PROTODECO: DECOUPLING ANOMALY RANKING AND NORMALITY MEASUREMENT FOR CROSS-DOMAIN ZERO-SHOT ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-domain zero-shot anomaly detection must rank target images without target-domain anomaly annotations. However, the representation that supports anomaly ranking need not provide a suitable metric for measuring target normality. We present ProtoDeCo, a post-hoc transductive calibration method that separates these roles: a fixed host detector selects low-score target images as pseudo-normal candidates, while frozen Contrastive Language-Image Pre-training (CLIP) patch features construct category-specific prototypes and measure deviations from them. The deviation score is fused with the host image score; the host pixel map remains unchanged. The method uses no target labels or masks and updates no model parameters. In the matched comparison, Detector Proto and Frozen Proto use the same host ranking, pseudo-normal candidates, memory construction, query reduction, and score fusion; they differ only in the representation used for metric measurement. Across 12 industrial host–target pairs, Frozen Proto improves the uncalibrated host by an average of 0.88 area under the receiver operating characteristic curve (AUROC) points, with positive gains on all pairs. On eight distinct-stream pairs, it improves over Detector Proto by 0.23 points; the 95% confidence interval is 0.14 to 0.36 points. On four medical image-level targets, applying ProtoDeCo to the uncalibrated BD-Host increases mean AUROC and average precision (AP) from 80.53 and 85.57 to 81.17 and 85.85, with gains on three targets. These results support separating anomaly ranking from target-side metric measurement under the evaluated transductive protocol.

1 INTRODUCTION

Cross-domain zero-shot anomaly detection evaluates whether a detector transfers to target categories or domains unavailable during training. This setting is challenging in industrial inspection, where shifts may involve object categories, acquisition conditions, defect types, and visual appearance Zhang et al. (2023); Pemula et al. (2025). Contrastive Language-Image Pre-training (CLIP)-based methods improve transfer through language-guided prompts, local matching, visual context, and task-specific adaptation Radford et al. (2021); Jeong et al. (2023); Zhou et al. (2024); Cao et al. (2024); Hou et al. (2026). Their representations are typically evaluated by anomaly ranking and localization, and may also be reused to measure distances to target images treated as normal.

These uses impose different requirements. A decision representation should rank anomalous images above normal ones, whereas a metric representation should preserve useful target-domain neighborhoods. Image-level area under the receiver operating characteristic curve (AUROC) evaluates ranking quality Fawcett (2006), but does not establish that feature distances are informative under domain shift. Source-task adaptation may preserve anomaly ranking while changing local feature geometry Zhou et al. (2022); Yao et al. (2023); Khattak et al. (2023); Zhu et al. (2023); Kumar et al. (2022). We refer to this difference between deployment roles as the *judge–ruler mismatch*; it does not imply that detector features are semantically corrupted.

We study this distinction with ProtoDeCo, a post-hoc transductive calibration method, as illustrated in Figure 1. For each target category, the fixed host ranks the available unlabeled images and selects low-score images as pseudo-normal candidates. A frozen CLIP pathway extracts patch features,

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.

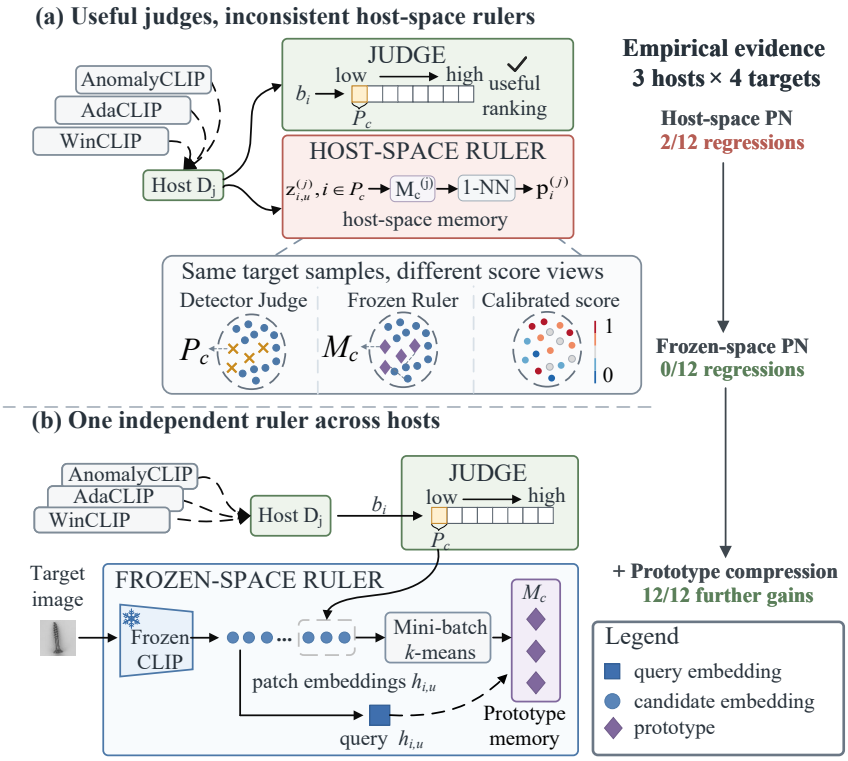


Figure 1: Decision and metric decoupling. The host ranks candidates, while frozen features construct and query pseudo-normal prototypes. PN denotes pseudo-normal; 1-NN denotes one-nearest-neighbor retrieval.

builds category-specific prototypes, and measures target deviations. The resulting score is normalized and fused with the host image score, while the host pixel map is retained.

The components used by ProtoDeCo, including pseudo-normal selection, frozen features, prototype memories, nearest-neighbor distances, and score fusion, are established techniques. Our contribution is to assign them to different deployment roles and evaluate this assignment with matched controls. Detector Proto and Frozen Proto use the same candidates, memory construction, query reduction, and score fusion; only the metric representation differs.

Our contributions are:

- We formulate the judge–ruler mismatch as a distinct transfer issue in cross-domain zero-shot anomaly detection, separating anomaly ranking from target-side normality measurement.
- We develop ProtoDeCo, a post-hoc transductive calibration method that retains the host detector for ranking and pseudo-normal selection while using frozen visual features for prototype-based measurement.
- We evaluate this role separation with matched Detector Proto and Frozen Proto controls across industrial and medical targets, while explicitly characterizing the transductive and localization limitations of the method.

2 Related Work

2.1 CLIP-based zero-shot anomaly detection

CLIP Radford et al. (2021) provides a frozen vision–language baseline for zero-shot anomaly detection. WinCLIP Jeong et al. (2023) adds prompt ensembles and local window matching. Anoma-

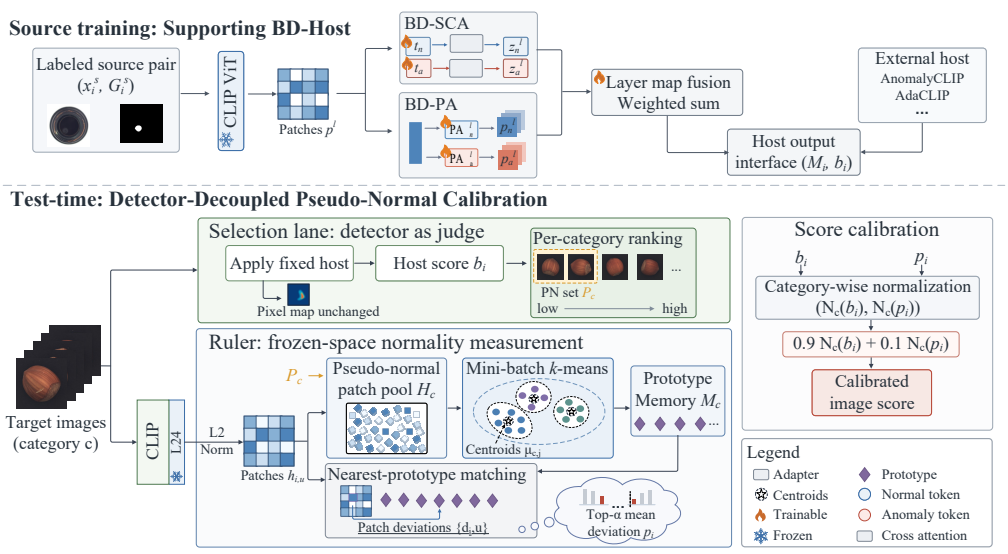


Figure 2: Host-agnostic test-time pipeline of ProtoDeCo. The fixed host acts as the judge by ranking target images and selecting category-wise pseudo-normal candidates, while an independent frozen visual pathway acts as the ruler by constructing prototypes and calibrating only the image-level score. ProtoDeCo does not modify the host parameters or pixel map.

lyCLIP Zhou et al. (2024) learns object-agnostic prompts, while AdaCLIP Cao et al. (2024) uses image-conditioned adaptation. S2SWCLIP Zhang et al. (2026) incorporates spatial-wavelet features. VisualAD Hou et al. (2026) and FB-CLIP Hu et al. (2026) use visual adaptation and foreground–background modeling, respectively. These methods primarily improve anomaly scores or localization maps; our study additionally asks whether their features provide a suitable metric for measuring target normality.

2.2 Memory-based anomaly detection and target-side calibration

PaDiM Defard et al. (2021) models normal patch-feature distributions, and PatchCore Roth et al. (2022) uses representative normal patches with nearest-neighbor retrieval. SoftPatch Jiang et al. (2022) reduces the effect of potentially anomalous samples during memory construction. These methods provide precedents for memory-based scoring, but generally assume access to normal data from the deployment domain or use the memory itself as the detector.

Recent methods also use unlabeled target collections or frozen representations at deployment. MuSc Li et al. (2024) constructs a transductive detector from relations among target images, while MRAD Xu et al. (2026) uses frozen visual representations for memory-driven image- and patch-level retrieval. DNPR Li et al. (2026) refines normal prototypes at deployment. These methods establish the usefulness of target-side memories or frozen features, but they do not define the same judge–ruler separation evaluated here. ProtoDeCo keeps the host detector fixed for target ranking and pseudo-normal selection, uses a separate frozen representation for prototype-based measurement, and retains the host score as part of the final image-level decision. It changes only the scalar image score and does not replace the host detector or its pixel map.

3 Method

3.1 Deployment Setting

Let D_θ be a source-trained anomaly detector whose parameters remain fixed at target deployment. The unlabeled target collection is $\mathcal{X} = \{(x_i, c_i)\}_{i=1}^{N_t}$, where x_i is an image and c_i is its category identifier. Category identifiers are used only to form within-category groups; target anomaly labels and segmentation masks are unavailable and are never used for calibration. For each image, the host

produces a pixel-level map M_i and an image-level score b_i:

$$(M_i, b_i) = D_\theta(x_i), \qquad M_i \in \mathbb{R}^{H \times W}. \tag{1}$$

The host score ranks target images; the map is retained as the localization output.

3.2 Pseudo-normal Selection and Frozen Features

For each category c, let $\mathscr{I}_c = \{i : c_i = c\}$ and $n_c = |\mathscr{I}_c|$. Sort its images by host score, so that

$$b_{\pi_c(1)} \le b_{\pi_c(2)} \le \cdots \le b_{\pi_c(n_c)}. \tag{2}$$

The lowest-scoring images form the pseudo-normal candidate set:

$$\mathscr{P}_c = \{\pi_c(j) : 1 \le j \le \max(1, \mathrm{round}(\rho n_c))\}, \qquad \rho = 0.1. \tag{3}$$

Selection uses only the host ranking, not target labels, masks, or metric-space distances. Host image scores are obtained by top-k aggregation of the host map, using ratios of 1% for industrial targets and 5% for medical targets.

The metric pathway uses a standard frozen Contrastive Language-Image Pre-training (CLIP) visual forward Radford et al. (2021). It does not inject the host's anomaly token, normal token, task token, adapter, or other detector-specific modules. We use patch features from layer 24, discard the class token, and retain at most $P_{\max} = 256$ patch positions per image. Each retained feature is ℓ_2-normalized:

$$h_{i,u} = \frac{f_{i,u}}{\max(\|f_{i,u}\|_2, \varepsilon)}, \qquad \varepsilon = 10^{-8}. \tag{4}$$

Patch subsampling uses a fixed per-image seed.

3.3 Prototype Memory and Deviation Score

For category c, pool the normalized patch features from its pseudo-normal candidates:

$$\mathscr{H}_c = \{h_{i,u} : i \in \mathscr{P}_c,\ u \in \mathscr{U}_i\}. \tag{5}$$

The memory contains at most $K_p = 128$ representatives. If the pool is larger, MiniBatchKMeans produces 128 centroids; otherwise, the pooled features are retained. Centroids are L2-normalized before scoring.

For a query patch, define its deviation from the nearest prototype by cosine distance:

$$d_{i,u} = 1 - \max_j \langle h_{i,u}, m_{c,j} \rangle. \tag{6}$$

Let $U_i = |\mathscr{U}_i|$ and $r_i = \max\{1, \mathrm{round}(\alpha U_i)\}$. Let Top_{r_i} return the r_i largest values in the multiset of patch deviations. The image-level prototype score is

$$p_i = \frac{1}{r_i} \sum_{d \in \mathrm{Top}_{r_i}(\{d_{i,u} : u \in \mathscr{U}_i\})} d, \qquad \alpha = 0.1. \tag{7}$$

3.4 Score Fusion and Matched Control

For a score vector s and category c, define

$$s_c^{\min} = \min_{k \in \mathscr{I}_c} s_k, \qquad s_c^{\max} = \max_{k \in \mathscr{I}_c} s_k.$$

The category-wise normalization is

$$\mathscr{N}_c(s_i) = \begin{cases} \dfrac{s_i - s_c^{\min}}{s_c^{\max} - s_c^{\min}}, & s_c^{\max} > s_c^{\min}, \\ 0.5, & s_c^{\max} = s_c^{\min}. \end{cases} \tag{8}$$

Host and prototype scores are normalized separately and fused as

$$\widetilde{b}_i = (1 - w)\mathscr{N}_{c_i}(b_i) + w\mathscr{N}_{c_i}(p_i), \qquad w = 0.1. \tag{9}$$

The main experiments use $w = 0.1$ as the conservative deployment setting, while the $w = 0.3$ results are used only for diagnostic stress tests. The host score therefore remains the dominant component of the final image-level score, and the host pixel map is unchanged.

Detector Proto is the matched metric-space control. It uses the same host ranking, pseudo-normal candidates, layer-24 patch sampling, memory construction, query reduction, normalization, and fusion as Frozen Proto, but obtains candidate and query embeddings from the trained detector's token-augmented patch-token stream. This stream is shaped by the learned task tokens used by the host, and no additional metric-specific projection is introduced. Frozen Proto changes only the ruler: it obtains layer-24 patch tokens from a standard frozen CLIP visual forward, discards the class token, L2-normalizes the tokens, and applies the same prototype construction and nearest-prototype scoring. The comparison therefore isolates the representation used for target-side normality measurement while holding candidate selection and score computation fixed.

The procedure is transductive within each target category: the unlabeled target collection is used for candidate selection, memory construction, and score normalization. Category identifiers are used only for grouping, whereas anomaly labels and segmentation masks are accessed only for evaluation. The procedure performs no gradient updates. The deployment pipeline is summarized in Figure 2.

4 Experiments

4.1 Experimental Setup

We evaluate ProtoDeCo under a transductive cross-domain zero-shot protocol. Target images and their category identifiers are used for pseudo-normal candidate selection, prototype construction, and category-wise score normalization; anomaly labels and segmentation masks are not used for calibration.

Industrial targets are MVTec AD Bergmann et al. (2019), VisA Zou et al. (2022), BTAD Mishra et al. (2021), KSDD2 Tabernik et al. (2020), DAGM2007 Wieler & Hahn (2007), and DTD-Synthetic Aota et al. (2023). Learned detectors use VisA-trained checkpoints on the other industrial targets and MVTec-trained checkpoints on VisA, without target fine-tuning. Medical targets are OCT2017 Kermany et al. (2018), BrainMRI Salehi et al. (2021), Brain-AD Baid et al. (2021); Bakas et al. (2017); Menze et al. (2014), and Histopathology Ehteshami Bejnordi et al. (2017); these experiments use the corresponding VisA-trained checkpoints.

The industrial comparison includes CLIP Radford et al. (2021), WinCLIP Jeong et al. (2023), AnomalyCLIP Zhou et al. (2024), AdaCLIP Cao et al. (2024), S2SWCLIP Zhang et al. (2026), VisualAD Hou et al. (2026), and FB-CLIP Hu et al. (2026). We report category-macro image-level area under the receiver operating characteristic curve (AUROC) and average precision (AP). Learned methods report the mean and population standard deviation over three aligned seeds when available; deterministic baselines are reported once. Unless stated otherwise, the final configuration uses $\rho = 0.1$, $P_{\max} = 256$, $K_p = 128$, top-10% patch reduction, one-nearest-neighbor (1-NN) retrieval, and fusion weight $w = 0.1$. These calibration settings are fixed without target validation.

4.2 Cross-Domain Generalization on Industrial Benchmarks

Table 1 reports image-level results on six industrial targets. BD-Host with ProtoDeCo obtains mean AUROC and AP of 94.92 and 96.23, respectively. It has the highest AUROC on five targets; FB-CLIP leads on KSDD2. It has the highest AP on five targets; VisualAD leads on BTAD. These are system-level results combining the supporting host and the calibration procedure; the matched host study below isolates the effect of ProtoDeCo from the effect of the host architecture.

