RETHINKING READOUTS IN CLIP-BASED ZERO-SHOT ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP-based zero-shot anomaly detectors build task-specific decision rules on top of frozen vision-language representations. However, a strong representation does not guarantee a strong decision: useful anomaly evidence may already be encoded in the representation but not fully used by the readout, i.e., the mechanism that turns the representation into an anomaly score. We study this gap in CLIP-based zero-shot anomaly detection. For a canonical binary CLIP readout, the anomaly score depends on a single decision direction. We show that substantial predictive information remains outside this direction, that the remaining information is structured rather than a by-product of high dimensionality, and that the useful directions vary across anomaly contexts. We then ask whether this unused information can be recovered with only a simple change to the detector. We keep the original detector frozen and add a lightweight, image-conditioned reader that makes only a small bounded correction to the existing decision pathway, using auxiliary anomaly data for training. Despite its limited capacity, this intervention improves a diverse set of CLIP-based anomaly detectors. Our results suggest that improving the representation is only part of the problem in zero-shot anomaly detection. It is also important to ask how much task-relevant information the final readout actually uses, and whether information left unused by the original decision can be recovered.

1 INTRODUCTION

Recent CLIP-based zero-shot anomaly detection (ZSAD) methods have made rapid progress by improving the information available to the detector. Existing work has explored better anomaly prompts, stronger visual features, and richer interaction between vision and language. Most of these efforts ask how to build a better representation of abnormality.

Yet representation is only one part of the problem. Many CLIP-based ZSAD detectors ultimately map rich visual features to a compact normal-versus-anomalous decision, while anomalies can vary greatly across objects and defect types. This raises a different question: *if useful anomaly information is already present in the representation, does the detector actually use all of it?*

A model can make an imperfect prediction even when its representation still contains information that could support a better one. We refer to the mismatch between what is encoded and what is used by the final decision mechanism, or readout, as a *representation–readout gap*. Rather than asking only how informative a representation is, we study how much task-relevant information is left unused by the final decision.

We find that this gap can be substantial. For a canonical binary CLIP detector, the anomaly score depends on a single effective direction. However, strong predictive information remains outside this direction. This information is structured rather than a simple consequence of high dimensionality, and the useful directions vary across anomaly contexts. As a result, a single fixed readout cannot capture the remaining evidence uniformly.

We next ask whether this unused information can be recovered with only a small change to the detector. We keep the original detector frozen and add a lightweight image-conditioned reader that makes a small bounded correction to the existing decision pathway. The reader is trained only on auxiliary anomaly data and does not modify the host model itself. Across a diverse set of CLIP-based

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

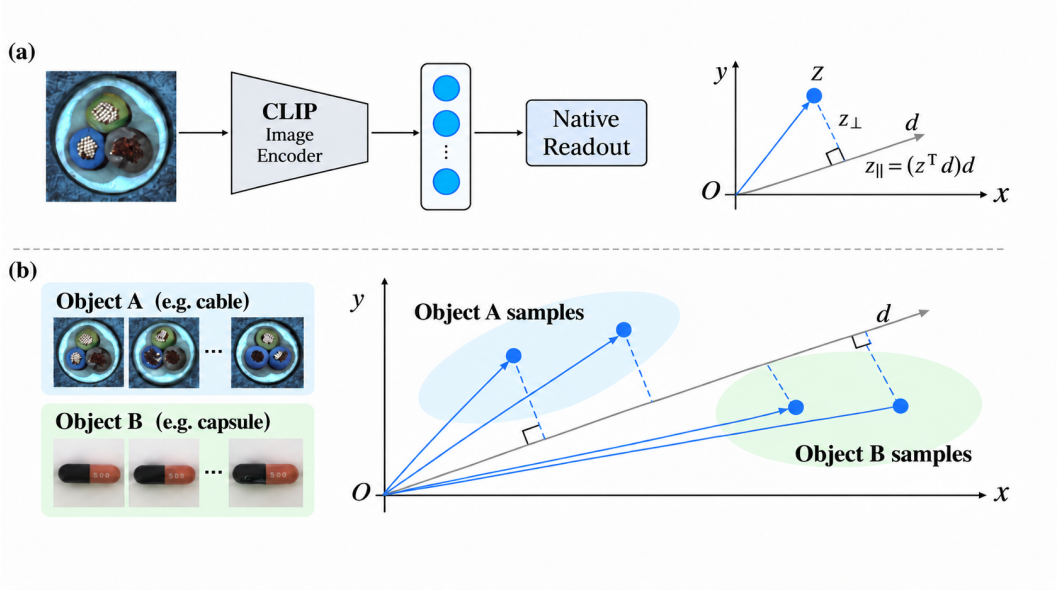


Figure 1: **Readout-limited anomaly evidence in CLIP-based ZSAD.** (a) A canonical binary CLIP readout uses only the projection of the visual representation z onto a single decision direction d, leaving the orthogonal component $z_\perp$ unused by the native score. (b) The residual component can still contain structured anomaly evidence whose useful direction varies across object contexts, motivating an image-conditioned readout.

anomaly detectors, this simple intervention consistently improves image-level anomaly detection, with the gains remaining stable across random seeds.

Together, these results show that representation quality and decision utilization should be considered separately. A detector may already encode useful anomaly evidence that its readout does not expose, and part of this evidence can be recovered without retraining the underlying model.

Our contributions are threefold:

- We formalize the *representation–readout gap* in CLIP-based ZSAD and show that substantial anomaly-predictive information can remain unused by the final decision.
- We show that the remaining information is structured and context-dependent, and characterize where it lies for a canonical binary CLIP readout.
- We introduce a simple bounded image-conditioned reader that keeps the host detector frozen and recovers part of this unused information across different ZSAD models.

2 Related Work

2.1 CLIP-based Zero-shot Anomaly Detection

Zero-shot anomaly detection (ZSAD) aims to detect anomalies from unseen object categories without target-domain training samples. CLIP Radford et al. (2021) is well suited to this setting because it aligns visual features with open-vocabulary text. Since WinCLIP Jeong et al. (2023) introduces CLIP into ZSAD task, most methods improve performance by strengthening the anomaly evidence available to the detector through training on an auxiliary dataset.

Anomaly semantics. Early methods rely on manually designed normal and abnormal descriptions. AnomalyCLIP Zhou et al. (2024) instead learns object-agnostic normal and abnormal prompts from auxiliary anomaly data. Later work further enriches these semantics with prompt distributions, fine-grained anomaly concepts, or image-dependent prompts Cao et al. (2024); Qu et al. (2024; 2025);

Zhu et al. (2025). These methods make the textual side more expressive and, in some cases, adapt the anomaly semantics to each input image.

