XPAD: LANGUAGE-FREE AND TRAINING-FREE ZERO-SHOT INDUSTRIAL ANOMALY DETECTION VIA CROSS-DATASET PROTOTYPES

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot anomaly detection (ZSAD) detects anomalies of unseen categories without target-domain supervision. Existing ZSAD methods generally fine-tune on an auxiliary dataset to learn transferable textual prompts or visual representations, incurring considerable training costs and leaving the question of what is actually transferred unresolved. Across six industrial benchmarks, we observe that a frozen encoder already organizes normal and anomalous patch features into a shared, dataset-agnostic direction, which can be estimated without any training. This suggests that an effective scoring function can be derived directly from this direction. We instantiate this idea as **XPAD** (Cross-dataset Prototypes for Anomaly Detection), a language-free and training-free framework. XPAD aggregates a single annotated auxiliary dataset into cross-dataset prototypes, and scores test patches by prototype similarity. To sharpen this comparison, it further separates the normal prototype into foreground and background sub-prototypes, and calibrates the aggregation-degree mismatch between dataset-level prototypes and individual query patches. Without textual prompts, additional training, or target-domain data, XPAD achieves state-of-the-art average performance on six industrial benchmarks, and runs directly on pretrained vision backbones such as the CLIP image encoder and DINOv2. Code will be made publicly available.

1 INTRODUCTION

Anomaly detection (AD) aims to identify samples or regions that deviate from the normal appearance, and is central to industrial visual inspection Bergmann et al. (2019). Traditional unsupervised methods Defard et al. (2021); Zavrtanik et al. (2021); Deng & Li (2022); Roth et al. (2022) achieve strong performance by training on large collections of normal images. Motivated by the cost of collecting and training on such data, recent research has shifted toward reducing data dependency: few-shot methods Huang et al. (2022); Xie et al. (2023); Gu et al. (2024) require only a limited number of normal samples, and zero-shot anomaly detection (ZSAD) methods Jeong et al. (2023); Zhou et al. (2023); He et al. (2025); Zhu et al. (2025) eliminate target-domain training altogether, enabling direct deployment on previously unseen categories.

Contemporary ZSAD is largely dominated by a cross-dataset fine-tuning paradigm, which follows two main routes: (i) employing vision-language models to learn transferable textual prompts that represent normal and anomalous semantics, and (ii) employing vision-only models to learn transferable visual representations of normal and anomalous patterns. Both routes aim to learn a universal normal–anomaly pattern from auxiliary data, yet both incur considerable training costs and pose the same fundamental question: what is actually transferred from the auxiliary domain?

We therefore conduct a systematic cross-dataset study across six industrial AD datasets with a frozen visual encoder, aggregating the normal and anomalous patch features of each dataset into dataset-level prototypes and measuring all pairwise similarities (Fig. 1). A consistent pattern emerges: normal prototypes of different datasets are mutually similar, and so are anomaly prototypes, while the normal–anomaly similarity is consistently lower than either. This is measured before any auxiliary-dataset training, so the direction is not learned but already imprinted in the feature space. It follows that the scoring function can be read out directly from the similarity to cross-dataset prototypes.

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.

Based on this insight, we propose XPAD, a language-free and training-free ZSAD framework. XPAD aggregates a single annotated auxiliary dataset into cross-dataset prototypes and scores test patches directly by prototype similarity. To strengthen this simple comparison in industrial scenes, it separates the normal prototype into foreground and background sub-prototypes with a model-free procedure, and calibrates the aggregation-degree mismatch between highly aggregated prototypes and individual query patches via query-side aggregation.

Our contributions are summarized as follows:

- We analyze the dataset-level prototype similarities across six industrial benchmarks and find that the cross-dataset prototypes can serve as effective zero-shot anomaly detectors.
- We propose XPAD, a language-free and training-free ZSAD framework that enhances cross-dataset prototypes through two components: Foreground–Background Normal Prototype Separation (FBNPS) and Aggregation-Degree Calibration (ADC).
- XPAD matches or surpasses the performance of recent training-based, language-dependent prompt-learning methods on both image-level and pixel-level metrics averaged over the six datasets, at no training cost.

2 Related Work

Industrial Anomaly Detection. Recent advances in anomaly detection have predominantly focused on unsupervised methods, also known as full-shot methods Defard et al. (2021); Zavrtanik et al. (2021); Deng & Li (2022); Roth et al. (2022), which rely on large collections of normal images during training. PatchCore Roth et al. (2022), for example, builds a normal-patch memory bank via greedy coreset sampling, thus reducing redundancy, storage, and inference time. In contrast, few-shot AD methods Huang et al. (2022); Xie et al. (2023); Gu et al. (2024) address the challenge of limited normal data. GraphCore Xie et al. (2023), for instance, leverages GNNs to capture visual isometric-invariant features, enabling an effective compact memory bank from few normal images.

Zero-shot Anomaly Detection. Leveraging the strong generalization capabilities of large pre-trained vision-language models Radford et al. (2021), zero-shot AD methods have emerged that bypass the need for any normal reference images. Most zero-shot AD methods Jeong et al. (2023); Chen et al. (2023); Zhou et al. (2023); Cao et al. (2024); He et al. (2025) compare test images against text prompts. WinCLIP Jeong et al. (2023) pioneered this direction by combining a prompt ensemble strategy with multiple window-based forward passes over image patches. Subsequent methods have sought to improve performance through cross-dataset fine-tuning. For instance, AnomalyCLIP Zhou et al. (2023) introduces object-agnostic learnable prompts along with a diagonal-prominent attention map to suppress category semantics. MRAD Xu et al. (2026) constructs a memory bank from an auxiliary dataset for retrieval and injects region priors from the retrieval model into the learnable prompts. VisualAD Hou et al. (2026) eschews textual prompts and instead inserts two learnable tokens (normal and abnormal) into a frozen ViT Dosovitskiy (2020), optimizing them on auxiliary data. While these methods attain strong accuracy, they still rely on textual prompts or require costly training. XPAD distinguishes itself from all of these methods by neither relying on language nor requiring training, yet it matches or surpasses their accuracy via cross-dataset prototypes.

3 Method

We present XPAD, a language-free and training-free ZSAD framework, built upon the observation of cross-dataset prototype similarity. We first formalize the motivation underlying our approach in Sec. 3.1. We then present the proposed method, which comprises Cross-Dataset Prototype Aggregation (Sec. 3.2), Foreground–Background Normal Prototype Separation (FBNPS; Sec. 3.3), and Aggregation-Degree Calibration (ADC; Sec. 3.4). Finally, we summarize the inference pipeline in Sec. 3.5. An overview of the framework is provided in Fig. 2.

