SIMILARITY IS NOT NORMALITY: REASONING OVER RETRIEVED EVIDENCE FOR ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-based anomaly detection retains normal exemplars at inference, yet most existing methods fuse retrieval discrepancies without preserving which descriptor produced each of them. Similarity alone does not explain normality. A cracked but visually plausible component may retrieve close appearance neighbors while violating part layout, whereas a normal object observed from an unseen viewpoint may exhibit the opposite pattern. The key question is not only how large a discrepancy is, but also which property causes it and whether the remaining properties agree. Once an aggregation removes that distinction, a predictor observing only the aggregation cannot recover it. We refer to this limitation as the retrieval-to-reasoning gap and show that it is measurable, since permutation-invariant summaries of retrieved evidence are markedly less separable than the evidence itself on real data. To address this, we propose **R2AD** that retrieves from four factorized partitions of a frozen prototype bank corresponding to appearance, geometry, structure, and semantics, selecting the partition of the category predicted from the class token without using ground-truth category labels. It calibrates each discrepancy against an empirical distribution of normal costs, placing all factors on a common rank scale while preserving their identities. A conflict-aware mixture of experts then reasons over the resulting evidence spectrum, retaining isolated violations while attenuating the coherent elevations that benign variation tends to produce. R2AD reaches 75.1 pixel AP and 96.6 AUPRO on MVTec-AD and 46.5 pixel AP and 94.1 AUPRO on VisA, and ablation studies show the contribution of the evidence operators, the category conditioning, and the rank calibration.

1 INTRODUCTION

Unsupervised anomaly detection aims to localize defects using only anomaly-free training data Cao et al. (2024a); Liu et al. (2024); Lin et al. (2025). The central challenge is to distinguish subtle defects from benign variations in pose, illumination, texture, and object configuration, which can produce deviations of comparable magnitude.

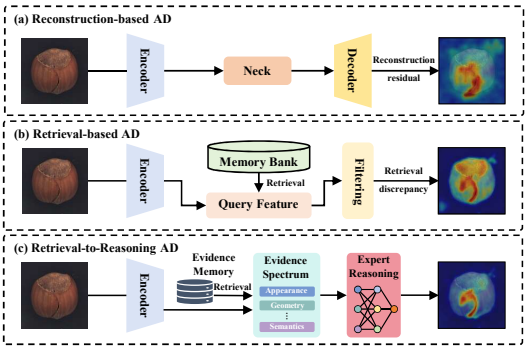

Figure 1: Comparison of reconstruction-based, retrieval-based, and retrieval-to-reasoning anomaly detection. Reconstruction-based methods identify anomalies from reconstruction residuals, while conventional retrieval-based methods directly convert nearest neighbor discrepancies into anomaly scores. R2AD instead interprets retrieved normal references as factor-specific evidence and reasons over their configuration before producing the final anomaly map.

Two widely used paradigms are reconstruction and retrieval. Reconstruction-based methods (Fig. 2a) detect anomalies through residuals between inputs and reconstructed images or features Zavrtanik et al. (2021); He et al. (2024b;a); Guo et al. (2025); Luo et al. (2025). However, expressive models may also reconstruct defects, while restrictive models may suppress normal details. Retrieval and embedding-based methods (Fig. 2b) compare query features with normal references Cohen & Hoshen (2020); Roth et al. (2022); Damm et al. (2025); Qu et al. (2025); Zhang et al. (2025); Cai

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.

et al. (2026). Despite advances in memory construction, feature correspondence, and discrepancy filtering, fusing retrieval responses can obscure which descriptor produced each discrepancy.

This distinction matters because similarity alone does not determine normality. A defective object may resemble normal examples in appearance yet violate structural consistency. Conversely, an unfamiliar viewpoint may alter a normal object's appearance while leaving other properties consistent. Interpreting a discrepancy therefore requires knowing its magnitude, its source, and its relation to the remaining evidence. Once aggregation discards source identity, a predictor receiving only the aggregate cannot recover it. We call this limitation the **retrieval-to-reasoning gap**. Sec. 2 formalizes this information loss and measures its effect on separability in real data.

We propose **R2AD** to address this gap by separating evidence acquisition from evidence reasoning (Fig. 1c). Using a frozen DINOv2 encoder Oquab et al. (2024), we construct a factorized prototype bank for appearance, geometry, structure, and semantics. Each descriptor independently retrieves normal references from the category predicted by the class token, without ground-truth category labels at inference. We then calibrate each discrepancy against its corresponding category-specific distribution of normal costs. This maps heterogeneous responses to a common rank scale while retaining source identity, yielding an **evidence spectrum**. Unlike prompt-based zero- and few-shot detectors that use vision-language priors Zhou et al. (2024); Cao et al. (2024b); Lv et al. (2025), R2AD requires no text prompts. It also requires no real anomaly samples.

A conflict-aware mixture of experts interprets the evidence spectrum. The router combines specialized experts using local evidence and image-wide context. An evidence-preserving path retains isolated violations, while a concordance-aware gate attenuates coherent increases across sources that benign variation tends to produce. Together, these components allow one response to change how another is interpreted. Sec. 4 evaluates R2AD on four industrial benchmarks in the multi-class setting, with additional few-shot experiments and controlled ablations.

Contributions. (i) We formalize the retrieval-to-reasoning gap and show how discarding source identity can reduce anomaly separability. (ii) We propose calibrated evidence spectrum retrieval, combining a factorized, category-conditioned prototype bank with empirical calibration that aligns discrepancies while preserving factor identity. (iii) We introduce conflict-aware evidence-specialized reasoning through source-aware routing, evidence preservation, and concordance-aware attenuation. (iv) We conduct controlled studies to isolate the contributions of descriptor design, category conditioning, scale alignment, factor identity, and configuration-dependent decision rules.