Visual and cross-modal adaptation. Another line of work improves the visual evidence used for anomaly detection. Recent methods use visual adapters, multi-scale features, local features, or learnable visual tokens. MoECLIP Park et al. (2026) adapts patch features with dynamically routed experts, while VisualAD Hou et al. (2026) introduces learnable normal and abnormal visual tokens. Other methods strengthen fine-grained anomaly evidence through patch-level, frequency-aware, or foreground-background visual adaptation, and increasingly couple these local representations with textual anomaly semantics Ma et al. (2025); Gong et al. (2025); Zhang & Zhang (2026); Hu et al. (2026). Together, these methods improve the information available for the final anomaly prediction.

Representation and readout. A different question is whether the detector fully uses the predictive information already present in its representation. A strong representation may still contain useful anomaly evidence that is not reflected in the final score, while our work studies this gap directly.

2.2 Representation Probing and Readout Use

Probing methods are commonly used to examine what information is retained in learned representations, often by training simple predictors on frozen features Alain & Bengio (2017); Hewitt & Liang (2019). At the same time, several studies have pointed out that information being decodable from a representation does not necessarily mean that it is used by the original model when making predictions Ravichander et al. (2021); Elazar et al. (2021). This distinction is also related to predictive $\mathcal{V}$-information, which characterizes information relative to a restricted family of predictors rather than assuming that all encoded information is equally accessible Xu et al. (2020).

We study a related question in CLIP-based zero-shot anomaly detection, but tie it directly to the detector's native decision. Given the native anomaly score, we ask whether the underlying frozen representation still contains label-predictive information that is not exposed by this score. For a binary CLIP readout, this question leads to an explicit geometric decomposition of the representation, which lets us identify where the residual predictive signal can reside and examine how it varies across anomaly contexts. We then ask whether a small auxiliary readout can recover part of this signal without updating the host detector.

3 A Representation–Readout Gap in Zero-Shot Anomaly Detection

We separate the representation available to a detector from the decision it exposes. Throughout the paper, Native denotes the original frozen detector before any auxiliary probe or intervention, and *decision pathway* denotes the frozen computation mapping representational states to the final anomaly score.

Let Z denote the frozen representation and $S = g(Z)$ its native anomaly score. We use *representation–readout gap* to describe predictive information retained in Z but not exposed by S. We ask three questions: (i) what remains unread by the native score; (ii) where it resides in a canonical CLIP readout; and (iii) whether it can be captured by one fixed readout.

3.1 What Does the Native Readout Leave Unused?

Let $Y \in \{0, 1\}$ denote the anomaly label. Because $S = g(Z)$ is deterministic, $I(Y; S) \leq I(Y; Z)$. We define the unread predictive information as

$$U_{\text{repr}} := I(Y; Z \mid S) = I(Y; Z) - I(Y; S). \tag{1}$$

Under Bayes-optimal log loss, $\mathcal{R}^*_{\log}(X) = H(Y \mid X)$, giving

$$U_{\text{repr}} = \mathcal{R}^*_{\log}(S) - \mathcal{R}^*_{\log}(Z). \tag{2}$$

Thus, U_{repr} is equivalently the optimal log-loss reduction still available beyond the native score.

We next instantiate this quantity for a canonical binary CLIP readout. Let $z \in \mathbb{R}^D$ be an ℓ_2-normalized visual feature and $t_N, t_A \in \mathbb{R}^D$ normalized normal and anomalous text prototypes. The anomalous probability is

$$p_A(z) = \frac{\exp(z^\top t_A/\tau)}{\exp(z^\top t_A/\tau) + \exp(z^\top t_N/\tau)} = \sigma\left(\frac{\|t_A - t_N\|_2}{\tau} z^\top d\right), \tag{3}$$

where

$$d = \frac{t_A - t_N}{\|t_A - t_N\|_2} \tag{4}$$

is the effective decision direction. Because the sigmoid is monotonic, the readout depends on z only through $z^\top d$, so we write

$$S = s_{\mathrm{nat}}(Z) = Z^\top d. \tag{5}$$

For the controlled analysis below, we use frozen CLIP image embeddings together with the text decision axis learned by AnomalyCLIP; full diagnostic details are given in Appendix B. Although Z is high-dimensional, this binary readout exposes only one decision coordinate.

3.2 Where Does the Unread Predictive Structure Reside?

The one-dimensional readout induces the decomposition

$$z_\parallel = (z^\top d)d, \qquad z_\perp = z - z_\parallel, \tag{6}$$

with $z_\perp^\top d = 0$. Since $S = Z^\top d$ and $\|d\|_2 = 1$,

$$Z = Sd + Z_\perp, \tag{7}$$

so Z and $(S, Z_\perp)$ are in one-to-one correspondence. Hence

$$I(Y; Z \mid S) = I(Y; Z_\perp \mid S), \tag{8}$$

and therefore

$$U_{\mathrm{repr}} = \mathcal{R}^*_{\log}(S) - \mathcal{R}^*_{\log}(S, Z_\perp). \tag{9}$$

For this binary readout, any predictive information beyond the native score must therefore reside in $Z_\perp$ once S is known.

We first test this with a matched target-supervised diagnostic on 48 defect-level tasks (Appendix B.1). For each task, we fit either a one-dimensional logistic probe on S or the same predictor family on the joint feature $[S, Z_\perp]$, using only the training partition for preprocessing and model selection. Adding $Z_\perp$ reduces macro-average held-out log loss from 0.1520 to 0.0553, improving $45/48$ tasks, while Image AUROC increases from 81.816 to 98.858, improving $41/48$ tasks. Thus, even after controlling for the native coordinate through the same predictor family, $Z_\perp$ provides substantial additional predictive signal. These target-supervised predictors are diagnostic witnesses, not zero-shot readouts or numerical estimators of conditional mutual information.

We next use a separate five-split oracle diagnostic to ask how much predictive structure $Z_\perp$ contains on its own. Here, the NATIVE baseline is the unfitted scalar score $z^\top d$, while the ORTHOGONAL ORACLE and FULL ORACLE are target-supervised linear classifiers trained on $Z_\perp$ and Z, respectively. The orthogonal oracle reaches an average Image AUROC of 0.988, compared with 0.851 for the native scalar score and 0.989 for the full representation (Fig. 2(a)).

