REFERENCE SATURATION: NORMAL EXEMPLARS BUY STRUCTURAL ANOMALY DETECTION TWICE AS FAST AS LOGICAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero- and few-shot anomaly detectors are almost always reported on MVTec LOCO as a single pooled AUROC, even though the benchmark separates *structural* anomalies, whose local appearance is atypical, from *logical* anomalies, whose every local region looks normal and only the composition is wrong. We re-measure the reference-based family under one protocol, split by subset, and find a stable asymmetry: going from zero to thirty-two normal exemplars closes 41.9% of the remaining headroom on structural anomalies but only 20.7% on logical ones, a ratio of 2.02 (95% CI $[1.67, 2.51]$), which an independent redraw of the thirty-two exemplars reproduces at 2.02 $[1.69, 2.51]$; it is 2.00 at sixteen exemplars and, for a second method with a different mechanism measured from its one-exemplar baseline, 1.75, and every interval excludes one. Reference *diversity* then saturates, and reference *depth* recovers little of what it leaves. Adding a third reference branch to a two-branch combination moves logical AUROC by at most 0.2 points while roughly doubling the exemplar budget, and spending the entire budget on one method's exemplar diversity buys nothing measurable: three independently seeded WinCLIP+ branches at 46.4 exemplars reach 66.4, against 66.1 for a single branch at 16 and 69.6 for the two-branch combination at that same 16; pooling the same budget into one 32-exemplar memory per branch instead, the strongest reference-only configuration we could build, reaches 70.5, a gain of 0.7 over the best diverse one that we cannot resolve. Replacing that third branch with a compositional part-count score — whose per-image scores on logical anomalies agree with the reference branches far more weakly than they agree with each other — raises logical AUROC from 69.6 to 73.3 ± 1.2 over three independent draws of the composition branch's own references, at no more than 32 images per category. Against that deeper control the paired bootstrap over images gives $+2.85$, 95% CI $[0.97, 4.79]$, and against the best three-branch reference configuration $+3.58$, 95% CI $[1.77, 5.40]$; structural AUROC does not move resolvably against either. With all three branches reading the same sixteen images, half the control's budget, the gain over it is $+3.74$, 95% CI $[1.88, 5.66]$. We report a pre-specified control that falsified our own initial two-branch claim, four protocol anchors against published numbers, and a spread of up to 17.8 points across three published tables for the same baseline at the same shot count. All per-image scores, exemplar manifests and per-category budgets are released. We make no state-of-the-art claim.

1 INTRODUCTION

MVTec LOCO (Bergmann et al., 2022) draws a distinction that most anomaly detection benchmarks do not. A *structural* anomaly is locally atypical: a scratch, a contamination, a wrong texture. A *logical* anomaly is not. Every local region of a logical anomaly is drawn from the normal distribution of local regions; what is wrong is the arrangement — a missing part, a duplicated part, an item in the wrong compartment, two components that should not appear together. The distinction is not taxonomic bookkeeping. It determines which scoring functions can work in principle.

The dominant approach to zero- and few-shot industrial anomaly detection is *reference-based*: given k normal exemplars of a category, score a test image by how far its local features are from the

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.

nearest normal feature. The two mechanisms we measure are the two that dominate it — a non-parametric visual patch memory and vision–language window retrieval — and our conclusions are about scoring by appearance distance to stored normal features, not about every method that consumes normal exemplars. WinCLIP+, PatchCore and PromptAD all instantiate this idea (Jeong et al., 2023; Roth et al., 2022; Li et al., 2024b). On MVTec LOCO this family is almost always reported as a single pooled AUROC — CLIP (Radford et al., 2021) 53.6, WinCLIP 59.1, AnomalyCLIP 61.9, AnomalyAgent 62.8, all four as tabulated by Zhang et al. (2026). We quote that table and do not read the four values as a homogeneous set: AnomalyCLIP learns its prompts on auxiliary anomaly data and is therefore not training-free in the sense the other rows are, and neither it nor CLIP evaluates on LOCO in its own paper, so those two entries originate somewhere other than the work being cited for them. Methods that use the full normal set do report the logical/structural split (Zhang et al., 2025a; Patra & Ben Taieb, 2024); in the few-shot regime the splits we are aware of are system-internal ablations of LogSAD and UniVAD (Appendix G), a reference-based baseline is split only by LogicAD and across two tables, and no work measures how either subset responds to exemplar count. Two basic questions therefore have no published answer under one controlled protocol. Does this family work on logical anomalies at all? And does adding normal exemplars help the two subsets equally?

We answer both by re-measuring the family under one protocol, split by subset, with the exemplar budget as the independent variable: $k \in \{0, 1, 2, 4, 8, 16, 32\}$ for WinCLIP/WinCLIP+ and $k \in \{1, 4, 16, 32\}$ for PatchCore, which has no zero-exemplar mode, three random draws each. Our protocol is checked against four published anchors and agrees with all of them to within 1.5 points (Section 3). Every main-text normalisation is inductive; no test-set statistic enters it.

Findings. **(i) Asymmetry.** Going from 0 to 32 exemplars absorbs 41.9% of the structural headroom and 20.7% of the logical headroom, a ratio of 2.02 (95% CI $[1.67, 2.51]$); at 16 exemplars the ratio is 2.00, and for a second method with a different mechanism it is 1.75 $[1.36, 2.35]$. The cumulative ratio barely moves as the budget doubles and survives an independent redraw, so it is not an artefact of one operating point; two methods and five categories bound how far it generalises (Section 7).

(ii) Saturation. Once two reference branches are combined, a *third* reference branch changes logical AUROC by at most 0.2 points while raising the unique exemplar budget from 16 to 31.5 images per category. Spending the entire budget on one method's exemplar diversity buys nothing measurable: three independently seeded WinCLIP+ branches at 46.4 exemplars reach 66.4 logical, against 66.1 for a single branch at 16 and 69.6 for the two-branch combination at that same budget. Spending it on *depth* instead — one 32-exemplar memory per branch, the strongest reference-only configuration we could build — reaches 70.5, a logical gain of 0.7 over the best diverse configuration that the bootstrap cannot resolve (95% CI $[-0.5, 2.0]$); depth does buy about a point on structural anomalies (Section 5).

(iii) Complementarity. Replacing that third branch with a compositional part-count score — whose per-image scores on the logical subset agree with either reference branch far more weakly than the two agree with each other — raises logical AUROC to 73.3 ± 1.2 over three draws of its own references, at 31.6 exemplars, above every reference-only configuration we measured at any budget: by 2.85 points (95% CI $[0.97, 4.79]$) against the deeper 32-exemplar control and by 3.58 ($[1.77, 5.40]$) against the best three-branch one, with structural AUROC not resolvably changed against either. Reading the same sixteen images as the reference branches, at half the control's budget, it is $+3.74$ ($[1.88, 5.66]$) above it.

A falsified claim, reported. Our initial claim concerned a *two*-branch fusion. A pre-specified control showed that substituting a second reference branch for the compositional branch reaches the same number, so a two-way fusion result cannot be attributed to composition; the claim was withdrawn and the three-branch controls in Section 5 were designed to answer what remains after saturation. The question, the decision rule and the outcome, as they were written down before the run, are reproduced in Appendix F. We also report two label-free tests that failed to support our preferred mechanistic explanation (Section 7). We state the effect as a ratio, not as a null effect, because that is what the data support.

Relation to the state of the art. LogSAD (Zhang et al., 2025a) reaches 86.3 with four normal exemplars and 90.2 with the full normal set, above every configuration in this paper. Its own ablation reports the same effect in the same direction at a much stronger operating point: adding

its composition branch to its patch branch gains 14.4 logical points and loses 1.5 structural ones, uncontrolled and on a single system; UniVAD's one-exemplar ablation (Gu et al., 2025) agrees in direction. ObjectCore (Fučka et al., 2026) reaches 80.8 at four exemplars with the same broad recipe executed far more carefully. Appendix G reads all three against this paper in detail. Our contribution is the controlled, cross-method, budget-matched generalisation of that observation, together with the point at which it stops.

2 Related Work

Reference-based zero- and few-shot detection. WinCLIP (Jeong et al., 2023) scores windows against text prompts and, in its WinCLIP+ variant, against a memory of normal exemplar features. PatchCore (Roth et al., 2022) uses no text at all: it stores a coreset of normal patch features and scores by nearest-neighbour distance. AnomalyCLIP (Zhou et al., 2024) learns object-agnostic prompts. WinCLIP+ and PatchCore, despite differing mechanisms, share the property that anomaly evidence is a distance to a normal reference, which is precisely the property we study; AnomalyCLIP and zero-shot WinCLIP compare against text rather than exemplars.

