WHICH NORMAL SAMPLES MATTER? QUERY-RELEVANT REFERENCE SELECTION FOR INDUSTRIAL ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent few-shot methods have achieved strong performance in industrial anomaly detection with only a few normal samples as references. However, in practical industrial scenarios, normal data is often abundant and can be continuously collected over time. This raises a practical problem: *when a large pool of normal samples is available, which samples are most suitable to serve as references for a given query image?* We find that normal references do not contribute equally, and their relevance to the query can substantially affect anomaly detection performance. Based on this observation, we propose a training-free anomaly detection framework built on a foundation visual encoder. For each query image, our method first retrieves a small set of normal samples with the highest retrieval scores for the query. We then use the retrieved references to model anomalies at the patch level by jointly considering how well each query patch matches normal features, and how well it can be explained by the normal variations across references. The resulting anomaly evidence is aggregated across multi-level features for anomaly detection. Experimental results show that our method achieves competitive performance across multi-class and single-class settings, with particularly strong results on pixel-level anomaly localization. Moreover, applying our reference retrieval strategy to existing few-shot methods consistently improves their performance, further demonstrating the importance of query-relevant normal references.

1 INTRODUCTION

Industrial anomaly detection (IAD) aims to identify defective samples and localize anomalous regions while usually relying on normal training data. Full-shot methods such as MambaAD (He et al., 2024), Dinomaly (Guo et al., 2025) and OmiAD (Feng et al., 2025) exploit all normal training samples to detect anomalies. Recent few-shot approaches such as AnomalyDINO (Damm et al., 2025), FoundAD (Zhai et al., 2026), SubspaceAD (Lendering et al., 2026) and ANoCo (Seo et al., 2026) have shown that strong anomaly detection performance can be achieved using only a few normal reference images. These advances substantially reduce the dependence on large training sets and make anomaly detection more flexible under limited-data conditions.

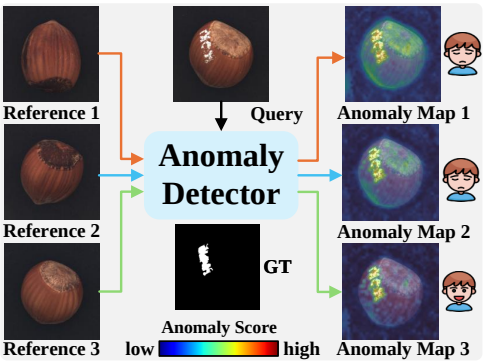


Figure 1: Different normal references can lead to different anomaly detection results.

However, in practical industrial systems, normal data is often abundant and can be continuously collected during production, forming a large pool of normal data that covers diverse poses, structures and appearances. This raises a practical question beyond the conventional few-shot setting: *when a large pool of normal samples is available, which samples are most suitable to serve as references for a given query image?* As illustrated in Figure 1, even with the same anomaly detector, different normal references can lead to noticeably different anomaly detection results for the same query image. This suggests that normal references may not contribute equally to anomaly inference. While

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.

many existing few-shot methods mainly focus on how to exploit visual feature information from a limited set of normal samples, the potential influence of reference variation and reference selection on detection performance has received comparatively less attention. For example, AnomalyDINO (Damm et al., 2025) detects anomalies through patch-level matching with few-shot normal references, while ANoCo (Seo et al., 2026) evaluates the non-conformity of query features with respect to the normal feature structure. Both focus on exploiting the given reference set, whereas the influence of which normal samples are selected as references remains less explored. This observation motivates us to further consider the relevance of each normal sample to the query.

Based on this observation, we propose a training-free anomaly detection framework following a retrieve-then-score paradigm. Given a query image, a foundation visual encoder retrieves highly relevant normal samples by considering both global image information and patch-level structural similarity. We then estimate anomaly evidence from the retrieved references using two complementary measures: nearest-neighbor distances obtained through spatially constrained patch matching and reconstruction residuals from local low-rank subspaces. The anomaly evidence from multiple encoder layers is fused to produce image- and pixel-level anomaly scores. Without task-specific training, our framework can be directly applied across different categories and datasets.

We evaluate our method on MVTec-AD (Bergmann et al., 2019) and VisA (Zou et al., 2022) under both multi-class and single-class settings. The proposed framework achieves competitive image-level detection performance and particularly strong pixel-level anomaly localization. We further conduct experiments on reference selection, reference budget, and different forms of normal variation. The results consistently show that query-relevant normal references provide a stronger basis for anomaly inference than arbitrary references under the same budget and that simply introducing more normal samples may not always be meaningful. These findings highlight reference relevance as an important consideration when there is abundant normal data.

Our main contributions are summarized as follows:

- We revisit IAD from the perspective of abundant normal data and observe that normal references do not contribute equally. Beyond the number of available references, their relevance to the query has a substantial impact on anomaly detection performance.
- We propose a training-free anomaly detection framework that first retrieves a compact set of query-relevant normal references from a large normal archive and then constructs local normality through spatially constrained matching and local normal subspaces.
- We demonstrate strong performance on MVTec-AD and VisA in both multi-class and single-class settings, especially for pixel-level anomaly localization. Moreover, applying the same reference retrieval strategy to existing few-shot methods consistently improves their performance, providing further evidence for the importance of reference relevance.

2 Related Work

Few-shot Anomaly Detection Few-shot anomaly detection aims to perform anomaly detection with only a small number of normal references. Early few-shot methods mainly exploit vision-language models and prompt learning (Jeong et al., 2023; Li et al., 2024; Lv et al., 2025; Tao et al., 2025), while recent training-free approaches increasingly rely on frozen visual foundation representations. AnomalyDINO (Damm et al., 2025) detects anomalies by matching query patches against features from a given few-shot reference set, while SubspaceAD (Lendering et al., 2026) models the normal variation of the given references with a subspace. These methods assume that the small reference set is already specified and mainly study how to exploit it for anomaly inference. In contrast, we study the practical problem of which normal samples should be selected as references when a much larger normal archive is available.

Retrieval-based anomaly detection provides a related but different perspective. RAID (Cai et al., 2026) retrieves normal representations to provide auxiliary evidence for a learned noise-suppression module, using retrieval as part of a trained anomaly inference pipeline. RAD (Zhang et al., 2026) performs training-free anomaly detection by retrieving and matching query features against a normal feature memory, showing that memory retrieval can effectively replace task-specific reconstruction. In these methods, retrieval primarily serves the downstream anomaly scoring process. Our work instead explicitly formulates normal reference selection as a separate query-adaptive stage: before

anomaly scoring, we select a compact set of query-relevant samples from a normal image pool according to global semantic and patch-level structural compatibility, and then construct query-conditioned local normality with these references.

Multi-class Anomaly Detection Conventional IAD methods (Deng & Li, 2022; Liu et al., 2023; Zhang et al., 2024) typically build category-specific detectors, whereas multi-class AD aims to handle multiple object categories within a unified system. UniAD (You et al., 2022) introduced the multi-class anomaly detection setting, aiming to handle multiple categories within a unified model. Subsequent methods such as Dinomaly, OmiAD, DecAD (Wang et al., 2025), and FoundAD have further improved multi-class anomaly detection by learning shared normal representations across categories. Most of these methods learn a unified representation of normality from all categories, which remains fixed during inference. In contrast, our method does not rely on a fixed normal representation for all categories. Instead, it keeps normal samples from different categories in a shared pool and selects a small set of relevant references for each query. As a result, the normal evidence used for anomaly detection can adapt to each query. This also enables multi-class anomaly detection without category-specific models or task-specific training.

3 METHOD

3.1 OVERVIEW

Given a query image $\mathbf{x}_q$, our goal is to estimate an image-level anomaly score and a pixel-level anomaly map using only normal reference images and a frozen foundation visual encoder. In the multi-class anomaly detection setting, normal training images from all categories are combined into a shared reference pool. As shown in Figure 2, for a query image, we first use the CLS token and patch tokens to retrieve K visually compatible normal images from the pool. At each scoring layer $l \in \mathcal{L}$ in the vision encoder, every query patch is then matched to spatially nearby patches in the retrieved references. The resulting local normal observations provide two complementary anomaly cues: nearest-neighbor discrepancy and deviation from a local normal subspace. These cues are combined across multi-level features to detect anomalies.

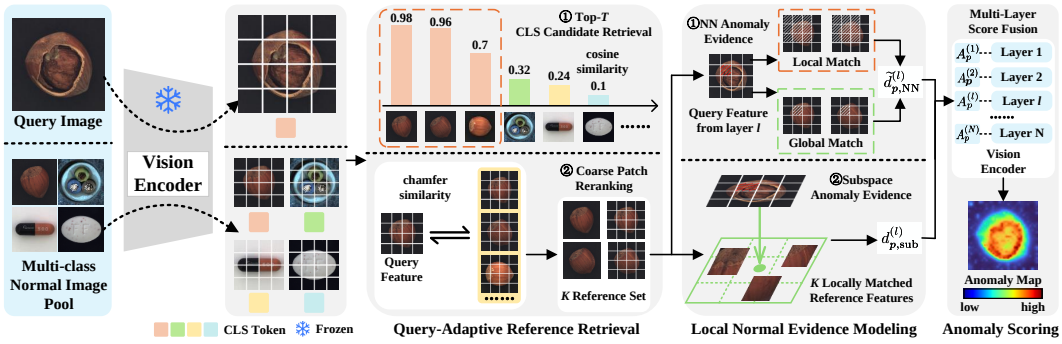


Figure 2: Anomaly detection pipeline of our method.

3.2 QUERY-ADAPTIVE REFERENCE RETRIEVAL STRATEGY

A query image may differ substantially from many normal images in the shared reference pool. Additional analysis is provided in Appendix A.1. Therefore, we first retrieve a compact query-specific reference set using global CLS similarity followed by coarse patch reranking. In the vision encoder, we use a CLS token from layer l_{cls} for global retrieval and patch tokens from layer l_{ret} for reference reranking.

CLS candidate retrieval. Let $\mathbf{cls}_q$ and $\mathbf{cls}_i$ denote the normalized CLS tokens of the query and the i-th normal image from layer l_{cls}. Their cosine similarity is

$$s_i^{\text{cls}} = \mathbf{cls}_q^\top \mathbf{cls}_i. \tag{1}$$

The T images with the highest CLS similarities form the candidate set $\mathcal{C}_q$.

Coarse patch reranking. The patch tokens from layer l_{ret} are reshaped into a spatial grid, adaptively pooled, and ℓ_2-normalized to a $G \times G$ representation. Let $\bar{\mathbf{Q}}_q, \bar{\mathbf{Q}}_i \in \mathbb{R}^{G^2 \times C}$ denote the coarse query and candidate tokens. Their pairwise similarity matrix is $\mathbf{M}_i = \bar{\mathbf{Q}}_q \bar{\mathbf{Q}}_i^\top$. We measure their correspondence using bidirectional Chamfer similarity:

$$s_{q\to i} = \frac{1}{G^2}\sum_{a=1}^{G^2} \max_b \mathbf{M}_i(a,b), \qquad s_{i\to q} = \frac{1}{G^2}\sum_{b=1}^{G^2} \max_a \mathbf{M}_i(a,b), \tag{2}$$

$$s_i^{\text{patch}} = \frac{1}{2}\left(s_{q\to i} + s_{i\to q}\right). \tag{3}$$

Since CLS and patch similarities have different numerical ranges, we fuse their relative rankings rather than their raw values. For a candidate ranked at position $\pi \in \{1, \dots, T\}$, its linear rank score is $\rho(\pi) = 1 - \frac{\pi-1}{T-1}$. The final retrieval score is

$$s_i^{\text{ref}} = \alpha\rho\big(\pi_i^{\text{cls}}\big) + (1-\alpha)\rho\Big(\pi_i^{\text{patch}}\Big), \tag{4}$$

where α controls the contribution of global appearance. The top K candidates according to Eq.(4) constitute the final query-adaptive reference set $\mathcal{R}_q$.