3.1 Motivation: Cross-Dataset Prototype Similarity

The direction that separates normal from anomalous content is shared across industrial datasets and needs no learning. We support this claim with a purely empirical study. For each dataset d of

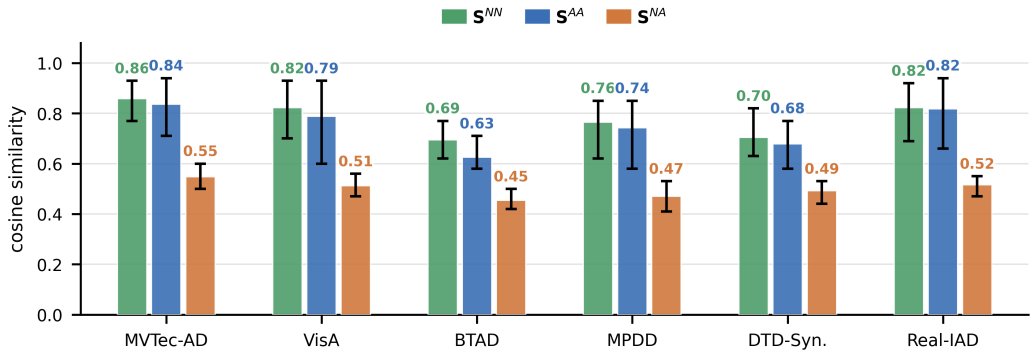

Figure 1: Cross-dataset prototype similarity (DINOv2-R-ViT-L/14). Per dataset, each bar is the mean cosine similarity to the prototypes of the other five datasets, and the black whisker denotes its observed $[\min, \max]$. The two intra-class similarities stay high, whereas normal–anomaly similarity is lower for every dataset without any learning, so the structure already exists in the frozen features.

the six benchmarks we aggregate its normal patch features into a dataset-level normal prototype $\mathcal{P}_N^d$ and its anomalous patch features into an anomaly prototype $\mathcal{P}_A^d$ (the aggregation follows Sec. 3.2), and compute the pairwise cosine similarities

$$\left[\mathbf{S}^{NN}\right]_{d,d'} = \cos\left(\mathcal{P}_N^d, \mathcal{P}_N^{d'}\right), \qquad \left[\mathbf{S}^{AA}\right]_{d,d'} = \cos\left(\mathcal{P}_A^d, \mathcal{P}_A^{d'}\right), \qquad \left[\mathbf{S}^{NA}\right]_{d,d'} = \cos\left(\mathcal{P}_N^d, \mathcal{P}_A^{d'}\right) \tag{1}$$

between the 30 cross-dataset pairs, $d \neq d'$ (Fig. 1). Two observations hold consistently:

1. **Dataset-agnostic intra-class structure.** The cross-dataset entries of both $\mathbf{S}^{NN}$ and $\mathbf{S}^{AA}$ are high: averaged over the 30 cross-dataset pairs, the mean similarity between normal prototypes of two different datasets is 0.78 and that between anomaly prototypes 0.75, so both classes point to a shared direction of the feature space.
2. **Stable inter-class separation.** Averaged over the same pairs, the mean normal–anomaly similarity is only 0.50, and $[\mathbf{S}^{NA}]_{d,d'} < \min([\mathbf{S}^{NN}]_{d,d'}, [\mathbf{S}^{AA}]_{d,d'})$ holds for all 30 cross-dataset pairs with a mean margin of 0.24. Moreover, this margin is more pronounced on more diverse datasets, such as MVTec-AD and Real-IAD.

Both follow from prototypes built on frozen features with a training-free aggregation operator, so they reflect a property of the encoder rather than an artifact of auxiliary training. Industrial anomalies are local deviations from the surrounding structural or textural regularity, a category-agnostic property, which may explain why the encoder organizes normal and anomalous features along a direction independent of the specific object categories or imaging domains. This finding implies that the scoring function can be read out directly, without any training. XPAD instantiates this idea.

3.2 Cross-Dataset Prototype Aggregation

Let $\mathcal{D}^{aux} = \{(I_i, M_i, y_i)\}_{i=1}^N$ be an annotated auxiliary dataset, where $I_i \in \mathbb{R}^{H\times W\times 3}$ is an image, $M_i \in \{0,1\}^{H\times W}$ is its pixel-level anomaly mask ($M_i = \mathbf{0}$ for normal images), and $y_i \in \{0,1\}$ is the image-level label. We extract patch features $\{f_{i,p}\}_{p=1}^P = \phi(I_i) \in \mathbb{R}^{P\times d}$ with a frozen encoder ϕ, where P is the number of patches and d the feature dimension. The anomaly mask is downsampled to the patch grid by average pooling, $\tilde{m}_i = \mathrm{Down}(M_i) \in [0,1]^P$, so that $\tilde{m}_{i,p}$ is the fraction of anomalous pixels covered by patch p and naturally serves as a soft weight.

Image-level prototypes. For an anomalous image, anomaly regions are emphasized by weighting patch features with the downsampled mask, yielding the image-level anomaly prototype

$$\mu_i^A = \frac{\sum_{p=1}^P \tilde{m}_{i,p}\, f_{i,p}}{\sum_{p=1}^P \tilde{m}_{i,p}}, \tag{2}$$

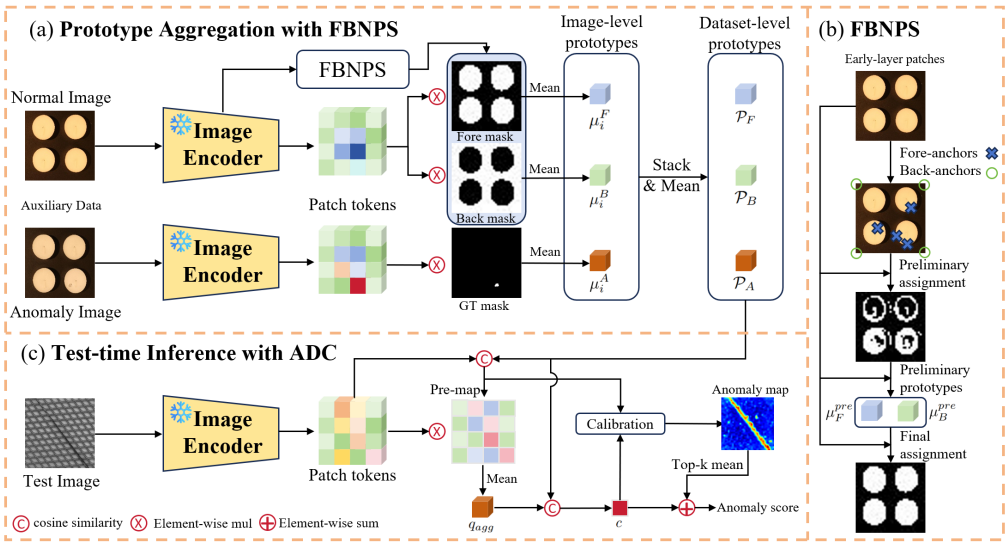

Figure 2: Overview of the proposed XPAD framework. It is language-free and training-free: dataset-level prototypes are aggregated from an auxiliary dataset with foreground–background normal prototype separation (FBNPS), and test images are scored by prototype similarity with aggregation-degree calibration (ADC).

