0-CLIP: TRAINING-FREE ANOMALY DETECTION HIDDEN IN CLIP'S PROJECTOR

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP-based zero-shot anomaly detection commonly improves dense discriminability through prompts or adapters trained on auxiliary data. However, reliable transfer to unseen domains remains a practical challenge. This motivates a closer examination of how much anomaly information is already present in CLIP itself. Our preliminary experiments show that CLIP's frozen visual projector $\boldsymbol{P}$ contains useful dense anomaly structure, but its patch-level scores tend to rank normal and anomalous regions in the wrong order. Based on this observation, we introduce **0-CLIP**, a training-free method that converts $\boldsymbol{P}$ into a patch-to-text anomaly operator. The construction reverses the normal–abnormal score polarity, forms a compact operator from the principal singular directions of $\boldsymbol{P}$, and selectively recovers lower-energy directions using a separation criterion computed on the first unlabeled test image of each category. Without any labeled data, the resulting operator yields both pixel-level anomaly maps and image-level anomaly scores. Across industrial and medical benchmarks, 0-CLIP achieves performance competitive with training-based methods, and ablation studies show the effectiveness of the proposed method. These results suggest that substantial dense anomaly discriminability is already latent in CLIP's frozen visual projector and can be exposed without training. Our code is available in the Supplementary Material.

1 INTRODUCTION

Anomaly detection (AD) aims to identify samples or regions that deviate from normal patterns (Pang et al., 2021; Ruff et al., 2021), and is critical in industrial inspection (Bergmann et al., 2019; Zou et al., 2022; Qu et al., 2023) and medical imaging (Huang et al., 2024; Bao et al., 2024). Conventional AD methods learn target-specific normality through reconstruction and generative modeling (Zavrtanik et al., 2021; He et al., 2024), teacher–student feature distillation (Bergmann et al., 2020; Deng & Li, 2022; Zhang et al., 2023), explicit distribution modeling (Rudolph et al., 2022; Gudovskiy et al., 2022), feature matching (Cohen & Hoshen, 2020; Defard et al., 2021; Roth et al., 2022), and self-supervised learning with synthetic anomalies (Li et al., 2021; Zavrtanik et al., 2021; Chen et al., 2024a). Labeled data in the target domain are often scarce or unavailable, motivating interest in zero-shot AD (ZSAD) (Zhou et al., 2024; Chen et al., 2026). However, recognizing and localizing previously unseen anomalies without target-domain supervision remains challenging.

Vision–language models such as CLIP (Radford et al., 2021) offer a natural foundation for ZSAD: their shared image–text space makes it possible to compare visual features directly with semantic prototypes such as "normal" and "abnormal" (Jeong et al., 2023). Yet CLIP is pretrained for global image–text retrieval, whereas anomaly detection demands fine-grained sensitivity to subtle local deviations. Early training-free methods bridge this gap by re-encoding multi-scale sliding windows (Jeong et al., 2023), at the cost of heavy inference overhead, or by modifying the encoder's attention (Chen et al., 2024b).

Recent methods turn to auxiliary labeled datasets to establish dense patch-level alignment, proposing various designs to enhance domain generalization, such as object-agnostic prompts (Zhou et al., 2024; Cao et al., 2024), controlled visual adaptation (Ma et al., 2025), foreground–background disentanglement (Hu et al., 2026), and retrieval from a feature-label memory bank (Xu et al., 2026). Despite these efforts, even the latest detectors still struggle with cross-domain generalization, suggesting that models trained on specific source datasets can inherit strong source-domain biases.

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.

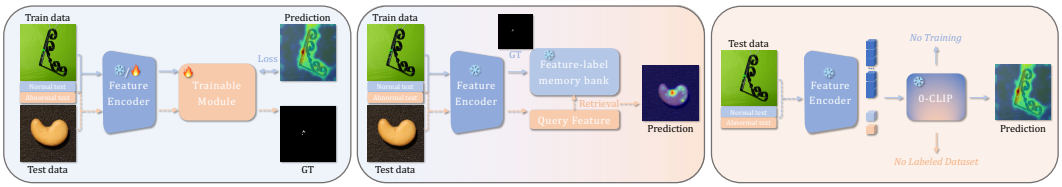


Figure 1: **Patch-to-text anomaly modeling with CLIP.** *(Left)* CLIP-based AD methods learn patch-to-text alignment with prompts or adapters using auxiliary data. *(Center)* MRAD (Xu et al., 2026) models anomalies by retrieving reference features from its feature-label memory bank, which is built from a separate labeled auxiliary dataset. *(Right)* 0-CLIP keeps the whole pipeline frozen and predicts directly from the test data, requiring neither training nor a labeled auxiliary dataset.

Indeed, our cross-source comparisons confirm that switching the training domain causes noticeable performance instability on unseen targets (Table 4 in Appendix B.1). This persistent source bias, along with the heavy annotation burden, motivates a training-free path and raises a fundamental question: **how much dense anomaly discriminability is already encoded by frozen CLIP itself?**

We answer this question at the level of CLIP's visual projector. CLIP contains a fixed linear map $\boldsymbol{P} \in \mathbb{R}^{d_t \times d_v}$ that projects its global visual token into the shared image–text space. When this map is reused for patch tokens, however, its anomaly scores tend to rank normal and anomalous regions in the wrong order. Rather than learning a replacement, we derive a patch-to-text anomaly operator directly from $\boldsymbol{P}$. The resulting method, **0-CLIP**, is training-free and requires neither parameter optimization nor any labeled data. Its construction progressively adapts the retrieval projector to dense AD (Figures 3–5). First, negating $\boldsymbol{P}$ reverses its cosine similarities to the text prototypes, correcting the normal–abnormal score polarity. Second, we keep only the leading singular directions of $\boldsymbol{P}$, obtained by a thin SVD, to form a compact operator. Third, we examine nearby lower-energy directions and retain only those that improve separation on the first unlabeled test image. This selection is performed separately for each feature layer, producing layer-specific operators whose scores are combined for pixel- and image-level detection.

In summary, our contributions are:

- We show empirically that CLIP's frozen visual projector already carries dense anomaly information: its patch-level scores are consistently polarity-reversed, and its singular directions can be reweighted and reselected into an operator suited to dense AD.
- Building on this, we propose **0-CLIP**, a training-free method that constructs layer-specific patch-to-text operators for AD: it reverses the projector's score polarity, forms a compact operator from the principal singular directions, and selectively recovers lower-energy directions using a label-free separation criterion on the first test image.
- Through extensive experiments on industrial and medical benchmarks, we show that **0-CLIP** achieves performance comparable to training-based methods in both offline and online settings, without training in either. Ablation studies further validate the effectiveness of each component in our construction.

2 Related Work

2.1 CLIP-based Anomaly Detection

Large vision–language models (Radford et al., 2021; Sun et al., 2023; Zhai et al., 2023) enable zero-shot AD by matching image regions with normal and abnormal text in a shared embedding space, as pioneered by WinCLIP (Jeong et al., 2023). The core difficulty is that CLIP is trained for image-level retrieval, so its patch tokens carry weak anomaly semantics. WinCLIP sidesteps this by re-encoding multi-scale windows as standalone images, keeping the model frozen but paying for it in stride-limited resolution and inference cost. CLIP Surgery (Li et al., 2025) and CLIP-AD (Chen et al., 2024b) observe reversed patch–text similarities and correct the patch features with value–value attention, leaving the projector unchanged. APRIL-GAN instead trains linear projectors that map patch tokens into the text space (Chen et al., 2023). Most subsequent work follows this route of scoring patch tokens directly, closing the alignment gap with components trained on auxiliary annotations: learned prompts (Zhou et al., 2024; Cao et al., 2024; Qu et al., 2024; Zhu et al., 2025;

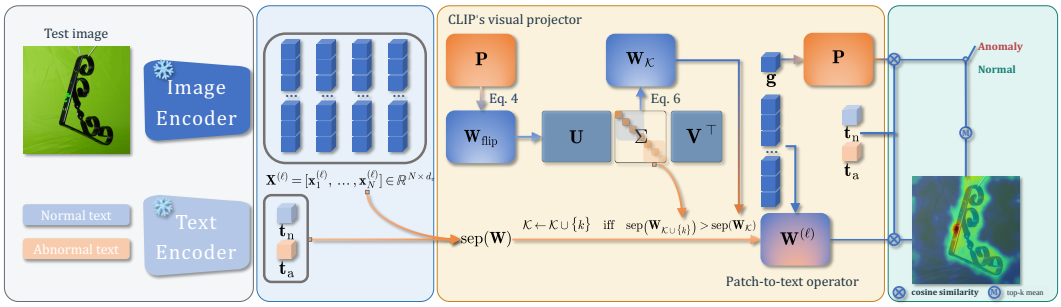


Figure 2: **Overview of 0-CLIP.** Three closed-form corrections turn CLIP's frozen visual projector into an anomaly operator, flipping the polarity (Sec. 3.2), truncating to the inverse-weighted principal directions (Sec. 3.3), and recovering a few weak directions selected without labels (Sec. 3.4). Scoring patch tokens against the text prototypes then gives the anomaly map and the image score.

Qu et al., 2025; Zhang & Zhang, 2026), lightweight adapters over frozen features (Ma et al., 2025; Fang et al., 2025; Gao et al., 2026), patch-specialized experts (Park et al., 2026), and foreground–background disentanglement (Hu et al., 2026). MRAD retrieves labeled reference features from its feature-label memory bank, built from a separate labeled auxiliary dataset (Xu et al., 2026). These approaches depend on auxiliary labeled data to establish patch-level alignment. In contrast, 0-CLIP derives the patch-to-text operator in closed form from CLIP's frozen projector.

2.2 Online Zero-Shot Anomaly Detection

Normal-reference AD is distinct from both zero-shot protocols considered here. PatchCore (Roth et al., 2022) constructs a memory bank from target-domain normal images and keeps it fixed at inference, as do few-shot variants (Zhu & Pang, 2024). Offline zero-shot AD scores images independently after initialization. Batch methods, including ACR (Li et al., 2023) and MuSc (Li et al., 2024; 2026), exploit cross-image rarity but need test samples in advance. Online zero-shot methods such as RareCLIP (He et al., 2025) and ReMem (Yi et al., 2026) instead update reference information from past unlabeled test images, requiring no pre-collected normal support set. Their goal is to estimate rarity in the test stream, not deviation from a fixed normal reference set. 0-CLIP supports offline and online zero-shot inference without retraining.

3 Method

3.1 Preliminaries

Given an image, CLIP's visual encoder produces a global token $\boldsymbol{g} \in \mathbb{R}^{d_v}$ and patch tokens,

$$\boldsymbol{X}^{(\ell)} = [\boldsymbol{x}_1^{(\ell)}, \ldots, \boldsymbol{x}_N^{(\ell)}] \in \mathbb{R}^{N \times d_v}, \qquad \ell \in \mathcal{L}, \tag{1}$$

where $\mathcal{L}$ denotes the selected backbone layers. CLIP's visual projector is a linear map $\boldsymbol{P} \in \mathbb{R}^{d_t \times d_v}$ that sends the global token into the shared image–text space, $\boldsymbol{Pg} \in \mathbb{R}^{d_t}$. This space also contains two prompt-ensemble prototypes, $\boldsymbol{t}_{\mathrm{n}}, \boldsymbol{t}_{\mathrm{a}} \in \mathbb{R}^{d_t}$, representing "normal" and "abnormal". Because CLIP provides no layer-specific projector for patch tokens, each layer requires a patch-to-text map $\boldsymbol{W}^{(\ell)} \in \mathbb{R}^{d_t \times d_v}$. We score each patch by contrasting its similarity to the two prototypes with a temperature-τ softmax, then average across layers:

$$s_i = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \frac{\exp(\left\langle \boldsymbol{W}^{(\ell)} \boldsymbol{x}_i^{(\ell)}, \boldsymbol{t}_{\mathrm{a}} \right\rangle / \tau)}{\exp(\left\langle \boldsymbol{W}^{(\ell)} \boldsymbol{x}_i^{(\ell)}, \boldsymbol{t}_{\mathrm{n}} \right\rangle / \tau) + \exp(\left\langle \boldsymbol{W}^{(\ell)} \boldsymbol{x}_i^{(\ell)}, \boldsymbol{t}_{\mathrm{a}} \right\rangle / \tau)} \in [0, 1], \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity. The resulting field $\{s_i\}$ forms the pixel-level anomaly map. Most CLIP-based methods learn each map $\boldsymbol{W}^{(\ell)} = \boldsymbol{W}_\theta^{(\ell)}$ as a linear layer using pixel-level labels from an auxiliary dataset. For image-level detection, the global cue s_{g} is computed with the same softmax using $\boldsymbol{Pg}$. It is commonly combined with the strongest localized cue:

$$S_{\mathrm{cls}} = \lambda\, s_{\mathrm{g}} + (1 - \lambda) \max_i s_i. \tag{3}$$

0-CLIP constructs these maps in a training-free manner. Sections 3.2, 3.3, and 3.4 describe the construction (Figure 2), and Section 3.5 presents the inference procedure for AD. The diagnostic figures accompanying these three steps follow the protocols in Appendix C.

3.2 FLIP THE POLARITY

CLIP's visual projector induces the wrong score polarity for AD, so we reverse it by negating the projector. CLIP's visual projector $\boldsymbol{P}$ was trained to map a global image feature near its matching caption in the shared embedding space. When it is reused at the patch level, we observe that anomalous patches tend to score below normal ones, as also shown in prior work (Li et al., 2025; Chen et al., 2024b). Thus, we initialize $\boldsymbol{W}_{\text{flip}}$ by negating $\boldsymbol{P}$,

$$\boldsymbol{W}_{\text{flip}} \;=\; -\boldsymbol{P}. \tag{4}$$

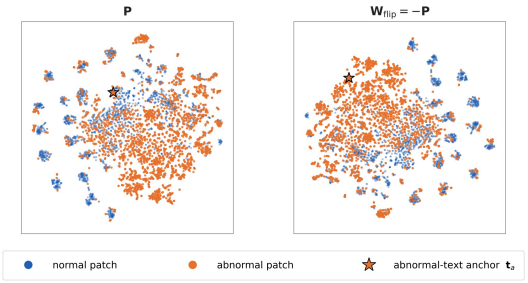


Figure 3: **Flip the polarity.** Negating $\boldsymbol{P}$ reverses patch alignment with the abnormal-text anchor $\boldsymbol{t}_{\text{a}}$, placing abnormal patches closer.