3.3 Local Normal Evidence Modeling

Spatially Constrained Local Matching. Let $\mathbf{q}_p^{(l)}$ from a scoring layer l denote a normalized query patch token at spatial position $p = (y, x)$, and let $\mathbf{s}_{k,j}^{(l)}$ denote patch j from the k-th retrieved reference in $\mathcal{R}_q$. For each query patch, we restrict correspondence search to a local neighborhood

$$\mathcal{N}_d(p) = \{(y', x') : |y' - y| \le d,\ |x' - x| \le d\}, \tag{5}$$

where d denotes the neighborhood radius, corresponding to a $(2d+1) \times (2d+1)$ local window and positions outside the feature grid are excluded. This constraint allows moderate spatial displacement while preventing matches to distant but visually similar structures. From each reference image, we retain the most similar patch within this neighborhood:

$$j_{k,p}^* = \arg\max_{j \in \mathcal{N}_d(p)} \left(\mathbf{q}_p^{(l)}\right)^\top \mathbf{s}_{k,j}^{(l)}, \qquad \mathbf{z}_{k,p}^{(l)} = \mathbf{s}_{k,j_{k,p}^*}^{(l)}. \tag{6}$$

The resulting local normal support is $\mathcal{Z}_p^{(l)} = \left\{\mathbf{z}_{1,p}^{(l)}, \dots, \mathbf{z}_{K,p}^{(l)}\right\}$.

Nearest-Neighbor Anomaly Evidence. The best local normal similarity is $s_{p,\text{local}}^{(l)} = \max_k \left(\mathbf{q}_p^{(l)}\right)^\top \mathbf{z}_{k,p}^{(l)}$, yielding the local nearest-neighbor distance

$$d_{p,\text{NN}}^{(l)} = 1 - s_{p,\text{local}}^{(l)}. \tag{7}$$

Although spatially constrained matching efficiently finds highly similar local correspondences, it may miss the globally appropriate match under large spatial displacement. We therefore additionally search all patches in the retrieved references:

$$s_{p,\text{global}}^{(l)} = \max_{k,j} \left(\mathbf{q}_p^{(l)}\right)^\top \mathbf{s}_{k,j}^{(l)}. \tag{8}$$

The improvement of the global match over the local one is

$$g_p^{(l)} = s_{p,\text{global}}^{(l)} - s_{p,\text{local}}^{(l)}. \tag{9}$$

We use a soft gate $b_p^{(l)} = \text{clip}\left(\frac{g_p^{(l)}}{\delta}, 0, 1\right)$ to obtain the corrected nearest-neighbor evidence

$$\widetilde{d}_{p,\text{NN}}^{(l)} = d_{p,\text{NN}}^{(l)} - b_p^{(l)} g_p^{(l)}. \tag{10}$$

Thus, the correction preserves the local distance when local and global correspondences agree, while progressively reducing it when a better global match is available.

Local Subspace Anomaly Evidence. Nearest-neighbor matching models normality through discrete exemplars but does not explicitly capture variations shared by the matched normal features. We therefore construct a query- and position-conditioned local affine model, where stable normal variations are represented by a ridge-regularized low-rank reconstruction and the residual serves as complementary anomaly evidence. The K locally matched normal features are stacked as $\mathbf{Z}_p^{(l)} \in \mathbb{R}^{K\times C}$. Their mean is $\boldsymbol{\mu}_p^{(l)} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{z}_{k,p}^{(l)}$. We center both the normal observations and the query feature:

$$\widehat{\mathbf{Z}}_p^{(l)} = \mathbf{Z}_p^{(l)} - \mathbf{1}\left(\boldsymbol{\mu}_p^{(l)}\right)^\top, \qquad \widehat{\mathbf{q}}_p^{(l)} = \mathbf{q}_p^{(l)} - \boldsymbol{\mu}_p^{(l)}. \tag{11}$$

To model local normal variation efficiently, we decompose the compact Gram matrix $\mathbf{G}_p^{(l)} = \widehat{\mathbf{Z}}_p^{(l)}\left(\widehat{\mathbf{Z}}_p^{(l)}\right)^\top$. Let $\mathbf{V}_{p,r}^{(l)}$ and $\mathbf{\Lambda}_{p,r}^{(l)}$ denote the eigenvectors and eigenvalues associated with the largest $r = \min(r_s, K-1)$ directions, where r_s is the maximum subspace rank. We construct ridge-scaled local normal directions as

$$\mathbf{U}_p^{(l)} = \left(\widehat{\mathbf{Z}}_p^{(l)}\right)^\top \mathbf{V}_{p,r}^{(l)}\left(\mathbf{\Lambda}_{p,r}^{(l)} + \epsilon\mathbf{I}\right)^{-\frac{1}{2}}, \tag{12}$$

where $\epsilon = 0.05$ is added to the eigenvalues to stabilize the decomposition and suppress weak local directions. The corresponding anomaly evidence is

$$d_{p,\mathrm{sub}}^{(l)} = \left\| \widehat{\mathbf{q}}_p^{(l)} - \mathbf{U}_p^{(l)}\left(\mathbf{U}_p^{(l)}\right)^\top \widehat{\mathbf{q}}_p^{(l)} \right\|_2. \tag{13}$$

We further measure the dispersion of the local normal support as

$$\sigma_p^{(l)} = \frac{1}{K}\sum_{k=1}^{K} \left\| \mathbf{z}_{k,p}^{(l)} - \boldsymbol{\mu}_p^{(l)} \right\|_2. \tag{14}$$

For $K = 1$, $r = 0$ and $\sigma_p^{(l)} = 0$, while the subspace residual reduces to $\|\mathbf{q}_p^{(l)} - \mathbf{z}_{1,p}^{(l)}\|_2$. A large $\sigma_p^{(l)}$ indicates that the retrieved normal observations are less consistent at this location and is therefore used to discount unreliable anomaly evidence.

Layer-Wise Scoring and Multi-Layer Fusion. For a patch token at position p from scoring layer l, nearest-neighbor and subspace evidence are combined as

$$A_p^{(l)} = \frac{(1-\beta)\widetilde{d}_{p,\mathrm{NN}}^{(l)} + \beta d_{p,\mathrm{sub}}^{(l)}}{1+\sigma_p^{(l)}}, \tag{15}$$

where β is used to balance the two anomaly cues. The global correction is applied only to the NN distance, while the subspace residual retains spatially constrained correspondences to preserve local structural information. The final patch anomaly score is obtained by averaging the scoring layers:

$$A_p = \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}} A_p^{(l)}. \tag{16}$$

Then, the scores of all query patches are rearranged according to their spatial positions to form the patch-level anomaly map $\mathbf{A}$.

3.4 Anomaly Detection and Localization

Image-Level. For image-level anomaly detection, we average the Top-m patch anomaly scores $S_{\mathrm{patch}} = \frac{1}{m}\sum_{i=1}^{m} A_{\mathrm{top},i}$. For the K retrieved references, we compute a global CLS anomaly score $S_{\mathrm{CLS}} = \frac{1}{K}\sum_{k=1}^{K}\left(1 - s_k^{\mathrm{cls}}\right)$. The final image-level anomaly score is

$$S_{\mathrm{img}} = S_{\mathrm{patch}} + S_{\mathrm{CLS}}. \tag{17}$$

Pixel-Level. For pixel-level anomaly localization, the fused patch scores are first arranged into the anomaly map $\mathbf{A}$. To reduce locally depressed responses inside anomalous regions, we compute a local average map $\overline{\mathbf{A}} = \mathrm{AvgPool}_{3\times3}(\mathbf{A})$ and apply a one-sided refinement $\mathbf{A}' = \mathbf{A} + \mathrm{ReLU}\left(\overline{\mathbf{A}} - \mathbf{A}\right)$. Finally, the refined patch map is bilinearly upsampled to the input resolution and smoothed with a local average filter:

$$\mathbf{A}_{\mathrm{pixel}} = \mathrm{AvgPool}_{3\times3}\left(\mathrm{Up}\left(\mathbf{A}'\right)\right). \tag{18}$$

4 Experiments

4.1 Experimental Setup

Datasets We conduct experiments on two widely used IAD benchmarks: MVTec-AD (Bergmann et al., 2019) and VisA (Zou et al., 2022). Dataset details are reported in Appendix A.2.1.

Evaluation Following standard protocols in IAD, we report Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPR) for image-level detection, and AUROC and Per-Region Overlap (PRO) for pixel-level localization.

Baselines We evaluate our method against state-of-the-art anomaly detection methods across multi-class and single-class settings. For multi-class scenarios, we compare with OmiAD (Feng et al., 2025), DecAD (Wang et al., 2025), Dinomaly (Guo et al., 2025), INP-Former (Luo et al., 2025), RAID (Cai et al., 2026), AnomalyDINO (Damm et al., 2025) and RAD (Zhang et al., 2026). For single-class scenarios, we compare with MambaAD (He et al., 2024), INP-Former, Dinomaly, PatchCore (Roth et al., 2022), AnomalyDINO, ANoCo (Seo et al., 2026) and SubspaceAD (Lendering et al., 2026). Note that all normal training images are available in our experimental setting. Therefore, for a fair comparison, we report the performance of competing methods under the full-shot setting whenever applicable.

Implementation Details Implementation details are reported in Appendix A.2.2.

4.2 Main Results

Tables 1 and 2 report the comparison with state-of-the-art methods under the multi-class and single-class settings, respectively. Here, K-Ref denotes using the K normal query-relevant references retrieved from all normal samples for the computation of the anomaly score.

Comparison with Multi-Class Baselines As shown in Table 1, our method achieves strong performance on both MVTec-AD and VisA. On MVTec-AD, Ours (16-Ref) obtains 99.4% I-AUROC and 99.8% I-AUPR, while achieving the best P-AUROC of 98.6% and PRO of 97.0% among training-free methods. In particular, its PRO also exceeds that of all compared training-based methods. On VisA, Ours (16-Ref) achieves 95.9% I-AUROC and 95.8% I-AUPR, together with the best P-AUROC of 99.2% and PRO of 96.0% among training-free methods; the two pixel-level metrics also surpass all training-based baselines. Notably, without access to category labels, all retrieved references belong to the same category as the query, yielding 100% category-level retrieval precision in the multi-class setting.