2 Why Retrieval Scores Are Not Enough

A large retrieval discrepancy need not indicate a defect: appearance variation may be benign, while a change in component layout may be anomalous. Its interpretation depends on which descriptor produced it and how other descriptors respond. We distinguish two limitations of score fusion: information lost in the representation and interactions excluded by the decision rule.

Retrieval-only anomaly detection. Let $\phi_p(\mathbf{I})$ be the descriptor from a frozen encoder Φ at patch p. A retrieval detector stores normal descriptors in a bank $\mathcal{M}$. With $\mathcal{N}_k(p)$ denoting the k nearest entries under cosine similarity, a common retrieval cost is

$$C_p(\mathbf{I}) = 1 - \frac{1}{k} \sum_{\mathbf{x} \in \mathcal{N}_k(p)} \cos\big(\phi_p(\mathbf{I}), \mathbf{x}\big). \tag{1}$$

Multiple descriptors provide various views of normality, but their usefulness depends on how their responses are combined. CostFilter (Zhang et al., 2025) and RAID (Cai et al., 2026) use feature guidance to filter matching costs while preserving spatial structure. We examine a problem: *whether fusion preserves source identity and allows one response to change the interpretation of another.*

Normality evidence. We describe responses by its source and its deviation from normal references.

Definition 1 (Normality evidence)**.** Let Ω denote the query image's patch index set and $\mathcal{S} = \{s_1, \ldots, s_K\}$ its descriptor sources. For source s, let $e_p^s \geq 0$ quantify the retrieval discrepancy at patch p relative to normal responses from that source. The ordered vector $\mathbf{e}_p = [e_p^1, \ldots, e_p^K]$ is the patch's normality evidence, and $\mathbf{E} = \{\mathbf{e}_p\}_{p \in \Omega}$ is its evidence spectrum.

Write a patch decision as $\hat{Y}_p = f(T(\mathbf{e}_p))$, where T determines the retained representation and f interprets it. A rule is **source-agnostic** if it depends only on permutation-invariant statistics, and **monotone** if increasing any response cannot change an anomaly decision into a normal one. Here, **factor** denotes a descriptor source, without assuming independent generative factors.

Two limits of fusion. The following constructions isolate these restrictions. They use an abstract vector $\mathbf{x} \in \mathbb{R}^K$, allowing signed coordinates in the Gaussian example. Risk denotes misclassification probability; $Q(t) = \Pr[Z > t]$ for $Z \sim \mathcal{N}(0,1)$, and Id is the identity matrix.

Proposition 1 (Information loss under source removal)**.** *Let $Y \in \{0,1\}$, with $Y = 1$ indicating an anomaly and $\Pr[Y = 1] = \frac{1}{2}$. For $K = 2$, let $\mathbf{x} \mid Y = 0 \sim \mathcal{N}((h,l)^\top, \sigma^2\mathrm{Id})$ and $\mathbf{x} \mid Y = 1 \sim \mathcal{N}((l,h)^\top, \sigma^2\mathrm{Id})$, where $h > l$ and $\sigma > 0$. For any measurable permutation-invariant $g : \mathbb{R}^2 \to \mathbb{R}^m$, $m \geq 1$, the conditional distributions of $g(\mathbf{x})$ are identical. Every classifier observing only this statistic has risk $\frac{1}{2}$, while the full vector has Bayes risk $Q\big(\frac{h-l}{\sigma\sqrt{2}}\big)$.*

The classes differ only in which source responds more strongly. The mean, maximum, and sorted coordinates erase this distinction, even as the ordered vector's Bayes risk approaches zero with increasing $(h - l)/\sigma$. Thus, *additional classifier capacity cannot compensate for source identity lost by aggregation*. Sorting is a particularly informative control: it retains every value and the input dimension, isolating the effect of source assignment. The limitation concerns the intermediate representation, not scalar prediction itself.

Proposition 2 (Decision error under monotone fusion)**.** *Let $Y \in \{0,1\}$ and let $\mathbf{x}$ be uniform on $\{l,h\}^2$ with $l < h$. Assign $Y = 1$ to $\{(h,l),(l,h)\}$ and $Y = 0$ to $\{(l,l),(h,h)\}$. Every coordinatewise monotone classifier has risk at least $\frac{1}{4}$, while a non-monotone rule attains zero risk.*

Here, the full vector remains available, but increasing the second response changes (h,l) from anomalous to (h,h), which is normal. Consequently, *preserving source identity is insufficient if the decision rule excludes such reversals*. Coordinatewise nondecreasing calibration followed by monotone fusion remains monotone, so scale alignment alone cannot remove this restriction. This motivates comparing decision rules on identical evidence, separately from testing source removal.

Together, the propositions motivate preserving source assignments and allowing non-monotone interactions. They establish failure cases for restricted representations and rules, without assuming these distributions describe real data or prescribing a particular architecture. Proofs are in App. A.

Testing the two limitations. We first compare probes mapping the ordered evidence vector, its mean, maximum, or sorted coordinates to patch-level anomaly scores against ground-truth masks. Ordered and sorted probes use identical architectures; scalar-control capacity and evaluation details are in App. B. No probe receives source-tagged auxiliary inputs. The ordered vector achieves 91.4 patch AUROC on MVTec-AD versus 88.7 for the strongest evaluated source-agnostic representation; the Real-IAD gap is 4.8 points (Fig. 2). These probes assess separability from at most four values per patch. The calibrated sources' mean absolute off-diagonal correlation is 0.20, indicating weak linear association without implying independence.