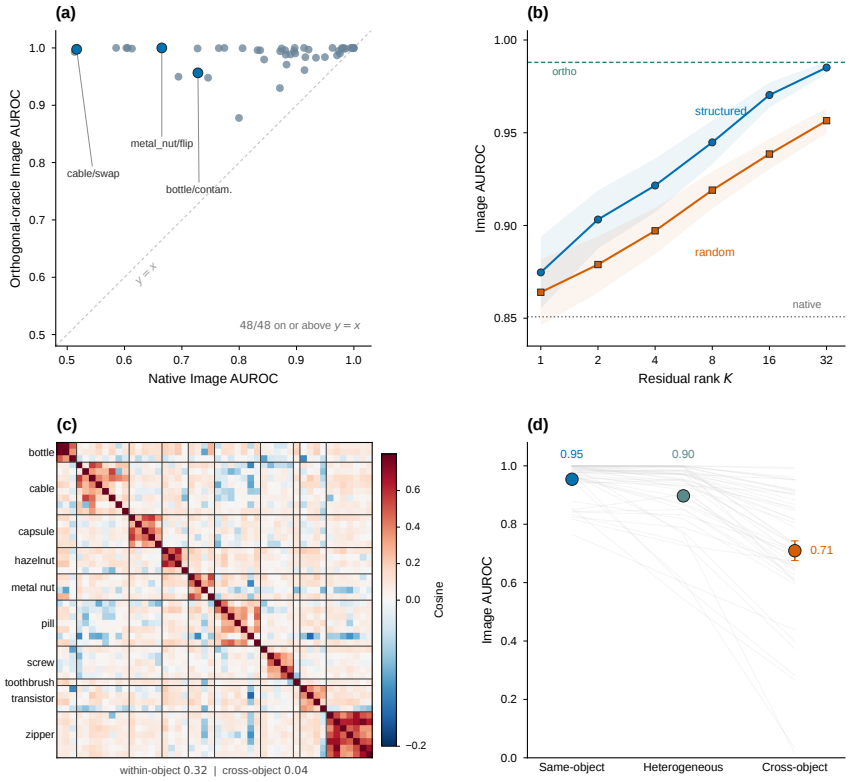


Figure 2: **Predictive evidence outside the native decision coordinate is strong, structured, and context-dependent.** (a) A target-supervised oracle using only $Z_\perp$ outperforms the native scalar score $z^\top d$ for all 48 defect types and nearly matches the full-feature oracle. (b) Low-rank structured residual subspaces recover the orthogonal-oracle performance faster than matched random subspaces. (c) Defect-specific residual directions are more aligned within the same object than across objects (mean cosine 0.32 vs. 0.04). (d) Transfer weakens as the training context becomes less compatible: mean Image AUROC decreases from 0.95 for same-object evidence to 0.90 for heterogeneous evidence and 0.71 for cross-object evidence. Shaded regions and error bars denote SEM across defect-level results where applicable.

To test whether this result is merely a consequence of high dimensionality, we compare low-rank PCA residual subspaces with dimension-matched random subspaces under the same held-out protocol (Appendix B.2). As shown in Fig. 2(b), structured residual subspaces recover the orthogonal-oracle performance substantially faster than matched random subspaces. Thus, useful anomaly evidence is preferentially concentrated in low-dimensional residual structure rather than arbitrary residual dimensions.

3.3 Why a Fixed Residual Readout Is Insufficient

The remaining question is whether one shared residual direction can capture the evidence identified above. Let $u_c \in d^\perp$ denote the target-supervised residual direction for anomaly context c; its estimation protocol is detailed in Appendix B.3.

Figure 2(c) shows that these directions are locally coherent but vary substantially across objects: mean cosine similarity is 0.32 within the same object and only 0.04 across objects. When heterogeneous residual directions are weakly aligned, averaging them into one shared correction attenuates their context-specific components.

The held-out-defect transfer experiment (Appendix B.4) shows the corresponding predictive effect. Using the same readout family and evaluation protocol, mean Image AUROC is approximately 0.95 with same-object training evidence, 0.90 with heterogeneous evidence, and 0.71 under cross-object

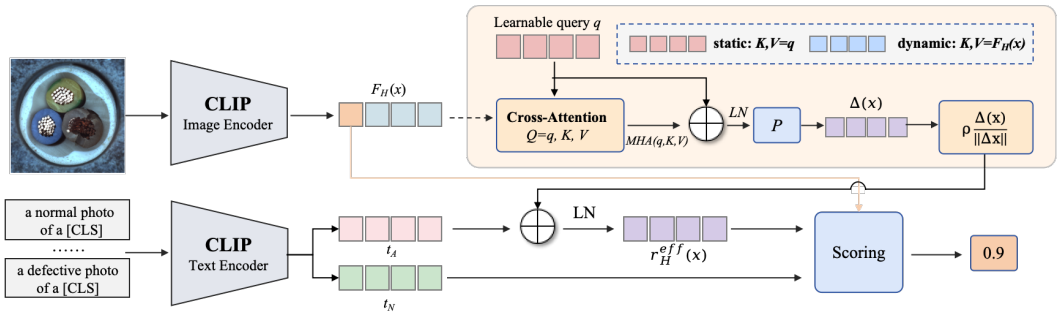


Figure 3: **Controlled readout probe on AnomalyCLIP.** The frozen host supplies image features to a lightweight probe. STATIC uses a shared context, while DYNAMIC conditions on the current image. The resulting residual is norm-bounded and injected only into the anomalous text prototype before native scoring.

transfer (Fig. 2(d)). Thus, residual evidence transfers substantially better within a compatible object context than across heterogeneous contexts.

Section 3 therefore leaves two observations for the intervention study: the frozen representation contains structured predictive evidence outside the coordinate exposed by the binary readout, and the useful residual direction varies with context. Section 4 asks whether an image-conditioned restricted probe can make part of this evidence operational using only auxiliary anomaly supervision.

4 A CONTROLLED PROBE FOR READOUT ACCESSIBILITY

Section 3 shows that predictive anomaly information can remain outside the native decision coordinate and that useful residual directions vary across contexts. We therefore ask whether a small change to the readout can recover part of this information without modifying the host representation. We freeze the detector and attach a lightweight residual probe to one designated readout state. Figure 3 illustrates the construction on AnomalyCLIP.

4.1 CONTROLLED RESIDUAL PROBE

For a frozen host H, let $r_H(x)$ denote the readout state being modified and $F_H(x) = \{f_i(x)\}_{i=1}^N$ the frozen features available to the probe. For AnomalyCLIP, $r_H(x) = t_A$; intervention points for other hosts are given in Appendix A.3.

