EVOSAE: EVOLVING SPARSE AUTOENCODER FOR FEW-SHOT INDUSTRIAL ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot anomaly detection (FSAD) aims to identify abnormal samples from only a limited number of normal examples, gaining increasing attention due to its effectiveness and flexibility for new industrial scenarios. However, it is usually hard for existing FSADs to discover sufficient knowledge from scarce reference samples. Moreover, most approaches focus on feature-level representation while overlooking the polysemanticity of neurons in foundation models. In the paper, we introduce `EvoSAE`, a self-improving FSAD framework built upon sparse autoencoders (SAE). Moving beyond feature-level memory construction, `EvoSAE` establishes an atom-level memory by disentangling features as sparse overcomplete directions. These semantic atoms provide more fine-grained representations of normal patterns and enhance the compositional capability for modeling unseen features. Importantly, we softly weight test patches according to their reconstruction errors, assigning larger weights to patches that are more likely to be normal. The weighted patches are then used to fine-tune the SAE during inference, enabling its semantic atoms to progressively adapt to the test distribution. Extensive experiments on three industrial anomaly detection benchmarks demonstrate that `EvoSAE` consistently outperforms recent state-of-the-art FSAD methods across different shot settings. Visualizations of the learned atoms further reveal their semantic interpretability in real-world anomaly detection scenarios.

1 INTRODUCTION

Industrial anomaly detection (IAD) aims to identify and localize defects in manufacturing processes Liu et al. (2024); Huang et al. (2025), playing a crucial role in automated inspection across diverse industrial applications Liu et al. (2023); Huang et al. (2024); Yang et al. (2024). Since industrial anomalies are rare and unpredictable, IAD is typically formulated as an unsupervised or one-class learning problem that models normal patterns from defect-free data. However, conventional IAD methods often rely on abundant normal samples, which can be challenging to collect under short production cycles and frequent manufacturing changeovers Tao et al. (2022); Bouman et al. (2024). To alleviate this dependency, few-shot anomaly detection (FSAD) has emerged as a practical alternative that aims to detect anomalies using only a handful of normal support images from each category Snell et al. (2017); Wang et al. (2020).

Generally, the core of FSAD lies in learning representative normal knowledge from only a few support samples. Reconstruction-based methods leverage Generative Adversarial Networks (GANs) Duan et al. (2023), Transformers Lu et al. (2023), or diffusion models Hu et al. (2024) to model normal patterns, and detect anomalies from the discrepancy between inputs and their reconstructions. Embedding-based methods Nafez et al. (2025); Wei et al. (2025), in contrast, perform feature matching against normal references stored in a memory bank, which may contain image-level features Guo et al. (2023) or patch-level embeddings Damm et al. (2025); Roth et al. (2022a). Motivated by the impressive success of vision foundation models (*e.g.*, ViT Dosovitskiy et al. (2021) and DINO Oquab et al. (2024)), recent advances further enhance both paradigms with stronger pretrained representations. For example, AnomalyDINO Damm et al. (2025) and SubSpaceAD Lendering et al. (2026) construct normal memories from DINOv2 features, while Dinomaly Guo et al. (2025) distills DINOv2 representations into a student decoder.

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.

Despite these advances, we argue that existing methods still face two fundamental limitations. First, they primarily operate at the feature level, either constructing feature memories or performing reconstruction in the feature space. Recent studies have shown that many neurons in foundation models are polysemantic, responding simultaneously to seemingly unrelated concepts Shu et al. (2025); Bussmann et al. (2024); Lim et al. (2025a). Such polysemanticity may entangle distinct normal patterns within dense feature representations, limiting their ability to capture normality in a fine-grained and compositional manner. Second, test samples often arrive in a streaming manner and are predominantly normal, yet existing methods typically keep their normal memory fixed after training, overlooking the opportunity to continuously refine normal knowledge from incoming data. Several recent works have explored test-time memory refinement Guo et al. (2025); Li et al. (2026); Luo et al. (2025b); however, they often rely on additional regularization objectives or learnable selection modules, leaving efficient and reliable test-time adaptation underexplored.

To address these limitations, we introduce `EvoSAE`, a self-improving FSAD framework built upon sparse autoencoders (SAEs). Instead of directly memorizing dense features, `EvoSAE` decomposes them into sparse combinations of overcomplete dictionary directions, which we regard as semantic atoms. By disentangling polysemantic representations into more elementary components, these atoms provide a finer-grained description of normal patterns and can be flexibly recombined to represent unseen normal variations. In this way, `EvoSAE` transforms conventional feature-level normal memory into an atom-level memory, enabling more expressive and compositional modeling under scarce support data.

Beyond constructing a more fine-grained normal memory, `EvoSAE` further allows the learned atoms to evolve during inference. Specifically, we exploit the reconstruction error of the SAE as an estimate of patch normality and softly weight patches in each incoming test sample, assigning larger weights to patches that are more likely to be normal. These weighted patches are then used to continually fine-tune the SAE, allowing reliable information from test data to progressively refine the learned semantic atoms while suppressing the influence of potential anomalies. Our contributions are threefold:

- We introduce an atom-level perspective for few-shot anomaly detection, using sparse autoencoders to decompose dense foundation-model features into sparse and compositional semantic atoms for modeling normal patterns.
- We propose a self-improving test-time adaptation strategy that softly weights incoming test patches according to their reconstruction errors and uses them to continually refine the SAE and its normal memory.
- Extensive experiments on three industrial anomaly detection benchmarks demonstrate consistent improvements over recent state-of-the-art FSAD methods across different shot settings, while visualizations further reveal the semantic interpretability of the learned atoms.