Logical anomaly detection. Methods designed for logical anomalies typically add an explicit representation of composition: GCAD's global branch (Bergmann et al., 2022), set-based descriptors in SINBAD (Cohen et al., 2023), component histograms in PSAD (Kim et al., 2024), component semantics in SALAD (Fučka et al., 2025), and, in the vision–language setting, LogSAD's checklist-driven composition matching (Zhang et al., 2025a), SAM-LAD's reference matching (Peng et al., 2025) and LogiCode's program synthesis (Zhang et al., 2025b). These works operate with full normal sets or with task-specific pipelines; none isolates how much of their advantage is attributable to composition as opposed to a stronger reference.

Few-shot logical anomaly detection. A third line attacks both problems at once, and all of it puts an explicit compositional representation on top of a reference: a vision–language description in LogicAD (Jin et al., 2025), component clustering and a graph over components in UniVAD (Gu et al., 2025), and bipartite matching of object representations in ObjectCore (Fučka et al., 2026), which reaches 80.8 on MVTec LOCO with four exemplars. LogicAD reports WinCLIP's logical subset in its one-shot table (64.3) and its structural subset in a separate table with no stated shot count (64.6); none of these works varies the exemplar budget with that budget matched across arms. A fourth line changes what the references are rather than how many: InCTRL (Zhu & Pang, 2024) learns in-context residuals from a support set rather than treating it as a memory. Our question sits before all of these — what normal references alone buy, and where they stop paying.

Evaluation practice. LogSAD, SALAD and ULSAD (Patra & Ben Taieb, 2024) report the LOCO split under the full-data protocol. In the few-shot regime, the subset-resolved numbers for a system we are aware of are the ablations of LogSAD's three detectors and UniVAD's two modules; for a reference-based baseline, only LogicAD prints both halves, in two tables; and no work measures how the split responds to exemplar count. Section 6 documents a further problem: the same baseline at the same shot count appears in three published tables with a spread of 17.8 points.

3 Setup

Data and metric. MVTec LOCO, five categories, 1568 test images (575 good, 561 logical, 432 structural). For each category we compute image-level AUROC of good versus each anomalous subset, and average over the five categories. We do not pool images across categories: doing so gives 54.7 for zero-shot WinCLIP (55.1 after per-category min–max normalisation), which is not comparable to the 59.1 reported in the literature, whereas the mean-over-categories protocol gives 58.6. This 3.9-point gap between two defensible aggregation rules reappears in Section 6. The pooled column of Table 2 is good against the union of both subsets, computed the same way: within each category it is the size-weighted mean of the two subset AUROCs, so it cannot be recovered from the printed columns and the global subset totals (that gives 59.2, not 58.6, for zero-shot WinCLIP), and the released audit script recomputes it exactly for every row.

Table 1: Protocol anchors. Our reproduction against published values, with the source of each published number.

Setting	Dataset	Published	Source	Ours
WinCLIP, 0 exemplars	MVTec LOCO	59.1	Zhang et al. (2026)	58.6
WinCLIP+, 1 exemplar	MVTec LOCO	61.2	Zhang et al. (2026)	62.3 ± 1.8
WinCLIP+, 4 exemplars	MVTec LOCO	63.9	Zhang et al. (2026)	65.3 ± 0.5
WinCLIP+, 1 exemplar	MVTec AD	93.1	Jeong et al. (2023)	93.2 ± 0.4

Implementation. WinCLIP and WinCLIP+ use the Anomalib implementation (v2.6.0) with the OpenCLIP ViT-B-16-plus-240 backbone pretrained on LAION-400M, at its native 240×240 input resolution and with that implementation's prompt ensemble. PatchCore is our own implementation of Roth et al. (2022): WideResNet50-2 with ImageNet weights, layer-2 and layer-3 features with 3×3 average pooling, bicubic resize to 256 followed by a 224 centre crop, and no coreset subsampling. The composition branch counts parts with OWLv2 (`owlv2-base-patch16-ensemble`) at a per-class confidence threshold of 0.15; it is the only branch with an external detector and, as Section 7 shows, it is the weakest link. Every normal image used anywhere in this work, as exemplar or for calibration, comes from the LOCO `train/good` split; the LOCO validation split is never touched, which matters because a competing convention does use it (Section 6). Package versions and the full command lines are in the released artefact.

Protocol validation. Table 1 checks our implementation against four published numbers. All four agree to within 1.5 points, including one on a different dataset, which controls for the few-shot code path separately from the LOCO scoring path. The three LOCO anchors come from a single source, so our LOCO numbers are calibrated to the convention that reproduces Zhang et al.'s table — one side of the disagreement documented in Section 6, and not a validation of it (Appendix A). All four anchor WinCLIP or WinCLIP+. Our PatchCore is unanchored: published four-exemplar LOCO values for it span 63.5 to 68.7 (Section 6), ours gives 62.5 ± 1.8 pooled, below all three, and 56.6 ± 2.0 at one exemplar, where its ratio in Section 4 is anchored, against 60.9 in Zhang et al. (2026), and there is no single value to anchor against. We therefore treat it as a second reference mechanism measured under our protocol, not as a reproduction of a published system.

Branches. *WinCLIP+* and *PatchCore* are the two reference branches, configured as above; PatchCore is left unsubsampled because for $k \leq 32$ the memory is already small and a coreset would add a hyper-parameter we would have to justify. The *composition* branch scores an image by the deviation of its open-vocabulary part counts from the counts observed on its own draw of 16 normal exemplars, over three draws. Drawn independently of the reference exemplars, the three-branch configuration costs 31.6 images per category (Appendix H); counting against the reference branches' own images (*shared*), it costs 16. Its score is λ times a composition term plus $1 - \lambda$ times a structural fallback, a patch-level appearance score that keeps the branch from being blind to structural anomalies; we fix $\lambda = 0.5$ a priori and Appendix C reports both endpoints, including the fact that $\lambda = 1$ scores higher on logical alone and was nonetheless not adopted. The part names the detector is prompted with are a short hand-written list per category, written once from the category description and never revised against any score; this is a prior the reference branches do not receive. The branch is deliberately simple, and Section 7 documents how weak it is and what the prior costs.

Normalisation and fusion. Every branch is standardised using 16 *held-out* normal images per category. These are disjoint from the exemplars of every configuration that standardisation actually enters, that is every branch at $k \leq 16$ and the composition branch, and are never placed in those memory banks. The transform is a per-category z-score and nothing else: a raw score s in category c becomes $(s - \mu_c)/\sigma_c$ for the sample mean and standard deviation of that branch on the 16 calibration images of c. Scores are not clipped; fusion is the unweighted mean of the standardised scores. One exception is worth stating: at $k = 32$ the exemplar draw is large enough to collide with the calibration set, and 14, 5 and 12 of the 160 exemplars per draw are also calibration images. This cannot affect any number we report. The two $k = 32$ rows of Table 2 are single-branch AUROCs computed from raw scores, and AUROC is invariant to the per-category affine rescaling that standardisation applies. The one fused configuration at $k = 32$, WC_{32} + PC_{32}, uses a separate redraw taken from the training normals minus the calibration images and is standardised by exactly the rule above (Appendix I). The

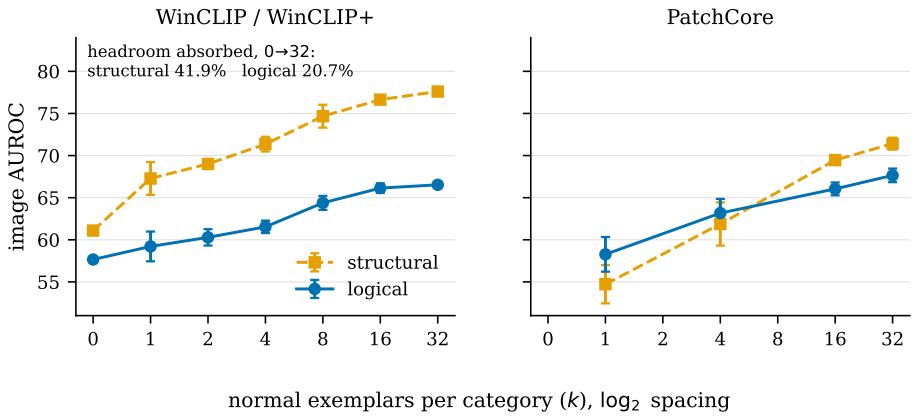

Figure 1: Adding normal exemplars helps structural anomalies about twice as much as logical ones, for two methods with different mechanisms. The horizontal axis is $\log_2 k$ in both panels, so slopes are comparable across panels; the zero-exemplar point is placed one unit to the left of $k{=}1$. Error bars are the sample standard deviation over three exemplar draws, which is the draw-to-draw component only; Section 7 discusses the larger between-category component. PatchCore has no zero-exemplar mode.

released manifests list every exemplar and calibration filename, so the overlap and its consequences can be checked directly. Fusion weights are fixed a priori at $1/n$ for n branches; no weight is searched, and no test label is used at any point. The transductive alternative, which normalises with test-split statistics, is better for us in 17 of 18 cells and is reported in Appendix B for completeness only.