where $f_{i,p}$ is the feature of the p-th patch. For a normal image, the image-level normal prototype is the plain average of its patch features,

$$\mu_i^N = \frac{1}{P}\sum_{p=1}^{P} f_{i,p}. \tag{3}$$

Dataset-level prototypes. Averaging the image-level prototypes of the same class over the whole auxiliary dataset gives the dataset-level anomaly and normal prototypes:

$$\mathcal{P}_A = \frac{1}{|\mathcal{D}_A^{aux}|}\sum_{i\in\mathcal{D}_A^{aux}} \mu_i^A, \qquad \mathcal{P}_N = \frac{1}{|\mathcal{D}_N^{aux}|}\sum_{i\in\mathcal{D}_N^{aux}} \mu_i^N, \tag{4}$$

where $\mathcal{D}_A^{aux}$ and $\mathcal{D}_N^{aux}$ denote the anomalous and normal subsets of the auxiliary dataset. Aggregating at two levels (first within an image, then across images) avoids the dominance of images with large defect regions and yields robust statistics.

Basic anomaly scoring. For a test patch with feature q_p, the basic anomaly score is computed from the excess similarity to the anomaly prototype over the normal prototype,

$$s_a = \cos(q_p, \mathcal{P}_A), \qquad s_n = \cos(q_p, \mathcal{P}_N), \tag{5}$$

$$s_p^{base} = \sigma\Big((s_a - s_n)/\tau\Big), \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function mapping the score to $(0, 1)$ and τ a temperature hyper-parameter.

3.3 Foreground–Background Normal Prototype Separation (FBNPS)

In industrial inspection, anomalies tend to occur on or around foreground objects, while backgrounds are usually normal. Thus the dataset-level anomaly prototype in Eq. (4) is biased toward foreground features. However, a single global normal prototype $\mathcal{P}_N$ mixes foreground and background normality and is often background-dominated, since background pixels constitute approximately 70% of the image area in MVTec-AD (75% in VisA), as measured with the FBNPS masks. For a foreground query patch, such a background-dominated $\mathcal{P}_N$ provides a poor estimate of foreground normality and may distort the anomaly score. For a background patch, the comparison is trivially passed

and provides little discriminative information. We therefore separate the normal reference into a foreground normal prototype $\mathcal{P}_F$ and a background normal prototype $\mathcal{P}_B$.

The separation requires knowing, for each patch, whether it belongs to the foreground or the background. We obtain this partition with a simple four-step procedure that uses early-layer ViT features $\{z_{i,p}\}_{p=1}^{P} = \phi^l(I_i)$ (which preserve the spatial layout better than later-layer features) and the mild assumption that the four image corners belong to the background. For every normal image of the auxiliary dataset:

1. **Fore- and back-anchors.** Take the early-layer patch features at the four corners of the patch grid, $\mathcal{B} = \{b_1, b_2, b_3, b_4\}$, as background anchors. Then, select k patch features $\mathcal{A} = \{a_1, \ldots, a_k\}$ that are farthest from the anchors selected so far by iterated farthest-point sampling (FPS) in cosine distance. As the corners are background, these farthest points tend to lie on the central foreground object.
2. **Preliminary assignment.** For each patch $z_{i,p}$, compute its similarities to the two anchor sets, and assign it to the pre-foreground set if $\max_{a\in\mathcal{A}} \cos(z_{i,p}, a) > \max_{b\in\mathcal{B}} \cos(z_{i,p}, b)$, and to the pre-background set otherwise.
3. **Preliminary prototypes.** Patches assigned to the pre-foreground set are averaged into a preliminary foreground prototype μ_F^{pre} and those assigned to the pre-background set into a preliminary background prototype μ_B^{pre}.
4. **Final assignment.** Patch $z_{i,p}$ is assigned to the foreground if it is more similar to the preliminary foreground prototype than to the preliminary background prototype,
$$\cos(z_{i,p}, \mu_F^{pre}) > \cos(z_{i,p}, \mu_B^{pre}), \tag{7}$$
and to the background otherwise. This final pass replaces the noisy point-wise decisions of step 2 with decisions based on aggregated statistics.

After all normal images are partitioned, the image-level foreground prototype μ_i^F and the image-level background prototype μ_i^B are the averages of the foreground and background patches of image i. The dataset-level prototypes

$$\mathcal{P}_F = \frac{1}{|\mathcal{D}_N^{aux}|} \sum_{i\in\mathcal{D}_N^{aux}} \mu_i^F, \qquad \mathcal{P}_B = \frac{1}{|\mathcal{D}_N^{aux}|} \sum_{i\in\mathcal{D}_N^{aux}} \mu_i^B \tag{8}$$

are obtained by averaging over all normal auxiliary images. The foreground and background partition is derived from the early-layer features, whereas $\mathcal{P}_F$ and $\mathcal{P}_B$ average the later-layer features of the assigned patches, so that all normal prototypes are defined in the same feature space as $\mathcal{P}_A$ and the query patch features.

Similarity with separated normal prototypes. Given the foreground and background normal prototypes, the similarity of a test patch q_p to the normal reference becomes

$$s_n^* = \max\big(\cos(q_p, \mathcal{P}_F), \cos(q_p, \mathcal{P}_B)\big), \tag{9}$$

i.e., the similarity to the normal reference is replaced by the maximum over the two normal prototypes. The maximum operation keeps the comparison anchored to whichever normal mode is closer, so a patch is judged normal only if it is close to some normal mode. This removes the bias that a single global mean introduces when one normal mode (e.g., the background) dominates it.

3.4 Aggregation-Degree Calibration (ADC)

Dataset-level prototypes are aggregated from a large number of patch features across an entire dataset. Such highly aggregated statistics are stable, lie in the high-density region of the feature distribution, and act as the expected directions of normality and abnormality. In contrast, individual test patches without any aggregation fluctuate around their class direction. Cosine similarities measured between entities of different aggregation degrees are therefore statistically biased: the distribution of similarities to a prototype is systematically shifted for individual patches compared with what would be obtained by comparing two equally aggregated features. Using the raw per-patch scores directly thus yields miscalibrated scores and confidence—especially for image-level evidence, where a few outlier patches can dominate.