2 Related works

Anomaly Detection & Few-Shot Anomaly Detection. Visual anomaly detection is commonly formulated as an unsupervised or one-class learning problem, where normality is learned from defect-free data and deviations are identified as anomalies. Existing methods mainly follow three paradigms. Reconstruction-based approaches Bergmann et al. (2018); Fang et al. (2023) recover normal images or features and use reconstruction residuals for anomaly scoring. Memory-based methods Deng & Li (2022); Roth et al. (2022b) model the normal feature space with pretrained representations. SPADE Cohen & Hoshen (2005) and PatchCore Roth et al. (2022a), for example, store normal patch embeddings and perform nearest-neighbor matching, while PaDiM Defard et al. (2021) models location-wise feature distributions using Gaussian statistics. A third line synthesizes pseudo-anomalies for discriminative learning. CutPaste Li et al. (2021) and DRAEM Zavrtanik et al. (2021) generate artificial defects and train models to distinguish or localize them. Despite their effectiveness, such methods depend on how well synthetic anomalies approximate real defects.

To extend AD to the few-shot setting, recent methods focus on learning representative normal patterns from only a few references. RegAD Huang et al. (2022) learns category-agnostic registration through meta-learning to transfer knowledge across products, while FastRecon Fang et al. (2023) derives query-specific transformations from limited support features. More recently, vision foun-

dation models have substantially strengthened FSAD. AnomalyDINO Damm et al. (2025) builds robust normal references from DINOv2 patch features with foreground masking and augmentation, whereas SubspaceAD Lendering et al. (2026) models aggregated DINOv2 features with a compact PCA subspace and detects anomalies through projection residuals. INP-Former Luo et al. (2025b) and FastRef Li et al. (2026) further incorporate test-time information by constructing query-adaptive prototypes, bridging the gap between training and testing stages. However, these test-time models adapt independently to each query and do not preserve an evolving normality model across successive test samples.

Sparse Autoencoders. Foundation models often exhibit *polysemanticity*, where individual neurons respond to multiple unrelated concepts, making their representations difficult to disentangle Shu et al. (2025); Bussmann et al. (2024); Lim et al. (2025a). Sparse autoencoders (SAEs) address this issue by decomposing dense activations into sparse combinations of overcomplete latent directions, yielding more interpretable and semantically coherent representations Sharkey et al. (2023); Cunningham et al. (2024). Following the standard ReLU-based SAE Bricken et al. (2023), recent studies have explored improved sparsification mechanisms, including gated activation Rajamanoharan et al. (2024a), learnable thresholds Rajamanoharan et al. (2024b), and Top-K selection Bussmann et al. (2024); Gao et al. (2025a), to improve the sparsity–reconstruction trade-off.

Beyond language models, SAEs have been extended to vision foundation models Qiang et al. (2025); Lim et al. (2025b); Stevens et al. (2025); Oquab et al. (2024), diffusion models Cywiński & Deja (2025); Gao et al. (2025a), and multimodal LLMs Zhang et al. (2025); Pach et al. (2025), demonstrating their ability to uncover semantically meaningful concepts across different modalities and tasks. Despite this progress, the potential of SAEs for fine-grained normal memory construction in few-shot anomaly detection remains largely unexplored. Our work explores this direction by using SAE dictionary directions as semantic atoms to construct an atom-level normal memory and progressively refine it with incoming test samples.

3 Background: Sparse Autoencoder

Dense features from foundation models can be polysemantic, with individual neurons encoding multiple concepts. SAEs alleviate this issue by decomposing dense features into sparse combinations of overcomplete latent directions. Mathematically, given a hidden feature $\boldsymbol{x} \in \mathbb{R}^d$ from pretrained models, the encoding and decoding processes can be formulated as:

$$f(\boldsymbol{x}) = \sigma(\mathbf{W}_{\text{enc}}\boldsymbol{x} + \boldsymbol{b}_{\text{enc}}), \quad \hat{\boldsymbol{x}} = f(\boldsymbol{x})\mathbf{W}_{\text{dec}} + \boldsymbol{b}_{\text{dec}}. \tag{1}$$

where $\mathbf{f}(\boldsymbol{x}) \in \mathbb{R}^m$ denotes the sparse latent representation, $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ is the reconstructed feature, and $m > d$ corresponds to an overcomplete latent dictionary. The activation function $\sigma(\cdot)$ controls the sparsity of the latent representation, and latent i is considered active when $f_i(\boldsymbol{x}) > 0$.

SAEs are trained on the activations of a foundation model at a particular site, such as the residual stream, using a loss function of the form:

$$\mathcal{L}_{\text{SAE}} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 + +\lambda\mathcal{E}(f(\boldsymbol{x})) + \alpha\mathcal{L}_{\text{aux}}, \tag{2}$$

where $\mathcal{E}$ is a function of the latent coefficients that penalizes non-sparse decompositions, and λ is a sparsity coefficient, where higher λ encourages sparsity at the cost of higher reconstruction error. Some SAEs also require the use of an auxiliary loss $\mathcal{L}_{\text{aux}}$, for example to recycle inactive latents in TopK SAEs Gao et al. (2025b).

In this paper, we employ the BatchTopK SAE Bussmann et al. (2024) as our atom-level memory constructor due to its simple implementation and flexible trade-off between reconstruction and sparsity. In detail, BatchTopK SAE specifies $f(\cdot) = \text{BatchTopK}_k(\cdot)$, where $\text{BatchTopK}_k(\cdot)$ retains the largest Bk activations across the batch and sets the rest to zero, yielding an average of k active latents per input. We regard each learned decoder direction as a semantic atom and use the resulting dictionary to represent normal patterns.

4 Proposed Model

We now introduce the details of our proposed `EvoSAE`, whose overall framework is illustrated in Fig. 1. During support training, `EvoSAE` constructs a multi-scale atom-level normal memory from

features extracted by a frozen DINOv2 encoder through sparse autoencoding. At test time, the current memory estimates patch reliability from reconstruction errors, which are then used to guide memory adaptation before anomaly scoring. The updated memory is retained for subsequent test images, allowing EvoSAE to progressively self-improve along the streaming test sequence.