Comparison with Single-Class Baselines For the single-class setting, we additionally evaluate Ours (Full-shot), where all available normal samples of each class are used for anomaly score computation. As shown in Table 2, on MVTec-AD, Ours (16-Ref) achieves 99.4% I-AUROC and matches the best I-AUPR of 99.8%, while attaining the best PRO of 97.0% among training-free methods. Ours (Full-shot) further achieves 98.7% P-AUROC and the same 97.0% PRO. On VisA, Ours (Full-shot) obtains 97.9% I-AUROC and 98.2% I-AUPR, together with the best P-AUROC of 99.3% and the second-best PRO of 96.7% among training-free methods. Its pixel-level performance also remains strong compared with the training-based baselines.

Table 1: Comparison with multi-class anomaly detection methods on MVTec-AD and VisA. Best and second-best results among training-free methods are highlighted.

Method	Venue	Training-Free	MVTec-AD					VisA				
			I-AUROC	I-AUPR	P-AUROC	PRO	Mean	I-AUROC	I-AUPR	P-AUROC	PRO	Mean
OmiAD	ICML 2025	✗	98.8	99.7	97.7	93.2	97.35	95.3	96.0	98.9	89.2	94.85
DecAD	ICCV 2025	✗	99.0	99.7	97.8	92.5	97.25	94.7	95.5	98.5	91.7	95.10
Dinomaly	CVPR 2025	✗	99.6	99.8	98.4	94.8	98.15	98.7	98.9	98.7	94.5	97.70
INP-Former	CVPR 2025	✗	99.7	99.9	98.5	94.9	98.25	98.9	99.0	98.9	94.4	97.80
RAID	CVPR 2026	✗	99.4	99.8	98.6	95.6	98.35	94.9	95.5	99.0	91.7	95.28
AnomalyDINO	WACV 2025	✓	96.8	98.6	98.1	93.6	96.78	90.5	91.4	97.5	86.3	91.43
RAD	ICML 2026	✓	**99.6**	**99.8**	98.5	94.9	98.20	**97.5**	**97.3**	99.1	94.6	**97.13**
Ours (4-Ref)	/	✓	99.2	99.7	98.4	96.8	98.53	94.7	94.5	99.0	95.5	95.95
Ours (8-Ref)	/	✓	99.3	99.7	98.5	**97.0**	98.63	95.3	95.0	99.1	95.9	96.33
Ours (16-Ref)	/	✓	99.4	**99.8**	**98.6**	**97.0**	**98.70**	95.9	95.8	**99.2**	**96.0**	96.73

Table 2: Comparison with single-class anomaly detection methods on MVTec-AD and VisA. Best and second-best results among training-free methods are highlighted.

Method	Venue	Training-Free	MVTec-AD					VisA				
			I-AUROC	I-AUPR	P-AUROC	PRO	Mean	I-AUROC	I-AUPR	P-AUROC	PRO	Mean
MambaAD	NeurIPS 2024	✗	98.7	99.6	97.8	94.0	97.53	95.3	95.3	98.8	91.9	95.33
INP-Former	CVPR 2025	✗	99.7	99.9	98.3	95.4	98.33	98.5	98.6	98.4	93.8	97.33
Dinomaly	CVPR 2025	✗	99.7	/	99.9	95.0	98.20	98.9	/	98.9	95.1	97.63
PatchCore	CVPR 2022	✓	99.2	**99.8**	**99.4**	94.3	98.18	95.1	96.2	98.8	91.2	95.33
AnomalyDINO	WACV 2025	✓	**99.5**	**99.8**	98.2	95.0	98.13	97.6	98.0	98.8	96.1	97.63
ANoCo	CVPR 2026	✓	99.4	99.7	98.2	96.4	98.43	**98.7**	**98.9**	99.2	95.3	98.03
SubspaceAD	CVPR 2026	✓	99.2	99.6	98.2	95.6	98.15	98.2	98.3	99.1	**96.9**	**98.13**
Ours (8-Ref)	/	✓	99.3	99.7	98.5	**97.0**	98.63	95.3	95.0	99.1	95.9	96.33
Ours (16-Ref)	/	✓	99.4	**99.8**	98.6	**97.0**	**98.70**	95.9	95.8	99.2	96.0	96.73
Ours (Full-shot)	/	✓	99.3	99.7	98.7	**97.0**	98.68	97.9	98.2	**99.3**	96.7	98.03

Overall, our method is entirely training-free and relies only on a foundation visual backbone, while still achieving strong performance in both multi-class and single-class settings. Quantitative class-wise results are reported in Appendix B.1.

Qualitative Results As shown in Figure 3, the retrieved normal references exhibit strong visual compatibility with their corresponding query images. The resulting anomaly maps provide clear responses around defective regions across the illustrated cases, including complex structures, small defects, and irregular anomaly patterns. More qualitative results are shown in Appendix B.2.

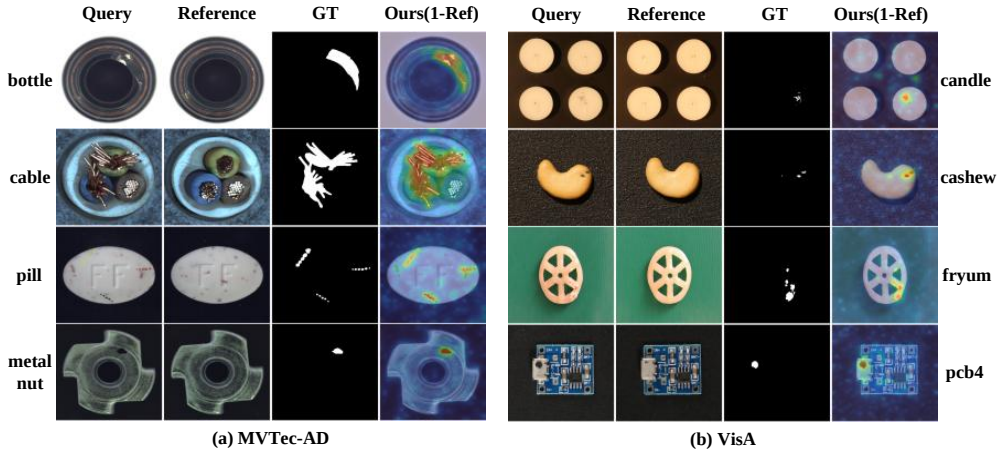


Figure 3: Qualitative results of our method under the 1-Ref setting on MVTec-AD and VisA.

Inference Efficiency Inference efficiency details are reported in Appendix A.2.3.

4.3 Generalization to Existing Methods

To investigate the importance of query-relevant normal samples and the generalizability of our retrieval strategy, we apply it to two existing few-shot methods, AnomalyDINO and SubspaceAD. We first evaluate all methods under the standard K-shot setting using five random seeds, and then replace the randomly sampled references with the K query-relevant samples retrieved by our strategy, while keeping each method's original anomaly scoring procedure unchanged. We compare the corresponding K-shot and K-Ref settings for $K \in \{1, 2, 4\}$. For clarity, Table 3 reports only the mean K-shot performance, while the complete results are provided in Appendix A.2.4. All experiments are conducted under the single-class setting, using DINOv3-B/16 for our method and SubspaceAD, and DINOv2-B/14 for AnomalyDINO. We further extend this evaluation to the real-world 3CAD dataset (Yang et al., 2025), with detailed results provided in Appendix A.2.4.

Table 3: Generalization of our query-adaptive reference retrieval strategy to existing methods on MVTec-AD and VisA. Higher results within each pair are shown in bold.

Method	Mode		MVTec-AD				VisA			
	K	Type	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
Ours	1	Shot	92.7	96.2	97.2	93.9	90.3	91.1	98.0	91.9
		Ref	**98.6**	**99.3**	**98.1**	**96.2**	**93.8**	**94.3**	**98.6**	**93.5**
	2	Shot	94.1	96.9	97.7	94.7	91.3	91.6	98.4	92.8
		Ref	**98.9**	**99.4**	**98.3**	**96.6**	**94.0**	**94.0**	**98.9**	**94.8**
	4	Shot	95.9	97.8	98.1	95.6	92.9	92.8	98.7	94.0
		Ref	**99.2**	**99.7**	**98.4**	**96.8**	**94.7**	**94.5**	**99.0**	**95.5**
SubspaceAD	1	Shot	94.9	97.5	97.1	93.6	89.7	89.7	97.6	93.0
		Ref	**97.6**	**98.6**	**97.4**	**94.4**	**91.4**	**90.8**	**98.5**	**94.5**
	2	Shot	95.4	97.8	97.3	93.9	91.0	90.6	97.7	93.4
		Ref	**97.9**	**98.8**	**97.6**	**94.6**	**92.6**	**91.6**	**98.7**	**95.0**
	4	Shot	96.5	98.4	97.4	94.2	92.3	91.8	97.8	93.7
		Ref	**98.2**	**98.9**	**97.7**	**94.8**	**93.7**	**92.4**	**98.8**	**95.4**
AnomalyDINO	1	Shot	95.7	97.3	95.8	91.7	84.3	83.5	96.0	85.6
		Ref	**97.3**	**98.0**	**96.3**	**92.7**	**85.7**	**85.6**	**96.7**	**87.7**
	2	Shot	96.1	97.6	96.0	91.9	85.4	84.7	96.5	86.8
		Ref	**98.1**	**98.5**	**96.7**	**93.1**	**87.4**	**86.6**	**97.3**	**89.0**
	4	Shot	97.1	98.1	96.1	92.1	88.2	87.6	97.0	88.4
		Ref	**98.4**	**98.6**	**96.9**	**93.4**	**89.7**	**88.4**	**97.6**	**90.1**

Results Analysis As shown in Table 3, our method already performs strongly under the conventional few-shot setting. More importantly, replacing random K-shot references with query-relevant references consistently improves all evaluated methods across both datasets, all reference budgets, and all metrics. For example, under 1-Ref, our I-AUROC/PRO improves from 92.7%/93.9% to 98.6%/96.2% on MVTec-AD and from 90.3%/91.9% to 93.8%/93.5% on VisA. Similar gains are observed for SubspaceAD and AnomalyDINO. Notably, 1-Ref often matches or outperforms 2-shot, and in several cases even 4-shot, despite using fewer references in modeling anomalies, suggesting that reference relevance can be more important than simply increasing the number of randomly sampled normal images. These consistent improvements, together with the gains obtained on SubspaceAD and AnomalyDINO without modifying their scoring procedures, support the effectiveness and generalizability of our query-adaptive reference retrieval strategy.

4.4 Ablation Study

We conduct ablation studies on the key components of our method, including the reference ranking strategy, anomaly scoring, and visual backbone. All experiments are performed on the MVTec-AD and VisA datasets. Additional ablation studies are provided in Appendix A.2.6.

4.4.1 Reference Ranking Strategy

We study the contribution of global CLS similarity and coarse patch similarity to reference preselection under the 1-Ref setting. As shown in Figure 4, the performance remains relatively stable over a broad range of fusion weights, while relying solely on CLS similarity generally leads to inferior pixel-level localization and weaker performance on VisA. In contrast, patch-dominant or mixed

configurations consistently maintain strong performance, suggesting that fine-grained patch-level compatibility is useful for reranking candidates, while global CLS similarity provides complementary semantic information.