Budgets. Because different branches may draw different exemplars, every configuration is reported with the size of the *union* of exemplar filenames it consumes, per category, averaged over the five categories, which is why some budgets are not integers. This matters: WinCLIP+ and PatchCore at the same seed draw the same 16 images, so their combination costs 16, whereas two differently seeded reference branches cost ≈ 31 and three cost 46.4. One cost sits outside the column because every configuration pays it alike: the 16 held-out images per category used for standardisation. Counting them adds a constant 16 to every row of Table 2 — 47.6 unique normal images per category for the three-branch composition configuration against 47.5 and 48 for the reference-only ones — and leaves every comparison unchanged. We state it because Section 6 holds other work to the same accounting.

4 What Normal Exemplars Buy

Figure 1 shows the two subsets as a function of exemplar count. Structural AUROC rises steeply and logical AUROC rises slowly, for both methods. Because the two subsets start at different levels, we quantify the effect as the fraction of remaining headroom absorbed, $(A_k - A_0)/(100 - A_0)$.[1]

For WinCLIP+, $0 \to 16$ absorbs 40.0% of the structural headroom and 20.0% of the logical headroom, a ratio of 2.00 (95% CI $[1.64, 2.53]$); $0 \to 32$ absorbs 41.9% and 20.7%, a ratio of 2.02 ($[1.67, 2.51]$). For PatchCore, $1 \to 16$ gives 32.4% and 18.5%, a ratio of 1.75 ($[1.36, 2.35]$). All three intervals exclude 1 and overlap one another; the bootstrap resamples images within category and subset and averages over the three exemplar draws. On WinCLIP+'s first draw from $k = 0$ to 4, a paired bootstrap over images ($B = 3000$, resampling within category and subset) puts the difference in raw AUROC gain at $+7.8$ points in favour of structural, 95% CI $[+4.9, +10.9]$, $p < 0.001$ ($+6.4$ averaged over the three draws); here and throughout, p-values below the resolution of $B = 3000$ resamples are obtained from a normal approximation to the bootstrap distribution and are reported only as bounds.

[1]The headroom normalisation is not what produces the asymmetry: the raw gains, $+16.3$ structural against $+8.7$ logical, are already in a ratio of 1.87. Percentages are computed from full-precision AUROC; recomputing them from the rounded values printed in Table 2 reproduces every percentage to within 0.2 points and every ratio to within 0.02. The PatchCore ratio is anchored at its one-exemplar operating point, which is the leftmost PatchCore point of Figure 1 and is in the released k-sweep rather than in Table 2.

A single-point ratio could reflect where each subset happens to sit on its own saturation curve. What we consider load-bearing is that the ratio exceeds one for both methods, at every cumulative endpoint and under an independent redraw; its near-constancy between 16 and 32 exemplars is weaker evidence, for the reasons below. The two methods are also anchored at different budgets above, WinCLIP+ at $k = 0$ and PatchCore at $k = 1$; over the matched $1 \to 16$ interval they agree more closely still, at 1.68 $[1.36, 2.15]$ and 1.75 $[1.36, 2.35]$. An independent redraw of the 32 exemplars (Appendix I) reproduces the WinCLIP+ ratio at 2.02 $[1.69, 2.51]$, and gives 1.93 $[1.50, 2.60]$ for PatchCore from $k = 1$ (1.65 $[1.31, 2.13]$ on the original draw). Two cautions belong with it: the measurements at 16 and 32 share seeds and overlapping exemplar sets, so they are not independent replicates, and both sit close to saturation, where stability is weaker evidence than it would be over a wider range. A third: every ratio above is cumulative from the anchor, and since more than four fifths of the absorption (nine tenths for WinCLIP+) happens by $k = 16$, the $k = 32$ value is dominated by the first interval. Measured *incrementally*, the ratio is 2.88 $[1.96, 5.38]$ then 1.54 $[1.20, 2.05]$ for WinCLIP+ over $0 \to 4$ and $4 \to 16$, and 1.35 $[0.91, 2.18]$ then 2.55 $[1.56, 6.55]$ for PatchCore over $1 \to 4$ and $4 \to 16$; over $16 \to 32$ neither is measurable, because under five per cent of the logical headroom is absorbed there and the ratio's denominator is near zero (Appendix I). We claim persistence of a cumulative asymmetry into the saturated regime, not a constant marginal rate.

What the error bars do and do not cover. Deviations are over exemplar draws with the categories fixed; the bootstrap resamples images within category and subset and macro-averages as the tables do. Neither measures how far the five-category mean would move on other categories, which is much larger: at $k = 16$ the per-category logical AUROCs have standard deviation 6.2, a standard error of 2.8 on their mean against a draw-to-draw deviation of 0.6. Every claim is conditional on these five categories.

The effect is visible per image, not only in aggregate (Figure 3, right). Adding four exemplars makes 44.1% of logical anomalies score *more* normal relative to the normal distribution, against 30.5% of structural anomalies; the median logical image does not move at all, while the median structural image moves $+5.3$ percentile points. The logical mean still improves because the minority of logical images that do move, move a long way.

5 Saturation and Complementarity

Reference *diversity* saturates, and depth recovers little. The five fusion rows of Table 2, from WC + PC down to PC + PC′ + PC″, are the control that a single-branch budget match does not provide. Adding a third independently seeded reference branch to WC + PC moves logical AUROC by -0.2 to $+0.2$ points while roughly doubling the exemplar budget. Three WinCLIP+ branches at 46.4 exemplars give 66.4 against 66.1 for one at 16: some thirty further images per category buy no logical evidence. The natural objection is that nearest-neighbour memories are scaled by pooling exemplars into one bank, not by averaging independently drawn banks. One 32-exemplar memory per branch, WC_{32} + PC_{32}, is the strongest reference-only configuration we could build: 70.5 ± 0.3 logical, $+0.73$ over the best diverse one (95% CI $[-0.51, +1.95]$, $p = 0.26$), so depth buys at most a fraction of a point on logical anomalies that we cannot resolve, though about a point on structural ones ($+1.16$, $[+0.01, +2.28]$). It is the strongest control below.

The composition branch is not redundant with references. Substituting the composition branch for that third reference branch raises logical AUROC from 69.6 to 73.3 ± 1.2, over three draws of the composition branch's own references each paired with the reference branches of the same seed, at 31.6 exemplars per category against 31.5 and 32 for the reference-only configurations. A paired bootstrap ($B = 3000$) that resamples images and averages over the same three assignments Table 2 averages over gives, against the strongest control, WC_{32} + PC_{32}, $+2.85$, 95% CI $[+0.97, +4.79]$, $p = 0.003$ on logical and -0.50, $[-2.28, +1.22]$, $p = 0.57$ on structural. We specified this contrast in advance as deciding the claim. Against WC + WC′ + PC it gives $+3.58$ $[+1.77, +5.40]$ logical and $+0.66$ $[-1.11, +2.36]$ structural, against WC + PC $+3.76$ $[+2.17, +5.33]$ and $+1.38$ $[-0.19, +2.91]$, and against WC + PC + PC′, whose second PatchCore branch costs it 1.1 structural points, $+3.95$ $[+2.22, +5.63]$ and $+2.48$ $[+0.88, +4.14]$. Every logical interval excludes zero; against the two

Table 2: Every configuration with the number of *unique* normal exemplars per category it consumes, averaged over the five categories. Mean ± sample s.d. over three exemplar draws, with the five categories held fixed (Section 4). WC = WinCLIP+, PC = PatchCore, Comp. = composition branch; a prime denotes an independently drawn exemplar set. The primary metric is logical AUROC. Three rows carry no deviation: the zero-shot row is deterministic and the two three-branch single-method rows consume all three draws at once. Composition rows are over three draws of that branch's own references, each paired with the reference draws of the same seed. Budgets exclude the 16 calibration images per category, which every row consumes alike (Section 3). Bold marks the best logical and pooled value; the structural column is left unbolded because that difference is not resolved (Section 5). "Shared": the composition branch counts parts against the same 16 images as the two reference branches, so the three branches together consume 16. †One 32-exemplar memory per branch, with the exemplars redrawn from the training normals minus the calibration images so that this fused row is standardised like every other (Section 3); the two single-branch $k = 32$ rows above use the original draw.