4.1 Multi-Scale Sparse Normality Modeling

Unlike previous works that focus on feature-level memory construction Lendering et al. (2026); Guo et al. (2025), our normal memory consists of the learned decoder directions of an SAE. Let $\mathcal{D}_{\boldsymbol{\theta}} = \{\boldsymbol{a}_j\}_{j=1}^{m}$ denote this dictionary, where $\boldsymbol{a}_j \in \mathbb{R}^d$ is an atom, $m > d$, and $\boldsymbol{\theta}$ contains the SAE encoder and decoder parameters. Using the sparse encoder $f_{\boldsymbol{\theta}}$ introduced in the background, we express reconstruction explicitly as a composition of atoms:

$$\hat{\boldsymbol{x}} = \mathbf{b}_{\text{dec}} + \sum_{j=1}^{m} f_{\boldsymbol{\theta},j}(\boldsymbol{x})\boldsymbol{a}_j, \qquad f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \text{BatchTopK}_k(\boldsymbol{x}). \tag{3}$$

Each patch is therefore described by the atoms it activates and their coefficients. Patches with different overall appearances can reuse common atoms, and new combinations of these atoms can represent normal variations absent from the support set. This compositional structure is the inductive bias we use to model normality from scarce examples.

Let $\mathcal{S} = \{I_n\}_{n=1}^{N}$ be the few-shot normal support set with N images. To expose the dictionary to geometric variations, we augment each image I_n with N_a randomly rotated views, following SubspaceAD Lendering et al. (2026). Together with the original images, these views form a new support set $\mathcal{S}$, containing $N(1+N_a)$ normal images (For simplicity, we continue to use $\mathcal{S}$ to denote the augmented support set when no ambiguity arises.). For each image $I \in \mathcal{S}$, we average the patch features from a set of DINOv2 layers $\mathcal{L}$:

$$\mathbf{F}(I) = \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}} \phi_l(I), \qquad \mathbf{F}(I) \in \mathbb{R}^{H\times W\times d}, \tag{4}$$

where ϕ_l is the l-th encoder layer of DINOv2. Let Ω be the $H \times W$ patch grid. Following the local neighborhood aggregation in MuSc Li et al. (2024b), for patch $p \in I$ we construct its features at multiple spatial scales:

$$\boldsymbol{x}_{I,p}^{(r)} = \frac{1}{|\mathcal{N}_r(p)|}\sum_{q\in\mathcal{N}_r(p)} \mathbf{F}(I)(q), \qquad p \in \Omega, \quad r \in \mathcal{R}, \tag{5}$$

where $\mathcal{N}_r(p)$ denotes the $r \times r$ neighborhood of p and $\mathcal{R}$ is the set of aggregation scales. Small neighborhoods retain local appearance, while larger neighborhoods encode surrounding structure. All scales remain in the same d-dimensional feature space and can therefore be reconstructed by the same dictionary. The final support features used to learn the memory are$\mathcal{X}_{\mathcal{S}} = \left\{\mathbf{x}_{I,p}^{(r)} \;\middle|\; I \in \mathcal{S},\ p \in \Omega,\ r \in \mathcal{R}\right\}$, which will be used to train the BatchTopK SAE with Eq. equation 2 under the framework of Eq. 3. This shared optimization enables the SAE to learn a unified sparse dictionary of normal patterns across different geometric variations and spatial contexts, thereby providing a multi-scale normality prior for subsequent continual test-time adaptation.

4.2 Spatially Calibrated Soft Normality Weighting

The support-trained memory provides an initial model of normality, but its coverage is limited by the available examples. Fortunately, Incoming test images offer additional normal appearances, together with possible defects. To identify the normal feature in the test image, we estimate a soft normality weight for every test patch based on its reconstruction error.

For brevity, let $\boldsymbol{x}_{t,p}^{(r)}$ denote the feature extracted from test image I_t at location p and scale r. Its reconstruction error under parameters $\boldsymbol{\theta}_{t-1}$ (with $\boldsymbol{\theta}_0$ denoting the pretrained parameters) is

$$e_{t,p}^{(r)}(\boldsymbol{\theta_{t-1}}) = \left\|\mathbf{x}_{t,p}^{(r)} - g_{\boldsymbol{\theta}}\left(\mathbf{x}_{t,p}^{(r)}\right)\right\|_2^2. \tag{6}$$

A low error indicates compatibility with the current normal memory. Instead of directly using Eq. 6 to measure the weighting of $\boldsymbol{x}_{t,p}^{(r)}$, we incorporate neighboring evidence by averaging the errors over

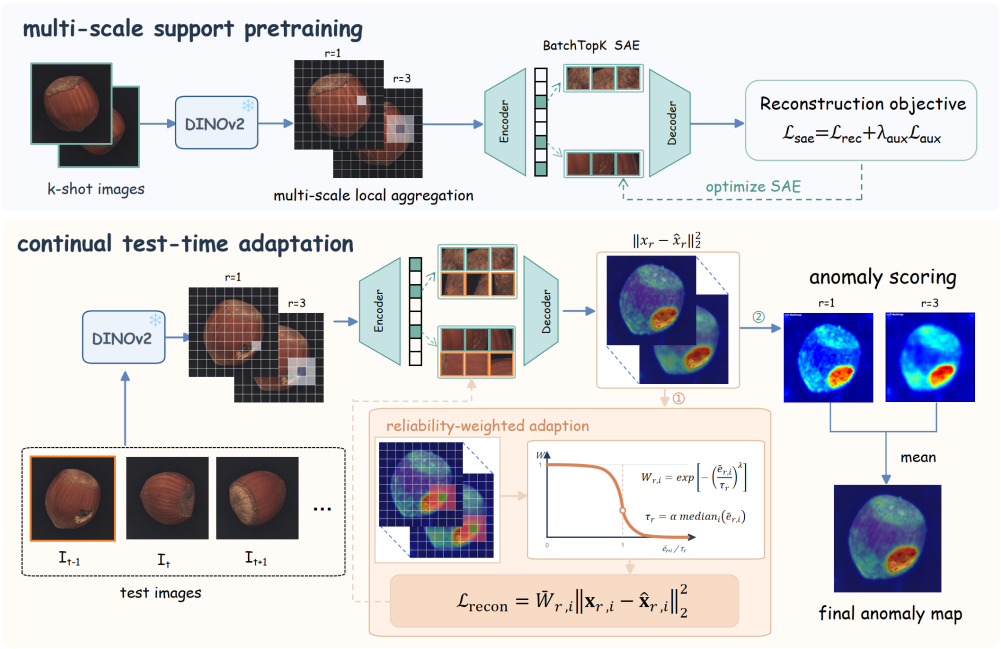


Figure 1: The overall framework of EvoSAE EvoSAE learns a set of multi-scale atom-level memory under the SAE framework at the support training stage. And self-imporving its memory according to the reconstruction error at the test time.

