied post hoc to the 66 printed comparisons we coded, it already finds verifiably different access under one setting label in 54. The four cards and the certificate list what a comparison leaves unmatched before any score is read. The measured prices say when an unmatched field is large enough to decide a ranking. We release the specification, the certificate and the audit chain so that both can be applied to tables we did not audit.

AI USE STATEMENT

In accordance with the ICLR 2027 policy on AI assistance, we disclose the categories of generative-AI use in the preparation of this paper. They correspond to the two categories the authors selected on the submission form.

Aiding and polishing the writing. We used large language model tools for English-language polishing, sentence-level rewriting of author-written text, and organizing the structure of the manuscript, including section ordering and headings.

Retrieval and discovery. We used large language model tools to search for and organize related work. Every cited work was then independently retrieved by the authors and checked against its official version of record, and the claims attributed to it in this paper were verified against its full text.

The authors reviewed all AI-assisted text and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies evaluation protocols for industrial visual anomaly detection. It involves no human subjects, no personally identifying information, and no user study. All four benchmarks (MVTec AD, VisA, BTAD and Real-IAD) are public industrial-inspection image datasets used with their official category structures, and no new data were collected. Raw datasets and encoder weights are identified by hash rather than redistributed. Because the original BTAD repository and secondary documentation disagree on the dataset license, the raw BTAD files are used locally for non-commercial research only and are deliberately excluded from the released artifact; this conservative boundary is stated in the appendix rather than resolved here. The bounded audit of reference access in 12 cited papers, and the row-by-row coding of the comparison tables of 12 zero- and few-shot papers, record which quantitative tables place different access protocols side by side. Both are descriptive codings of published reporting practice on purposive samples, not allegations of misconduct, assessments of those methods' validity, or prevalence estimates for the field.

REPRODUCIBILITY STATEMENT

Section 3 defines the controlled fixed-k estimand and separates it from the complete-protocol full-pool comparison. The v4 study and the separately frozen BTAD endpoint stress test each use an immutable registry and support manifest in which every cell binds dataset, query, support, encoder, configuration, code, environment, and metric identities (Appendices A and B). Appendices C, I, J, and H state the fairness checks, recovery tests, mask semantics, and reporting checklist; Appendix W reports the cross-hardware rerun, the independent raw-to-metric recomputation, and the one numerical context the identity does not bind; Appendix M binds every reported number to its source, Appendix K adds an artifact-only check on a second implementation, and Appendix R states the registered decomposition protocol, its seam control, its deviations and its one interim look, Appendix X gives the registration status of every protocol and of the post hoc analyses, and Appendix Y lists the registered protocols not used for its main-text claims. The public datasets are used with their official category structures. The accompanying anonymous artifact contains the protocol-control code, frozen registries and support manifests, environment captures, tests, table generator, and hash-pinned BTAD aggregate evidence; its reporting pipeline accepts only the closed sets of verified cells. Raw datasets, model weights and the raw-score archive are identified by hash rather than redistributed.

REFERENCES

Ayusha Abbas, Saram Abbas, and Kabita Adhikari. Training-free human-in-the-loop anomaly detection via memory bank correction. *arXiv preprint arXiv:2608.17775*, 2026.

Mohammed Ahmed Mohammed Eltoum Mohammed Ali, Ehtesham Iqbal, Samee Ullah Khan, and Yusra Abdulrahman. RadioCore: Few-shot industrial anomaly segmentation with multi-scale Radio ViT features. In *CVPRW*, 2026.

Yuhu Bai, Jiangning Zhang, Yunkang Cao, Guangyuan Lu, Qingdong He, Xiangtai Li, and Guanzhong Tian. Bridge feature matching and cross-modal alignment with mutual-filtering for zero-shot anomaly detection. In *ICCVW*, 2025.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *CVPR*, pp. 4182–4191, 2020. doi: 10.1109/CVPR42600.2020.00424.

John Bronskill, Jonathan Gordon, James Requeima, Sebastian Nowozin, and Richard Turner. TaskNorm: Rethinking batch normalization for meta-learning. In *ICML*, volume 119 of *PMLR*, pp. 1153–1164, 2020.

Naci Cankaya. Bit-exact AI inference verification without performance tradeoffs. *arXiv preprint arXiv:2606.00279*, 2026.

Joongwon Chae, Runming Wang, and Peiwu Qin. What memory composition does not tell us about anomaly detection. *arXiv preprint arXiv:2608.23295*, 2026a.

Joongwon Chae, Runming Wang, and Peiwu Qin. What remains normal? Clean images miss useful near-defect normal patches for anomaly detection. *arXiv preprint arXiv:2608.23299*, 2026b.

Xuhai Chen, Yue Han, and Jiangning Zhang. APRIL-GAN: A zero-/few-shot anomaly classification and segmentation method for CVPR 2023 VAND workshop challenge tracks 1&2: 1st place on zero-shot AD and 4th place on few-shot AD. *arXiv preprint arXiv:2305.17382*, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023.

Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. AnomalyDINO: Boosting patch-based few-shot anomaly detection with DINOv2. In *WACV*, 2025.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges*, LNCS, pp. 475–489, 2021. doi: 10.1007/978-3-030-68799-1_35.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *EMNLP-IJCNLP*, pp. 2185–2194, 2019. doi: 10.18653/v1/D19-1224.

Alejandro Galan-Cuenca, Marcelo Saval-Calvo, and Antonio Javier Gallego. Are we really doing few-shot learning? A critical examination of pre-training assumptions. *arXiv preprint arXiv:2609.10851*, 2026.

Alexander Gammerman, Vladimir Vovk, and Vladimir Vapnik. Learning by transduction. In *UAI*, 1998.

Avijit Ghosh, Anka Reuel-Lamparth, Matthew Kennedy, Jenny Chim, Yilin Huang, Srishti Yadav, Anastassia Kornilova, Damian Stachura, Kevin Klyman, Jennifer Mickel, Felix Friedrich, Max Lamparth, Jan Batzner, Anoop Mishra, Jeba Sania, Eliya Habba, Nathan Heath, Shalaleh Rismani, Usman Gohar, Andrea Loehr, David Manheim, Ruchira Dhar, Sree Harsha Nelaturu, Aarush Sinha, Leshem Choshen, Yixiong Hao, Yanan Long, Andrew Tran, Drishti Sharma, Ishan Khire, Amit Saha, Subramanyam Sahoo, Michael Hardy, Michael A. Riegler, Kabir Manghnani,

Michelle Lin, Nancy Jiang, Asaf Yehudai, Nuno Moniz, Yacine Jernite, Zeerak Talat, Stella Biderman, Mubashara Akhtar, Jessica Ji, Aris Hofmann, Mykel J. Kochenderfer, Sanmi Koyejo, and Irene Solaiman. Evaluation Cards: An interpretive layer for AI evaluation reporting. In *NeurIPS (Evaluations and Datasets Track)*, 2026.

Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, Rama Chellappa, Andrew Gordon Wilson, and Tom Goldstein. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. In *NeurIPS*, 2023. Datasets and Benchmarks Track.

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 *CVPR*, pp. 15194–15203, 2025. doi: 10.1109/CVPR52734.2025.01415.

Seokhee Han, Seungjun Chu, Mateusz Nowak, and Peter Chin. SPARC: Subspace position-aware robust few-shot calibration for distribution-shifted industrial anomaly detection. *arXiv preprint arXiv:2608.18585*, 2026.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. ADBench: Anomaly detection benchmark. In *NeurIPS*, 2022. Datasets and Benchmarks Track.

Horace He and Thinking Machines Lab. Defeating nondeterminism in LLM inference. Thinking Machines Lab: Connectionism, September 2025. Technical blog post; not peer reviewed.

Jianfang He, Min Cao, Silong Peng, and Qiong Xie. RareCLIP: Rarity-aware online zero-shot industrial anomaly detection. In *ICCV*, 2025.

Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. The MVTec AD 2 dataset: Advanced scenarios for unsupervised anomaly detection. *IJCV*, 134(4): 175, 2026. doi: 10.1007/s11263-026-02743-0.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Yanning Hou, Peiyuan Li, Zirui Liu, Yitong Wang, Yanran Ruan, Jianfeng Qiu, and Ke Xu. VisualAD: Language-free zero-shot anomaly detection via vision transformer. In *CVPR*, 2026.

Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *ECCV*, pp. 303–319, 2022. doi: 10.1007/978-3-031-20053-3_18.

Po-Han Huang, Jeng-Lin Li, Po-Hsuan Huang, Ming-Ching Chang, and Wei-Chao Chen. PatchEAD: Unifying industrial visual prompting frameworks for patch-exclusive anomaly detection. In *WACV*, 2026. arXiv:2509.25856.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, 2023.

Hannaneh Kalantary and Javad Khoramdel. When more references hurt: Contamination-aware DINOv2 memory banks for few-shot steel defect detection. *arXiv preprint arXiv:2608.22082*, 2026.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Sayash Kapoor, Emily M. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica Hullman, Michael A. Lones, Momin M. Malik, Priyanka Nanayakkara, Russell A. Poldrack, Inioluwa Deborah Raji, Michael Roberts, Matthew J. Salganik, Marta Serra-Garcia, Brandon M. Stewart, Gilles Vandewiele, and Arvind Narayanan. REFORMS: Consensus-based recommendations for machine-learning-based science. *Science Advances*, 10(18):eadk3452, 2024. doi: 10.1126/sciadv.adk3452.

Tai Le-Gia. Toward training-free zero-shot anomaly detection in 3D medical images: A batch-based approach using 2D foundation models. *arXiv preprint arXiv:2606.18749*, 2026.

Tai Le-Gia and Jaehyun Ahn. On the problem of consistent anomalies in zero-shot industrial anomaly detection. *TMLR*, 2025.

Teven Le Scao and Alexander Rush. How many data points is a prompt worth? In *NAACL-HLT*, pp. 2627–2636, 2021. doi: 10.18653/v1/2021.naacl-main.208.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Camile Lendering, Erkut Akdag, and Egor Bondarev. SubspaceAD: Training-free few-shot anomaly detection via subspace modeling. In *CVPR*, 2026.

Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. Zero-shot anomaly detection via batch normalization. In *NeurIPS*, 2023. arXiv:2302.07849; earlier versions titled "Zero-Shot Anomaly Detection without Foundation Models" and "Zero-Shot Batch-Level Anomaly Detection".

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. PromptAD: Learning prompts with only normal samples for few-shot anomaly detection. In *CVPR*, pp. 16848–16858, 2024a. doi: 10.1109/CVPR52733.2024.01594.

Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. MuSc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *ICLR*, 2024b.

Xurui Li, Feng Xue, and Yu Zhou. MuSc-V2: Zero-shot multimodal industrial anomaly classification and segmentation with mutual scoring of unlabeled samples. *TPAMI*, 48(9):10713–10727, 2026. doi: 10.1109/TPAMI.2026.3688174.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: A vision transformer network for image anomaly detection and localization. In *ISIE*, 2021.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *FAT**, 2019.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *ECCV*, 2020.

OpenReview. Official reviews and author responses for "MuSc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images". OpenReview, ICLR 2024 public discussion, https://openreview.net/forum?id=AHgc5SMdtd, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

Jun Yeong Park, JunYoung Seo, Minji Kang, and Yu Rang Park. MoECLIP: Patch-specialized experts for zero-shot anomaly detection. In *CVPR*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021. Extended version: arXiv:2103.00020.

Fahim Rahman, S. M. Tanjeeb Meheran Rohan, Md. Taimum Ibne Sayed, Asaduzzaman Herok, and Md. Bakhtiar Hasan. Training-free spectral transductive refinement for cross-domain few-shot classification. *arXiv preprint arXiv:2609.23758*, 2026.

Philipp Röchner, Simon Klüttermann, Kevin Kammler, Franz Rothlauf, Emmanuel Müller, and Daniel Schlör. We need to rethink benchmarking in anomaly detection. In *NeurIPS (Position Paper Track)*, 2026.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022.

Peter J. Rousseeuw and Mia Hubert. Anomaly detection by robust statistics. *WIREs Data Mining and Knowledge Discovery*, 8(2):e1236, 2018. doi: 10.1002/widm.1236.

Lijun Sheng, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. The illusion of progress? A critical look at test-time adaptation for vision-language models. In *NeurIPS*, volume 38, pp. 138586–138614, 2025. doi: 10.52202/085713-4167. Datasets and Benchmarks Track.

Anthony F. Shorrocks. Decomposition procedures for distributional analysis: A unified framework based on the Shapley value. *The Journal of Economic Inequality*, 11(1):99–126, 2013. doi: 10.1007/s10888-011-9214-z.

Prabhant Singh, Sibylle Hess, and Joaquin Vanschoren. How NOT to benchmark your SITE metric: Beyond static leaderboards and towards realistic evaluation. In *ICLR*, 2026.

Harish Varma Siravuri and Hamed Alhoori. Cross-paper result comparability: A dataset, protocol-partitioned leaderboards, and a census. Zenodo dataset, doi:10.5281/zenodo.21715276, 2026.

Anna Sokol, Elizabeth Daly, Michael Hind, David Piorkowski, Xiangliang Zhang, Nuno Moniz, and Nitesh Chawla. BenchmarkCards: Standardized documentation for large language model benchmarks. In *NeurIPS*, 2025. Datasets and Benchmarks Track.

VAND Workshop Organizers. VAND 2025 evaluation framework. https://github.com/cvpr-vand/challenge, commit 6e42e42, 2025.

Olivier Veilleux, Malik Boudiaf, Pablo Piantanida, and Ismail Ben Ayed. Realistic evaluation of transductive few-shot learning. In *NeurIPS*, volume 34, pp. 9290–9302, 2021.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *CVPR*, pp. 22883–22892, 2024. doi: 10.1109/CVPR52733.2024.02159.

Guoyang Xie, Jinbao Wang, Jiaqi Liu, Yaochu Jin, and Feng Zheng. Pushing the limits of fewshot anomaly detection in industry vision: GraphCore. In *ICLR*, 2023.

Guoyang Xie, Jinbao Wang, Jiaqi Liu, Jiayi Lyu, Yong Liu, Chengjie Wang, Feng Zheng, and Yaochu Jin. IM-IAD: Industrial image anomaly detection benchmark in manufacturing. *IEEE Transactions on Cybernetics*, 54(5):2720–2733, 2024. doi: 10.1109/TCYB.2024.3357213.

Chaoran Xu, Chengkan Lv, Qiyu Chen, Feng Zhang, and Zhengtao Zhang. MRAD: Zero-shot anomaly detection with memory-driven retrieval. In *ICLR*, 2026.

Jiayi Yuan, Hao Li, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, and Zirui Liu. Understanding and mitigating numerical sources of nondeterminism in LLM inference. In *NeurIPS*, volume 38, pp. 188045–188077, 2025. doi: 10.52202/085713-5653.

Keqin Zhang and Sharon Lee. Beyond leaderboards: Protocol cards for trustworthy coding-agent evaluation. In *KDD 2026 Workshop on Evaluation and Trustworthiness of Agentic AI*, Jeju, Republic of Korea, August 2026. URL https://kdd-eval-workshop.github.io/agenticai-evaluation-kdd2026/.

Yi Zhang, Jiawen Zhu, Lele Fu, and Guansong Pang. AnomalyAgent: Training-free agentic models for zero-/few-shot anomaly detection. *arXiv preprint arXiv:2605.30140*, 2026a.

Yunbei Zhang, Janet Wang, Yingqiang Ge, Weijie Xu, Jihun Hamm, and Chandan K. Reddy. Stop comparing LLM agents without disclosing the harness. *arXiv preprint arXiv:2605.23950*, 2026b.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *ICML*, 2023.

Zhen Zhao, Fan Song, Xinyun Wang, Tianshun Yuan, and Jiali Zhou. Class-sensitive TPB-guided memory refinement for online zero-shot anomaly detection. *Sensors*, 26(11):3537, 2026. doi: 10.3390/s26113537.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *ICLR*, 2024.

Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *CVPR*, pp. 17826–17836, 2024. doi: 10.1109/CVPR52733.2024.01688.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. SPot-the-Difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*, 2022.

A FROZEN EXPERIMENT CONTRACT

Table 4: Cell counts of the two frozen protocols: v4 on MVTec AD and VisA, and the endpoint-only BTAD stress test, which is outside the v4 total. "Paths" are deterministic, predeclared support selections, not random seeds. Budgets and paths are per-category counts and are not summed in the Total row; – marks the absent full-pool arm.