The probe consists of one learnable query q, a single cross-attention layer, and a linear projection:

$$h(x) = q + \mathrm{MHA}(\mathrm{LN}(q), \mathrm{LN}(K), \mathrm{LN}(V)), \qquad \Delta_\theta(x) = P(\mathrm{LN}(h(x))). \tag{10}$$

We use no feed-forward block.

To isolate the value of image conditioning, we compare two matched variants:

$$(K, V) = \begin{cases} (q, q), & \text{STATIC}, \\ (F_H(x), F_H(x)), & \text{DYNAMIC}. \end{cases} \tag{11}$$

STATIC therefore produces one shared correction, whereas DYNAMIC can adapt the correction to the current image.

4.2 BOUNDED READOUT INTERVENTION

The probe modifies only the designated readout state:

$$r_H^{\mathrm{eff}}(x) = \mathrm{Norm}\left(r_H(x) + \rho \frac{\Delta_\theta(x)}{\|\Delta_\theta(x)\|_2 + \epsilon}\right). \tag{12}$$

Table 1: Comparison of native, static, and dynamic readouts across six frozen ZSAD hosts. Reader results are mean ± standard deviation over three seeds.

Host	Native	Static	Dynamic	Δ_{shared}	Δ_{cond}	Δ_{total}
AnomalyCLIP	91.55	90.98 ± 0.25	$\mathbf{92.19 \pm 0.33}$	-0.57 ± 0.25	$+1.21 \pm 0.51$	$\mathbf{+0.64 \pm 0.33}$
APRIL-GAN	86.16	87.34 ± 0.39	$\mathbf{88.63 \pm 1.09}$	$+1.18 \pm 0.39$	$+1.30 \pm 1.03$	$\mathbf{+2.47 \pm 1.09}$
VisualAD	92.23	92.23 ± 0.00	92.23 ± 0.00	$+0.00 \pm 0.00$	$+0.00 \pm 0.00$	$+0.00 \pm 0.00$
MoECLIP	90.72	90.37 ± 0.34	$\mathbf{91.21 \pm 0.07}$	-0.35 ± 0.34	$+0.83 \pm 0.41$	$\mathbf{+0.49 \pm 0.07}$
FAPrompt	91.80	91.84 ± 0.19	$\mathbf{92.03 \pm 0.12}$	$+0.04 \pm 0.19$	$+0.19 \pm 0.10$	$\mathbf{+0.23 \pm 0.12}$
VCP-CLIP	89.79	89.45 ± 0.20	$\mathbf{89.92 \pm 0.23}$	-0.34 ± 0.20	$+0.48 \pm 0.27$	$\mathbf{+0.13 \pm 0.23}$
Avg.	**90.38**	$\mathbf{90.37 \pm 0.04}$	$\mathbf{91.04 \pm 0.07}$	$\mathbf{-0.01 \pm 0.04}$	$\mathbf{+0.67 \pm 0.06}$	$\mathbf{+0.66 \pm 0.07}$

The bound ρ limits the intervention magnitude, while all downstream host computations remain unchanged. For AnomalyCLIP, Eq. 12 modifies only the anomalous text prototype t_A.

All host parameters are frozen. The projection P is zero-initialized, and only the probe parameters are optimized on auxiliary anomaly data with an image-level classification loss. No target-domain samples or labels are used. Across hosts, the probe architecture and training protocol are fixed; only the intervention state and native scoring function remain host-specific.

5 Experiments

Section 3 showed that predictive anomaly evidence can remain outside the coordinate exposed by a binary CLIP readout, and that useful residual directions vary across contexts. We now ask whether a restricted reader trained only on auxiliary anomalies can make part of this evidence operational under the zero-shot protocol, where its gain comes from, and whether that gain requires the context of the current image.

5.1 Experimental Setup

We train the reader only on Real-IAD-SV Wang et al. (2024) and evaluate it directly on MVTec AD Bergmann et al. (2019). No MVTec image or label is used for probe training, checkpoint selection, or target-domain adaptation, and all host detectors remain frozen. We evaluate six CLIP-based ZSAD hosts: AnomalyCLIP Zhou et al. (2024), APRIL-GAN Chen et al. (2023), VisualAD Hou et al. (2026), MoECLIP Park et al. (2026), FAPrompt Zhu et al. (2025), and VCP-CLIP Qu et al. (2024).

We compare the frozen detector (Native) with two matched readers. Static learns one shared residual correction, whereas Dynamic conditions the correction on frozen features of the current image. The two variants use the same host checkpoint, auxiliary data, optimization budget, host-specific objective, and residual bound. Unless otherwise stated, we use one shared residual budget $\rho = 0.025$ for all hosts and report results over three training seeds. Our primary metric is macro image-level AUROC over the 15 MVTec categories.

Full implementation and host-specific integration details are given in Appendix A. Seed-level results are reported in Appendix C.1, and sensitivity to the shared residual budget is reported in Appendix C.2.

5.2 Can the Reader Recover Unused Predictive Evidence?

We decompose the change in performance as

$$\Delta_{\text{shared}} = M_{\text{Static}} - M_{\text{Native}}, \qquad \Delta_{\text{cond}} = M_{\text{Dynamic}} - M_{\text{Static}}, \qquad \Delta_{\text{total}} = M_{\text{Dynamic}} - M_{\text{Native}},$$

where M denotes macro image-level AUROC.

Table 1 shows a clear difference between a shared correction and an image-conditioned one. Dynamic improves five of the six frozen hosts and leaves VisualAD unchanged, yielding an average gain of $+0.66 \pm 0.07$ AUROC points over Native. By comparison, Static changes the cross-host average by only -0.01 ± 0.04, while conditioning contributes $+0.67 \pm 0.06$ relative to Static.

Table 2: Decomposition of the Dynamic gain on AnomalyCLIP. Only the indicated component of the reader-induced margin change is retained. Results are macro image-level AUROC (%).

Variant	Image AUROC
Native	91.55
Parallel-only	91.54
Orthogonal-only	91.95
Full Dynamic	**91.95**

The effect is therefore not explained, on average, by simply learning one shared residual correction. The seed-level results in Appendix C.1 show the same overall pattern, while Appendix C.2 shows that the cross-host effect is not tied to a single narrowly tuned residual magnitude. In particular, $\rho = 0.025$ is shared across hosts rather than selected separately for each detector.

These improvements quantify the utility realized by the restricted reader; they are not numerical estimates of U_{repr}. We next test whether the realized gain passes through the representation component that is invisible to the binary native coordinate.