Negating a projected patch reverses its cosine similarity to both text prototypes. Writing $p_i(\boldsymbol{W})$ for the single-layer anomaly probability under the binary normal–abnormal softmax in equation 2, negation gives $p_i(-\boldsymbol{W}) = 1 - p_i(\boldsymbol{W})$ for every patch, so patches that the forward map scores as most normal become the most anomalous, without learned parameters (Figure 3). In our ablations, this sign flip provides the largest individual improvement, suggesting that $\boldsymbol{P}$ already contains useful discriminative structure but with a polarity poorly matched to anomaly detection. This polarity reversal persists across layers and classes on both MVTec-AD and VisA (Appendix Figures 8–11).

3.3 REDUCE TO THE PRINCIPAL DIRECTIONS

We retain a compact set of rebalanced principal directions. Negation changes only the sign of $\boldsymbol{P}$: every direction of the retrieval map survives, carrying the same weight it had for retrieval. Those weights are highly uneven, so a few directions drive most of each patch score, while the many weak ones contribute unstable, task-irrelevant variation. We expose this spectral structure through the thin SVD

$$\boldsymbol{P} \;=\; \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}, \qquad \boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1 \geq \cdots \geq \sigma_r). \tag{5}$$

For any retained index set $\mathcal{K}$, we construct the orientation-preserving, rank-selective operator

$$\boldsymbol{W}_{\mathcal{K}} \;=\; -\sum_{k\in\mathcal{K}} \frac{1}{\sigma_k}\, \boldsymbol{u}_k \boldsymbol{v}_k^{\top}, \tag{6}$$

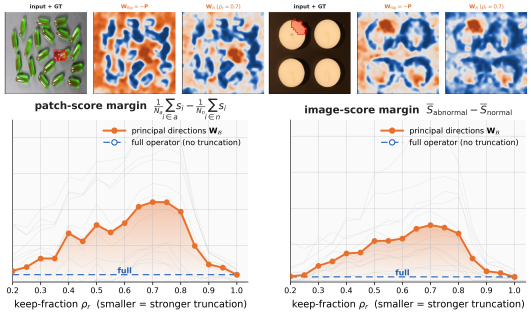


Figure 4: **Reduce to the principal directions.** Retaining the leading components suppresses diffuse background responses and widens patch- and image-score gaps.

where $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are the k-th columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively, and σ_k is the corresponding singular-value gain. This operator has two knobs. The index set $\mathcal{K}$ selects which singular directions survive; we initialize it with the leading block $\mathcal{B} = \{1, \ldots, n_r\}$, where $n_r = \lfloor \rho_r\, r \rfloor$, and $\boldsymbol{W}_{\mathcal{B}}$ for the operator it induces. The reciprocal weights $1/\sigma_k$ then rebalance the survivors, reversing the original gain ordering so that the largest retrieval components no longer dominate the retained subspace. If all directions are retained, $\boldsymbol{W}_{\mathcal{K}}$ is the negated transpose of the pseudoinverse $\boldsymbol{P}^{\dagger} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\top}$. Unlike $\boldsymbol{P}^{\dagger}$, which maps text back to the visual space, $\boldsymbol{W}_{\mathcal{K}}$ keeps the visual-to-text direction of $\boldsymbol{P}$ and only replaces each σ_k by $1/\sigma_k$. The resulting compact operator markedly suppresses the diffuse background responses of $\boldsymbol{W}_{\text{flip}}$ and concentrates high scores on the defective regions (Figure 4). These cleaner maps reflect a wider score gap: over a broad range of retention ratios, anomalous patches and images score farther above their normal counterparts than under the full operator, confirming the benefit of selecting and reweighting principal directions for localized AD.

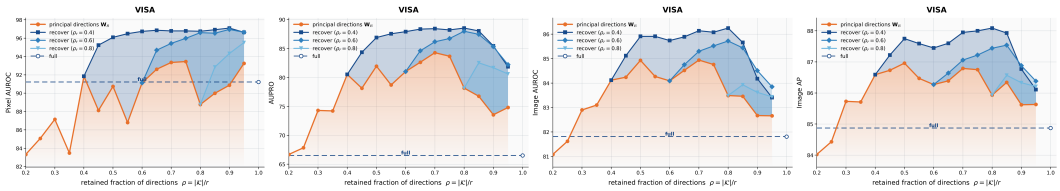


Figure 5: **Recover lower-energy directions on VisA.** Orange uses the principal block of Section 3.3 alone. The three blue curves recover directions that improve separation on the first unlabeled test image, starting from base ratios ρ_r of 0.4, 0.6 or 0.8. Shaded regions show the recovery gains, and the dashed line marks the untruncated flipped operator.

3.4 Recover lower-energy directions

We recover low-energy directions that improve separation. Truncating at n_r is convenient but blunt: σ_k ranks directions by their role in CLIP's retrieval geometry, and this ordering need not match their usefulness for AD. Some directions just beyond the retained block may still help AD despite low retrieval energy. To evaluate these candidates, we calibrate each class on the first unlabeled image of its test stream. This image is not selected by label and may be normal or anomalous. Since each evaluation seed reshuffles the stream, the calibration image varies randomly across seeds. Let $\{\boldsymbol{x}_i\}_{i=1}^{N}$ collect the patches of this calibration image. For $\boldsymbol{W}$, we define separation as

$$\mathrm{Sep}(\boldsymbol{W}) \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big(\langle \boldsymbol{W}\boldsymbol{x}_i,\, \boldsymbol{t}_{\mathrm{a}}\rangle - \langle \boldsymbol{W}\boldsymbol{x}_i,\, \boldsymbol{t}_{\mathrm{n}}\rangle\Big), \tag{7}$$

where the two terms are the cosine similarities to the abnormal and normal text prototypes, respectively. As the patches are unlabeled and may contain no defect, Sep measures the average prototype contrast that the score uses.

Starting from $\mathcal{K} = \mathcal{B}$, we scan the next $n_e = \lfloor \rho_e\, r \rfloor$ candidates in descending singular-value order and retain k iff its inclusion improves separation:

$$\mathcal{K} \leftarrow \mathcal{K} \cup \{k\} \quad \text{iff} \quad \mathrm{Sep}\big(\boldsymbol{W}_{\mathcal{K}\cup\{k\}}\big) > \mathrm{Sep}\big(\boldsymbol{W}_{\mathcal{K}}\big), \qquad k = n_r{+}1, \ldots, n_r{+}n_e. \tag{8}$$

The final 0-CLIP construction gives $\boldsymbol{W} = \boldsymbol{W}_{\mathcal{K}}$, where $\mathcal{K}$ contains the base indices and the accepted candidate directions. Thus, 0-CLIP first forms a compact principal operator whose inverse singular-value weights prevent dominant retrieval components from overwhelming the retained subspace, then augments it with lower-energy directions only when they improve separation. Figure 5 verifies this selection empirically: across all four metrics, the blue selection curves lie above the orange principal-direction baseline over most retention ratios, while the blue shaded regions mark the improvement produced by the selected weak components. These gains build on the principal block rather than replace it, since recovering from the leading direction alone saturates early (Figure 6).

3.5 Inference implementation

Anomaly score. The basis of $\boldsymbol{P}$ and the leading block $\mathcal{B}$ are shared across layers, but Sep is evaluated on the calibration patches of each layer, so equation 8 can produce a distinct index set $\mathcal{K}^{(\ell)}$ and operator $\boldsymbol{W}^{(\ell)}$. Averaging their anomaly probabilities with uniform weights as in equation 2 gives one text-branch score $s_i \in [0, 1]$ per patch. For localization these scores are reshaped to the $\sqrt{N} \times \sqrt{N}$ patch grid, bilinearly upsampled and Gaussian smoothed to obtain a map $\boldsymbol{A} \in \mathbb{R}^{H\times W}$. For classification we combine the global-token score s_{g} with the mean of the top-k pixels of $\boldsymbol{A}$,

$$S \;=\; \lambda\, s_{\mathrm{g}} \;+\; (1-\lambda)\,\frac{1}{k}\sum_{u\in \text{top-}k(\boldsymbol{A})} \boldsymbol{A}_u, \tag{9}$$

so that a cluster of confidently anomalous pixels suffices, while isolated outliers are suppressed.

Offline and online inference. Both settings are training-free. After the per-class calibration fixes $\{\boldsymbol{W}^{(\ell)}\}$, each test image is scored independently in the offline setting. For online inference, we additionally use the unmodified rarity branch of RareCLIP (He et al., 2025) to combine our text scores with rarity evidence from previously observed unlabeled images of the same class. This

Algorithm 1 0-CLIP. Training-free AD for one class.

```
# P: frozen CLIP projector; t_n, t_a: normal / abnormal text prototypes
# layers: selected layers; rho_r, rho_e: retention / recovery ratios
U, sig, Vt = svd(P); r = len(sig)                  # Eq. (5)
n_r, n_e = floor(rho_r * r), floor(rho_e * r)
def op(K): return -(U[:, K] / sig[K]) @ Vt[K]      # Eq. (6)

_, x0 = encode(stream[0]); W = {}  # calibrate on the first unlabeled image
for l in layers:
    W[l] = op(range(n_r))        # principal directions
    for k in range(n_r, n_r + n_e):
        W_try = W[l] + op([k])
        if sep(W_try, x0[l]) > sep(W[l], x0[l]): W[l] = W_try  # Eq. (8)

for img in stream:
    g, x = encode(img)
    s = mean([prob(W[l] @ x[l], t_n, t_a) for l in layers])
    A, s_g = smooth(upsample(s)), prob(P @ g, t_n, t_a)
    score = lam * s_g + (1 - lam) * mean(topk(A))  # Eq. (9)
```

branch only stores features of past test images and is itself training-free. What RareCLIP trains on labeled auxiliary data is its text branch, which our operator replaces. The constructed weights remain fixed, and online inference uses neither test labels nor future images nor a pre-collected normal support set. This optional branch is not our contribution, and its reference updates, score fusion, and default settings are detailed in Appendix A.

4 Experiments

4.1 Experimental Setup

Datasets. Following recent ZSAD work, we evaluate on 15 industrial and medical benchmarks. For the industrial domain we use MVTec-AD (Bergmann et al., 2019), VisA (Zou et al., 2022), BTAD (Mishra et al., 2021), MPDD (Ježek et al., 2021), DAGM (Wieler et al., 2007), DTD-Synthetic (Aota et al., 2023), SDD (Tabernik et al., 2020), KSDD2 (Božič et al., 2021), and Real-IAD (Wang et al., 2024). For the medical domain we use four subsets from BMAD (Bao et al., 2024): Brain MRI, Liver CT, Retina RESC, and Retina OCT, together with the brain-tumour set Br35H (Hamada, 2020) and TN3K (thyroid ultrasound) (Gong et al., 2021). Datasets are summarized in Table 8.

Evaluation metrics and baselines. We report pixel-level AUROC (P-AUROC) and per-region overlap (PRO) (Bergmann et al., 2020) for localization, and image-level AUROC (I-AUROC) and average precision (I-AP) for classification, following prior ZSAD work (Zhou et al., 2024; Jeong et al., 2023). We compare 0-CLIP with seven CLIP-based ZSAD methods: the training-free WinCLIP (Jeong et al., 2023), and six training-based methods that learn prompts or adapters on a fully labeled auxiliary dataset, namely AdaCLIP (Cao et al., 2024), AnomalyCLIP (Zhou et al., 2024), AA-CLIP (Ma et al., 2025), RareCLIP (He et al., 2025), MRAD (Xu et al., 2026) and FB-CLIP (Hu et al., 2026). Details are given in Appendix B.

Implementation details. Algorithm 1 states the whole procedure. We use a frozen OpenCLIP `ViT-L/14-336` backbone (Radford et al., 2021) with 518×518 inputs. Features from layers $\{12, 16, 20, 24\}$ are equally averaged. We set $\rho_r = 0.6$, $\rho_e = 0.1$, and construct and freeze the class-specific weights using the first test image. The global cue s_g uses CLIP's unmodified $\boldsymbol{Pg}$. The online setting† follows the official rarity-aware protocol of RareCLIP (He et al., 2025). We average results over five seeds (0–4), each randomly shuffling the test stream within each class and using its first image for calibration. All runs use a single NVIDIA RTX 4090 (24 GB). Full implementation details are provided in Appendix A.

4.2 Main Results

Tables 1 and 2 report the performance of 0-CLIP under offline and online† inference settings.

Table 1: **Pixel-level anomaly detection.** We report **P-AUROC / PRO** (%) on industrial (⚙) and medical (🏥) datasets, highlighting the best, second and third per metric. "–" marks unreported entries, excluded from averages. 0-CLIP: five-seed means, stds in Table 7.