its $\rho \times \rho$ neighborhood:

$$\bar{e}_{t,p}^{(r)} = \frac{1}{|\mathcal{N}_\rho(p)|} \sum_{q \in \mathcal{N}_\rho(p)} e_{t,q}^{(r)}(\boldsymbol{\theta}_{t-1}). \tag{7}$$

This makes reliability depend on local spatial evidence and reduces the influence of isolated residual fluctuations. Here, r controls the spatial context of the input features, whereas ρ controls the spatial context used to estimate their reliability. The final weighting score for $\boldsymbol{x}_{t,p}^{(r)}$ is defined as:

$$w_{t,p}^{(r)} = \exp\left\{ -\left(\frac{\bar{e}_{t,p}^{(r)}}{\tau(c_t^{(r)} + \epsilon)} \right)^\gamma \right\}, \tag{8}$$

where $\tau > 0$ sets the error scale of the decay, $\gamma > 0$ controls its shape, $\epsilon > 0$ ensures numerical stability. $c_t^{(r)} = \underset{p \in \Omega}{\text{median}}\, \bar{e}_{t,p}^{(r)}$ denotes the median value among all patches. Eq. 8 assigns greater weight to patches that the current memory explains well relative to other patches in the same image and scale, which will be used as the most normal sample to fine-tune our memory.

4.3 Evolving the Memory through Weighted Adaptation

The normality weights determine how much each test patch contributes to learning. With the weights in Eq. equation 8 calculated, the reconstruction term in Eq. 2 is formulated as:

$$\mathcal{L}_{\text{recon}} = \bar{w}_{t,p}^{(r)} \left\| \mathbf{x}_{t,p}^{(r)} - \hat{\mathbf{x}}_{t,p}^{(r)} \right\|_2^2, \tag{9}$$

where $\bar{w}_{t,p}^{(r)}$ is the normalized weighting across all scales and patches. Intuitively, patches assigned higher trust contribute more strongly to the reconstruction objective, while those assigned lower trust have reduced relative influence. Starting from $\boldsymbol{\theta}_{t-1}$, we optimize Eq. equation 9 for J adaptation steps and obtain $\boldsymbol{\theta}_t$. Updating the sparse encoder and decoder adjusts both the assignment of features to atoms and the atoms themselves, allowing normal evidence from the current image to refine the shared memory.

The parameters $\boldsymbol{\theta}_t$ are then used to score I_t and initialize adaptation to I_{t+1}. This persistence is central to EvoSAE: information incorporated from earlier images can affect both reconstruction and

reliability estimation for later images. The reconstruction residual therefore has two roles in the pipeline: before adaptation it guides which patches influence the memory, and after adaptation it measures their remaining discrepancy from that memory. The pixel-level and image-level anomaly detection are then calculated according to the second reconstruction error with θ_t for test image I_t.

5 EXPERIMENTS

5.1 EXPERIMENT SETUP

Datasets. Following previous works Li et al. (2026); Lendering et al. (2026), three benchmarks: MVTec AD Bergmann et al. (2019), VisA Zou et al. (2022), and MPDD Jezek et al. (2021) are used to test our EvoSAE. MVTec-AD contains 15 object and texture categories covering diverse industrial defects. VisA comprises 12 object categories with complex structures and various types of surface and structural anomalies. MPDD focuses on six categories of metal parts captured under varying poses, scales, and illumination conditions. All three datasets provide pixel-level annotations for anomalous regions, supporting both image-level anomaly detection and pixel-level anomaly localization.

Evaluation Metrics. For pixel-level anomaly localization, we bilinearly upsample the reconstruction error to the input resolution and apply Gaussian smoothing with standard deviation $\sigma = 4$ to suppress high-frequency noise while preserving localization accuracy. We report AUROC and the per-region overlap (PRO) score. For image-level anomaly detection, we calculate the average values in the set containing the top 1% pixel-level anomaly scores. Following the standard evaluation protocol Damm et al. (2025); Roth et al. (2022a), PRO is computed as the area under the PRO curve for false-positive rates up to 0.3. Higher values indicate better performance for all metrics.

Implementation Details. In this paper, we use the frozen DINOv2-G model Oquab et al. (2023) as our feature extractor, with features averaged across layers 22-28, as described in Sec. 4.1. All images are resized to 672×672 pixels before feature extraction, and the spatial aggregation degrees are set to $\mathcal{R} = \{1, 3\}$. Please refer to the Appendix A.1 for more detailed implementation.

Baselines. We compare our method against the most recently representative few-shot methods, including RegAD Huang et al. (2022), GraphCore Xie et al. (2023), PatchCore Roth et al. (2022a), WinCLIP Jeong et al. (2023), PromptAD Li et al. (2024a), AnomalyDINO Damm et al. (2025), DCP-SFR Jiang et al. (2026), SubspaceAD Lendering et al. (2026), and methods refine their normal memory at the test time: FOCT Tian et al. (2024), FastRecon Fang et al. (2023), INP-Former Luo et al. (2025a), FastRef Li et al. (2026). For baseline models, we either reported the results as presented in the original papers or reran their official code using new datasets.

5.2 QUANTITATIVE COMPARISON

Table 1 compares our method with recent few-shot anomaly detection methods on MVTec-AD, VisA, and MPDD under the 1-, 2-, and 4-shot settings. Across the 36 dataset–shot–metric combinations in Table 1, EvoSAE overall achieves the best result in 28 cases and the second-best result in 7 cases, ranking among the top two in all but one case. In the particularly challenging 1-shot setting, our method achieves an image-level AUROC of 98.4% and a PRO of 95.9% on MVTec-AD, outperforming all previous methods. On VisA, it obtains image- and pixel-level AUROC scores of 94.8% and 98.9%, respectively, exceeding the previous state-of-the-art methods by a clear margin. On MPDD, our method achieves image- and pixel-level AUROC scores of 78.6% and 98.0%, respectively, demonstrating competitive anomaly classification and superior localization performance. As the number of support images increases, our method consistently maintains its advantage in both image-level classification and pixel-level localization. Under the 4-shot setting, it achieves an image AUROC of 99.1% on MVTec-AD, a PRO of 96.6% on VisA, and a pixel AUROC of 98.5% on MPDD. Overall, these results demonstrate the effectiveness and robustness of our method across datasets with diverse anomaly characteristics, highlighting its ability to exploit additional normal reference information at test time while remaining highly competitive in very low-data regimes.