4.3 Cross-Domain Transfer to Medical Imaging

Table 2 compares image-level performance on four medical targets. BD-Host with ProtoDeCo reaches mean AUROC and AP of 81.17 and 85.85. The paired calibration comparison shows gains on OCT2017, BrainMRI, and Histopathology, while Brain-AD does not improve. We therefore treat this as an aggregate transfer result, not a uniform advantage across medical targets.

Metric	Method	MVTec	VisA	BTAD	KSDD2	DAGM	DTD Synthetic
AUROC	CLIP	88.57	82.52	60.78	90.96	90.83	96.54
	WinCLIP	90.32 ± 0.02	75.45 ± 0.02	67.42 ± 0.04	93.54 ± 0.03	91.67 ± 0.02	94.97 ± 0.01
	AnomalyCLIP	91.52 ± 0.16	81.48 ± 0.38	89.00 ± 0.10	92.31 ± 0.32	98.03 ± 0.05	94.04 ± 0.30
	AdaCLIP	91.36 ± 2.52	81.51 ± 2.93	91.19 ± 3.59	95.01 ± 0.88	98.22 ± 0.14	91.78 ± 2.56
	S2SWCLIP	90.80 ± 0.36	82.83 ± 0.12	88.40 ± 0.22	91.73 ± 0.33	97.80 ± 0.16	96.87 ± 0.05
	VisualAD	92.00 ± 0.16	81.90 ± 0.08	94.33 ± 0.82	96.27 ± 0.74	98.30 ± 0.14	94.83 ± 0.74
	FB-CLIP	90.30 ± 0.33	83.38 ± 0.68	92.13 ± 0.83	**98.06 ± 0.26**	97.64 ± 0.55	92.93 ± 1.71
	Ours†	**92.65 ± 0.13**	**86.04 ± 0.66**	**95.71 ± 0.44**	97.97 ± 0.33	**99.22 ± 0.34**	**97.93 ± 0.49**
AP	CLIP	94.71	85.57	64.72	92.54	77.12	98.43
	WinCLIP	95.53 ± 0.02	78.53 ± 0.03	70.17 ± 0.11	94.24 ± 0.02	79.05 ± 0.03	97.61 ± 0.01
	AnomalyCLIP	96.23 ± 0.11	84.90 ± 0.34	91.49 ± 0.31	93.78 ± 0.23	92.43 ± 0.05	97.51 ± 0.14
	AdaCLIP	95.84 ± 1.40	84.90 ± 2.45	92.19 ± 6.63	95.98 ± 0.77	93.42 ± 0.90	96.57 ± 1.46
	S2SWCLIP	95.63 ± 0.12	85.43 ± 0.09	89.97 ± 0.45	93.60 ± 0.24	92.17 ± 0.37	98.67 ± 0.05
	VisualAD	96.63 ± 0.05	85.70 ± 0.16	**97.87 ± 0.62**	96.80 ± 0.54	94.47 ± 0.68	97.73 ± 0.40
	FB-CLIP	95.67 ± 0.14	85.79 ± 0.33	94.02 ± 1.38	98.30 ± 0.21	93.42 ± 1.68	97.45 ± 0.66
	Ours†	**96.74 ± 0.18**	**88.52 ± 0.62**	97.72 ± 0.46	**98.35 ± 0.24**	**96.75 ± 1.39**	**99.32 ± 0.10**

Table 1: Industrial image-level results (%). Ours† denotes the proposed calibration method evaluated using BD-Host as the fixed supporting host. BD-Host provides image ranking and pseudo-normal candidate selection; the proposed method performs target-side metric calibration. Learned methods report three-seed mean ± population standard deviation when available; other baselines are reported once. Bold and underline mark the top two.

Metric	Method	OCT2017	BrainMRI	Brain-AD	Histopathology	Mean
AUROC	CLIP	26.06	86.10	66.27	53.40	57.96
	WinCLIP	55.49 ± 0.01	86.54 ± 0.04	72.39 ± 0.01	**76.91 ± 0.00**	72.83
	AnomalyCLIP	66.53 ± 3.70	**96.28 ± 0.14**	69.46 ± 0.93	59.07 ± 0.74	72.84
	AdaCLIP	68.66 ± 9.01	90.89 ± 0.57	**79.36 ± 1.76**	63.42 ± 1.14	75.58
	S2SWCLIP	72.63 ± 0.31	95.73 ± 0.05	68.23 ± 0.86	58.57 ± 1.05	73.79
	VisualAD	**91.57 ± 0.88**	94.87 ± 0.39	77.77 ± 1.39	54.73 ± 0.88	79.73
	FB-CLIP	81.16 ± 2.27	90.16 ± 1.29	71.90 ± 2.21	57.87 ± 2.01	75.28
	Ours†	89.65 ± 2.29	96.11 ± 0.57	78.38 ± 1.29	60.52 ± 1.79	**81.17**
AP	CLIP	63.75	84.47	89.44	51.03	72.17
	WinCLIP	81.30 ± 0.00	91.25 ± 0.05	91.51 ± 0.01	**70.70 ± 0.01**	83.69
	AnomalyCLIP	87.56 ± 1.52	**97.27 ± 0.10**	90.28 ± 0.30	58.99 ± 0.62	83.53
	AdaCLIP	87.38 ± 4.48	93.85 ± 0.74	**94.38 ± 0.61**	59.04 ± 1.52	83.66
	S2SWCLIP	89.70 ± 0.16	97.07 ± 0.09	89.53 ± 0.50	57.93 ± 0.65	83.56
	VisualAD	**97.17 ± 0.31**	94.13 ± 0.31	93.47 ± 0.66	51.90 ± 0.78	84.17
	FB-CLIP	92.37 ± 1.22	91.43 ± 1.07	91.20 ± 1.03	54.41 ± 0.70	82.35
	Ours†	96.69 ± 0.74	96.96 ± 0.58	93.76 ± 0.89	56.01 ± 1.46	**85.85**

Table 2: Medical image-level results (%). Ours† denotes the proposed calibration method evaluated using the same fixed BD-Host and protocol as Table 1. BD-Host provides image ranking and pseudo-normal candidate selection; the proposed method performs target-side metric calibration. Bold and underline mark the top two.

Localization scope. Table 4 makes explicit which part of the system is affected by ProtoDeCo. Since ProtoDeCo changes only the scalar image-level score, the pixel output of Ours is exactly the unchanged BD-Host map. On the industrial targets, this map obtains 94.82 pixel AUROC and 48.04 pixel AP; it is not the best pixel AUROC, but it obtains the highest pixel AP in the table. On the medical localization targets, it obtains 85.24 pixel AUROC and 40.68 pixel AP, whereas FB-CLIP obtains 89.89 and 52.64, respectively. These results are consistent with the intended scope of the method: ProtoDeCo improves target-side image-level normality measurement while leaving localization to the supporting host. Thus, the image-level gains should not be interpreted as gains from a modified pixel-level detector.

4.4 Ablation Studies

The ablations examine host design, candidate selection, metric space, prototype construction, and query inclusion. Unless noted otherwise, comparisons use the same target images, host scores, normalization, and score fusion.

(a) Host comparison

Method	Industrial	Medical	Overall
baseline	92.96 ± 0.21	79.73 ± 0.58	87.66 ± 0.35
BD-SC	94.54 ± 0.32	**81.37 ± 1.00**	89.27 ± 0.31
BD	**94.92 ± 0.19**	81.17 ± 0.55	**89.42 ± 0.27**

(b) ProtoDeCo calibration

Scope	BD (w/o)	BD (w/)
Industrial mean	94.72/96.03	94.92/96.23
Medical mean	80.53/85.57	81.17/85.85

Table 3: Host-architecture and ProtoDeCo ablations with aligned seeds. Panel (a) reports image AUROC as mean ± population standard deviation. Industrial, Medical, and Overall denote the unweighted means over six industrial, four medical, and ten image-level datasets, respectively. BD-SC is the semi-decoupled host, whereas BD is the fully decoupled host used in the main results. Panel (b) reports BD image AUROC/AP, in that order, without and with ProtoDeCo.

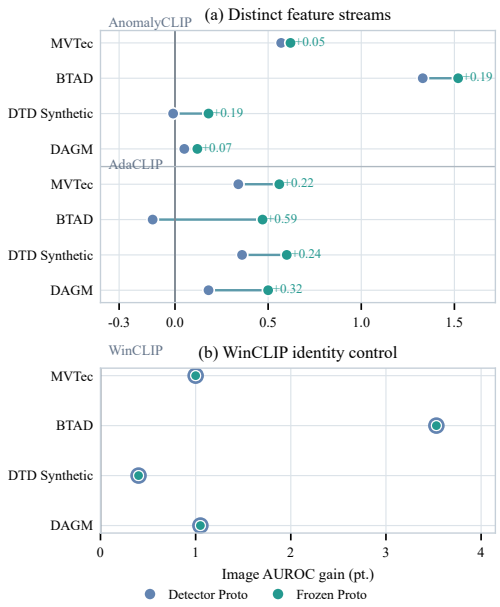


Figure 3: Paired image AUROC gains relative to the uncalibrated host at $w = 0.1$. Panel (a) compares Detector Proto and Frozen Proto on the eight distinct-stream host–target pairs; panel (b) is the shared-stream WinCLIP identity control. Colors denote Detector Proto and Frozen Proto, respectively.

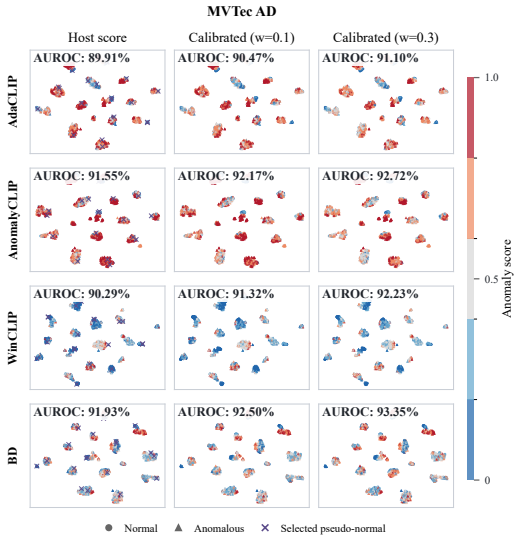


Figure 4: Fixed-coordinate t-distributed stochastic neighbor embedding (t-SNE) visualization of MVTec scores for BD-Host seed 111. The three columns show the host score and the calibrated scores at $w = 0.1$ and $w = 0.3$ using the same two-dimensional coordinates; marker shapes identify normal, anomalous, and selected pseudo-normal samples, while colors encode normalized anomaly scores. The visualization is qualitative, and the quantitative results use $w = 0.1$.

Host and calibration. Table 3 reports the BD-SC and BD host comparison and the image-level results before and after ProtoDeCo. Calibration changes image-level scores only; it does not alter the host pixel map.

Candidate source. At the diagnostic weight $w = 0.3$, low-score selection improves mean AUROC by 2.03 points, whereas random selection has a positive mean gain of 1.17 points despite mixed per-case results; high-score selection decreases mean AUROC by 1.82 points. The label-informed oracle gain is 4.09 points and is not available at deployment. The final protocol uses $w = 0.1$. The compact candidate-source and query-inclusion controls are summarized in Table 9; detailed candidate-source and cross-fitting results are reported in the supplementary material.

Metric space. Figure 3 compares the paired industrial host–target results. In the matched external-host comparison, Frozen Proto improves over the paired uncalibrated host on all 12 industrial point estimates, with a mean gain of 0.88 AUROC points. Across the eight distinct-stream pairs, it improves over Detector Proto by 0.23 points; the 95% confidence interval is 0.14 to 0.36 points. The

(a) Industrial localization

Method	Pixel AUROC	Pixel AP
CLIP	36.65	1.58
WinCLIP	77.33 ± 0.00	11.75 ± 0.00
AnomalyCLIP	**95.50 ± 0.03**	41.04 ± 0.12
AdaCLIP	94.42 ± 1.67	46.92 ± 4.61
S2SWCLIP	95.38 ± 0.07	39.95 ± 0.44
VisualAD	94.74 ± 0.20	47.97 ± 0.66
FB-CLIP	94.68 ± 0.91	44.82 ± 2.82
Ours	94.82 ± 0.16	**48.04 ± 1.81**

(b) Medical localization

Method	AUROC	AP
CLIP	56.43	15.05
WinCLIP	72.98 ± 0.00	19.40 ± 0.00
AnomalyCLIP	87.46 ± 0.09	48.23 ± 0.20
AdaCLIP	83.73 ± 4.21	38.67 ± 7.63
S2SWCLIP	86.56 ± 0.06	45.89 ± 0.47
VisualAD	86.65 ± 0.15	40.78 ± 0.83
FB-CLIP	**89.89 ± 0.11**	**52.64 ± 0.86**
Ours	85.24 ± 0.80	40.68 ± 1.79

Table 4: Macro pixel-level localization results (%). Learned methods report three-seed mean ± population standard deviation. BD-Host is the unchanged supporting-host pixel output; ProtoDeCo changes only the image-level score. Bold and underline denote the best and second-best results, respectively.

Table 5: BTAD feature-space diagnostic. PN std is pseudo-normal score standard deviation; Raw uses the uncompressed memory and Prototype uses $K_p = 128$. Δ is AUROC gain over the host. Diagnostic $w = 0.3$; final $w = 0.1$.

Host	PN std		Raw Δ		Prototype Δ	
	Detector	Frozen	Detector	Frozen	Detector	Frozen
WinCLIP	0.1034	0.1034	+7.08	+7.26	+9.97	+9.43
AnomalyCLIP	0.0334	0.0729	+3.06	+2.64	+3.34	+3.66
AdaCLIP	0.0041	0.0781	−0.45	+1.14	−1.74	+1.66

Table 6: Target-batch stress test. Mean and Worst are AUROC gains over the host; Negative is the fraction of resamples below zero.

Condition	Mean	Worst	Negative
Batch size 8	−0.05	−6.25	0.02
Anomaly ratio 75%	+0.08	−2.86	0.21
Mixed category batches	+0.06	−5.56	0.13
Single category batches	+0.17	−6.25	0.22

Table 7: DINOv2 ViT-L/14 portability on BD-Host. Entries are image AUROC (%, mean ± standard deviation (SD) over three seeds); other settings follow the final protocol.

Dataset	Base	Calibrated	Δ
MVTec	86.94 ± 1.33	87.85 ± 1.47	+0.91 ± 0.19
BTAD	85.50 ± 0.63	86.64 ± 0.61	+1.14 ± 0.20
KSDD2	98.95 ± 0.40	98.88 ± 0.42	−0.07 ± 0.03
DAGM	88.87 ± 2.83	90.15 ± 2.27	+1.28 ± 0.55
DTD Synthetic	93.65 ± 2.11	94.02 ± 1.97	+0.37 ± 0.14
VisA	82.48 ± 0.71	83.15 ± 0.72	+0.67 ± 0.02
Macro mean	89.40	90.12	**+0.72**

Table 8: Prototype compression at $w = 0.1$. Macro image AUROC/AP (%) over 12 industrial pairs at seed 111. Gain is Proto minus Raw in points; Positive pairs counts cases in which Proto exceeds Raw. Single-seed descriptive summary.

Metric	Frozen Raw	Frozen Proto	Gain	Positive pairs
AUROC	90.98	91.33	+0.35	12/12
AP	91.38	91.69	+0.31	12/12

four WinCLIP Jeong et al. (2023) pairs use a shared score cache and are identity controls. Figure 4 provides a qualitative fixed-coordinate view of how the calibrated score changes the score ordering, while the quantitative results use the conservative setting $w = 0.1$. The complete paired intervals and medical paired results are reported in Appendix Table 21.

Backbone portability. Table 7 tests whether the role separation depends on Contrastive Language–Image Pre-training (CLIP) Radford et al. (2021). Replacing the frozen CLIP ruler with DINOv2 Oquab et al. (2023) ViT-L/14 increases the mean image AUROC from 89.40 to 90.12, a gain of 0.72 points. Fifteen of the 18 seed–target pairs improve, and five of the six industrial targets have positive mean gains. KSDD2 decreases by 0.07 points, so this result supports portability but not a backbone-independent non-regression claim.

Mechanistic interpretation. The feature-space diagnostics in Table 5 provide a possible explanation for the judge–ruler mismatch. On AdaCLIP Cao et al. (2024) with BTAD, the pseudo-normal score standard deviation is 0.0041 in the detector space but 0.0781 in the frozen space. Under the same diagnostic setting with $w = 0.3$, Detector Proto decreases AUROC by 1.74 points, whereas Frozen Proto improves it by 1.66 points. The same direction is observed without prototype compression: Detector Raw decreases AUROC by 0.45 points, while Frozen Raw improves it by 1.14 points. This controlled reversal is consistent with the judge–ruler mismatch: detector adaptation can preserve task-specific ranking while producing a representation that is less suitable for distance-

based target calibration. These diagnostics are interpretive rather than causal evidence and do not imply that detector features are semantically corrupted.

Prototype construction and query inclusion. Table 8 isolates prototype compression under the final setting $w = 0.1$. With identical candidates, frozen features, query reduction, retrieval, and fusion, Frozen Proto increases the mean AUROC from 90.98 to 91.33 and the mean AP from 91.38 to 91.69 over 12 industrial host–target pairs. Both metrics improve in all 12 pairs. Because this audit uses seed 111 for stochastic prototype construction, it is a descriptive point-estimate check rather than a cross-seed significance claim. We also test whether a query benefits simply from being included in its own reference memory. Leave-one-out scoring retains 95%, 107%, and 107% of the self-memory gains for AnomalyCLIP Zhou et al. (2024), AdaCLIP Cao et al. (2024), and BD-Host, respectively. On BD-Host, the self-memory, two-fold cross-fitting, and leave-one-out protocols improve 14, 13, and 14 of the 18 seed–target pairs. These controls indicate that neither prototype compression nor direct query inclusion alone explains the observed calibration effect.