Dataset	TRAINING-BASED						TRAINING-FREE		
	AdaCLIP ECCV'24	AnomCLIP ICLR'24	AA-CLIP CVPR'25	RareCLIP† ICCV'25	MRAD‡ ICLR'26	FB-CLIP CVPR'26	WinCLIP CVPR'23	**0-CLIP**	**0-CLIP†**
⚙ MVTec-AD	86.8 / 33.8	91.1 / 81.4	91.9 / 84.6	97.7 / 93.5	93.0 / 86.8	91.9 / 85.7	85.1 / 64.6	94.9 / 87.3	97.8 / 94.0
⚙ VisA	95.1 / 71.3	95.5 / 87.0	95.5 / 83.0	98.8 / 93.5	95.9 / 88.0	96.3 / 91.4	79.6 / 56.8	95.3 / 86.2	98.9 / 93.7
⚙ BTAD	87.7 / 17.1	93.3 / 69.3	94.0 / 69.0	97.5 / 82.2	95.4 / 72.8	95.8 / 80.0	71.4 / 32.8	96.5 / 81.7	98.0 / 83.9
⚙ MPDD	95.2 / 10.8	96.2 / 79.7	96.7 / 88.4	97.3 / 93.3	97.9 / 90.6	96.9 / 91.1	71.2 / 48.9	96.2 / 91.1	97.5 / 93.5
⚙ DTD-Syn	94.1 / 24.9	97.6 / 88.3	96.4 / 85.9	98.1 / 94.8	98.1 / 89.8	98.3 / 92.6	78.4 / 51.0	98.0 / 94.0	98.3 / 95.2
⚙ SDD	79.5 / 4.9	90.1 / 62.9	86.8 / 53.1	89.1 / 71.4	93.0 / 72.0	90.4 / 64.0	55.9 / 14.7	91.9 / 78.0	93.5 / 77.8
⚙ DAGM	76.2 / 56.3	96.5 / 88.4	91.6 / 76.5	97.0 / 94.0	97.4 / 90.3	97.3 / 94.1	75.5 / 44.4	96.9 / 94.1	98.2 / 96.4
⚙ KSDD2	85.8 / 72.9	97.9 / 94.9	99.3 / 96.1	98.5 / 96.3	98.9 / 95.6	99.1 / 97.6	75.4 / 69.2	97.3 / 94.8	98.5 / 96.2
⚙ Real-IAD	– / –	93.9 / 72.0	93.1 / 70.9	97.0 / 86.3	94.5 / 78.6	95.9 / 88.2	– / –	95.3 / 87.4	97.4 / 88.6
🏥 Liver CT	94.5 / –	94.0 / 73.4	96.8 / 92.9	97.3 / 91.7	96.2 / 83.6	96.9 / 88.2	96.2 / –	95.9 / 90.9	97.5 / 92.0
🏥 Brain MRI	93.9 / –	96.2 / 72.6	95.9 / 79.2	96.8 / 82.7	96.4 / 78.8	96.4 / 82.1	86.0 / –	93.7 / 76.3	96.9 / 81.2
🏥 Retina RESC	88.5 / –	92.5 / 70.5	94.3 / 76.6	95.9 / 84.1	93.5 / 73.9	94.8 / 76.9	80.6 / –	87.4 / 59.0	95.1 / 81.7
🏥 TN3K	77.2 / –	81.5 / 50.4	83.9 / 52.7	82.2 / 51.1	78.0 / 49.6	79.0 / 46.8	70.7 / 39.8	82.4 / 53.0	80.0 / 57.9
Average	87.9 / 36.5	93.6 / 76.2	93.6 / 77.6	95.6 / 85.8	94.5 / 80.8	94.5 / 83.0	77.2 / 46.9	94.0 / 82.6	96.0 / 87.1

† Online inference using unlabeled test-stream observations. ‡ Feature-label memory bank built from a separate labeled auxiliary dataset.

Table 2: **Image-level anomaly detection.** Cells report **I-AUROC / I-AP** (%) on industrial (⚙) and medical (🏥) datasets, highlighting the best, second and third per metric. "–" marks unreported entries, excluded from averages. 0-CLIP: five-seed means, stds in Table 7.

Dataset	TRAINING-BASED						TRAINING-FREE		
	AdaCLIP ECCV'24	AnomCLIP ICLR'24	AA-CLIP CVPR'25	RareCLIP† ICCV'25	MRAD‡ ICLR'26	FB-CLIP CVPR'26	WinCLIP CVPR'23	**0-CLIP**	**0-CLIP†**
⚙ MVTec-AD	92.0 / 96.4	91.5 / 96.2	90.5 / 94.9	98.2 / 99.3	94.0 / 97.4	92.4 / 96.6	91.8 / 95.1	94.4 / 97.4	98.2 / 99.3
⚙ VisA	83.0 / 84.9	82.1 / 85.4	84.6 / 82.2	94.4 / 95.3	85.7 / 88.3	89.5 / 90.7	78.1 / 77.5	85.3 / 87.0	92.9 / 93.7
⚙ BTAD	91.6 / 92.4	89.1 / 91.1	93.8 / 97.9	95.9 / 98.8	92.8 / 94.2	93.2 / 96.4	83.3 / 84.1	95.0 / 94.6	96.5 / 97.9
⚙ MPDD	76.4 / 80.4	73.7 / 76.5	73.9 / 79.7	75.8 / 81.3	81.8 / 83.4	79.1 / 86.2	61.5 / 69.2	76.8 / 81.8	73.3 / 77.5
⚙ DTD-Syn	92.8 / 97.0	94.5 / 97.7	93.3 / 97.8	98.1 / 99.3	96.0 / 98.4	97.9 / 99.1	63.8 / 83.5	98.0 / 99.3	98.0 / 99.3
⚙ SDD	77.5 / 65.7	82.7 / 73.8	63.8 / 41.0	83.4 / 72.6	83.9 / 76.2	80.8 / 72.0	54.8 / 37.7	84.5 / 78.4	85.6 / 76.2
⚙ DAGM	77.6 / 72.7	98.0 / 97.7	93.9 / 84.5	98.8 / 93.6	98.4 / 98.6	99.0 / 96.5	58.5 / 59.1	99.5 / 98.3	99.6 / 96.7
⚙ KSDD2	90.5 / 77.7	90.8 / 76.5	94.1 / 87.1	94.4 / 88.0	95.1 / 88.9	97.3 / 90.5	77.5 / 43.9	95.5 / 89.6	96.4 / 91.8
⚙ Real-IAD	– / –	68.7 / 64.2	65.8 / 61.4	76.7 / 72.5	73.8 / 70.9	80.6 / 78.4	– / –	76.3 / 74.2	80.8 / 79.0
🏥 Brain MRI	– / –	69.8 / 90.2	75.0 / 92.6	69.3 / 89.4	84.6 / 95.9	71.1 / 90.7	– / –	80.4 / 94.7	74.0 / 91.1
🏥 Br35H	95.7 / 95.7	97.3 / 96.1	89.4 / 91.0	91.6 / 88.0	97.9 / 97.6	97.1 / 97.4	80.5 / 82.2	96.1 / 94.6	89.5 / 83.8
🏥 Retina OCT	82.7 / –	63.5 / 86.4	79.5 / 93.5	92.3 / 97.3	86.8 / 95.3	84.2 / 94.7	42.5 / –	79.3 / 92.0	88.1 / 95.9
Average	86.0 / 84.8	83.5 / 86.0	83.1 / 83.6	89.1 / 89.6	89.2 / 90.4	88.5 / 90.8	69.2 / 70.3	88.4 / 90.2	89.4 / 90.2

† Online inference using unlabeled test-stream observations. ‡ Feature-label memory bank built from a separate labeled auxiliary dataset.

Competitive performance without auxiliary training. Without any auxiliary training data, 0-CLIP† performs comparably to the strongest training-based methods and obtains the highest average on three of the four metrics. Against RareCLIP†, which shares its rarity branch but trains its text branch on labeled auxiliary data, our training-free operator slightly improves all four averages. Qualitative comparisons in Appendix Figure 12 and E.5 show that 0-CLIP responds to the annotated defects, while the online variant further suppresses spurious responses on normal backgrounds. These results suggest that competitive ZSAD does not necessarily require thousands of labeled anomaly samples.

The offline operator alone is competitive. In the offline setting, 0-CLIP outperforms established training-based detectors such as AnomalyCLIP and AA-CLIP across all four average metrics, while trailing the leading FB-CLIP by less than one point. Among training-free methods (Appendix Table 5), offline 0-CLIP leads WinCLIP, CLIP-AD and MRAD-TF on all four average metrics.

Table 3: **Ablation studies**. (a) adds construction steps. (b) compares weights σ_k, 1, and $1/\sigma_k$ for fixed retained directions. (c) compares recovery objectives; random recovery is count-matched to separation (MVTec / VisA). Results use seed 0 in (a,b) and means over seeds 0–4 in (c).

(a) **Operator construction**

Data	Operator	P-AUROC	PRO	I-AUROC	I-AP
MVTec	CLIP projector	14.35	1.22	90.52	95.73
	Flipped polarity	85.65	84.22	92.66	96.74
	+ principal block	92.48	82.94	94.07	97.19
	+ recovered dir.	**95.16**	**87.78**	**94.59**	**97.22**
VisA	CLIP projector	15.94	2.02	79.57	83.51
	Flipped polarity	84.06	79.25	84.24	86.67
	+ principal block	91.13	81.02	84.09	86.27
	+ recovered dir.	**95.43**	**86.17**	**85.30**	**87.06**

(b) **Inverse singular-value reweighting**

Weight	P-AUROC	PRO	I-AUROC	I-AP
σ_k	93.98 / 91.73	84.90 / 81.95	94.33 / 84.56	96.93 / 86.64
1	94.80 / 93.92	87.13 / 84.37	94.50 / 84.67	97.17 / 86.75
$1/\sigma_k$	**95.16 / 95.43**	**87.78 / 86.17**	**94.59 / 85.30**	**97.22 / 87.06**

(c) **Direction-recovery objective**

Data	Objective	P-AUROC	PRO	I-AUROC	I-AP
MVTec	**Separation**	**94.91**	**87.30**	**94.44**	97.35
	Abnormal pull	94.63	86.80	94.42	**97.38**
	Normal push	93.42	85.35	94.40	97.35
	Cosine prox.	94.46	86.49	94.29	97.33
	Euclid. prox.	94.64	86.81	94.33	97.31
	Low entropy	93.30	85.28	94.21	97.32
	Random	93.68	85.46	94.23	97.26
VisA	**Separation**	**95.32**	**86.16**	**85.29**	**86.98**
	Abnormal pull	94.39	85.18	84.99	86.79
	Normal push	91.44	81.91	84.62	86.63
	Cosine prox.	93.82	84.74	84.86	86.68
	Euclid. prox.	92.85	83.97	84.89	86.69
	Low entropy	81.43	73.57	83.75	86.09
	Random	93.07	83.40	84.68	86.57

4.3 Method Ablations

Operator construction. Table 3a shows that flipping the polarity (Sec. 3.2) yields the largest gain, while retaining the principal directions (Sec. 3.3) and recovering lower-energy directions (Sec. 3.4) each provide substantial further improvements.

Inverse singular-value reweighting. Table 3b fixes the retained directions and varies only their weights. On both datasets, every metric improves monotonically from CLIP's original gains σ_k to equal weights and then to the inverse weights $1/\sigma_k$ of Sec. 3.3. Equalizing the gains removes the dominance of high-gain retrieval components, and inverting them lets retained lower-energy directions contribute even more to anomaly discrimination. Thus, selecting useful directions and adjusting their relative influence are both important.

Direction-recovery objective. Table 3c shows that the separation criterion (Sec. 3.4) delivers the strongest overall performance relative to other criteria. This supports using the difference between cosine similarities to the abnormal and normal text prototypes, consistent with the anomaly score. Dropping the sign and merely pulling patches towards the text prototypes (cosine and Euclidean proximity) recovers part of the gain but no alternative is consistent across both datasets. Random recovery already improves localization over the principal block, supporting a benefit from restored spectral capacity; separation improves further at the same recovery count through direction selection. Appendix A.1 defines the alternative criteria.

Recovery without the principal block. Figure 6 shows that recovery using the separation criterion (Sec. 3.4) from the leading direction alone yields strong initial results but saturates early. Starting from the principal block (Sec. 3.3) instead enables recovery to reach higher performance (Figure 5). The two mechanisms are therefore complementary: the principal block provides a compact, rebalanced reference operator, while recovery adds selected directions beyond the principal block. Direction recovery benefits from this structured starting point and does not replace it.

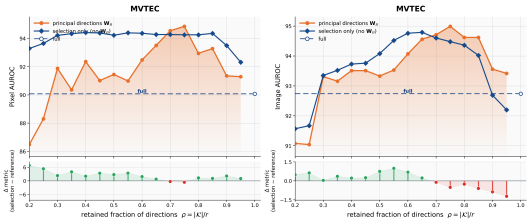


Figure 6: **Recovery without the principal block on MVTec-AD.** Orange is the principal block of Sec. 3.3, blue recovers from the leading direction alone using the separation criterion (Sec. 3.4), and the lower panels show their difference. Recovery is strong at small ρ but saturates.

4.4 Parameter Sensitivity

Retention and recovery ratios. Figure 5 shows that recovering lower-energy directions (Sec. 3.4) from the candidate window controlled by ρ_e improves on the principal block (Sec. 3.3) at all three

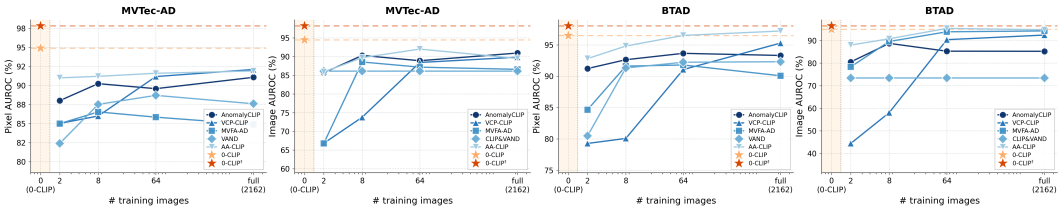


Figure 7: **Training-free 0-CLIP versus training-based methods.** Image- and pixel-level AUROC on MVTec-AD and BTAD as the auxiliary training set grows from 2 labeled images to the full set of 2162. Stars and dashed lines mark 0-CLIP, which uses none.

initial retention ratios ρ_r, indicating that singular-value ordering alone does not capture the usefulness of a direction for anomaly detection. Appendix C.1 details the sweep setup.

4.5 Dependence on Auxiliary Training Data

Auxiliary training data scale. Figure 7 plots detection performance on MVTec-AD and BTAD as a function of the auxiliary training set size for leading CLIP detectors (Chen et al., 2023; Zhou et al., 2024; Qu et al., 2024; Huang et al., 2024; Ma et al., 2025). While training-based methods rely on hundreds or thousands of annotated images to saturate, 0-CLIP matches or surpasses their fully-trained counterparts without using any training data.

Sensitivity to the training source. For FB-CLIP and MRAD-CLIP, switching between MVTec-AD- and VisA-trained checkpoints causes marked performance swings across target datasets (Table 4, Appendix B.1), with neither source consistently superior. This indicates that auxiliary-trained detectors can inherit source-domain biases, which 0-CLIP avoids because it uses no source data.

5 Discussion

By deriving anomaly operators directly from frozen CLIP and the first unlabeled test image, 0-CLIP avoids both auxiliary data collection and source checkpoint selection (Section 4.5). Its competitive performance with auxiliary-trained detectors suggests that frozen CLIP already encodes substantial dense anomaly discriminability, without requiring auxiliary supervision. The construction also carries over to SigLIP (Zhai et al., 2023), which lacks a dedicated visual projector, when the token-to-embedding map of its attention-pooling head serves as $\boldsymbol{P}$ (Appendix E.2).

To prioritize practical deployment, we maintain a single, fixed configuration ($\rho_r = 0.6$, $\rho_e = 0.1$) across all datasets, classes, and random seeds. Developing fully unsupervised, sample-adaptive mechanisms to dynamically calibrate these ratios, as well as designing more defect-specific criteria for recovering lower-energy directions in equation 8, are promising directions for future research.