Configuration	Exemplars	Logical	Structural	Pooled
WinCLIP (zero-shot)	0	57.7	61.1	58.6
WinCLIP+	16	66.1 ± 0.6	76.6 ± 0.2	70.2 ± 0.2
WinCLIP+	32	66.4 ± 0.4	77.4 ± 0.8	70.7 ± 0.6
PatchCore	16	66.0 ± 0.8	69.5 ± 0.3	67.2 ± 0.6
PatchCore	32	67.6 ± 0.8	71.4 ± 0.7	69.0 ± 0.3
Composition (ours)	16	65.7 ± 1.0	62.3 ± 0.7	63.9 ± 0.8
WC + PC	16	69.6 ± 0.8	78.1 ± 0.3	72.9 ± 0.5
WC + PC + PC′	31.5	69.4 ± 0.6	77.0 ± 0.2	72.4 ± 0.3
WC + WC′ + PC	31.5	69.8 ± 0.5	78.8 ± 0.3	73.3 ± 0.4
WC + WC′ + WC″	46.4	66.4	76.9	70.5
PC + PC′ + PC″	46.4	67.0	70.7	68.3
WC_{32} + PC_{32} †	32	70.5 ± 0.3	80.0 ± 0.3	74.3 ± 0.2
WC + PC + Comp.	31.6	73.3 ± 1.2	79.5 ± 0.6	75.6 ± 0.9
WC + PC + Comp., shared	16	$\mathbf{74.2 \pm 0.9}$	79.7 ± 0.5	$\mathbf{76.2 \pm 0.5}$

strongest controls no structural one does, and we read those neither as an effect nor as its absence.[2] The gain we can resolve against the strongest controls is confined to the logical subset, and the intervals are not corrected for multiple contrasts. A secondary configuration specified in advance lets the composition branch count against the *same* 16 images as the reference branches, half the control's budget. It reaches 74.2 ± 0.9 logical, $+3.74$ $[+1.88, +5.66]$ over WC_{32} + PC_{32} and $+4.65$ $[+3.03, +6.26]$ over WC + PC at equal budget: the evidence it adds does not come from extra images. Nor does it hinge on the composition draw: the branch's own logical AUROC has s.d. 1.0 over its three draws, under a third of the 3.2 a first-order argument puts as needed to overturn the primary contrast (Appendix H).

Why the branches are not redundant. Figure 3 (left) gives the mechanism-level reason. The two reference branches agree with each other at Spearman $\rho = 0.22$ to 0.34 on the anomalous subsets, despite one using vision–language windows and the other a purely visual memory. The composition branch agrees with either far more weakly on the logical subset, where the gain appears, at $+0.01$ and $+0.03$; on structural anomalies its agreement with PatchCore, 0.18, nears their 0.22, as expected of a branch with a patch-level structural fallback. An ensemble gains from disagreement, and there is much less of it left inside the reference family.

Is the gain carried by one category? The composition branch is far stronger on breakfast box than on the other four categories (Table 4), so the three-branch gain could in principle be a single-category effect. Against WC + PC it is positive in all five; against the stronger controls it is weaker, positive in four of five against WC + WC′ + PC and three of five against WC_{32} + PC_{32}, and $+1.40$ against the latter once breakfast box is removed. The claim rests on the image-level bootstrap; five categories cannot support a sign test (Appendix E).

[2]The bootstrap resamples images while the table averages per-category AUROC over draws; the two estimands are close, not identical, and with the assignments matched agree to within 0.1. One image resample per replicate is shared by the three assignments, which differ in exemplars, not in test images; resampling independently per assignment would narrow the interval by about $\sqrt{3}$.

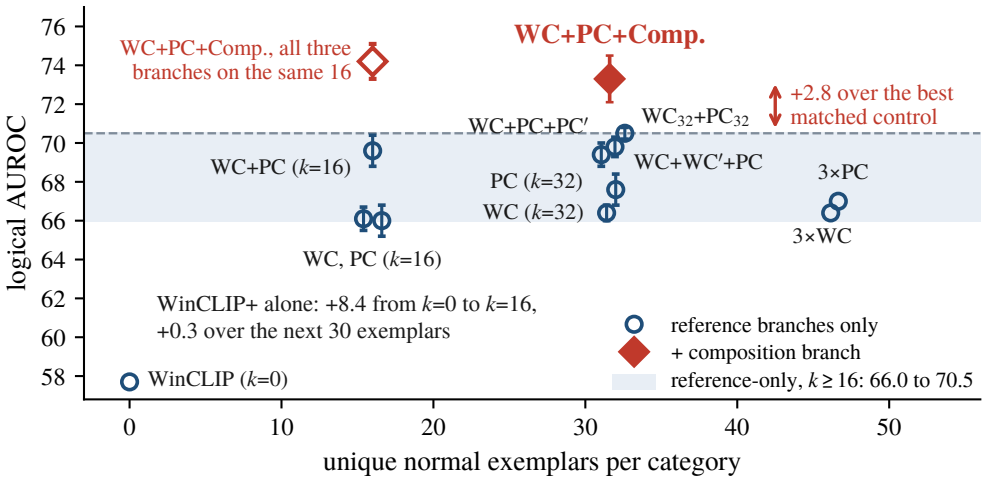

Figure 2: Logical AUROC against the unique normal-exemplar budget. Every reference-only configuration of Table 2 with $k \geq 16$ falls in the shaded band regardless of how many exemplars it is given; spending the budget on reference diversity does not escape it, and spending it on one deeper memory per branch, WC_{32} + PC_{32} at the top of the band, escapes it by a margin we cannot resolve. Replacing one reference branch with the composition branch does, at 31.6 against 32 (filled diamond, three draws of its own references), and so does letting all three branches read the same 16 images (open diamond). The zero-exemplar point is plotted so that the rise which the band flattens is visible: along the WinCLIP+ ladder alone, the first 16 exemplars are worth 8.4 points as tabulated (8.48 unrounded) and the next 30 are worth 0.3. The arrow is the Table 2 difference against the strongest budget-matched reference-only configuration, WC_{32} + PC_{32} (paired bootstrap $+2.85$, 95% CI $[+0.97, +4.79]$). Error bars are the sample standard deviation over three exemplar draws, absent for the three configurations that use all three draws at once or none. Coincident points are jittered horizontally; between 31.5 and 32.0 the configurations are closer together than one marker, so left-to-right order inside that cluster carries no information.

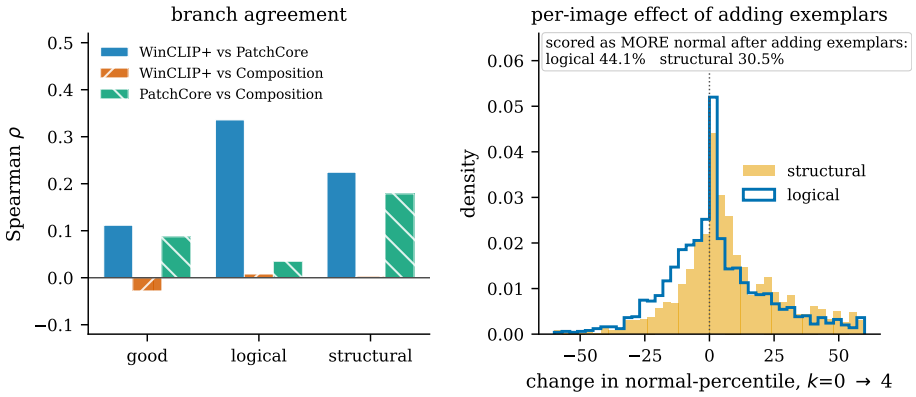

Figure 3: Left: per-image rank agreement between branches (Spearman ρ; mean over categories and matched draws). The two reference branches correlate; the composition branch is far less correlated with either of them on the logical subset, and with WinCLIP+ on every subset. Right: distribution of the change in normal-percentile when four exemplars are added, for WinCLIP $\rightarrow$ WinCLIP+, pooled over three draws; the dotted line marks no change. Mass to the left of zero means the image became harder to detect.