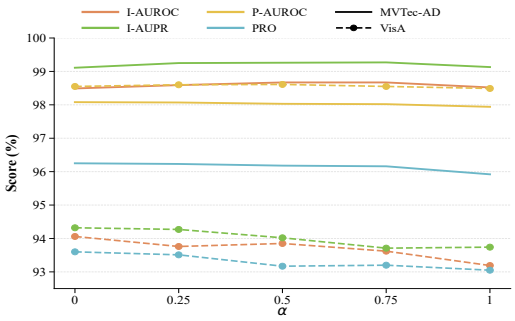


Figure 4: Effect of the reference ranking coefficient α under the 1-Ref setting.

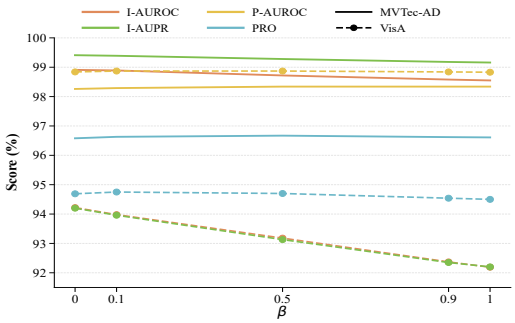


Figure 5: Effect of the anomaly score fusion coefficient β under the 2-Ref setting.

4.4.2 Complementary Anomaly Scoring

We further examine the fusion between nearest-neighbor and subspace anomaly evidence under the 2-Ref setting. Figure 5 shows that nearest-neighbor evidence plays the dominant role, particularly for image-level discrimination, whereas introducing subspace evidence preserves highly competitive performance and slightly benefits some pixel-level metrics on MVTec-AD. Mixed configurations retain stable localization performance while avoiding the image-level degradation observed when the subspace cue dominates, supporting its role as complementary evidence.

4.4.3 Comparison of Different Visual Backbones

We compare different visual backbones of comparable model scale under the 1-Ref setting. We choose DINO (Caron et al., 2021), DINOv2 (Oquab et al., 2024), DINOv3, SigLIP (Zhai et al., 2023) and SigLIP2 (Tschannen et al., 2025) for this study. As shown in Table 4, DINOv3 achieves the strongest overall performance among backbones of comparable scale, with the best results on six metrics across MVTec-AD and VisA. More detailed results are reported in Table 12.

Table 4: Comparison of different backbones under the 1-Ref setting on MVTec-AD and VisA.

Backbone	MVTec-AD				VisA			
	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
DINOv3 B/16	**98.6**	99.3	**98.1**	**96.2**	**93.8**	**94.3**	**98.6**	93.5
DINOv2 B/14	**98.6**	**99.5**	98.0	95.7	90.9	91.7	98.2	**93.8**
DINO B/16	96.2	98.0	97.5	93.3	86.3	86.8	97.8	85.9
SigLIP B/16	97.3	98.4	96.0	92.0	89.8	90.3	97.6	90.8
SigLIP2 B/16	97.0	97.8	95.9	91.8	89.6	90.3	96.8	87.0

5 Conclusion

In this work, we revisit industrial anomaly detection from the perspective of abundant normal data, highlighting the importance of reference relevance beyond the sample quantity. We propose a training-free retrieve-then-score framework that selects query-relevant references from a large normal data pool and combines nearest-neighbor and query-conditioned subspace evidence for anomaly inference. Our method achieves competitive performance in both multi-class and single-class settings, particularly in pixel-level anomaly localization. Furthermore, applying our retrieval strategy to existing few-shot methods consistently improves their performance, demonstrating the importance of query-relevant normal references. We hope this work encourages further study of how abundant normal data should be effectively utilized.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the manuscript. We did not use generative AI tools for any tasks requiring disclosure. All AI-assisted work was carefully reviewed and revised by the authors. We take responsibility for the final content of this work.

REFERENCES

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

Mingxiu Cai, Zhe Zhang, Gaochang Wu, Tianyou Chai, and Xiatian Zhu. RAID: Retrieval-augmented anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21367–21378, June 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. AnomalyDINO: Boosting patch-based few-shot anomaly detection with DINOv2. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 1319–1329, February 2025.

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 (CVPR)*, pp. 9737–9746, June 2022.

Yaoxuan Feng, Wenchao Chen, Yuxin Li, Bo Chen, Yubiao Wang, Zixuan Zhao, Hongwei Liu, and Mingyuan Zhou. OmiAD: One-step adaptive masked diffusion model for multi-class anomaly detection via adversarial distillation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=859NdHQv0Z.

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20405–20415, June 2025.

Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. MambaAD: Exploring state space models for multi-class unsupervised anomaly detection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 71162–71187. Curran Associates, Inc., 2024. doi: 10.52202/079017-2274. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/833b21da1956c6b92f6df253bf655cf5-Paper-Conference.pdf.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19606–19616, Los Alamitos, CA, USA, June 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01878. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01878.

Camile Lendering, Erkut Akdag, and Egor Bondarau. SubspaceAD: Training-free few-shot anomaly detection via subspace modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28557–28566, June 2026.

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. PromptAD: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16838–16848, June 2024.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. SimpleNet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20402–20411, June 2023.

Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9974–9983, June 2025.

Wenxi Lv, Qinliang Su, and Wenchao Xu. One-for-all few-shot anomaly detection via instance-induced prompt learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Zzs3JwknAY.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

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 (CVPR)*, pp. 14318–14328, June 2022.

Jungwook Seo, Minjeong Kim, Younkwan Lee, Seungho Shin, and Sungyong Baik. Anomaly as non-conformity via training-free graph Laplacian energy minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21336–21345, June 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Fenfang Tao, Guo-Sen Xie, Fang Zhao, and Xiangbo Shu. Kernel-aware graph prompt learning for few-shot anomaly detection, 2025. URL https://arxiv.org/abs/2412.17619.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Xiaolei Wang, Xiaoyang Wang, Huihui Bai, Eng Gee Lim, and Jimin Xiao. DecAD: Decoupling anomalies in latent space for multi-class unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21568–21577, October 2025.

Enquan Yang, Peng Xing, Hanyang Sun, Wenbo Guo, Yuanwei Ma, Zechao Li, and Dan Zeng. 3CAD: A large-scale real-world 3C product dataset for unsupervised anomaly detection. 2025.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 4571–4584. Curran Associates, Inc., 2022. doi: 10.52202/068431-0330. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/1d774c112926348c3e25ea47d87c835b-Paper-Conference.pdf.

Guangyao Zhai, Yue Zhou, Xinyan Deng, Lars Heckler-Kram, Nassir Navab, and Benjamin Busam. Foundation visual encoders are secretly few-shot anomaly detectors. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YRrlJ8oVEH.

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

Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16699–16708, June 2024.

Xingwu Zhang, Guanxuan Li, Paul Henderson, Gerardo Aragon-Camarasa, and ZIJUN LONG. Is task-specific training necessary for anomaly detection? In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=0Qudez0gLp.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 392–408, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20056-4.

A Appendix

A.1 Analysis of Query-Adaptive Reference Retrieval

We further analyze the motivation for the two-stage retrieval strategy through the t-SNE visualization of CLS representations in Figure 6. Two observations emerge. First, normal samples within the same category are not equally informative for a given query. For example, in the bottle category, one anomalous sample exhibits noticeably different distances to different normal samples, motivating further investigation of query-adaptive reference selection. Second, the normal–anomaly separability in CLS space varies across categories: categories such as leather and carpet show clearer separation, while in categories such as pill, the two distributions can be much closer. These observations suggest that CLS similarity alone is insufficient to consistently characterize reference compatibility. We therefore use CLS features for global candidate retrieval, followed by coarse patch reranking to incorporate local structural correspondence.

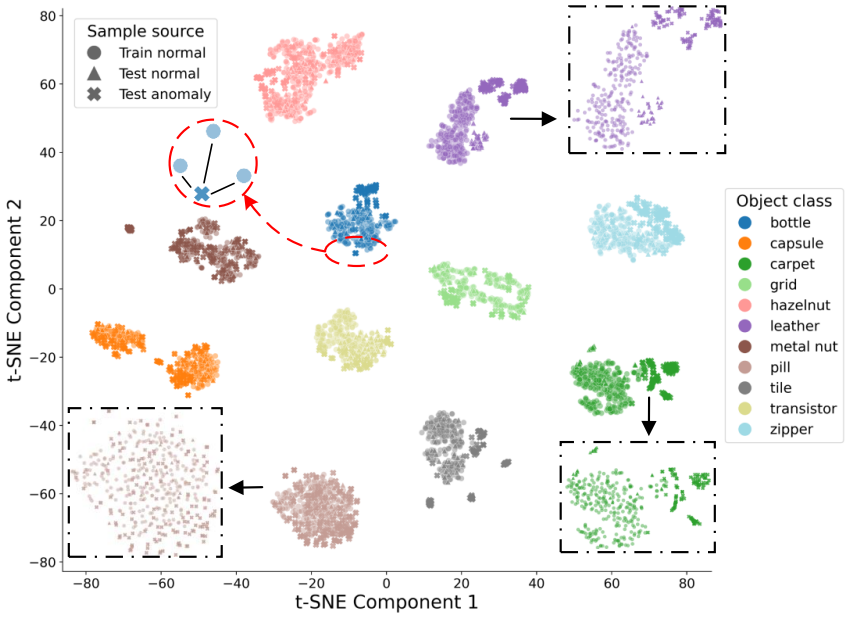


Figure 6: t-SNE visualization of CLS tokens from a subset of categories in the MVTec-AD dataset.

A.2 Experimental Details

A.2.1 Dataset Details

MVTec-AD contains 15 categories with 3,629 normal training images and 1,725 test images. VisA contains 12 object categories with 8,659 normal training images and 2,162 test images. Both datasets provide pixel-level masks for all anomalous samples. These two benchmarks cover various types of defects and visual complexities, providing a comprehensive evaluation of image-level and pixel-level anomaly detection.

A.2.2 Implementation Details

All experiments are conducted on a single RTX 3090 GPU. We use the DINOv3-B/16 (Siméoni et al., 2025) as our visual encoder. The input images are resized to 512×512. The CLS token from the last layer and the patch tokens from the $9th$ layer are used for query-adaptive reference retrieval. The number of candidate images, T, is set to 32. The reranking grid size G is set to 8. The coarse patch reranking coefficient α is 0.25. The neighborhood radius d is set to 5. The scoring coefficient β is 0.1. We use layers $[7, 8, 9, 10]$ to compute the anomaly score. r_s is set to 4 and δ is set to 0.05. Following common practice in recent IAD methods, we use dataset-specific Top-m aggregation for the image-level anomaly score, with $m = 12$ for MVTec-AD and $m = 6$ for VisA.

A.2.3 Inference Efficiency

Let N denote the number of available normal images and $K \ll N$ the number of references used for anomaly scoring. During inference, memory-based full-shot methods exploit all available normal samples, but their reference-dependent anomaly scoring typically scales with N, resulting in relatively high inference cost. Conventional K-shot methods reduce this cost by using only K normal samples, but the remaining normal data are excluded from inference. In contrast, our method first retrieves K query-relevant references from the pre-built feature bank and performs the anomaly scoring only on these selected samples, providing a practical trade-off between normal-data coverage and inference efficiency.