Our online variant uses RareCLIP's rarity branch (He et al., 2025) unmodified, incurring only a $1.2\times$ per-image runtime overhead with near-offline peak GPU memory. While effective overall, image-level performance declines on MPDD, Brain MRI, and Br35H. The analysis in Appendix E.6 calls for streaming references that stay reliable when normal appearance varies more than defects.

6 Conclusion

In this paper, we investigate dense anomaly discriminability in CLIP's frozen visual projector and find that it already contains useful anomaly structure, but its patch-level scores tend to rank normal and anomalous regions in the wrong order. Building on this finding, we propose **0-CLIP**, a training-free method for ZSAD. It constructs layer-specific patch-to-text operators through polarity reversal, inverse-weighted principal-direction selection, and selective recovery of lower-energy directions guided by the first unlabeled test image of each class. Experiments across 15 industrial and medical benchmarks demonstrate performance competitive with auxiliary-trained methods. Ablation studies validate the effectiveness of each component. These results highlight the anomaly information already present in CLIP's frozen projector and establish 0-CLIP as a practical alternative that requires neither auxiliary anomaly annotations nor source-specific checkpoint selection.

AI USE STATEMENT

We used generative AI tools to assist with implementing a small portion of our methods, to create the code for data visualization, to assist with translation, and to draft and edit parts of the manuscript for readability. We did not use generative AI tools for designing or providing feedback on research methodology or experiments, developing theoretical models or conceptual frameworks, formulating mathematical claims, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, proposing or refining hypotheses, or interpreting results. Generating synthetic datasets, providing critical ingredients for proving mathematical claims and assisting in the writing of proofs are not applicable to this work, and we did not use generative AI tools for any of these purposes. We have carefully reviewed all LLM-generated content: all LLM-assisted text was examined and revised in detail by the authors, and all LLM-assisted code was manually inspected and tested for correctness. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The operator construction and inference procedure are described in Section 3 and Algorithm 1. Implementation settings, baseline evaluation protocols, and variability across seeds are provided in Appendices A, B, and E.3, respectively. Our code is available in the Supplementary Material.

REFERENCES

Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5564–5572, 2023.

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. BMAD: Benchmarks for medical anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4042–4053, 2024.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4183–4192, 2020.

Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459, 2021.

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*, pp. 55–72. Springer, 2024.

Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In *European Conference on Computer Vision*, pp. 37–54. Springer, 2024a.

Qiyu Chen, Zhen Qu, Wei Luo, Haiming Yao, Yunkang Cao, Yuxin Jiang, Yinan Duan, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. CoPS: Conditional prompt synthesis for zero-shot anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, 2026.

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

Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, and Yong Liu. CLIP-AD: A language-guided staged dual-path model for zero-shot anomaly detection. In *Communications in Computer and Information Science (ADFM Workshop at IJCAI)*, pp. 17–33, 2024b.

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

Thomas Defard, Aleksandr Setkov, Angélique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pp. 475–489. Springer, 2021.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9737–9746, 2022.

Qingqing Fang, Wenxi Lv, and Qinliang Su. AF-CLIP: Zero-shot anomaly detection via anomaly-focused CLIP adaptation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM)*, pp. 4846–4855, 2025.

Bin-Bin Gao, Yue Zhou, Jiangtao Yan, Yuezhi Cai, Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, and Chengjie Wang. AdaptCLIP: Adapting CLIP for universal visual anomaly detection. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 4095–4103, 2026.

Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 257–261, 2021. doi: 10.1109/ISBI48211.2021.9434087.

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 98–107, 2022.

Ahmed Hamada. Br35H :: Brain tumor detection 2020. Kaggle, 2020. URL `https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection`.

Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. DiAD: A diffusion-based framework for multi-class anomaly detection. In *AAAI Conference on Artificial Intelligence*, 2024.

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

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. *arXiv preprint arXiv:2603.19608*, 2026.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11375–11385, 2024.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19606–19616, 2023.

Štěpán Ježek, Martin Jonák, Radim Burget, Pavel Dvořák, and Miloš Škoták. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pp. 66–71. IEEE, 2021.

Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. Zero-shot anomaly detection via batch normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9664–9674, 2021.

Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. MuSc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *International Conference on Learning Representations (ICLR)*, 2024.

Xurui Li, Feng Xue, and Yu Zhou. MuSc-V2: Zero-shot multimodal industrial anomaly classification and segmentation with mutual scoring of unlabeled samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48:10713–10727, 2026.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025.

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

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: A vision transformer network for image anomaly detection and localization. In *IEEE International Symposium on Industrial Electronics (ISIE)*, pp. 01–06, 2021.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

Jun Yeong Park, JunYoung Seo, Minji Kang, and Yu Rang Park. MoECLIP: Patch-specialized experts for zero-shot anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Zhen Qu, Xian Tao, Fei Shen, Zhengtao Zhang, and Tao Li. Investigating shift equivalence of convolutional neural networks in industrial defect segmentation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–17, 2023.

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*, pp. 301–317. Springer, 2024.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14318–14328, 2022.

Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1088–1097, 2022.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020.

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.

Matthias Wieler, Tobias Hahn, and Fred A. Hamprecht. Weakly supervised learning for industrial optical inspection. DAGM 2007 Competition Dataset, 29th Annual Symposium of the German Association for Pattern Recognition, 2007.

Chaoran Xu, Chengkan Lv, Qiyu Chen, Feng Zhang, and Zhengtao Zhang. MRAD: Zero-shot anomaly detection with memory-driven retrieval. In *International Conference on Learning Representations (ICLR)*, 2026.

Ling Yi, Zhe Chen, Gaochang Wu, Jinliang Ding, Xiaojie Wang, and Zhaolong Ning. ReMem: A dynamic memory evolution detector for zero-shot anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, 2026.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM: A discriminatively trained model for surface anomaly detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Gaowei Zhang and Lihe Zhang. DLVP-CLIP: Enhancing fine-grained zero-shot anomaly detection via dynamic local visual prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. DeSTSeg: Segmentation guided denoising student-teacher for anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3914–3923, 2023.

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

Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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

A Implementation Details

This appendix collects the settings of the inference procedure summarized in Section 3.5.

Backbone and visual projector. We load the original OpenAI CLIP weights for `ViT-L/14-336` through OpenCLIP and keep the backbone frozen. We use its visual projector as $\boldsymbol{P}$.[1] Each 518×518 input produces a 37×37 grid of 1,369 patches. We extract features from layers $\{12, 16, 20, 24\}$

[1] In OpenCLIP, this map is the transpose of the stored `visual.proj` matrix.

and average their anomaly probabilities uniformly. For each layer, we retain the leading $\rho_r = 0.6$ fraction of singular directions and examine the next $\rho_e = 0.1$ fraction using the class-specific separation criterion in equation 8. From the calibration image of each class, which is the first image of its test stream and is also scored like any other test image, we use six views (the original image, horizontal and vertical flips, and rotations by 90°, 180°, and 270°) and pool their patches when evaluating Sep, which makes component selection less sensitive to a single view. Augmentation is used only for this one-image calibration, and the constructed weights are frozen thereafter.

Text prototypes. For $\boldsymbol{t}_{\mathrm{n}}, \boldsymbol{t}_{\mathrm{a}}$ we reuse the prompt ensemble of RareCLIP (He et al., 2025) unchanged: the templates of April-GAN (Chen et al., 2023) with the class name replaced by the generic terms `object` and `texture`. No prompt is learned or tuned.

Patch scores. For scoring, each patch token is blended with the mean of its 3×3 neighbourhood (weights 0.6 and 0.4) before projection to suppress isolated noisy patches. Following equation 2, the projected tokens are ℓ_2-normalized before their cosine similarities to $\boldsymbol{t}_{\mathrm{n}}$ and $\boldsymbol{t}_{\mathrm{a}}$ are computed. The class token is excluded, the softmax uses a logit scale of 20 (i.e. $\tau = 1/20$), and the per-layer probabilities are averaged with uniform weights to give s_i.

Pixel map and image score. The anomaly map $\boldsymbol{A}$ in Section 3.5 is obtained by reshaping $\{s_i\}$ to the patch grid, followed by bilinear upsampling and Gaussian smoothing with standard deviation 4 and kernel size 25. For the image score in equation 9, we set $\lambda = 1/2$ and $k = 100$, the top-k is taken over pixels of the upsampled map. Evaluation follows the official code of RareCLIP (He et al., 2025), which resizes predicted maps and ground-truth masks to 256×256. PRO is computed as in AnomalyCLIP (Zhou et al., 2024) and April-GAN (Chen et al., 2023).

Online variant. The online setting adds the rarity mechanism of RareCLIP (He et al., 2025), which we use as published, with all of its default hyper-parameters and no normal reference images (k-shot $= 0$). RareCLIP scores each patch by how unlike it is to the patches of the images already seen in the stream, and fuses that rarity evidence with the output of its own text branch. That text branch relies on projection layers trained on a labeled auxiliary dataset. We replace its patch map with our training-free score s_i, so the online variant adds no training to our construction. Everything else in their inference pipeline is left untouched, namely the refinement of the text map, the patch- and image-level rarity estimates, the fusion of the two branches into a combined score $\tilde{s}_i$, and the final image-level re-scoring against similar past images. We refer to their paper for the details of these stages. Both $\boldsymbol{A}$ and the image score are then built from $\tilde{s}_i$ in place of s_i. Their fusion gives the rarity term no weight on the first image of a class and increases it over the next few, reflecting that the online reference memory is sparse and less reliable early in the stream. All incoming patch features update the stored references, so no hard normal/anomalous selection is performed. The patch feature bank of each layer and scale nevertheless has a fixed capacity. Whenever an update exceeds it, RareCLIP's Sequential Coreset Sampling keeps the features that are least similar to earlier entries and discards the redundant ones, so the bank does not grow with the stream. The patch-to-image similarities and the image-level references are likewise kept only for a fixed number of recent images. The offline setting disables this branch, keeps no test-stream references, and uses $\tilde{s}_i = s_i$. No image is excluded from evaluation in the online setting: the first image of the stream, which serves as the calibration image and is scored by the text branch alone (empty reference sets), still contributes its prediction to the reported metrics like every other test image.

Per-class reset. All state above (including the stored references, the cumulative text map, the stream position, and the calibrated $\{\boldsymbol{W}^{(\ell)}\}$) is reset before each class is processed. The evaluation order within a class is a fixed shuffle determined by the seed. Class grouping follows the existing benchmark protocol. The calibration image is whichever image the shuffle places first, normal or anomalous, taken without reading any label and still scored as an ordinary test image. This initialization adds no supervision and does not change the evaluated ZSAD setting.

A.1 Direction-Recovery Objectives

Table 3c keeps the acceptance rule of equation 8 and changes only the quantity it maximizes. For one layer, write $c_i^{(\mathrm{n})} = \langle \boldsymbol{W}\boldsymbol{x}_i, \boldsymbol{t}_{\mathrm{n}} \rangle$ and $c_i^{(\mathrm{a})} = \langle \boldsymbol{W}\boldsymbol{x}_i, \boldsymbol{t}_{\mathrm{a}} \rangle$ for the two prototype cosines of a calibration patch. Their margin m_i and the single-layer anomaly probability p_i in equation 2 satisfy

$$m_i \;=\; c_i^{(\mathrm{a})} - c_i^{(\mathrm{n})}, \qquad p_i \;=\; \frac{1}{1+\exp(-m_i/\tau)}. \tag{10}$$

For $\tau > 0$, the anomaly probability increases monotonically with this margin. Separation in equation 7, $\mathrm{Sep} = \frac{1}{N}\sum_i m_i$, therefore selects directions using the same normal–abnormal prototype contrast that determines the anomaly score. The five alternatives below replace that contrast with something weaker. None of them reads a classification or a segmentation label. Because every criterion, separation included, averages over unlabeled calibration patches, none of them can single out directions that respond specifically to defects. The comparison therefore asks which label-free proxy is most useful for recovery, not which one identifies defect-sensitive directions.

Abnormal pull and normal push keep one of the two similarities and drop the other,

$$\mathrm{Pull} = \frac{1}{N}\sum_i c_i^{(\mathrm{a})}, \qquad \mathrm{Push} = \frac{1}{N}\sum_i \left(1 - c_i^{(\mathrm{n})}\right). \tag{11}$$

The first rewards a direction for moving calibration patches towards $\boldsymbol{t}_{\mathrm{a}}$ and the second for moving them away from $\boldsymbol{t}_{\mathrm{n}}$, so comparing both with Sep asks whether the useful signal is the contrast between the two prototypes or the response to either one on its own.

Cosine and Euclidean proximity drop the normal–abnormal sign altogether and reward a direction for placing the calibration patches close to the text prototypes, regardless of which one denotes normality or abnormality. The first averages both similarities and the second takes the distance to the nearer prototype,

$$\mathrm{Prox}_{\cos} = \frac{1}{2N}\sum_i \left(c_i^{(\mathrm{n})} + c_i^{(\mathrm{a})}\right), \qquad \mathrm{Prox}_{\mathrm{euc}} = -\frac{1}{N}\sum_i \min_{\boldsymbol{t}\in\{\boldsymbol{t}_{\mathrm{n}},\boldsymbol{t}_{\mathrm{a}}\}} \left\|\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{t}}\right\|_2, \tag{12}$$

where $\hat{\boldsymbol{z}}_i = \boldsymbol{W}\boldsymbol{x}_i / \|\boldsymbol{W}\boldsymbol{x}_i\|_2$ and $\hat{\boldsymbol{t}}$ are the ℓ_2-normalized patch and prototype. Both accept a direction when the patches become better grounded in the text subspace, since raising both similarities by the same amount improves them while leaving the margin m_i unchanged. They therefore ask whether alignment with the text prototypes as a whole is enough. The two differ in how they weight the patches. $\mathrm{Prox}_{\cos}$ treats the two prototypes symmetrically, whereas the concave map $\|\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{t}}\|_2 = \sqrt{2 - 2\left\langle \hat{\boldsymbol{z}}_i, \hat{\boldsymbol{t}}\right\rangle}$ and the $\min$ make $\mathrm{Prox}_{\mathrm{euc}}$ follow whichever prototype a patch is already nearer to.

Low entropy maximizes negative binary entropy,

$$\mathrm{Conf} = \frac{1}{N}\sum_i \left[p_i \log p_i + (1 - p_i)\log(1 - p_i)\right], \tag{13}$$

using the same anomaly probabilities and temperature as equation 10. It accepts a direction when the average prediction entropy decreases, so probabilities near either zero or one are favored over uncertain predictions near $1/2$ and confident normal and confident abnormal predictions are rewarded equally. It thus asks whether greater confidence alone is sufficient.