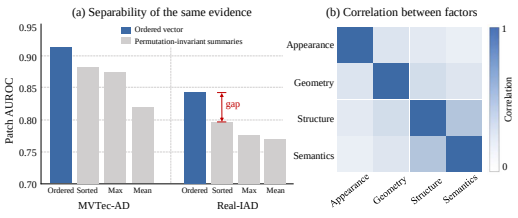

Figure 2: Retrieval evidence diagnostics. **Left:** patch AUROC for source-tagged evidence and the evaluated permutation-invariant representations, with error bars over repeated fits. **Right:** pairwise correlations between calibrated sources.

The complementary comparison holds the evidence fixed and evaluates conditional versus monotone fusion (Sec. 4.3). Conditional fusion improves P-AP by 1.4 points on MVTec-AD and 2.5 on Real-IAD. These comparisons assess the practical relevance of the two restrictions: which information is retained, and which decisions the fusion rule can express.

3 The Proposed R2AD

Sec. 2 shows how discarding source identity and restricting interactions between responses can impair detection. R2AD addresses these limitations through four stages: extracting descriptors of

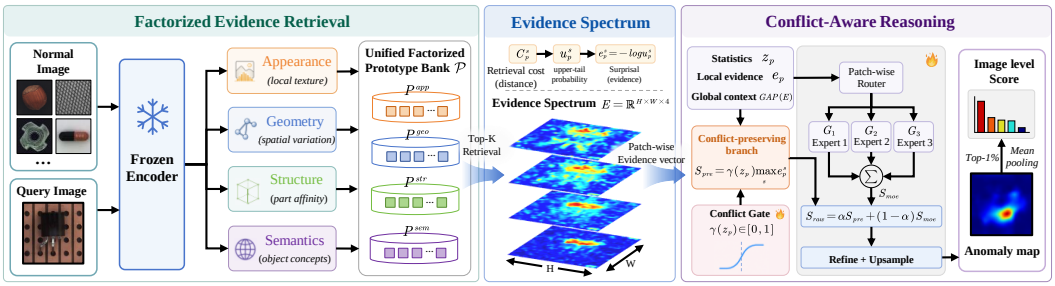

Figure 3: Overview of **R2AD**. Four descriptor views retrieve from separate normal-reference partitions of the category predicted from the class token. Empirical calibration places the four responses on a common rank scale while preserving which view produced each of them. A router combines expert predictions using the local evidence vector and global context, while a gated path carries the strongest response with a weight that depends on how the remaining responses are distributed. No ground-truth category label is used at inference.

complementary properties, retrieving references for the object category, calibrating discrepancies to a common scale, and interpreting their joint configuration. Fig. 3 summarizes the framework. We describe each stage below, followed by the self-supervised training procedure.

3.1 Factorized Evidence Operators

Patch appearance, local neighborhoods, region layout, and image context provide complementary cues for assessing normality. We capture them with four operators applied to a frozen DINOv2-B encoder with registers (Oquab et al., 2024; Darcet et al., 2024). Let $L = 12$ denote the final block, $\mathbf{f}_p^\ell$ the patch token at block ℓ, and $\mathbf{f}^{\text{cls}}$ the class token. All four views use a single encoder pass, drawing from blocks $\ell_1 = 8$, $\ell_2 = 10$, and $L = 12$. The operators encode distinct properties of a match; App. E.2 separates their contribution from the effect of depth alone.

Appearance. We use $\psi_p^{\text{app}} = \mathbf{f}_p^{\ell_1}$ as the patch-appearance descriptor. This intermediate representation retains the local texture and surface regularity needed for appearance matching.

Geometry. Similar patch tokens can occur in different surroundings, and a scratch may be more apparent as a break in local continuity. Let $\mathcal{V}(p)$ be the 3×3 neighborhood of p excluding p, let $\mathbf{\Delta}_{up}$ be the normalized offset from p to u, and define weights $\omega(\mathbf{\Delta}_{up}) \propto (\|\mathbf{\Delta}_{up}\|_2 + \epsilon_{\text{geo}})^{-1}$ that sum to one over u. We concatenate the token with a weighted neighborhood residual:

$$\psi_p^{\text{geo}} = \Big[\mathbf{f}_p^{\ell_2}; \sum_{u \in \mathcal{V}(p)} \omega(\mathbf{\Delta}_{up})(\mathbf{f}_u^{\ell_2} - \mathbf{f}_p^{\ell_2})\Big].$$

Retaining the center token lets retrieval judge whether a local deviation is normal for that texture. The radial weights introduce no signed directional preference.

Structure. A missing or displaced component may leave locally plausible patches while disrupting object layout, requiring comparisons beyond a fixed window. We cluster final-block tokens into $R = 50$ regions, each with pooled feature $\bar{\mathbf{f}}_r$, normalized area a_r, and centroid $\mathbf{c}_r$. Let $\rho(p)$ denote the region containing p, and let π_p order the remaining regions by the polar angle of their centroid offset from $\mathbf{c}_{\rho(p)}$, with distance and area breaking ties:

$$\psi_p^{\text{str}} = \Big[\bar{\mathbf{f}}_{\rho(p)}; a_{\rho(p)}; \big[a_{\pi_p(r)}; \mathbf{c}_{\pi_p(r)} - \mathbf{c}_{\rho(p)}\big]_{r=1}^{R-1}\Big].$$

Angular ordering removes dependence on arbitrary k-means cluster indices, making descriptors comparable across images while retaining relative pose. Centroid offsets also remove explicit dependence on absolute object position. This view has two limitations: patches within a region share a descriptor, providing coarse support rather than precise boundaries, and per-image clustering may merge a defect with surrounding normal content, weakening its structural response.

Semantics. A locally familiar pattern may still be inappropriate for its object. We use $\psi_p^{\text{sem}} = [\mathbf{f}_p^L; \mathbf{f}^{\text{cls}}]$ to compare patch appearance jointly with image context.

3.2 Category-Conditioned Reference Sets

A mixed-category bank can match a query patch to a different object, assigning a low cost because the patch looks familiar somewhere in the bank. Category conditioning instead evaluates whether it is normal for the object under inspection, an important distinction in the multi-class setting.