We thus compensate for this bias by repeating the aggregation operator on the query side, producing a pseudo-prototype with a comparable aggregation degree. Specifically, for a test image with patch features $\{q_p\}_{p=1}^{P}$:

1. Compute the per-patch scores $s_p = \sigma\Big((s_a - s_n^*)/\tau\Big)$ to construct a preliminary map and select the subset $\mathcal{T}$ of the top $r\%$ patches by score, mimicking the way in which prototype extraction aggregates only the relevant (anomalous) regions.
2. Aggregate the selected patches with their anomaly scores as weights,
$$q_{agg} = \frac{\sum_{p\in\mathcal{T}} s_p\, q_p}{\sum_{p\in\mathcal{T}} s_p}. \tag{10}$$
3. Evaluate the anomaly score of the pseudo-prototype,
$$c = \sigma\Big(\frac{\cos(q_{agg}, \mathcal{P}_A) - \max\big(\cos(q_{agg}, \mathcal{P}_F), \cos(q_{agg}, \mathcal{P}_B)\big)}{\tau}\Big), \tag{11}$$
and treat c as the calibration coefficient of the test image.

By virtue of the cross-dataset similarity shown in Sec. 3.1, the pseudo-prototype of a normal test image stays close to the shared normal direction and yields a small c, whereas that of an anomalous test image is pulled toward the anomaly direction and yields a large c. The similarity of each patch with the anomaly prototype becomes

$$s_a^* = (c + \lambda)\cos(q_p, \mathcal{P}_A), \tag{12}$$

where λ is a hyper-parameter controlling the weight of the uncalibrated anomaly similarity. When c is large, the anomaly-similarity evidence of every patch is amplified, boosting genuine defect responses; when c is small, the evidence is damped, suppressing false alarms. c acts as an adaptive gain (an image-level confidence weight) on the anomaly-similarity term. It calibrates the confidence of per-patch responses with image-level evidence obtained at a matched aggregation degree, correcting the bias caused by the aggregation-degree mismatch.

3.5 Inference and Anomaly Scoring

At inference time, for a query image I^t we extract patch features $\{q_p\}_{p=1}^{P} = \phi(I^t) \in \mathbb{R}^{P\times d}$ and compute the calibrated per-patch scores

$$s_p^{final} = \sigma\Big((s_a^* - s_n^*)/\tau\Big), \tag{13}$$

which are upsampled to the input resolution to obtain the pixel-level anomaly map. The image-level anomaly score combines the strongest pixel-level evidence with the calibrated global evidence:

$$S^{img} = \mathrm{Mean}\big(\mathrm{TopK}(\hat{S})\big) + c, \tag{14}$$

where $\hat{S}$ is the upsampled anomaly map and $\mathrm{TopK}(\cdot)$ selects the pixels with the highest scores. The entire pipeline is fully deterministic: once the prototypes $\mathcal{P}_A, \mathcal{P}_F, \mathcal{P}_B$ are computed from the auxiliary dataset, per-image inference requires only $\mathcal{O}(P)$ similarity computations, TopK selection, and averaging. No textual prompt, learnable parameter, target data, or auxiliary learner is involved.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate ZSAD performance on six public industrial benchmarks covering multiple objects, textures, and diverse imaging conditions: MVTec-AD Bergmann et al. (2019) (15 object/texture categories), VisA Zou et al. (2022) (12 complex-object categories), BTAD Mishra et al. (2021) (3 industrial products), MPDD Jezek et al. (2021) (6 metal-part categories), DTD-Synthetic Aota et al. (2023) (12 texture categories with synthetic anomalies), and Real-IAD Wang et al. (2024) (30 real-world multi-view categories).

Table 1: Image-level ZSAD comparison (I-AUROC%, I-AP%). Bold indicates the best performance, while underline denotes the second-best result.

Method	Venue	MVTec-AD	VisA	BTAD	MPDD	DTD-Syn.	Real-IAD	Avg.
Training-based methods								
AnomalyCLIP	ICLR'24	(91.5, 96.2)	(82.1, 85.4)	(88.3, 87.3)	(77.0, 82.0)	(93.5, 97.0)	(69.4, 64.2)	(83.6, 85.4)
RareCLIP	ICCV'25	(91.5, 96.5)	(86.1, 89.0)	(91.7, 93.4)	(78.1, 83.5)	(95.0, 98.5)	(75.7, 74.4)	(86.4, 89.2)
FAPrompt	ICCV'25	(91.9, 95.7)	(84.6, 86.8)	(92.2, 92.5)	(80.1, 83.9)	(96.2, 98.1)	(77.3, 74.8)	(87.1, 88.6)
MRAD-CLIP	ICLR'26	(**94.0**, **97.4**)	(85.7, 88.3)	(92.1, 92.8)	(76.7, 77.2)	(95.8, 98.3)	(73.1, 70.1)	(86.2, 87.4)
FB-CLIP	CVPR'26	(92.4, 96.6)	(89.5, 90.7)	(93.2, 96.4)	(79.1, 86.2)	(97.9, 99.1)	(80.6, 78.4)	(88.8, 91.2)
VisualAD-CLIP	CVPR'26	(92.2, 96.7)	(84.7, 87.6)	(94.3, 96.5)	(71.4, 75.3)	(97.7, 99.2)	(77.1, 74.4)	(86.2, 88.3)
VisualAD-DINOv2	CVPR'26	(90.1, 94.8)	(83.1, 86.8)	(93.3, 93.8)	(73.9, 77.3)	(97.2, 98.9)	(80.0, 77.8)	(86.3, 88.2)
Training-free methods								
WinCLIP	CVPR'23	(91.8, 95.1)	(78.1, 77.5)	(83.3, 84.1)	(61.5, 69.2)	(63.8, 83.5)	(66.8, 62.7)	(74.2, 78.7)
MRAD-TF	ICLR'26	(79.0, 90.0)	(75.0, 79.1)	(88.2, 90.3)	(77.5, 82.2)	(93.1, 97.1)	(65.6, 61.1)	(79.7, 83.3)
XPAD-CLIP (Ours)	-	(93.2, 96.9)	(89.4, 91.1)	(96.7, **98.0**)	(79.0, 86.5)	(**98.0**, **99.3**)	(79.9, 78.3)	(89.4, 91.7)
XPAD-DINOv2 (Ours)	-	(92.3, 96.3)	(**92.1**, **93.7**)	(**97.3**, 96.9)	(**90.7**, **93.8**)	(96.4, 99.0)	(**84.3**, **82.3**)	(**92.2**, **93.7**)

Table 2: Pixel-level ZSAD comparison (P-AUROC%, AUPRO%). Bold indicates the best performance, while underline denotes the second-best result.