Random recovery samples the same candidate window uniformly using the run seed and keeps the number of directions that separation accepts at each layer, separating the benefit of recovering weak directions at all from the benefit of choosing which ones. All non-random criteria share the candidate window, acceptance rule of equation 8, calibration image and seed, and differ only in the quantity being compared.

Separation gives the best localization on both datasets and no alternative matches it on either. Abnormal pull is the closest competitor overall. The two unsigned proximity criteria recover part of the gain on MVTec-AD but lose more of it on VisA, where Euclidean proximity even falls below random recovery. Normal push and low entropy localize worst, and low entropy is far behind on VisA, so more confident predictions alone do not give better localization. Image-level differences between the criteria are small throughout. Unsigned grounding in the text subspace is therefore a usable but unreliable proxy, helping where the prototypes already sit close to the patch distribution and degrading where they do not, whereas the signed contrast transfers across both datasets.

B Baselines and Reproduction Protocol

Sources of baseline results. We use results from the original papers whenever available. Missing results are reproduced using the authors' official implementations. We use the released code, check-

points, and default configurations, changing only the evaluation dataset. Empty entries indicate results not reported in our comparison.

Baselines. The seven baselines are described in Section 4; all of them except WinCLIP require a fully labeled auxiliary dataset and are grouped as training-based in Tables 1 and 2. Three need a note here. Among the main-table baselines, *WinCLIP* (Jeong et al., 2023) is the only one that, like 0-CLIP, optimizes no parameters; its official implementation is unavailable, so quoted values are used wherever possible and the widely used community re-implementation otherwise. For *MRAD* (Xu et al., 2026) we report its strongest variant, MRAD-CLIP, which, at inference time, retrieves labeled reference features from a feature-label memory bank built from a separate labeled auxiliary dataset and is marked with ‡ in the main tables. *RareCLIP* (He et al., 2025) and our online variant are marked with † because both use unlabeled test-stream observations. Our online variant follows RareCLIP's inference protocol unchanged (Appendix A).

Batch zero-shot methods. We do not compare with batch methods such as ACR (Li et al., 2023) and MuSc (Li et al., 2024; 2026), which jointly process a large collection of test images. These methods trade computational cost against accuracy. Cross-image rarity becomes more reliable as more test images are available (He et al., 2025), but the cost grows accordingly, since MuSc scores each image against all other unlabeled test images (Li et al., 2024). For an online variant of MuSc, He et al. (2025) report that the per-image time grows from 455 ms to 4055 ms and the GPU memory from 4.7 GB to 19.9 GB as the number of stored images grows from 50 to 1000. In practical deployment, the batch size must therefore be balanced against speed and computational resources. Moreover, the online protocol of RareCLIP already outperforms these batch methods on MVTec-AD and VisA (He et al., 2025), and RareCLIP is included in our comparison. Both of our settings score each image as it arrives, using at most previously observed images.

B.1 Sensitivity to the Auxiliary Training Source

Setup. For targets other than MVTec-AD and VisA, the main tables report FB-CLIP with an MVTec-AD-trained checkpoint and MRAD with a VisA-trained checkpoint. We swap the two sources, re-evaluating FB-CLIP with a VisA-trained checkpoint and MRAD with an MVTec-AD-trained checkpoint, and leave every other component of each method unchanged. MVTec-AD and VisA are excluded as evaluation targets so that neither auxiliary source is also a test set. Table 4 reports the reference values (REF, copied from Tables 1 and 2), the swapped-source values (ALT), and $\Delta = \text{ALT} - \text{REF}$ in percentage points.

Results. Neither source is uniformly preferable. The sign of Δ flips across both targets and metrics for both methods, and the largest shifts reach several points in either direction. Without target labels, which are unavailable in zero-shot deployment, the choice of which checkpoint transfers better cannot be directly validated. 0-CLIP removes this source-specific choice, without claiming independence from CLIP pretraining or invariance to the target distribution.

C Diagnostic Figure Protocols

Figures 3–6 illustrate the three construction steps rather than report benchmark numbers. All of them are produced by the inference code of Section 3.5 with the rarity bank disabled, so they show the text branch alone and are unaffected by the online setting. Patch labels come from the ground-truth mask pooled onto the 37×37 grid, where a patch counts as abnormal if any of its pixels is.

Patch-feature embeddings (Figures 3 and 8–11). Panels show a two-dimensional t-SNE embedding of the ℓ_2-normalized projected patch features, pooled over the classes of one dataset (Figure 3: MVTec-AD, deepest of the four layers) and subsampled for legibility. Each operator is embedded independently, so the two panels do not share a coordinate frame and only the relation of the patch cloud to the text anchor is comparable between them, not the absolute positions. The anchor is embedded jointly with the patches it is compared against. For the pooled views we average the per-class abnormal prototypes and renormalize, which yields one representative anchor.

Score-margin diagnostics (Figure 4). The upper row shows one automatically selected VisA example, chosen to have a small defect and a large localization gain of $\boldsymbol{W}_{\mathcal{B}}$ over $\boldsymbol{W}_{\text{flip}}$, its maps are smoothed as in Section 3.5 and then contrast-normalized per map. The lower row plots two mar-

Table 4: **Auxiliary-source checkpoint comparison** on industrial (⚙) and medical (🏥) datasets. Values are percentages; Δ = ALT − REF in percentage points, shaded blue for a gain and red for a loss with intensity increasing in $|\Delta|$. REF: FB-CLIP trained on MVTec-AD, MRAD trained on VisA. ALT: FB-CLIP trained on VisA, MRAD trained on MVTec-AD. "–" marks unreported comparisons; Brain MRI is omitted because its main-table subset cannot be matched unambiguously to the additional logs.

		P-AUROC		PRO		I-AUROC		I-AP	
Method	Dataset	Ref → Alt	Δ	Ref → Alt	Δ	Ref → Alt	Δ	Ref → Alt	Δ
FB-CLIP	⚙ BTAD	95.8 → 96.4	+0.6	80.0 → 77.5	−2.5	93.2 → 95.1	+1.9	96.4 → 97.9	+1.5
	⚙ MPDD	96.9 → 97.0	+0.1	91.1 → 88.7	−2.4	79.1 → 75.3	−3.8	86.2 → 82.2	−4.0
	⚙ DTD-Syn	98.3 → 97.8	−0.5	92.6 → 91.3	−1.3	97.9 → 95.8	−2.1	99.1 → 98.6	−0.5
	⚙ SDD	90.4 → 89.4	−1.0	64.0 → 60.0	−4.0	80.8 → 82.0	+1.2	72.0 → 73.3	+1.3
	⚙ DAGM	97.3 → 95.8	−1.5	94.1 → 91.9	−2.2	99.0 → 98.8	−0.2	96.5 → 96.5	0.0
	⚙ KSDD2	99.1 → 98.5	−0.6	97.6 → 95.3	−2.3	97.3 → 95.9	−1.4	90.5 → 86.3	−4.2
	🏥 Liver CT	96.9 → 97.3	+0.4	88.2 → 85.4	−2.8	–	–	–	–
	🏥 Retina RESC	94.8 → 95.2	+0.4	76.9 → 79.8	+2.9	–	–	–	–
	🏥 TN3K	79.0 → 75.0	−4.0	46.8 → 37.9	−8.9	–	–	–	–
	🏥 Br35H	–	–	–	–	97.1 → 96.2	−0.9	97.4 → 96.6	−0.8
	🏥 Retina OCT	–	–	–	–	84.2 → 86.1	+1.9	94.7 → 95.1	+0.4
MRAD	⚙ BTAD	95.4 → 95.4	0.0	72.8 → 72.9	+0.1	92.8 → 92.2	−0.6	94.2 → 94.4	+0.2
	⚙ MPDD	97.9 → 97.8	−0.1	90.6 → 90.1	−0.5	81.8 → 81.5	−0.3	83.4 → 83.0	−0.4
	⚙ DTD-Syn	98.1 → 98.4	+0.3	89.8 → 89.7	−0.1	96.0 → 95.9	−0.1	98.4 → 98.6	+0.2
	⚙ SDD	93.0 → 93.1	+0.1	72.0 → 72.2	+0.2	83.9 → 82.6	−1.3	76.2 → 72.7	−3.5
	⚙ DAGM	97.4 → 97.7	+0.3	90.3 → 91.4	+1.1	98.4 → 98.2	−0.2	98.6 → 93.7	−4.9
	⚙ KSDD2	98.9 → 98.9	0.0	95.6 → 95.6	0.0	95.1 → 93.0	−2.1	88.9 → 86.3	−2.6
	🏥 Liver CT	96.2 → 96.5	+0.3	83.6 → 83.4	−0.2	–	–	–	–
	🏥 Retina RESC	93.5 → 93.3	−0.2	73.9 → 73.1	−0.8	–	–	–	–
	🏥 TN3K	78.0 → 82.6	+4.6	49.6 → 56.4	+6.8	–	–	–	–
	🏥 Br35H	–	–	–	–	97.9 → 97.9	0.0	97.6 → 97.7	+0.1
	🏥 Retina OCT	–	–	–	–	86.8 → 83.8	−3.0	95.3 → 94.2	−1.1

gins against ρ_r in steps of 0.05, as class averages with individual classes drawn faintly behind. The dashed line is the $\rho_r = 1$ end of the same curve, i.e. the untruncated operator. Both margins split patches, respectively images, by ground truth: the left panel is the mean patch score s_i over abnormal patches minus the mean over normal ones, and the right panel the analogous difference of image scores, where for this diagnostic the image score is the mean of the highest-scoring patches without the global-token term of equation 9. The shading under the curves is a visual guide only.

C.1 RETENTION AND RECOVERY SWEEPS

Figure 5 in the main text reports all four metrics on VisA, namely P-AUROC, PRO, I-AUROC and I-AP, as averages over classes in the offline setting. The horizontal axis is the fraction $\rho = |\mathcal{K}|/r$ of singular directions the operator keeps. Along the orange curve $\mathcal{K}$ is the principal block of Section 3.3 on its own, so ρ is just ρ_r. Each blue curve instead fixes ρ_r at one of 0.4, 0.6 and 0.8 and widens the candidate window ρ_e of Section 3.4, so it leaves the orange curve at $\rho = \rho_r$ and moves right as more weak directions are examined. A blue curve is plotted at the retained fraction it actually reaches rather than at the window it was offered. The dashed line and the open marker at $\rho = 1$ are the full operator, i.e. equation 6 with all directions retained, $-\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^\top = -(\boldsymbol{P}^\dagger)^\top$. Shaded blue is the gain that separation-guided recovery adds to the principal block it starts from. Figure 6 follows the same protocol on MVTec-AD, with the principal block replaced by the leading direction alone as the starting point for recovery.

The recovery curves improve on their starting principal blocks at all three initial retention ratios. Keeping more directions is not uniformly beneficial: P-AUROC largely plateaus at high retention, while PRO and the image-level metrics decline toward the right end of the sweep. The different

trends show why retained rank alone is insufficient. Selecting useful directions matters, and no single retention ratio maximizes every metric.

D PATCH-FEATURE VISUALIZATIONS

Figures 8–11 visualize the projector's polarity reversal across layers and classes on MVTec-AD and VisA, following the embedding protocol of Appendix C. The all-layer views cover all 24 layers, while the per-class views show the four layers used by 0-CLIP. The amount of normal–abnormal overlap varies across classes and depths. Negation changes the patches' relation to the fixed text anchor but does not remove this overlap. This supports polarity correction as a starting point, with the further gains from spectral construction quantified in Table 3a.

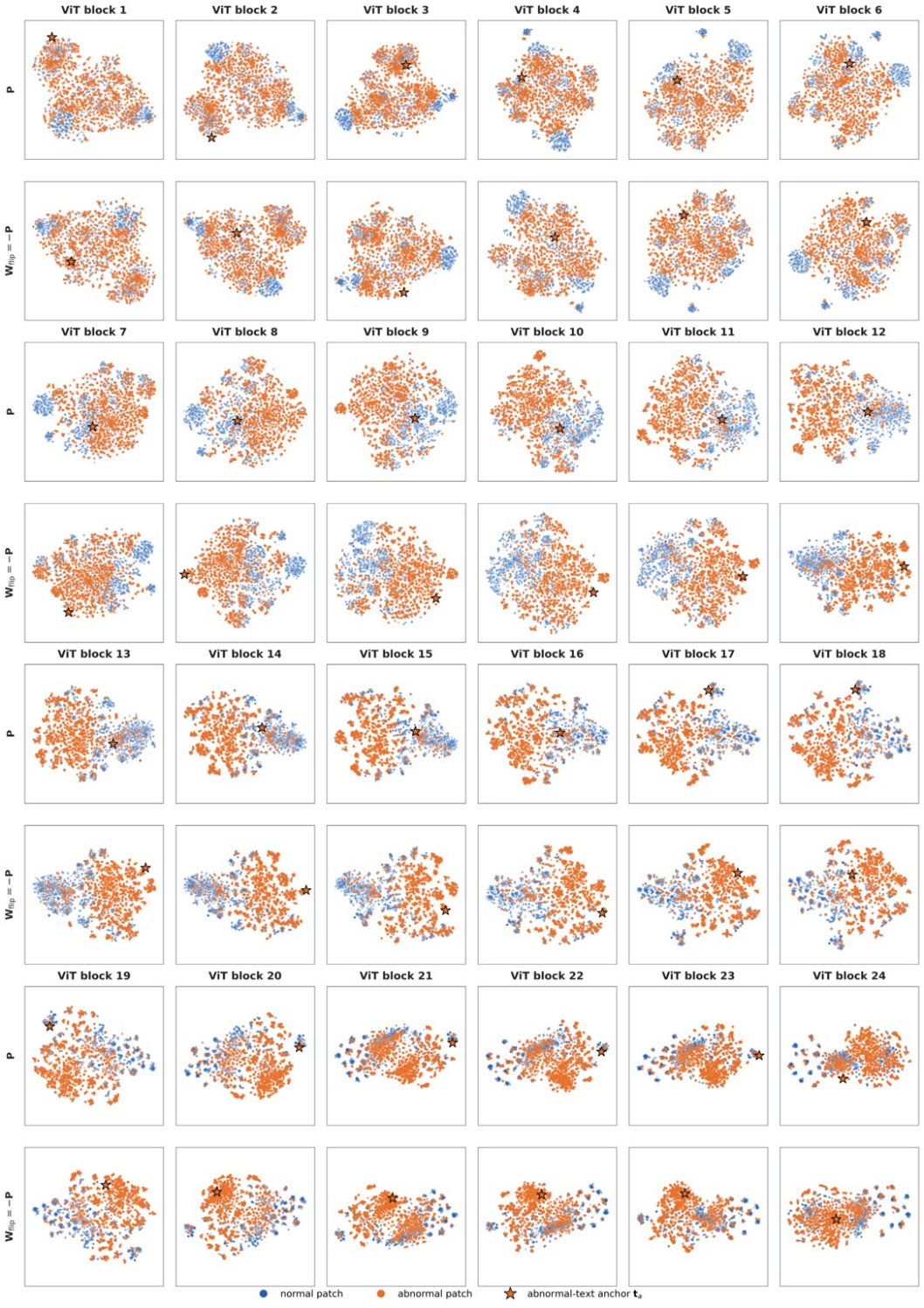


Figure 8: **All-layer patch features on MVTec-AD.** Panels cover ViT layers 1–24, with each pair of rows using $\boldsymbol{P}$ and $-\boldsymbol{P}$. Blue and orange points are normal and abnormal patches, and the star is $\boldsymbol{t}_{\mathrm{a}}$.