Bank construction. Let $\mathcal{B}^c$ contain M randomly selected normal images of category $c \in \mathcal{C}$. For each $\mathbf{I}_i \in \mathcal{B}^c$ and source s, we compute $\widetilde{\mathbf{f}}^s_{i,p} = \mathrm{Norm}(\psi^s_p(\mathbf{I}_i)\mathbf{T}^s)$, where $\mathbf{T}^s \in \mathbb{R}^{D_s \times d}$ is a fixed Gaussian projection and $d = 1024$. The projections map heterogeneous operator outputs to a common indexing dimension d, enabling a shared retrieval implementation and memory budget. Their purpose is dimensional alignment; Gaussian projections approximately preserve inner products between normalized vectors at this width.

Category selection. We select $\hat{c}$ by matching the query class token to the nearest category centroid computed from bank images. Retrieval and calibration use $\mathcal{P}^{s,\hat{c}}$ and the corresponding reference list, while the shared fusion module receives no category label. This assumes a known category set and separable class-token representations. These conditions hold for the evaluated inspection benchmarks but need not hold in open-set deployment. App. C.4 compares predicted-category retrieval with mixed-category and oracle-category references.

Reference exclusion. When computing reference costs for $\mathbf{I}_i \in \mathcal{B}^c$, we exclude all its source-s descriptors from the bank. Self-matches would otherwise bias reference costs downward. This exclusion does not remove near-duplicate patches from other images, so reference costs may remain optimistic relative to unseen normals, shifting their calibrated evidence upward. App. H.1 describes the within-image exclusion used with a single few-shot support image.

3.3 Calibration into an Evidence Spectrum

Raw cosine costs are not directly comparable across descriptor spaces: a value typical for one source may be unusual for another. Direct fusion must therefore learn these scale differences alongside anomaly discrimination, as examined in Sec. 4.3. We instead express each cost by its empirical rank among normal costs from the same source and category. This requires no parametric assumption and accommodates the different, non-Gaussian cost distributions observed across sources.

For source s and category c, let $C^{s,c}_p$ be the cost in Eq. 1, using $k = 5$ neighbors retrieved by FAISS IVF-Flat from $\mathcal{P}^{s,c}$; at inference, $c = \hat{c}$. Averaging neighbors reduces sensitivity to individual matches, particularly for structure descriptors affected by unstable region assignments. A normal patch from bank image $\mathbf{I}_i$ is queried against its image-excluded bank to obtain $C^{s,c,\mathrm{loo}}_{i,p}$. We sort $n = 50000$ such costs, retaining repeated values, into the reference list $(t^{s,c}_j)^n_{j=1}$ and define

$$u^{s,c}_p = \frac{1 + \sum_{j=1}^{n} \mathbf{1}[t^{s,c}_j \geq C^{s,c}_p]}{n+1}, \qquad e^s_p = -\log u^{s,c}_p. \tag{2}$$

Large costs yield small upper-tail ranks and high evidence values. The negative logarithm expands differences among small ranks that would otherwise be compressed near zero. The added count bounds e^s_p within $[0, \log(n+1)]$. A sorted lookup costs $O(\log n)$ and introduces no learned parameters. Resampling the four maps to a common grid and stacking them in a fixed channel order yields $\mathbf{E} \in \mathbb{R}^{H \times W \times K}$ with $K = 4$. The channel order preserves source identity.

This calibration has finite resolution: costs between adjacent references receive identical values, and costs above the largest reference saturate, losing their relative ordering. App. C.5 further discusses the finite resolution and saturation behavior of this empirical rank representation. We use the construction only for empirical normalization. Reference costs share banks and include spatially dependent patches, so we claim neither conformal coverage nor image-level error control. Each reference pool also mixes foreground and background positions; ranks are therefore relative to this mixture, not to position-matched normals.

3.4 Reasoning via a Conflict-Aware MoE

After calibration, the remaining task is to interpret the joint evidence. A fixed weighted sum assigns each source the same importance across patches and images, limiting the conditional behavior motivated in Sec. 2.

Configuration statistics. An isolated defect may produce a strong response in one source, while viewpoint or illumination changes may elevate several sources together. To distinguish these patterns, let $\bar{\mathbf{e}}_p$ be the ϵ_{norm}-smoothed normalization of $\mathbf{e}_p$ and Ent the Shannon entropy. We define

$$\mathbf{z}_p = \Big[\max_{s\in\mathcal{S}} e_p^s - \text{med}_{s\in\mathcal{S}}\, e_p^s;\, \frac{1}{K}\sum_{s\in\mathcal{S}} e_p^s;\, \text{Ent}(\bar{\mathbf{e}}_p) \Big], \tag{3}$$

where med is the lower median. These terms measure the prominence of the strongest response, overall magnitude, and uniformity across sources. Using the median in the first term prevents an isolated large coordinate from raising its own baseline. Together, the statistics distinguish uniformly weak, strong, and concentrated responses, while their association with anomalies is learned.

Context-dependent routing. Global context helps distinguish an isolated local violation from widespread elevation under unfamiliar imaging conditions. With GAP denoting global average pooling of $\mathbf{E}$ and $J = 3$ experts, the router computes

$$\mathbf{r}_p = \text{softmax}(\text{MLP}([\mathbf{e}_p; \mathbf{z}_p; \text{GAP}(\mathbf{E})])) \in \mathbb{R}^J. \tag{4}$$

Expert j is a learned map $\Psi_j : \mathbb{R}^{H\times W\times K} \to \mathbb{R}^{H\times W}$, and the mixture is $m_{\text{moe},p} = \sum_j r_{j,p} m_{j,p}$. The nonlinear experts can represent non-monotone interactions over the full evidence field; $\mathbf{r}_p$ adapts their contributions at each patch.