6 Baseline Numbers on LOCO Are Not Comparable Across Papers

WinCLIP+ with four normal exemplars on MVTec LOCO is reported as 63.9 by AnomalyAgent (Zhang et al., 2026) and as 71.3 by LogSAD (Zhang et al., 2025a): a 7.4-point disagreement for the same method, dataset and shot count. PatchCore under the same setting appears as 64.2 and 68.7. A third table widens the spread rather than settling it. Fučka et al. (2026) list a WinCLIP row in each of their zero-, one-, two- and four-exemplar blocks at 53.2, 51.6, 52.6 and 53.5; the value moves with the shot count from the one-exemplar block onward, so the row in the four-exemplar block

is the few-shot variant. They give 63.5 for PatchCore in that same block. Across the three tables, WinCLIP+ at four exemplars spans 17.8 points, from 53.5 to 71.3, and PatchCore spans 5.2, from 63.5 to 68.7. Our own reproduction, anchored as in Table 1, gives 65.3 ± 0.5.

We do not attempt to adjudicate, but two different questions are tangled together in that spread and separating them is most of the work. *Why two reproductions of the same baseline disagree*: aggregation, which moves our own zero-shot number by 3.9 points (Section 3); the variant a row labelled WinCLIP holds, which Fučka et al. (2026)'s blocks show going both ways; and implementation — backbone, pretraining corpus, input resolution, coreset ratio — none of which the shot count records. *Why a multi-stage system cannot be read against a budget-matched baseline*: LogSAD standardises two detectors with the LOCO validation split, which we never touch, and its offline rule generation consumes normal images the shot count omits (Appendix G). The second group cannot explain the first, since WinCLIP+ and PatchCore use neither. We release ours in full so that it can be checked rather than trusted.

7 Limitations

One benchmark. MVTec LOCO is, to our knowledge, the only industrial benchmark that labels logical and structural anomalies separately, so we cannot test external validity and our conclusions should be read as being about this benchmark.

Five categories and three draws. The between-category standard error is several times the draw-to-draw deviation (Section 4), and a deviation over three draws is itself noisy. Two methods are not a family either, although the asymmetry surviving the removal of the vision–language component narrows what it can be attributed to. The intervals describe a fixed benchmark, not a population of categories or methods.

Our composition branch is weak. It reaches 65.7 ± 1.0 logical alone on sixteen exemplars, against 78.2 for LogSAD's on four plus the offline budget of Appendix G. The bottleneck is the open-vocabulary detector, whose counts are misaligned with what a specification declares: one portion of banana chips is detected as 17.8 ± 4.1 regions, two cable connectors as 22.1 ± 2.5, and fifteen pushpins, one per compartment, as exactly 1.00 ± 0.00 in every normal image. We therefore treat the branch as a probe, not a method.

We could not localise the effect to an anomaly subtype. Our preferred explanation — that references are blind specifically to quantity anomalies, whose every patch is on the normal manifold — makes a testable prediction, and two label-free tests failed to support it: the mixture structure in the per-image exemplar response is not specific to the logical subset, and exemplars help the count-deviant half of that subset *more*, not less. Appendix D reports both tests in full. We therefore state the result as a ratio rather than as blindness, and leave subtype attribution open.

The budget match counts images, not priors. The budgets match normal images, not everything a configuration is given: the composition branch also receives a hand-written part vocabulary and an external detector. That is intrinsic to the question, but the result is about what normal images cannot buy, not a contest between methods given equal information.

Fusion weights are fixed, not tuned. We fix $1/n$ a priori. This is a deliberate restriction that avoids a search, but it means the reported numbers are not the best achievable by fusion. Table 2 accordingly compares configurations within a controlled study, not systems on a leaderboard.

8 Conclusion

On MVTec LOCO, normal exemplars buy structural detection at about twice the rate they buy logical detection, a ratio that holds as the budget doubles and under an independent redraw. Past two branches, neither more diverse nor deeper references buy logical evidence we can resolve, while a compositional score nearly uncorrelated with them on logical anomalies does, at a matched budget and against the strongest reference-only control we could build. We supply the controls and budget accounting that the state of the art's single-system ablations lack — on one benchmark, five categories.

Ethics Statement

This work uses MVTec LOCO, a public benchmark of photographs of manufactured objects, under its published licence for non-commercial research; it involves no human subjects and no personally identifying data. The released artefact contains per-image scores and exemplar filename manifests derived from that benchmark, not the images themselves. Section 6 reports that published numbers for the same baseline disagree with one another. We report that as a property of current reporting conventions on this benchmark, we do not attribute error to any author, and every number we attribute to another paper is quoted from its printed table.

Reproducibility Statement

Section 3 specifies the metric, the branch configurations, the normalisation and the fusion rule. Table 1 reports the four external anchors used to validate the implementation. The supplementary archive contains per-image scores for every configuration reported in this paper (130,656 rows), the exemplar filename manifest for every branch, seed and category, the per-category unique budget of every configuration in Table 2, the scripts that regenerate every table and figure from those files, and SHA-256 checksums for all of the above. Every MVTec LOCO table, figure, interval and in-text AUROC is computed from those files by the included scripts; the exceptions, which need data outside the archive, are listed in Appendix J.

AI Use Statement

This section is required by the ICLR 2027 AI policy and does not count toward the page limit. We disclose the following uses of generative AI tools in the preparation of this work. *Used:* drafting and copy-editing the prose of this manuscript; writing and refactoring the analysis, plotting and consistency-checking scripts; searching for and formatting related work; and, in discussion with the authors, proposing candidate framings of the results and candidate control experiments, including the three-branch control of Section 5. *Not used:* generating or synthesising any data; running any experiment; translating text; cleaning or reformatting the dataset; and formulating or proving any mathematical claim. *Not applicable:* this work contains no theorems, no proofs and no generated data. All measurements reported here come from runs the authors specified and executed on real data; the decision criteria and stopping rules, including those of the pre-specified control in Appendix F, were fixed by the authors before the corresponding runs and were applied as written, including in the case where the outcome falsified the authors' own initial claim. Every reported number that the released files support (Appendix J lists the exceptions) was recomputed by the authors from the released score files using the released scripts, and was independently reproduced line for line by a second party in a clean environment. The authors have read and verified all AI-assisted text and code and take full responsibility for the content of this work, including all text, claims, figures and released artefacts.

References

P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. MVTec AD — a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, pp. 9592–9600, 2019.

P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130:947–969, 2022.

N. Cohen, I. Tzachor, and Y. Hoshen. Set features for anomaly detection. arXiv:2311.14773, 2023.

M. Fučka, V. Zavrtanik, and D. Skočaj. SALAD — semantics-aware logical anomaly detection. In *ICCV*, 2025.

M. Fučka, V. Zavrtanik, and D. Skočaj. ObjectCore — efficient few-shot logical anomaly detection using object representations. In *WACV*, pp. 3857–3867, 2026.

Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang. UniVAD: A training-free unified model for few-shot visual anomaly detection. In *CVPR*, 2025.

J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, pp. 19606–19616, 2023.

E. Jin, Q. Feng, Y. Mou, S. Decker, G. Lakemeyer, O. Simons, and J. Stegmaier. LogicAD: Explainable anomaly detection via VLM-based text feature extraction. In *AAAI*, 2025.

S. Kim, S. An, P. Chikontwe, M. Kang, E. Adeli, K. M. Pohl, and S. H. Park. Few shot part segmentation reveals compositional logic for industrial anomaly detection. In *AAAI*, volume 38, pp. 8591–8599, 2024.

X. Li, Z. Huang, F. Xue, and Y. Zhou. MuSc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *ICLR*, 2024.

X. Li, Z. Zhang, X. Tan, C. Chen, Y. Qu, Y. Xie, and L. Ma. PromptAD: Learning prompts with only normal samples for few-shot anomaly detection. In *CVPR*, pp. 16838–16848, 2024.

S. Patra and S. Ben Taieb. Revisiting deep feature reconstruction for logical and structural industrial anomaly detection. *Transactions on Machine Learning Research*, 2024.

Y. Peng, X. Lin, N. Ma, J. Du, C. Liu, C. Liu, and Q. Chen. SAM-LAD: Segment anything model meets zero-shot logic anomaly detection. *Knowledge-Based Systems*, 314:113176, 2025.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, pp. 14318–14328, 2022.