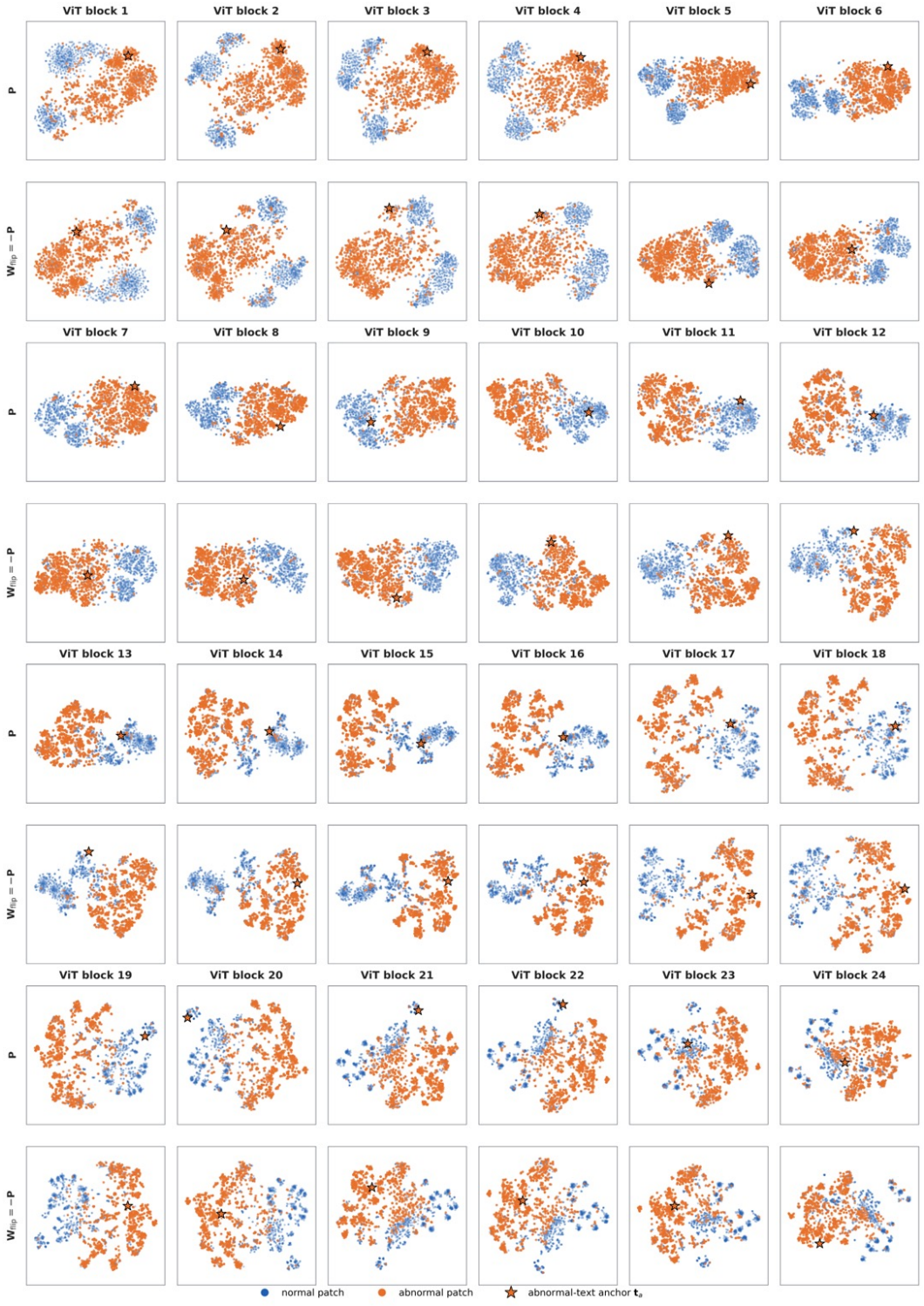


Figure 9: **All-layer patch features on VisA.** Setup as in Figure 8.



Figure 10: **All-class patch features on MVTec-AD.** Each class shows layers 12, 16, 20 and 24, with colors and symbols as in Figure 8. The reversal repeats across all 15 object and texture classes.

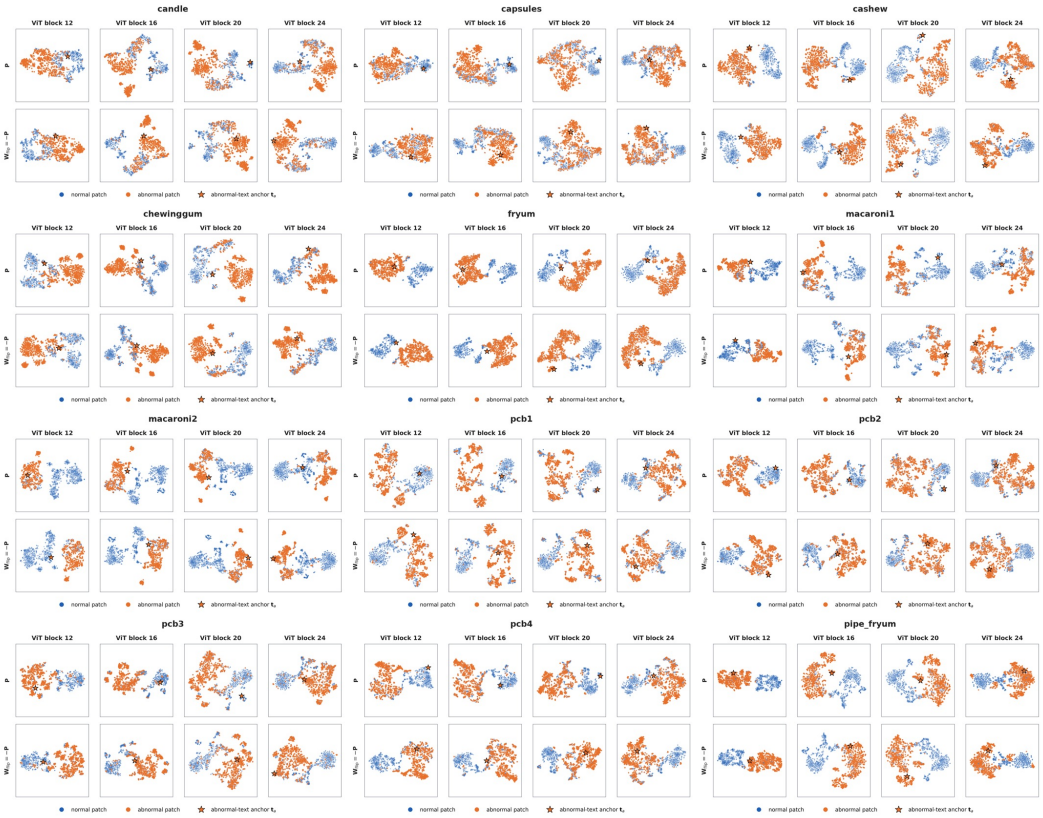


Figure 11: **All-class patch features on VisA.** Panels as in Figure 10, with the reversal repeating across all 12 object classes.

Table 5: **Training-free comparison on shared benchmarks.** Industrial (⚙) and medical (🏥) datasets; best and second results are highlighted per metric. Averages use the listed eight pixel-level and nine image-level datasets.

Pixel level: P-AUROC / PRO				
Dataset	WinCLIP	CLIP-AD	MRAD-TF‡	0-CLIP
⚙ MVTec-AD	85.1 / 64.6	87.5 / 75.0	86.7 / 63.5	94.9 / 87.3
⚙ VisA	79.6 / 56.8	88.1 / 68.4	91.0 / 71.0	95.3 / 86.2
⚙ BTAD	71.4 / 32.8	72.4 / 35.7	80.5 / 36.3	96.5 / 81.7
⚙ MPDD	71.2 / 48.9	62.6 / 38.7	92.7 / 76.2	96.2 / 91.1
⚙ DTD-Syn	78.4 / 51.0	92.5 / 80.8	90.7 / 73.8	98.0 / 94.0
⚙ SDD	55.9 / 14.7	84.5 / 50.1	85.0 / 63.0	91.9 / 78.0
⚙ DAGM	75.5 / 44.4	84.8 / 65.1	86.9 / 66.1	96.9 / 94.1
⚙ KSDD2	75.4 / 69.2	97.1 / 93.0	94.9 / 87.1	97.3 / 94.8
Average	74.1 / 47.8	83.7 / 63.4	88.5 / 67.1	95.9 / 88.4

Image level: I-AUROC / I-AP				
Dataset	WinCLIP	CLIP-AD	MRAD-TF‡	0-CLIP
⚙ MVTec-AD	91.8 / 95.1	90.5 / 95.6	79.0 / 90.0	94.4 / 97.4
⚙ VisA	78.1 / 77.5	78.7 / 81.5	75.0 / 79.1	85.3 / 87.0
⚙ BTAD	83.3 / 84.1	78.3 / 82.3	86.3 / 90.3	95.0 / 94.6
⚙ MPDD	61.5 / 69.2	61.5 / 69.4	67.9 / 73.8	76.8 / 81.8
⚙ DTD-Syn	63.8 / 83.5	94.3 / 97.5	79.1 / 89.6	98.0 / 99.3
⚙ SDD	54.8 / 37.7	81.6 / 72.2	80.0 / 66.3	84.5 / 78.4
⚙ DAGM	58.5 / 59.1	85.7 / 65.5	78.7 / 80.6	99.5 / 98.3
⚙ KSDD2	77.5 / 43.9	94.5 / 80.3	90.4 / 82.6	95.5 / 89.6
🏥 Br35H	80.5 / 82.2	85.2 / 86.8	91.0 / 92.1	96.1 / 94.6
Average	72.2 / 70.3	83.4 / 81.3	80.8 / 82.7	91.7 / 91.2

‡ Uses a labeled auxiliary memory bank.

E Additional Experimental Results

E.1 Comparison with Training-Free Methods

In Table 5, WinCLIP and MRAD-TF results are taken directly from Table 1 of Xu et al. (2026). 0-CLIP results are from Tables 1 and 2. For CLIP-AD (Chen et al., 2024b), we use its training-free version.

Offline 0-CLIP obtains the best average on all four metrics and the best result on every dataset. These gains require neither online test-stream references nor the labeled auxiliary memory bank used by MRAD-TF, showing that the advantage is already present in the constructed operator.

E.2 Transfer to SigLIP

Using the `google/siglip-large-patch16-384` checkpoint, we also apply 0-CLIP to SigLIP ViT-L/16 (Zhai et al., 2023). Unlike CLIP, SigLIP has neither a class token nor a separate visual projector. Its image embedding is produced by a multi-head attention-pooling (MAP) head on the layer-normalized final tokens, $\boldsymbol{e} = \boldsymbol{h} + \mathrm{MLP}(\mathrm{LN}(\boldsymbol{h}))$, where $\boldsymbol{h}$ is the output of multi-head attention with a learned query. Each head aggregates the value projections of the tokens with its own attention weights, and the output projection $\boldsymbol{W}_o$ maps the concatenated heads to the embedding. If all heads assigned a token $\boldsymbol{x}$ the same weight a, its contribution to $\boldsymbol{h}$ would be $a\,\boldsymbol{W}_o\boldsymbol{W}_v\boldsymbol{x}$ up to bias terms, where $\boldsymbol{W}_v$ stacks the value projections of all heads. We use this token-to-embedding map, $\boldsymbol{P} = \boldsymbol{W}_o\boldsymbol{W}_v$, as the substitute projector. The MLP acts on the aggregated vector rather than on individual tokens and is not included. The rest of the construction is unchanged. The output of

Table 6: **0-CLIP with CLIP and SigLIP backbones** on VisA (%, offline setting, seed 0).

Backbone	P-AUROC	PRO	I-AUROC	I-AP
CLIP ViT-L/14-336	**95.43**	**86.17**	85.30	87.06
SigLIP ViT-L/16	95.10	77.05	**86.57**	**87.82**

Table 7: **0-CLIP over five seeds** (%) on industrial (⚙) and medical (🏥) datasets. Cells report the mean with its standard deviation as a grey subscript ($97.83_{\pm 0.25}$ reads as 97.83 ± 0.25), and "–" marks unreported results. Means match Tables 1 and 2 rounded to one decimal.