5.3 QUALITATIVE ANALYSIS

In addition to the quantitative comparison, we also provide visualizations of our EvoSAE. Fig. 2 compares the anomaly maps of various models on the 1-shot settings. We find that our EvoSAE

Table 1: Few shot performance comparison on the MVTec-AD, VisA and MPDD datasets with DINOv2-G. Best and second-best results are highlighted in bold and underlined, respectively. Results on different version of DINOv2 can be found in the Appendix A10.

Setup	Method	MVTec-AD		VisA		MPDD	
		Image-level AUROC/AUPR	Pixel-level AUROC/PRO	Image-level AUROC/AUPR	Pixel-level AUROC/PRO	Image-level AUROC/AUPR	Pixel-level AUROC/PRO
1-shot	RegAD	82.9/–	92.5/–	–/–	–/–	60.9/–	92.6/–
	GraphCore	89.9/–	95.6/–	–/–	–/–	**84.7**/–	95.2/–
	PatchCore	83.4/92.2	92.0/79.7	79.9/82.8	95.4/80.5	69.9/70.7	95.3/83.9
	WinCLIP	93.1/96.5	95.2/87.1	83.8/85.1	96.4/85.1	64.0/70.4	92.6/78.4
	PromptAD	94.6/97.1	95.9/87.9	86.9/88.4	96.7/85.1	74.7/78.1	95.8/87.2
	AnomalyDINO	96.6/98.2	96.8/92.7	87.4/89.0	97.8/92.5	75.9/77.1	95.4/87.8
	DCP-SFR	97.3/98.8	96.9/92.7	92.0/93.7	98.2/93.3	–/–	–/–
	SubspaceAD	98.0/99.0	**97.6**/93.7	93.3/93.2	98.3/93.4	75.2/74.9	97.7/94.0
	FOCT	87.1/–	94.4/–	–/–	–/–	78.9/–	96.2/–
	FastRecon	–/–	–/–	–/–	–/–	67.6/69.3	96.4/87.4
	INP-Former	96.6/98.2	97.0/92.6	91.4/92.2	96.3/89.5	–/–	–/–
	FastRef+AnomalyDINO	97.2/–	97.1/–	89.8/–	98.7/–	77.7/–	97.1/–
	EvoSAE (Ours)	**98.4/99.3**	97.2/**95.9**	**94.8/94.4**	**98.9/95.9**	78.7/**78.4**	**98.0/94.7**
2-shot	RegAD	85.7/–	94.6/–	–/–	–/–	63.4/–	93.2/–
	GraphCore	91.9/–	96.9/–	–/–	–/–	**85.4**/–	95.4/–
	PatchCore	86.3/93.8	93.3/82.3	81.6/84.8	96.1/82.6	77.9/80.7	95.9/85.8
	WinCLIP	94.4/97.0	96.0/88.4	84.6/85.8	96.8/86.2	64.2/70.2	92.6/78.7
	PromptAD	95.7/97.9	96.2/88.5	88.3/90.0	97.1/85.8	80.9/82.7	96.1/87.1
	AnomalyDINO	96.9/98.2	97.0/93.1	89.7/90.7	98.0/93.4	79.1/80.2	96.8/91.5
	DCP-SFR	97.7/98.9	97.1/93.3	92.5/94.2	98.4/93.8	–/–	–/–
	SubspaceAD	98.1/99.1	**97.8**/94.0	94.1/93.7	98.4/93.5	81.5/82.5	98.1/95.2
	FOCT	90.5/–	94.8/–	–/–	–/–	82.4/–	96.5/–
	FastRecon	91.0/–	95.9/–	–/–	–/–	71.2/74.2	96.7/89.1
	INP-Former	97.0/98.2	97.2/93.1	94.6/**94.9**	97.2/91.8	–/–	–/–
	FastRef+AnomalyDINO	97.4/–	97.5/–	92.0/–	98.6/–	82.3/–	97.9/–
	EvoSAE (Ours)	**98.7/99.5**	97.5/**96.3**	**95.3/94.9**	**99.0/96.3**	82.5/**83.6**	**98.2/95.4**
4-shot	RegAD	88.2/–	95.8/–	–/–	–/–	68.3/–	93.9/–
	GraphCore	92.9/–	97.4/–	–/–	–/–	**85.7**/–	95.7/–
	PatchCore	88.8/94.5	94.3/84.3	85.3/87.5	96.8/84.9	79.2/82.3	96.8/88.8
	WinCLIP	95.2/97.3	96.2/89.0	87.3/88.8	97.2/87.6	64.7/71.2	92.6/79.0
	PromptAD	96.6/98.5	96.5/90.5	89.1/90.8	97.4/86.2	82.1/86.7	96.6/86.5
	AnomalyDINO	97.7/98.7	97.2/93.4	92.6/92.9	98.2/94.1	78.6/80.5	97.0/92.5
	DCP-SFR	98.0/98.9	97.5/93.5	92.8/94.7	98.6/94.2	–/–	–/–
	SubspaceAD	98.4/99.3	**98.0**/94.2	94.5/93.8	98.5/94.1	82.7/82.5	98.3/95.4
	FOCT	93.2/–	96.2/–	–/–	–/–	83.2/–	96.7/–
	FastRecon	94.2/–	97.0/–	–/–	–/–	74.8/77.0	97.2/91.1
	INP-Former	97.6/98.6	97.0/92.9	**96.4/96.0**	97.7/93.1	–/–	–/–
	FastRef+AnomalyDINO	98.0/–	97.6/–	92.9/–	99.0/–	82.1/–	97.7/–
	EvoSAE (Ours)	**99.1/99.7**	97.8/**96.5**	95.6/95.1	**99.1/96.6**	85.1/**88.0**	**98.5/95.9**