(a) Candidate source
BrainMRI AUROC (%), $w = 0.3$

Host	Base	Rand.	High	Low	Oracle†
WinCLIP	86.5	86.8	85.1	87.2	87.7
AnomalyCLIP	96.1	94.4	93.0	96.3	96.3
AdaCLIP	95.3	94.1	92.5	95.6	95.6

(b) Query inclusion
Mean AUROC gain, $w = 0.1$

Host	*n*	Self	2-fold	LOO
AnomalyCLIP	4	0.61	0.63	0.58
AdaCLIP	4	0.53	0.55	0.57
BD-Host	18	0.20	0.20	0.21

Table 9: Target-side controls. Panel (a) reports BrainMRI image AUROC; the candidate-source aggregate gains across evaluated cases are reported in the text. Panel (b) reports image-AUROC gains over the uncalibrated host. Here, *n* is four industrial targets for AnomalyCLIP and AdaCLIP, and 18 seed–target pairs for BD-Host. LOO denotes leave-one-out. † Oracle selection uses anomaly labels and is not deployable.

4.5 Further Analysis and Limitations

ProtoDeCo is transductive and requires the unlabeled target collection for candidate selection and category-wise score normalization. It changes only the scalar image-level score and leaves the host pixel map unchanged. On 1,725 MVTec images with batch size one, adding the frozen metric pathway increases end-to-end score generation time from 233.59 ± 0.16 to 354.72 ± 0.18 ms per image, an overhead of 121.13 ± 0.26 ms per image, while peak allocated GPU memory remains 2340.6 MiB. These measurements exclude model loading and offline metric construction. The target-batch stress test summarized in Table 6 exposes an operating limitation: with a batch size of eight images, the mean AUROC gain becomes -0.05 points, with a worst-case loss of 6.25 points; at a 75% anomaly ratio, the mean gain remains positive at 0.08 points, but 21% of resamples produce a negative gain. These results indicate that candidate selection and category-wise normalization require a sufficiently large and reasonably representative target batch.

5 Conclusion

We studied whether the representation used to rank anomalies should also measure target normality. ProtoDeCo retains the fixed host for pseudo-normal selection and uses frozen Contrastive Language-Image Pre-training (CLIP) features Radford et al. (2021) for prototype construction and distance measurement. In matched industrial experiments, Frozen Proto outperformed the uncalibrated hosts on all 12 host-target AUROC point estimates, and in eight pairs of different flow combinations, its performance was on average better than Detector Proto. Medical results show an aggregate gain across four image-level targets, but not a uniform improvement.

These findings support treating anomaly ranking and target-side metric measurement as distinct transfer requirements. They do not imply that detector features are semantically corrupted or that frozen features are universally superior. The method depends on the evaluated transductive protocol, requires the unlabeled target collection, and changes only image-level scores; improving pixel-level localization remains outside its scope.

AT USE STATEMENT

Generative AI tools were used solely for language editing of the manuscript, including improving grammar, clarity, readability, and sentence flow. They were not used to develop the research methodology, formulate hypotheses, design or conduct experiments, prepare or analyze data, interpret results, or generate research claims, tables, or figures. The authors reviewed and revised all AI-assisted edits to ensure that the scientific meaning was preserved. The authors take full responsibility for the methodology, results, interpretations, and final content of this work.

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.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4 (1):170117, 2017.

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.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

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.

Thomas Defard, Aleksandr Setkov, Angelique 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.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, CAMELYON16 consortium, Meyke Hermsen, Quirine F Manson, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

Steven A Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo L Hammer, and Michael A Riegler. The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In *International Conference on Pattern Recognition*, pp. 263–274. Springer, 2021.

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

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

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19606–19616, 2023.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pp. 451–462. Springer, 2019.

Xi Jiang, Jianlin Liu, Jinbao Wang, Qiang Nie, Kai Wu, Yong Liu, Chengjie Wang, and Feng Zheng. Softpatch: Unsupervised anomaly detection with noisy data. *Advances in Neural Information Processing Systems*, 35:15433–15445, 2022.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15190–15200, 2023.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Shuyun Li, Zhi Li, Weidong Wang, Long Zheng, and Yu Lu. Dnpr: Zero-shot industrial anomaly detection via dynamic normal prototype refinement. *Expert Systems with Applications*, pp. 131331, 2026.

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 *The Twelfth International Conference on Learning Representations*, 2024.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2014.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. *arXiv preprint arXiv:2104.10036*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Robust ad: A real world benchmark dataset for robustness in industrial anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4047–4057, 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*, pp. 8748–8763. PmLR, 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.

Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14902–14912, 2021.

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.

Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, volume 6, pp. 11, 2007.

Chaoran Xu, Chengkan Lv, Qiyu Chen, Feng Zhang, and Zhengtao Zhang. Mrad: Zero-shot anomaly detection with memory-driven retrieval. *arXiv preprint arXiv:2602.00522*, 2026.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6757–6767, 2023.

Huan Zhang, Chunlei Wu, Jing Lu, and Mengyuan Jing. S2SWCLIP: Semantic-optimized prompts with spatial-wavelet synergy for zero-shot anomaly detection. *Scientific Reports*, 16(1):13062, 2026. doi: 10.1038/s41598-026-43044-3.

Zilong Zhang, Zhibin Zhao, Xingwu Zhang, Chuang Sun, and Xuefeng Chen. Industrial anomaly detection with domain shift: A real-world dataset and masked multi-scale reconstruction. *Computers in Industry*, 151:103990, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022.

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*, volume 2024, pp. 49705–49737, 2024.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15659–15669, 2023.

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*, pp. 392–408. Springer, 2022.

A Scope and Protocol Notes

This supplement provides the supporting protocol details, diagnostic analyses, and run-level audits for ProtoDeCo. The main paper presents the problem formulation, the proposed judge–ruler separation, the deployment procedure, the primary industrial and medical comparisons, and the central ablations. The supplement gives the aligned dataset statistics, host and calibration controls, candidate-source and query-inclusion tests, parameter sensitivity, backbone portability, batch-size stress tests, feature-space diagnostics, localization consistency, runtime measurements, and comparisons with related transductive methods. All supplementary experiments follow the same transductive setting as the main paper: target images and category identifiers may be used to construct pseudo-normal references and category-wise score normalization, whereas target anomaly labels and pixel masks are never used for calibration. Unless explicitly stated otherwise, the final deployment configuration uses pseudo-normal ratio $\rho = 0.1$, at most $P_{\max} = 256$ retained patch features per image, $K_p = 128$ prototype patches, top-10% query reduction, one-nearest-neighbor retrieval, and fusion weight $w = 0.1$. Diagnostic experiments use $w = 0.3$ only when a larger calibration signal is needed to expose the behavior of a controlled variant; this diagnostic setting does not replace the final deployment setting. Aggregate learned-method results are reported as the mean and population standard deviation over the aligned seeds when the corresponding runs are available, while the seed-wise audit tables report the individual records directly. To fit the ICLR single-column layout, the long seed-wise audit tables are displayed as two vertically stacked six-column panels within one table: panel (a) contains WinCLIP, AnomalyCLIP, and AdaCLIP, and panel (b) contains VisualAD, FB-CLIP, and Ours. The two panels repeat the Domain, Metric, and Dataset columns and preserve the same row order; this is only

Table 10: Test-split composition of the six industrial and seven medical datasets. The number in parentheses after each dataset gives its category count. Total, Normal, and Anomalous report the total number of test images and their normal/anomalous counts, respectively. The MVTec AD and VisA test splits are also used as source training data for the corresponding cross-domain runs; the other datasets are evaluation-only. Label indicates the available annotation type: image-level labels (Image) or pixel-level masks (Mask).

Dataset	Test total	Normal	Anomalous	Image size	Label
Industrial					
MVTec AD (15)	1,725	467	1,258	Variable	Mask
VisA (12)	2,162	962	1,200	Variable	Mask
BTAD (3)	741	451	290	Variable	Mask
KSDD2 (1)	712	356	356	Variable	Mask
DAGM2007 (10)	8,050	6,996	1,054	512×512	Mask
DTD Synthetic (12)	1,304	357	947	Variable	Mask
Medical					
OCT2017 (1)	968	242	726	$512/768 \times 496$	Image
BrainMRI (1)	253	98	155	Variable	Image
Brain-AD (1)	3,715	640	3,075	240×240	Mask
Histopathology (1)	1,997	1,003	994	256×256	Image
CVC-ClinicDB (1)	612	0	612	384×288	Mask
EndoTect (1)	200	0	200	Variable	Mask
Kvasir (1)	1,000	0	1,000	Variable	Mask

a presentation split and does not introduce additional experiments or aggregation. Each audit cell reports the ordered tuple (AUROC, maximum F1 score (F1-max), AP) in percent. Bold and underline identify the best and second-best values for each metric in the original full-method comparison, respectively. A dash denotes an unavailable source record, and no missing value is reconstructed from an aggregate result. Pixel-area under the per-region overlap curve (AUPRO) is omitted from the seed-wise audit because the available AUPRO fields are placeholders rather than valid measurements. Medical results are included as exploratory transfer evidence under the same protocol and should not be interpreted as clinical validation or as a claim of modality-independent performance.

B Dataset and Evaluation Structure

We use aligned evaluation splits to separate source training from target-domain evaluation. The industrial targets are MVTec AD, VisA, BTAD, KSDD2, DAGM2007, and DTD Synthetic. MVTec AD and VisA are used as source datasets in the corresponding cross-domain runs, while the remaining industrial datasets are evaluation-only targets. The medical targets are OCT2017, BrainMRI, Brain-AD, Histopathology, CVC-ClinicDB, EndoTect, and Kvasir; these datasets are evaluated using the corresponding VisA-trained checkpoints without target fine-tuning. The number in parentheses after each dataset name denotes the number of categories. The test counts report the total number of images and the corresponding normal and anomalous counts. Image sizes refer to the native resolution before common model preprocessing, and "variable" indicates that the dataset contains multiple native resolutions. The available annotation type is listed as Image for image-level labels and Mask for pixel-level masks.

The first four medical datasets support image-level evaluation. Brain-AD also provides pixel-level masks and is therefore included in both image-level and pixel-level analyses. CVC-ClinicDB, EndoTect, and Kvasir contain anomalous test images with pixel-level masks but no normal test images in the aligned protocol, so they contribute only to pixel-level evaluation. ProtoDeCo uses the unlabeled target images grouped by category to select pseudo-normal candidates, construct prototype memories, and normalize scores; the labels and masks summarized in Table 10 are never used during calibration.

C SUPPORTING BD-HOST FORMULATION

ProtoDeCo is host-agnostic. The BD-Host used in the main results is a supporting source-trained detector whose anomaly and normal branches use separate evidence transformations. We denote this fully decoupled host as BD. The semi-decoupled control is denoted BD-SC and differs only in the parameter sharing of the spatial cross-attention module. Let $P^l = [p_1^l, \ldots, p_{T_l}^l]^\top \in \mathbb{R}^{T_l \times d}$ denote the patch tokens at layer l, and let $t_b \in \mathbb{R}^d$ denote the final task token for branch $b \in \{a, n\}$, where a and n denote the anomaly and normal branches, respectively.

Spatial cross-attention. For the fully decoupled BD host, each branch and layer has its own R learnable anchors A_b^l, positional encoding E_b^l, projections, normalization layers, output projection, gate, and residual scale. The branch-specific spatial cross-attention is

$$Q_b^l = \mathrm{LN}_{q,b}^l(W_{q,b}^l A_b^l), \qquad K_b^l = \mathrm{LN}_{k,b}^l\left(W_{k,b}^l(P^l + E_b^l)\right), \tag{10}$$

$$V_b^l = W_{v,b}^l P^l, \qquad H_b^l = \mathrm{softmax}\left(\frac{Q_b^l (K_b^l)^\top}{\sqrt{d}}\right) V_b^l. \tag{11}$$

The task token controls the contribution of the anchors through

$$g_b^l = \sigma(G_b^l t_b), \tag{12}$$

and the enhanced task token is computed as

$$z_b^l = t_b + \gamma_b^l W_{o,b}^l \sum_{r=1}^{R} g_{b,r}^l H_{b,r}^l. \tag{13}$$

Thus, BD allows the anomaly and normal branches to learn independent spatial evidence transformations.

BD-SC retains the same computation but shares the SCA core across the two branches. In particular, W_q^l, W_k^l, W_v^l, E^l, LN_q^l, LN_k^l, and W_o^l are shared, while the branches retain separate anchor offsets, token-guided gates, and residual scales. Its anchors can be written as

$$A_b^l = A_0^l + \Delta A_b^l, \tag{14}$$

where A_0^l is shared and ΔA_b^l is branch-specific. BD-SC is therefore semi-decoupled: the two branches can attend to different spatial evidence, but they remain in a shared projected and normalized feature space. All other host components, including the branch-specific patch transformations, training data, optimization schedule, and random seeds, are matched between BD-SC and BD.

Table 11: Pixel level results and consistency checks for the final BD-Host. Values are mean ± population standard deviation over seeds 111, 222, and 333 with input resolution 518, metric map size 256, and test time augmentation. Panel (a) reports medical localization. Each entry in panel (b) is shared by BD-Host and BD-Host with ProtoDeCo because calibration leaves the industrial pixel maps unchanged.

(a) Medical localization

Dataset	Pixel AUROC	F1 max	AP
Brain-AD	94.70 ± 0.50	44.80 ± 1.43	40.40 ± 0.94
CVC-ClinicDB	85.23 ± 1.05	46.50 ± 2.22	38.20 ± 2.43
EndoTect	83.40 ± 0.92	50.00 ± 1.56	44.70 ± 2.06
Kvasir	77.63 ± 1.33	45.13 ± 1.78	39.43 ± 2.18
Mean	85.24	46.61	40.68

(b) Industrial map consistency

Dataset	Pixel AUROC	F1 max	AP
MVTec	90.63 ± 0.12	44.30 ± 0.37	41.13 ± 0.66
VisA	95.64 ± 0.10	34.47 ± 0.36	27.22 ± 0.49
BTAD	92.90 ± 0.65	48.97 ± 1.68	44.07 ± 2.12
KSDD2	99.23 ± 0.09	59.70 ± 2.43	58.00 ± 3.48
DAGM2007	92.87 ± 0.21	55.10 ± 2.33	52.63 ± 3.07
DTD Synthetic	97.63 ± 0.12	65.50 ± 0.78	65.20 ± 1.27

Branch-specific patch transformation. For each patch p_u^l, BD applies a separate gated residual adapter to the anomaly and normal branches. The transformation for branch b is

$$q_{b,u}^l = W_{\mathrm{down},b}^l \mathrm{LN}_b^l(p_u^l), \tag{15}$$

$$c_{b,u}^l = \sigma\left(W_{2,b}^l \mathrm{SiLU}\left(W_{1,b}^l q_{b,u}^l\right)\right), \tag{16}$$

$$p_{b,u}^l = p_u^l + \eta_b^l \mathrm{Drop}\left[W_{\mathrm{up},b}^l\left(\mathrm{GELU}(q_{b,u}^l) \odot c_{b,u}^l\right)\right]. \tag{17}$$

The up-projection is initialized to zero and the residual scale is initialized to a small value, keeping both transformed branches close to the frozen patch representation at initialization. The same branch-specific patch transformation is used in BD-SC so that the host comparison isolates the SCA sharing pattern.

Host score and source objective. With $\hat{v} = v/\max(\|v\|_2, \varepsilon)$, the branch-contrastive patch logit is

$$r^l_{i,u} = \tau \left[\left\langle \hat{p}^l_{a,i,u}, \hat{z}^l_{a,i} \right\rangle - \left\langle \hat{p}^l_{n,i,u}, \hat{z}^l_{n,i} \right\rangle \right]. \tag{18}$$

The layer-wise logits are reshaped into spatial maps, resized to the input resolution, and fused to form the host pixel map $M_\theta(x_i)$. The host image score averages the largest βHW values in this map, with $\beta = 0.01$ for industrial targets and $\beta = 0.05$ for medical image-level targets. These settings are fixed across the three seeds and are used identically before and after ProtoDeCo calibration. The source-training objective is

$$\mathscr{L}_{\text{src}} = \lambda_{\text{seg}} (\mathscr{L}_{\text{focal}} + \mathscr{L}_{\text{dice}}) + \lambda_{\text{img}} \mathscr{L}_{\text{bce}} + \lambda_{\text{rel}} \mathscr{L}_{\text{token}}. \tag{19}$$

No target-domain sample participates in source optimization. The final-block V–V attention setting is held fixed across the host comparisons and is treated as part of the supporting host rather than as a separate contribution. ProtoDeCo is applied only after source training and changes the scalar image-level score; it does not update the host parameters or modify the host pixel map.

D Calibration Audit and Method Positioning

This section audits the representation used for target-side normality measurement and separates the effects of feature space from the effects of prototype compression. The main paper reports the final calibration results at $w = 0.1$, while the fixed-coordinate t-SNE visualization and the stronger mechanism diagnostics use $w = 0.3$ to make score changes easier to observe. The deployment setting and diagnostic setting therefore serve different purposes: $w = 0.1$ is used for the final quantitative comparison, whereas $w = 0.3$ is used only to expose the behavior of controlled variants. Table 21 reports the complete paired confidence intervals for the industrial host comparison and the medical BD-Host calibration results.