Dataset	0-CLIP†				0-CLIP			
	Pixel level		Image level		Pixel level		Image level	
	P-AUROC	PRO	I-AUROC	I-AP	P-AUROC	PRO	I-AUROC	I-AP
⚙ MVTec-AD	$97.83_{\pm 0.25}$	$93.95_{\pm 0.21}$	$98.22_{\pm 0.12}$	$99.34_{\pm 0.07}$	$94.91_{\pm 0.20}$	$87.30_{\pm 0.47}$	$94.44_{\pm 0.32}$	$97.35_{\pm 0.21}$
⚙ VisA	$98.90_{\pm 0.03}$	$93.73_{\pm 0.23}$	$92.92_{\pm 0.31}$	$93.73_{\pm 0.23}$	$95.32_{\pm 0.10}$	$86.16_{\pm 0.04}$	$85.29_{\pm 0.16}$	$86.98_{\pm 0.11}$
⚙ BTAD	$98.03_{\pm 0.03}$	$83.89_{\pm 0.78}$	$96.54_{\pm 0.36}$	$97.90_{\pm 0.30}$	$96.48_{\pm 0.23}$	$81.74_{\pm 0.57}$	$94.98_{\pm 0.64}$	$94.59_{\pm 0.35}$
⚙ MPDD	$97.47_{\pm 0.36}$	$93.54_{\pm 0.68}$	$73.27_{\pm 2.12}$	$77.48_{\pm 1.60}$	$96.17_{\pm 0.07}$	$91.10_{\pm 0.19}$	$76.79_{\pm 0.26}$	$81.84_{\pm 0.13}$
⚙ DTD-Syn	$98.33_{\pm 0.01}$	$95.19_{\pm 0.10}$	$98.03_{\pm 0.14}$	$99.25_{\pm 0.03}$	$97.99_{\pm 0.05}$	$93.99_{\pm 0.17}$	$98.01_{\pm 0.05}$	$99.28_{\pm 0.01}$
⚙ SDD	$93.51_{\pm 0.27}$	$77.84_{\pm 0.64}$	$85.56_{\pm 0.89}$	$76.24_{\pm 1.98}$	$91.88_{\pm 0.24}$	$77.97_{\pm 0.26}$	$84.52_{\pm 0.16}$	$78.41_{\pm 0.29}$
⚙ DAGM	$98.19_{\pm 0.03}$	$96.38_{\pm 0.07}$	$99.56_{\pm 0.08}$	$96.72_{\pm 0.23}$	$96.91_{\pm 0.05}$	$94.08_{\pm 0.12}$	$99.53_{\pm 0.02}$	$98.30_{\pm 0.12}$
⚙ KSDD2	$98.52_{\pm 0.29}$	$96.17_{\pm 0.32}$	$96.42_{\pm 0.18}$	$91.78_{\pm 0.23}$	$97.28_{\pm 0.30}$	$94.81_{\pm 0.25}$	$95.48_{\pm 0.09}$	$89.57_{\pm 0.15}$
⚙ Real-IAD	$97.38_{\pm 0.05}$	$88.59_{\pm 0.24}$	$80.84_{\pm 0.06}$	$78.99_{\pm 0.11}$	$95.34_{\pm 0.05}$	$87.37_{\pm 0.08}$	$76.29_{\pm 0.12}$	$74.23_{\pm 0.11}$
🏥 Liver CT	$97.53_{\pm 0.13}$	$92.01_{\pm 0.54}$	–	–	$95.85_{\pm 0.07}$	$90.87_{\pm 0.17}$	–	–
🏥 Brain MRI	$96.88_{\pm 0.06}$	$81.16_{\pm 0.36}$	$73.96_{\pm 0.71}$	$91.13_{\pm 0.14}$	$93.74_{\pm 0.06}$	$76.28_{\pm 0.18}$	$80.42_{\pm 0.21}$	$94.72_{\pm 0.05}$
🏥 Retina RESC	$95.11_{\pm 0.24}$	$81.66_{\pm 0.80}$	–	–	$87.36_{\pm 0.14}$	$59.01_{\pm 0.63}$	–	–
🏥 TN3K	$80.01_{\pm 0.45}$	$57.87_{\pm 0.34}$	–	–	$82.44_{\pm 0.18}$	$52.99_{\pm 1.04}$	–	–
🏥 Br35H	–	–	$89.49_{\pm 0.73}$	$83.75_{\pm 0.79}$	–	–	$96.06_{\pm 0.05}$	$94.59_{\pm 0.05}$
🏥 Retina OCT	–	–	$88.07_{\pm 1.32}$	$95.92_{\pm 0.45}$	–	–	$79.29_{\pm 0.16}$	$91.99_{\pm 0.06}$

† Online inference using unlabeled test-stream observations.

the MAP head serves as the global embedding for s_{g}, and the text prototypes are computed with SigLIP's text encoder and the same prompt ensemble.

Table 6 compares both backbones on VisA in the offline setting. SigLIP reaches a similar P-AUROC and higher image-level scores. The construction thus carries over to a backbone without a dedicated visual projector once the token-to-embedding map of its pooling head takes the place of $\boldsymbol{P}$.

E.3 Standard Deviations

Table 7 reports the five-seed means and standard deviations of 0-CLIP behind Tables 1 and 2. Each seed fixes the shuffle of every class stream, and the first image of that shuffle, normal or anomalous, calibrates $\boldsymbol{W}$ (Appendix A). Offline, the seed therefore changes only the calibration image. Online, it also changes the order in which the rarity branch accumulates its references. Offline standard deviations stay within about one point, so the operator hardly depends on which image calibrates it. Online, pixel-level metrics remain equally stable, while image-level metrics vary more on a few datasets, by up to about two points on MPDD. For a given seed both settings calibrate the same operator, so the extra spread comes from the stream order seen by the rarity branch, in line with the stream dependence analyzed in Appendix E.6.

E.4 Qualitative Comparison

Figure 12 shows the anomaly maps behind the numbers of Tables 1 and 2 on four MVTec-AD and four VisA classes. Appendix E.5 extends the coverage of our own maps to 37 categories from twelve datasets. The online variant concentrates its responses near the annotated defects on Capsule, Pill,

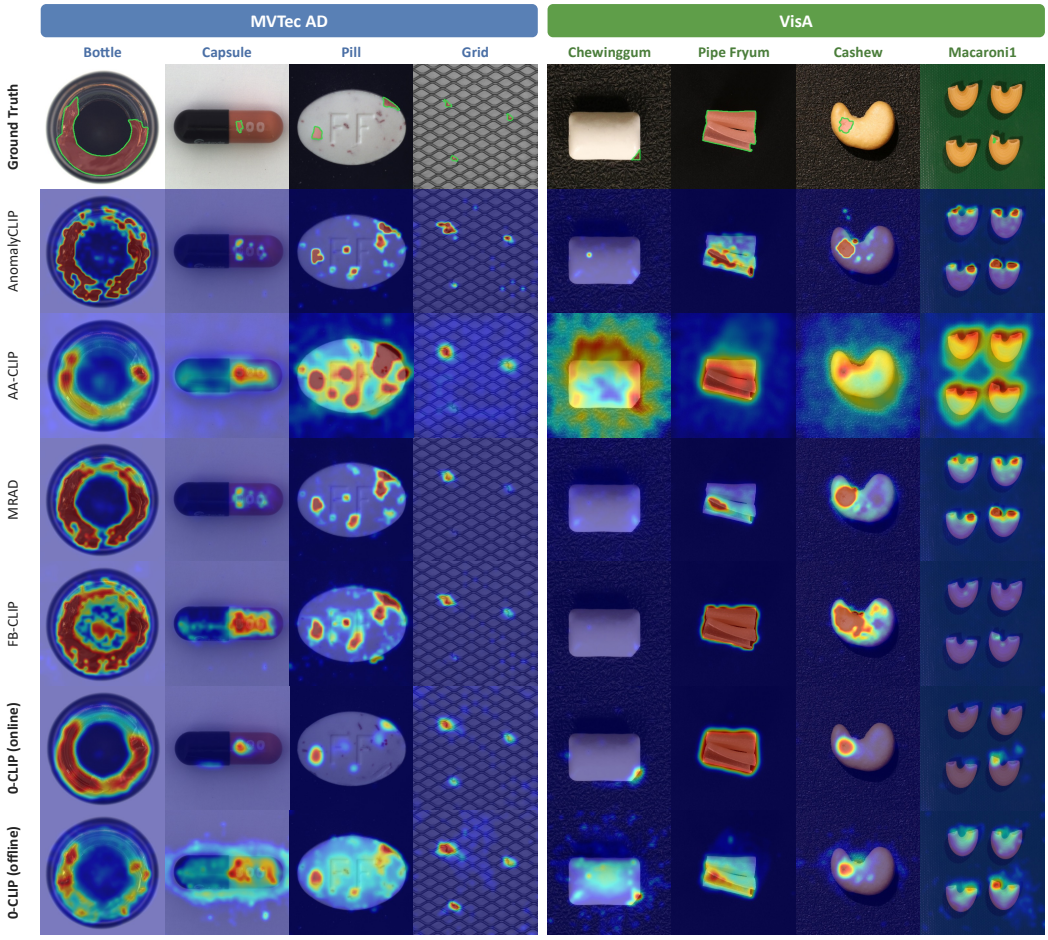


Figure 12: **Qualitative comparison on MVTec-AD and VisA.** Each column is one test image and the top row draws the ground-truth mask in green over it. The remaining rows are the anomaly maps of AnomalyCLIP, AA-CLIP, MRAD, FB-CLIP, 0-CLIP† (online) and 0-CLIP (offline). The online variant suppresses normal-region responses more effectively than the offline variant on Capsule and Pill, while both highlight the small defects on Grid and Cashew.

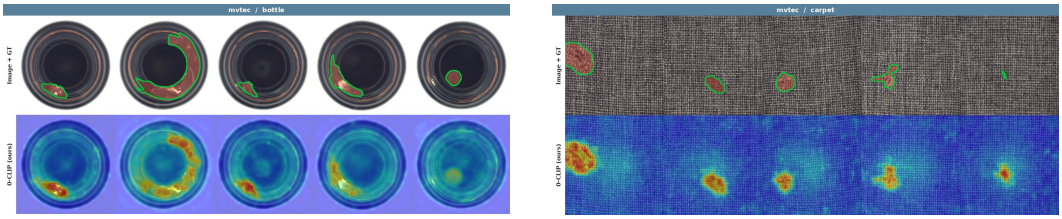
Chewinggum and Cashew, while suppressing much of the activation on normal regions. Offline 0-CLIP also highlights the defects but retains more diffuse responses on Capsule and Pill, illustrating the additional refinement provided by the rarity branch. These examples support the localization gains without implying uniformly clean maps for every image or either setting.

E.5 Per-Category Qualitative Results

Figure 12 compares methods on eight MVTec-AD and VisA classes. Figures 13–16 instead widen the coverage of our own maps to 37 categories from twelve datasets, all produced by offline 0-CLIP. Within each panel the top row overlays the ground-truth mask in green on the input and the bottom row is the anomaly map for the same five test images. The banner repeats the dataset and category so the panels stay readable on their own. The panels are drawn from the test streams to span the categories rather than selected for map quality, and they are not uniformly clean.

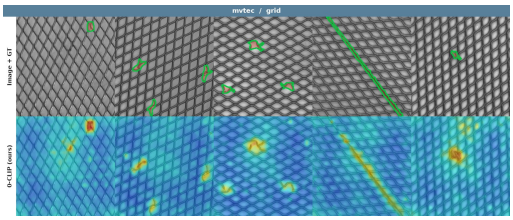
E.6 Where Online Inference Declines

On three datasets the online variant ranks images worse than the offline one (Table 2): MPDD (73.3 versus 76.8 I-AUROC), Brain MRI (74.0 versus 80.4) and Br35H (89.5 versus 96.1). Yet on MPDD and Brain MRI, the two with masks, localization still improves (97.5 versus 96.2 and 96.9 versus