We further evaluate the computational efficiency of our method on MVTec-AD. With a batch size of 16, constructing the normal feature bank from 3,629 normal images takes approximately 192.5 s. During inference, the average latency is 79.76, 144.10, 224.31, 355.19 and 716.01 ms per image for the 1-, 2-, 4-, 8-, and 16-Ref settings, respectively. As expected, the inference cost increases with the number of selected references, since anomaly scoring is performed only on the retrieved K references. Therefore, our method can exploit the entire normal pool for reference retrieval while keeping the expensive scoring stage restricted to a small subset, providing a practical trade-off between normal-data coverage and inference efficiency.

A.2.4 Details of Generalization Experiment

The complete comparison between K-shot and K-Ref on MVTec-AD and VisA, including the mean and standard deviation over five random seeds, is reported in Table 5. Notably, the K-shot results exhibit standard deviations exceeding 1% in some cases, indicating that the selection of normal reference samples can have a non-negligible impact on inference performance. This further motivates the need for selecting query-relevant references rather than relying on random sampling.

Generalization to Real-World Industrial Data. To further evaluate the importance of query-relevant normal samples and generalizability of our retrieval strategy in practical industrial scenarios, we perform experiments on 3CAD (Yang et al., 2025), an industrial anomaly detection dataset collected from actual manufacturing processes. The dataset comprises 8 object categories, 47 anomaly types, 10,493 normal training images, and 16,546 test images. We use Top-12 aggregation for the image-level anomaly score on 3CAD.

The experimental results are reported in Table 6. All three methods exhibit relatively weak performance under the conventional few-shot setting. Substantial variations in object pose, viewpoint, and appearance within 3CAD may cause randomly sampled normal references to be poorly matched to the query images. Moreover, the standard deviations exceed 1 percentage point in several cases, in-

Table 5: Complete comparison between K-shot and K-Ref settings on MVTec-AD and VisA. The K-shot results are reported as mean $\pm$ standard deviation over five random seeds. Higher results within each pair are shown in bold.

Method	Mode		MVTec-AD				VisA			
	K	Type	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
Ours	1	Shot	92.7 ± 0.5	96.2 ± 0.3	97.2 ± 0.1	93.9 ± 0.1	90.3 ± 0.3	91.1 ± 0.4	98.0 ± 0.0	91.9 ± 0.3
		Ref	**98.6**	**99.3**	**98.1**	**96.2**	**93.8**	**94.3**	**98.6**	**93.5**
	2	Shot	94.1 ± 0.4	96.9 ± 0.2	97.7 ± 0.0	94.7 ± 0.2	91.3 ± 0.1	91.6 ± 0.3	98.4 ± 0.1	92.8 ± 0.2
		Ref	**98.9**	**99.4**	**98.3**	**96.6**	**94.0**	**94.0**	**98.9**	**94.8**
	4	Shot	95.9 ± 0.5	97.8 ± 0.3	98.1 ± 0.0	95.6 ± 0.1	92.9 ± 0.3	92.8 ± 0.3	98.7 ± 0.0	94.0 ± 0.2
		Ref	**99.2**	**99.7**	**98.4**	**96.8**	**94.7**	**94.5**	**99.0**	**95.5**
SubspaceAD	1	Shot	94.9 ± 1.4	97.5 ± 0.7	97.1 ± 0.1	93.6 ± 0.1	89.7 ± 1.4	89.7 ± 1.2	97.6 ± 0.1	93.0 ± 0.2
		Ref	**97.6**	**98.6**	**97.4**	**94.4**	**91.4**	**90.8**	**98.5**	**94.5**
	2	Shot	95.4 ± 1.3	97.8 ± 0.6	97.3 ± 0.1	93.9 ± 0.1	91.0 ± 0.5	90.6 ± 0.5	97.7 ± 0.0	93.4 ± 0.2
		Ref	**97.9**	**98.8**	**97.6**	**94.6**	**92.6**	**91.6**	**98.7**	**95.0**
	4	Shot	96.5 ± 1.0	98.4 ± 0.4	97.4 ± 0.1	94.2 ± 0.2	92.3 ± 0.7	91.8 ± 0.8	97.8 ± 0.0	93.7 ± 0.1
		Ref	**98.2**	**98.9**	**97.7**	**94.8**	**93.7**	**92.4**	**98.8**	**95.4**
AnomalyDINO	1	Shot	95.7 ± 0.9	97.3 ± 0.6	95.8 ± 0.2	91.7 ± 0.2	84.3 ± 0.4	83.5 ± 1.0	96.0 ± 0.1	85.6 ± 0.4
		Ref	**97.3**	**98.0**	**96.3**	**92.7**	**85.7**	**85.6**	**96.7**	**87.7**
	2	Shot	96.1 ± 0.9	97.6 ± 0.5	96.0 ± 0.2	91.9 ± 0.2	85.4 ± 0.6	84.7 ± 0.7	96.5 ± 0.1	86.8 ± 0.2
		Ref	**98.1**	**98.5**	**96.7**	**93.1**	**87.4**	**86.6**	**97.3**	**89.0**
	4	Shot	97.1 ± 0.1	98.1 ± 0.3	96.1 ± 0.2	92.1 ± 0.2	88.2 ± 0.6	87.6 ± 1.1	97.0 ± 0.0	88.4 ± 0.3
		Ref	**98.4**	**98.6**	**96.9**	**93.4**	**89.7**	**88.4**	**97.6**	**90.1**

Table 6: Comparison between K-shot and K-Ref settings on 3CAD. The K-shot results are reported as mean $\pm$ standard deviation over five random seeds. Higher results within each pair are shown in bold.

Method	Mode		3CAD			
	K	Type	I-AUROC	I-AUPR	P-AUROC	PRO
Ours	1	Shot	61.3 ± 1.3	75.6 ± 0.8	86.1 ± 0.9	69.1 ± 1.5
		Ref	**84.3**	**89.8**	**94.2**	**85.6**
	2	Shot	62.8 ± 0.8	76.1 ± 1.2	87.4 ± 0.6	69.0 ± 1.0
		Ref	**84.0**	**89.7**	**94.6**	**86.0**
	4	Shot	63.6 ± 1.4	76.7 ± 0.9	88.4 ± 0.8	70.9 ± 1.6
		Ref	**83.8**	**89.6**	**94.8**	**86.2**
SubspaceAD	1	Shot	64.9 ± 1.1	77.0 ± 0.7	87.5 ± 0.7	71.9 ± 1.1
		Ref	**80.5**	**87.3**	**94.4**	**86.1**
	2	Shot	66.5 ± 1.0	78.2 ± 1.1	89.0 ± 0.3	74.2 ± 1.2
		Ref	**80.8**	**87.6**	**94.6**	**86.8**
	4	Shot	68.3 ± 1.5	79.7 ± 1.2	90.1 ± 0.4	76.6 ± 1.3
		Ref	**80.6**	**87.8**	**94.6**	**87.0**
AnomalyDINO	1	Shot	61.8 ± 0.9	75.6 ± 1.4	79.0 ± 1.3	40.0 ± 2.7
		Ref	**83.6**	**89.4**	**93.9**	**79.6**
	2	Shot	64.8 ± 1.1	77.9 ± 0.7	83.0 ± 0.8	47.4 ± 2.4
		Ref	**84.3**	**89.7**	**94.1**	**79.8**
	4	Shot	66.7 ± 0.8	78.8 ± 0.8	84.9 ± 1.5	51.0 ± 2.5
		Ref	**84.8**	**90.0**	**94.3**	**79.7**

dicating that reference selection has a non-negligible impact on detection performance. By replacing random references with query-relevant samples, our retrieval strategy consistently improves all four metrics across the three methods and all reference budgets. For instance, under 1-Ref, our I-AUROC and PRO increase from 61.3% to 84.3% and from 69.1% to 85.6%, respectively. Similar gains are observed for SubspaceAD and AnomalyDINO. Notably, 1-Ref even outperforms 4-shot across all four metrics for all three methods, suggesting that reference relevance can be more important than simply increasing the number of randomly sampled normal images. These results further highlight the importance of reference relevance and support the generalizability of our retrieval strategy to real-world industrial data.

A.2.5 Generalization to Complex Industrial Scenarios

Robustness to a Larger Multi-class Normal Data Pool To better approximate practical industrial scenarios with a large number of normal samples and object categories, we merge all normal training images from MVTec-AD and VisA into a shared reference pool and perform retrieval directly from this mixed pool. This shared pool contains 12,288 normal images that span 27 object categories. As shown in Table 7, the resulting performance remains comparable to that obtained when constructing a separate normal pool for each dataset. Notably, without access to category labels during retrieval, all selected references still belong to the same category as their corresponding query images, yielding 100% category-level retrieval precision. These results suggest that our retrieval strategy remains robust to a substantially enlarged and more diverse normal reference pool, supporting its applicability and generalization to more complex industrial settings.

We also evaluate the inference efficiency under this setting. With a batch size of 16, constructing the feature bank for 12,288 normal images takes approximately 481 s. The average inference latency is 78.63, 124.04, 201.47, 361.00, and 713.28 ms per image for the 1-, 2-, 4-, 8- and 16-Ref settings, respectively. Despite the substantially enlarged normal pool, the online inference latency remains primarily governed by the number of retrieved references, since only the selected K samples participate in the subsequent anomaly scoring. This result further supports the scalability of our retrieval-based design to larger and more diverse normal reference pools.

Table 7: Robustness evaluation under a large mixed normal reference pool. All normal training images from MVTec-AD and VisA are merged into a shared pool for reference retrieval.

K-Ref	MVTec-AD				VisA			
	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
1	98.6	99.3	98.1	96.2	93.8	94.3	98.6	93.5
2	98.9	99.4	98.3	96.6	94.0	93.8	98.9	94.8
4	99.2	99.7	98.4	96.8	94.7	94.5	99.0	95.5

Evaluation on a Large-Scale Real-World Industrial Dataset To further evaluate the robustness of our method in realistic industrial settings, we conduct multi-class experiments on 3CAD, where all 8 object categories share a unified normal reference pool. As shown in Table 8, our method maintains performance close to its single-class counterpart, with slight declines in most metrics. The category-level retrieval precision reaches 97.1%, 97.0%, and 96.8% under the 1-, 2-, and 4-Ref settings, respectively. The substantial variations in object pose, shape, and appearance may complicate the retrieval of category-consistent and visually compatible references, potentially contributing to the observed performance degradation. These results demonstrate the potential of our method in realistic multi-class industrial scenarios while revealing limitations in reference retrieval, which we aim to address in future work. We further evaluate the computational efficiency of our method on 3CAD. With a batch size of 16, constructing the feature bank from 10,493 normal images takes approximately 414 s, representing a one-time offline cost. The average inference latency is 75.3, 128.3, 217.5, 398.8, and 803.8 ms per image under the 1-, 2-, 4-, 8-, and 16-Ref settings, respectively.

Table 8: Robustness evaluation on the real-world 3CAD dataset under the multi-class setting.