Gated preservation. A learned mixture may suppress a strong response confined to one source. To preserve such evidence, we add a direct path carrying the maximum response with a configuration-dependent weight:

$$m_{\text{pre},p} = \gamma(\mathbf{z}_p) \max_s e_p^s, \qquad \gamma(\mathbf{z}_p) = \text{sigm}(\mathbf{v}^\top \mathbf{z}_p + b). \tag{5}$$

The gate γ adjusts the contribution of the strongest response according to its agreement with the others. It is intended to preserve isolated peaks while attenuating peaks accompanied by coherent elevation across sources. It uses only symmetric statistics, leaving source-specific interactions to the router and experts. App. E.3 evaluates the contribution of the conflict gate through controlled ablations, with further discussion in App. F.

Combination. With a learned $\alpha \in (0, 1)$, we combine the two paths as $m_{\text{raw},p} = \alpha m_{\text{pre},p} + (1-\alpha) m_{\text{moe},p}$. Upsampling to the pixel grid and residual convolutional refinement yield $\mathbf{A} = \text{Refine}(\text{Upsample}(\mathbf{m}_{\text{raw}}))$. Refinement receives only the fused score map and cannot recover information discarded during fusion. Following RAID (Cai et al., 2026), the image score is the mean of the top 1% of responses. App. E.3 reports component ablations.

Self-Supervised training. Normal exemplars define the retrieval reference but provide no defect masks for training fusion. Following RAID (Cai et al., 2026), we generate synthetic image and mask pairs by blending textures through multi-scale Perlin masks. The encoder, bank, and calibration references remain frozen, so variants differing only in fusion use identical evidence. Only the router, experts, gate, mixing coefficient, and refinement blocks are trained. We minimize $\mathcal{L} = \mathcal{L}_{\text{focal}}(\mathbf{A}^{\text{syn}}, \mathbf{Y}^{\text{syn}}) + \lambda_{\text{bal}}\mathcal{L}_{\text{bal}}$ with $\lambda_{\text{bal}} = 0.005$. The focal term downweights easy background pixels under severe foreground imbalance, while the balance term discourages routing collapse onto a subset of experts. All controlled trainable variants receive the same pre-generated pairs in the same order. App. C.6 provides optimization details.

4 Experiments

4.1 Experimental Settings

Datasets and protocols. We evaluate multi-class UAD on MVTec-AD Bergmann et al. (2019), VisA Zou et al. (2022), MPDD Jezek et al. (2021), and Real-IAD Wang et al. (2024). MVTec-AD and VisA cover diverse objects and textures; MPDD emphasizes metallic parts and geometric variation; Real-IAD includes viewpoint, background, and intra-class changes. We also evaluate 1-, 2-, and 4-shot settings on MVTec-AD and VisA.

Baselines. The reconstruction-based baselines are UniAD You et al. (2022), HVQ-Former Lu et al. (2023), DiAD He et al. (2024b), MambaAD He et al. (2024a), Dinomaly Guo et al. (2025), INP-Former Luo et al. (2025), and InvAD Sakai et al. (2026). Retrieval-based baselines are PatchCore Roth et al. (2022), AnomalyDINO Damm et al. (2025), CostFilter-AD Zhang et al. (2025),

Table 1: Multi-class UAD results (%) on MVTec-AD and VisA. † denotes results reproduced with official code under our evaluator at 256^2 resolution. Best results are **bold**; second-best are underlined.

Methods ↓	MVTec-AD (15 classes)							VisA (12 classes)						
	Image-level			Pixel-level				Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
Reconstruction-based methods														
UniAD *NeurIPS'22*	96.5	98.8	96.2	96.8	43.4	49.5	90.7	85.5	85.5	84.4	95.9	21.0	27.0	75.6
HVQ-Former *NeurIPS'23*	98.0	99.5	97.5	97.3	48.2	53.3	91.4	93.2	92.8	87.6	98.7	35.0	39.6	86.3
DiAD *AAAI'24*	97.2	99.0	96.5	96.8	52.6	55.5	90.7	86.8	88.3	85.1	96.0	26.1	33.0	75.2
MambaAD *NeurIPS'24*	98.6	99.6	97.8	97.7	56.3	59.2	93.1	94.3	94.5	89.4	98.5	39.4	44.0	91.0
Dinomaly† *CVPR'25*	99.3	99.7	98.9	98.0	58.6	61.5	92.3	**96.9**	97.3	**93.7**	98.8	39.5	44.7	92.0
INP-Former† *CVPR'25*	99.4	99.8	98.9	98.1	59.4	61.6	93.0	**96.9**	**97.4**	93.3	98.8	39.1	43.7	91.6
InvAD *CVPR'26*	99.0	99.6	98.5	97.5	46.5	52.3	92.7	**96.9**	97.2	**93.7**	**99.1**	39.2	43.1	92.9
Retrieval-based methods														
PatchCore *CVPR'22*	96.4	–	–	95.7	–	–	–	–	–	–	–	–	–	–
AnomalyDINO *WACV'25*	96.8	98.6	97.1	98.1	61.3	60.8	93.6	90.5	91.4	86.2	97.5	39.6	40.4	86.3
CostFilter-AD *ICML'25*	99.0	99.7	98.6	98.0	58.1	61.2	93.2	93.4	95.2	89.3	98.6	41.4	45.0	86.8
RAID *CVPR'26*	99.4	99.8	98.7	98.6	71.7	68.5	95.6	94.9	95.5	90.5	99.0	45.2	49.2	91.7
RAD† *ICML'26*	**99.5**	99.8	98.8	98.2	69.0	66.5	94.0	95.1	95.6	91.9	98.3	45.2	49.9	88.5
R2AD (Ours)	**99.5**	**99.9**	**99.1**	**98.8**	**75.1**	**71.1**	**96.6**	95.8	96.1	92.3	**99.1**	**46.5**	**50.2**	**94.1**

Table 2: Multi-class UAD results (%) on MPDD and Real-IAD. † denotes reproduction with official code under our evaluator. Both benchmarks contain substantial geometric, viewpoint, and intra-class variation, so they stress configuration-dependent reasoning more heavily than MVTec-AD. Best in **bold**, second underlined.