Dataset	categories	budgets	paths	fixed-k cells	full-pool cells
MVTec AD	15	5	5	375	15
VisA	12	5	5	300	12
Total	27	5	5	675	27
BTAD stress test	3	2	5	30	–

Table 5: Complete fixed-k macro image-AUROC curves and the deterministic full-pool protocol (percent). Full-pool values are repeated only to make each path comparison readable.

Dataset	path	k=1	k=2	k=4	k=8	k=16	full pool
MVTec AD	p0	93.30	94.77	96.71	97.52	98.27	96.01
	p1	92.27	93.32	94.00	97.28	97.73	96.01
	p2	92.53	95.37	95.83	96.83	98.20	96.01
	p3	92.83	93.46	94.71	95.66	98.81	96.01
	p4	92.26	94.68	95.87	97.75	98.18	96.01
VisA	p0	85.41	87.92	90.01	91.95	92.72	87.99
	p1	87.96	88.57	90.92	92.53	93.08	87.99
	p2	86.52	90.19	91.37	92.55	93.62	87.99
	p3	86.66	89.66	91.93	92.86	93.51	87.99
	p4	87.90	90.29	91.03	92.41	93.06	87.99

The formal protocol identifier is `P000018-T1-fixed1nn-access-v4`. The immutable support manifest is checked before every execution. Its rows bind dataset, category, support path, rank, relative path, file size, and file SHA-256. Registry validation rejects missing, duplicate, extra, or identity-inconsistent rows and requires exactly the 675 fixed-k cells and 27 full-pool cells described in Table 4.

The frozen configuration uses OpenCLIP `ViT-L-14-336` (the open_clip name of ViT-L/14-336) with `openai` weights, input size 518×518, feature layers $\{5, 11, 17, 23\}$, and LNAMD radii $\{1, 3, 5\}$. Scoring is `MSM-only` with absolute q=1; interpolation is bilinear with aligned corners, and masks are binarized after resizing at $m > 0.5$. The encoder SHA-256 is `3035c92b350959924f9f00213499208652fc7ea050643e8b385c2dac08641f02`; the formal code and environment hashes are `2d0acfec58e704615d336e7d3d5b63ef2a29dc4fbdbc0ea02bf6b33a183b3d94` and `82c8274d260175d289838f43706f138d4aff3387d3d43ce6767219dc7c8de132`, respectively.

The frozen functional is a deliberately reduced instance of the antecedent it borrows from, MuSc (Li et al., 2024b), and the full-pool arm should not be read as a reproduction of any published headline number. Two declared choices separate them. First, this study fixes the absolute q=1 minimum reduction over reference images, whereas the antecedent aggregates a proportion of the lowest scores and reports its headline results under that proportional setting. Second, this study uses the mutual scoring stage alone and omits the antecedent's image-level re-scoring stage. The antecedent's own ablations report that each choice lowers image-level AUROC, and that both cost more on VisA than on MVTec AD. Both choices are frozen identically across the fixed-k and full-pool arms, so they cannot move the within-study contrast; they only mean that the absolute level of either arm is below what a complete published pipeline reports.

Table 6: Mean and minimum of the five predeclared path-wise k=16 minus k=1 differences (percentage points); every path-wise difference is positive. All metrics are recomputed from the same accepted raw artifacts; F1-max uses an oracle threshold and is diagnostic.

	MVTec AD		VisA	
Metric	mean	min	mean	min
Image AUROC	+5.60	+4.97	+6.31	+5.12
Image AP	+2.73	+2.39	+5.28	+4.35
Image F1-max	+2.90	+2.55	+5.80	+4.58
Pixel AUROC	+1.42	+1.24	+0.98	+0.74
Pixel AP	+8.85	+7.70	+7.02	+4.46
Pixel F1-max	+8.06	+6.67	+6.15	+3.77
AUPRO	+3.56	+2.95	+3.86	+3.22

The primary endpoint is dataset-specific equal-category macro image AUROC. Secondary metrics are image AP, image-level F1-max over thresholds, pixel AUROC, pixel AP, pixel-level F1-max over thresholds, and AUPRO, the area under the per-region-overlap curve (Bergmann et al., 2020). They are all recomputed from the same raw query-score and anomaly-map artifact rather than from separate runs. Full-pool results are deterministic single values and receive no artificial standard deviation. Table 6 summarizes these metrics, and Table 7 counts the category-level signs of the primary contrast.

Table 7: Category-level signs of the image-AUROC k=16 minus k=1 contrast. Counts use exact archived values before display rounding; a tie is an exactly zero difference, mostly where both budgets already reach 100 percent.

Dataset	path	positive	tie	negative
MVTec AD	p0	11	4	0
	p1	11	3	1
	p2	11	3	1
	p3	12	2	1
	p4	12	3	0
VisA	p0	12	0	0
	p1	12	0	0
	p2	12	0	0
	p3	11	0	1
	p4	12	0	0

Table 8 records the observed category-execution resources of this accepted v4 run. The logs are implementation-specific measurements of the exact runner rather than a general complexity benchmark, and the separately frozen BTAD extension below is not included in them.

Table 8: Observed category-execution resources for the accepted 702-cell run. Minutes are totals over executions, each covering all 26 cells of a category; the time range is the shortest and longest wall-clock time of one execution; host and device GiB are the largest measured peaks of CPU and accelerator memory over executions.

Dataset	executions	minutes	time range (s)	host GiB	device GiB
MVTec AD	15	34.45	46.21–215.76	16.62	5.47
VisA	12	52.83	192.30–337.65	19.51	5.83

B PROSPECTIVE BTAD ENDPOINT STRESS TEST

BTAD (Mishra et al., 2021) was selected after the v4 results closed, but its dataset manifest, three categories, five support paths, two endpoint budgets, scorer, metric code and all 30 registry rows were frozen before any BTAD performance was inspected, and the closed aggregate verifies them without failure or retry. All five path-level macro contrasts are positive (Table 9), averaging $+3.43$ percentage points (SD 1.19, range $+1.43$ to $+4.48$); of the 15 category–path contrasts 14 are positive and 1 is negative, category 02 on p2 by -0.73 points, and it is retained. The post hoc dataset choice keeps this a stress test on a third public dataset, not a replication.

Table 9: Separately frozen prospective BTAD endpoint stress test (macro image AUROC, percent; Δ in percentage points, from unrounded values). Each path uses the same three categories and the predeclared k=1 and k=16 clean-support prefixes; BTAD is outside the 702-cell v4 matrix.

Path	k=1	k=16	Δ
p0	92.66	96.00	+3.33
p1	91.66	95.45	+3.79
p2	92.26	96.75	+4.48
p3	93.73	95.16	+1.43
p4	91.63	95.73	+4.10
Mean $\pm$ SD (five paths)			+3.43 $\pm$ 1.19
Range (five paths)			[+1.43, +4.48]

The extension protocol is `P000018-T1-btad-access-v1`. It binds the verified BTAD dataset manifest, three official product categories, five deterministic hash-ranked support paths, the two endpoint budgets $k \in \{1, 16\}$, and 30 registry identities. It preserves the v4 absolute q=1 global reference-patch scorer, but uses a separate code/environment identity and a separate append-only operational ledger. The completed archive records 13/13 stages, 30/30 verified cells, and no failure or retry. The archive also keeps the source uploaded before execution; against it, the final code differs only in the orchestrator's download command, its virtual-environment creation and one added package, all edited before the first BTAD cell was written, while every scorer, model, data-loading and metric file is byte-identical to the upload. Table 10 gives the category-level differences behind that aggregate.

Table 10: BTAD category-level image-AUROC k=16 minus k=1 differences (percentage points). Exact archived values are used for sign accounting; displayed values are rounded only after subtraction.

Category	p0	p1	p2	p3	p4
01	+1.55	+1.36	+2.04	+0.78	+2.72
02	+5.83	+7.40	-0.73	+1.22	+4.38
03	+2.61	+2.60	+12.14	+2.30	+5.19

The evidence package has SHA-256 `14f764b0b6dd6546a4afe091c52c2e58898a1d7a aff2cafea156cf60e5f86c10`.

Its accepted aggregate, registry, support manifest, and dataset manifest are hash-pinned by the same manuscript generator as the v4 exports. The BTAD data archive and OpenCLIP weights remain external and are recoverable by recorded hashes. Because the original repository and secondary documentation disagree on the dataset license (CC BY-SA in the original repository,[1] CC BY-NC-

[1] https://github.com/pankajmishra000/VT-ADL

SA 4.0 in the Anomalib documentation[2]), raw BTAD files are used locally for non-commercial research and are not redistributed. This boundary also prevents the manuscript artifact from being described as a self-contained runtime.

C Access and fairness checklist

For every fixed-k cell, the verifier checks identical ordered queries; a clean training-normal support prefix; the frozen encoder and preprocessing; the same feature layers, radii, distance, functional, and metrics; and absence of unselected-support or other-query influence. For every full-pool cell, it checks complete category-cohort visibility, exact self-exclusion, and the cohort identity/order hash. Comparisons between these arms are labeled complete-protocol comparisons.

D Extended related work

The access question elsewhere. Test-time adaptation benchmarks for vision-language models separate episodic from online methods by the test data each may use (Sheng et al., 2025), and few-shot results shift with whether pre-training uses labeled in-domain data (Galan-Cuenca et al., 2026). One recent few-shot paper that ranks a transductive rule against inductive baselines states that its table is not "a fully setting-controlled ranking" (Rahman et al., 2026). FiSeCLIP adopts batch-level mutual references because full-dataset joint testing may be impractical (Bai et al., 2025); these are useful systems, but "training free" alone does not determine what references a method may inspect.

Documentation layers. Model Cards and BenchmarkCards ask authors to state intended use, benchmark provenance and known limits (Mitchell et al., 2019; Sokol et al., 2025), and REFORMS asks authors to justify that modeling uses no test-set information (Kapoor et al., 2024). A concurrent position paper asks LLM-agent evaluations to disclose their harness on a structured card and to compare models only under a locked harness or a factorial harness-by-model design (Zhang et al., 2026b), and a concurrent workshop paper shows that context construction, metric family and cost accounting change coding-agent conclusions, while verification depth has no reliable effect, and proposes protocol cards that record the information, procedure and measurement choices behind each score (Zhang & Lee, 2026). BudgetMatch is narrower than these layers and asks a different question of them.

Benchmark discipline. The fine-tuning protocol changes which backbones benefit (Goldblum et al., 2023); a same-venue evaluation audit converts such a critique into matched-resource recommendations and a reporting checklist (Singh et al., 2026). A position paper instead regroups anomaly-detection benchmarks into application scenarios (Röchner et al., 2026), which changes what belongs on one benchmark rather than whether two rows on it may be compared.

Inside mutual scoring. CoDeGraph finds and filters repeated, mutually reinforcing anomaly patterns in the test cohort (Le-Gia & Ahn, 2025). That fixes how one detector builds a usable reference set; we ask instead which comparisons between declared protocols are admissible at all. We do not claim every test cohort contains such patterns, and we do not use them to explain the RQ2 gap.

Published tables and reference quality. The second set of 12 coded papers shares six with the first. In eleven of the twelve comparisons whose top row sits outside the largest same-access group under every coding, that row trains on labeled auxiliary data rather than scoring against the test cohort, consistent with only 6 of the 54 verified differences lying in test-time access (Appendix G). Among the concurrent reference-quality studies, one controls for budget to isolate which candidates were chosen (Chae et al., 2026b); another holds a patch budget fixed and reports that unverified extra references can score below a clean bank of the same capacity (Kalantary & Khoramdel, 2026), while a companion study finds that the least contaminated memory is not the best-scoring one, so contamination alone does not order memories (Chae et al., 2026a); our purity terms are measured under named reductions and do not contradict that finding elsewhere. We hold the clean-support selection

[2] https://anomalib.readthedocs.io/en/latest/markdown/guides/reference/data/datamodules/image/btech.html

rule fixed for the primary contrast, because we do not intervene separately on which candidates, how clean, how visible, or how many. Han et al. (2026) set clean-reference detectors beside MuSc and concede that their comparison with purpose-built few-shot detectors "is not an information-matched head-to-head", and Abbas et al. (2026) report that corrected evaluation images entering the bank push AUROC toward one, so they hold out a disjoint split; both handle the access difference in prose rather than in a checkable field.

E BOUNDED AUDIT OF REFERENCE-ACCESS PROTOCOL SPECIFICATIONS

Table 11 records a post hoc, purposive audit of 12 methods already used to motivate or delimit the paper. Eight papers place more than one reference-access specification in the same quantitative comparison table, counting appendix tables. The audit therefore establishes the existence of mixed-access comparison practice in the cited literature, but its $8/12$ fraction is not a systematic prevalence estimate. The machine-readable rows, source locators, and scope notes are distributed with the artifact.

Table 11: Bounded audit of reference-access protocol specifications in 12 cited method papers. "Mixed" means that a quantitative table in the paper, appendices included, places methods with different reference-access specifications side by side; it is not a judgment of invalidity. Reference access records test-time references only; training data are coded separately, in Table 1 for its four methods and in Appendix G for the rest.

Method	Reference access	Query visibility	Evidence	Mixed
SPADE	clean training-normal bank	current query only	pp. 2–5; method and experiments	no
PaDiM	clean training-normal distribution	current query only	pp. 1–4; method and experiments	no
PatchCore	clean training-normal coreset	current query only	pp. 1–5; method and experiments	no
WinCLIP	no target-image reference or clean few-shot reference	current query only	pp. 6–7 and Tables 1, 2 and 4	yes
APRIL-GAN	no target-image reference or clean few-shot reference	current query only	p. 1 and Tables 3–4	yes
AnomalyCLIP	no target-image reference	current query only	pp. 1–6 and appendix Table 12	yes
MuSc	unlabeled complete test cohort	all category test images jointly	Table 1 and pp. 1–3	yes
RareCLIP	historical test-observation memory	current and past queries	Figure 1 and Tables 1–2	yes
AnomalyDINO	clean one/few-shot reference or unlabeled test cohort	current query or complete test cohort	pp. 2–3 and Table 4	yes
SubspaceAD	clean few-shot reference or unlabeled complete category test set	current query or complete test cohort	p. 6 and Tables 1–2	yes
FiSeCLIP	other images in the current test batch	current batch	pp. 1–4 and Table 1	yes
What Remains Normal	fixed-size clean or near-defect candidate bank	current query only	preprint pp. 2–4, p. 27 and Table 5	no

F Coding a published comparison

Table 1 codes the ten rows of the two MVTec AD blocks of Table 1 in Li et al. (2024b), which prints MuSc once in each block. Each row was coded on the access and resource cards of Section 3 from its own paper—Jeong et al. (2023), Chen et al. (2023), Li et al. (2023), Huang et al. (2022), Roth et al. (2022), Xie et al. (2023) and Li et al. (2024b)—with one evidence locator per field; the functional card is the method itself, and the evaluation card is shared by every row. Data a method draws on during training, beyond its pretrained weights, is coded under the access card's source field, because the specification defines source as the data a method may draw on; Li et al. (2024b) themselves note that APRIL-GAN trains on a labeled auxiliary dataset and that RegAD and ACR train on other categories of the evaluated dataset. The four-shot PatchCore row of that table repeats the values of WinCLIP's reimplementation, so it is coded with WinCLIP's few-shot protocol rather than with PatchCore's own.

A field a paper does not state is coded as unknown and counts as a difference. RegAD's support-selection rule and the data on which GraphCore's graph encoder is trained are such fields, and the pairs whose difference consists only of them are reported as undecided rather than resolved either way. ACR is coded as seeing its category's whole test set through batch statistics, following its

description; had it used smaller test batches, its difference from MuSc would grow by one field without changing any grouping.

The coding describes what each paper reports and passes no judgment on any method: a shared setting label is a reporting convention that the table makes visible, not an error of the paper that printed it. The APRIL-GAN row of Table 11 records test-time reference access only; coded on all eight fields here, it also carries the labeled auxiliary data used to train its projection. The machine-readable coding, the evidence for every field and the script that recomputes the table are distributed with the artifact.