Method	Venue	MVTec-AD	VisA	BTAD	MPDD	DTD-Syn.	Real-IAD	Avg.
Training-based methods								
AnomalyCLIP	ICLR'24	(91.1, 81.4)	(95.5, 87.0)	(94.2, 74.8)	(96.5, 88.7)	(97.9, 92.3)	(95.1, 84.0)	(95.1, 84.7)
RareCLIP	ICCV'25	(91.5, 86.2)	(95.7, 90.2)	(91.7, 73.2)	(95.8, 88.4)	(97.6, 93.2)	(95.9, 85.5)	(94.7, 86.1)
FAPrompt	ICCV'25	(90.6, 83.3)	(95.9, 87.7)	(95.6, 75.1)	(96.5, 87.9)	(98.3, 93.3)	(95.0, 82.1)	(95.3, 84.9)
MRAD-CLIP	ICLR'26	(**93.0**, 86.8)	(95.9, 88.0)	(96.1, 77.0)	(**97.3**, 89.1)	(98.4, 92.0)	(94.8, 80.8)	(95.9, 85.6)
FB-CLIP	CVPR'26	(91.9, 85.7)	(**96.3**, 91.4)	(95.8, 80.0)	(96.9, 91.1)	(98.3, 92.6)	(95.9, 88.2)	(95.9, 88.2)
VisualAD-CLIP	CVPR'26	(90.8, 87.5)	(95.8, 91.0)	(92.6, 78.5)	(96.8, 90.8)	(98.6, 96.5)	(97.4, 90.9)	(95.3, 89.2)
VisualAD-DINOv2	CVPR'26	(91.3, 88.6)	(95.3, 88.2)	(94.6, 79.0)	(95.8, 88.9)	(98.0, 94.8)	(96.4, 89.9)	(95.2, 88.2)
Training-free methods								
WinCLIP	CVPR'23	(85.1, 64.6)	(79.6, 56.8)	(71.4, 32.8)	(71.2, 48.9)	(78.4, 51.0)	(84.5, 54.6)	(78.4, 51.5)
MRAD-TF	ICLR'26	(86.7, 63.5)	(91.0, 71.0)	(83.0, 41.3)	(93.7, 79.7)	(93.4, 80.7)	(88.0, 54.5)	(89.3, 65.1)
XPAD-CLIP (Ours)	-	(89.9, 87.6)	(95.1, 91.2)	(93.4, 79.7)	(96.4, **91.3**)	(96.9, 92.8)	(96.6, 91.2)	(94.7, 89.0)
XPAD-DINOv2 (Ours)	-	(91.6, **89.9**)	(95.9, **91.9**)	(**98.5**, **88.5**)	(96.4, 90.3)	(**98.9**, **96.6**)	(**97.9**, **95.4**)	(**96.5**, **92.1**)

Evaluation protocol. Following the cross-dataset ZSAD protocol of prior work Zhou et al. (2023); Hu et al. (2026); Zhu et al. (2025), we use MVTec-AD as the auxiliary dataset by default to build the prototypes and test directly on the remaining datasets; when MVTec-AD itself is the target, VisA serves as the auxiliary dataset. For image-level anomaly classification we report the Area Under the ROC Curve (I-AUROC) and the Average Precision (I-AP). For pixel-level anomaly segmentation we report the pixel-level AUROC (P-AUROC) and the Area Under the Per-Region Overlap curve (AUPRO).

Implementation details. We instantiate XPAD on two frozen ViT-based visual backbones: **XPAD-CLIP** uses CLIP-ViT-L/14-336 Radford et al. (2021), and **XPAD-DINOv2** uses DINOv2-R-ViT-L/14 Darcet et al. (2024). All images are resized to 518×518. Patch features for prototype aggregation and scoring come from layers 12, 16, and 20, while FBNPS in Sec. 3.3 uses layer-6 features. The number of fore-anchors in FBNPS is fixed to $k = 4$. The aggregation ratio of ADC is fixed to $r = 50\%$, and the balance coefficient $\lambda = 0.6$. The temperature is fixed to $\tau = 0.1$ and $\mathrm{TopK}(\cdot)$ in anomaly scoring selects the top 0.1% of pixels by score.

4.2 Comparison with State-of-the-Art Methods

We compare XPAD (instantiated as XPAD-CLIP and XPAD-DINOv2) with representative ZSAD methods: WinCLIP Jeong et al. (2023), MRAD-TF / MRAD-CLIP Xu et al. (2026), AnomalyCLIP Zhou et al. (2023), RareCLIP He et al. (2025), FAPrompt Zhu et al. (2025), FB-CLIP Hu et al. (2026), and VisualAD-CLIP / VisualAD-DINOv2 Hou et al. (2026). To ensure a protocol-consistent comparison, we evaluate all cross-dataset methods under the same protocol as XPAD. Tab. 1 reports the image-level comparison (I-AUROC / I-AP) and Tab. 2 the pixel-level comparison (P-AUROC / AUPRO) on the six datasets plus the average.

Against training-free methods. Against the strongest training-free baseline MRAD-TF, XPAD-CLIP improves by $+9.7$ I-AUROC, $+8.4$ I-AP, $+5.4$ P-AUROC and $+23.9$ AUPRO on average, showing that cross-dataset prototypes are a stronger reference than memory-bank retrieval.

Against training-based methods. At the image level XPAD-DINOv2 reaches 92.2 / 93.7 on average, ahead of the best training-based method FB-CLIP by $+3.4$ / $+2.5$, and XPAD-CLIP (89.4 / 91.7) surpasses it as well. This gain comes mainly from ADC, which calibrates the image-level decision with evidence aggregated at a matched degree. At the pixel level XPAD-DINOv2 attains the best average of all compared methods, with $+2.9$ AUPRO over the second-best method VisualAD-CLIP. XPAD-CLIP is weaker in P-AUROC but still reaches a competitive AUPRO which we attribute to FBNPS suppressing background false positives. Both instantiations perform consistently across the two backbones, indicating that XPAD does not depend on a specific encoder.

Training cost comparison. Tab. 3 compares the one-off prototype aggregation of XPAD-CLIP with the training stages of MRAD-CLIP and VisualAD-CLIP. XPAD uses no learnable parameter, requires $5\times$ less GPU memory than MRAD-CLIP and $8\times$ less than VisualAD-CLIP, and yields a model file of only 0.04 MB, while obtaining the highest average I-AUROC among the three compared methods. The comparison shows that comparable accuracy can be obtained without training any parameters.

Table 3: Training cost comparison.

Method	Learnable Params	GPU (GB)	Model Size (MB)	I-AUROC
MRAD-CLIP	9,491,968	9.4	54	86.2
VisualAD-CLIP	31,522,836	15.0	150	86.2
XPAD-CLIP (Ours)	**0**	**1.8**	**0.04**	**89.4**

4.3 Ablation Studies

Module ablation. Tab. 4 reports the effect of removing each of the two task-specific modules from the full model under both backbones, using the six-dataset averages. Removing either module degrades the metrics at all levels for both backbones, most notably at the image level. FBNPS sharpens the normal reference, raising I-AUROC and AUPRO by around one point, while ADC alone already recovers most of the image-level gap to the full model ($+3.3$ I-AUROC for XPAD-CLIP and $+3.4$ for XPAD-DINOv2). Removing both causes the largest drop, and the full model is best on both backbones, so the two gains are complementary under either encoder.