Methods ↓	MPDD (6 classes)							Real-IAD (30 classes)						
	Image-level			Pixel-level				Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
Reconstruction-based methods														
UniAD *NeurIPS'22*	80.1	83.2	85.1	95.4	19.0	25.6	83.8	83.0	80.9	74.3	97.3	21.1	29.2	86.7
HVQ-Former *NeurIPS'23*	86.5	88.1	85.8	96.7	27.6	31.4	86.9	86.6	84.9	79.4	98.0	27.6	34.4	88.7
DiAD *AAAI'24*	85.8	89.2	86.5	91.4	15.3	19.2	66.1	75.6	66.4	69.9	88.0	2.9	7.1	58.1
MambaAD *NeurIPS'24*	89.2	93.1	90.3	97.7	33.5	38.6	92.8	86.3	84.6	77.0	98.5	33.0	38.7	90.5
Dinomaly† *CVPR'25*	94.4	96.8	94.5	98.5	**50.4**	**51.2**	95.0	89.0	86.5	79.8	98.6	27.8	35.3	90.0
INP-Former† *CVPR'25*	93.0	95.4	93.2	98.2	47.7	49.1	94.7	89.4	87.3	80.1	98.6	31.1	37.7	90.9
InvAD *CVPR'26*	96.5	96.5	94.4	98.3	40.1	43.9	91.2	–	–	–	–	–	–	–
Retrieval-based methods														
PatchCore *CVPR'22*	83.5	–	–	97.7	–	–	–	–	–	–	–	–	–	–
AnomalyDINO *WACV'25*	–	–	–	–	–	–	–	–	–	–	–	–	–	–
CostFilter-AD *ICML'25*	93.1	95.4	90.3	97.5	34.1	37.0	82.9	–	–	–	–	–	–	–
RAID† *CVPR'26*	96.3	97.6	95.0	**98.9**	47.0	46.9	**96.6**	84.1	84.2	80.9	97.0	31.0	37.4	83.2
RAD† *ICML'26*	94.2	95.2	93.2	98.4	46.4	48.9	94.3	89.4	87.6	81.7	98.1	38.0	42.4	90.6
R2AD (Ours)	**97.6**	**98.2**	**95.7**	**98.9**	47.3	48.2	**96.6**	**90.8**	**88.9**	**83.5**	**98.9**	**40.6**	**44.8**	**92.1**

RAID Cai et al. (2026), and RAD Zhang et al. (2026). AnomalyDINO and CostFilter-AD are particularly relevant: the former uses frozen DINOv2 features for patch retrieval, while the latter retains and filters multiple matching costs. R2AD and all reproduced baselines use 256×256 inputs. A † marks results reproduced with official code under our evaluator. We reproduce RAID on MPDD and Real-IAD because its published results use a different resolution. Entries marked "–" are unavailable.

Metrics and variance. We report image-level AUROC, AP, and F_1-max, and pixel-level AUROC, AP, F_1-max, and AUPRO. Pixel AP captures performance under foreground-background imbalance, while AUPRO measures region-level overlap.

Implementation. R2AD uses a frozen DINOv2-B encoder Darcet et al. (2024) with $K = 4$ factors, projection dimension $d = 1024$, $k = 5$ neighbors per factor, $J = 3$ experts, and $R = 50$ structural regions across all datasets. Only the router, experts, conflict gate, mixing coefficient, and refinement module are trained. Following RAID Cai et al. (2026), all trainable variants use identical synthetic image-mask pairs in the same order and the objective in Sec. 3.4. Training and calibration details are in Apps. C.5 and C.6.

4.2 Comparison to the State of the Art

As shown in Tab. 1 and 2, R2AD achieves the best or joint-best results on all localization metrics for MVTec-AD, VisA, and Real-IAD. On MPDD, it ties the best P-AUROC and AUPRO but trails Dinomaly in P-AP and P-F_1-max. Relative to the strongest retrieval baseline for each metric, R2AD improves P-AP and AUPRO by 3.4 and 1.0 points on MVTec-AD, and by 2.6 and 1.5 points on

Table 3: Few-shot results on MVTec-AD and VisA, reported as I-AUROC / P-AUROC.

Setup	Datasets	PatchCore	Win-CLIP	FastRecon	PromptAD	IIPAD	DFM	RAID	R2AD (Ours)
1-shot	MVTec-AD	86.3 / 93.3	92.6 / 91.6	83.7 / 93.9	93.0 / 95.2	94.2 / 96.4	87.2 / 95.2	95.1 / 96.6	**95.5 / 96.9**
	VisA	79.9 / 95.4	84.8 / 95.3	80.1 / 96.5	85.2 / 97.2	85.4 / 96.9	84.0 / 96.4	85.8 / 97.7	**86.5 / 98.0**
2-shot	MVTec-AD	83.4 / 92.0	93.8 / 91.9	88.9 / 95.3	95.4 / 95.6	95.7 / 96.7	90.2 / 96.0	96.6 / 97.1	**96.9 / 97.4**
	VisA	81.6 / 96.1	83.5 / 95.7	84.6 / 97.5	85.1 / 97.7	86.7 / 97.2	86.0 / 96.8	88.5 / 97.9	**89.0 / 98.2**
4-shot	MVTec-AD	88.8 / 94.3	95.5 / 92.4	94.2 / 95.9	95.9 / 96.0	96.1 / 97.0	92.0 / 96.2	96.9 / 96.9	**97.3 / 97.4**
	VisA	85.3 / 96.8	85.7 / 96.0	68.5 / 96.0	87.5 / 97.9	88.3 / 97.4	89.8 / 97.1	89.3 / 98.2	**90.0 / 98.5**

Table 4: Comparison of evidence operators, post-retrieval representations, and decision rules under identical bank images, and training schedules. Rows 1–2 test capacity on a collapsed cost; rows 1, 3, 4, and 8 test operator–reasoning interactions; rows 5–8 compare representations and decision rules using identical evidence.