Table 12: AdaCLIP feature-space diagnostics. PN denotes pseudo-normal. Normal dist. is the within-normal feature distance, and PN score std. is the standard deviation of pseudo-normal scores; these metrics summarize feature dispersion and score variability. PN AUROC is a secondary metric. centered kernel alignment (CKA) compares frozen and projection patch representations over the same pseudo-normal samples.

Dataset	Feature space	Normal dist.	PN score std.	PN AUROC	CKA
Brain AD	Frozen	0.013700	0.1078	0.6699	
	Detector	0.002200	0.0207	0.6857	0.392
	Projection	0.000046	0.0062	0.6535	
BrainMRI	Frozen	0.015600	0.1183	0.8694	
	Detector	0.002500	0.0604	0.5688	0.128
	Projection	0.000100	0.0164	0.5906	
BTAD	Frozen	0.018300	0.1199	0.8640	
	Detector	0.001600	0.0087	0.8471	0.439
	Projection	0.000026	0.0001	0.7624	
MVTec	Frozen	0.029600	0.1049	0.7403	
	Detector	0.005300	0.0344	0.7772	0.150
	Projection	0.000160	0.0225	0.7289	

The frozen representation has larger within-normal feature dispersion and pseudo-normal score variability than the detector and projection representations on all four diagnostic datasets. This difference provides more dynamic range for distance-based target calibration, but the dispersion statistics and

CKA values are interpretive diagnostics rather than causal evidence. They support the judge–ruler mismatch hypothesis without implying that the detector representation is semantically corrupted.

At the final setting $w = 0.1$, every industrial Base-to-Frozen point estimate is positive, and every corresponding paired confidence interval lies above zero. Across the eight distinct-stream pairs, Frozen Proto improves over Detector Proto by a mean of 0.23 AUROC points, with a 95% confidence interval from 0.14 to 0.36 points. Three of the eight individual intervals exclude zero. The four WinCLIP pairs are identity controls because both arms use the same score cache. On the four medical image-level targets, ProtoDeCo improves 10 of 12 seed–target pairs and increases mean AUROC by 0.64 points. The gain is positive on OCT2017, BrainMRI, and Histopathology, whereas Brain-AD changes by -0.06 points.

Prototype compression improves both image metrics in every evaluated host–target cell at the deployment weight. Because this audit uses one fixed stochastic seed for prototype construction, it establishes consistency of the point estimates across hosts and targets but is not a cross-seed significance claim.

D.1 Candidate and Query Controls

These controls examine two transductive concerns: whether the host ranking provides useful pseudo-normal candidates and whether including a query in its own reference memory explains the calibration gain. The candidate-source experiment changes only the source of the pseudo-normal candidates while keeping frozen-space scoring fixed. The redraw experiment measures the stability of random candidate selection. The self-inclusion audit compares the deployed memory construction with cross-fitting and leave-one-out scoring. KSDD2 is reported separately as a near-ceiling stress case.

Candidate-source control. We hold the frozen feature space, prototype construction, query reduction, retrieval rule, and fusion weight fixed, and change only how pseudo-normal candidates are selected. Low-score selection follows the detector ranking, high-score selection reverses it, and random selection uses one fixed random draw. Oracle selection uses labeled normal images and is included only as a label-informed reference that is unavailable at deployment. The experiment uses the stronger diagnostic setting $w = 0.3$; random-selection stability at the deployment setting is reported separately in Table 15.

At $w = 0.3$, low-score selection improves mean AUROC by 2.03 points, whereas random selection gives a mean gain of 1.17 points with variation across cases. High-score selection decreases mean AUROC by -1.82 points, while the label-informed oracle gives a gain of 4.09 points. At the deployment setting, random candidate selection produces positive gains in 78 of 80 redraws, with a mean gain of 0.51 points. The contrast between low- and high-score selection supports the direction of the host ranking, while the random control shows that the frozen metric space contributes even when candidate selection is not detector-guided.

Self-inclusion audit. A query may be selected into the pseudo-normal memory used to score that same query. We therefore compare the deployed self-memory protocol with two-fold cross-fitting and leave-one-out (LOO), while keeping the frozen metric space, candidate ratio, prototype count, query reduction, retrieval rule, and fusion weight fixed. Table 16 reports AnomalyCLIP, AdaCLIP, and the three-seed BD-Host results at the final setting $w = 0.1$. LOO removes every query from its own reference memory, yet retains 95%, 107%, and 107% of the AnomalyCLIP, AdaCLIP, and BD-Host gains, respectively. All three AdaCLIP protocols improve every evaluated target. For BD-Host, the self-memory, two-fold, and LOO protocols improve 14, 13, and 14 of the 18 seed–target pairs, respectively.

Additional KSDD2 ceiling-stress test. KSDD2 Tabernik et al. (2020) follows the same VisA-source direction as the primary diagnostic targets but was not part of the original four-target panel. We report it separately because its uncalibrated AUROC is close to saturation, making scriptsize calibration-induced ranking changes more visible. These results are not included in the primary 12-pair mean.

Table 13: Frozen Raw versus Frozen Proto under the final deployment protocol. Both arms use identical low-score candidates, frozen CLIP Radford et al. (2021) layer-24 features, $\rho = 0.1$, $\alpha = 0.1$, 1-NN, and $w = 0.1$; Frozen Proto uses $K_p = 128$. Values are category-macro image metrics (%). Δ denotes Proto minus Raw. All point estimates use seed 111 for stochastic prototype construction.

Host	Dataset	Raw AUROC	Proto AUROC	Δ AUROC	Raw AP	Proto AP	Δ AP
AnomalyCLIP	MVTec	91.93	92.17	0.24	96.56	96.65	0.09
	BTAD	89.93	90.64	0.71	91.89	92.44	0.55
	DTD Synthetic	94.50	94.64	0.14	97.71	97.78	0.07
	DAGM	98.03	98.09	0.07	92.70	92.99	0.30
AdaCLIP	MVTec	90.18	90.47	0.30	95.85	96.01	0.15
	BTAD	90.28	90.48	0.20	90.69	90.92	0.22
	DTD Synthetic	91.90	92.31	0.42	96.33	96.55	0.21
	DAGM	96.75	96.96	0.21	89.49	90.04	0.55
WinCLIP	MVTec	90.95	91.32	0.37	95.91	96.07	0.16
	BTAD	69.93	70.66	0.72	71.39	71.63	0.24
	DTD Synthetic	95.23	95.44	0.21	97.80	97.90	0.10
	DAGM	92.15	92.73	0.58	80.26	81.30	1.04
Mean over 12 pairs		90.98	91.33	**0.35**	91.38	91.69	**0.31**
Positive point estimates			12/12			12/12	

Table 16: Self-inclusion audit at $w = 0.1$. Entries are image AUROC gains over the paired uncalibrated host. BD-Host values are averaged over seeds 111, 222, and 333; the final row averages all 18 seed–target pairs.

Host	Dataset	Self	Two fold	LOO
AnomalyCLIP	MVTec	0.61	0.77	0.66
	BTAD	1.52	1.39	1.37
	DTD Synthetic	0.18	0.21	0.17
	DAGM	0.12	0.13	0.12
	Mean	0.61	0.63	0.58
AdaCLIP	MVTec	0.56	0.56	0.55
	BTAD	0.47	0.60	0.53
	DTD Synthetic	0.60	0.59	0.69
	DAGM	0.50	0.45	0.51
	Mean	0.53	0.55	0.57
BD-Host	MVTec	0.58	0.63	0.66
	BTAD	−0.01	−0.03	−0.02
	KSDD2	−0.11	−0.08	−0.11
	DAGM	0.17	0.17	0.17
	DTD Synthetic	0.20	0.22	0.19
	VisA	0.35	0.32	0.37
	Mean over 18 pairs	0.20	0.20	0.21

Table 17: BD-Host adapter parameterization check. Values are mean image and pixel AUROC over the eight available tasks. The final row reports the gain of BD-PA over the shared adapter.

Adapter configuration	Image AUROC	Pixel AUROC
Shared MLP	89.46	90.79
Shared Adapter	88.18	89.75
BD-PA	88.16	90.54
BD-PA gain over Shared Adapter	−0.01	0.79

Table 14: Candidate-source ablation with frozen CLIP layer 24, $\rho = 0.1$, $K_p = 128$, $\alpha = 0.1$, 1-NN, and $w = 0.3$. Values are image AUROC (%). Medical rows are exploratory diagnostics.

Host	Dataset	Base	Random	High	Low	Oracle
WinCLIP	BrainMRI	86.5	86.8	85.1	87.2	87.7
	Brain-AD	72.4	70.4	67.5	73.3	77.8
	BTAD	67.5	78.9	69.9	76.9	84.4
AnomalyCLIP	BrainMRI	96.1	94.4	93.0	96.3	96.3
	Brain-AD	69.8	67.3	66.9	69.3	71.6
	BTAD	89.1	92.5	89.1	92.8	93.3
AdaCLIP	BrainMRI	95.3	94.1	92.5	95.6	95.6
	BTAD	90.0	91.7	88.1	91.7	92.6
Mean gain over Base		N/A	1.17	−1.82	**2.03**	4.09

Table 15: Random candidate redraw control at $w = 0.1$. Each cell uses 10 redraws. Mean, standard deviation, and worst are image-AUROC gains over the host. The dagger denotes the mean standard deviation within a cell.

Host	Dataset	Mean	Std.	Worst	Negative
AnomalyCLIP	MVTec	0.45	0.12	0.27	0/10
	BTAD	1.14	0.24	0.75	0/10
	DTD Synthetic	0.41	0.14	0.18	0/10
	DAGM	0.14	0.02	0.12	0/10
AdaCLIP	MVTec	0.58	0.10	0.40	0/10
	BTAD	0.36	0.29	−0.16	2/10
	DTD Synthetic	0.45	0.10	0.33	0/10
	DAGM	0.56	0.07	0.44	0/10
All 80 redraws		0.51	0.14†	−0.16	2/80

Table 18: KSDD2 control at $w = 0.1$. Entries are image AUROC gains over each host. WinCLIP uses one shared run for both identical streams.

Host	Detector Raw	Detector Proto	Frozen Raw	Frozen Proto	Frozen minus Detector
AnomalyCLIP	0.24	0.29	0.25	0.45	0.16
AdaCLIP	0.02	−0.04	0.01	0.05	0.09
WinCLIP	0.78	0.81	0.78	0.81	0.00

The two distinct stream hosts favor Frozen Proto by $+0.16$ and $+0.09$ points. The WinCLIP row is an identity check, not evidence for either space. The WinCLIP row reuses one score cache for both identical streams, so the Frozen Proto minus Detector Proto contrast is exactly zero.

D.2 Supporting Host and Parameter Checks

These checks concern the supporting BD-Host and fixed ProtoDeCo settings rather than the central cross-host claim. They are placed after the external-host controls so that host-specific evidence does not precede the host-agnostic result.

Table 19: End-to-end score generation efficiency for BD-Host on 1,725 MVTec images. Values are mean ± population standard deviation over five paired trials; peak denotes allocated GPU memory. Here, BD (w/o) denotes BD-Host without ProtoDeCo, BD (w/) denotes BD-Host with ProtoDeCo, and overhead is the difference between the two variants.

Mode	ms/img	Total (s)	Peak (MiB)
BD (w/o)	233.59 ± 0.16	402.95 ± 0.27	2340.6
BD (w/)	354.72 ± 0.18	611.90 ± 0.30	2340.6
Overhead	121.13 ± 0.26	208.95 ± 0.45	0.0

Within BD-Host, replacing the shared adapter with the branch-specific adapter changes image AUROC by only -0.01 points but improves pixel AUROC by 0.79 points. The branch-specific adapter is therefore retained mainly for localization. The parameter sweeps are post hoc robustness checks and are not used for target-specific selection.

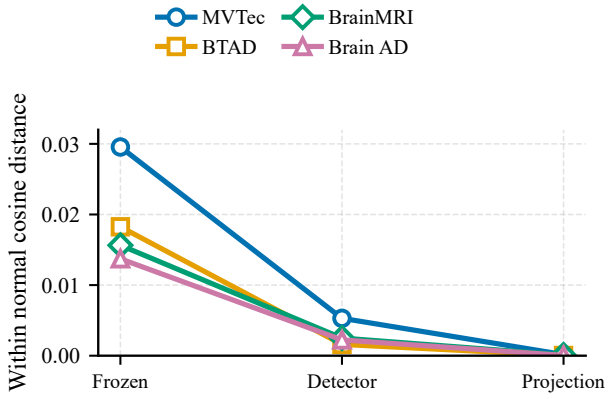
E Backbone Portability with DINOv2

Table 7 reports the DINOv2 Oquab et al. (2023) portability diagnostic. We instantiate both BD-Host and its task-token-free metric path with DINOv2 ViT-L/14. The metric path uses the public pretrained output normalization rather than the source-trained detector normalization. Across seeds 111, 222, and 333, 15 of 18 seed–target pairs improve at $w = 0.1$, producing positive mean gains on five of six industrial targets. KSDD2 changes by -0.07 points.

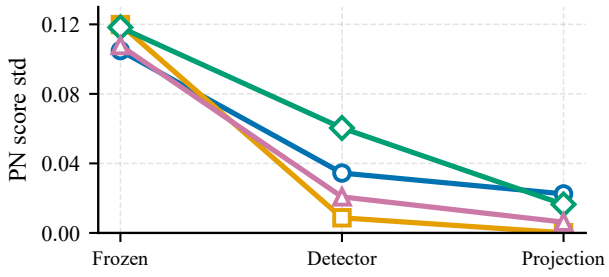
The 0.72-point macro improvement indicates that the decision–metric separation is not tied to CLIP alone. However, the KSDD2 regression rules out a backbone-independent non-regression claim. We therefore treat DINOv2 as portability evidence while retaining frozen CLIP as the evaluated default configuration.

F Target Batch Stress Test

ProtoDeCo is transductive over the available target batch. Using the seed-222 BD-Host checkpoints, we therefore vary one factor at a time on six industrial targets using mixed-category and single-category batches. The batch-size sweep uses 8, 16, 32, 64, and 128 images. The anomaly-ratio sweep uses 5, 10, 25, 50, and 75%. The pseudo-normal ratio sweep uses 5, 10, 20, and 30%. Repeated default conditions occur across sweeps, so the pooled rows below are descriptive summaries rather than counts of independent configurations.



(a) Absolute distance scale



(b) Calibration signal dynamic range

Figure 5: Feature-space diagnostics for AdaCLIP on MVTec, BTAD, BrainMRI, and Brain AD. The frozen-to-detector and projection transformations reduce within-normal cosine distance and pseudo-normal score variability. These diagnostic statistics are not used to form the main paper's reported target averages and are descriptive rather than causal because normalization within each category removes positive affine rescaling.

Table 20: One-factor ProtoDeCo sensitivity on MVTec and BTAD using the fixed seed-111 checkpoint. Entries are image AUROC values (%). Bold entries indicate the fixed default setting.

(a) MVTec

Parameter	Values	AUROC (%)
ρ	.05/.10/.20/.30	92.3/**92.5**/92.6/92.4
K_p	32/64/128/256	92.4/92.4/**92.5**/92.4
Layer	12/18/24	92.4/92.7/**92.5**
w	.05/.10/.20/.30	92.2/**92.5**/93.0/93.3

(b) BTAD

Parameter	Values	AUROC (%)
ρ	.05/.10/.20/.30	95.7/**95.7**/95.7/95.7
K_p	32/64/128/256	95.7/95.7/**95.7**/95.8
Layer	12/18/24	95.9/95.9/**95.7**
w	.05/.10/.20/.30	95.8/**95.7**/95.5/95.3

Table 21: Final paired calibration results at $w = 0.1$. In panel (a), $\Delta_{\mathrm{B,F}}$ denotes the image-AUROC gain of Frozen Proto relative to Base, and $\Delta_{\mathrm{D,F}}$ denotes the gain relative to Detector Proto. In panel (b), $\Delta_{\mathrm{B,P}}$ denotes the gain of ProtoDeCo relative to Base. Here, B, D, F, and P denote Base, Detector Proto, Frozen Proto, and ProtoDeCo, respectively. Parenthesized values are 95% paired bootstrap confidence intervals.

(a) Industrial external host comparison

Host	Dataset	$\Delta_{\mathrm{B,F}}$	$\Delta_{\mathrm{D,F}}$
AnomalyCLIP	MVTec	0.62 (0.35, 0.91)	0.05 (−0.11, 0.22)
	BTAD	1.52 (0.80, 2.39)	0.19 (−0.34, 0.69)
	DTD Synthetic	0.18 (0.01, 0.37)	0.19 (0.05, 0.35)
	DAGM	0.12 (0.07, 0.18)	0.07 (0.04, 0.11)
AdaCLIP	MVTec	0.56 (0.28, 0.86)	0.22 (−0.09, 0.51)
	BTAD	0.47 (0.12, 0.88)	0.59 (−0.01, 1.29)
	DTD Synthetic	0.60 (0.35, 0.87)	0.24 (−0.05, 0.54)
	DAGM	0.50 (0.39, 0.61)	0.32 (0.21, 0.44)
WinCLIP	MVTec	1.00 (0.69, 1.34)	0.00 (0.00, 0.00)
	BTAD	3.53 (2.15, 4.89)	0.00 (0.00, 0.00)
	DTD Synthetic	0.40 (0.14, 0.69)	0.00 (0.00, 0.00)
	DAGM	1.05 (0.82, 1.29)	0.00 (0.00, 0.00)
Mean over 12 pairs		0.88 (0.48, 1.44)	0.16 (0.07, 0.26)

(b) Medical BD-Host ablation

Host	Dataset	$\Delta_{\mathrm{B,P}}$	Positive seeds
BD-Host	OCT2017	0.87 ± 0.23	3/3
	BrainMRI	0.50 ± 0.22	3/3
	Brain-AD	-0.06 ± 0.25	1/3
	Histopathology	1.24 ± 0.15	3/3
Mean over 12 seed–target pairs		$\mathbf{0.64 \pm 0.53}$	10/12

The pooled mean remains modestly positive in most tested regimes, but individual resamples can degrade sharply. Batch size 8 gives a negative mean, and heavy contamination increases the frequency of negative outcomes. These results define an operating limitation rather than a robustness claim. The method should use the largest available target batch, while very small or anomaly-dominated batches require caution.