Table 4: Module ablation of XPAD-CLIP and XPAD-DINOv2. Each cell reports the six-dataset average of the corresponding metric. Image (I-AUROC%, I-AP%) and Pixel (P-AUROC%, AUPRO%).

FBNPS	ADC	XPAD-CLIP		XPAD-DINOv2	
		Image	Pixel	Image	Pixel
×	×	(85.4, 88.3)	(94.2, 87.4)	(87.3, 88.9)	(96.2, 91.2)
×	√	(88.7, 90.4)	(94.6, 87.4)	(90.7, 92.5)	(96.4, 91.3)
√	×	(86.4, 89.1)	(94.5, **89.0**)	(88.3, 89.4)	(96.3, **92.1**)
√	√	(**89.4**, **91.7**)	(**94.7**, **89.0**)	(**92.2**, **93.7**)	(**96.5**, **92.1**)

Within FBNPS. Tab. 5 reports the effect of four choices of the normal reference within FBNPS, using the six-dataset averages. The foreground-only reference is already close to the global prototype reference: at the image level it matches or exceeds the global prototype on both backbones, which is expected since defects typically occur on the foreground object. The background-only reference is the weakest of the four at image level and AUPRO, confirming that background statistics carry less discriminative evidence for anomalies. Combining the two references with the max operator is what yields the gain: FBNPS exceeds the single global prototype by $+0.7$ I-AUROC and $+1.6$ AUPRO at XPAD-CLIP and by $+1.5$ I-AUROC and $+0.8$ AUPRO at XPAD-DINOv2.

Table 5: Ablation on the normal-prototype decomposition inside FBNPS.

Normal reference	XPAD-CLIP		XPAD-DINOv2	
	Image	Pixel	Image	Pixel
$\mathcal{P}_N$ (no FBNPS)	(88.7, 90.4)	(94.6, 87.4)	(90.7, 92.5)	(96.4, 91.3)
$\mathcal{P}_F$ only	(89.3, 91.7)	(94.4, 88.0)	(90.7, 92.7)	(95.4, 91.3)
$\mathcal{P}_B$ only	(87.8, 89.5)	(94.3, 86.0)	(89.3, 91.1)	(96.3, 90.5)
$\mathcal{P}_F$ & $\mathcal{P}_B$	(89.4, 91.7)	(94.7, 89.0)	(92.2, 93.7)	(96.5, 92.1)

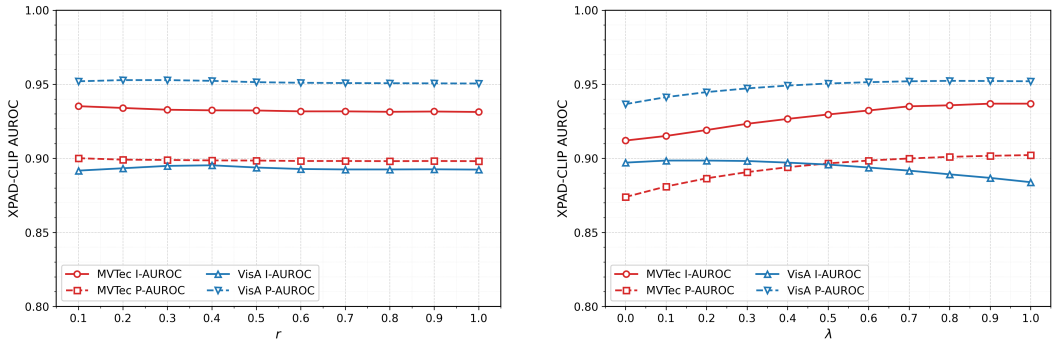

Figure 3: Sensitivity of XPAD-CLIP to the ADC hyper-parameters: the aggregation ratio r (left) and the balance coefficient λ (right). I-AUROC and P-AUROC are reported on MVTec-AD and VisA.

Table 6: Ablation of the auxiliary dataset on XPAD-CLIP. Each row is the auxiliary dataset used to aggregate the prototypes and each column the test dataset. (I-AUROC%, P-AUROC%) are reported, and gray marks the degenerate settings in which the auxiliary dataset and the target coincide.

Test→ ↓Auxiliary	MVTec-AD	VisA	BTAD	MPDD	DTD-Syn.	Real-IAD
MVTec-AD	(94.8, 90.1)	(89.4, 95.1)	(96.7, 93.4)	(79.0, 96.4)	(98.0, 96.9)	(79.9, 96.6)
VisA	(93.2, 89.9)	(90.1, 95.0)	(95.0, 93.1)	(78.6, 96.9)	(95.2, 96.8)	(78.9, 96.6)
BTAD	(87.0, 88.1)	(81.6, 91.2)	(94.4, 94.3)	(70.5, 91.8)	(75.3, 91.1)	(70.5, 95.7)
MPDD	(90.1, 89.9)	(85.4, 94.8)	(94.1, 89.9)	(78.4, 97.8)	(91.5, 95.5)	(73.2, 95.8)
DTD-Syn.	(90.3, 88.9)	(83.3, 92.8)	(84.4, 82.4)	(64.8, 94.4)	(98.7, 97.3)	(75.3, 95.1)
Real-IAD	(92.8, 88.8)	(89.0, 93.4)	(95.5, 93.0)	(74.9, 96.2)	(97.2, 97.0)	(80.1, 96.7)

Within ADC. ADC adds two hyper-parameters, the aggregation ratio r and the balance coefficient λ, swept in Fig. 3 on MVTec-AD and VisA. Both image-level and pixel-level accuracy are essentially flat in r. In contrast, λ trades VisA image-level accuracy against MVTec-AD image-level accuracy; it increases monotonically on MVTec-AD and decreases on VisA. We choose $\lambda = 0.6$ as a compromise rather than an optimum.

Ablation of auxiliary dataset. We conduct an ablation of auxiliary dataset input on XPAD-CLIP in Tab. 6, which shows that performance does not depend on any particular auxiliary dataset. In most cases, changing the auxiliary dataset affects the results slightly. At the image level, using MVTec-AD as the auxiliary dataset even yields higher I-AUROC than using the target dataset's own data on BTAD and MPDD ($+2.3$ and $+0.6$ I-AUROC). The exceptions are the two least diverse auxiliary datasets: BTAD, which contains only three classes and a limited variety of anomaly types, and DTD-Synthetic, a texture-only synthetic dataset. When BTAD serves as the auxiliary dataset, the I-AUROC on DTD-Synthetic drops to 75.3; even in this case, however, the model does not collapse, as it still achieves a P-AUROC of 91.1. This finding is consistent with our design intuition: XPAD reads out a shared anomaly direction whose estimate depends on how diverse the auxiliary dataset is, not on which particular dataset is used.