G CODING EVERY PRINTED COMPARISON OF THE AUDITED PAPERS

One coded table shows that the check can be run; it does not show how often it matters. This post hoc extension, not registered, codes the main image-AUROC comparison tables on MVTec AD and VisA of 12 zero- and few-shot papers: six whose own methods appear in the table coded above, and six more of the same family that print such comparisons (AnomalyCLIP, AnomalyDINO, PromptAD, InCTRL, SubspaceAD and UniVAD; Zhou et al., 2024; Damm et al., 2025; Li et al., 2024a; Zhu & Pang, 2024; Lendering et al., 2026; Gu et al., 2025). For SubspaceAD we code its few-shot and batched blocks, not its zero-shot rows. The papers were chosen, not sampled. Their tables hold 66 comparisons, one per (paper, table, setting, dataset), and 374 rows. Each row is coded on the access and resource cards under the protocol of the run that produced its number—the method's own paper, or the paper that re-ran it—with an evidence locator for every field, and the unchanged code of Appendix F computes the difference set of every pair of rows in a comparison. As there, a field a paper does not state is unknown: it can make a pair undecided, but never certifies that two rows are equal. The coding of Table 1 is recomputed as a regression anchor and returns the same pair counts.

Table 12: The coding extended from one published table to every image-AUROC comparison in the main MVTec AD and VisA tables of 12 zero- and few-shot papers, a set that shares six papers with the access audit of Table 11 (post hoc). A comparison is one (paper, table, setting, dataset) block; each row is coded on the access and resource cards under the protocol of the run that produced its value, with an evidence locator per field. Verified: the block holds two rows under its one setting label whose recorded fields differ in at least one stated value. Reversal: the top row lies outside the largest group of rows with identical signatures under that label, differs from that group's best row in a stated field, and no reading of the unstated fields could make its own group as large. Robust: a reversal in the primary coding and in all three sensitivity codings (values traced to another shot count, anomaly-type names ignored, the two draw rules merged).

Paper	comparisons	rows	verified	reversals	robust
ACR NeurIPS2023	1	4	1	0	0
APRIL-GAN arXiv2305.17382v3	8	34	8	5	3
AnomalyCLIP ICLR2024	2	12	2	0	0
AnomalyDINO WACV2025	14	80	13	4	3
GraphCore ICLR2023	4	41	4	0	0
InCTRL CVPR2024	6	42	6	6	6
MuSc ICLR2024	4	17	4	0	0
PromptAD CVPR2024	6	38	6	6	0
RegAD ECCV2022	3	9	0	0	0
SubspaceAD CVPR2026	8	53	8	0	0
UniVAD CVPR2025	2	14	2	0	0
WinCLIP CVPR2023	8	30	0	0	0
Total	66	374	54	21	12

In 54 of the 66 comparisons, two rows printed under the same setting label differ in at least one stated access or resource field (Table 12). Of the 1012 pairs of rows that share a comparison, 244 (24.1%) differ in the method alone; the rest differ in a stated field, or in a field that is unstated or that we could not code from the method's own paper. The stronger question is whether the top row of a comparison owes its place to a difference in access. We call a comparison a reversal when its top row lies outside the largest group of rows that share one signature, differs from that

group's best row in a stated field, and no reading of the unstated fields could make the top row's own group as large. In detail, a signature is the exact tuple of the eight coded fields; a comparison has no reversal when every signature group among the rows under its label is a singleton, when two groups tie for largest, or when the top row's own label differs from the comparison's; tied top rows must each qualify; and the unstated-field condition compares the number of label rows verifiably compatible with the top row against the largest group's certain size (the full rule is stated in the artifact's `recompute_multitable.py`). In the primary coding 21 comparisons meet this, and 12 do so in all four codings (19, 15 and 20 in the three sensitivity codings). Eleven of the twelve are of one kind: the top row trains on labeled data beyond the evaluated category's normal images—APRIL-GAN in three VisA few-shot comparisons of its own paper and two of AnomalyDINO's, InCTRL (Zhu & Pang, 2024) in six comparisons of its own Table 1—while the best row of the largest same-signature group does not. In the twelfth, a zero-shot comparison in AnomalyDINO's Table 4, the top row draws on less data than the group it outranks.

Three qualifications belong with these counts. A reversal says that the ordering printed under one label is not a single-factor comparison; it says nothing about which method is better, and the margins are often small (InCTRL leads the re-run WinCLIP rows of its own Table 1 by 0.5 to 1.9 points). In the two AnomalyDINO comparisons the rows repeat values first printed by APRIL-GAN, so the difference originates in that earlier table and is carried forward under a new setting label. And thirteen methods could not be coded from their own papers; recoding each of their rows as the signature of any other row in its comparison, or as a new signature, removes no robust reversal in any of the four codings. The 75 values of the robust comparisons were re-read against the source PDFs and all matched. Together with Appendix E, whose 12 papers share six with these, this bounds the practice within the audited papers; it is still not a prevalence estimate for the field.

The verified differences are mostly not the ones this paper measures. Of the 54 comparisons with verifiably different access under one label, 48 differ in the source or purity of the data a method trains on or draws references from but in no test-time field (some of them also in the selection or size of the references), and 6 in test-time access (query visibility, self-exclusion or cohort coupling); none of the 21 reversals involves a test-time field. Test-time access is instead undeterminable in 30 comparisons, each of which lists a method that could not be coded from its own paper; for those rows the coding cannot say whether they score against the test cohort, and the count measures the reach of this coding, not what the papers leave unstated. One printed value could not be traced: the 8-shot PaDiM row of GraphCore's Table 2 is attributed to Roth et al. (2022), whose low-shot study reports no 8-shot setting; the row is coded under the protocol it claims.

H MACHINE-READABLE PROTOCOL SPECIFICATION AND REPORTING CHECKLIST

The anonymous artifact includes `protocol/protocol_signature.json` and `protocol/verify_protocol_signature.py`. The validator flattens the four cards, computes the exact field-difference set, confirms that the k=1 versus k=16 contrast changes only the declared clean-reference-budget field, and confirms that fixed-k versus full-pool is labeled a complete-protocol comparison. For the released specification it returns a single field for RQ1, `resource.reference_budget`, and the following eight for RQ2:

- `access.source`, `access.purity`, `access.query_visibility`, `access.self_exclusion`, `access.cohort_coupling`;
- `resource.reference_budget`, `resource.selection`, `resource.reuse_update`.

Table 13: Reference-access protocol specification of the two arms, card by card: the first five rows are the access card, the next three the resource card, then the functional and evaluation cards. Equality in the final column refers to the frozen implementation; "no" marks a field in the RQ2 difference set D.

Field	Fixed-k clean support	Self-excluded full pool	Equal?
Source	training-normal split	category test cohort	no
Purity	split-verified normal	unlabeled mixed cohort	no
Query visibility	current query only	complete category cohort	no
Self-exclusion	disjoint train/test splits	query-specific leave-one-out	no
Cohort coupling	none	other query identities affect score	no
Reference budget	k nested images	$\lvert Q_{dc}\rvert - 1$ images	no
Support selection	hash-ranked path prefix	all other category queries	no
Reuse/update	fixed bank reused across queries	query-specific bank	no
Encoder/scorer	frozen encoder and global reference-patch 1-NN	identical	yes
Evaluation	same test cohort, aggregation, and metrics	identical	yes

These are exactly the unequal rows of Table 13, so the printed specification and the executable certificate report the same difference set. The certificate is a reporting and consistency control; it does not replace implementation tests or establish causal identification by itself.

For a reference-based detector, a leaderboard row should report at least:

- reference source, split, and purity guarantee;
- query visibility, order assumptions, self-exclusion, and cohort coupling;
- reference-image and patch budgets and storage;
- support-selection rule and any outcome-dependent filtering;
- reuse and update policy;
- encoder, preprocessing, scoring functional, and aggregation;
- evaluation cohort, metric version, reporting unit, and uncertainty unit; and
- the predeclared intervention fields, observed difference set, and immutable identities.

I CORRECTNESS AND RECOVERY TESTS

The released test suite contains ten protocol-level checks, from a hand-computable NumPy oracle of the scoring functional upward, and four formal-operation checks. The latter validate the exact

registry matrix, the shape of the formal code identity, an interruption drill using the real scheduler with a deterministic fake runner, and an independent recomputation of the formal raw metrics. The drill first completes one category, interrupts the next category during execution, restarts from the same append-only ledger, and verifies one execution for the completed category and two for the interrupted category. The scheduler permits at most two attempts per category, and every attempt is written to that ledger. A scheduler cannot infer success from a process exit alone: it invokes the runner's artifact verifier before skipping or accepting a category.

J MASK SEMANTICS

MVTec segmentation masks are resized with bilinear interpolation in the upstream data path. Casting the resulting fractional values directly to integers silently removes boundary pixels. The formal metric version instead applies the frozen predicate $m > 0.5$. The threshold is part of the configuration and metric identities, and the raw binary mask is retained for independent recomputation. Earlier engineering artifacts produced with integer truncation remain archived but are explicitly superseded and excluded from formal evidence.

K ARTIFACT-ONLY COMPLETE-PROTOCOL EVIDENCE

An archived AnomalyDINO run family, outside every registered protocol, provides a post hoc, deliberately bounded cross-implementation check. An independent, artifact-only recomputation matched all 459 metric rows and all 34 macro rows within 10^{-12}, using 66,079 raw NumPy score files. Its official few-shot slices are deterministic but generally non-nested: 216 of 324 adjacent slice pairs violate nesting. Accordingly, Table 14 is complete-protocol evidence only: the arms are compared as they were run rather than by intervening on one field. It neither identifies reference-source causality nor independently confirms the formal 702-cell result, and it does not rescue pixel-level segmentation.

Table 14: Artifact-only complete-protocol classification evidence from archived AnomalyDINO outputs (macro image AUROC, percent): the arms are compared as they were run, not by intervening on one field. The transductive column scores each test image against all other test images of its category; full normal uses the complete training-normal split. The k columns average three deterministic, generally non-nested support slices; they are not random-seed uncertainty estimates.

Dataset	transductive	k=1	k=16	full normal
MVTec AD	92.15	96.25	98.00	98.87
VisA	82.25	81.78	90.71	95.11

L SECONDARY AGGREGATION SENSITIVITY CHECK

The formal v4 scorer uses the minimum over reference images for each query patch. To test sensitivity to this single reduction choice, we reran the same frozen encoder, preprocessing, feature layers, support/query identities, five deterministic paths, and $k \in \{1, 2, 4, 8, 16\}$ grid, changing only the reduction over reference images, under the separately identified protocol `P000018-T1-secondagg-median-image-v1`; its runner survives and is shipped, archived after the run and not hash-registered, but no separate registration document does (Appendix X). With the per-reference distances $a^{\lambda\rho}_{ipj}$ and upsampling Up of Section 3.2, the secondary score is

$$\widetilde{s}_{ip}(J) = \frac{1}{|\Lambda||\Gamma|} \sum_{\lambda,\rho} \mathrm{median}_{j \in J}\, a^{\lambda\rho}_{ipj}, \qquad \widetilde{S}_i(J) = \max \mathrm{Up}\big(\widetilde{s}_{i\cdot}(J)\big).$$

The sample median is an order-based location statistic that resists almost half of the values being outliers (Rousseeuw & Hubert, 2018). For an even number of values the textbook median is the mean of the two middle order statistics; the implementation instead returns the lower of the two, which

is what PyTorch's `torch.median` is documented to return,[3] and that is therefore the reduction actually applied at every even support size $k \in \{2, 4, 8, 16\}$. We state the implemented convention rather than the textbook one because the two differ and the released code fixes the former. At k=1 the median equals the minimum, yet this rerun groups the distance computation into different tensor shapes from v4, so float32 accumulation order differs and its k=1 category AUROC is not always bit-identical to v4: 66 of the 135 category–path cells agree exactly and the largest absolute disagreement is 0.28 percentage points. The differences in Table 15 are therefore paired within this branch. This branch was executed after the formal v4 closeout and is therefore an exploratory, image-level secondary check rather than a new primary estimand. Pixel metrics were intentionally not run; the branch does not establish pixel-level robustness, category-wise monotonicity, or source-only causality.

In Table 15, the direction of the macro image-AUROC change is positive for every planned path on both datasets, with mean changes of $+1.41 \pm 0.64$ percentage points on MVTec AD and $+4.12 \pm 1.14$ percentage points on VisA. Image AP and image-level F1-max over thresholds show the same directional pattern in the macro summaries (Table 16). These values are reported as a sensitivity diagnostic only: they reuse the public benchmark cohorts and the same support paths as v4, and they must not be read as an independent confirmation run.

Table 15: Secondary image-level check using the median of per-reference nearest-patch distances (the lower median at even k, as `torch.median` returns). Values are k=16 minus k=1 differences in macro image AUROC, in percentage points, computed from the same five predeclared support paths. This branch is exploratory and is not part of the primary v4 estimand.

Dataset	p0	p1	p2	p3	p4	mean ± SD	range
MVTec AD	+0.56	+1.59	+1.66	+1.01	+2.22	+1.41 ± 0.64	[+0.56, +2.22]
VisA	+5.34	+2.84	+4.59	+4.84	+2.98	+4.12 ± 1.14	[+2.84, +5.34]

Table 16: Secondary image-level metrics for the median aggregation check. Entries are mean k=16 minus k=1 differences across the five predeclared support paths, in percentage points. F1-max uses an oracle threshold and is diagnostic.

Dataset	Image AP	Image-level F1-max
MVTec AD	+0.87	+1.43
VisA	+3.48	+3.15

M Data provenance and reporting closure

The accepted formal inputs are the clean 702/702 artifact audit, the registry-exact macro and cell-level metric exports, and the terminal formal summary. The separate BTAD inputs are its 30-row closed aggregate, decision, registry, support manifest, verified dataset manifest, and archive closeout record. Their SHA-256 digests are recorded in `paper/audit/provenance/canonical-results.json`; the frozen registry digest begins `ecbacfd27acd5e24`. The archived AnomalyDINO evidence is separately bound to its rescue audit, metric recomputation, and run-summary digests. A project-owned generator checks all input hashes, recomputes the displayed transformations, applies decimal round-half-up at two decimals, and emits the v4, BTAD, access-audit and AnomalyDINO tables; companion generators, one per analysis, emit Table 1, Table 12, Tables 15–27, and Tables 28–38 (the decomposition, its registered follow-ups and the post hoc analyses) with their own cell provenance. The frozen analysis scripts these generators read store statistics at three decimals; to avoid rounding twice, a stored value whose third decimal is 5 is replaced by its unrounded value, from a rerun of the same script, before it is rounded once. Tables 3, 39, 4,

[3] https://docs.pytorch.org/docs/stable/generated/torch.median.html: "In this case the lower of the two medians is returned."

13, 40 and 41 are set by hand; the only numbers in Tables 3 and 39 are macros emitted by those generators, the counts in Table 4 are checked against the registry by the non-result provenance audit, and Table 40 records evidence states and hash prefixes, not results.

Every result-bearing table cell and quantitative-figure coordinate has a project-owned provenance row with its result ID, source, full hash, selector, transformation, and rounding rule. The result–claim ledger binds each displayed occurrence to generated TeX, and numeric inventories cover the assembled source and final PDF. Two figures are active. In Figure 1, panels (a) and (b) are a non-result schematic, while panel (c) is generated from the shot-equivalent ledger behind Table 20 with its own coordinate sidecar; the budget plot in Figure 2 is generated from a hash-bound coordinate sidecar that uses the same canonical results as Table 5. Historical unused figures remain excluded.

No result cell is outcome-filtered, and neither frozen protocol adds a replacement backbone, scorer, budget, path, or metric after its own outcomes. BTAD was selected after v4 but frozen before its own performance was inspected, so it is reported only as a prospective external stress test.

N Reduction sensitivity of the full-pool arm

Appendix L varies the reduction inside the clean fixed-k arm. It leaves open the complementary question, which a reader is entitled to ask of Table 5: is the position of the full-pool arm an artifact of the minimum, the statistic most exposed to a cohort that contains anomalies? Under the separately registered protocol `P000018-T1-fullpool-reductions-v1`, we recomputed the self-excluded full-pool arm for all 27 categories with the encoder, preprocessing, feature layers, cohort and self-exclusion unchanged, emitting three reductions from one and the same per-reference-image distance tensor: the minimum, the sample median, and the interval average that MuSc publishes as its own default, transcribed from the released implementation at its pinned commit with its published parameters $(0, 0.3)$ (Table 17). The registration fixed the four arms, a per-cell first-crossing shot-equivalent within each reduction and a reduction-robustness verdict, and it barred any contrast between the interval-average arm and a clean curve; Appendix O uses a different, post hoc definition. Applied after the run, and without a registered tolerance, the verdict step judged the two reductions' median shot-equivalents to agree on both datasets; the within-reduction shot-equivalents, including how many cells never cross, and the fp32-versus-fp16 precision contrast are archived with the registration rather than reported here.

Three readings follow, and the third bounds the first two.