K-Ref	3CAD				Retrieval Precision
	I-AUROC	I-AUPR	P-AUROC	PRO	
1	84.0	89.7	94.0	85.3	97.1
2	83.6	89.6	94.3	85.4	97.0
4	83.5	89.6	94.6	85.9	96.8

A.2.6 Additional ablation study

Number of Reference Candidates As shown in Table 9, the performance remains relatively stable across different candidate-set sizes, while increasing the number of pre-selected references beyond 32 does not yield consistent improvements. In particular, several metrics decrease at $T = 48$ and

$T = 64$, whereas a larger candidate set also incurs additional retrieval time. We therefore use $T = 32$ as a practical balance between retrieval coverage, performance, and efficiency.

Table 9: Ablation on the number of retrieved reference candidates under the 1-Ref setting.

T	MVTec-AD				VisA			
	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
8	98.55	99.21	98.05	96.21	93.84	94.24	98.58	93.25
16	**98.59**	99.21	**98.07**	**96.24**	93.80	94.19	98.58	93.36
32	**98.59**	**99.25**	**98.07**	96.23	93.76	94.27	**98.60**	**93.51**
48	98.56	99.22	98.05	96.22	**93.89**	**94.36**	98.59	93.42
64	98.53	99.20	98.05	96.23	93.82	94.30	98.56	93.31

Neighborhood Radius We study the spatial neighborhood radius d used for local patch matching under different reference budgets. As shown in Table 10, the performance is generally stable over $d \in [1, 5]$, indicating that the method is not highly sensitive to the exact neighborhood size. A small radius performs competitively in the 1-Ref setting, whereas larger neighborhoods become more favorable as the reference budget increases; in particular, $d = 5$ achieves the strongest overall results under 2-Ref and 4-Ref. This suggests that a moderate spatial search range provides useful tolerance to local displacement while preserving sufficiently constrained normal correspondences.

Table 10: Ablation on the local neighborhood radius under different reference budgets on MVTec-AD and VisA. The best results under each reference budget are highlighted in bold.

Radius d	K-Ref	MVTec-AD				VisA			
		I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
1	1	**98.63**	**99.31**	**98.08**	**96.25**	**94.02**	**94.41**	**98.64**	93.31
	2	98.80	99.37	98.26	96.58	93.61	93.75	98.80	94.17
	4	99.20	99.66	98.40	96.79	94.13	94.03	99.00	95.08
2	1	**98.63**	99.30	98.07	96.24	93.88	94.36	98.61	93.32
	2	98.85	99.37	98.27	96.60	93.56	93.67	98.80	94.45
	4	99.22	99.66	98.42	96.81	94.13	93.98	98.99	95.26
3	1	98.60	99.26	98.07	96.24	93.78	94.26	98.60	93.40
	2	**98.89**	**99.40**	**98.29**	**96.63**	93.79	93.85	98.83	94.59
	4	**99.24**	99.66	**98.44**	96.83	94.43	94.25	99.00	95.38
4	1	98.59	99.26	98.07	96.23	93.75	94.26	98.60	93.47
	2	98.86	99.38	**98.29**	**96.63**	93.91	93.87	98.86	94.71
	4	99.23	99.65	**98.44**	**96.84**	94.58	94.40	**99.02**	95.48
5	1	98.59	99.25	98.07	96.23	93.76	94.27	98.60	**93.51**
	2	**98.89**	99.39	**98.29**	**96.63**	**93.98**	**93.96**	**98.87**	**94.75**
	4	**99.24**	**99.67**	**98.44**	**96.84**	**94.73**	**94.55**	**99.02**	**95.52**

Layer Aggregation We evaluate different combinations of the 12 DINOv3 layers under 1-, 2-, and 4-Ref settings. As shown in Table 11, using only shallow layers $\{0, 1, 2, 3\}$ leads to substantially weaker performance, while middle-to-late representations consistently provide stronger results across different reference budgets. Among them, $\{7, 8, 9, 10\}$, $\{8, 9, 10, 11\}$, and $\{1, 2, 9, 10\}$ are generally competitive, but $\{7, 8, 9, 10\}$ provides a particularly balanced performance across metrics and reference budgets, with consistently strong pixel-level localization. We therefore adopt $\{7, 8, 9, 10\}$ as the default scoring layers in our experiments.

Backbone Ablation Details We further evaluate different backbone families and model scales. As shown in Table 12, the DINOv3 family achieves the strongest overall performance across the evaluated configurations. Smaller backbones generally perform worse, possibly because their limited feature dimensionality restricts visual representation capacity, whereas excessively large backbones do not consistently yield further gains, possibly due to redundant visual information that is less beneficial for anomaly discrimination. Overall, increasing visual backbone size does not consistently improve performance across the evaluated configurations.

Table 11: Ablation on different combinations of scoring layers under different reference budgets on MVTec-AD and VisA. The best results under each reference budget are highlighted in bold.

Layers	K-Ref	MVTec-AD				VisA			
		I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
$\{7, 8, 9, 10\}$	1	98.59	99.25	98.07	96.23	93.76	94.27	98.60	**93.51**
	2	**98.89**	99.39	98.29	96.63	93.98	93.96	98.87	**94.75**
	4	**99.24**	**99.67**	98.44	96.84	94.73	94.55	**99.02**	**95.52**
$\{0, 1, 2, 3\}$	1	94.29	97.35	95.01	86.83	81.58	84.86	93.16	74.45
	2	96.44	98.44	96.19	89.94	85.35	87.35	94.57	79.22
	4	97.06	98.70	96.88	91.72	89.68	90.55	95.49	82.75
$\{4, 5, 6, 7\}$	1	98.25	99.03	97.94	95.67	91.40	91.88	97.98	91.75
	2	98.84	99.37	98.24	96.27	92.59	92.49	98.43	93.37
	4	99.21	99.63	98.44	96.61	93.77	93.80	98.69	94.57
$\{8, 9, 10, 11\}$	1	**98.64**	99.33	**98.21**	**96.31**	**93.82**	94.20	**98.67**	93.10
	2	98.84	99.39	98.38	96.59	93.90	93.72	**98.94**	94.33
	4	99.15	99.62	98.50	96.72	94.50	94.23	99.01	95.19
$\{1, 2, 5, 6\}$	1	97.49	98.79	97.70	94.86	89.18	90.48	97.28	88.85
	2	98.55	99.35	98.12	95.79	91.50	91.95	97.95	91.43
	4	99.02	99.56	98.40	96.31	93.59	94.00	98.38	93.27
$\{1, 2, 9, 10\}$	1	**98.64**	**99.39**	98.17	96.19	93.52	**94.34**	98.35	91.87
	2	**98.89**	**99.48**	**98.41**	**96.66**	**94.51**	**94.72**	98.75	93.71
	4	99.22	99.65	**98.57**	**96.89**	**95.55**	**95.69**	99.00	94.78

B Additional Results

B.1 Class-Wise Quantitative Results

We report the class-wise quantitative results of our method under the 1-Ref, 2-Ref, 4-Ref, 8-Ref, 16-Ref and full-shot settings on MVTec-AD and VisA in Tables 13–16. Note that K-Ref experiments are conducted under the multi-class setting, while full-shot experiments are conducted under the single-class setting. We report mean I-AUROC (%), I-AUPR (%), P-AUROC (%), and PRO (%) results.

B.2 Additional Qualitative Results

Figures 7–12 present qualitative results for representative samples from the MVTec-AD and VisA datasets. These examples further illustrate the localization performance of our method across diverse object categories and anomaly types.

Table 12: Detailed backbone ablation under different reference budgets on MVTec-AD and VisA. The best results are highlighted in bold.

Backbone	Scale	K-Ref	MVTec-AD				VisA			
			I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
DINOv3	S	1	97.7	98.6	97.5	94.5	90.6	91.6	97.8	89.4
		2	98.2	98.8	97.9	95.4	91.5	91.6	98.3	91.8
		4	98.7	99.2	98.1	95.9	92.4	92.2	98.7	93.1
	B	1	98.6	99.3	98.1	96.2	93.8	94.3	98.6	93.5
		2	98.9	99.4	98.3	96.6	94.0	94.0	98.9	94.8
		4	**99.2**	**99.7**	**98.4**	**96.8**	94.7	94.5	99.0	**95.5**
	L	1	98.1	98.6	97.9	95.9	93.9	93.6	98.5	93.5
		2	98.6	98.9	98.1	96.4	94.2	94.0	98.8	94.7
		4	99.1	99.4	98.3	96.7	**95.1**	**94.8**	**99.1**	95.4
	H^+	1	97.8	98.3	97.8	95.9	92.0	92.7	98.1	92.5
		2	98.5	98.8	98.1	96.4	92.7	93.2	98.6	93.8
		4	98.9	99.3	98.3	**96.8**	94.3	94.7	98.8	94.6
DINOv2	S	1	97.8	98.6	97.8	95.5	88.0	87.7	97.9	92.1
		2	98.0	98.7	98.0	95.9	89.2	88.5	98.4	93.3
		4	98.6	99.2	98.2	96.2	91.3	91.0	98.6	94.0
	B	1	98.6	99.5	98.0	95.7	90.9	91.7	98.2	93.8
		2	98.7	99.4	98.2	96.2	92.8	93.1	98.5	94.8
		4	99.0	99.5	**98.4**	96.5	94.0	94.3	98.8	**95.5**
	L	1	98.4	99.3	97.6	95.4	91.7	92.3	98.2	92.3
		2	98.8	99.4	98.0	95.8	92.7	93.1	98.5	93.5
		4	99.1	**99.7**	98.2	96.2	94.2	94.2	98.8	94.2
DINO	S	1	96.9	98.3	97.6	93.5	85.3	86.5	97.2	85.6
		2	97.4	98.6	98.0	94.1	88.0	88.1	97.9	87.9
		4	98.0	98.9	98.2	94.5	90.4	90.5	98.2	88.9
	B	1	96.2	98.0	97.5	93.3	86.3	86.8	97.8	85.9
		2	97.0	98.4	97.9	94.0	88.2	88.5	98.3	88.4
		4	97.5	98.7	98.0	94.4	90.0	90.0	98.6	89.8
SigLIP 2	B	1	97.0	97.8	95.9	91.8	89.6	90.3	96.8	87.0
		2	97.9	98.4	96.3	92.1	90.9	91.5	97.4	88.7
		4	98.5	98.8	96.6	92.5	92.8	93.0	97.7	89.4
	L	1	97.1	97.9	95.2	90.5	90.0	90.6	96.6	85.8
		2	97.9	98.6	95.7	90.8	91.4	91.4	97.1	87.6
		4	98.2	99.0	96.0	91.3	93.0	93.4	97.4	88.6
SigLIP	B	1	97.3	98.4	96.0	92.0	89.8	90.3	97.6	90.8
		2	98.1	98.9	96.4	92.2	91.1	91.5	98.0	91.6
		4	98.7	99.2	96.6	92.3	92.9	93.2	98.2	92.1

Table 13: Per-class performance on the MVTec-AD dataset under the 1-Ref, 2-Ref and 4-Ref settings.