5 Conclusion

In this paper we ask what cross-dataset fine-tuning actually transfers, and find that the answer is largely a structure that already exists in the frozen features: a dataset-agnostic direction, which is measurable before any auxiliary-dataset training. We therefore propose XPAD, a language-free and training-free zero-shot anomaly detection framework that reads out this direction with three prototypes. XPAD keeps the backbone frozen and can be instantiated on any pretrained encoder. Its two instantiations, XPAD-CLIP and XPAD-DINOv2, match or surpass recent training-based, language-dependent prompt-learning methods on average across six industrial benchmarks. We hope this study will encourage the community to rethink the role of language and training in ZSAD: for industrial anomalies, which are essentially local structural and textural deviations, pure visual statistics may already suffice.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and to edit the manuscript for improved readability. We did not use these tools to generate synthetic datasets; develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for proving mathematical claims; assist in the writing of proofs; propose or refine hypotheses; design or provide feedback on research methodology or experiments; implement methods; clean or reformat datasets; support qualitative or thematic data analysis; or interpret results. All AI-assisted text was reviewed, verified, and revised by the human authors. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

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

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.

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.

Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 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.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9737–9746, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1932–1940, 2024.

Jianfang He, Min Cao, Silong Peng, and Qiong Xie. Rareclip: Rarity-aware online zero-shot industrial anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24478–24487, 2025.

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 (CVPR)*, pp. 21346–21356, June 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 (CVPR)*, pp. 35659–35669, June 2026.

Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, pp. 303–319. Springer, 2022.

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.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pp. 66–71, 2021. doi: 10.1109/ICUMT54235.2021.9631567.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pp. 01–06. IEEE, 2021.

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.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22883–22892, June 2024.

Guoyang Xie, Jinbao Wang, Jiaqi Liu, Feng Zheng, and Yaochu Jin. Pushing the limits of few-shot anomaly detection in industry vision: Graphcore. In *International Conference on Learning Representations*, 2023.

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

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8330–8339, 2021.

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*, 2023.

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

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pp. 392–408. Springer, 2022.

APPENDIX

A IMPLEMENTATION DETAILS

XPAD is instantiated on two frozen ViT backbones with 24 transformer blocks: CLIP-ViT-L/14-336 Radford et al. (2021) for **XPAD-CLIP** and the self-supervised DINOv2-R-ViT-L/14 Darcet et al. (2024) for **XPAD-DINOv2**. Inputs are resized to 518×518 and the patch grid is 37×37, i.e. $P = 1369$ tokens per image. All backbones are kept frozen and used in inference mode, and no test-time adaptation is performed. We use MVTec-AD as the default auxiliary dataset for cross-dataset evaluation, and VisA as the auxiliary dataset when MVTec-AD is the target. Patch features for prototype aggregation and scoring come from layers 12, 16, and 20, giving 9 prototypes in total to store, and the final anomaly map is the average over these three layers. FBNPS uses layer-6 features, which carry more spatial information, and the number of fore-anchors is fixed to $k = 4$. Within ADC, the aggregation ratio is fixed to $r = 50\%$, and the balance coefficient $\lambda = 0.6$. The upsampled anomaly map $\hat{S}$ is smoothed using a Gaussian filter with $\sigma = 4$. The temperature is fixed to $\tau = 0.1$, and $\mathrm{TopK}(\cdot)$ in anomaly scoring selects the top 0.1% of pixels by score. Furthermore, in constructing cross-dataset prototypes, we exclude anomalous images whose anomaly ratio is below 0.1%: such a region covers only a small fraction of any patch, so the resulting patch features would be dominated by normal content. We report results averaged over all categories of each target dataset.

B DATASETS

Tab. 7 summarizes our six benchmarks: five real industrial inspection datasets and one synthetic texture benchmark, covering objects, textures, and mixed categories.

MVTec-AD Bergmann et al. (2019) is the standard benchmark of the field with 15 categories (10 objects and 5 textures). Its categories differ widely in appearance and imaging setup, which makes it a natural source of cross-category variability.

VisA Zou et al. (2022) is a challenging industrial dataset that includes 12 object subsets spanning PCBs, capsules, and foodstuffs. Compared with MVTec-AD it has more complex object structures and greater variation in pose and position.

BTAD Mishra et al. (2021) includes three real industrial products, with pixel-level masks for anomalous images. Its limited variety of anomaly types makes it a less diverse auxiliary dataset.

MPDD Jezek et al. (2021) covers six categories of painted metal parts, captured under varied orientations, positions, distances, and illumination, with non-uniform backgrounds. Its unconstrained object layout makes it a difficult target in our evaluation.

DTD-Synthetic Aota et al. (2023) inserts synthetic defects into 12 texture classes, yielding 357 normal images and 947 anomalous images. Because it is purely textural and contains no foreground object, this benchmark tests FBNPS outside the object scenes for which it was designed.

Real-IAD Wang et al. (2024) is the largest benchmark in our study, with high-resolution images of 30 object categories, each photographed from five viewpoints. Its defect area ratio spans a much wider range than those of the earlier datasets, making it the most challenging target.

Table 7: The six benchmarks used in this work. "Type" distinguishes object from texture data.

Dataset	Classes	Normal Images	Anomaly Images	Type
MVTec-AD	15	467	1,258	object & texture
VisA	12	962	1,200	object
BTAD	3	451	290	object & texture
MPDD	6	176	282	object
DTD-Syn.	12	357	947	texture
Real-IAD	30	63,256	51,329	object

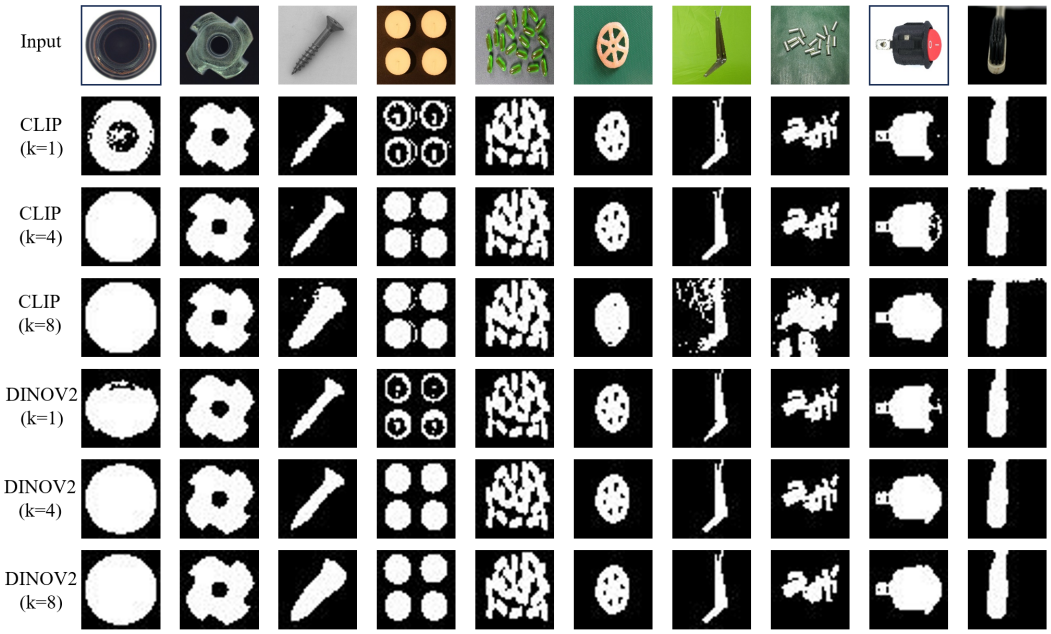

Figure 4: Foreground masks produced by FBNPS on the CLIP and DINOv2 backbones for different k.