First, the rerun reproduces the closed v4 full-pool cells exactly. Its fp16 minimum path, which uses the precision of the original full-pool computation, reproduces all 27 category AUROC values of the closed v4 cells bit for bit (listed per category in Table 18), and all 5,321,303 underlying patch scores are bit-identical as well, although the rerun used a different accelerator, driver, CUDA release and PyTorch build, and the machine that produced the original cells had already been decommissioned. We report this as reproducibility evidence and claim nothing further from it.

Second, at the level of the reported estimand the minimum is not a handicap, read post hoc for the median and as the registered analysis also prints for the interval average: replacing it with the contamination-resistant median costs the full-pool arm 3.06 points on MVTec AD and 3.69 on VisA, and MuSc's own interval average gains only 1.38 and 1.57. The ordering is mechanical rather than surprising. In a cohort that contains anomalies the smallest distance is what identifies a normal patch, since a normal patch finds a close counterpart somewhere; a median over every other cohort image is diluted by the many dissimilar ones. MuSc's interval average is built from the smallest distances, so it refines the minimum rather than the median, which is consistent with its macro value sitting above both.

Third, a post hoc per-category view shows that the macro statement hides its own exceptions and we report them rather than summarize them away. The median is in fact better in 14 of the 27 categories, by as much as 18.48 points on `metal_nut`; the macro favors the minimum only because a few categories collapse under the median, by 48.80 points on `pcb1` and 39.31 on `screw`. The interval average is worse than the minimum in 9 of 27. No claim that one reduction dominates survives the per-category view, and none is made here; Table 18 lists all 27.

Table 17: Reduction sensitivity of the self-excluded full-pool arm. Equal-category macro image AUROC (percent) with the encoder, preprocessing, feature layers, cohort and self-exclusion held fixed; only the reduction over reference images changes. The minimum (rerun) column is computed in fp32 like the other two reductions; a separate fp16 minimum path of the same rerun, not shown, reproduces the closed v4 full-pool cells bit for bit on different hardware and a different PyTorch build. The interval average is MuSc's own published default $(0, 0.3)$, which averages the per-reference distances ranked in the lowest 30 percent of the bank, applied to the full-pool bank; it changes the scoring functional, so that column is a complete-protocol reference point, not a term in any single-factor contrast.

Dataset	v4 minimum	minimum (rerun)	median	interval average
MVTec AD	96.01	96.01	92.95	97.39
VisA	87.99	87.99	84.30	89.56

Table 18: Per-category values behind the macro of Table 17. The minimum column is category image AUROC (percent) under the minimum; the other two columns are differences from it in percentage points. The median is better in 14 of 27 categories even though the macro favors the minimum, and the interval average is worse in 9 of 27. No category is removed.

Dataset	Category	minimum	median − min	interval avg. − min
MVTec AD	bottle	99.92	−0.24	−0.16
MVTec AD	cable	97.32	−0.92	+1.74
MVTec AD	capsule	95.57	−30.39	+1.64
MVTec AD	carpet	97.35	+2.45	+2.13
MVTec AD	grid	99.67	−1.42	−1.17
MVTec AD	hazelnut	94.82	+4.21	+4.68
MVTec AD	leather	99.42	+0.58	+0.58
MVTec AD	metal_nut	81.52	+18.48	+18.08
MVTec AD	pill	93.37	+2.35	+3.27
MVTec AD	screw	95.47	−39.31	−13.75
MVTec AD	tile	98.99	+1.01	+1.01
MVTec AD	toothbrush	100.00	−5.83	−1.11
MVTec AD	transistor	97.46	−1.21	+0.29
MVTec AD	wood	91.23	+5.88	+3.51
MVTec AD	zipper	98.08	−1.55	−0.08
VisA	candle	90.96	−0.12	+1.72
VisA	capsules	82.25	+4.87	+3.60
VisA	cashew	88.40	+5.10	+3.92
VisA	chewinggum	95.24	+2.30	+2.26
VisA	fryum	95.06	+2.58	+2.44
VisA	macaroni1	83.46	+4.24	+4.20
VisA	macaroni2	52.89	+10.34	+7.83
VisA	pcb1	90.21	−48.80	−2.02
VisA	pcb2	89.87	−8.96	−2.78
VisA	pcb3	96.81	−8.05	−1.90
VisA	pcb4	97.36	−12.72	−4.94
VisA	pipe_fryum	93.32	+4.98	+4.54

The placement below is post hoc and falls outside the registration, which barred it. The interval average changes the scoring functional and therefore cannot enter a single-factor contrast with the clean curves. There is exactly one budget at which the two arms can be compared under an identical declared functional; the comparison remains complete-protocol, because the eight access and resource fields of Table 13 still differ. With the published parameters the admitted interval at budget k holds $\lfloor 0.3k \rfloor$ images, so it is empty for $k \in \{1, 2\}$ and holds a single image at $k{=}4$, where the interval average therefore coincides with the minimum and the frozen $k{=}4$ column already is the clean arm's interval-average value.

Table 19: Functional-matched placement at k=4. Under MuSc's interval average with its published $(0, 0.3)$, the clean arm's admitted interval at k=4 contains $\lfloor 0.3 \cdot 4 \rfloor = 1$ image, so the reduction coincides with the minimum there and the frozen k=4 values are already the clean arm's interval-average values. This is the one budget at which the two arms can be compared under the identical declared functional; $k \in \{1, 2\}$ are degenerate because the admitted interval is empty. Values are equal-category macro image AUROC (percent). Only the declared functional is matched: the aggregation width is not, and the eight access and resource fields still differ, so the comparison remains complete-protocol. Post hoc; the registration barred this comparison.

Dataset	full pool	p0	p1	p2	p3	p4
MVTec AD	97.39	96.71	94.00	95.83	94.71	95.87
VisA	89.56	90.01	90.92	91.37	91.93	91.03

In Table 19 the full-pool arm on VisA lies below every one of the five predeclared clean paths at that matched budget. Under the audited method's own reduction, applied to each arm's own bank, access to the whole self-excluded cohort of 149 to 200 unlabeled images is therefore worth less than four verified-normal training images on VisA.

Two properties of that statement should be read off it rather than into it. First, matched here means that both arms declare the same functional with the same parameters, not that the operator does the same amount of averaging on each side: because the admitted interval is a fixed fraction of the bank, it spans a single image at k=4 and between 44 and 60 images across the VisA cohorts. The declaration is matched; the aggregation width is not, and it cannot be, since bank size is the field under comparison. Second, on MVTec AD the arm lies above the k=4 column, so the matched comparison there brackets its position only from below. Closing that bracket would require the clean arm at $k \in \{8, 16\}$ under the interval average, and we deliberately did not run it: at those budgets the admitted interval holds two and four images, so reference count and aggregation width would move together and the resulting placement would confound the two, which is the same defect that keeps a 0.3-fraction sweep out of the registered primary contrast. A published-parameter caveat belongs here as well. The upstream implementation computes the interval size as $\lfloor 0.3M \rfloor$ for a bank of M images without a lower guard, so for $M \leq 3$ it takes an empty interval and returns a non-finite score; our transcription raises instead of returning silently. That bound never binds in what we report, because the interval average is applied only to the full-pool arm, whose banks hold between 41 and 200 images, and to the k=4 clean column, where it reduces to the minimum. The clean arm at $k \in \{1, 2\}$ is reported under the minimum only.

Three limits apply to this appendix. The interval-average arm omits MuSc's image-level re-scoring stage, so it reproduces that method's patch-level component and not the method; its values remain below MuSc's published figures. Every quantity is image-level, since pixel metrics were deliberately not computed. And the comparison is between reductions of one frozen representation on two public benchmarks, not between methods.

O PLACING THE FULL-POOL ARM ON THE CLEAN BUDGET GRID

The comparison this paper audits has a natural unit. If a detector that scores against the whole unlabeled cohort is placed in the same table as detectors given k clean reference images, the obvious question is what k the first one is worth. The construction is not new: Radford et al. (2021) express zero-shot performance as the number of labeled examples per class a linear probe on the same features needs to match it, interpolating log-linearly over one to sixteen shots and fully supervised results (Figure 7 of the extended version, arXiv:2103.00020), and related units price a prompt or pretraining in data points (Le Scao & Rush, 2021; Hernandez et al., 2021). We apply it to reference access. We call the crossing the *shot-equivalent* of the full-pool arm, and compute it for every category from the closed v4 cells. The declared definition averages the category's image AUROC over the five support paths at each budget and interpolates linearly in k to the budget at which that curve first reaches the category's full-pool value; it is 1 when the one-image average already reaches it, and "not reached" when the sixteen-image average does not. It is a per-category quantity, unlike

Table 20: Shot-equivalent of the self-excluded full-pool arm: the interpolated clean budget k at which the path-averaged clean curve first reaches the full-pool value of the same category (declared definition: average the five support paths at each k, then interpolate linearly in k; 1.00 means the k=1 average already reaches it). The k=1 clean and full-pool columns give category image AUROC (percent), the former averaged over the five paths. The distribution is bimodal. In most categories a single clean reference image already matches the arm that sees the whole cohort; in a minority, sixteen do not. The quantity is post hoc and descriptive and carries the same reporting status as the full-pool arm itself. A dagger marks categories whose path-averaged clean value at any budget $k \leq 16$, or whose full-pool value, reaches 99.5, where the metric has little headroom left to resolve a difference.

Dataset	Category	k=1 clean	full pool	shot-equivalent
MVTec AD	bottle†	99.83	99.92	8.00
MVTec AD	cable†	93.64	97.32	3.80
MVTec AD	capsule	71.50	95.57	13.24
MVTec AD	carpet†	99.86	97.35	1.00
MVTec AD	grid†	97.33	99.67	not reached
MVTec AD	hazelnut†	96.19	94.82	1.00
MVTec AD	leather†	100.00	99.42	1.00
MVTec AD	metal_nut†	99.85	81.52	1.00
MVTec AD	pill	95.90	93.37	1.00
MVTec AD	screw	56.13	95.47	not reached
MVTec AD	tile†	100.00	98.99	1.00
MVTec AD	toothbrush†	92.67	100.00	not reached
MVTec AD	transistor	92.89	97.46	8.34
MVTec AD	wood	97.91	91.23	1.00
MVTec AD	zipper	95.90	98.08	6.35
VisA	candle	92.42	90.96	1.00
VisA	capsules	85.07	82.25	1.00
VisA	cashew	94.53	88.40	1.00
VisA	chewinggum	97.59	95.24	1.00
VisA	fryum	95.17	95.06	1.00
VisA	macaroni1	84.15	83.46	1.00
VisA	macaroni2	58.98	52.89	1.00
VisA	pcb1	89.91	90.21	3.29
VisA	pcb2	78.48	89.87	11.09
VisA	pcb3	75.09	96.81	not reached
VisA	pcb4	93.53	97.36	6.32
VisA	pipe_fryum†	97.73	93.32	1.00

the dataset-level crossing k^*_{dr} of Section 3. No new measurement is involved: the cells were already accepted, and the accompanying artifact rederives every number below from them.

Two things make the quantity worth reporting. The first is that the question is asked in practice and left unanswered. In the public review of Li et al. (2024b), three of the five reviewers questioned scoring against unlabeled test images (OpenReview, 2024). One objected that using them to score collectively differs from evaluating each image independently, and that this "can result in a somewhat unfair comparison"; another asked whether the method could be extended to the few-shot setting. The authors replied that the difference "may lead to a potentially unfair comparison", added a row for another method that also scores a set of unlabeled test images, and provided a few-shot appendix whose ladder, read against their zero-shot row, already implies a crossing. The public record of that review contains no such computation. The second is that when we do compute it, the answer is not a constant.

The distribution is bimodal rather than centered. Under the declared definition, one clean reference image already matches or beats the full-pool arm in 15 of 27 categories, while in 4 categories sixteen clean images still do not reach it. The categories that need more than one image are structured rather than random: all four VisA printed-circuit-board categories do, pcb3 never crosses, and neither do

screw, grid and toothbrush on MVTec AD. The same heterogeneity appears in the paired differences of Section 4; the shot-equivalent simply gives it a unit.

The unit measures the full-pool arm of this paper, a reduced MuSc that keeps its mutual scoring over the cohort but takes the minimum distance and drops the image-level re-scoring stage. To see where the method itself lands, we re-analyzed, post hoc and without any new run, archived runs of the unmodified official MuSc program with the same encoder, feature layers, radii and image size (Table 21). They reproduce its published macro image AUROC closely, 97.72 on MVTec AD and 92.52 on VisA against 97.8 and 92.8 (Li et al., 2024b). Placed on the same clean curves, the official method is worth one clean image or fewer in only 4 of 27 categories and is not reached by sixteen in 8; it scores above the reduced arm in 21. The two placements differ in the scoring functional as well as in access, so neither is a single-factor comparison. What they show together is the point of this appendix: the number of clean images the unlabeled cohort is worth is a property of the scorer that uses it, and a table that prints a cohort-scoring row beside few-shot rows implies a shot-equivalent it never declares.

Table 21: The official MuSc method (archived runs of its unmodified program: divide_num $= 1$, the same encoder, layers, radii and image size, batch size four), placed post hoc on the frozen 1-NN clean curve. Official: its image AUROC (percent); 1-NN pool: the self-excluded full-pool arm of this paper; the last two columns give each one's shot-equivalent under the declared definition (1 = one clean image suffices; nr = not reached by k=16). The official method's functional differs from the curve's, so every placement here is complete-protocol.

Dataset	Category	official	1-NN pool	shot-eq. (official)	shot-eq. (1-NN pool)
MVTec AD	bottle	99.92	99.92	8.00	8.00
MVTec AD	cable	99.03	97.32	12.20	3.80
MVTec AD	capsule	95.65	95.57	13.33	13.24
MVTec AD	carpet	99.88	97.35	1.75	1.00
MVTec AD	grid	98.75	99.67	8.24	nr
MVTec AD	hazelnut	99.61	94.82	4.34	1.00
MVTec AD	leather	100.00	99.42	1.00	1.00
MVTec AD	metal_nut	96.87	81.52	1.00	1.00
MVTec AD	pill	96.40	93.37	1.88	1.00
MVTec AD	screw	82.05	95.47	11.83	nr
MVTec AD	tile	100.00	98.99	1.00	1.00
MVTec AD	toothbrush	100.00	100.00	nr	nr
MVTec AD	transistor	99.37	97.46	nr	8.34
MVTec AD	wood	98.51	91.23	nr	1.00
MVTec AD	zipper	99.79	98.08	nr	6.35
VisA	candle	96.75	90.96	nr	1.00
VisA	capsules	88.32	82.25	2.07	1.00
VisA	cashew	98.40	88.40	nr	1.00
VisA	chewinggum	98.12	95.24	12.14	1.00
VisA	fryum	98.62	95.06	nr	1.00
VisA	macaroni1	89.53	83.46	10.01	1.00
VisA	macaroni2	68.64	52.89	9.94	1.00
VisA	pcb1	88.77	90.21	1.00	3.29
VisA	pcb2	93.15	89.87	nr	11.09
VisA	pcb3	93.21	96.81	6.64	nr
VisA	pcb4	98.49	97.36	7.65	6.32
VisA	pipe_fryum	98.28	93.32	2.20	1.00

A median over categories would hide the shape, and it is also the wrong summary here for a reason worth stating. We evaluated nine definitions rather than one: three ways to place the crossing between two grid points, including interpolation in $\log_2 k$, which is the spacing the budget grid actually uses, crossed with three ways to combine the five support paths.

Under all nine definitions of Table 22 the median over the categories that do cross is 1.00 on both datasets. The crossing rule is not what makes a median unstable; the censoring is. Four categories never reach the full-pool value inside the budget, and folding them in as though they merely exceeded

Table 22: Sensitivity of the shot-equivalent summary to its definition: three ways to place the crossing between two grid points (linear in k, linear in $\log_2 k$, or the next grid point), crossed with three ways to combine the five support paths (mean first, median first, or the median of per-path crossings), so nine definitions. The median over the categories that do cross is 1.00 under every one of them. What moves is not the crossing rule but the censoring: four categories never reach the full-pool value within the budget, and folding them in as though they exceeded sixteen moves the MVTec AD median as far as 4.00. For this reason the text reports the two counts rather than a median.

Quantity over the nine definitions	MVTec AD	VisA
median over categories that cross	1.00	1.00
median with the four censored categories folded in	1.59–4.00	1.00
categories at one image or fewer (of 27)	14–15	
categories not reached by k=16 (of 27)	4	

sixteen moves the MVTec AD median as far as 4.00. We therefore report the two counts, which are 14 to 15 of 27 at one image or fewer and exactly 4 of 27 not reached under every definition, and decline to report a headline median at all. The dagger in Table 20 marks the categories where the metric is close enough to its ceiling that the crossing is weakly identified; on MVTec AD these are the majority, which is why the VisA rows carry more of the evidence.