	Operators, post-retrieval representation, and rule	MVTec-AD		Real-IAD	
		P-AP	AUPRO	P-AP	AUPRO
1	Scalar cost, no learning	69.8	94.2	35.7	89.6
2	Scalar cost + capacity-matched filter	71.1	95.1	31.8	84.5
3	Depth-only features + scalar cost	66.1	94.5	32.2	87.8
4	Depth-only features + full reasoning	68.7	95.4	34.6	89.5
5	Our operators + concatenated raw costs	72.9	95.8	36.4	88.7
6	Our operators + calibrated, identity removed	72.1	95.5	35.9	88.9
7	Our operators + calibrated, monotone fusion	73.7	96.0	38.1	90.7
8	**Our operators + evidence spectrum, conditional reasoning**	**75.1**	**96.6**	**40.6**	**92.1**

Real-IAD. Image-level performance is competitive, although reconstruction-based methods lead on VisA. The results thus show more consistent gains in localization than in image-level ranking.

Few-shot evaluation. Table 3 reports the 1-, 2-, and 4-shot results. For 2- and 4-shot evaluation, reference costs are constructed with leave-one-image-out retrieval over the available support images; for 1-shot, where image-level exclusion would leave an empty bank, we use leave-one-patch-out retrieval. R2AD achieves the best listed I-AUROC and P-AUROC in all six dataset–shot settings, reaching 95.5/96.9 on MVTec-AD and 86.5/98.0 on VisA in the 1-shot setting. Protocol details are provided in App. H.1.

4.3 Ablation Study

Tab. 4 compares evidence operators, post-retrieval representations, and decision rules under matched bank images, synthetic pairs, and training schedules. Row 1 applies the scalar retrieval cost of Eq. 1 to our operators without learning; row 8 is the full method. For more ablation results, please refer to the Appendix.

Model capacity. Adding a capacity-matched filter to the scalar cost (row 2) improves P-AP by 1.3 points on MVTec-AD but reduces it by 3.9 on Real-IAD. In contrast, rows 5–8 retain multiple responses and outperform this filter on both benchmarks under the same parameter budget and training schedule. These results support the importance of the retained representation, although they do not rule out sensitivity to optimization or tuning.

Evidence operators. Rows 3–4 replace our operators with plain tokens from the same encoder depths. Full reasoning improves the depth-only features from 66.1 to 68.7 P-AP on MVTec-AD and from 32.2 to 34.6 on Real-IAD, corresponding to gains of 2.6 and 2.4 points, respectively. In contrast, reasoning over our operators yields gains of 5.3 and 4.9 points. The resulting operator–reasoning interactions are therefore +2.7 P-AP on MVTec-AD and +2.5 on Real-IAD, indicating that conditional reasoning benefits more from structured factor-specific evidence than from generic multi-depth features.

Calibration. Using raw costs with the same reasoning module (row 5) reduces P-AP by 2.2 points on MVTec-AD and 4.2 on Real-IAD. This supports aligning descriptor scales before learning the decision rule.

Factor identity. Sorting the calibrated coordinates at each patch (row 6) retains every magnitude but removes its factor assignment. P-AP decreases by 3.0 points on MVTec-AD and 4.7 on Real-IAD.

This complements the diagnostic in Sec. 2 and supports preserving source identity, as motivated by Prop. 1.

Configuration dependence. Replacing the router and gate with monotone fusion over the same calibrated spectrum (row 7) reduces P-AP by 1.4 points on MVTec-AD and 2.5 on Real-IAD. The comparison supports learning interactions between responses, consistent with Prop. 2. This effect is smaller than those of calibration and factor identity. The effect is larger on Real-IAD (2.5 versus 1.4 P-AP), where the evidence configurations are more heterogeneous. Together, these comparisons support contributions from the operators, calibration, factor identity, and the decision rule.

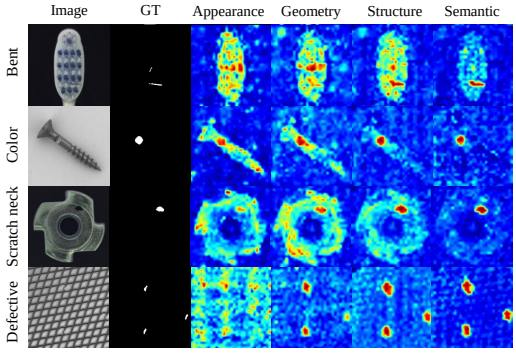

Figure 4: Per-factor evidence maps under the shared range of Eq. 2, with a common colorbar. Columns show inputs, masks, and the appearance, geometry, structure, and semantic evidence maps. The structure map is piecewise constant over regions by construction.

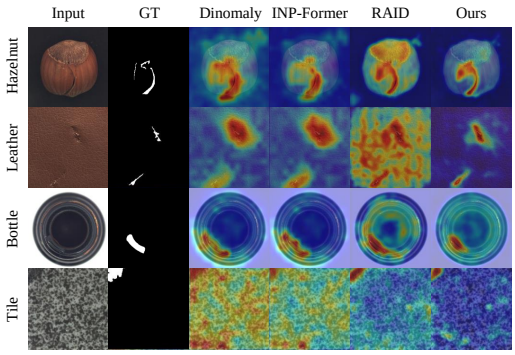

Figure 5: Qualitative localization comparison with representative reconstruction- and retrieval-based baselines.