Table 22: Stronger weight stress diagnostic across four hosts. Base is the host image AUROC and the remaining columns are gains at $w = 0.3$, used to magnify the calibration signal rather than as the final deployment protocol. Raw and Proto denote uncompressed and $K_p = 128$ memories in detector or frozen CLIP space.

Host	Dataset	Base	Detector Raw	Detector Proto	Frozen Raw	Frozen Proto
AdaCLIP	MVTec	89.93	0.62	0.57	0.83	1.20
	BTAD	90.02	−0.45	−1.74	1.14	1.66
	DTD Synthetic	91.71	0.72	0.53	0.52	1.48
	DAGM	96.54	0.54	0.16	0.69	1.16
AnomalyCLIP	MVTec	91.55	1.00	0.90	0.96	1.15
	BTAD	89.12	3.06	3.34	2.64	3.66
	DTD Synthetic	94.46	−0.23	−0.16	0.04	0.47
	DAGM	97.97	0.11	0.07	0.16	0.20
WinCLIP	MVTec	90.31	1.49	1.71	1.53	1.78
	BTAD	67.46	7.08	9.97	7.26	9.43
	DTD Synthetic	94.98	0.38	0.47	0.39	0.72
	DAGM	91.68	1.23	2.11	1.22	2.17
VisualAD Hou et al. (2026)	MVTec	92.23	0.87	1.09	0.84	0.96
	BTAD	94.93	0.59	0.47	0.47	0.41
	DTD Synthetic	97.46	0.46	0.70	0.41	0.59
	DAGM	99.54	0.17	0.13	0.16	0.14
Mean gain over 16 pairs		N/A	1.10	1.27	1.20	**1.70**
Negative cases		N/A	2	2	0	0
Worst gain		N/A	−0.45	−1.74	0.04	**0.14**

G Additional Diagnostics and Method Positioning

This section collects secondary analyses that contextualize, but do not define, the final deployment result: end-to-end cost, the stronger $w = 0.3$ mechanism stress test, and comparison with MuSc as a standalone transductive detector.

Runtime details. Table 19 reports five paired end-to-end score-generation trials on 1,725 MVTec images using an RTX 3080, batch size one, test-time augmentation, and five warm-up images. Timing includes data loading and every score-generation stage, but excludes model loading and offline metric computation. Both variants use 2340.6 MiB peak allocated GPU memory; reserved memory is 2478/2486 MiB for BD-Host and BD-Host with ProtoDeCo, respectively. In an isolated memory benchmark, prototype compression reduces lookup from 3.137 to 1.320 ms/image and total lookup time from 5.41 to 2.28 s.

Interpreting the stress diagnostic. The larger fusion weight separates the effects of metric-space choice and memory compression. Frozen Raw removes the two detector-space regressions, while Frozen Proto raises the mean gain from 1.20 to 1.70 points over the 16 cells. These magnitudes characterize a stress setting rather than deployment performance; Table 13 provides the corresponding compression comparison at the final $w = 0.1$.

Relation to MuSc. MuSc Li et al. (2024) is the closest standalone transductive detector and is evaluated separately because it replaces, rather than calibrates, the host. It globally mutual-scores the target set and generates new image and pixel anomaly outputs without using a host detector score. In contrast, ProtoDeCo uses the host ranking to select a low-score subset, measures deviation in a separate frozen space, calibrates only the scalar image score, and preserves the host pixel map. Table 23 therefore reports performance context and an attached-calibrator negative control rather than treating MuSc as another host baseline.

Table 23: MuSc performance context and negative control. Panel (a) reports image AUROC/AP (%) on the two targets in MuSc's published native protocol; this partial coverage does not form a cross-target mean. Panel (b) reports reproduced MuSc scores and calibration gains at $w = 0.1$.

(a) Published-protocol overlap

Method	MVTec	VisA
MuSc (published)	97.80/99.10	92.80/93.50
MuSc (reproduced)	97.77/99.07	92.42/93.17
BD-Host with ProtoDeCo	92.65/96.74	86.04/88.52

(b) Attached-calibrator negative control

Dataset	Base	Detector Raw Δ	Detector Proto Δ	Frozen Raw Δ	Frozen Proto Δ
MVTec	97.77/99.07	$0.02/-0.01$	$0.04/-0.01$	$-0.02/-0.02$	$-0.05/-0.05$
BTAD	95.89/97.75	$0.03/-0.07$	$0.02/-0.38$	$0.05/0.17$	$-0.05/-0.09$
VisA	92.42/93.17	$0.01/0.02$	$0.03/0.06$	$-0.03/-0.02$	$-0.01/0.00$
KSDD2	95.01/96.13	$-0.03/-0.02$	$-0.04/-0.02$	$-0.07/-0.06$	$-0.02/-0.01$
DTD Synthetic	96.35/98.64	$0.00/-0.02$	$-0.03/-0.04$	$0.00/-0.01$	$0.16/0.05$
Five-target mean	95.49/96.95	$0.01/-0.02$	$0.00/-0.08$	$-0.01/0.01$	$0.01/-0.02$

MuSc is substantially stronger on the two-target overlap, which is reported for transparency rather than folded into the complete six-target comparison. The reproduced MVTec and VisA bases are within 0.03 and 0.38 AUROC points of the published values. In panel (b), all four five-target mean AUROC changes are within 0.01 points of zero, consistent with redundancy between MuSc's target-set mutual normality and the attached pseudo-normal signal. The DAGM run returned no category-level scores and produced undefined aggregate metrics, so it is excluded and no six-target MuSc mean is reported.

H Feature Space Scale and Ranking Diagnostics

Figure 5 summarizes secondary feature-dispersion diagnostics across four targets. The controlled BTAD mechanism diagnostic is reported in main-paper Table 5, while Table 12 provides complementary AdaCLIP statistics, including within-normal distance, pseudo-normal score standard deviation, and the secondary pseudo-normal AUROC diagnostic.

Interpretation. The diagnostics show an association between detector-side feature contraction and unreliable distance-based calibration, not a sufficient causal explanation. AdaCLIP on BTAD provides the clearest controlled reversal: both detector-space memories regress while the matched frozen-space memories improve. Because category-wise normalization can preserve rankings despite large scale changes, dispersion and CKA are used to describe the geometry change rather than to predict gains by themselves.

I Pixel Level Results and Consistency

Tables 24 and 25 provide the complete cross-model, per-dataset comparison corresponding to the pixel-level results in the main paper, including S2SWCLIP Zhang et al. (2026) and FB-CLIP Hu et al. (2026). The medical localization targets are Brain-AD, CVC-ClinicDB Bernal et al. (2015), EndoTect Hicks et al. (2021), and Kvasir Jha et al. (2019). Table 11 reports Pixel F1 for the final BD-Host and verifies that calibration leaves its industrial pixel maps unchanged. Learned-method cells use completed runs from the fixed seed set $\{111, 222, 333\}$; deterministic methods are shown once. Because ProtoDeCo changes only the scalar image score, the BD-Host rows are not presented as a localization contribution of ProtoDeCo.

BD-Host ranks first in mean industrial Pixel AP (48.04), with the best results on MVTec AD and DTD Synthetic and second-best results on VisA and DAGM2007. Its industrial Pixel AUROC ranks

Table 24: Industrial pixel-level localization results (%). Learned methods report mean ± population standard deviation over completed runs. CLIP is the local deterministic baseline, and WinCLIP follows the deterministic results reported with the VisualAD benchmark. The Mean column is the arithmetic mean over the six displayed datasets, computed before rounding. Bold and underline mark the best and second-best results, respectively.

Metric	Method	MVTec AD	VisA	BTAD	KSDD2	DAGM2007	DTD Synthetic	Mean
Pixel AUROC	CLIP	44.50	51.90	30.70	41.90	24.00	26.90	36.65
	WinCLIP	82.30	73.20	72.70	94.10	87.60	79.50	81.57
	AnomalyCLIP	90.94 ± 0.10	95.40 ± 0.07	93.24 ± 0.01	99.04 ± 0.04	**96.86 ± 0.03**	97.52 ± 0.01	**95.50**
	AdaCLIP	89.19 ± 1.84	**95.96 ± 0.15**	95.36 ± 1.33	98.64 ± 0.69	93.93 ± 1.43	93.41 ± 4.89	94.42
	S2SWCLIP	**90.97 ± 0.12**	95.07 ± 0.09	92.03 ± 0.17	99.20 ± 0.00	96.83 ± 0.05	**98.20 ± 0.08**	95.38
	VisualAD	90.20 ± 0.09	95.25 ± 0.05	93.63 ± 0.37	99.06 ± 0.25	92.88 ± 0.60	97.45 ± 0.08	94.74
	FB-CLIP	88.91 ± 0.62	95.22 ± 0.57	**96.96 ± 0.27**	98.70 ± 0.32	93.12 ± 2.30	95.22 ± 2.06	94.68
	BD-Host	90.63 ± 0.12	95.64 ± 0.10	92.90 ± 0.65	**99.23 ± 0.09**	92.87 ± 0.21	97.63 ± 0.12	94.82
Pixel AP	CLIP	3.40	2.00	2.10	0.40	0.20	1.40	1.58
	WinCLIP	18.20	5.40	12.90	17.40	6.80	9.80	11.75
	AnomalyCLIP	34.35 ± 0.20	20.78 ± 0.27	40.97 ± 0.68	43.71 ± 0.39	**54.05 ± 0.31**	52.40 ± 0.15	41.04
	AdaCLIP	40.94 ± 4.47	**30.70 ± 1.39**	46.60 ± 5.41	62.54 ± 2.14	49.52 ± 4.81	51.23 ± 12.43	46.92
	S2SWCLIP	34.37 ± 0.42	19.60 ± 0.22	38.73 ± 0.37	43.03 ± 0.73	52.33 ± 0.33	51.63 ± 0.82	39.95
	VisualAD	39.21 ± 0.87	25.92 ± 0.28	**48.45 ± 1.02**	**63.41 ± 1.02**	49.05 ± 3.57	61.78 ± 0.56	47.97
	FB-CLIP	40.83 ± 0.66	25.88 ± 2.28	45.34 ± 2.05	59.96 ± 5.55	44.99 ± 8.06	51.94 ± 4.37	44.82
	BD-Host	**41.13 ± 0.66**	27.22 ± 0.49	44.07 ± 2.12	58.00 ± 3.48	52.63 ± 3.07	**65.20 ± 1.27**	**48.04**

third and attains the best KSDD2 result. On the four medical targets, BD-Host ranks fifth in both mean Pixel AUROC and Pixel AP. These results support competitive industrial localization by the supporting host, not a universal pixel-level advantage and not a localization gain from ProtoDeCo.

Figure 6 complements the aggregate industrial results with thresholded pixel-error decompositions. VisualAD is used as the controlled reference because BD-Host is derived from the same host formulation. A direct qualitative comparison with FB-CLIP would additionally conflate the proposed host changes with its distinct vision-language localization pipeline and text branch; its quantitative results remain included in Tables 24 and 25. The examples illustrate error types and do not imply uniform per-sample improvement.

Table 26, Table 27, and Table 28 provide the seed-level audit in the same format for image-level and pixel-level metrics. Values are read from the source-backed per-seed records and rounded directly to two decimal places for presentation; they are not reconstructed from one-decimal aggregate values. Ours denotes BD-Host with ProtoDeCo. Seed entries are reported only when the original run log was available; missing cells are intentionally left blank rather than reconstructed from aggregate summaries.

The seedwise audit exposes run-level variation rather than defining an additional aggregate ranking. Ours has the highest industrial image AUROC on VisA, BTAD, DAGM2007, and DTD-Synthetic in all three reported seeds, while the ordering on MVTec AD and KSDD2 varies across methods. The medical image and pixel rows are more heterogeneous, which is consistent with the aggregate results above and supports reporting seed means instead of selecting individual runs. Missing cells remain visible so that this audit does not imply measurements that were not recorded.

Panel (a) characterizes the localization ability of the supporting host rather than an effect of ProtoDeCo. The exact equality in panel (b) is expected from the method definition and serves as an implementation check: calibration changes the scalar image score but never reads, regenerates, or post-processes the pixel map. Image-level gains should therefore not be interpreted as localization gains.

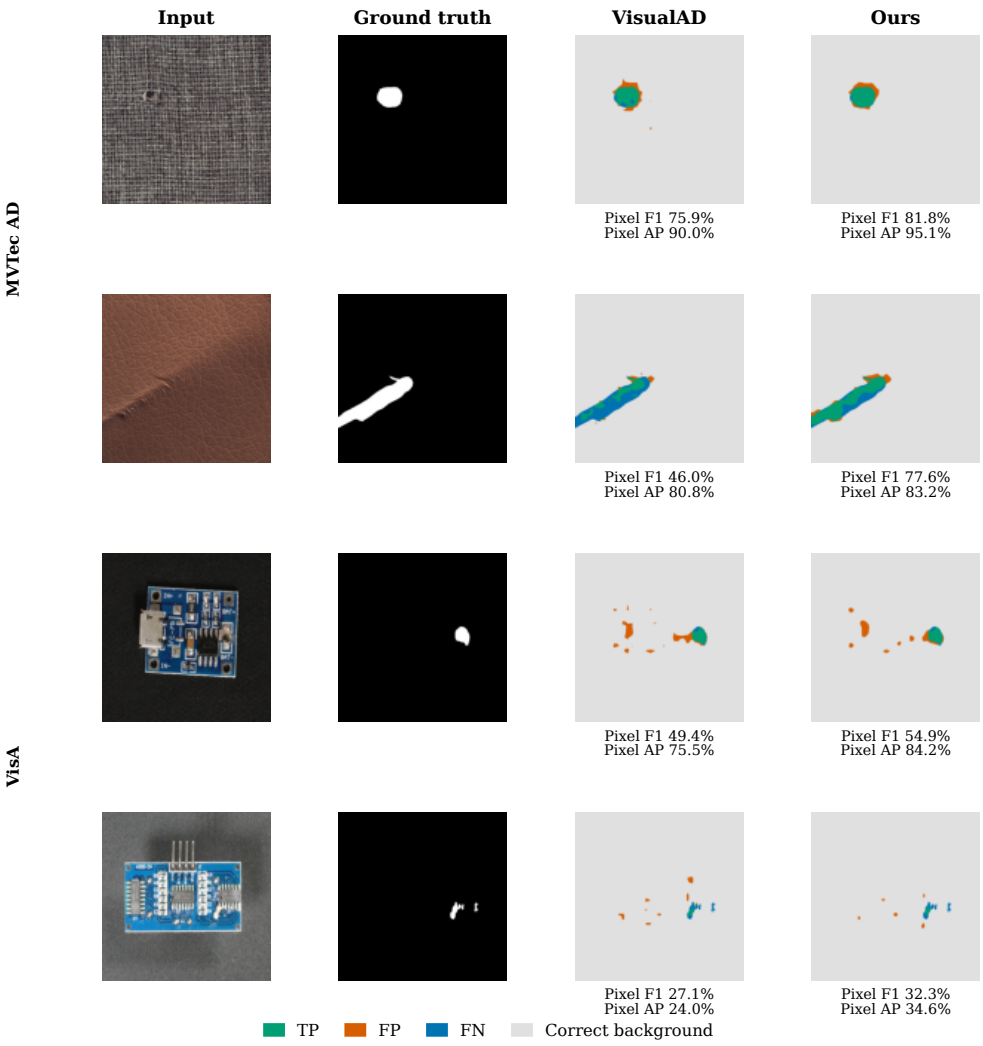


Figure 6: Industrial pixel-error decompositions at 518×518 resolution: (a) MVTec AD and VisA. Each dataset contributes two print-legible samples near the fixed 75th percentile of paired localization-improvement ranks, rather than maximum-gain cases. Green, orange, and blue denote true positives, false positives, and false negatives; gray denotes correct background. Per-image Pixel F1 is computed using each method's dataset-level max-F1 threshold; no image-specific threshold is selected. Pixel F1 and Pixel AP are in percent. Ours denotes BD-Host; ProtoDeCo leaves its pixel map unchanged.

Table 25: Medical pixel-level localization results (%) on the four targets with masks. Reporting and highlighting follow Table 24. † denotes that the S2SWCLIP Brain-AD pixel cell has two completed seeds; all other learned-method cells use three.