Sixteen images is a property of the grid, not of the category. Four categories do not reach the full-pool value inside the registered budget, which on its own says only that the grid stops where it stops. To separate the two we extended the frozen support manifest from rank sixteen to rank thirty-two under the same selection key. The extension is strictly nested, so the first sixteen references of every support path are the frozen ones and no already-reported value is replaced; overlapping cells are recomputed only as a seam check. We ran the 7 categories whose conclusion could change, by a criterion fixed before the run: shot-equivalent above eight, or not reached by k=16. The other twenty cross by k=8, where a larger budget carries no information about them.

Reading two protocols on one curve is legitimate only if the overlap is unchanged, so the registration made equality at k=16 a precondition, and the seam checker, fixed before the run, added every full-pool cell: all 42 overlapping cells, which is every k=16 cell and every full-pool cell of these categories, reproduced the frozen run bitwise in the adopted run. The registration voided the protocol if the k=16 seam failed; we depart from that rule and disclose it as a deviation. The precondition did its job once. In the first smoke run two of the five k=16 cells of `toothbrush` missed bitwise equality in the sixth decimal of the patch scores (image AUROC unchanged), because the runner computed distances against the union of all supports up to rank thirty-two, which changed the fp16 blocking; computing each budget against its own union, as the frozen design intended, restored equality. The failed outputs are archived, and no k=32 value was read before the seam held.

Table 23 reports the outcome: 1 category crosses once the budget is doubled, and 3 do not: for those, thirty-two clean images, twice the registered budget, still do not buy what the unlabeled cohort provides. That is a sharper statement than the frozen grid could support, and it holds for `screw`, which no frozen encoder brings within the registered budget (Appendix P). Because the roster was deliberately restricted to the categories that could change, no average over it would mean anything, and none is reported.

Published scorers disagree on the shot-equivalent. Published tables contain the same quantity without naming it, and the values range from below one image to above eight. Damm et al. (2025) report a batched zero-shot arm of their own method alongside its k-shot ladder; on MVTec AD the batched arm falls below the one-shot entry, and on VisA it lands between two and four images. Huang et al. (2026) evaluate one frozen backbone in both regimes and their batched arm lands between one and two images on MVTec AD with their DINOv2 backbone (between two and four with their CLIP backbone). Read against the MuSc+ ladder in its own appendix (their Table 16), which scores patches only against the clean references (its image-level re-scoring still uses the test images), the zero-shot row of Li et al. (2024b) instead implies a crossing above eight images; read against their Table 15, which adds the clean references to the unlabeled cohort, it is matched at one.

Table 23: Clean budget extended from sixteen to thirty-two images, for the seven categories whose conclusion could change. Selection was frozen before the run: shot-equivalent above eight, or not reached by k=16. The remaining twenty cross by k=8 and are unaffected. The k=16, k=32 and full-pool columns give category image AUROC (percent), averaged over the five support paths for the clean arm; the last two columns give the shot-equivalent computed on the grid up to k=16 and up to k=32. Every k=16 cell and every full-pool cell of these categories reproduced the frozen run bitwise, which is what licenses reading the two protocols on one curve; the roster is a deliberate subset, so no average over it is meaningful.

Dataset	Category	k=16	k=32	full pool	within 16	within 32
MVTec AD	`capsule`	97.97	99.03	95.57	13.24	13.24
MVTec AD	`grid`	99.30	99.52	99.67	not reached	not reached
MVTec AD	`screw`	86.05	93.24	95.47	not reached	not reached
MVTec AD	`toothbrush`	97.22	97.61	100.00	not reached	not reached
MVTec AD	`transistor`	98.59	99.30	97.46	8.34	8.34
VisA	`pcb2`	90.36	91.21	89.87	11.09	11.09
VisA	`pcb3`	95.52	97.01	96.81	not reached	29.91

We transcribed these from the published tables rather than reproducing them, and the artifact does not rederive them. The spread is the point: a number that ranges from below one to above eight, depending on which scorer is asked, cannot be treated as a property of the dataset, and a ranking that mixes the two access protocols is comparing arms whose reference budgets differ by an amount nobody has measured.

Why this is a disclosure problem and not a misconduct claim. Some of this literature already discloses reference access precisely. Li et al. (2023) state batch-level prediction and a normal majority in every test batch as numbered assumptions, alongside a meta-training set, and report AUC for test batch sizes from three to one hundred. Damm et al. (2025) code each compared method by setting in a table and decline to compare across settings on that basis. Others place methods with different access in one ranking without a field for it. The inconsistency, not any individual paper, is what an evaluation contract can fix, and the shot-equivalent is the missing field expressed in the units a reader already understands.

Limits. The quantity is post hoc: it was not part of the frozen RQ1 contrast, and it carries the same descriptive status as the full-pool arm itself. It is defined only where the clean curve crosses within $k \leq 16$, and four categories are reported as not reached rather than extrapolated. It is a property of this frozen scorer on these benchmarks, and the cross-method readings above are transcriptions of other authors' published numbers under other scorers, protocols and, in the case of Huang et al. (2026), a reference pool of sixty-three images rather than a whole cohort.

P A SECOND FROZEN ENCODER

Appendix O places the full-pool arm on the clean budget grid under one frozen encoder. The obvious question is whether that placement is a property of the comparison or of the features, and it is the question a reader should ask before any of it is believed. We registered the test before the adopted runs. Protocol `P000018-T1-dinov2-budget-v1` reuses the v4 registry, the five frozen support paths byte for byte, the budget grid, the self-exclusion rule and the scorer, and changes only the frozen encoder, from OpenCLIP ViT-L/14-336 to DINOv2 ViT-L/14 (Oquab et al., 2024), together with what an encoder change cannot avoid. Each backbone keeps the preprocessing its upstream implementation prescribes, so the DINOv2 branch of the audited code normalizes with ImageNet statistics, and a short adapter loads the backbone and, as the upstream code does, prepends the placeholder class token that the aggregation stage removes. Aggregation, scoring and metrics are the shared frozen code. The support manifest can be reused unchanged because its selection key is a hash of the image bytes and is therefore independent of the encoder. Two arms were run, with and without the four register tokens (Darcet et al., 2024); they share one code hash and one environment hash, so the checkpoint is the only difference between them. Each arm is 27 categories and 702 cells. The frozen part of the registration names the checkpoint without registers; an implementation note

Table 24: The registered second-encoder comparison: the count condition, for both DINOv2 checkpoints. Protocol, support sets, budget grid, scorer and batch size are held fixed; only the frozen encoder, with its upstream preprocessing, changes. The adjudication rule, fixed before the adopted runs, has two conditions reported on their own terms: a count moving by at least four categories, and a per-category rank correlation below one half. In neither DINOv2 arm does a count move that far; the correlation condition, reported in the text, is met in both. Counts are over 27 categories with the five support paths averaged; the per-path ranges give the same counts computed on each frozen support path separately, which is the variation the protocol itself carries. For the at-most-one-image count that range spans three to four categories, comparable to the four-category threshold, so the registered test could not have resolved an effect of this size.

	at one image or fewer		not reached by k=16	
Frozen encoder	paths averaged	per-path range	paths averaged	per-path range
OpenCLIP ViT-L/14-336	15/27	12–16	4/27	3–5
DINOv2 ViT-L/14	15/27	14–18	4/27	1–3
DINOv2 ViT-L/14-reg	16/27	16–19	2/27	2–3

appended to the same document, which states that it was written before any result, made the register checkpoint the primary arm and kept the other as an ablation. We therefore adjudicate the registered rule for both arms and report both. The register arm was launched twice before the adopted run, under earlier code hashes: the first launch hit repeated transient network errors while loading the backbone, and the second stalled on a network connection at the same stage; the adopted code pins the hub reference. Both partial launches and a one-category pilot of the checkpoint without registers are archived, and no value from them is used. The adopted register run was interrupted once by a stale job and resumed on the same code hash.

The adjudication rule is in the frozen part of the dated registration document, which the file record places before both adopted runs; the pilot and superseded launches ran earlier the same day, and no hash earlier than adjudication orders them against the rule. It has two conditions, each with its own label: an endpoint count moving by at least four of 27 categories is recorded as a material change of the counts, and a per-category Spearman correlation between the two encoders below 0.5 is recorded as the per-category ordering not being preserved. If neither condition is met the result is negative; if one is met and the other is not, each is reported on its own terms and neither is chosen over the other. Two implementation notes were later appended to the same registration document. Its frozen part, which contains this rule, is unchanged: it reproduces the hash prefix we recorded at adjudication. That is the earliest hash on record, so what dates the rule before the run is the dated document itself, not a hash (Appendix X).

The two conditions disagree, identically in both arms, so we report them separately, which is what we registered to do. In Table 24 the count condition is not met: with registers the at-most-one-image count moves from 15 to 16 and the not-reached count from 4 to 2, and without them neither count moves, against a threshold of four. The correlation condition is met in both. With registers, over the 22 categories finite under both encoders, with 5 excluded as the registration requires us to state, Spearman's r_s is 0.4946; without them, over 22 categories, it is 0.4285. Both lie below the registered threshold of 0.5, and Kendall's τ_b (not registered), which also corrects for ties, is below 0.5 as well, at 0.4524 and 0.3614. Our registration anticipated this split and fixed what to do with it: report each condition on its own terms and do not choose between them. The remainder of this appendix uses the register arm, which the appended note designated primary; the arm without registers gives the same split.

A correction, and why the registered statistic was the wrong instrument. Our first computation of this correlation returned 0.711, which would have made the result negative. That number came from breaking ties by position rather than giving tied observations their average rank, which is the standard definition. The distinction is decisive here and we did not foresee it when registering: the shot-equivalent saturates at the bottom of the budget grid, so 12 of the 22 compared categories carry exactly the same value under both encoders. A rank correlation computed over a vector that is more than half constant is determined by how the ties are broken, and the tie-broken variant ranges from

Table 25: Register ablation, descriptive; equal-category macro image AUROC (percent), clean values averaged over the five support paths. The two DINOv2 arms differ only in whether the published checkpoint carries the four register tokens; they share one code hash and one environment hash. Registers raise the clean arm at k=1 and k=4 on MVTec AD, leave VisA nearly unchanged at those budgets, and lower the full-pool arm on both benchmarks. The text below discusses one possible reading; this protocol identifies no mechanism.

Dataset	Encoder	k=1	k=4	k=16	full pool
MVTec AD	DINOv2 ViT-L/14	91.86	94.46	98.29	95.94
MVTec AD	DINOv2 ViT-L/14-reg	93.45	96.08	98.49	95.70
VisA	DINOv2 ViT-L/14	86.79	90.23	93.07	87.58
VisA	DINOv2 ViT-L/14-reg	86.84	90.20	92.61	86.25

0.4037 to 0.8814 across four orderings of the same data—straddling the threshold. It is a property of the listing order, not of the encoders, and it is not a result. We found this while binding every number in this paper to its provenance, which is the check this paper argues for, and we report the correction rather than the corrected number alone.

What the ties hide is worth more than the coefficient. The two encoders agree exactly on 12 of 22 categories and put 16 on the same side of the one-image line, so whether a single clean image suffices is largely, though not entirely, encoder-invariant. On the 10 categories where at least one encoder needs more than one image, the ordering is not preserved at all ($r_s = 0.2547$, exploratory). The sharper reading is therefore not that the encoder does nothing, but that it changes *whether* one clean image is enough in a minority of categories while it does not preserve *how many* are needed when one is not—which is the regime where published shot-equivalents disagree. We consequently do not claim that the disagreement across published tables is independent of the frozen features; our experiment does not exclude the encoder.

One sensitivity, labeled exploratory because it was not registered: dropping the categories whose one-image clean value or full-pool value reaches 99.5 under either encoder leaves 17 of the 22 categories and raises r_s to 0.5426, above the threshold. We report it because it exists and would otherwise be found by a reader, and we decline to adopt it, because choosing after the fact the analysis that returns the preferred label is exactly what registering the rule was meant to prevent.

Two qualifications belong with the count condition rather than after it. The first is power. The five support paths were predeclared for selection sensitivity, not for statistical power, and computing the same counts on each path separately (post hoc) gives a spread of up to four categories in the at-most-one-image count, which is the size of the effect the registered threshold was set to detect. A test built on five deterministic paths cannot resolve a shift of this size, so the counts failing to move is not evidence that nothing moved. The second is that the two encoders are not interchangeable underneath: the per-path counts differ, and so do many per-category values. Both qualifications point the same way as the correlation condition, which is part of why we decline to read the count condition alone as a null.

The register ablation of Table 25 is descriptive. Registers raise the clean arm at k=1 and k=4 on MVTec AD, leave VisA nearly unchanged at those budgets, and lower the full-pool arm on both benchmarks. The direction is consistent with the high-norm artifact tokens that Darcet et al. (2024) find in low-information background regions, and that registers remove, if a clean reference bank is thin there and a whole-cohort bank is not; this protocol isolates no mechanism and we claim none. The token-norm diagnostic planned in the first implementation note has no recorded run.

Post hoc, Table 26 is the most useful thing the second encoder buys. Where the two arms differ substantially, the conclusion does not depend on the encoder at all: 11 categories need a single clean image under all three encoders, and `screw` is not reached by sixteen under any of them, with the full-pool arm ahead by 9.42, 13.33 and 6.72 percentage points respectively. The categories whose crossing status does depend on the encoder are exactly the near-tied ones, with a mean absolute gap of 0.62 points against 6.10 for the categories that need one image, or are not reached, under all three encoders. A threshold-crossing statistic is undetermined when the two sides are tied; that is a

Table 26: Which per-category conclusions survive the encoder change, descriptive; the three encoders are OpenCLIP and the two DINOv2 checkpoints, and the rows partition the 27 categories. The last column is the mean absolute gap between the full-pool arm and the path-averaged clean arm at k=16, in percentage points, averaged over the three encoders. The categories whose crossing status depends on the encoder are those whose two arms are nearly tied, which is where a threshold-crossing statistic is undetermined by construction rather than unreliable.

Group	categories	mean |gap|
one clean image suffices, all three encoders	11	5.76
reached within $k\leq$16, more than one image under at least one	10	5.25
not reached by k=16, all three encoders	1	9.83
crossing status depends on the encoder	5	0.62

property of the statistic, not a defect in the measurement, and it bounds where the shot-equivalent should be read at all.

That bound also reaches back to Section 4. The `screw` exception, where sixteen clean images do not reach the full-pool arm, survives a complete change of frozen features spanning two pretraining paradigms and two preprocessing conventions. We have no explanation for it, and none of the three encoders we tried removes it.

Q How many unlabeled images the full-pool arm needs

The remaining factor we can control is the size of the reference pool itself. Each category's test cohort, in its frozen order (its queue), is split into disjoint blocks of at most n images and every test image is scored against its own block, so the evaluation set is identical across cardinalities and identical to the full-pool arm; no image is discarded and none is scored twice. A short final block is kept rather than merged, so at n=16 the smallest block holds three images (`bottle` and `metal_nut`), that is, two references per query. Partition rather than sampling follows Li et al. (2024b), who divide the test set into s subsets, and the partition is deliberately not stratified by anomaly type, because a paper auditing reference access should not build its pools from test labels. The realized anomaly ratio of every block is therefore reported with the data, since the composition of a test cohort is known to move transductive scores (Veilleux et al., 2021). Because unstratified blocks let that ratio vary, the registration fixed in advance a trigger for a stratified control: a within-category range of block anomaly ratios above one half at n=16. The largest observed range, on `pipe_fryum`, sits exactly at that threshold, so the control was not triggered and is no part of the registered result; an exploratory stratified rerun of that category, made outside the registration, is archived and not used here. One category has fewer test images than the larger cardinality, so its n=64 arm is a single block covering the whole queue; that arm reproduced the frozen full-pool cell bitwise. We call such an overlap with the frozen study a *seam control*; here it was a consequence of the design rather than a check added afterwards. One gap in the record belongs here: the registered records of the 27 categories (54 cells) carry the hash of the runner that produced them, but that runner version was not archived; the archived runner is the later version that produced the exploratory stratified cell. The frozen base code, the aggregate and every cell record are archived, so the numbers below can be re-derived from the aggregate but not re-executed from the exact runner. Table 27 gives category image AUROC at the three cardinalities and the n=16 and full-pool shot-equivalents with their ratio.