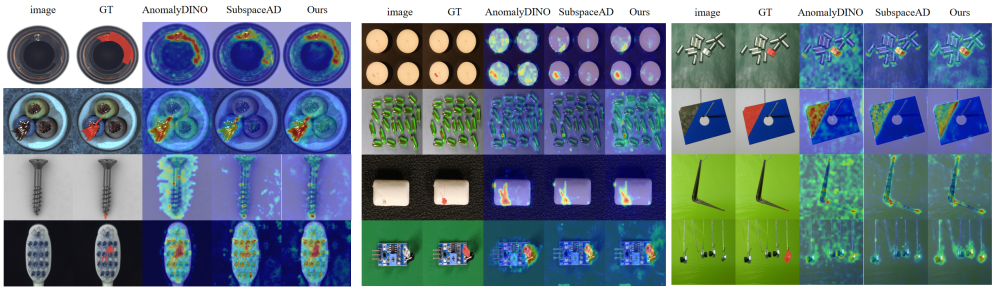


Figure 2: Qualitative comparison on MVTec-AD, VisA and MPDD under 1-shot setting. Our EvoSAE achieves sharper, cleaner anomaly map compared to baselines.

produces sharper, cleaner, and more spatially precise anomaly maps across benchmarks, showing the effectiveness of the atom-level memory bank, which tends to mine refined concepts of normal patterns. Moreover, Fig. 3 visualizes the comparison of reconstruction errors before and after normality-guided test-time adaptation. Normal patches generally exhibit larger error reductions after adaptation, whereas anomalous patches retain relatively high errors or decrease to a lesser extent. This selective adaptation enlarges the reconstruction error gap between normal and anomalous regions, leading to more discriminative anomaly maps. More results are included in Appendix A.2.

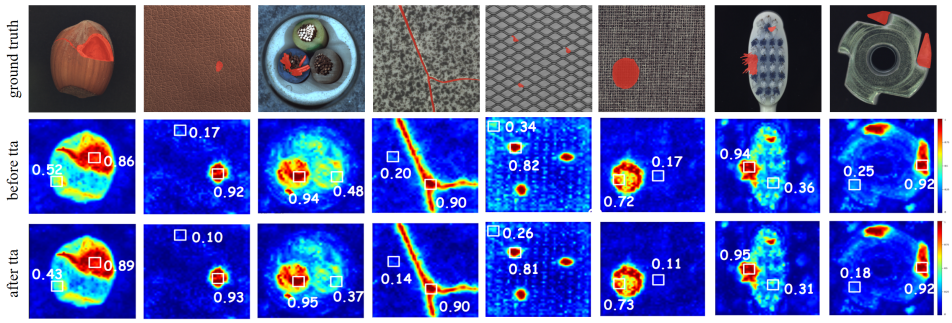


Figure 3: Reconstruction-error comparison before and after normality-guided test-time learning on MVTec-AD dataset, respectively. Normal or anomalous patches are highlighted by box with reconstruction errors.

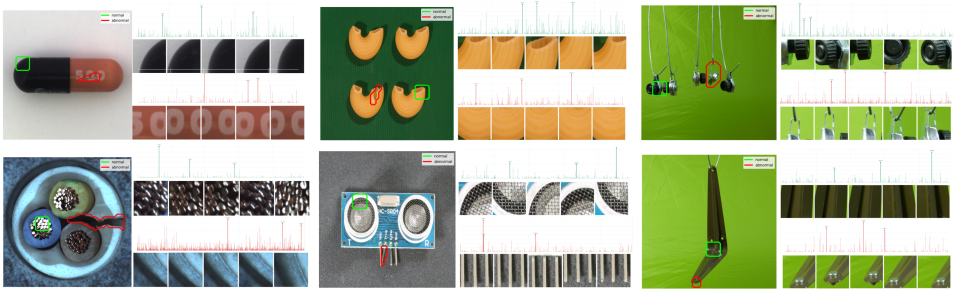


Figure 4: Interpretability of EvoSAE concepts.

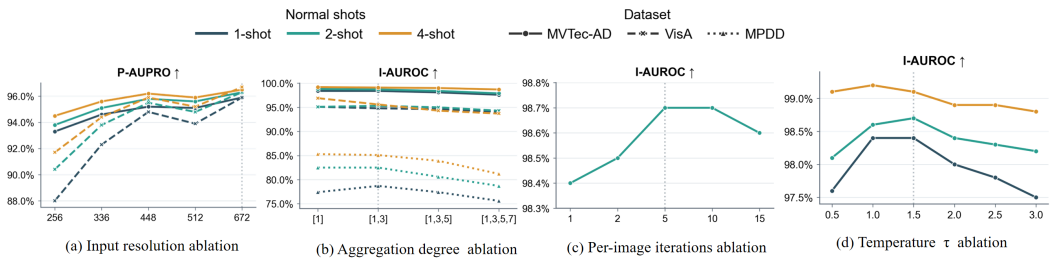
Table 2: Module ablation under the 2-shot setting. Best results are highlighted in bold. SC denotes Spatial Calibration.

Setup	MVTec-AD				VisA				MPDD			
	Image-level		Pixel-level		Image-level		Pixel-level		Image-level		Pixel-level	
	AUROC	AUPR	AUROC	PRO	AUROC	AUPR	AUROC	PRO	AUROC	AUPR	AUROC	PRO
frozen SAE	98.1	99.2	**97.9**	95.8	91.4	91.4	98.7	95.2	82.7	84.6	97.9	94.9
query-specific TTA	98.5	99.4	**97.9**	96.0	92.3	92.2	98.8	95.6	**82.9**	**84.8**	98.0	95.1
continual TTA w/o SC	**98.7**	**99.5**	97.4	96.1	95.0	94.7	98.9	96.0	81.6	82.5	98.1	95.2
Ours	**98.7**	**99.5**	97.5	**96.2**	**95.3**	**94.9**	**99.0**	**96.3**	82.5	83.6	**98.2**	**95.4**

To investigate the interpretability of our EvoSAE, we present a concept-guided region retrieval analysis in Fig. 4. For each query image, a representative normal region and an anomalous region are marked by green and red boxes, respectively. We identify the top-1 concept and retrieve the top-five corresponding patches from the normal support set. For normal regions, the retrieved patches generally exhibit similar textures, contours, or material structures to those observed in the query image. In contrast, anomalous regions produce different activation profiles and retrieved appearances, highlighting their deviation from the normal visual patterns. The concentrated peaks in the activation histograms further suggest that only a sparse subset of concepts is responsible for characterizing each region, providing an interpretable distinction between normal structures and anomaly-related deviations.