Metric	Method	Brain-AD	CVC-ClinicDB	EndoTect	Kvasir	Mean
Pixel AUROC	CLIP	42.20	59.20	64.00	60.30	56.43
	WinCLIP	87.60	70.30	68.20	69.70	73.95
	AnomalyCLIP	**96.07 ± 0.10**	85.06 ± 0.09	86.56 ± 0.22	82.13 ± 0.20	87.46
	AdaCLIP	92.30 ± 3.07	82.69 ± 4.19	81.19 ± 5.09	78.72 ± 4.53	83.73
	S2SWCLIP	95.75 ± 0.05†	83.77 ± 0.05	85.37 ± 0.12	81.40 ± 0.08	86.57
	VisualAD	93.67 ± 0.06	85.89 ± 0.09	85.74 ± 0.45	81.35 ± 0.25	86.65
	FB-CLIP	94.97 ± 0.11	**88.46 ± 0.39**	**89.37 ± 0.34**	**86.72 ± 0.39**	**89.89**
	BD-Host	94.70 ± 0.50	85.23 ± 1.05	83.40 ± 0.92	77.63 ± 1.33	85.24
Pixel AP	CLIP	2.50	13.40	22.10	22.20	15.05
	WinCLIP	13.30	19.40	23.80	27.80	21.08
	AnomalyCLIP	**56.18 ± 0.52**	41.95 ± 0.26	51.12 ± 0.60	43.69 ± 0.43	48.23
	AdaCLIP	30.34 ± 9.24	37.98 ± 6.80	42.92 ± 7.49	43.45 ± 7.08	38.67
	S2SWCLIP	52.45 ± 0.85†	39.37 ± 0.42	49.73 ± 0.33	42.43 ± 0.33	46.00
	VisualAD	35.60 ± 1.27	37.80 ± 0.57	47.00 ± 1.26	43.20 ± 0.78	40.78
	FB-CLIP	48.55 ± 1.51	**48.62 ± 1.78**	**58.59 ± 2.28**	**54.83 ± 1.62**	**52.64**
	BD-Host	40.40 ± 0.94	38.20 ± 2.43	44.70 ± 2.06	39.43 ± 2.18	40.68

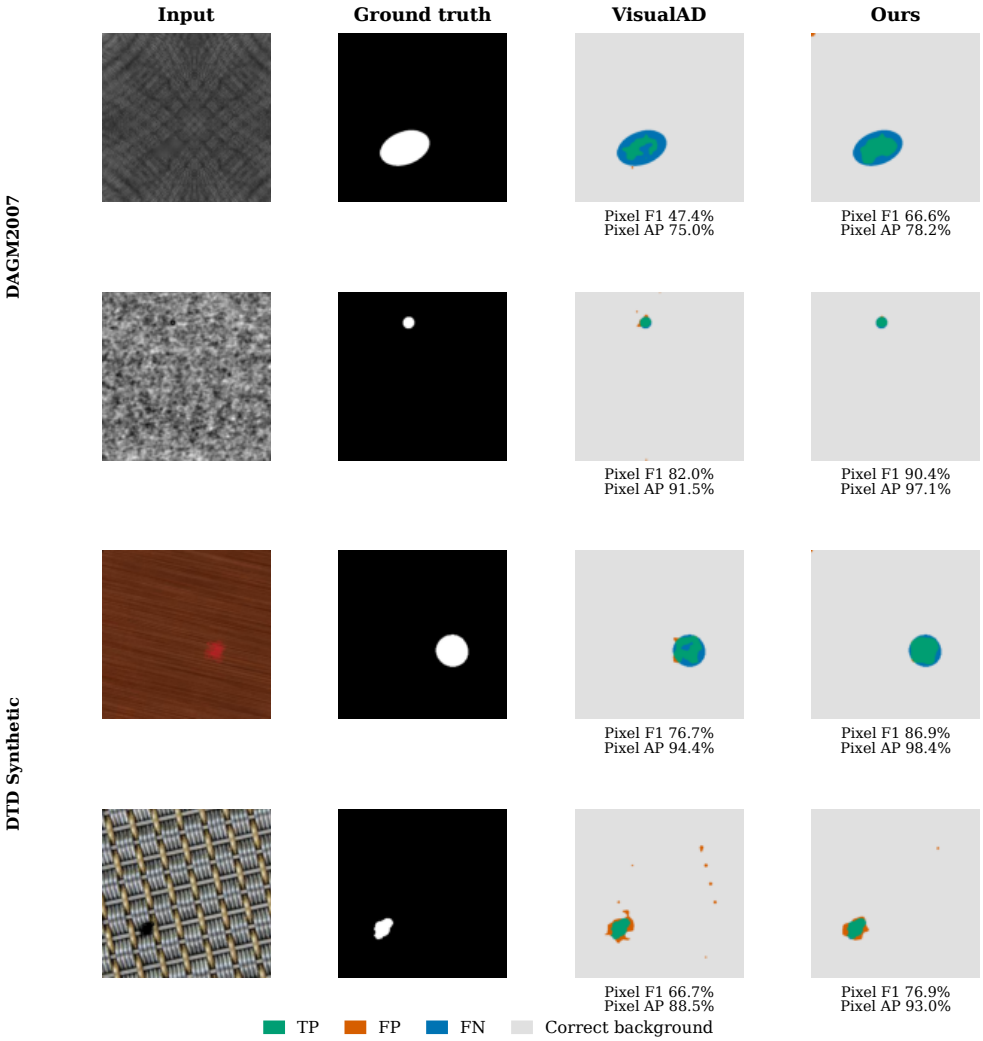


Figure 6: Industrial pixel-error decompositions (continued): (b) DAGM2007 and DTD Synthetic. Notation and sample selection follow panel (a).

Table 26: Per-seed image-level and pixel-level audit for seed 111. Each cell reports (AUROC, F1-max, AP) in percent. Bold and underline mark the best and second-best result for each metric, respectively. Pixel-AUPRO is omitted because the available AUPRO fields are placeholders rather than valid measurements. Ours denotes BD-Host with ProtoDeCo; missing Ours cells indicate that the original seed log was unavailable in the audit source, so no aggregate value is backfilled.

(a) WinCLIP, AnomalyCLIP, and AdaCLIP

Domain	Metric	Dataset	WinCLIP	AnomalyCLIP	AdaCLIP
Industrial	Image-level (AUROC, F1-max, AP)	MVTec AD	(70.23, 86.24, 84.63)	(91.55, 92.75, 96.36)	(**92.69**, 93.17, 96.71)
		VisA	(64.29, 74.22, 70.32)	(81.97, 80.44, 85.34)	(78.11, 79.79, 81.97)
		BTAD	(57.33, 64.86, 63.64)	(89.12, 85.99, 91.05)	(94.62, 91.28, 96.21)
		KSDD2	(78.67, 72.58, 80.09)	(92.67, 85.28, 94.04)	(95.20, 89.49, 96.16)
		DAGM2007	(59.11, 28.70, 21.59)	(97.97, 90.71, 92.48)	(98.38, 93.00, 94.60)
		DTD-Synthetic	(62.96, 85.63, 82.69)	(94.46, 94.53, 97.70)	(93.19, 93.58, 97.42)
	Pixel-level (AUROC, F1-max, AP)	MVTec AD	(80.69, 22.98, 16.53)	(**91.07**, 39.08, 34.51)	(90.49, **45.64**, **43.15**)
		VisA	(70.05, 9.51, 6.09)	(95.49, 28.18, 21.15)	(**96.09**, **36.49**, **30.30**)
		BTAD	(64.37, 12.26, 8.07)	(93.25, 47.68, 41.90)	(96.19, 51.72, **49.44**)
		KSDD2	(95.58, 27.09, 17.61)	(98.99, 50.13, 43.19)	(98.91, **60.79**, **62.38**)
		DAGM2007	(76.35, 6.97, 2.80)	(**96.90**, **57.64**, **54.30**)	(94.20, 55.34, 51.05)
		DTD-Synthetic	(77.88, 15.67, 9.41)	(97.54, 55.60, 52.21)	(96.74, 59.65, 58.05)
Medical	Image-level (AUROC, F1-max, AP)	OCT17	(63.64, 85.71, 84.29)	(63.51, 85.71, 86.41)	–
		BrainMRI	(68.56, 78.42, 77.87)	(**96.10**, **94.04**, **97.22**)	–
		Brain_AD	(47.33, 90.60, 80.80)	(69.83, 90.70, 90.24)	(**80.29**, **92.05**, 94.49)
		Histopathology_AD	(**61.58**, 68.26, **59.88**)	(58.60, 66.71, 58.50)	–
	Pixel-level (AUROC, F1-max, AP)	Brain_AD	(86.19, 19.42, 11.40)	(**96.19**, **53.77**, **56.81**)	(94.43, 40.60, 37.52)
		CVC-ClinicDB	(70.05, 25.90, 18.02)	(85.06, 45.62, 41.85)	(84.75, 43.86, 41.62)
		EndoTect	(61.79, 29.39, 18.50)	(86.29, 53.50, 50.49)	–
		Kvasir	(66.42, 33.49, 24.24)	(81.87, 49.90, 43.32)	–

(b) VisualAD, FB-CLIP, and Ours

Domain	Metric	Dataset	VisualAD	FB-CLIP	Ours
Industrial	Image-level (AUROC, F1-max, AP)	MVTec AD	(92.24, 92.45, **96.74**)	(89.83, 91.13, 95.59)	(92.46, **93.23**, 96.48)
		VisA	(81.85, 80.17, 85.70)	(82.68, 81.10, 85.64)	(**86.89**, **83.59**, **89.18**)
		BTAD	(93.23, 92.43, 96.97)	(90.95, 88.52, 92.09)	(**95.72**, **96.00**, **98.27**)
		KSDD2	(95.34, 89.99, 96.14)	(97.87, 92.63, 98.12)	(**98.28**, **93.60**, **98.56**)
		DAGM2007	(98.50, 92.88, 95.40)	(97.43, 89.89, 93.60)	(**99.47**, **95.10**, **97.66**)
		DTD-Synthetic	(94.86, 93.35, 97.54)	(94.47, 94.40, 97.99)	(**98.23**, **97.18**, **99.38**)
	Pixel-level (AUROC, F1-max, AP)	MVTec AD	(90.09, 41.46, 37.99)	(89.77, 42.42, 40.85)	(90.62, 44.25, 41.16)
		VisA	(95.21, 32.16, 25.52)	(95.98, 33.32, 27.73)	(95.61, 34.65, 27.27)
		BTAD	(93.43, **52.88**, 47.69)	(**96.66**, 48.25, 44.80)	(93.45, 48.15, 43.49)
		KSDD2	(98.71, 60.24, 62.13)	(99.09, 53.29, 56.22)	(**99.27**, 60.23, 58.76)
		DAGM2007	(92.16, 53.04, 50.14)	(95.25, 53.70, 51.37)	(92.87, 56.02, 53.90)
		DTD-Synthetic	(97.42, 61.17, 61.42)	(**97.79**, 55.29, 54.92)	(97.62, **66.13**, **66.15**)
Medical	Image-level (AUROC, F1-max, AP)	OCT17	(90.79, **90.10**, 96.89)	(78.11, 86.69, 90.69)	(**90.90**, 89.50, **97.20**)
		BrainMRI	(95.42, 92.95, 94.29)	(89.35, 88.40, 91.35)	(95.30, 92.30, 96.10)
		Brain_AD	(75.98, 91.13, 92.61)	(71.79, 91.77, 91.02)	(79.60, 91.20, **94.50**)
		Histopathology_AD	(53.46, 67.83, 50.77)	(57.17, 67.61, 54.86)	(60.30, **69.20**, 56.30)
	Pixel-level (AUROC, F1-max, AP)	Brain_AD	(93.74, 39.03, 33.80)	(95.12, 45.28, 47.44)	(95.00, 45.70, 41.20)
		CVC-ClinicDB	(86.00, 46.19, 37.50)	(**87.91**, **49.40**, **47.04**)	(85.70, 46.90, 37.70)
		EndoTect	(85.15, 51.40, 45.30)	(**89.80**, **59.84**, **60.23**)	(83.40, 49.70, 44.30)
		Kvasir	(81.01, 48.85, 42.20)	(**87.28**, **57.66**, **56.06**)	(78.00, 45.30, 39.50)

Table 27: Per-seed image-level and pixel-level audit for seed 222. Each cell reports (AUROC, F1-max, AP) in percent. Bold and underline mark the best and second-best result for each metric, respectively. Pixel-AUPRO is omitted because the available AUPRO fields are placeholders rather than valid measurements. Ours denotes BD-Host with ProtoDeCo; missing Ours cells indicate that the original seed log was unavailable in the audit source, so no aggregate value is backfilled.

(a) WinCLIP, AnomalyCLIP, and AdaCLIP

Domain	Metric	Dataset	WinCLIP	AnomalyCLIP	AdaCLIP
Industrial	Image-level (AUROC, F1-max, AP)	MVTec AD	(70.23, 86.24, 84.63)	(91.69, 92.55, 96.26)	(87.84, 91.24, 93.87)
		VisA	(64.29, 74.22, 70.32)	(81.06, 80.16, 84.52)	(85.27, **82.92**, **87.97**)
		BTAD	(57.33, 64.86, 63.64)	(88.99, 87.27, 91.78)	(86.23, 81.21, 82.85)
		KSDD2	(78.67, 72.58, 80.09)	(91.89, 84.51, 93.49)	(93.85, 88.27, 94.96)
		DAGM2007	(59.11, 28.70, 21.59)	(98.08, 90.93, 92.44)	(98.03, 88.15, 92.40)
		DTD-Synthetic	(62.96, 85.63, 82.69)	(93.78, 94.18, 97.38)	(88.19, 90.18, 94.52)
	Pixel-level (AUROC, F1-max, AP)	MVTec AD	(80.69, 22.98, 16.53)	(**90.85**, 38.38, 34.07)	(86.59, 38.74, 34.70)
		VisA	(70.05, 9.51, 6.09)	(95.39, 27.77, 20.61)	(**96.04**, **38.05**, **32.56**)
		BTAD	(64.37, 12.26, 8.07)	(93.25, 46.74, 40.73)	(93.48, 42.97, 39.03)
		KSDD2	(95.58, 27.09, 17.61)	(99.07, 50.65, 43.81)	(97.70, 59.91, 60.00)
		DAGM2007	(76.35, 6.97, 2.80)	(**96.85**, **57.71**, 54.24)	(92.06, 48.73, 43.01)
		DTD-Synthetic	(77.88, 15.67, 9.41)	(97.51, 55.52, 52.43)	(86.50, 38.63, 33.79)
Medical	Image-level (AUROC, F1-max, AP)	OCT17	(63.64, 85.71, 84.29)	(71.74, 85.71, 89.71)	(57.56, 85.71, 81.75)
		BrainMRI	(68.56, 78.42, 77.87)	(**96.45**, 94.08, **97.41**)	(90.47, 87.37, 94.77)
		Brain_AD	(47.33, 90.60, 80.80)	(70.37, 90.64, 90.66)	(76.89, 90.78, 93.58)
		Histopathology_AD	(61.58, 68.26, **59.88**)	(58.50, 66.76, 58.62)	(**63.28**, **70.21**, 57.16)
	Pixel-level (AUROC, F1-max, AP)	Brain_AD	(86.19, 19.42, 11.40)	(**96.06**, **53.14**, **56.19**)	(87.95, 25.11, 17.29)
		CVC-ClinicDB	(70.05, 25.90, 18.02)	(84.95, 45.41, 41.69)	(76.85, 33.34, 28.46)
		EndoTect	(61.79, 29.39, 18.50)	(86.59, 53.87, 50.92)	(74.01, 37.70, 32.33)
		Kvasir	(66.42, 33.49, 24.24)	(82.16, 50.27, 43.45)	(72.31, 38.90, 33.45)

(b) VisualAD, FB-CLIP, and Ours

Domain	Metric	Dataset	VisualAD	FB-CLIP	Ours
Industrial	Image-level (AUROC, F1-max, AP)	MVTec AD	(92.02, 92.31, 96.60)	(90.54, 91.93, 95.87)	(**92.75**, **93.07**, **96.88**)
		VisA	(81.96, 79.99, 85.87)	(84.30, 82.28, 86.26)	(**85.29**, 82.51, 87.70)
		BTAD	(94.74, **95.53**, **98.27**)	(92.71, 90.68, 95.15)	(**96.24**, 95.34, 97.75)
		KSDD2	(96.38, 90.69, 96.87)	(**97.87**, 92.54, **98.20**)	(97.52, **92.70**, 98.01)
		DAGM2007	(98.23, 91.17, 93.80)	(98.40, 92.63, 95.37)	(**99.46**, **95.48**, **97.80**)
		DTD-Synthetic	(95.74, 94.15, 98.26)	(93.76, 93.67, 97.84)	(**97.24**, **96.57**, **99.18**)
	Pixel-level (AUROC, F1-max, AP)	MVTec AD	(90.21, 42.50, 39.69)	(88.39, 42.09, 40.01)	(90.78, **44.84**, **41.85**)
		VisA	(95.23, 32.64, 26.08)	(95.05, 31.51, 27.23)	(95.78, 34.80, 27.79)
		BTAD	(94.15, **54.18**, **49.90**)	(**96.91**, 46.86, 43.13)	(93.19, 51.27, 46.88)
		KSDD2	(**99.29**, 61.06, **63.46**)	(98.30, 54.23, 55.86)	(99.27, **62.35**, 61.76)
		DAGM2007	(93.63, 54.88, 52.77)	(89.92, 40.73, 33.62)	(93.06, 57.43, **55.60**)
		DTD-Synthetic	(**97.57**, 60.90, 61.35)	(92.74, 50.24, 45.75)	(97.49, **65.99**, **66.09**)
Medical	Image-level (AUROC, F1-max, AP)	OCT17	(91.11, 90.25, 96.99)	(83.55, 86.56, 93.54)	(**92.00**, **90.70**, **97.40**)
		BrainMRI	(94.48, 91.64, 93.72)	(91.97, 90.28, 92.79)	(95.70, **94.60**, 96.30)
		Brain_AD	(79.42, 91.37, 94.19)	(74.67, **91.77**, 92.55)	(**80.10**, 91.40, **94.20**)
		Histopathology_AD	(55.21, 67.87, 52.43)	(60.60, 69.21, 54.96)	(58.70, 68.50, 55.10)
	Pixel-level (AUROC, F1-max, AP)	Brain_AD	(93.59, 40.39, 36.50)	(94.87, 48.49, 50.68)	(94.90, 45.10, 40.50)
		CVC-ClinicDB	(85.89, 46.22, 38.60)	(**88.68**, **50.94**, **47.73**)	(85.60, 47.50, 39.30)
		EndoTect	(85.84, 52.80, 47.40)	(**88.96**, **56.77**, **55.37**)	(83.00, 49.70, 44.40)
		Kvasir	(81.41, 49.25, 43.30)	(**86.42**, **54.80**, **52.54**)	(77.00, 44.50, 38.70)

Table 28: Per-seed image-level and pixel-level audit for seed 333. Each cell reports (AUROC, F1-max, AP) in percent. Bold and underline mark the best and second-best result for each metric, respectively. Pixel-AUPRO is omitted because the available AUPRO fields are placeholders rather than valid measurements. Ours denotes BD-Host with ProtoDeCo; missing Ours cells indicate that the original seed log was unavailable in the audit source, so no aggregate value is backfilled.