The registered summary, and what it cannot see. We registered the summary as the median over categories of the ratio between the shot-equivalent at the full pool and at n=16 (an arm not reached by k=16 has no shot-equivalent and is marked – in Table 27), with a shot-equivalent that changes monotonically with cardinality and a median ratio of at least two counting as an operational effect, and a median below 1.5 or an inconsistent direction as a negative result. The registration did not say how to treat an undefined ratio; we take the median over the categories where it is defined, and every other treatment gives the same median. That median is 1.000, so by the rule we registered this is a negative result and we report it as one: by the registered summary, cardinality does not

Table 27: Reference-pool cardinality. Each test image is scored against its own disjoint block of at most n images, so the evaluation set is identical across cardinalities and to the full-pool arm. The n=16, n=64 and full columns give category image AUROC (percent). The shot-eq. columns give the shot-equivalent of the n=16 and full-pool arms, and ratio is the second divided by the first; – marks an arm not reached by k=16, for which the ratio is undefined. A dagger marks the categories whose shot-equivalent is already at the floor of one image in the frozen run: their ratio is one by construction, because a quantity censored at one image cannot record a change. The registered summary, the median ratio, here over the categories where it is defined (every other treatment gives the same median), and the exploratory median over the eight undaggered categories with a finite ratio are given in the text.

Dataset	Category	n=16	n=64	full	shot-eq. (16)	shot-eq. (full)	ratio
MVTec AD	bottle	98.81	96.75	99.92	1.00	8.00	8.00
MVTec AD	cable	96.96	95.71	97.32	3.45	3.80	1.10
MVTec AD	capsule	96.45	97.73	95.57	14.25	13.24	0.93
MVTec AD	carpet†	98.39	97.75	97.35	1.00	1.00	1.00
MVTec AD	grid	97.99	97.91	99.67	1.61	–	–
MVTec AD	hazelnut†	92.25	93.79	94.82	1.00	1.00	1.00
MVTec AD	leather†	99.90	99.59	99.42	1.00	1.00	1.00
MVTec AD	metal_nut†	89.88	85.14	81.52	1.00	1.00	1.00
MVTec AD	pill†	95.04	94.30	93.37	1.00	1.00	1.00
MVTec AD	screw	73.74	87.62	95.47	6.21	–	–
MVTec AD	tile†	99.82	99.17	98.99	1.00	1.00	1.00
MVTec AD	toothbrush	94.17	100.00	100.00	3.69	–	–
MVTec AD	transistor	96.79	96.79	97.46	3.98	8.34	2.09
MVTec AD	wood†	92.19	91.49	91.23	1.00	1.00	1.00
MVTec AD	zipper	97.48	97.56	98.08	3.67	6.35	1.73
VisA	candle†	89.35	91.38	90.96	1.00	1.00	1.00
VisA	capsules†	81.25	83.05	82.25	1.00	1.00	1.00
VisA	cashew†	86.90	88.48	88.40	1.00	1.00	1.00
VisA	chewinggum†	95.42	95.00	95.24	1.00	1.00	1.00
VisA	fryum†	94.88	94.16	95.06	1.00	1.00	1.00
VisA	macaroni1†	83.95	82.10	83.46	1.00	1.00	1.00
VisA	macaroni2†	52.36	52.98	52.89	1.00	1.00	1.00
VisA	pcb1	85.08	88.53	90.21	1.00	3.29	3.29
VisA	pcb2	87.66	87.35	89.87	5.36	11.09	2.07
VisA	pcb3	94.12	96.01	96.81	7.90	–	–
VisA	pcb4	91.69	95.85	97.36	1.00	6.32	6.32
VisA	pipe_fryum†	92.60	94.20	93.32	1.00	1.00	1.00

move the shot-equivalent. The registration also named the shot-equivalent at the middle cardinality, a statement on the encoder, block anomaly ratios as a covariate and a prior expectation; none enters the registered rule, so none changes the verdict. The block ratios are archived with the aggregate; the other items are not reported.

The reason it cannot is worth more than the verdict. The shot-equivalent is censored at one image: once a category's one-image clean value exceeds the arm's value, the answer is one image however far the arm moves. Of the 27 categories, 15 sit at that floor in the frozen run, so their ratio is one by construction. Their image AUROC is not still—it moves by a median of 1.58 and by as much as 8.36 percentage points across the three cardinalities while the registered summary records nothing at all. A median taken over a quantity that is censored for more than half its support is uninformative about the half it censors, and we registered one anyway.

Two readings we report and do not mix. Restricted to the 8 categories whose shot-equivalent the frozen run could already resolve, the median ratio is 2.081, which would have met the registered threshold. That stratum is fixed by the frozen run, published before this protocol was designed, and it admits no tuning: the frozen shot-equivalents jump from one image straight to more than three, so every cut in that interval selects the same categories. We nonetheless label this estimate exploratory,

because the decision to stratify followed the pooled null even though the criterion could not have been chosen to suit it, and because eight categories is a small base.

Measured directly (post hoc, on the registered cells), in the form the comparable literature uses, the effect is consistent: image AUROC averaged over all 27 categories of both datasets rises monotonically from 90.93 at n=16 to 92.45 at the full pool, a gain of 1.51 points computed from unrounded values. Li et al. (2024b) divide each test set into up to three subsets, down to 14 to 56 images per subset on MVTec AD and 50 to 67 on VisA (their Table 7), and report that classification performance falls by at most 1.1 points as the subsets shrink. Our n=16 lies at the small end of their MVTec AD range; what agrees is the direction and the order of magnitude. The disagreement is therefore between the direct measurement and our own censored summary, not between us and the literature. The conclusion is narrower than either reading alone: pool size is an operational factor for the metric, and it is invisible to the shot-equivalent wherever the shot-equivalent is already at its floor—which is most of the benchmark.

R WHAT THE FULL-POOL GAP IS MADE OF

The full-pool arm and a clean arm differ in eight recorded fields at once, which is why their gap is reported as a complete-protocol placement and nothing more. Three of those fields can be moved one at a time without changing the scorer: how many references a query sees (budget), whether they include anomalies (purity), and whether the normal references come from the test cohort or from the training set (source). Protocol `P000018-T1-decomp-v1` builds, for every category and inside a single distance computation, arms that share the full pool's size or its anomaly set and differ from it, or from each other, in one of these respects (Table 28). Every arm scores the same queries under the v4 encoder, preprocessing, feature layers, radii and q=1 functional, with self-exclusion; the clean training references follow the v4 selection key, extended past rank sixteen under the same key, and the first sixteen are checked against the frozen manifest before any distance is computed. The two arms that sample the cohort by label (A2, A4) and the arm that mixes the cohort's anomalies into a clean bank (A6) use test labels; they are oracle diagnostics that isolate a field, not protocols anyone could deploy or rank.

The registration, as amended before any value existed, fixed two interval rules and one count rule. H-budget is read on A1 − A0: does a clean bank of the full pool's own size beat the full pool? H-source is read on A0 − A6, which keeps the anomaly set and swaps the normal references from test to training images, with A2 − A3, two clean banks of equal size from the two sources, reported alongside; both also change how the normal references were selected, so we call them source-and-selection contrasts. A contrast is *present* if the 95% lower bound of its category-stratified bootstrap interval exceeds $+0.5$ and at least two thirds of the categories are positive, *equivalent to zero* if its 90% interval lies inside ± 0.5, and *undetermined* otherwise. H-purity is a count: among the fifteen categories where one clean image already matches the full pool, it is supported if at least ten show both A2 − A4 > 0.5 and A6 − A1 < -0.5. The other rows of Table 28 are descriptive. The categories `grid` and `toothbrush` sit at the ceiling and were declared uninformative in advance; they stay in the macro.

The run is tied to the frozen study by a seam control: every category recomputes its k=1 and k=16 cells on path p0 and its full-pool cell, and these must lie within 0.5 percentage points of the frozen values; 37 of the 81 seam cells reproduce them bitwise and the largest difference is 0.08 points. Three events are disclosed. Before any analysis cell completed, the second-scorer algebra was moved from double precision on the CPU to double precision on the GPU, with the same formulas, and thread pools were set to the container's CPU quota, because the CPU path ran effectively single-threaded; an operational overlap of two jobs on one GPU exhausted memory for one category twice, and that category received a third attempt alone, beyond the registered two; and the frozen analysis script was run once on the sixteen categories complete at that point as a pipeline check, with no change to any rule, arm, category list or stopping decision afterwards.

Table 28: Decomposition of the full-pool gap (image AUROC, percentage points; macro over categories with a category-stratified 95% bootstrap interval). Every arm scores the same queries against a self-excluded bank built from one distance computation per category: A0 the whole cohort; A1 the first $N-1$ clean training images of path p0; A2 the other test normals (oracle labels); A3 the first b_0 clean training images, b_0 the number of test normals minus one; A4 a hash-drawn cohort subsample of size b_0 at the cohort's anomaly ratio (mean of five draws); A6 the cohort's anomalies plus clean training normals, size $N-1$; Y_{16} the k=16 clean arm recomputed in the same run, the seam control of the frozen cell. Arms A2, A4 and A6 use test labels and are oracle diagnostics, not rankable protocols. The two source contrasts also change how the normal references were selected. Rules fixed before any value: an effect is present if the 95% lower bound exceeds $+0.5$ and at least two thirds of the categories are positive, equivalent to zero if the 90% interval lies inside ± 0.5, and undetermined otherwise; rows without a registered rule are descriptive. The categories `grid` and `toothbrush` were declared uninformative in advance (ceiling) and are kept in the macro. A5 was deferred by the registration and is added in Table 30.

Contrast	Arms	MVTec AD	VisA	Rule and verdict
Budget at the full pool's size	A1 − A0	+3.49 [+1.47, +6.16]	+7.07 [+4.16, +11.01]	H-budget: MVTec AD present; VisA present
Source of the normal references	A0 − A6	+1.50 [+0.52, +2.92]	+0.60 [+0.20, +0.98]	H-source: MVTec AD present; VisA undetermined
Source, both banks clean	A2 − A3	+0.69 [+0.03, +1.68]	+0.67 [+0.28, +1.23]	H-source (alongside): MVTec AD undetermined; VisA undetermined
Purity, test-cohort references	A2 − A4	+3.58 [+2.09, +5.31]	+7.45 [+4.44, +11.48]	descriptive
Purity, training references	A6 − A1	−4.99 [−7.77, −2.78]	−7.67 [−11.88, −4.61]	descriptive
Clean budget beyond sixteen	A1 − Y_{16}	+1.23 [+0.30, +2.82]	+2.35 [+0.73, +4.59]	descriptive
Total, full pool against k=16	A0 − Y_{16}	−2.26 [−5.27, +0.30]	−4.72 [−7.11, −2.50]	descriptive

The registered verdicts are read off Table 28 and reported one by one, without a combined verdict. H-budget is present on both datasets: at the full pool's own size, clean training references beat the whole cohort by $+3.49$ [$+1.47$, $+6.16$] points on MVTec AD and $+7.07$ [$+4.16$, $+11.01$] on VisA, positive in 14 of 15 and 12 of 12 categories (the k=32 cells of `screw` and `pcb3`, named as background by the amendment, are in Table 23 as five-path averages; the per-path values are archived). What the cohort lacks is therefore not size. H-purity is supported, in 14 of 15 categories; because those categories were fixed by the earlier, post hoc shot-equivalent, we also report, descriptively, that the same condition holds in 25 of all 27. Both purity contrasts point the same way on both datasets: anomalies in the bank change image AUROC by -4.99 and -7.67 points when they replace clean training references at the same bank size (A6 − A1, descriptive), and removing them from a test-cohort subsample of equal size is worth $+3.58$ and $+7.45$. H-source is present on MVTec AD ($+1.50$ [$+0.52$, $+2.92$]) and undetermined on VisA ($+0.60$ [$+0.20$, $+0.98$]), whose interval excludes zero but not the 0.5 threshold; the two clean banks of equal size (A2 − A3) are undetermined on both. The source contrast also changes the selection rule, so a test-cohort advantage on MVTec AD may reflect how close its test normals sit to the queries rather than where they come from.

The rows telescope: the full-pool arm's distance from the frozen k=16 arm, -2.26 on MVTec AD and -4.72 on VisA, is on path p0 the sum of a clean-budget term ($+1.23$, $+2.35$), a purity term (-4.99, -7.67) and a source term ($+1.50$, $+0.60$). Read descriptively, since these terms carry no registered rule, the purity term is the largest on both datasets and by itself exceeds the whole gap, which the clean-budget and source terms partly offset. The registration also asked for the reverse order but fixed no arms for it; the frozen script lists the same three terms in reverse, which changes no value, so one set is reported and order dependence is not tested. Where the fields interact, the share attributed to each depends on which arms define the terms, and the source $\times$ purity design

below, at a common bank size, shows, post hoc, the interaction to be small. The complete-protocol gap of Section 4.2 is therefore chiefly a statement about anomalies in the reference set, measured under a reduction that takes the nearest reference; the follow-ups below repeat it on all five support paths and at the pixel level, and Appendix S asks whether the budget effect survives a scorer that does not take a minimum. Whether the decomposition reads the same under that scorer is not tested here.

The decomposition at the pixel level. A follow-up registered after these results, before any of its own values, scores the archived arm maps of Table 28 with the unmodified v4 pixel metrics and applies the same analysis script (Table 29). The registration names pixel AUROC as the primary pixel metric and AUPRO as secondary and descriptive. Under pixel AUROC the budget row is present on MVTec AD ($+4.27$ points) and undetermined on VisA ($+0.74$), where pixel AUROC sits near its ceiling (an interpretation, not a registered test), and the registered purity count is 4 of 15, not supported. The source rows are equivalent to zero on both datasets. Descriptively, AUPRO reads like image AUROC: the clean bank leads by $+3.78$ and $+2.23$ points, the purity count reaches 11 of 15, and only its clean-bank source contrast on MVTec AD ($+0.30$) is not equivalent to zero. The pixel level therefore supports the budget effect on MVTec AD and the null source effect, but not, under its registered metric, the purity count.

Table 29: The registered decomposition at the pixel level (follow-up registered after the decomposition results of Table 28 were known; support path p0). The archived arm maps of Table 28 are scored with the unmodified v4 pixel metrics; entries are the macro contrast in percentage points with its 95% category-stratified bootstrap interval, and, for the rows with a rule, the verdict under the rule of Table 28 applied unchanged (undet.: undetermined; equiv. 0: equivalent to zero). The registration names pixel AUROC as the primary pixel metric; AUPRO is secondary and descriptive, and its verdicts are read against the same rule for comparison only. Every category's k=1 and k=16 pixel AUROC reproduces the v4 cell within 0.0087 points.

Metric	Arms	MVTec AD	VisA
pixel AUROC	A1 − A0	+4.27 [+0.86, +8.80] present	+0.74 [+0.25, +1.42] undet.
	A0 − A6	+0.14 [+0.02, +0.32] equiv. 0	+0.00 [−0.02, +0.03] equiv. 0
	A2 − A3	+0.11 [−0.02, +0.30] equiv. 0	+0.01 [−0.04, +0.05] equiv. 0
	A2 − A4	+3.28 [+0.64, +6.82]	+0.56 [+0.23, +0.95]
	A6 − A1	−4.40 [−8.87, −0.90]	−0.75 [−1.41, −0.25]
AUPRO	A1 − A0	+3.78 [+1.52, +6.55] present	+2.23 [+1.01, +3.60] present
	A0 − A6	+0.21 [+0.03, +0.44] equiv. 0	−0.07 [−0.22, +0.07] equiv. 0
	A2 − A3	+0.30 [+0.02, +0.71] undet.	−0.05 [−0.23, +0.13] equiv. 0
	A2 − A4	+3.45 [+1.39, +5.88]	+2.16 [+1.00, +3.51]
	A6 − A1	−3.99 [−6.85, −1.58]	−2.17 [−3.53, −0.97]

All five support paths, and the source × purity design. Another follow-up, registered after the decomposition results were known and before any of its own values, repeats the decomposition on all five support paths and adds the arm the first registration deferred, A5: training normals plus the cohort anomalies of A4, at the same size b_0 (Table 30). Its p0 column reproduces Table 28 bitwise in 189 of 189 overlapping cells. The budget row is path-robust present on both datasets, $+3.43$ and $+7.12$ points on the path mean, and the registered purity count holds in 14 of 15 categories. Replacing training references with the cohort's anomalies at the same size b_0 changes image AUROC by -3.64 and -7.01 points, present (negative, as predicted) on both datasets (the registration text states this rule with the lower bound; the frozen script, applied as frozen, uses the upper bound, and the verdict is the same under either), so the purity effect does not depend on where the normal references come from. The source rows are undetermined on the path mean: on MVTec AD the registered p0 verdict of Table 28 stands, but across paths the interval's lower bound falls just short of the rule. The four arms A2, A4, A3 and A5 form a 2×2 design at one bank size, so, post hoc and without registration, we also report its main effects on the path mean with category-stratified intervals: purity $+3.61$ [$+2.19$, $+5.25$] and $+7.23$ [$+4.43$, $+10.96$], source $+0.41$ [$+0.18$, $+0.73$] and $+0.50$ [$+0.20$, $+0.84$], interaction -0.06 [-0.40, $+0.34$] and $+0.44$ [$+0.00$, $+1.05$]. With two factors the Shapley share of each equals its main effect (Shorrocks, 2013), and with an interaction

small relative to the purity effect the ranking of the terms does not depend on which arms define them.