4.4 Efficiency and Qualitative Comparison

R2AD requires 32.8 G FLOPs and 5.1 GB of bank memory, with a latency of 8.1 ms per image. Among retrieval methods with reported measurements, it has the lowest latency and bank memory but the highest parameter count. Fig. 4 visualizes calibrated evidence on a shared scale. The factors capture complementary properties, with the structure map showing the region-level plateaus implied by its construction. Fig. 5 illustrates more compact anomaly responses, sharper object boundaries, and fewer spurious background activations than representative reconstruction- and retrieval-based baselines.

Table 5: Efficiency at a unified 256×256 resolution. Inference time is measured with batch size 1 on one RTX 4090 and includes per-image clustering and retrieval; one-time bank construction is excluded.

Method	Params↓	FLOPs↓	Lat.(ms)↓	Mem.(GB)↓
Reconstruction-based methods				
DiAD	1331.3	451.5	15567.0	–
Dinomaly	132.8	34.6	134.0	–
INP-Former	139.8	32.3	126.1	–
InvAD	1355.9	476.9	61.5	–
Retrieval-based methods				
PatchCore	<u>68.9</u>	**9.2**	11.1	28.3
RAID	**36.0**	36.2	12.1	<u>6.5</u>
RAD	85.7	<u>30.6</u>	<u>10.8</u>	14.2
R2AD	129.2	32.8	**8.1**	**5.1**

5 Conclusion

We presented R2AD, a retrieval-to-reasoning framework for industrial anomaly detection built on a single premise, that retrieval should not terminate in a nearest-neighbor score. Retrieved normal references are evidence, and that evidence must retain factor identity and be interpreted before an anomaly decision is made. R2AD combines a factorized normality prototype bank, empirical-null calibration, and conflict-aware reasoning to separate evidence retrieval from anomaly decision-making. Experiments on MVTec-AD, VisA, MPDD, and Real-IAD demonstrate strong localization performance. **Limitations and future work.** R2AD relies on predefined evidence operators from a frozen foundation encoder and patch-wise empirical calibration. Future work includes learning evidence operators end to end and developing image-wise error-controlled calibration.

AI USE STATEMENT

We used generative AI tools only to aid and polish the writing of this paper, including language refinement and clarity improvements. All AI-assisted text was reviewed and revised by the authors to ensure technical accuracy and consistency with the underlying methods and results. We take full responsibility for the final content of this work, including any text produced with the aid of generative AI.

REFERENCES

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, pp. 9592–9600, 2019. 6

Mingxiu Cai, Zhe Zhang, Gaochang Wu, Tianyou Chai, and Xiatian Zhu. Raid: Retrieval-augmented anomaly detection. *arXiv preprint arXiv:2602.19611*, 2026. 1, 2, 6, 7

Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. A survey on visual anomaly detection: Challenge, approach, and prospect. *arXiv preprint arXiv:2401.16402*, 2024a. 1

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *ECCV*, pp. 55–72, 2024b. 2

Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 1

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In *WACV*, pp. 1319–1329. IEEE, 2025. 1, 6

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 4, 7

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *CVPR*, pp. 20405–20415, 2025. 1, 6

Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. In *NeurIPS*, volume 37, pp. 71162–71187, 2024a. 1, 6

Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *AAAI*, volume 38, pp. 8472–8480, 2024b. 1, 6

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*, pp. 66–71, 2021. 6

Yuxuan Lin, Yang Chang, Xuan Tong, Jiawen Yu, Antonio Liotta, Guofan Huang, Wei Song, Deyu Zeng, Zongze Wu, Yan Wang, et al. A survey on rgb, 3d, and multimodal approaches for unsupervised industrial image anomaly detection. *Information Fusion*, 121:103139, 2025. 1

Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1): 104–135, 2024. 1

Ruiying Lu, Yujie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. In *NeurIPS*, volume 36, pp. 8487–8500, 2023. 6

Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *CVPR*, pp. 9974–9983, 2025. 1, 6

Wenxi Lv, Qinliang Su, and Wenchao Xu. One-for-all few-shot anomaly detection via instance-induced prompt learning. In *ICLR*, 2025. 2

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. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification. 2, 4

Zhen Qu, Xian Tao, Xinyi Gong, ShiChen Qu, Xiaopei Zhang, Xingang Wang, Fei Shen, Zhengtao Zhang, Mukesh Prasad, and Guiguang Ding. Dictas: A framework for class-generalizable few-shot anomaly segmentation via dictionary lookup. In *ICCV*, pp. 20519–20528, 2025. 1

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, pp. 14318–14328, 2022. 1, 6

Shunsuke Sakai, Xiangteng He, Chunzhi Gu, Leonid Sigal, and Tatsuhito Hasegawa. Invad: Inversion-based reconstruction-free anomaly detection with diffusion models. In *CVPR*, pp. 21389–21398, 2026. 6

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. 6

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *NeurIPS*, volume 35, pp. 4571–4584, 2022. 6

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, pp. 8330–8339, 2021. 1

Xingwu Zhang, Guanxuan Li, Paul Henderson, Gerardo Aragon-Camarasa, and Zijun Long. Is training necessary for anomaly detection? *arXiv preprint arXiv:2601.22763*, 2026. 7

Zhe Zhang, Mingxiu Cai, Hanxiao Wang, Gaochang Wu, Tianyou Chai, and Xiatian Zhu. CostFilter-AD: Enhancing anomaly detection through matching cost filtering. In *ICML*, volume 267, pp. 74540–74564, 13–19 Jul 2025. 1, 2, 6

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *ICLR*, 2024. 2

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*, pp. 392–408, 2022. 6