(a) WinCLIP, AnomalyCLIP, and AdaCLIP

Domain	Metric	Dataset	WinCLIP	AnomalyCLIP	AdaCLIP
Industrial	Image-level (AUROC, F1-max, AP)	MVTec AD	(70.23, 86.24, 84.63)	(91.31, 92.59, 96.08)	(**93.56**, **93.29**, **96.94**)
		VisA	(64.29, 74.22, 70.32)	(81.39, 80.60, 84.84)	(81.14, 81.03, 84.76)
		BTAD	(57.33, 64.86, 63.64)	(88.89, 86.53, 91.62)	(92.71, 92.92, 97.52)
		KSDD2	(78.67, 72.58, 80.09)	(92.38, 85.08, 93.82)	(95.97, 90.99, 96.81)
		DAGM2007	(59.11, 28.70, 21.59)	(98.03, 90.78, 92.37)	(98.25, 89.98, 93.27)
		DTD-Synthetic	(62.96, 85.63, 82.69)	(93.87, 94.46, 97.47)	(93.95, 93.66, 97.77)
	Pixel-level (AUROC, F1-max, AP)	MVTec AD	(80.69, 22.98, 16.53)	(**90.89**, 38.87, 34.46)	(90.49, **46.82**, **44.96**)
		VisA	(70.05, 9.51, 6.09)	(95.32, 27.62, 20.57)	(**95.76**, **34.56**, **29.23**)
		BTAD	(64.37, 12.26, 8.07)	(93.22, 46.23, 40.29)	(96.41, 52.28, **51.33**)
		KSDD2	(95.58, 27.09, 17.61)	(99.07, 50.80, 44.12)	(**99.32**, 60.88, 65.24)
		DAGM2007	(76.35, 6.97, 2.80)	(**96.83**, 57.06, 53.62)	(95.54, **57.21**, **54.50**)
		DTD-Synthetic	(77.88, 15.67, 9.41)	(97.51, 55.71, 52.57)	(96.99, 61.74, 61.84)
Medical	Image-level (AUROC, F1-max, AP)	OCT17	(63.64, 85.71, 84.29)	(64.33, 85.71, 86.56)	(79.63, 85.80, 92.70)
		BrainMRI	(68.56, 78.42, 77.87)	(96.29, 94.38, 97.17)	(90.51, 87.18, 93.82)
		Brain_AD	(47.33, 90.60, 80.80)	(68.19, 90.64, 89.93)	(80.89, 91.57, 95.06)
		Histopathology_AD	(61.58, 68.26, **59.88**)	(60.12, 66.76, 59.86)	(**62.10**, 68.32, 59.10)
	Pixel-level (AUROC, F1-max, AP)	Brain_AD	(86.19, 19.42, 11.40)	(**95.95**, **52.49**, **55.54**)	(94.51, 40.62, 36.21)
		CVC-ClinicDB	(70.05, 25.90, 18.02)	(85.18, 45.99, 42.30)	(86.47, 45.24, 43.87)
		EndoTect	(61.79, 29.39, 18.50)	(86.82, 54.40, 51.93)	(85.25, 50.72, 48.29)
		Kvasir	(66.42, 33.49, 24.24)	(82.36, 50.64, 44.30)	(82.05, 49.85, 48.87)

(b) VisualAD, FB-CLIP, and Ours

Domain	Metric	Dataset	VisualAD	FB-CLIP	Ours
Industrial	Image-level (AUROC, F1-max, AP)	MVTec AD	(91.80, 92.35, 96.57)	(90.53, 91.79, 95.57)	(92.74, 93.26, 96.85)
		VisA	(82.03, 80.07, 85.45)	(83.16, 81.83, 85.49)	(**85.95**, **82.81**, **88.69**)
		BTAD	(95.05, **95.42**, **98.45**)	(92.72, 89.06, 94.83)	(**95.16**, 93.33, 97.14)
		KSDD2	(97.15, 91.78, 97.44)	(**98.42**, **93.64**, **98.60**)	(98.11, 93.48, 98.47)
		DAGM2007	(98.18, **91.61**, 94.17)	(97.10, 88.12, 91.28)	(**98.74**, 91.21, **94.79**)
		DTD-Synthetic	(93.92, 93.04, 97.39)	(90.54, 92.83, 96.52)	(**98.32**, **97.16**, **99.41**)
	Pixel-level (AUROC, F1-max, AP)	MVTec AD	(90.31, 42.52, 39.95)	(88.56, 42.71, 41.63)	(90.53, 43.88, 40.29)
		VisA	(95.32, 32.88, 26.15)	(94.62, 26.70, 22.67)	(95.54, 33.97, 26.59)
		BTAD	(93.30, **52.73**, 47.77)	(**97.31**, 50.24, 48.07)	(91.97, 47.37, 41.84)
		KSDD2	(99.19, 61.58, 64.63)	(98.69, **61.90**, **67.80**)	(99.10, 56.49, 53.35)
		DAGM2007	(92.84, 48.64, 44.23)	(94.20, 53.48, 49.97)	(92.65, 51.93, 48.39)
		DTD-Synthetic	(97.37, 61.47, 62.57)	(95.14, 54.85, 55.14)	(**97.80**, **64.45**, **63.42**)
Medical	Image-level (AUROC, F1-max, AP)	OCT17	(**92.78**, **91.46**, **97.60**)	(81.83, 87.35, 92.88)	(90.10, 89.30, 96.80)
		BrainMRI	(94.73, 91.48, 94.34)	(89.15, 89.22, 90.16)	(**96.80**, **94.50**, **97.50**)
		Brain_AD	(77.87, 91.46, 93.64)	(69.24, 91.20, 90.04)	(**82.20**, **91.60**, **95.20**)
		Histopathology_AD	(55.47, 67.75, 52.53)	(55.84, 67.66, 53.43)	(**62.10**, **69.80**, 57.40)
	Pixel-level (AUROC, F1-max, AP)	Brain_AD	(93.67, 40.37, 36.50)	(94.92, 46.45, 47.52)	(95.30, 47.10, 42.80)
		CVC-ClinicDB	(85.78, 45.49, 37.30)	(**88.79**, **52.10**, **51.10**)	(86.00, 48.70, 40.70)
		EndoTect	(86.23, 53.55, 48.30)	(**89.35**, **59.00**, **60.18**)	(83.70, 51.20, 46.30)
		Kvasir	(81.62, 49.91, 44.10)	(**86.47**, **56.41**, **55.88**)	(78.30, 46.40, 41.00)

Dataset	BD		BD + PN	
	AUROC	AP	AUROC	AP
MVTec	92.07 ± 0.11	96.44 ± 0.17	92.65 ± 0.13	96.74 ± 0.18
VisA	85.69 ± 0.59	88.39 ± 0.56	86.04 ± 0.66	88.52 ± 0.62
BTAD	95.71 ± 0.44	97.58 ± 0.50	95.71 ± 0.44	97.72 ± 0.46
KSDD2	98.08 ± 0.31	98.43 ± 0.23	97.97 ± 0.33	98.35 ± 0.24
DAGM	99.05 ± 0.43	96.07 ± 1.70	99.22 ± 0.34	96.75 ± 1.39
DTD Synthetic	97.73 ± 0.61	99.27 ± 0.14	97.93 ± 0.49	99.32 ± 0.10
Industrial mean	94.72 ± 0.20	96.03 ± 0.33	94.92 ± 0.18	96.23 ± 0.28

Table 29: Detailed industrial calibration ablation. BD denotes the fully decoupled host without pseudo-normal calibration. BD + PN applies the proposed pseudo-normal calibration to the same host. Values are image AUROC and AP, reported as mean ± population standard deviation over seeds 111, 222, and 333.