Table 30: Decomposition contrasts on all five support paths (follow-up registered after the decomposition of Table 28 was known; image AUROC, percentage points). Columns p0–p4 are macros over categories per path; the last two columns average the paths within each category first and give the category-stratified 95% bootstrap interval, and the follow-up rule: path-robust if at least four of five path macros exceed $+0.5$, the lower bound exceeds $+0.5$ and two thirds of the categories are positive; A5 $-$ A3 has its own one-sided rule, present if the 95% upper bound is below -0.5 (the predicted sign is negative) and equivalent to zero if the 90% interval lies inside ± 0.5, with no path or category condition; robust: path-robust present; undet.: undetermined; descr.: descriptive, no rule. The p0 column repeats the run of Table 28, which it reproduces bitwise.

Dataset	Arms	p0	p1	p2	p3	p4	mean [95% CI]	rule
MVTec AD	A1 − A0	+3.49	+3.41	+3.44	+3.45	+3.37	+3.43 [+1.40, +6.06]	robust
MVTec AD	A0 − A6	+1.50	+1.35	+1.34	+1.45	+1.51	+1.43 [+0.48, +2.78]	undet.
MVTec AD	A2 − A3	+0.69	+0.39	+0.15	+0.17	+0.52	+0.38 [+0.04, +0.82]	undet.
MVTec AD	A2 − A4	+3.58	+3.58	+3.58	+3.58	+3.58	+3.58 [+2.08, +5.30]	descr.
MVTec AD	A6 − A1	−4.99	−4.77	−4.78	−4.90	−4.89	−4.86 [−7.52, −2.77]	descr.
MVTec AD	A5 − A3	−3.25	−3.55	−4.21	−3.91	−3.28	−3.64 [−5.18, −2.24]	present (< 0)
VisA	A1 − A0	+7.07	+7.22	+7.14	+7.21	+6.98	+7.12 [+4.29, +10.93]	robust
VisA	A0 − A6	+0.60	+0.25	+0.60	+0.87	+0.76	+0.62 [+0.17, +1.10]	undet.
VisA	A2 − A3	+0.67	+0.76	+0.65	+0.62	+0.91	+0.72 [+0.24, +1.35]	undet.
VisA	A2 − A4	+7.45	+7.45	+7.45	+7.45	+7.45	+7.45 [+4.45, +11.45]	descr.
VisA	A6 − A1	−7.67	−7.47	−7.74	−8.08	−7.74	−7.74 [−11.59, −4.77]	descr.
VisA	A5 − A3	−7.01	−7.12	−6.96	−7.16	−6.78	−7.01 [−10.40, −4.36]	present (< 0)

The decomposition under the second encoder. A further follow-up, registered after the decomposition was known and before any of its own values, reruns it on path p0 with the DINOv2 encoder of Appendix P, every other field unchanged, and applies the same analysis script (Table 31); its k=1, k=16 and full-pool cells reproduce the frozen DINOv2 cells within 0.06 points. The budget row is present on both datasets, $+3.92$ and $+9.22$ points, the registered purity count holds in 15 of 15 categories, and the source rows are undetermined, as they are across paths under OpenCLIP. What the cohort lacks is therefore not a property of one frozen encoder either.

Table 31: The decomposition of Table 28 under the second frozen encoder (follow-up registered after that decomposition was known; support path p0; image AUROC, percentage points, macro with 95% category-stratified bootstrap interval). Each cell gives the contrast and, where a rule exists, its verdict under the same rule and the same, unchanged analysis script. OpenCLIP repeats Table 28; DINOv2 is ViT-L/14 with registers.

Dataset	Arms	OpenCLIP ViT-L/14-336	DINOv2 ViT-L/14-reg
MVTec AD	A1 − A0	+3.49 [+1.47, +6.16], present	+3.92 [+2.09, +6.07], present
MVTec AD	A0 − A6	+1.50 [+0.52, +2.92], present	+1.02 [+0.03, +2.20], undetermined
MVTec AD	A2 − A3	+0.69 [+0.03, +1.68], undetermined	+0.63 [+0.03, +1.73], undetermined
MVTec AD	A2 − A4	+3.58 [+2.09, +5.31]	+3.66 [+2.07, +5.68]
MVTec AD	A6 − A1	−4.99 [−7.77, −2.78]	−4.95 [−7.32, −2.89]
VisA	A1 − A0	+7.07 [+4.16, +11.01], present	+9.22 [+6.08, +12.98], present
VisA	A0 − A6	+0.60 [+0.20, +0.98], undetermined	+0.25 [−0.38, +0.77], undetermined
VisA	A2 − A3	+0.67 [+0.28, +1.23], undetermined	+0.70 [+0.25, +1.30], undetermined
VisA	A2 − A4	+7.45 [+4.44, +11.48]	+9.05 [+5.85, +12.97]
VisA	A6 − A1	−7.67 [−11.88, −4.61]	−9.47 [−13.47, −6.36]

The decomposition under MuSc's own reduction. The minimum over a reference set is the reduction most exposed to anomalous references, so a follow-up, registered with the pixel-level one, reruns the decomposition on path p0 with MuSc's interval average in its place, the mean of each patch's smallest 30% per-reference distances, and applies the same analysis script (Table 32). The

same pass reproduces every minimum arm of Table 28 bitwise (162 of 162 cells), and its full-pool arm reproduces the interval-average reduction of Appendix N within 0.08 points. The reduction changes the answer where the answer is about anomalies. Purity still costs: removing the anomalies from a cohort subsample gains $+2.35$ and $+4.19$ points, and the registered purity count holds in 12 of 15 categories. But the cost is smaller than under the minimum, and at the full pool's own size the clean bank's margin over the cohort is $+1.00$ points on MVTec AD, undetermined, and $+4.01$ on VisA, present. A reduction that averages the closest 30% of references, so that a few similar ones cannot decide a score (Li et al., 2024b), thus recovers part of what the cohort's anomalies cost the minimum, which is one more way in which the worth of the cohort depends on the scorer that uses it.

Table 32: The decomposition of Table 28 under MuSc's own reduction (follow-up registered after that decomposition was known; support path p0). Each arm's reference set is reduced by the interval average, the mean of its smallest 30% per-reference distances, instead of the minimum; everything else is unchanged. Entries are the macro image-AUROC contrast in percentage points with its 95% category-stratified interval and, where a rule exists, the verdict under the same rule and unchanged analysis script (undet.: undetermined; equiv. 0: equivalent to zero). The same pass reproduces every minimum arm of Table 28 bitwise, and its interval-average full-pool arm reproduces the full-pool reduction of Appendix N within 0.08 points.

Dataset	Arms	minimum (Table 28)	interval average
MVTec AD	A1 − A0	+3.49 [+1.47, +6.16] present	+1.00 [+0.21, +2.11] undet.
	A0 − A6	+1.50 [+0.52, +2.92] present	+0.56 [+0.11, +1.17] undet.
	A2 − A3	+0.69 [+0.03, +1.68] undet.	+0.99 [+0.13, +2.48] undet.
	A2 − A4	+3.58 [+2.09, +5.31]	+2.35 [+0.79, +4.86]
	A6 − A1	−4.99 [−7.77, −2.78]	−1.57 [−3.14, −0.54]
VisA	A1 − A0	+7.07 [+4.16, +11.01] present	+4.01 [+2.31, +5.95] present
	A0 − A6	+0.60 [+0.20, +0.98] undet.	−0.26 [−0.72, +0.14] undet.
	A2 − A3	+0.67 [+0.28, +1.23] undet.	−0.06 [−0.44, +0.32] equiv. 0
	A2 − A4	+7.45 [+4.44, +11.48]	+4.19 [+2.50, +6.20]
	A6 − A1	−7.67 [−11.88, −4.61]	−3.75 [−5.79, −2.09]

Both reductions on a benchmark without a ceiling. On MVTec AD most categories sit at or near the ceiling in both arms, so the room for any contrast is small; the interval average also narrows the gap from both sides, raising the cohort arm and lowering the clean one, on both datasets. A further follow-up (`realiad-ia-f9`), registered after the results above were known and before any of its own values, therefore asks the budget and purity questions on the 30 Real-IAD categories of Appendix T, none of which reaches the ceiling. Real-IAD's clean pools hold only 158 to 263 images, so the clean arm C is the whole pool and always smaller than the self-excluded cohort A0 of 299; CX adds every anomalous cohort image except the query and uses test labels. This is not a replication of the equal-size contrast above, and applying the interval average to a clean bank handicaps it, since MuSc itself takes the minimum over clean references (Li et al., 2024b). Under the registered rules (Table 33) the smaller clean pool beats the cohort by $+8.66$ points under the minimum and $+3.64$ under the interval average, both present, and adding the cohort's anomalies changes it by -6.93 and -2.19, both present. Two descriptive analyses, added by an amendment frozen before any contrast value was read, give the reading closest to MuSc's own practice, a clean pool under the minimum against the cohort under the interval average, at $+6.43$ $[+3.89, +9.23]$ points, and, within Real-IAD, a Spearman correlation of 0.78 $[0.56, 0.90]$ between a category's interval-average margin and its headroom; the registered descriptive k=16 contrasts and per-category values are archived rather than reported. As the registration fixed in advance for this case, the Real-IAD result is compatible with the undetermined MVTec AD margin reflecting its ceiling and is not read causally; post hoc, the margin there is also small in several categories that have headroom, and the reduction itself narrows it. The minimum arms reproduce the Real-IAD cells of Appendix T bitwise in 169 of 180 cells and within 0.01 points in all.

Table 33: Budget and purity on Real-IAD under both reductions (follow-up registered after the decomposition and its interval-average version were known; 30 categories, none at the ceiling). C is the whole clean training pool, 158 to 263 images and always smaller than the self-excluded full pool A0 of 299; CX adds every anomalous cohort image except the query, so it uses test labels. Entries are the macro image-AUROC contrast in percentage points with its 95% category-stratified interval and verdict (official labels), and the same contrast without the normal-labeled views of defective products (the images of Real-IAD's NG folder labeled normal). The minimum arms reproduce the Real-IAD cells of Appendix T within 0.01 points.

Reduction	Arms	official labels	without NG normal views
Interval average	C − A0	+3.64 [+1.73, +5.76] present	+5.17
	CX − C	−2.19 [−3.67, −0.75] present	−3.11
Minimum	C − A0	+8.66 [+6.71, +10.74] present	+10.20
	CX − C	−6.93 [−8.76, −5.20] present	−7.50

S A SECOND SCORER FAMILY

Every number above comes from one nearest-neighbor scorer. To ask whether the budget effect is a property of that mechanism, the same protocol scores the same features with a scorer from a different family: the residual of each query patch from the principal subspace of the reference patches, adapted from SubspaceAD (Lendering et al., 2026), with the explained-variance threshold τ=0.99 as the registered variant and τ=0.95 and the squared residual as sensitivities, and a Ledoit–Wolf (Ledoit & Wolf, 2004) Mahalanobis distance as a secondary family. Each is fitted per feature layer and radius from the k clean images of each path, or leave-one-out from the whole cohort for the full-pool arm, and aggregated and upsampled exactly as the nearest-neighbor score. The registered rule, as amended before any value, reads the direction of k=16 minus k=1 on the ten dataset–path pairs and records the conclusion as not specific to the nearest-neighbor scorer if at least eight are positive; endpoint counts are descriptive.

Table 34: Second scorer family on the same features. Entries are the macro k=16 minus k=1 image AUROC difference (percentage points) of each scorer, per dataset and support path; the full-pool arm of every scorer is fitted leave-one-out. The registered rule is on the first column: the conclusion is recorded as not specific to the nearest-neighbor scorer if at least eight of the ten differences are positive, as they are for the 1-NN scorer. The other columns are two sensitivities (PCA .95, PCA sq.) and a secondary family (Maha. LW). PCA (principal component analysis) .99 and PCA .95: residual from the principal subspace that keeps 99% or 95% of the variance; PCA sq.: the squared residual at 99%; Maha. LW: Mahalanobis distance under Ledoit–Wolf shrinkage.

Dataset	Path	PCA .99	PCA .95	PCA sq.	Maha. LW
MVTec AD	p0	+6.00	+5.44	+5.94	+6.16
MVTec AD	p1	+5.72	+5.07	+5.77	+5.77
MVTec AD	p2	+5.88	+5.59	+5.59	+5.92
MVTec AD	p3	+5.68	+5.33	+5.68	+6.07
MVTec AD	p4	+7.37	+6.75	+7.35	+7.98
VisA	p0	+7.63	+7.14	+7.20	+8.42
VisA	p1	+5.35	+4.83	+4.80	+6.25
VisA	p2	+6.62	+6.64	+5.89	+7.77
VisA	p3	+6.43	+6.41	+6.24	+6.74
VisA	p4	+5.55	+4.90	+4.77	+6.39
categories at one image or fewer (of 27)		14	17	17	15
categories not reached by k=16 (of 27)		1	1	1	1

The registered rule is met: k=16 beats k=1 on 10 of 10 dataset–path pairs under the subspace scorer, and in 30 of the 30 pairs of the three variants (Table 34). The clean-budget effect is therefore not specific to the nearest-neighbor mechanism. The endpoint counts move: under the subspace scorer one clean image reaches the full pool in 14 of 27 categories and sixteen fail to reach it in 1, against

15 and 4 for the nearest-neighbor scorer (post hoc shot-equivalent). One reading, consistent with the decomposition but not tested here, is that a subspace fitted on the whole cohort is less exposed to a single nearby anomaly than a minimum over it, so fewer categories stay out of reach of sixteen clean images. As with the second encoder, whether one image suffices is largely stable, and how many are needed beyond that depends on the scorer.

T A BENCHMARK THAT IS NOT SATURATED

MVTec AD and VisA leave little headroom: several categories sit at the ceiling under one clean image. Real-IAD (Wang et al., 2024) does not: under this scorer none of its categories reaches 99.5 at any clean budget (0 of 30). We use its single-view split for camera C1 and all 30 categories, with the queue of each category capped at 300 test images taken in a label-blind hash order, a clean pool drawn only from the training images of the OK folder, and the five support paths built with the v4 selection key. Its registered endpoint is the v4 rule: the k=16 minus k=1 contrast must be positive on all five paths. A sensitivity drops the normal-labeled views of defective products, which Real-IAD keeps in its test lists.

Table 35: Real-IAD single view (camera C1, 30 categories). Test queues are the first 300 images of each category in a label-blind hash order, with realized anomaly ratios from 12.3% to 67.3%; the clean pool holds only training images of the OK folder. Entries are macro image AUROC (percent) of the 1-NN scorer at each clean budget, and the k=16 minus k=1 difference (percentage points), with and without the normal-labeled views of defective products. The self-excluded full-pool arm has macro 84.20.

Path	k=1	k=2	k=4	k=8	k=16	Δ	Δ, sens.
p0	84.22	85.80	87.73	89.63	90.54	+6.32	+6.50
p1	83.86	86.34	88.40	89.30	90.56	+6.70	+7.01
p2	84.74	86.78	87.93	89.10	90.72	+5.98	+6.31
p3	84.75	86.38	87.81	89.03	90.60	+5.86	+6.22
p4	84.56	86.02	87.45	89.32	90.37	+5.81	+6.26

The registered endpoint holds: the contrast is positive on all five paths, by +6.13 points on average (Table 35), and it holds as well without the normal-labeled views of defective products (positive on all five paths, +6.46). The clean curve rises at every budget on every path. The self-excluded full-pool arm is reached by one clean image in 13 categories and not reached by sixteen in 3. The budget effect is therefore not a product of benchmarks that saturate: on a larger benchmark whose categories stay below the ceiling, it is of the same size as on MVTec AD and VisA. As there, the realized anomaly ratio varies by category, here from 12.3% to 67.3%, and we report rather than control it. The follow-up registered with the five-path decomposition, after the decomposition results were known and before any of its own values, repeats this endpoint under the subspace scorer of Appendix S, on the same queues, pools and paths: it holds, by +5.56 points on average (+5.93 without the normal-labeled views), with the recomputed 1-NN arms reproducing the values above (Table 36).