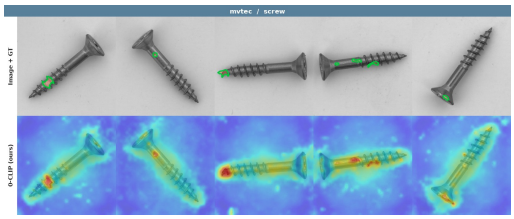


(a) MVTec-AD / Bottle

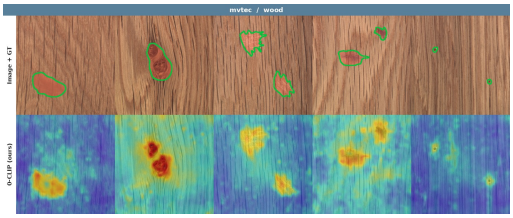
(b) MVTec-AD / Carpet



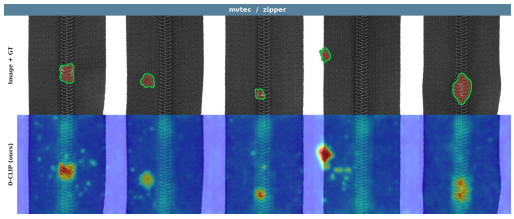
(c) MVTec-AD / Grid



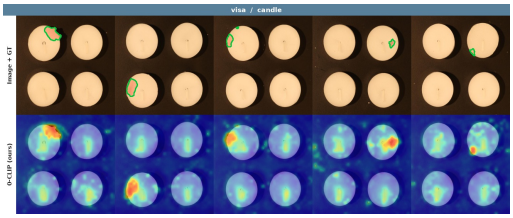
(d) MVTec-AD / Screw



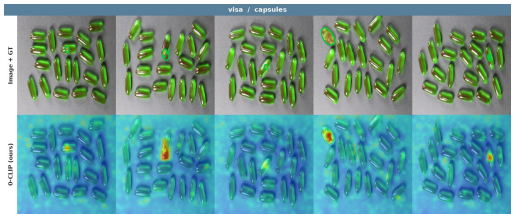
(e) MVTec-AD / Wood



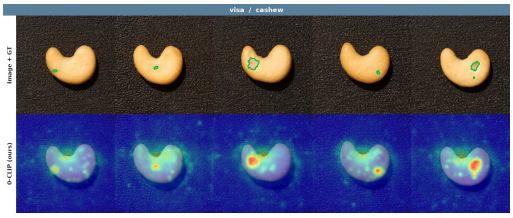
(f) MVTec-AD / Zipper



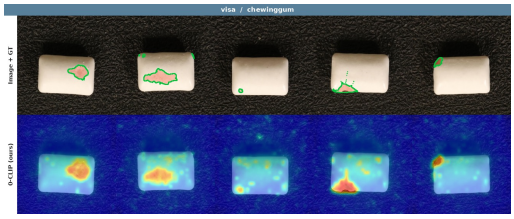
(g) VisA / Candle



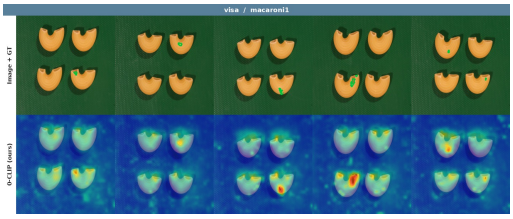
(h) VisA / Capsules



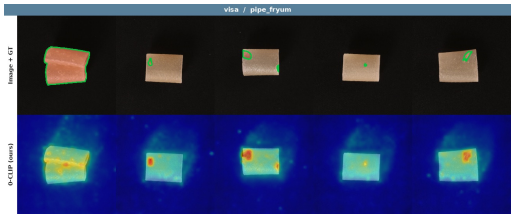
(i) VisA / Cashew



(j) VisA / Chewinggum

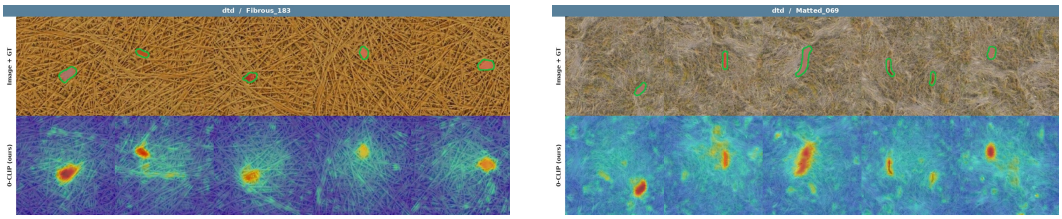


(k) VisA / Macaroni1

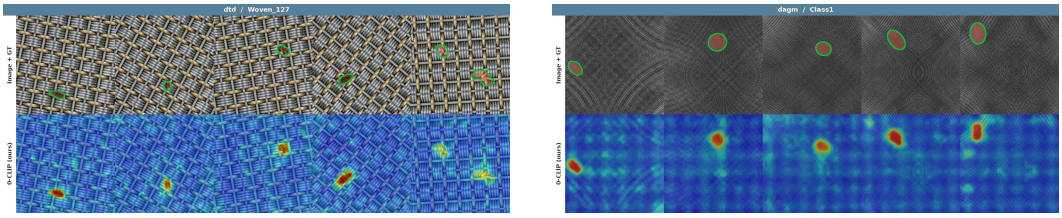


(l) VisA / Pipe_fryum

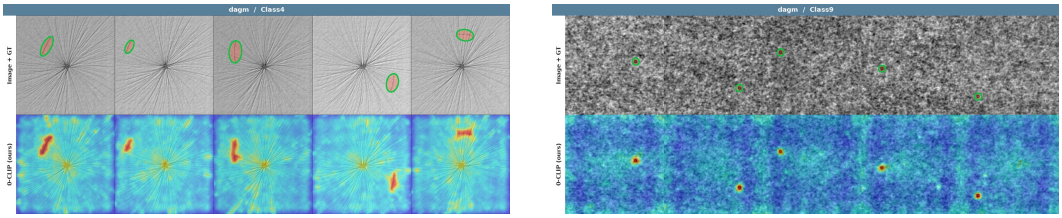
Figure 13: **Per-category anomaly maps on MVTec-AD and VisA.** Each panel covers one category: the upper row draws the ground-truth mask in green over five test images and the lower row is the offline 0-CLIP map for the same images. Maps are not normalized across panels.



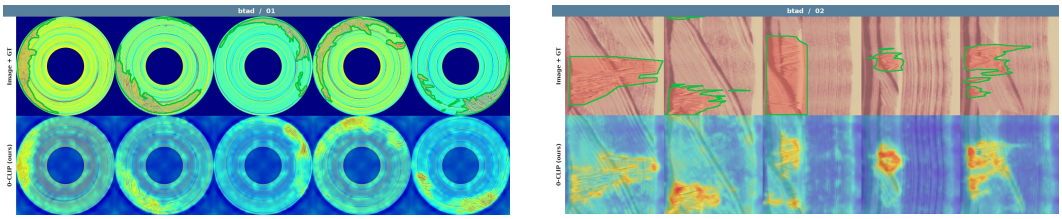
(a) DTD-Syn / Fibrous_183 (b) DTD-Syn / Matted_069



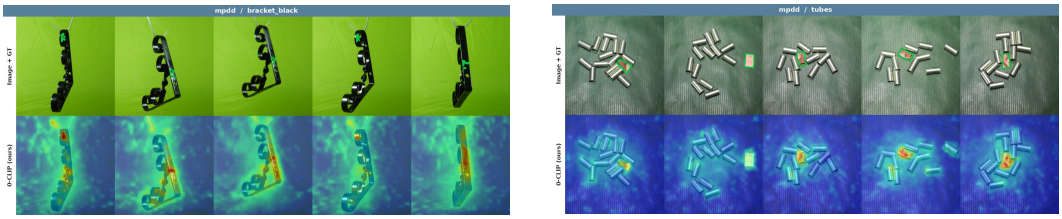
(c) DTD-Syn / Woven_127 (d) DAGM / Class1



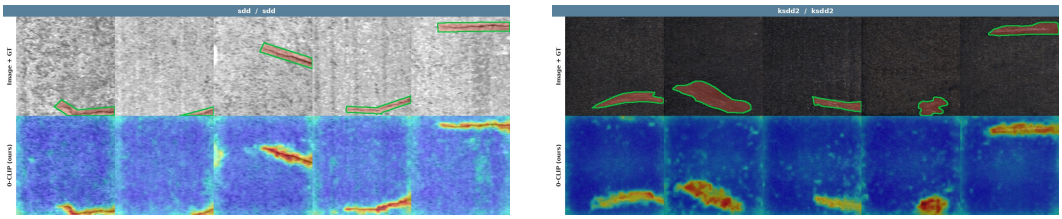
(e) DAGM / Class4 (f) DAGM / Class9



(g) BTAD / 01 (h) BTAD / 02

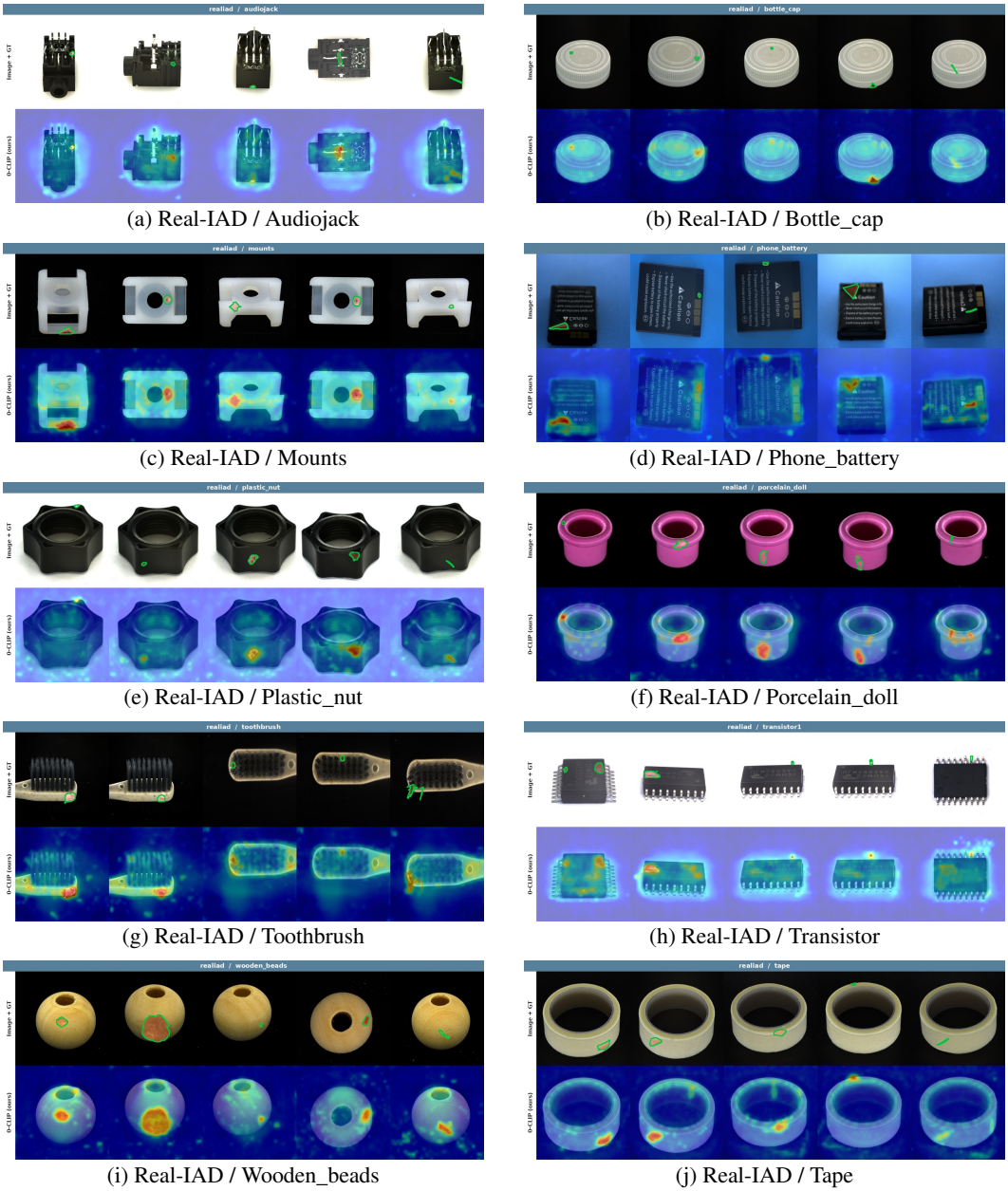


(i) MPDD / Bracket_black (j) MPDD / Tubes



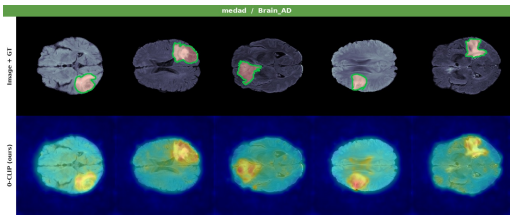
(k) SDD (l) KSDD2

Figure 14: **Per-category anomaly maps on the remaining industrial datasets.** Panels as in Figure 13, covering DTD-Synthetic, DAGM, BTAD, MPDD, SDD and KSDD2. SDD and KSDD2 are single-category datasets, so their panels carry no category name. On DAGM the maps also respond to the radial background texture away from the defect, and on SDD they brighten along the image borders.

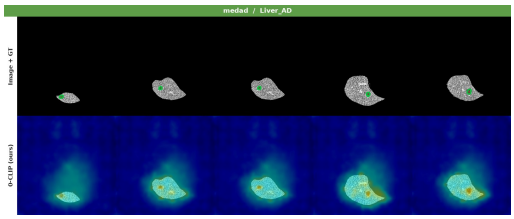


(a) Real-IAD / Audiojack
(b) Real-IAD / Bottle_cap
(c) Real-IAD / Mounts
(d) Real-IAD / Phone_battery
(e) Real-IAD / Plastic_nut
(f) Real-IAD / Porcelain_doll
(g) Real-IAD / Toothbrush
(h) Real-IAD / Transistor
(i) Real-IAD / Wooden_beads
(j) Real-IAD / Tape

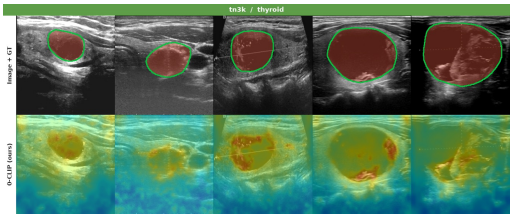
Figure 15: **Per-category anomaly maps on Real-IAD.** Panels as in Figure 13. The defects of Real-IAD are small relative to the object, and the maps spread over nearby object structure instead of tracking the mask boundary.



(a) Brain MRI (BMAD)



(b) Liver CT (BMAD)



(c) TN3K / Thyroid

Figure 16: **Per-category anomaly maps on the medical datasets.** Panels as in Figure 13. On Brain MRI the response spreads over nearby brain structure instead of tracking the mask boundary. On Liver CT it covers the organ broadly and peaks at the annotated lesion. On TN3K the nodules are large and low in contrast against the surrounding tissue, and the map responds over much of the field rather than resolving the nodule boundary.

93.7 P-AUROC). Looking at the misranked images, we found the same pattern each time: the rarity branch picks up something other than the defect.

On MPDD, nearly all of the drop comes from `bracket_white`. Its defects are small scratches and coating damage, but its normal parts vary a lot in pose, reflection and visible surfaces. The rarity branch follows that variation, so a normal bracket can light up along a whole edge while a scratched one stays dim. `metal_plate`, whose defects are large corroded areas, benefits from the same branch.

On Brain MRI, much of the test split is anomalous and several slices come from the same patient (Table 8). Slices from different positions of one brain always look different, whereas a small lesion often resembles ones already seen and stops being rare as the stream grows. Large lesions score about as well as offline, and online P-AUROC and PRO are both higher (Table 7). The stream helps locate a lesion but makes it harder to tell whether one is there.

On Br35H, images differ in contrast, cropping, orientation and watermarks. Among the misranked images, tumour-free ones with an unusual acquisition style score high, while ones with a clearly visible tumour score low.

The common cause is that rarity is measured by what recurs in the stream, and recurrence follows appearance, not labels. On most datasets this works in favour of the online variant. It works against it when normal images vary more than the anomalies do, or when anomalous appearance repeats.

Table 8: **The 15 evaluation datasets**, industrial (⚙) or medical (🏥). *Normal* and *Anomaly* count the test images used here. *Mask* and *Labels* mark the available ground truth, the latter being undefined on all-anomaly sets.

Dataset	Classes	Normal Samples	Anomaly Samples	Data type	Mask	Labels
⚙ *Industrial*						
MVTec-AD	15	467	1258	Obj. + Tex.	✓	✓
VisA	12	962	1200	Object	✓	✓
BTAD	3	451	290	Obj. + Tex.	✓	✓
MPDD	6	176	282	Object	✓	✓
DAGM	10	6996	1054	Texture	✓	✓
DTD-Synthetic	12	357	947	Texture	✓	✓
SDD	1	181	74	Texture	✓	✓
KSDD2	1	894	110	Texture	✓	✓
Real-IAD	30	63256	51329	Object	✓	✓
🏥 *Medical*						
Brain MRI (BMAD)	1	640	3075	Brain MRI	✓	✓
Liver CT (BMAD)	1	833	660	Liver CT	✓	✓
Retina RESC (BMAD)	1	1041	764	Retina OCT	✓	✓
Retina OCT (BMAD)	1	242	726	Retina OCT	✗	✓
Br35H	1	1500	1500	Brain MRI	✗	✓
TN3K	1	0	614	Thyroid ultrasound	✓	✗

F Dataset Statistics

Table 8 summarizes the evaluation datasets reported in Tables 1 and 2, including the number of sub-categories, per-image normal/anomaly counts of the test split, data type/modality, and the availability of pixel-level masks and image-level labels.