J. Zhang, G. Wang, Y. Jin, and D. Huang. Towards training-free anomaly detection with vision and language foundation models. In *CVPR*, pp. 15204–15213, 2025. Four-exemplar structural and logical AUROC of the individual detectors: Table 6; of the patch + composition combination: supplementary Table 8.

Y. Zhang, Y. Cao, X. Xu, and W. Shen. LogiCode: An LLM-driven framework for logical anomaly detection. *IEEE Transactions on Automation Science and Engineering*, 22:7712–7723, 2025.

Y. Zhang, J. Zhu, L. Fu, and G. Pang. AnomalyAgent: Training-free agentic models for zero-/few-shot anomaly detection. arXiv:2605.30140, 2026.

Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *ICLR*, 2024.

J. Zhu and G. Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *CVPR*, 2024.

Table 3: Composition-branch configurations under both normalisation rules (logical / structural / pooled). Configuration names are those used in the released artefact: `ref2` denotes the sixteen-image count reference, `batch` a reference taken from the test batch, `nocount` the branch with the count term removed, `nominal_band` a reference taken from declared counts, and `fallback_only` the structural fallback with no composition term.

	Inductive (used)			Transductive		
Configuration	Log.	Str.	Pool.	Log.	Str.	Pool.
`trainnormal_ref2` (main)	64.5	61.5	62.9	68.6	62.8	65.7
`composition_only`	67.8	56.3	62.2	68.6	56.5	62.7
`batch_ref2`	66.0	61.7	63.8	68.5	62.6	65.5
`trainnormal_nocount`	53.1	57.6	55.1	56.3	57.9	57.3
`nominal_band`	53.0	57.4	55.0	53.4	59.0	55.9
`fallback_only`	50.7	58.5	54.2	50.6	58.8	54.3

A Protocol anchors in detail

All four anchors in Table 1 use our own loader, our own subset definitions and our own AUROC implementation; only the exemplar-selection seed differs from the published setting, which is not reported by the sources. The three LOCO anchors are all drawn from the table of Zhang et al. (2026); they check that our pipeline reproduces one published convention, not that the convention is correct, and Section 6 explains why this distinction matters on this benchmark. The MVTec AD (Bergmann et al., 2019) anchor is included deliberately: it exercises the few-shot code path on a dataset with no logical subset, so a failure there would be attributable to the few-shot machinery rather than to the LOCO subset split. Our WinCLIP+ reproduction on MVTec AD at one exemplar is 93.2 ± 0.4 against a published 93.1.

B Inductive versus transductive normalisation

Table 3 reports every composition-branch configuration under both normalisations, all on draw 0 of the branch's references (Appendix H). Transductive normalisation uses the min and max of the test split, which is what an earlier version of this work did; inductive normalisation uses only held-out normal images. Batched zero-shot methods such as MuSc (Li et al., 2024a), which score unlabelled test images against one another, are transductive in the same sense, so their published numbers are not on the same footing as the inductive ones reported here. The main composition-branch configuration loses 2.8 pooled points under the inductive rule (first row of Table 3). All numbers in the main text are inductive.

C Composition-branch ablation

Every ablation here uses draw 0 of the branch's references. Removing the count term (`trainnormal_nocount`) drops logical AUROC from 64.5 to 53.1, close to the structural-fallback-only score of 50.7: essentially all of the branch's logical signal comes from part counts. Replacing the count reference with the counts declared in a natural-language specification (`nominal_band`) gives 53.0, which is the same as having no count term at all — the declared counts and the detector's counts are not on the same scale (Section 7).

Two observations deserve comment. First, `batch_ref2`, which takes the reference counts from the median of the test batch and is therefore transductive by construction, reaches 66.0 inductive-normalised, only 1.5 points above the main configuration; the 16 held-out normal images recover almost all of what the test batch would provide. Second, `composition_only` ($\lambda = 1$, no structural fallback) has a *higher* logical score alone (67.8 versus 64.5). We did not adopt it. $\lambda = 0.5$ was fixed a priori, before any of these numbers existed, so that the branch would carry some structural signal of its own rather than being a pure part counter, and we report the $\lambda = 1$ endpoint because the choice should be visible rather than because the data selected it. A transductive preview suggests substituting it would change fused logical by $+0.4$ and fused structural by -2.0, but the corresponding inductive run is not in the frozen experiment set, and we do not rest the choice of λ on a transductively

Table 4: Per-category logical / structural AUROC. Reference rows are the mean over three exemplar draws; fusion rows are the mean over the same assignments as Table 2. Category names are abbreviated from breakfast box, juice bottle, pushpins, screw bag and splicing connectors. Differences quoted in the text are computed before rounding and can therefore differ by 0.1 from differencing the printed entries.

	breakfast	juice	pushpins	screw	splicing
WinCLIP k=0	62.8/44.0	59.4/63.6	57.7/60.2	57.4/74.1	51.1/63.5
WinCLIP+ k=16	71.3/61.5	71.2/88.7	59.1/74.1	59.7/85.6	69.4/73.3
WinCLIP+ k=32	71.3/62.4	72.4/90.2	59.6/75.2	59.9/86.0	69.0/73.2
PatchCore k=16	78.4/69.7	80.7/84.2	56.7/67.3	53.6/68.4	60.8/57.7
PatchCore k=32	80.8/72.6	81.6/86.2	57.7/66.5	55.3/72.2	62.8/59.5
Composition, 3 draws	81.4/65.3	68.3/64.8	60.9/56.8	61.1/62.5	56.6/61.9
WC + PC	76.9/69.0	84.1/92.5	60.6/77.3	58.2/83.5	68.0/68.0
WC + PC + PC′	78.8/71.3	83.8/91.5	61.0/77.5	57.5/80.1	65.8/64.6
WC + WC′ + PC	75.0/66.6	83.5/92.9	60.4/77.0	59.0/86.3	70.9/71.2
WC + WC′ + WC″	71.6/61.8	71.7/89.1	59.2/74.2	59.6/85.7	69.9/73.7
PC + PC′ + PC″	79.7/70.2	81.3/85.3	58.2/69.4	54.8/70.5	61.2/57.9
WC_{32} + PC_{32}	78.4/72.7	86.0/94.3	59.6/77.8	59.2/86.1	69.2/69.0
WC + PC + Comp.	86.9/74.6	85.1/92.0	62.5/77.7	63.0/82.4	69.0/70.5
WC + PC + Comp., shared	88.4/75.2	85.5/92.3	63.8/76.9	62.7/83.0	70.6/70.9

normalised comparison when the rest of the paper argues against exactly that. The honest statement is that $\lambda = 0.5$ is an a priori design decision whose inductive alternative we have not measured.

D TWO LABEL-FREE TESTS OF THE SUBTYPE EXPLANATION

Section 7 reports that both tests failed to support our preferred mechanism. This section gives them in full.

Test 1: is the exemplar response bimodal on the logical subset only? If references were blind to a specific kind of logical anomaly, the per-image change in normalised score from $k = 0$ to $k = 4$ should separate into a responsive and an unresponsive mode on the logical subset and not on the structural one. A two-component Gaussian mixture does fit the logical subset better than a single Gaussian (ΔBIC $= 248$), but it fits the structural subset at least as well (ΔBIC $= 275$), so the structure is not specific to logical anomalies and does not support the prediction.

Test 2: do exemplars help count-deviant logical anomalies less? Splitting the logical subset at the median absolute deviation between the composition branch's detected part count and its reference count — a split that uses no anomaly-subtype label — the count-deviant half gains *more* from exemplars than the other half, not less: $+11.1$ against $+6.2$ in raw AUROC gain from $k = 0$ to $k = 4$. The two halves are disjoint sets of images, so this is a stratified two-sample bootstrap ($B = 3000$, images resampled within category, subset and half, single exemplar draw), not a paired one, and the contrast is formed in the order Δ = gain(non-deviant) $-$ gain(count-deviant) $= -4.35$, 95% CI $[-7.71, -1.26]$, $p = 0.010$; being a bootstrap mean on a single exemplar draw, Δ need not equal the difference of the two gains just quoted (-4.9); the negative sign is the count-deviant half gaining *more*, which is the opposite of what our explanation predicts. The direction survives headroom normalisation, 24.9% against 15.1%. Both tests use the composition branch's own counts, so they inherit its detector errors — OWLv2 miscounts these categories badly enough that the partition itself is noisy (Section 7) — and a cleaner version would split on the ground-truth defect-type annotations MVTec LOCO provides. We report them as failed tests of our explanation rather than as evidence against any alternative.