Table 36: Second scorer family on Real-IAD (follow-up registered after the decomposition results were known, before any of its own values): macro k=16 minus k=1 image AUROC (percentage points) per support path, on the queues, clean pools and support paths of Table 35. The follow-up rule is on the first column: it holds if all five paths are positive. The 1-NN arms were recomputed in the same pass and reproduce the values of Table 35 within the seam tolerance.

Path	PCA .99	PCA .95	PCA sq.	Maha. LW
p0	+5.88	+5.57	+5.70	+6.78
p1	+6.22	+5.76	+6.12	+6.68
p2	+5.05	+4.61	+4.80	+5.87
p3	+5.39	+4.86	+5.12	+6.17
p4	+5.24	+4.91	+5.08	+5.87

U CATEGORY-LEVEL INFERENCE AND THE SIZE OF THE BUDGET EFFECT

The registered endpoint is a sign over five deterministic paths per dataset. The paths are nearly disjoint draws from one training pool, but all five are scored on the same test cohort, so they measure sensitivity to the support draw, not variation across categories or test sets. The unit that does vary independently is the category, so we add, post hoc and without registration, the tests that treat it as the unit (Table 37). The path-mean contrast is positive in 12 of 15 MVTec AD categories and all 12 VisA categories; the only negative category, `carpet`, moves by less than a tenth of a point, and the two zeros are categories at ceiling. The bootstrap intervals exclude zero on both datasets. The shot-equivalent counts carry wide exact intervals: at most one image in 15 of 27 categories is compatible with 35.3% to 74.5%, and not reached by sixteen in 4 of 27 with 4.2% to 33.7%. All of this is conditional on these 27 categories and says nothing about categories not in the benchmarks.

Table 37: Category-level inference for the registered $k{=}16$ minus $k{=}1$ contrast (post hoc, not registered). The unit is the category; Δ_c is the path-mean image AUROC difference in percentage points. Two-sided exact sign test with zeros dropped, Wilcoxon signed-rank test with Pratt's zero handling, and a 95% percentile bootstrap of the dataset macro that resamples categories (20,000 draws, fixed seed). Inference is conditional on these categories.

Dataset	categories	+/−/0	macro Δ [95% CI]	min Δ_c	sign test p	Wilcoxon p
MVTec AD	15	12/1/2	+5.60 [+1.63, +10.78]	−0.06	0.003	0.002
VisA	12	12/0/0	+6.31 [+3.54, +9.72]	+0.64	$4.9{\times}10^{-4}$	$4.9{\times}10^{-4}$

A budget effect of a few points can look either trivial or decisive, depending on what it is compared with. The comparison that matters for a ranking table is the gap that separates methods in one. Table 38 sets the frozen scorer's $k{=}1 \to 2$, $1 \to 4$ and $1 \to 16$ effects against the gaps printed in the 66 comparisons of Appendix G. Of the 46 positive gains that a paper's own method achieves over the best prior method family in the same comparison, excluding the two papers whose proposal scores against the test cohort, 36 (78%) are smaller than the effect of moving this scorer from one clean image to sixteen. Measured with the archived AnomalyDINO ladder instead, whose one-to-sixteen effect is $+1.74$ points on MVTec AD and $+8.93$ on VisA, 33 of the 46 are smaller: fewer on MVTec AD and more on VisA, so the count belongs to the scorer as well as to the budget. The comparison is descriptive and crosses scorers, protocols and papers; what it shows is that a difference in reference budget that a table does not declare is of the same size as the differences such tables are used to establish.

Table 38: The clean-budget effect against method-to-method gaps printed in published tables (post hoc). We transcribed 374 image-AUROC values from 66 printed comparisons in 12 papers with page and table locators; the 75 values of the comparisons with a robust reversal in Table 12 were re-read against the source PDFs and all matched. Adjacent gap: difference between consecutive method families after sorting one printed comparison, variants of one method collapsed to their best row. Proposal gain: a paper's own method minus the best prior family in the same comparison, positive gains only, excluding the two papers whose proposal scores against the test cohort; GraphCore's own Aug-PatchCore ablation rows count as variants of its proposal, not as a prior family. Quartiles use the exclusive $(n{+}1)$ rule. The last three columns count gaps smaller than the path-mean $k{=}1 \to 2$, $1 \to 4$ and $1 \to 16$ effect of the frozen scorer (percentage points: 1.68, 2.78, 5.60 on MVTec AD; 2.44, 4.16, 6.31 on VisA).

Dataset	Gap	n	median [IQR]	$< 1 \to 2$	$< 1 \to 4$	$< 1 \to 16$
MVTec AD	adjacent gap	167	2.70 [1.20, 4.60]	56	86	135
MVTec AD	proposal gain	26	1.55 [1.13, 5.95]	14	14	19
VisA	adjacent gap	118	3.00 [1.10, 5.30]	52	84	93
VisA	proposal gain	20	3.05 [1.90, 5.83]	8	14	17

V AUDITING THE CONTRACT AGAINST ITSELF

A contract of this kind invites an objection: that pinning code, data, and configuration by hash, already standard practice, catches anything this contract would catch. We tested that on ourselves, and it does not. Table 39 lists what the self-audit found, in two kinds; none of it changes a conclusion of the main text, and it is reported here as a limit of the record rather than as a result.

Table 39: What the self-audit of our own study found. The first row, found by changing one field while every recorded field stayed fixed, is a numerical context that no recorded identity names; the remaining three, found by the provenance audit, are summary statistics, the first two registered in advance, that cannot be read on the data they summarize. Every number here is bound to an accepted artifact and re-derived by its own audit.

Kind	Instance	What the self-audit showed
Identity does not bind the numerical context	inference batch size	changing it from two to four alters between 99.86% and 99.99% of patch scores per category and image AUROC by up to 0.2778 percentage points, while the recorded identity is unchanged in 27 of 27 categories
Statistic unreadable on its own data	second-encoder rank correlation	12 of 22 categories take the same value under both encoders, so a tie-broken coefficient ranges from 0.4037 to 0.8814 with the listing order alone
	reference-pool cardinality ratio	15 of 27 categories sit at the grid floor, fixing the registered median at 1.000 while the underlying AUROC moves by up to 8.36 points
	shot-equivalent median (post hoc)	censored at both ends of the grid; its appendix had already declined to report it

The batch-size row is the one that answers the objection directly. Every hash a reader could check agrees, in every category, while almost every patch score underneath has changed. Whether that matters depends on what the number is used for: measured against each category's own gap between its full-pool and one-image values, the movement is a small fraction at the median and reaches a tenth or more in 3 of 27 categories, all three with a gap below one point, so the share is large there because the gap is small; in absolute terms image AUROC never moves by more than 0.2778 points. One instance is enough to settle the point, because a reproducibility record that cannot see this difference cannot see any difference of the same kind.

The other three rows are a different failure, found by the audit that binds every number in this paper to its provenance. A rank correlation over a vector that is more than half tied, or a median of a quantity censored at its grid floor, is a property of the statistic rather than of the data. Every registered verdict is reported as registered; the reading failure is reported beside it, in Appendices P and Q.

W REPRODUCIBILITY CHECKS

The self-excluded full-pool arm was later recomputed in full on different hardware, under a different accelerator driver, CUDA release and deep learning framework build, after the machine that produced the original cells had been decommissioned. All 27 category AUROC values and all $5{,}321{,}303$ underlying patch scores were bit-identical to the committed ones; Appendix N reports that rerun and the reduction sensitivity it was designed to measure. An independent raw-to-metric recomputation over all 702 registered cells reproduced every committed per-cell image metric with zero absolute difference, and re-deriving the equal-category macro table from those cells reproduced every published aggregate with zero absolute difference. The supporting raw-score archive is identified by SHA-256 in the artifact rather than redistributed, so the artifact alone re-derives the reported tables from the accepted aggregates, while recomputing metrics from raw scores additionally requires that archive.

Auditing the contract against its own assumptions exposed one degree of freedom it does not bind. The encoder runs under `float16` autocast, where the reduction order of a kernel is known to depend on the batch size it is called with (He & Thinking Machines Lab, 2025; Yuan et al., 2025), and recent work on inference verification treats the accumulated rounding of a given software and hardware setup, including unreported batch elements, as something an auditor must be able to recompute (Cankaya, 2026); the per-cell identity binds code, configuration, data, encoder, query set, support set and environment, but not the inference batch size. Re-running, post hoc, all 27 categories on one machine with the batch size changed from two to four left the recorded cell identity unchanged in 27 of 27 categories while altering between 99.86% and 99.99% of patch scores, by a per-category median relative amount between 2.0×10^{-4} and 4.6×10^{-4}. Over all 702 registered cells, each scored with the $q{=}1$ functional, the absolute change in image AUROC averaged 0.0144 percentage points and reached 0.2778 on MVTec AD `toothbrush`. Whether that is negligible depends on the category rather than on the movement. Measured against each category's own access effect, the gap between its full-pool and one-image values (median 2.82 percentage points), the movement is 1.3% at the median, but reaches a tenth or more in 3 of 27 categories and 55.6% on VisA `fryum`. A first, smaller probe had put the effect two orders of magnitude below the access effects of Section 4. Its figures reproduce exactly, but it covered a single cell in three categories; across every cell and every category the bound holds at the median and fails in the tail. The lesson runs in both directions: an unbound degree of freedom matters most exactly where the effect being measured is smallest, and a measurement whose scope is unstated will be read as wider than it is. No reported value of our scorer is affected, because its frozen runs and every reported rerun of it used the same batch size (the archived official MuSc runs of Table 21, a complete-protocol placement, used four), and the batch-size-two rerun reproduced the image AUROC of all 702 registered cells exactly. We record this as a defect of the contract rather than of the runs, and draw two corrections from it: the batch size belongs in the cell identity, and the environment should be stored in plaintext rather than only as a one-way hash, because a hash cannot be inspected after the fact when a cross-environment reproduction fails.

X Registration status of every protocol and of the post hoc analyses

A paper about what may be compared owes its reader the same record for its own evidence. Table 40 lists every protocol in this paper, with the evidence that it was fixed before its own values existed, and Table 41 lists the post hoc and exploratory analyses. All commitments are internal: none carries a third-party timestamp. The registration documents are withheld from the anonymous artifact because they name internal paths and hosts; the hashes below identify them for release after review.

Table 40: Registration status. *Before its values*: the evidence that the protocol was fixed before any of its values existed. Hash prefixes identify documents or code; "dated document" means that only the document's date, not a hash, predates the values.

Protocol or analysis	Before its values	Role here
v4 (`fixed1nn-access-v4`)	registry and support manifest pinned before execution; code hash in each cell identity; code `2d0acfec`; a locked replication of a direction explored on earlier runs	primary endpoint (C1)
BTAD (`btad-access-v1`)	registry, manifests and code frozen before any BTAD value; upload and final code compared (Appendix B)	stress test, chosen after v4
Median (`secondagg-median-image-v1`)	no registration document survives; runner `64f74225` shipped, archived after the run	exploratory sensitivity
Full-pool reductions	registration record with runner and analysis hashes (`62b9fc5c`); one amendment, to a tolerance in its code tests (not in the verdict rule), before any value	reduction sensitivity
Fixed width (`fixedq-width-v1`)	dated contract, no hash recorded; runner `6e7132f0`	batch-size-two arm only
Fixed width on BTAD	runner hash in the registration record	not used
DINOv2 (`dinov2-budget-v1`)	dated document; frozen part matches the prefix `f9f11cb6` recorded at adjudication; code `d1bd8c1b`; earlier launches archived	second encoder
k=32 extension	document hash chain recorded before the run (`10630df5`); code `e52d2585`; first smoke run failed the seam, rerun after a fix (deviation)	extended grid
Cardinality	same hash chain; base code `2d0acfec`; runner of the registered cells recorded (`e79491d5`) but not archived	negative registered result
Decomposition and Real-IAD (`decomp-v1`)	registration `4a5763bd` and amendment `ecd0ef7e` before any value; code frozen; two deviations and one interim look disclosed (Appendix R)	C1 decomposition, scorer family, Real-IAD
Five-path follow-up (`decomp-v2-followup`)	registered after the decomposition results above were known, before any of its own values; registration `3ae39457`	five-path decomposition, A5 at b_0, scorer grid (archived), Real-IAD under the subspace scorer; reported beside, not instead of, the verdicts above
Decomposition under DINOv2 (`decomp-dinov2-f5`)	registered after the decomposition results above were known, before any of its own values; registration `dbf4d2ff`; a first launch was refused before any value because its frozen inputs had not been placed on the host	second-encoder decomposition
Pixel-level decomposition (`decomp-pixel-f7`)	registered after the decomposition results were known, before any of its own values; registration `a9d92ca6`; two input-plumbing defects caught by its seam gate and fixed before any accepted value	pixel metrics of the decomposition
Interval-average decomposition (`decomp-ia-f6`)	same registration, `a9d92ca6`; code frozen before any value	decomposition under MuSc's reduction
Real-IAD, both reductions (`realiad-ia-f9`)	registered after the results above were known, before any of its own values; registration `3d6564a3`; a descriptive amendment (`6a9eba37`) frozen before any contrast value was read	budget and purity without a ceiling
Official MuSc sweep (`musc-official-f4`)	registered, then withdrawn before any value when earlier archived runs of the same program were found	not run

Table 41: Registration status, continued: the post hoc and exploratory analyses. None was registered before its values.

Analysis	Status	Role here
Shot-equivalent	post hoc	descriptive unit
Official MuSc placement	post hoc, on archived runs of the unmodified program	Appendix O
Batch-size probe	post hoc, on the fixed-width code	self-audit (Appendix V)
Category-level inference; published gaps	post hoc	Appendix U
Access audit of 12 papers; multi-table coding of 12 papers, six shared	post hoc, descriptive	C3 evidence
Functional-matched k=4 placement	post hoc; the registration barred this comparison	Section 4.2; Appendix N
Median-versus-minimum and per-category contrasts of the full-pool arm	post hoc, descriptive	Appendix N
AnomalyDINO artifact-only recomputation and its k ladder	post hoc, on archived runs	Appendices K and U
Coding of one published table; published shot-equivalents	post hoc, descriptive; transcribed	Appendices F and O
Second-encoder sensitivities (Kendall τ_b, tie-broken orderings, non-trivial and headroom subsets, per-path counts, encoder invariance)	post hoc, exploratory or descriptive	Appendix P
Cardinality: resolvable stratum, floor movement, AUROC by cardinality; stratified rerun	post hoc, exploratory; rerun outside the registration	Appendix Q
Source $\times$ purity main effects and interaction	post hoc, on arms of the registered follow-up	C1 attribution; Appendix R
Purity condition in all categories; post hoc reading of the MVTec AD interval-average margin in categories with headroom	post hoc, descriptive	Appendix R

Y REGISTERED PROTOCOLS NOT USED FOR MAIN-TEXT CLAIMS

Two further protocols were registered and run to completion; we summarize their registered outcomes but draw no main-text claim from those outcomes, and we list them so that the record is complete. The first fixes the absolute aggregation width, rather than a fraction of the bank, on both arms across MVTec AD and VisA. On VisA the position of the full-pool arm is unchanged at the macro level, though not in every category; on MVTec AD it depends on the chosen width, which is the limit Appendix N already reaches: no single aggregation setting matches both arms at once. Its $q{=}1$ cells, which repeat the frozen v4 functional, are used once: they are the batch-size-two arm of the probe in Appendix W. The second adds a full-pool arm, the intermediate budgets and a registered purity prediction to BTAD. The prediction held at every registered width, but with three categories it is directional only. A reproduction probe in the same run matched 10 of the 30 earlier BTAD cells bitwise in image AUROC and 13 within the last floating-point digit (the 10 included); the others differ by at most 0.10 percentage points, no category or path contrast changes sign, and the mean $k{=}16$ minus $k{=}1$ contrast is $+3.43$ in both runs. We attribute the remaining differences to the accelerator model but have not confirmed it. Apart from the batch-size arm, neither protocol changes a main-text number. Both registrations required a per-cell reproduction table in an appendix; we give the reproduction counts instead, here and in Appendix W, and the per-cell values behind them and behind the per-category exceptions the first registration named are archived with the registrations, outputs and attempt ledgers.