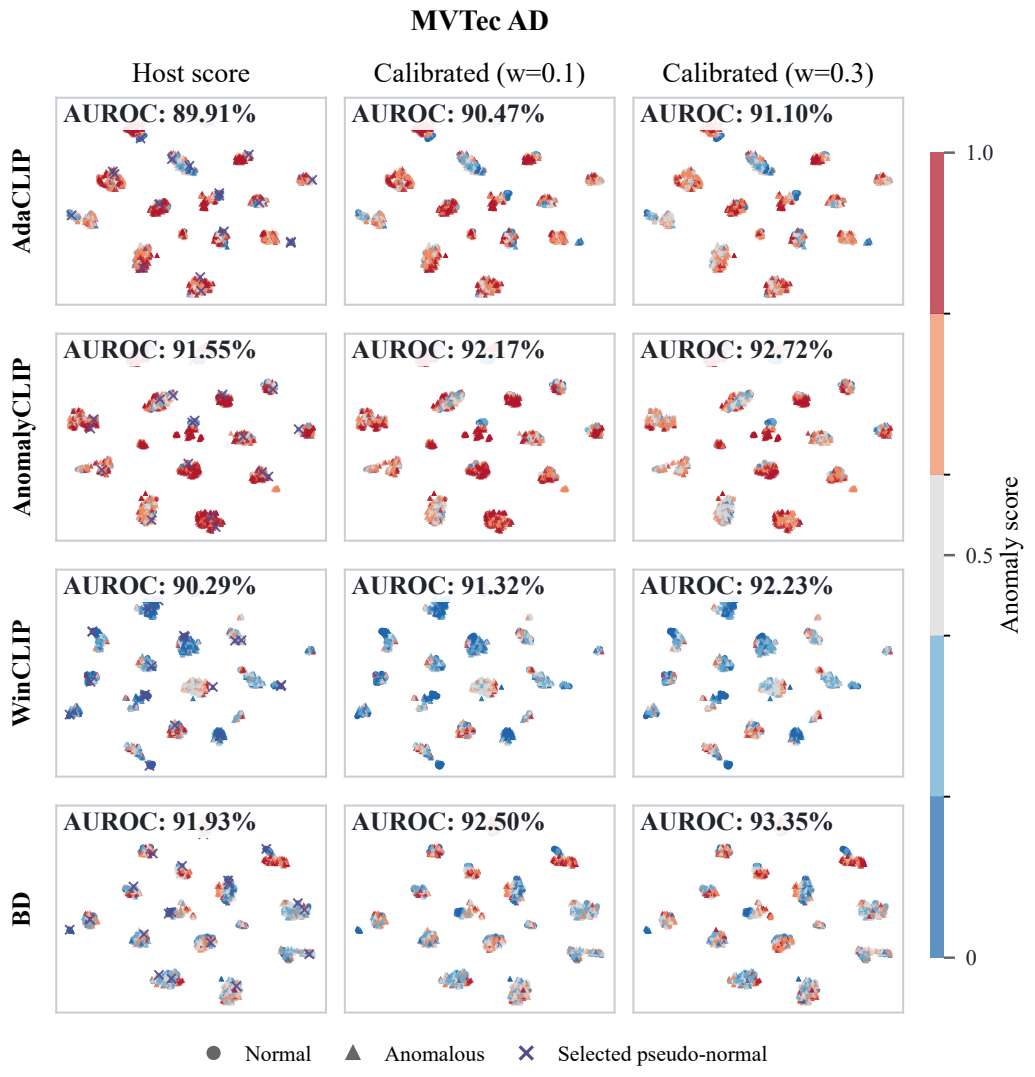


Figure 7: Mvtec

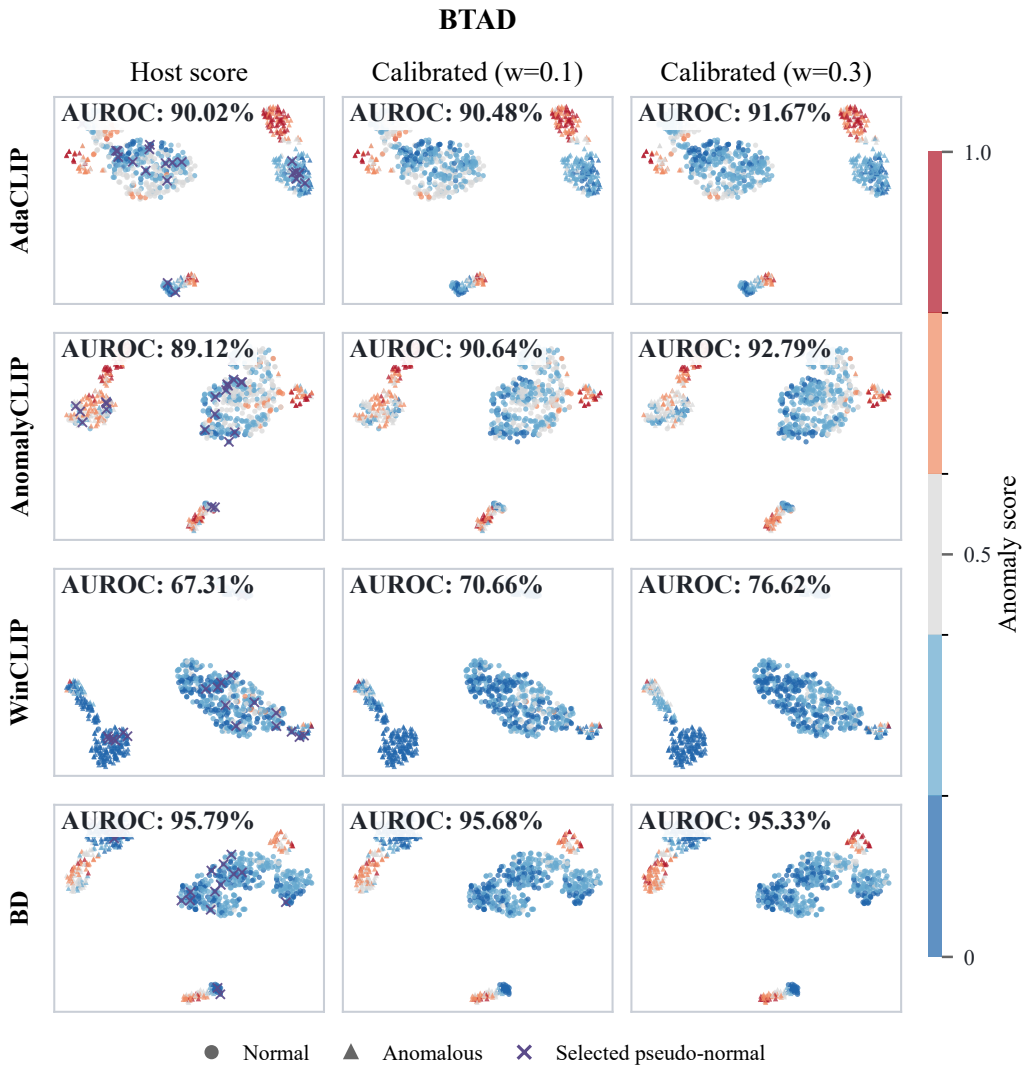


Figure 8: BTAD

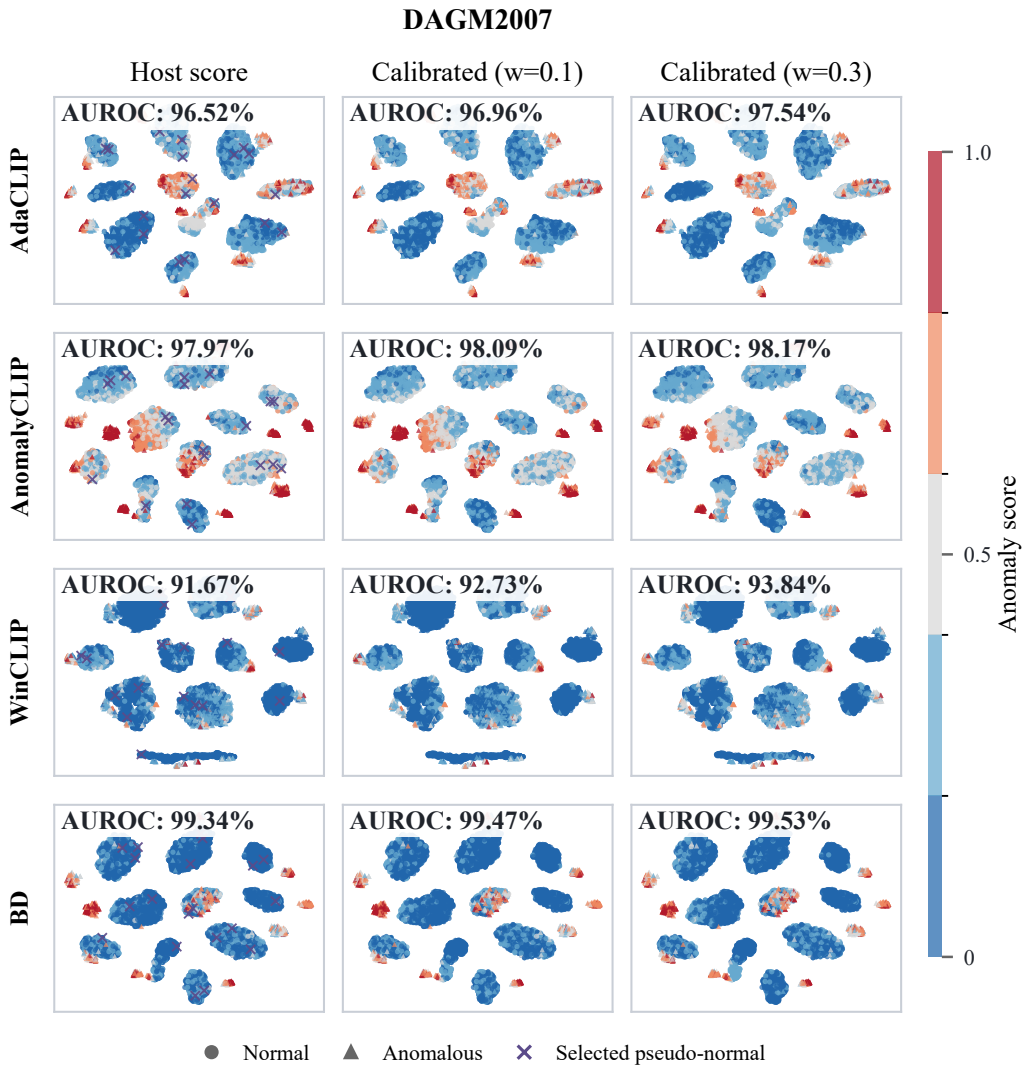


Figure 9: DAGM

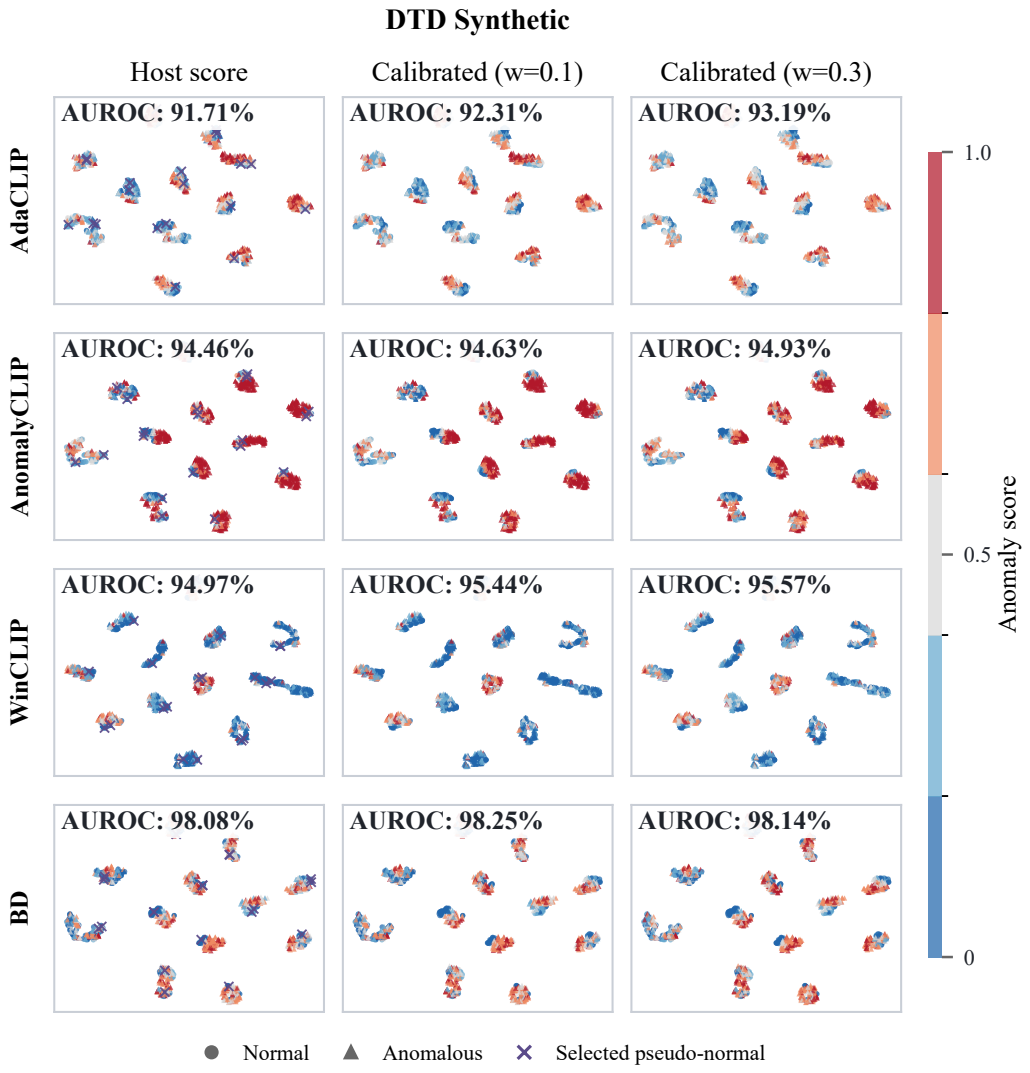


Figure 10: DTD

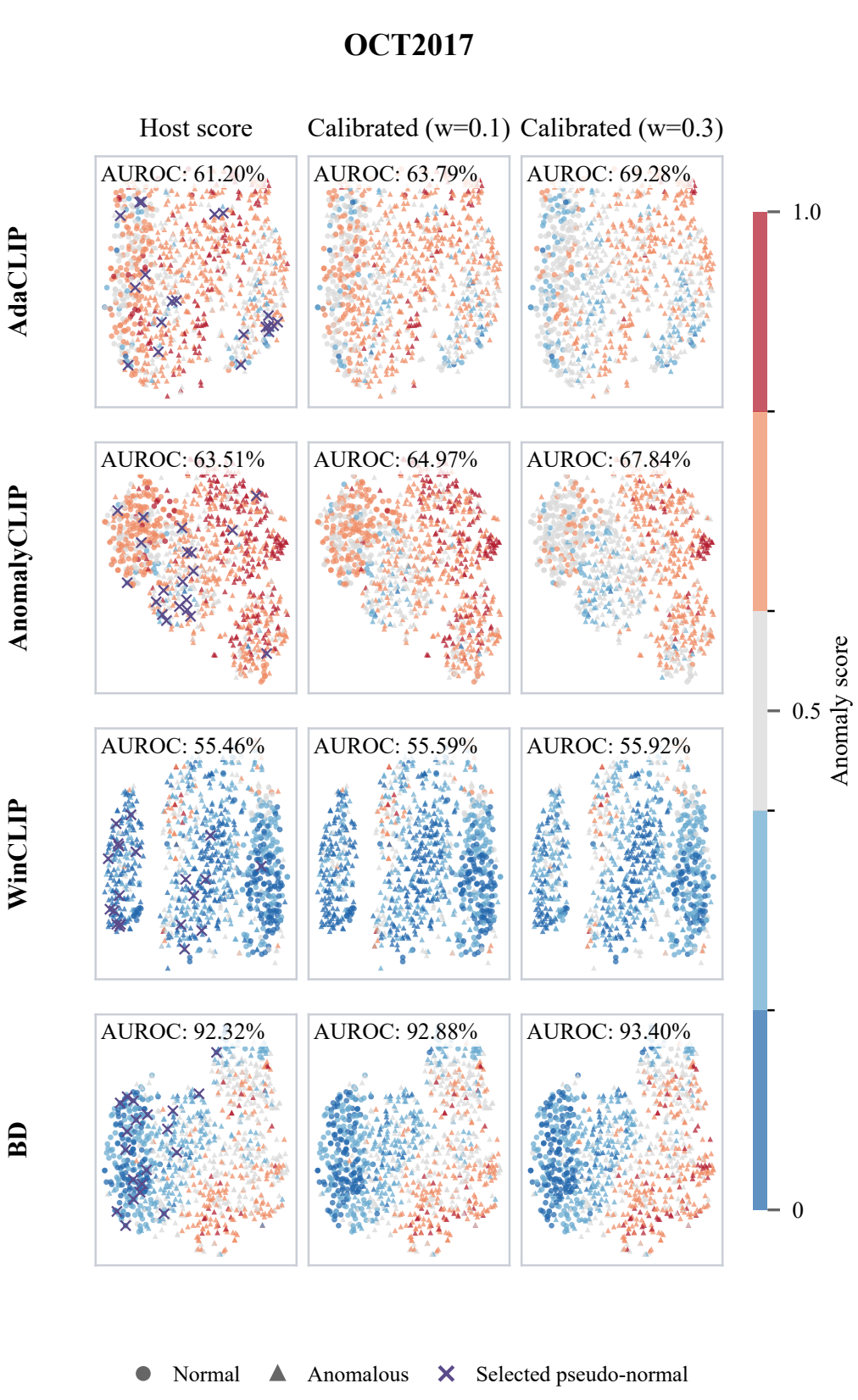


Figure 11: Retin OCT2017

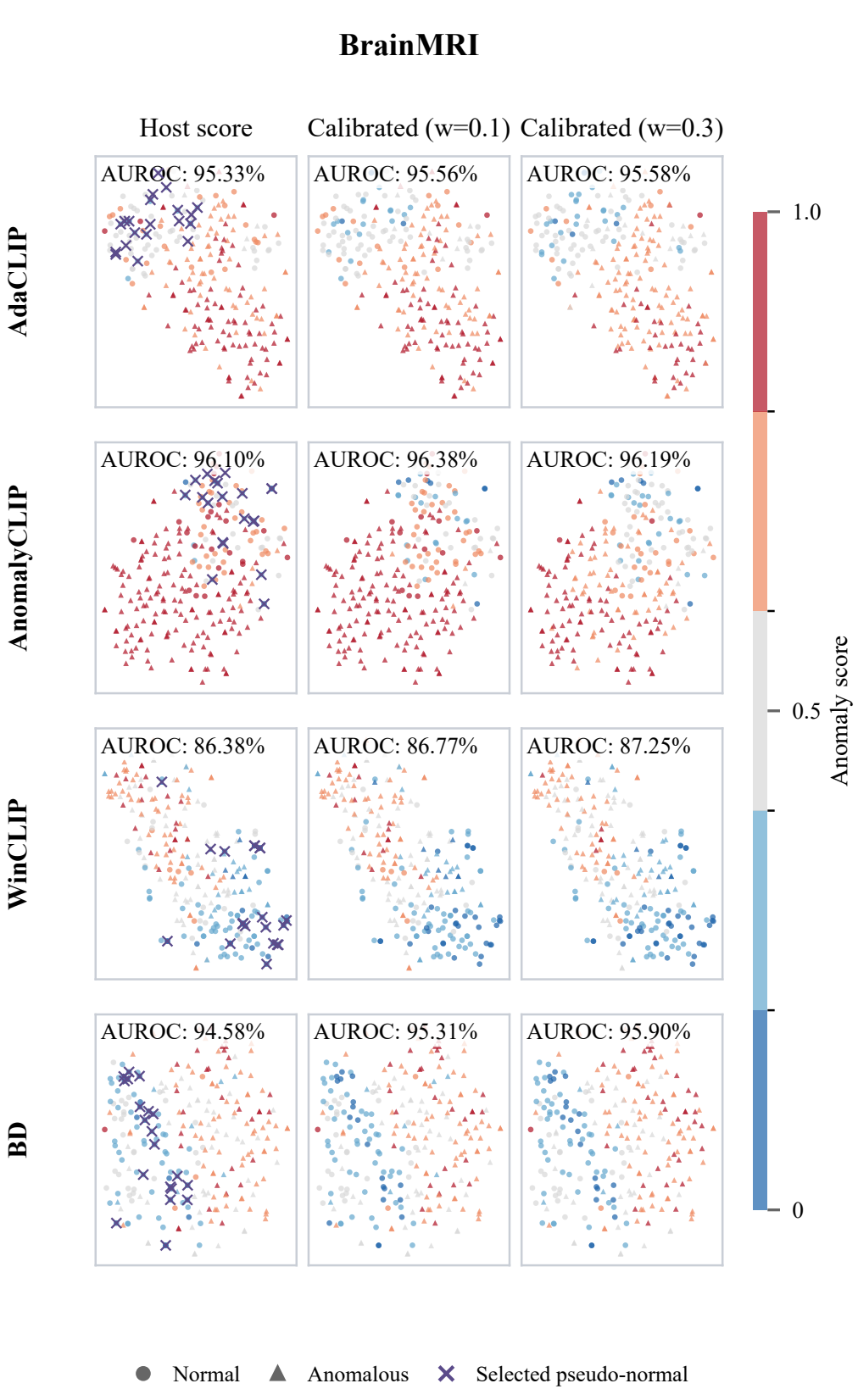


Figure 12: BrainMRI

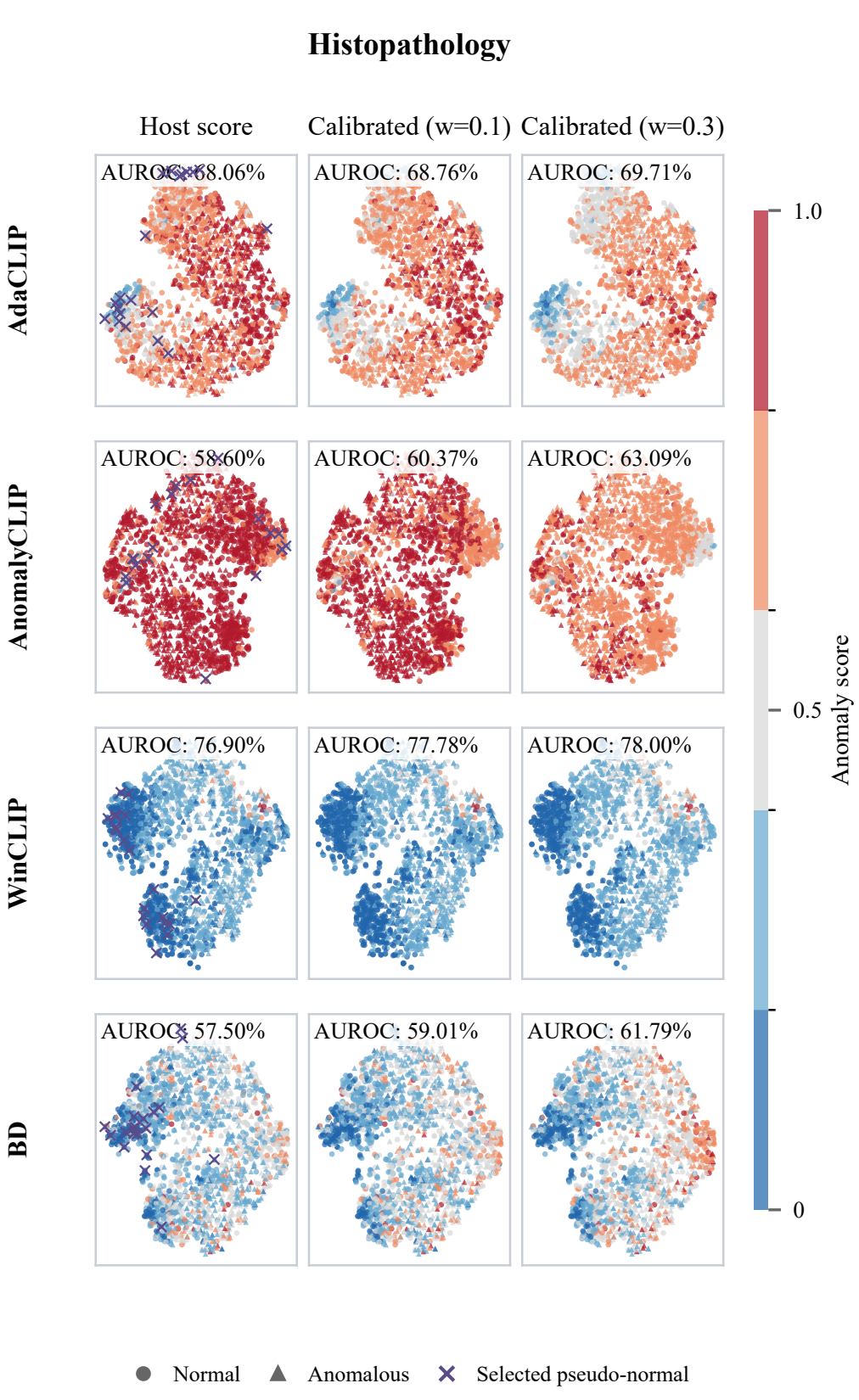


Figure 13: Histopathology