E PER-CATEGORY RESULTS

The per-category view shows how much the mean hides. Zero-shot WinCLIP scores 44.0 on breakfast-box structural anomalies — below chance — while reaching 62.8 on the logical subset of the same category; screw-bag is the reverse. The two reference methods also disagree strongly by category despite similar means: PatchCore is 7 points better than WinCLIP+ on breakfast-box logical and 15.6

points worse on splicing-connectors structural. This heterogeneity is one reason we report the split rather than a single pooled number.

The fusion rows address the question raised in Section 5: whether the three-branch gain is carried by breakfast box, the one category where the composition branch alone is strong. With the composition branch over its three draws, the per-category logical differences against WC + PC are $+10.1$, $+1.0$, $+1.9$, $+4.9$ and $+1.0$ in the order of the table, mean $+3.74$, positive in all five; removing each category in turn leaves the four-category mean between $+2.17$ and $+4.44$. Against WC + WC′ + PC they are $+12.0$, $+1.6$, $+2.1$, $+4.0$ and -1.9, mean $+3.56$, positive in four of five, with the four-category mean falling to $+1.45$ without breakfast box and rising to $+4.92$ without splicing connectors; against WC_{32} + PC_{32} they are $+8.6$, -0.9, $+2.8$, $+3.8$ and -0.2, positive in three of five (Appendix I). Five positives from five is an exact one-sided sign test at $p = 0.031$; four from five is $p = 0.19$, and with five categories no outcome short of five positives can reach $p < 0.05$, so this test carries little weight either way. The honest reading is that the gain is real in the image-level bootstrap, concentrated in the two categories with the most explicit part structure, and negative in two categories, by less than a point, against the control the argument rests on.

F Pre-specified control

The control that falsified our initial two-branch claim was specified in writing, with its decision rule, before it was run. We did not deposit it in a public registry, so it is pre-specified rather than pre-registered in the formal sense; we describe it in full here so that the rule can be checked against the outcome.

Question. Our two-branch result fused one reference branch with the composition branch. Does fusing *two reference branches* instead, under the identical calibration, weighting and evaluation rule, reach the same logical AUROC?

Decision rule, fixed before the run. If the two-reference fusion left logical AUROC near the single-branch level, the two-branch claim would stand, because the gain could then be attributed to the composition branch. If it clearly exceeded that level — we wrote down 68 as the threshold — the claim would be withdrawn, because a two-way fusion number could no longer be attributed to composition. Either outcome was to be reported; no weight, threshold or model was to be adjusted in response.

Outcome. The two-reference fusion reached 69.6 logical, above the threshold. The two-branch candidates it was compared against reached 70.4 (WC + Comp.) and 70.8 (PC + Comp.) under the identical rule, so the reference-only control came within 0.8 and 1.2 points of them; that closeness, not a failure of the candidates, is what made the two-branch attribution untenable. The two-branch claim was withdrawn and is not defended anywhere in this paper. The three-branch controls of Section 5 were designed afterwards to ask what remains once reference diversity is exhausted, and the result reported there is the answer to that second question, not a rescue of the first.

G Relation to the state of the art, in detail

LogSAD (Zhang et al., 2025a) reports 86.3 on MVTec LOCO with four normal exemplars and 90.2 with the full normal set, well above every configuration in Table 2. We are not competitive and do not claim to be.

We nevertheless read LogSAD as independent support. Its ablation (its Table 6, and supplementary Table 8 for the pair) under the four-exemplar protocol separates a patch matching branch from a composition matching branch and reports, in structural/logical order, $87.3/67.6$ for patch alone, $58.0/78.2$ for composition alone, and $85.8/82.0$ for both. Adding composition to patch matching gains 14.4 logical points and loses 1.5 structural points — the same direction and the same subset specificity we measure. Its patch baseline sits slightly below our two-branch reference baseline on logical anomalies (67.6 against 69.6) and its fused logical score about nine points above ours (82.0 against 73.3). That ablation is a single uncontrolled comparison within one system: it does not test whether a second patch branch would do as well, vary the exemplar budget, report variability, or match budgets across arms — exactly the controls we supply — and its fusion rule, a maximum over calibrated scores rather than a fixed $1/n$ average, puts the two effect sizes on different scales. UniVAD

(Gu et al., 2025) gives a second single-system ablation, at one exemplar, in logical/structural order: $59.9/82.2$ for its component-aware patch matching alone, $65.0/56.9$ for its graph over components alone, and $65.9/82.6$ for both. Adding the component graph to patch matching gains 0.0 logical points and 0.4 structural ones, again the direction and the subset specificity we measure, with the same missing controls. The three operating points, $+3.7$ here, $+6.0$ for UniVAD and $+14.4$ for LogSAD, do not order by the strength of the composition branch alone: UniVAD's is no stronger than ours (65.0 against 65.7 logical) and adds more, from a much weaker patch baseline (59.9 against 69.6). We therefore claim no trend in effect size, only a shared direction. ObjectCore (Fučka et al., 2026) reaches 80.8 at four exemplars with a finetuned open-vocabulary detector and bipartite object matching — the same broad recipe as our composition branch, executed far more carefully, and one more reason to read ours as a probe. It also sharpens what we claim. With only four exemplars it shows that exemplars *can* buy logical detection, so what saturates is not the exemplars but the reading of them: appearance distance to the nearest normal feature, the property defining the family we measure, extracts almost no further logical evidence past two branches at any budget. ObjectCore reads the same exemplars as objects to be matched — the change of scoring rule our three-branch result isolates in much weaker form.

LogSAD's four-exemplar configuration additionally samples three canonical normal images per category to generate its matching rules offline, and calibrates two of its three detectors on anomaly-free images from the LOCO validation split, so the unique normal images it consumes exceed four by an amount that depends on how one counts that split. The step is clearly described there, so this is not an error; it is the accounting any budget-matched comparison against it needs, and the kind this study is built to make explicit.

H THE COMPOSITION BRANCH OVER THREE DRAWS

The composition branch counts parts against the statistics of 16 normal exemplars of its own. We report it over three draws of those exemplars. Draw 0 is the one used throughout the study; draws 1 and 2 are new and are disjoint from the 16 calibration images, from draw 0, and from the reference branches' exemplars of the same seed, so each three-branch assignment consumes exactly 32 images per category. Draw 0 was not drawn under that last constraint and happens to share 2 and 4 images with the reference branches on breakfast box and juice bottle, so its assignment consumes 30.8 on average; the configuration's budget is therefore reported as 31.6, never above the 32 of the control it is compared with. In the *shared* configuration the branch uses the reference branches' own 16 images at every seed, and the three-branch budget is 16 everywhere.

Alone, the branch reaches 64.5, 65.9 and 66.5 logical over the three independent draws (61.5, 62.8 and 62.5 structural), and 67.6, 67.4 and 68.0 logical when it shares the reference images. Before draws 1 and 2 existed we asked how large its draw-to-draw variation would have to be for the three-branch conclusion to fail. With a bootstrap standard error of 0.81 on the single-draw logical difference of $+3.42$ against WC + PC, the 95% interval reaches zero once an independent component of 1.55 is added on the fused score; since the branch enters with weight $1/3$, and treating AUROC as locally linear in the fused score, its own logical AUROC would have had to vary by about 4.6 across draws. The measured deviation is 1.0. That bound was set against WC + PC, the only control that existed at the time. The same first-order argument applied to the primary contrast of Section 5, $+2.85$ with an implied bootstrap standard error of 0.98, gives about 3.2; it credits nothing to the averaging over three draws, which would raise it, and the measured 1.0 is under a third of it.

The 80 calibration images were re-scored against every reference set in this run, including draw 0. For that draw 26 of the 80 re-scored values differ from the ones every earlier number used, by at most 0.10, with per-category means shifting by at most 0.013; the test-image scores of draw 0 are identical to those used before. We keep the original calibration scores for draw 0, so that it is byte-identical to the branch reported elsewhere, and the released scripts report what changes otherwise: the fused logical AUROC of that assignment moves by 0.04 and the primary contrast of Section 5 by 0.01.

I REFERENCE DEPTH, REFERENCE DIVERSITY, AND THE SHAPE OF THE RATIO

Two controls that Section 5 and Section 4 rely on are reported here in full.