5.3 Does the Gain Come from the Unread Orthogonal Component?

We test this directly on AnomalyCLIP using the binary margin decomposition from Sec. 3. Let d denote the decision direction and $z = z_{\parallel} + z_{\perp}$. For the trained DYNAMIC reader, let $\Delta t = t_A^{\mathrm{eff}} - t_A$ denote the effective change to the anomalous prototype after normalization. The induced change in the image–text margin decomposes as

$$z^{\top}\Delta t = z_{\parallel}^{\top}\Delta t + z_{\perp}^{\top}\Delta t.$$

We evaluate counterfactual scores retaining only the parallel or orthogonal contribution. No model is retrained. This diagnostic uses reader seed 222, whose full Dynamic result corresponds to the seed-222 entry reported in Appendix C.1.

The orthogonal-only counterfactual preserves essentially the entire Dynamic gain, whereas the parallel-only contribution provides no improvement over NATIVE. For this AnomalyCLIP intervention, the useful change in the decision margin is therefore mediated almost entirely through the component orthogonal to the original decision direction. This provides an operational link to the residual component identified in Sec. 3.

We next ask whether access to that component is sufficient with arbitrary image context, or whether the correction must be conditioned on the image being scored.

5.4 Does the Reader Need the Correct Image Context?

A conditional reader could benefit from generic image-dependent features without using information specific to the image being scored. We test this possibility on AnomalyCLIP by keeping the trained reader and host input fixed while replacing only the reader's key and value features at inference time.

For a test image x_i, we compare: (i) CORRECT, using $F_H(x_i)$; (ii) SAME-OBJECT, using features from another image of the same MVTec category; (iii) CROSS-OBJECT, using features from a different category; and (iv) the image-independent STATIC reader. No reader is retrained. Same- and cross-object conditions are each averaged over 16 random donor pairings. Object identity is used only to construct the swap; anomaly labels and defect types are not used for donor selection.

Correct conditioning performs best. Replacing the current-image features with same-object features reduces AUROC by 0.75 points and nearly removes the advantage over STATIC; cross-object conditioning reduces performance further to 90.95. The ordering

$$\text{CORRECT} > \text{SAME-OBJECT} > \text{CROSS-OBJECT}$$

holds for both reader seeds used in this diagnostic.

Thus, most of the Dynamic gain requires conditioning on the particular image being scored. The smaller advantage of same-object over cross-object conditioning is consistent with the stronger

Table 3: Effect of mismatched conditioning on AnomalyCLIP. Only the conditioning features are replaced at inference time. Results are macro image-level AUROC (%).

Condition	Image AUROC
Correct	$\mathbf{92.01 \pm 0.08}$
Same-object	91.26 ± 0.07
Cross-object	90.95 ± 0.03
Static	91.27 ± 0.03

within-object residual alignment observed in Sec. 3.3. Together with Table 2, these counterfactuals support two distinct claims: the useful AnomalyCLIP correction acts primarily through the orthogonal component of the binary decision geometry, and its utility depends on matching the correction to the current image context.

6 Conclusion

A frozen anomaly representation having "something left to read" is not a single yes–or–no property. Our results suggest that three questions should be separated: whether task-relevant predictive information remains encoded beyond the native decision statistic, whether that information is accessible to a restricted readout, and whether auxiliary supervision can turn that accessibility into target-domain utility. In this sense, what is encoded, what is readable, and what is practically realized are distinct quantities.

We made this distinction concrete in CLIP-based zero-shot anomaly detection. For a canonical binary prompt readout, the native decision collapses to a single direction, while strong predictive structure remains in orthogonal coordinates that are invisible to the native score. This residual evidence is structured and context-dependent: a shared correction provides essentially no average benefit, whereas an image-conditioned reader recovers positive utility across diverse frozen hosts. When the conditioning context is replaced, most of this gain disappears, linking the behavior of the reader back to the residual geometry identified in our analysis. The absolute gains are modest, and we do not present the reader as a new performance-oriented ZSAD module. Its role is instead that of a controlled intervention, showing that part of the information missed by the native readout can be made operational under auxiliary supervision.

These observations motivate treating representation quality and readout utilization as separate parts of the anomaly-detection problem. A strong representation is not sufficient if the final decision exposes only a restricted view of its task-relevant information. This shifts the question from how much information a representation contains to what the native decision leaves unread, what forms of readout can access it, and how much of that accessibility can be realized under zero-shot transfer.

STATEMENTS

AI USE STATEMENT

Large language models and other generative AI tools were used to assist with drafting, organizing, and polishing the manuscript, preparing LaTeX content, and checking formatting and references. They also supported research ideation and execution and assisted in checking mathematical claims. All resulting content, experiments, and conclusions were independently reviewed and verified by the authors. AI-assisted checks were supplementary to this author review and were not treated as independent evidence for the validity of mathematical claims or empirical conclusions. The authors take responsibility for the final manuscript, including all AI-assisted text, claims, and artifacts.

ETHICS STATEMENT

This work studies anomaly detection using existing industrial visual benchmarks and pretrained models. Our evaluation concerns the use of predictive information in frozen representations; benchmark performance alone does not establish reliability in operational inspection systems. False negatives may allow defective products to pass inspection, while false positives may cause unnecessary rejection or additional manual review. Deployment therefore requires validation under the intended operating conditions and appropriate human oversight. We also distinguish target-supervised diagnostic analyses from the auxiliary-trained probe to avoid presenting oracle performance as deployable zero-shot performance. Use and redistribution of datasets, pretrained models, and associated artifacts should respect their respective licenses and access conditions.

REPRODUCIBILITY STATEMENT

The probe formulation and optimization objective are described in Section 4. Appendix A documents the auxiliary-data construction, target evaluation protocol, frozen host checkpoints, optimization settings, and host-specific integration. Appendix B details the diagnostic analyses of residual predictive information, including the use of target supervision for diagnostic oracles. Appendix C reports random-seed results and residual-budget sensitivity. These descriptions separate diagnostic and deployable settings and specify the comparisons used to assess the reader. Checkpoint provenance is reported where known, with unavailable provenance identified explicitly. The source code, configuration files, and evaluation scripts needed to reproduce the reported experiments will be released in an anonymous repository during review and made publicly available upon acceptance.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*, 2017.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592, 2019. doi: 10.1109/CVPR.2019.00982.

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 *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-72761-0_4.

Xuhai Chen, Yue Han, and Jiangning Zhang. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1 and 2: 1st place on zero-shot ad and 4th place on few-shot ad, 2023. URL https://arxiv.org/abs/2305.17382.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

Tao Gong, Qi Chu, Bin Liu, Wei Zhou, and Nenghai Yu. Fe-clip: Frequency enhanced clip model for zero-shot anomaly detection and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21220–21230, 2025.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275.