Category	1-Ref				2-Ref				4-Ref			
	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
Bottle	99.9	100.0	98.4	96.7	100.0	100.0	98.7	97.1	100.0	100.0	98.8	97.1
Cable	98.6	99.3	98.4	94.8	99.2	99.6	98.7	95.9	99.3	99.6	98.8	96.5
Capsule	97.6	99.4	98.8	96.8	98.1	99.6	99.0	97.2	98.4	99.6	99.1	97.6
Carpet	100.0	100.0	99.6	99.0	100.0	100.0	99.6	99.0	100.0	100.0	99.6	99.0
Grid	100.0	100.0	99.5	97.7	100.0	100.0	99.5	97.6	100.0	100.0	99.5	97.6
Hazelnut	100.0	100.0	99.6	96.8	100.0	100.0	99.6	97.3	100.0	100.0	99.6	97.1
Leather	100.0	100.0	99.5	99.2	100.0	100.0	99.5	98.9	100.0	100.0	99.5	98.8
Metal Nut	100.0	100.0	97.7	96.2	100.0	100.0	97.9	96.7	100.0	100.0	98.1	97.0
Pill	97.0	99.4	93.7	96.8	98.0	99.6	94.6	97.2	98.6	99.7	95.1	97.5
Screw	94.6	98.0	99.0	95.3	93.9	97.1	99.1	96.0	95.5	98.2	99.2	96.4
Tile	99.9	100.0	97.3	93.7	100.0	100.0	97.5	93.9	100.0	100.0	97.6	94.0
Toothbrush	100.0	100.0	99.3	97.2	100.0	100.0	99.4	97.4	100.0	100.0	99.4	97.4
Transistor	94.1	93.5	94.5	89.3	96.5	95.7	95.2	91.0	98.9	98.6	96.0	92.2
Wood	98.9	99.6	97.0	97.1	98.6	99.6	97.1	96.8	98.3	99.4	97.3	96.9
Zipper	98.3	99.6	98.9	96.9	98.9	99.7	99.0	97.2	99.5	99.9	99.1	97.4
Mean	**98.6**	**99.3**	**98.1**	**96.2**	**98.9**	**99.4**	**98.3**	**96.6**	**99.2**	**99.7**	**98.4**	**96.8**

Table 14: Per-class performance on the MVTec-AD dataset under the 8-Ref, 16-Ref and full-shot settings.

Category	8-Ref				16-Ref				Full-shot			
	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
Bottle	100.0	100.0	98.8	97.1	100.0	100.0	98.9	97.2	100.0	100.0	99.0	97.3
Cable	99.4	99.6	98.9	96.7	99.4	99.6	98.9	96.4	99.0	99.4	98.9	96.1
Capsule	98.7	99.7	99.2	97.8	98.8	99.7	99.2	97.9	99.2	99.8	99.3	98.1
Carpet	100.0	100.0	99.6	99.0	100.0	100.0	99.6	99.0	100.0	100.0	99.6	98.9
Grid	100.0	100.0	99.5	97.7	100.0	100.0	99.5	97.7	100.0	100.0	99.5	97.6
Hazelnut	100.0	100.0	99.6	97.2	100.0	100.0	99.6	97.3	99.9	100.0	99.6	97.3
Leather	100.0	100.0	99.5	98.7	100.0	100.0	99.5	98.7	100.0	100.0	99.5	98.8
Metal Nut	100.0	100.0	98.1	97.1	100.0	100.0	98.2	97.1	100.0	100.0	98.2	97.1
Pill	99.0	99.8	95.4	97.6	99.2	99.9	95.7	97.6	99.3	99.9	96.3	97.9
Screw	95.3	98.1	99.3	96.8	96.2	98.7	99.3	96.9	94.3	98.2	99.1	96.4
Tile	100.0	100.0	97.6	94.1	100.0	100.0	97.7	94.2	100.0	100.0	97.7	94.2
Toothbrush	100.0	100.0	99.4	97.4	100.0	100.0	99.4	97.5	100.0	100.0	99.4	97.5
Transistor	99.2	99.0	96.6	93.2	99.3	99.1	96.8	93.5	99.4	99.2	97.3	93.9
Wood	98.2	99.4	97.3	96.8	98.1	99.4	97.3	96.8	98.1	99.4	97.1	96.6
Zipper	99.9	100.0	99.1	97.4	99.9	100.0	99.1	97.5	100.0	100.0	99.2	97.6
Mean	**99.3**	**99.7**	**98.5**	**97.0**	**99.4**	**99.8**	**98.6**	**97.0**	**99.3**	**99.7**	**98.7**	**97.0**

Table 15: Per-class performance on the VisA dataset under the 1-Ref, 2-Ref and 4-Ref settings.

Category	1-Ref				2-Ref				4-Ref			
	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
Candle	94.5	94.8	99.4	97.8	95.0	95.4	99.5	97.9	95.4	95.8	99.6	97.9
Capsules	97.0	98.4	98.1	94.1	97.2	98.4	98.5	96.0	98.2	98.9	98.6	97.1
Cashew	96.3	98.3	99.3	98.2	95.4	97.8	99.4	98.1	96.8	98.3	99.4	98.2
Chewing Gum	98.2	99.2	99.6	92.0	98.5	99.4	99.5	92.8	99.2	99.6	99.6	93.3
Fryum	98.1	99.1	96.6	93.7	97.1	98.6	96.9	94.5	98.0	99.0	97.2	94.9
Macaroni1	92.6	92.6	99.4	94.7	93.1	93.7	99.5	95.6	90.3	90.9	99.5	96.6
Macaroni2	70.7	71.8	97.9	88.6	71.1	67.7	98.6	92.5	75.5	71.8	99.1	94.4
PCB1	92.8	91.6	99.7	95.1	93.2	91.4	99.8	96.1	94.0	91.9	99.8	97.1
PCB2	92.4	91.7	98.3	88.9	93.1	91.1	98.7	91.1	94.7	93.8	98.9	92.9
PCB3	94.4	95.4	98.3	91.7	96.5	96.7	98.7	93.3	96.6	96.7	99.0	94.0
PCB4	98.9	98.8	97.7	89.6	98.3	97.7	98.0	91.2	98.5	98.0	98.3	92.0
Pipe Fryum	99.3	99.6	99.1	97.8	99.2	99.6	99.2	97.8	99.6	99.8	99.2	97.9
Mean	**93.8**	**94.3**	**98.6**	**93.5**	**94.0**	**94.0**	**98.9**	**94.8**	**94.7**	**94.5**	**99.0**	**95.5**

Table 16: Per-class performance on the VisA dataset under the 8-Ref, 16-Ref and full-shot settings.

Category	8-Ref				16-Ref				Full-shot			
	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
Candle	96.4	96.3	99.6	97.7	96.4	96.3	99.7	97.8	98.9	98.9	99.7	98.1
Capsules	98.5	99.0	98.8	97.5	99.0	99.4	99.0	97.9	99.0	99.3	99.1	98.0
Cashew	98.2	99.0	99.5	98.2	98.2	99.0	99.5	98.2	97.9	98.9	99.6	98.4
Chewing Gum	98.9	99.5	99.6	93.4	98.8	99.4	99.5	93.4	99.0	99.6	99.6	93.9
Fryum	98.3	99.2	97.4	95.0	98.4	99.2	97.5	95.0	98.8	99.5	97.6	95.3
Macaroni1	92.7	92.6	99.6	97.1	93.5	93.4	99.7	97.4	96.2	96.4	99.8	98.2
Macaroni2	76.2	72.7	99.3	95.0	78.1	78.1	99.5	95.7	88.4	88.3	99.7	97.8
PCB1	93.4	91.5	99.8	97.4	94.8	92.4	99.8	97.6	99.6	99.6	99.8	97.5
PCB2	95.5	94.3	99.1	93.6	96.2	94.8	99.1	94.1	98.4	98.8	99.2	94.8
PCB3	97.3	97.6	99.1	94.5	98.1	98.3	99.2	94.7	99.3	99.4	99.3	95.5
PCB4	99.0	98.9	98.6	92.7	99.6	99.5	98.6	92.3	99.8	99.8	98.9	94.1
Pipe Fryum	99.2	99.7	99.3	98.1	99.3	99.7	99.3	98.1	99.3	99.6	99.4	98.2
Mean	**95.3**	**95.0**	**99.1**	**95.9**	**95.9**	**95.8**	**99.2**	**96.0**	**97.9**	**98.2**	**99.3**	**96.7**

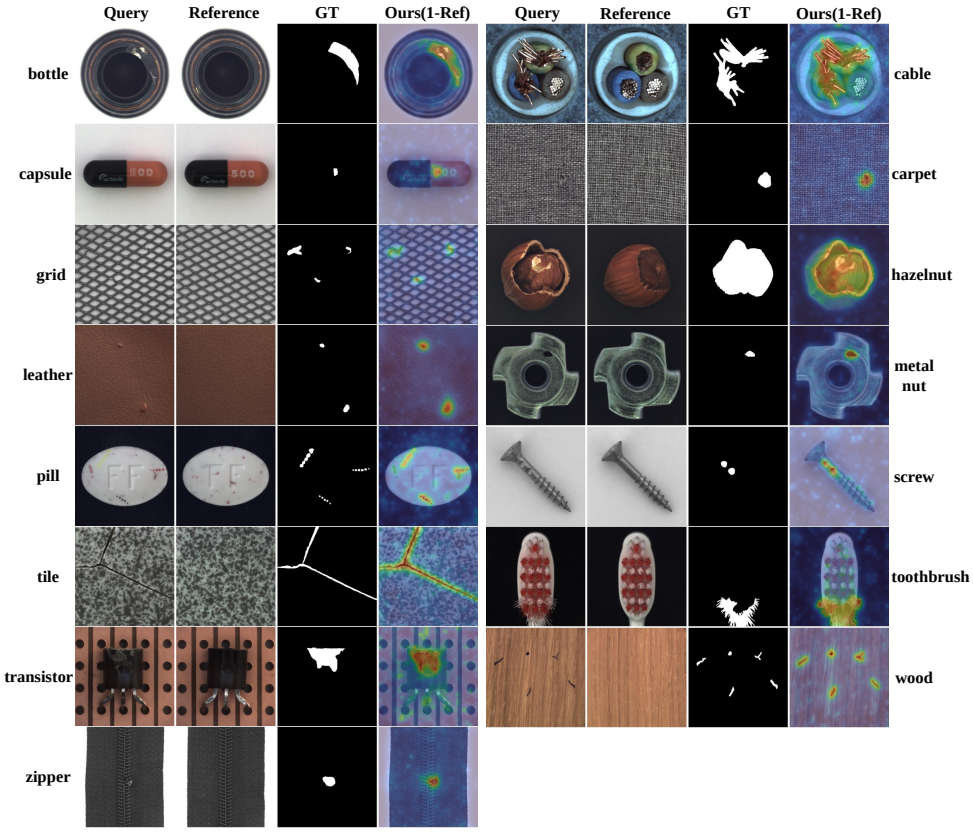


Figure 7: Qualitative results of our method under the 1-Ref setting on MVTec-AD.