Table 5: Unified-memory control, inductively standardised. The $k = 32$ rows use the calibration-disjoint redraw. Mean $\pm$ sample s.d. over three exemplar assignments. The paired bootstraps for the last two contrasts in the text are in the released `run_stats.py`.

Configuration	Exemplars	Logical	Structural	Pooled
WinCLIP+ 32 alone, redrawn	32.0	66.5 ± 0.3	77.6 ± 0.2	70.9 ± 0.1
PatchCore 32 alone, redrawn	32.0	66.5 ± 0.6	71.8 ± 0.3	68.5 ± 0.4
WC + WC′ + PC	31.5	69.8 ± 0.5	78.8 ± 0.3	73.3 ± 0.4
WC_{32} + PC_{32}	32.0	70.5 ± 0.3	80.0 ± 0.3	74.3 ± 0.2
WC + PC + Comp.	31.6	73.3 ± 1.2	79.5 ± 0.6	75.6 ± 0.9
WC + PC + Comp., shared	16.0	74.2 ± 0.9	79.7 ± 0.5	76.2 ± 0.5

A unified memory at the same budget. Every multi-branch configuration in the upper part of Table 2 spends its budget on branches that draw their exemplars independently and are averaged after standardisation. The standard way to scale a nearest-neighbour memory is instead to pool more exemplars into one bank. The original $k = 32$ draw cannot be fused under our inductive rule, because it overlaps the calibration set and no calibration scores exist for its memory banks, so we drew a second set of 32 exemplars per category and seed from the training normals minus the 16 calibration images, used the same set for WinCLIP+ and PatchCore at each seed as every other configuration does, and scored the calibration images against those banks. Table 5 reports the result under exactly the rule of Section 3.

Three things can be read off it. *The redraw is clean*: its two single branches land within 1.1 points of the original $k = 32$ rows of Table 2, and the calibration images score like test normals under every redrawn memory bank (the AUROC of one against the other averages 50.6 for WinCLIP+ and 49.5 for PatchCore across the fifteen category–draw cells), which rules out a calibration image having entered a bank. *Depth buys little on logical anomalies*: WC_{32} + PC_{32} is $+0.73$ over WC + WC′ + PC, 95% CI $[-0.51, +1.95]$, $p = 0.26$, and $+1.16$, $[+0.01, +2.28]$, $p = 0.048$ on structural. *The composition branch survives the strongest control*: over its three draws, $+2.85$, $[+0.97, +4.79]$, $p = 0.003$ logical and -0.50, $[-2.28, +1.22]$, $p = 0.57$ structural, and at half the budget with shared references $+3.74$, $[+1.88, +5.66]$ logical. Per category the three-draw gain against this control is $+8.6$, -0.9, $+2.8$, $+3.8$ and -0.2 in the order of Table 4, so it is carried by three categories of five, most of all by breakfast box.

How large a $k = 32$ draw effect is. Before the redraw existed we compared the same configurations under the transductive rule of Appendix B, which needs no calibration scores, and the original draw gave WC_{32} + PC_{32} 71.6. The released script separates the two things that changed: on the redraw, the transductive rule gives 70.8 and the inductive one 70.5, so the normalisation accounts for 0.3 points and the choice of exemplars for 0.8. That 0.8 between two three-draw means at the same budget is a direct measurement of how much a $k = 32$ exemplar choice can move this configuration, and it is the reason we report the inductive, calibration-disjoint number rather than the higher transductive one.

Incremental and cumulative headroom. The ratios in Section 4 are cumulative from each method's anchor, and every headroom percentage is computed per exemplar draw and then averaged over the draws. Because by $k = 16$ more than nine tenths of the absorption has happened for WinCLIP+ and more than four fifths for PatchCore, the cumulative ratio at $k = 32$ is arithmetically dominated by the first interval, and the stability we report is the persistence of the asymmetry into the saturated regime rather than a constant marginal rate. Table 6 gives each interval on its own, with a bootstrap interval on the ratio so that the noise is measured rather than asserted. Two things follow. Every cumulative ratio excludes 1, with $P(> 1) = 1.000$ to three decimals, and so do WinCLIP+'s two early incremental ratios ($P \geq 0.999$) and PatchCore's $4 \to 16$ one (0.998); PatchCore's first increment, $1 \to 4$, does not ($[0.91, 2.18]$, $P = 0.935$). The asymmetry is therefore present in the intervals that carry most of the absorption, but not resolvably in every one of them. The $16 \to 32$ intervals carry almost none of it — 0.9 and 4.7 per cent of the logical headroom — and their ratios are correspondingly unmeasurable, with intervals of $[-25.6, 39.6]$ and $[0.36, 4.91]$; we print them so that the reader can see they are uninformative rather than supporting. Within WinCLIP+ the two measurable incremental ratios, 2.88 and 1.54, have barely overlapping intervals, which is weak evidence that the marginal ratio falls as the budget grows; we do not claim it.

Table 6: Headroom absorbed per interval, with a bootstrap interval on the ratio ($B = 2000$, images resampled within category and subset, one resample shared by the three exemplar draws). Incremental rows measure each interval on its own; cumulative rows measure from the method's anchor. $1 \to 16$ is the only interval the two methods share. The last two rows repeat the $k = 32$ endpoint on the calibration-disjoint redraw of Appendix I, an independent set of 32 exemplars per category.

Method	Interval	Logical %	Struct. %	Ratio	95% CI	$P(>1)$
WinCLIP+	$0 \to 4$ incr.	9.2	26.4	2.88	$[1.96, 5.38]$	1.000
WinCLIP+	$4 \to 16$ incr.	11.9	18.4	1.54	$[1.20, 2.05]$	0.999
WinCLIP+	$16 \to 32$ incr.	0.9	3.1	3.58	$[-25.58, 39.58]$	0.757
WinCLIP+	$0 \to 16$ cum.	20.0	40.0	2.00	$[1.64, 2.53]$	1.000
WinCLIP+	$0 \to 32$ cum.	20.7	41.9	2.02	$[1.67, 2.51]$	1.000
WinCLIP+	$1 \to 16$ cum.	16.9	28.4	1.68	$[1.36, 2.15]$	1.000
PatchCore	$1 \to 4$ incr.	11.7	15.8	1.35	$[0.91, 2.18]$	0.935
PatchCore	$4 \to 16$ incr.	7.7	19.7	2.55	$[1.56, 6.55]$	0.998
PatchCore	$16 \to 32$ incr.	4.7	6.4	1.37	$[0.36, 4.91]$	0.754
PatchCore	$1 \to 16$ cum.	18.5	32.4	1.75	$[1.36, 2.35]$	1.000
PatchCore	$1 \to 32$ cum.	22.3	36.8	1.65	$[1.31, 2.13]$	1.000
WinCLIP+	$0 \to 32$ cum., redrawn	21.0	42.4	2.02	$[1.69, 2.51]$	1.000
PatchCore	$1 \to 32$ cum., redrawn	19.6	37.7	1.93	$[1.50, 2.60]$	1.000

Over the one interval the two methods share, $1 \to 16$, they agree more closely than the headline numbers suggest: 1.68 $[1.36, 2.15]$ for WinCLIP+ against 1.75 $[1.36, 2.35]$ for PatchCore, with almost coincident intervals. This is the cleanest cross-method statement in the paper, because it holds the interval fixed as well as the protocol, and we report it alongside the anchored ratios rather than in place of them.

J RELEASED ARTEFACT

The supplementary archive contains: per-image scores for every configuration reported here (130,656 rows across fifteen files, each with category, subset, image name and score: 128,576 of them are 82 scorer configurations over the 1568 test images, and the remaining 2,080 are 26 scorer settings over the 80 held-out calibration normals); the exemplar filename manifests for every branch, budget and draw, including the held-out calibration set and the calibration-disjoint $k = 32$ redraw; the per-category unique-exemplar budget of every configuration in Table 2; the scripts that regenerate every table and figure from those files, together with their expected output; a protocol description; and SHA-256 checksums for all files. Every MVTec LOCO number in the tables, figures and intervals, and the in-text quantities of Sections 3 and 4 and Appendix I, is recomputed from those files by the included scripts. Four kinds of number are not: the MVTec AD anchor of Table 1, whose per-image scores are not in the archive; the part-count statistics of Section 7 and the count-deviation split and mixture fits of Appendix D, which need the composition branch's per-part detections rather than its image scores; two historical quantities fixed before the frozen experiment set, the transductive $\lambda = 1$ preview of Appendix C and the pre-draw sensitivity bound of Appendix H; and the published numbers of other papers, which are quoted from their printed tables.