Yanning Hou, Peiyuan Li, Zirui Liu, Yitong Wang, Yanran Ruan, Jianfeng Qiu, and Ke Xu. Visualad: Language-free zero-shot anomaly detection via vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21346–21356, 2026.

Ming Hu, Yongsheng Huo, Mingyu Dou, Jianfu Yin, Peng Zhao, Yao Wang, Cong Hu, Bingliang Hu, and Quan Wang. Fb-clip: Fine-grained zero-shot anomaly detection with foreground-background disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35659–35669, 2026.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation, 2023. URL https://arxiv.org/abs/2303.14814.

Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S. Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4744–4754, 2025. doi: 10.1109/CVPR52734.2025.00447.

Jun Yeong Park, JunYoung Seo, Minji Kang, and Yu Rang Park. Moeclip: Patch-specialized experts for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35534–35544, 2026.

Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-72890-7_18.

Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30398–30408, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22883–22892, 2024.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

Gaowei Zhang and Lihe Zhang. Dlvp-clip: Enhancing fine-grained zero-shot anomaly detection via dynamic local visual prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35524–35533, 2026.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22241–22251, 2025.

Table 4: Frozen host checkpoints used for MVTec AD evaluation. Host training and probe training are separate stages.

Host	Checkpoint	Host source	Provenance
AnomalyCLIP	`epoch_15.pth`	VisA	local VisA→MVTec checkpoint
APRIL-GAN	`visa_pretrained.pth`	VisA	repository-provided checkpoint
VisualAD	`train_on_visa/CLIP.pth`	VisA	repository-provided checkpoint
MoECLIP	`VisA_fullshot/moe_last.pth`	VisA	locally trained final checkpoint
FAPrompt	`train_on_visa.pth`	VisA (indicated)	repository checkpoint; exact provenance unavailable
VCP-CLIP	`train_visa.pth`	VisA	repository-provided checkpoint

A Experimental Protocol and Implementation

A.1 Zero-Shot Protocol and Auxiliary Data

We follow the cross-dataset zero-shot anomaly detection protocol used in AnomalyCLIP Zhou et al. (2024) and related CLIP-based methods. The detector may use labeled auxiliary anomaly data, but no sample or label from the target dataset is used for training or target-domain adaptation. In our main experiments, MVTec AD Bergmann et al. (2019) is the target dataset.

The frozen host detector and the probe are trained at two separate stages. For MVTec evaluation, the host checkpoints predominantly follow the VisA-to-MVTec settings of the corresponding methods, whereas the probe is trained separately on Real-IAD-SV Wang et al. (2024). We therefore do not interpret Native–Probe comparisons as matched-supervision comparisons between two training procedures. The Static and Dynamic probes, however, use the same frozen host, auxiliary data, and optimization protocol.

Real-IAD-SV is a single-view subset of Real-IAD containing only the C1 camera view. Our auxiliary set combines 10,408 normal records from the original training partition with 10,387 normal and 9,415 anomalous records from the original test partition, giving 30,210 images in total. These Real-IAD partition names are internal to the auxiliary dataset and are unrelated to the MVTec target evaluation.

We report the unweighted macro average over the 15 MVTec categories. Image-level AUROC is the primary cross-host metric.

A.2 Frozen Hosts and Optimization

Table 4 summarizes the frozen checkpoints used for MVTec evaluation.

Unless otherwise specified, the probe is trained for 20,295 updates with batch size 8 using Adam with learning rate 10^{-3} and $\beta = (0.5, 0.999)$. We use no learning-rate scheduler, explicit weight decay, gradient clipping, or early stopping, and evaluate the final training checkpoint. The main residual budget is fixed to $\rho = 0.025$. The reader uses eight attention heads. Its learned query is initialized from a normal distribution with standard deviation 0.02, while the output projection is zero-initialized. All host parameters remain frozen and only probe parameters are optimized.

Static and Dynamic probes share the same frozen host checkpoint, Real-IAD-SV data, preprocessing, objective, and optimization budget. Static uses a learned global context, whereas Dynamic conditions on frozen features from the current image.

A.3 Host-Specific Integration

The six hosts expose heterogeneous decision pathways. For each host, the probe reads a frozen feature sequence $F_H(x)$ and applies a bounded residual to a designated representation $r_H(x)$ inside the original decision pathway. All remaining host computations are unchanged.

For AnomalyCLIP, APRIL-GAN, MoECLIP, and FAPrompt, the intervention acts on an abnormal text representation before the original downstream scoring pathway. VisualAD instead modifies its anomaly representation before anomaly-map generation, while VCP-CLIP modifies the final dense

Table 5: Host-specific probe integration. All host parameters remain frozen.

Host	Reader input F_H	Intervention state r_H	Native image score	Probe loss
AnomalyCLIP	non-CLS visual patches	abnormal text prototype t_A	image–text probability	CE + focal + Dice
APRIL-GAN	final projected patches	abnormal text embedding t_A	image–text probability	CE
VisualAD	transformed spatial features	anomaly representation	top-k anomaly-map score	BCE
MoECLIP	normalized spatial tokens	adapted abnormal text t_A	detection-branch score	BCE
FAPrompt	non-CLS spatial patches	DAP abnormal prototype t_A	fused text/map score	BCE
VCP-CLIP	VCP dense features	final new-map branch	fused raw/new map score	BCE

new-map branch. The feature dimensions are 768 for AnomalyCLIP, APRIL-GAN, MoECLIP, and FAPrompt, and 1024 for VisualAD and VCP-CLIP.

B Diagnostic Protocol for the Representation–Readout Analysis

The experiments in this section provide the diagnostic evidence underlying the representation–readout analysis in Sec. 3. Target-domain labels used in these analyses are never used to train the reader or any zero-shot deployment model. We analyze the 10 object categories of MVTec AD, excluding the five texture categories, yielding 48 defect types in total. Images are collected from both the training and test partitions specified by the MVTec metadata. For each defect-level task, anomalous images of the current defect form the positive class, while all normal (`good`) images of the same object form the negative class.

The diagnostic representation z is the ℓ_2-normalized image embedding produced by a frozen CLIP ViT-L/14@336 encoder. The text decision axis is derived from the normal and anomalous text representations learned by an AnomalyCLIP prompt learner trained on VisA:

$$d = \frac{t_A - t_N}{\|t_A - t_N\|_2}. \tag{13}$$

We then decompose the frozen CLIP representation as

$$z_{\|} = (z^\top d)d, \qquad z_{\perp} = z - z_{\|}. \tag{14}$$

Thus, the diagnostics in this section instantiate the canonical binary CLIP readout using a frozen CLIP image representation together with an AnomalyCLIP-derived text decision axis. They should not be interpreted as reproducing the complete AnomalyCLIP visual inference pathway.