5.4 Further Analysis

In this section, we conduct ablation experiments to verify the effectiveness of the proposed elements. Table 2 investigates the stepwise contributions of our test-time adaptation components under the 2-shot setting. Without test-time adaptation, the frozen SAE already establishes a strong reconstruction-based baseline, demonstrating that sparse feature reconstruction can effectively model normal patterns from limited training samples. Building upon this strong foundation, query-specific test-time adaptation improves performance by adapting to each query image and restoring the model weights before processing the next one. Continual adaptation further amplifies



(a) Input resolution ablation (b) Aggregation degree ablation (c) Per-image iterations ablation (d) Temperature τ ablation

Figure 5: Ablation study on input resolution, aggregation degree $\mathcal{R}$, exponent λ, and temperature τ. The vertical dashed lines indicate the default settings adopted in our method.

Table 3: Update parameter ablation under the 2-shot setting.

		Image-level		Pixel-level	
Dataset	Updated	AUROC	AUPR	AUROC	AUPRO
MVTec-AD	Enc. only	**98.7**	**99.5**	**98.0**	96.1
	Dec. only	98.4	99.4	97.8	96.2
	Enc.+Dec.	**98.7**	**99.5**	97.5	**96.3**
VisA	Enc. only	94.0	93.5	98.9	95.8
	Dec. only	94.6	94.2	98.9	96.0
	Enc.+Dec.	**95.3**	**94.9**	**99.0**	**96.3**

Table 4: Exponent λ ablation on MVTec-AD and MPDD under the 2-shot setting.

λ	MVTec-AD		MPDD	
	Image-level	Pixel-level	Image-level	Pixel-level
0	97.0/98.5	97.3/95.3	65.5/69.6	97.0/91.9
1	97.7/99.0	97.4/95.6	72.0/72.7	97.8/94.0
2	98.4/99.4	**97.5/96.3**	79.9/79.6	**98.2/95.5**
4	**98.7/99.5**	**97.5/96.3**	82.5/83.6	**98.2**/95.4
6	**98.7/99.5**	97.4/96.1	**83.0/84.7**	98.1/95.3

these gains by accumulating normal patterns across the test sequence without weight reset, boosting image-level AUROC on VisA from 91.4% to 95.0%. Finally, incorporating spatial calibration suppresses patch-level noise, enabling the full framework to achieve the best overall performance and consistently highest P-AUPRO scores across all three benchmarks.

Fig. 5(a) shows that increasing the input resolution generally improves pixel-level localization, with 672px achieving the best overall performance across different shot settings. The aggregation degree $\mathcal{R}$ controls the neighborhood sizes used for local feature aggregation. As shown in Fig. 5(b), incorporating $r = 3$ generally improves pixel-level localization while maintaining competitive image-level performance. However, excessively large neighborhoods may oversmooth local discriminative patterns. Fig. 5(c)-(d) studies the effect of the number of per-image optimization iterations and the temperature parameter τ on normality weighting. We find that 1), Increasing the number of iterations to 5 improves image-level performance while maintaining the highest P-AUPRO. and 2), a moderate temperature provides a better balance between learning reliable normal patterns and suppressing potential anomalies.

Table 3 compares the updating strategies of SAE. we find that Jointly updating both components consistently outperforms single-component adaptation on both datasets in terms of image-level detection and pixel AUPRO. Table 4 studies the performance with various weight decay exponent λ. According to Eq. 8, $\lambda = 0$ assigns equal weights to all patches, causing severe performance degradation compared with the frozen SAE reported in Table 2, especially on MPDD, where image AUROC and PRO drop from 82.7% to 65.5% and 94.9% to 91.9%. Increasing λ progressively improves performance by assigning lower weights to anomalous patches and suppressing abnormal pattern learning. We therefore set $\lambda = 4$, which provides the best balance between detection and localization.

6 Conclusion

We introduced EvoSAE, a few-shot industrial anomaly detection framework that models normality through sparse combinations of semantic atoms. Reconstruction-based normality weights guide continual test-time adaptation, allowing the atom-level memory to evolve with incoming data. Experiments on three industrial benchmarks demonstrate consistent gains over recent state-of-the-art methods across shot settings, while visualizations support the interpretability of learned atoms. These results highlight the potential of combining compositional normal representations with continual memory refinement.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, LaTeX formatting, and the drafting and debugging of auxiliary visualization code. We have not used generative AI tools to generate datasets, conduct or replace experiments, select or alter experimental results, generate scientific claims, or make final research decisions; the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for improving the clarity and organization of the manuscript. We have reviewed, corrected, and verified all AI-assisted text and code, and tested the visualization code before use. We take responsibility for the final content of this work, including all claims, analyses, experimental results, and artifacts produced with the aid of generative AI.

ACKNOWLEDGMENTS

REFERENCES

Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.

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)*, 2019.

Roel Bouman, Zaharah Bukhsh, and Tom Heskes. Unsupervised anomaly detection algorithms on real-world data: how many do we need? *Journal of Machine Learning Research*, 25(105):1–34, 2024.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. In *NeurIPS Workshop on Scientific Methods for Understanding Deep Learning*, 2024.

N Cohen and Y Hoshen. Sub-image anomaly detection with deep pyramid correspondences. arxiv 2020. *arXiv preprint arXiv:2005.02357*, 2005.

Hoagy Cunningham, Aidan Ewart, Logan Riggs Smith, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *International Conference on Machine Learning*, 2025.