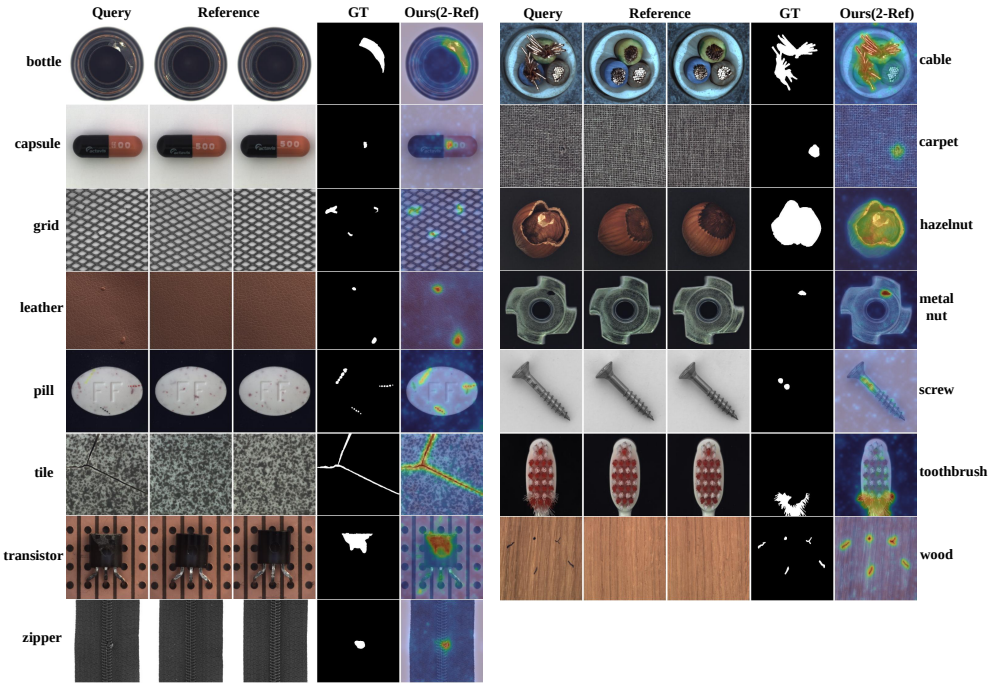


Figure 8: Qualitative results of our method under the 2-Ref setting on MVTec-AD.

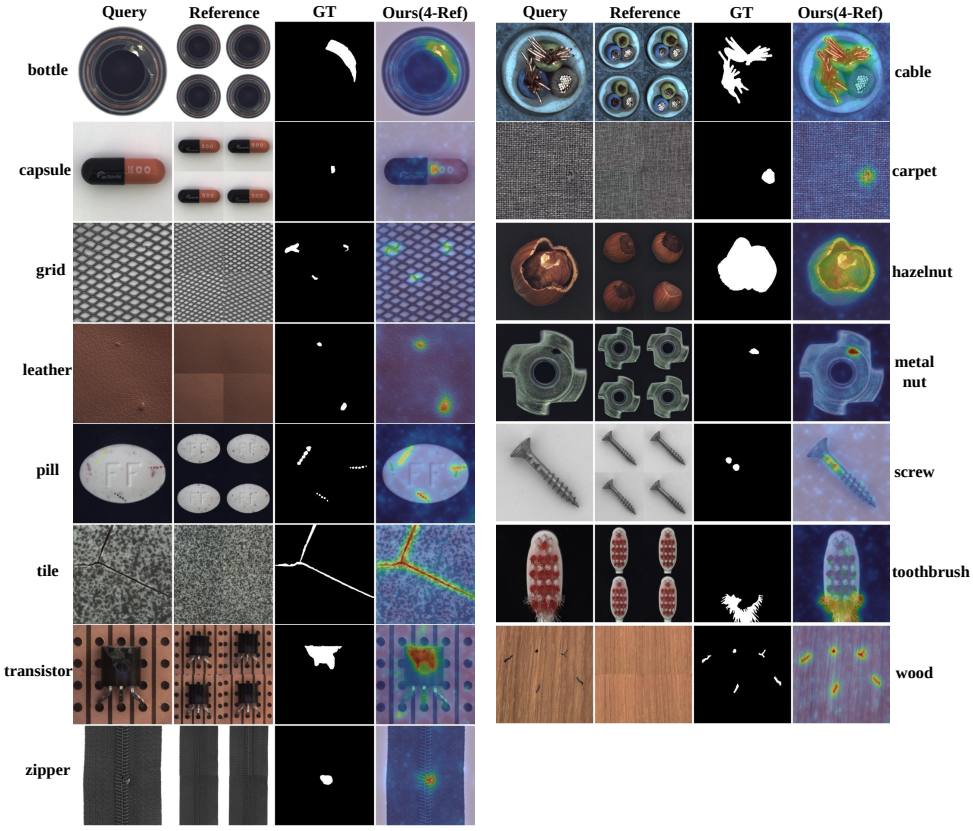


Figure 9: Qualitative results of our method under the 4-Ref setting on MVTec-AD.

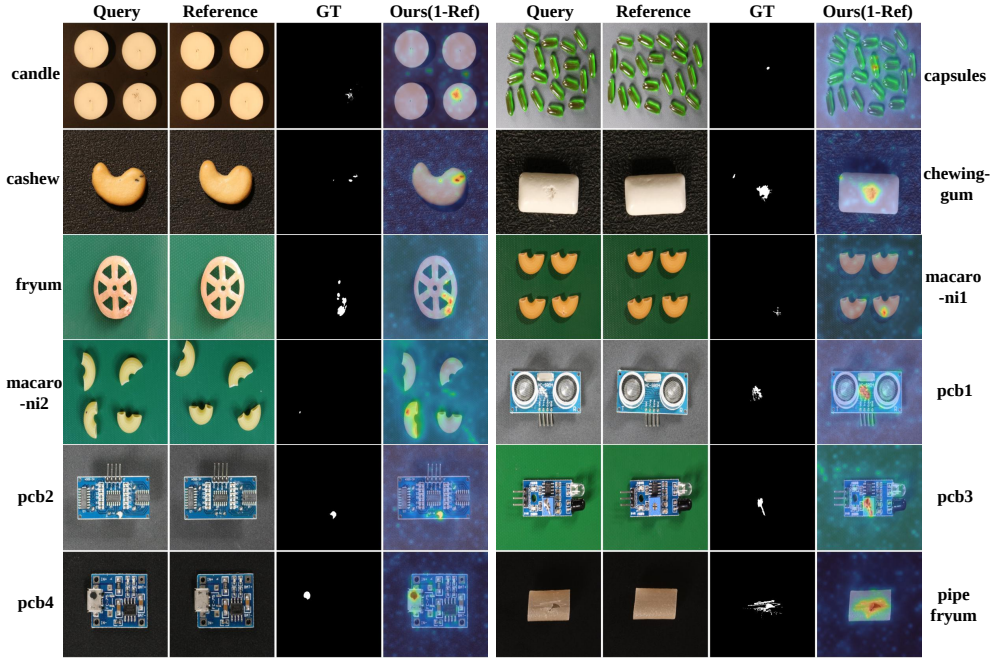


Figure 10: Qualitative results of our method under the 1-Ref setting on VisA.

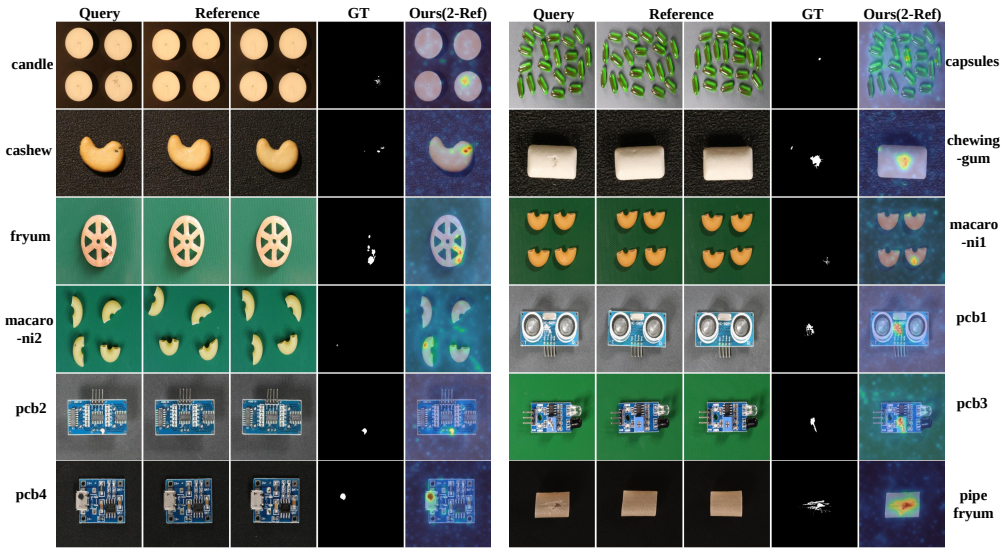


Figure 11: Qualitative results of our method under the 2-Ref setting on VisA.

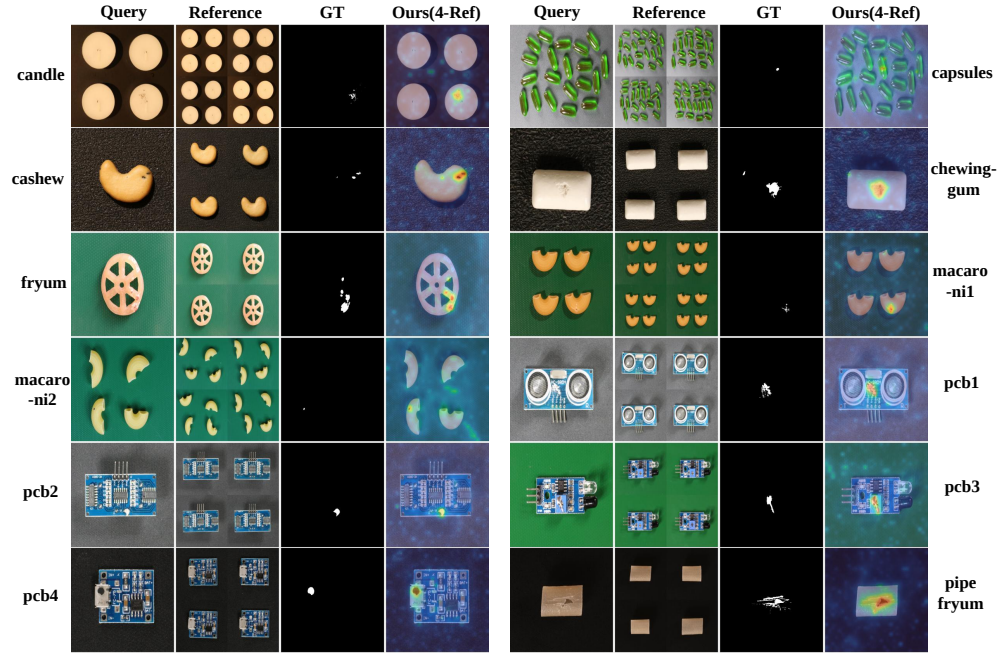


Figure 12: Qualitative results of our method under the 4-Ref setting on VisA.