Unless otherwise stated, each defect-level diagnostic treats images of the given defect as positives and normal (`good`) images of the same object as negatives.

Unless otherwise stated, learned diagnostic predictors use an ℓ_2-regularized logistic classifier implemented with `LogisticRegressionCV`. The inverse regularization strength is selected from

$$C \in \mathrm{logspace}(-2, 2, 8)$$

using stratified inner cross-validation with AUROC as the selection metric. The inner number of folds is $\min(5, n_+, n_-)$, with a minimum of two folds. We use the `lbfgs` solver, refit the selected model on the full outer training partition, and use neither class weighting nor sample weighting.

Before fitting each classifier, features are standardized using `StandardScaler` with mean centering and variance normalization. The scaler is fitted exclusively on the corresponding outer training partition and is then applied unchanged to the test partition. All other learned preprocessing, including PCA where applicable, is likewise fitted using training data only.

Unless otherwise specified, experiments are repeated over five outer partitions with test fraction 0.30 and deterministic random states $100, 101, 102, 103$, and 104. For the defect-wise bottleneck experiments, these partitions are stratified by the binary anomaly label. Randomness internal to classifier cross-validation is generated from a fixed NumPy random generator initialized with seed 0.

B.1 Residual Predictive Signal

We first test whether the frozen CLIP representation retains predictive signal beyond the scalar coordinate exposed by the binary text decision axis.

For each defect-level task, anomalous images of the current defect form the positive class and normal images of the same object form the negative class. All 48 defect types satisfy the inclusion criterion.

Conditional residual diagnostic. We first ask whether $Z_\perp$ provides additional predictive information once the native coordinate $S = z^\top d$ is already available. For each of the 48 tasks, we compare a one-dimensional logistic probe on S with the same predictor family applied to the concatenated feature $[S, Z_\perp]$.

This diagnostic uses a single stratified 70/30 train–test split with random seed 111. Both predictors use a `StandardScaler`→`LogisticRegression` pipeline. The scaler and classifier are selected entirely within the outer training partition. The regularization strength is chosen from

$$C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\}$$

using inner stratified cross-validation with negative log loss as the selection criterion, after which the selected pipeline is refit on the full outer training partition and evaluated on the held-out test partition.

Adding $Z_\perp$ reduces mean held-out log loss from 0.1520 to 0.0553, improving 45 of 48 tasks. Mean Image AUROC increases from 81.816 to 98.858, with improvements on 41 of 48 tasks. Both predictors use target labels and therefore serve only as diagnostic oracles; they are not zero-shot readouts or numerical estimators of $I(Y; Z_\perp \mid S)$.

Residual oracle diagnostic. We separately evaluate how much predictive structure $Z_\perp$ contains on its own using the default five-split protocol described above. Each task is evaluated over five stratified 70/30 train–test partitions.

The Native baseline is the unfitted scalar score

$$s_{\mathrm{nat}}(z) = z^\top d, \tag{15}$$

and is evaluated on exactly the same held-out images as the corresponding oracles.

The Orthogonal Oracle is an ℓ_2-regularized logistic classifier trained only on the standardized residual representation $z_\perp$. The Full Oracle uses the same scaler, model-selection procedure, and classifier family but receives the complete frozen representation z. All feature scaling and hyperparameter selection are performed using only the corresponding outer training partition.

For each defect, Image AUROC is first averaged over the five outer test partitions. The 48 points in Fig. 2(a) are these defect-level averages. The Orthogonal Oracle reaches a mean Image AUROC of 0.988, compared with 0.851 for the unfitted Native scalar score and 0.989 for the Full Oracle.

B.2 Low-Rank Structure of the Residual Signal

We next test whether the residual predictive signal is concentrated in low-dimensional structure or can be recovered equally well from arbitrary residual dimensions.

For each outer training partition and each rank

$$K \in \{1, 2, 4, 8, 16, 32\},$$

the Structured condition uses

$$\left[z^\top d,\ \mathrm{PCA}_K(z_\perp)\right]. \tag{16}$$

The PCA basis is fitted exclusively on the residual features of the outer training partition and uses no anomaly labels. The resulting features are subsequently standardized using the same train-only scaling procedure before fitting the logistic classifier.

The dimension-matched Random condition uses

$$\left[z^\top d,\ z_\perp Q_K\right], \tag{17}$$

where Q_K contains K random orthonormal directions in the subspace orthogonal to d. To construct Q_K, we sample an isotropic Gaussian matrix, remove its component along d, and orthonormalize the remaining columns using QR decomposition. For each outer partition and rank, we sample 20 independent random bases and average their test AUROC before aggregating across outer partitions.

All downstream scaling, classifier fitting, and regularization selection are identical between Structured and Random conditions. For each (defect, K) pair, results are first averaged over the five outer partitions. Figure 2(b) then reports the mean over the 48 defect types, with shaded regions showing the standard error across defects rather than across data splits.

Structured residual subspaces outperform dimension-matched random subspaces at every tested rank. At $K = 32$, the Structured condition reaches approximately 0.985 Image AUROC, close to the full orthogonal-oracle result of 0.988, whereas the matched Random condition reaches approximately 0.957.

B.3 Context Dependence of Residual Directions

We next examine whether the residual predictive direction is globally shared or varies across anomaly contexts.

For each defect type c, we construct the same binary task consisting of that defect versus normal images of the corresponding object. For each of five stratified 70/30 outer partitions, a logistic classifier is trained on standardized $z_\perp$ features.

Let $\beta_c^{(r)}$ denote the classifier coefficient in standardized coordinates for repeat r, and let $\sigma_c^{(r)}$ denote the feature-wise scale learned by the corresponding `StandardScaler`. We first map the coefficient back to the original feature coordinates:

$$w_c^{(r)} = \beta_c^{(r)} \oslash \sigma_c^{(r)}, \tag{18}$$

where $\oslash$ denotes element-wise division. Because only the direction of the classifier is used, the centering offset does not affect the resulting direction.

We then remove any numerical component along the native axis and normalize:

$$\hat{w}_c^{(r)} = \frac{w_c^{(r)} - (w_c^{(r)\top} d)d}{\left\| w_c^{(r)} - (w_c^{(r)\top} d)d \right\|_2}. \tag{19}$$

The final defect-specific residual direction is obtained by averaging the five unit directions and normalizing once more:

$$u_c = \frac{\sum_{r=1}^{5} \hat{w}_c^{(r)}}{\left\| \sum_{r=1}^{5} \hat{w}_c^{(r)} \right\|_2}. \tag{20}$$

Figure 2(c) reports the pairwise cosine similarities between these 48 defect-specific directions. The matrix is ordered by object category, and its diagonal entries are one. Summary statistics use only off-diagonal pairs: 111 defect pairs belong to the same object and 1,017 pairs belong to different objects.