C Additional Experimental Results

FBNPS with texture scenes. FBNPS is designed for object scenes, so we also evaluate its behavior in pure texture scenes. Tab. 8 shows that with MVTec-AD as the auxiliary dataset, FBNPS improves the performance of XPAD-CLIP on the DTD-Synthetic dataset, especially at the pixel level ($+0.7$ AUPRO). Tab. 9 shows that with the DTD-Synthetic dataset as the auxiliary dataset, FBNPS slightly decreases the image-level performance of XPAD-CLIP while increasing its pixel-level performance. This indicates that FBNPS mainly brings pixel-level gains, and the method remains stable even in pure texture scenes.

Table 8: Performance of XPAD-CLIP on the DTD-Synthetic dataset.

Auxiliary	I-AUROC	I-AP	P-AUROC	AUPRO
MVTec-AD (w/o FBNPS)	97.9	99.2	96.7	92.1
MVTec-AD (with FBNPS)	98.0	99.3	96.9	92.8

Table 9: Performance of XPAD-CLIP with the DTD-Synthetic dataset as the auxiliary dataset (I-AUROC%, AUPRO%).

Auxiliary	MVTec-AD	VisA	BTAD	MPDD	Avg.
DTD-Synthetic (w/o FBNPS)	(90.5, 82.5)	(83.3, 83.8)	(84.7, 63.4)	(65.8, 84.6)	(81.1, 78.6)
DTD-Synthetic (with FBNPS)	(90.3, 82.7)	(83.3, 84.1)	(84.4, 63.5)	(64.8, 84.9)	(80.7, 78.8)

Ablation on the number of fore-anchors k. FBNPS uses farthest-point sampling to find the fore-anchors, and we conduct an ablation study on their number k. Fig. 4 shows that the area of the foreground mask tends to increase with k. A small k may fail to capture the complete foreground region, while a large k may allow background patches to contaminate the foreground mask. Tab. 10 shows that the choice of k has only a minor effect on XPAD's performance, as the impact of occasional foreground/background misclassifications is diluted by averaging over many patches.

Visualization results. Qualitative zero-shot anomaly detection results on six industrial benchmarks are shown in Fig. 5. The figure displays the input images, ground-truth masks, and anomaly score maps obtained by XPAD using CLIP and DINOv2.

Table 10: Ablation on the number of fore-anchors k, evaluated on the MVTec-AD, VisA and BTAD datasets (I-AUROC%, P-AUROC%).

k	XPAD-CLIP			XPAD-DINOv2		
	MVTec-AD	VisA	BTAD	MVTec-AD	VisA	BTAD
1	(93.2, 89.9)	(89.5, 95.0)	(96.5, 93.3)	(91.9, 91.6)	(91.6, 95.8)	(96.5, 98.4)
4	(93.2, 89.9)	(89.4, 95.1)	(96.7, 93.4)	(92.3, 91.6)	(92.1, 95.9)	(97.3, 98.5)
8	(93.1, 89.8)	(89.2, 95.2)	(96.7, 93.4)	(92.2, 91.7)	(92.0, 95.9)	(97.5, 98.4)

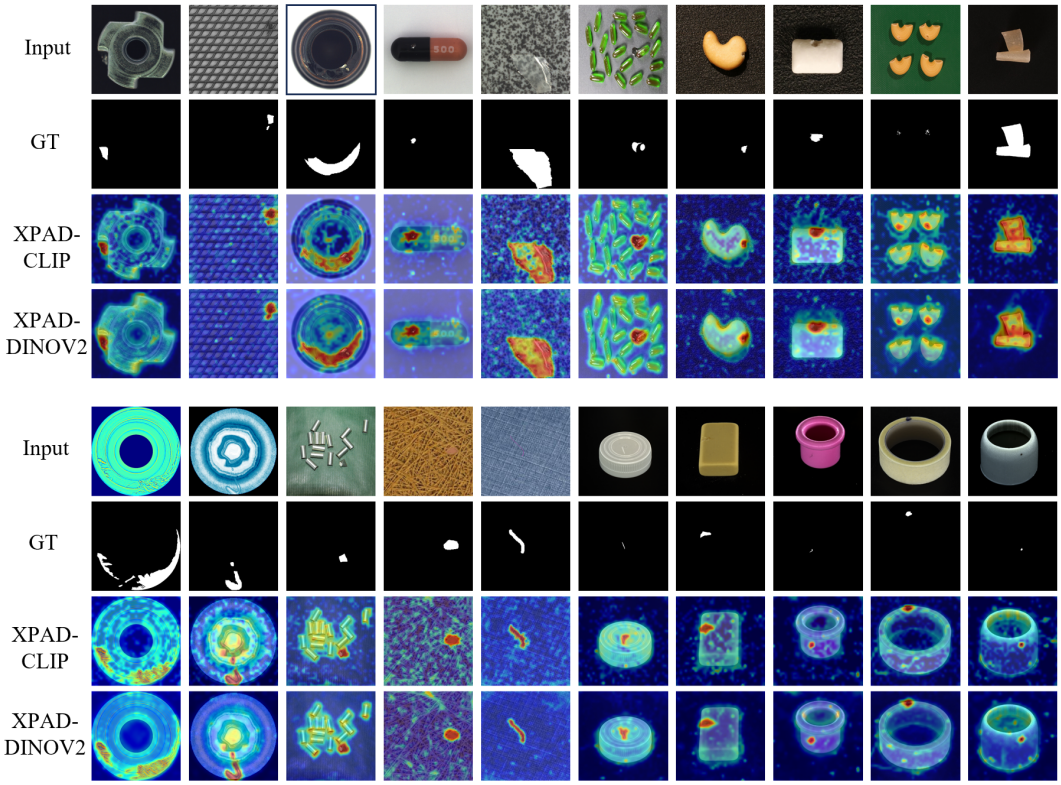

Figure 5: Anomaly score maps for samples from six datasets.