The mean cosine similarity is 0.32 for within-object pairs and 0.04 for cross-object pairs. Thus, residual directions show substantially stronger local alignment within an object context than across objects.

B.4 Transfer Across Anomaly Contexts

Finally, we test whether residual evidence learned from one anomaly context transfers to a previously unseen defect. This experiment uses a leave-one-defect-out protocol that differs from the within-defect splits above.

For each held-out defect, normal images of its object are split 70/30 using the same five deterministic random states. Because these samples are all normal, this split is not stratified. All anomalous images of the held-out defect are excluded from training and are used at test time. The test set

therefore consists of the held-out 30% normal images together with all anomalous images of the held-out defect.

We compare three anomaly-training pools:

- **Same-object (LODO):** the 70% training normal images from the target object together with all anomalous images from the other defect types of the same object;
- **Heterogeneous:** the Same-object pool augmented with all images from the remaining object categories, using their original normal/anomalous labels;
- **Cross-object:** only images from the remaining object categories, again using their original normal/anomalous labels.

All three conditions use the same StandardScaler and ℓ_2-regularized logistic-classifier pipeline described above. We use neither class weighting nor sample weighting, and the three training pools are not matched for sample count. In particular, the Same-object condition uses the smallest anomaly training pool.

For Same-object and Heterogeneous conditions, the five repeats differ in the 70/30 split of target-object normal images. For Cross-object, the source training pool itself is unchanged across repeats; the repeats differ through the target normal test split and the random state used by the classifier's inner cross-validation.

The analysis includes the 47 defect types for which at least one sibling defect exists in the same object category. `toothbrush/defective` is excluded because no same-object sibling defect is available.

For each held-out defect and transfer condition, Image AUROC is first averaged over the five repeats. Figure 2(d) shows the resulting paired defect-level values. Aggregate means and error bars are then computed across the 47 held-out defects; the error bars denote the standard error across defects, not across the five repeats.

Mean Image AUROC is 0.954 for Same-object transfer, 0.897 for the Heterogeneous condition, and 0.709 for Cross-object transfer. Although the training pools are not sample-count matched, the Same-object condition uses the least training data while achieving the strongest transfer. We therefore interpret the result as evidence that residual evidence transfers more effectively within a compatible object context, rather than as a strictly sample-matched causal isolation of context mismatch.

C Probe Ablations and Robustness

C.1 Random-Seed Robustness

Static and Dynamic probes use identical frozen host checkpoints, Real-IAD-SV data, training objectives, optimization settings, and $\rho = 0.025$. The only controlled difference is whether the reader uses a shared learned context or frozen features from the current image.

Table 6 reports all three training seeds underlying the mean and standard deviation values in the main paper.

Across seeds, Dynamic improves over the frozen Native detector for AnomalyCLIP, APRIL-GAN, MoECLIP, and FAPrompt in every run. VCP-CLIP shows positive gains in one seed and essentially neutral results in the other two, while VisualAD remains unchanged. The larger seed variation of APRIL-GAN is reflected in the standard deviation reported in the main table.

C.2 Residual-Budget Sensitivity

The residual budget ρ controls the maximum magnitude of the readout correction. Small values enforce a conservative intervention, whereas overly large corrections may interfere with a competent native decision pathway. We therefore evaluate several residual budgets without changing the remaining training protocol.

Table 6: Seed-level Image AUROC (%) on MVTec AD. Mean and standard deviation are computed over seeds 111, 222, and 333.

		Static				Dynamic			
Host	Native	111	222	333	Mean±Std	111	222	333	Mean±Std
AnomalyCLIP	91.55	90.77	90.91	91.25	90.98 ± 0.25	92.56	91.95	92.06	92.19 ± 0.33
APRIL-GAN	86.16	87.20	87.03	87.78	87.34 ± 0.39	87.37	89.23	89.30	88.63 ± 1.09
VisualAD	92.23	92.23	92.23	92.23	92.23 ± 0.00	92.23	92.23	92.23	92.23 ± 0.00
MoECLIP	90.72	90.31	90.74	90.07	90.37 ± 0.34	91.22	91.13	91.27	91.21 ± 0.07
FAPrompt	91.80	91.99	91.90	91.62	91.84 ± 0.19	92.16	92.01	91.92	92.03 ± 0.12
VCP-CLIP	89.79	89.50	89.61	89.23	89.45 ± 0.20	90.19	89.79	89.79	89.92 ± 0.23

Table 7: Sensitivity to the residual budget ρ using reader seed 111. Values report Image AUROC (%) on MVTec AD. Parentheses denote the change relative to the frozen Native host.

Host	Native	$\rho = 0.01$	$\rho = 0.025$	$\rho = 0.05$	$\rho = 0.10$
AnomalyCLIP	91.550	91.954 $(+0.404)$	92.560 $(+1.010)$	91.594 $(+0.044)$	92.669 $(+1.119)$
APRIL-GAN	86.157	90.179 $(+4.022)$	87.370 $(+1.213)$	88.747 $(+2.590)$	89.711 $(+3.554)$
MoECLIP	90.719	91.353 $(+0.634)$	91.220 $(+0.501)$	91.696 $(+0.976)$	91.105 $(+0.386)$
FAPrompt	91.795	91.916 $(+0.121)$	92.160 $(+0.365)$	92.052 $(+0.256)$	91.701 (-0.094)
VisualAD	92.227	92.227 $(+0.000)$	92.230 $(+0.003)$	92.228 $(+0.001)$	92.225 (-0.002)
VCP-CLIP	89.785	89.773 (-0.012)	90.190 $(+0.405)$	89.730 (-0.055)	89.784 (-0.001)

Table 7 reports the sensitivity analysis using reader seed 111. The $\rho = 0.025$ column corresponds to the seed-111 setting used in the main experiment.

The effect of ρ is host dependent rather than monotonic. APRIL-GAN and MoECLIP remain above their native baselines throughout the tested range, while FAPrompt benefits from small-to-moderate corrections and slightly degrades at $\rho = 0.10$. VisualAD remains effectively unchanged across the sweep. VCP-CLIP exhibits only a small and non-monotonic response outside the main $\rho = 0.025$ setting. These results support treating ρ as a bound on intervention strength rather than as a universally optimal performance parameter.