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*, 2025.

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*, 2021.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Yuxuan Duan, Yan Hong, Li Niu, and Liqing Zhang. Few-shot defect image generation via defect-aware feature manipulation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 571–578, 2023.

Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025a.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025b.

Hewei Guo, Liping Ren, Jingjing Fu, Yuwang Wang, Zhizheng Zhang, Cuiling Lan, Haoqian Wang, and Xinwen Hou. Template-guided hierarchical feature restoration for anomaly detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6424–6435. IEEE, 2023.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20405–20415. IEEE, 2025.

Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 8526–8534, 2024.

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*, 2022.

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

Haoqi Huang, Ping Wang, Jianhua Pei, Jiacheng Wang, Shahen Alexanian, and Dusit Niyato. Deep learning advancements in anomaly detection: A comprehensive survey. *IEEE Internet of Things Journal*, 12(21):44318–44342, 2025.

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 (CVPR)*, 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)*, 2021.

Le Jiang, Yan Huang, Zhen Xu, Yong Xu, Hau-San Wong, and Si Wu. Defect cue-preserved structural feature refinement for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Camile Lendering, Erkut Akdag, and Egor Bondarev. Subspacead: Training-free few-shot anomaly detection via subspace modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

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

Yufei Li, Long Tian, Yuyang Dai, Wenchao Chen, Liang Bao, and Xiyang Liu. Fastref: Fast prototype refinement for few-shot industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 24444–24469, 2025a.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *International Conference on Learning Representations*, 2025b.

Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1): 104–135, 2024.

Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12147–12156. IEEE, 2023.

Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 36:8487–8500, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9974–9983. IEEE, 2025b.

Mojtaba Nafez, Amirhossein Koochakian, Arad Maleki, Jafar Habibi, and Mohammad Hossein Rohban. Patchguard: Adversarially robust anomaly detection and localization through vision transformers and pseudo anomalies. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20383–20394. IEEE, 2025.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*, 2025.

Yao Qiang, Chengyin Li, Hui Zhu, Prashant Khanduri, and Dongxiao Zhu. Interpretability-aware vision transformer. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024a.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024b.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Lee Sharkey, Dan Braun, and Beren Millidge. Taking features out of superposition with sparse autoencoders. 2023.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. In *EMNLP (Findings)*, pp. 1690–1712, 2025.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 2017.

Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically rigorous interpretation of vision models. *arXiv preprint arXiv:2502.06755*, 2025.

Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 71:1–21, 2022.

Long Tian, Hongyi Zhao, Ruiying Lu, Rongrong Wang, Yujie Wu, Liming Wang, Xiongpeng He, and Xiyang Liu. Foct: Few-shot industrial anomaly detection with foreground-aware online conditional transport. In *Proceedings of the 32nd ACM international conference on multimedia*, 2024.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 2020.

Shun Wei, Jielin Jiang, and Xiaolong Xu. Uninet: A contrastive learning-guided unified framework with feature selection for anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9994–10003. IEEE, 2025.

Guoyang Xie, Jinbao Wang, Jiaqi Liu, Feng Zheng, and Yaochu Jin. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore. *ICLR*, 2023.

Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18899–18908. IEEE, 2024.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dræm–a discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8310–8319. IEEE, 2021.

Kaichen Zhang, Yifei Shen, Bo Li, and Ziwei Liu. Large multi-modal models can interpret features in large multi-modal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3650–3661, 2025.

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*, 2024.

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*, 2022.

A Additional Experimental Results

A.1 Implementation Details.

In this paper, we use the frozen DINOv2-G model Oquab et al. (2023) as our feature extractor, with features averaged across layers 22-28, as described in Sec. 4.1. All images are resized to 672×672 pixels before feature extraction, and the spatial aggregation degrees are set to $\mathcal{R} = \{1, 3\}$. For the few-shot setting, we randomly sample $N \in \{1, 2, 4\}$ normal images and generate $N_a = 30$ rotated variants per image using angles from 0° to 345°, except for the *transistor* category, for which rotation itself constitutes an anomaly. For our BatchTopK SAE, we set $k = 32$ and the overcomplete dictionary size to $m = 16d$, where d denotes the dimension of input features. We use a learning rate of 3×10^{-4} for SAE pretraining. For spatially calibrated normality weighting, we set the spatial calibration degree p to 3, the temperature parameter τ to 1.5 and the decay-sharpness parameter λ to 4. During test-time adaptation, we adopt a learning rate of 1×10^{-4} to update both the encoder and decoder of SAE, with 5 optimization iterations for each test image. Each few shot configuration averages results over randomly sampled normal images per category, following prior protocols Zhou et al. (2024). All experiments are performed on a single 48G RTX 6000 Ada GPU.

A.2 Category-wise Performance

Table A1: Category-wise performance on MVTec-AD under different few-shot settings.

Category	1-shot				2-shot				4-shot			
	Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO
Bottle	98.6	99.5	99.0	97.4	98.4	99.4	99.1	97.6	99.7	99.9	99.2	97.6
Cable	99.8	99.9	97.9	94.9	99.5	99.7	97.9	95.1	99.4	99.6	97.9	95.0
Capsule	97.3	99.4	98.7	97.4	97.5	99.5	98.8	97.6	97.8	99.5	99.0	98.0
Carpet	100.0	100.0	99.5	98.7	100.0	100.0	99.5	98.7	100.0	100.0	99.5	98.7
Grid	100.0	100.0	99.7	98.7	100.0	100.0	99.7	98.7	100.0	100.0	99.7	98.6
Hazelnut	98.3	98.8	99.5	96.3	100.0	100.0	99.6	97.2	100.0	100.0	99.6	97.7
Leather	100.0	100.0	99.1	98.5	100.0	100.0	99.1	98.6	100.0	100.0	99.2	98.8
Metal Nut	100.0	100.0	89.5	91.0	100.0	100.0	90.5	91.8	100.0	100.0	91.3	92.2
Pill	99.4	99.9	94.2	97.9	99.4	99.9	94.7	98.0	99.4	99.9	95.2	98.0
Screw	84.6	93.7	99.6	98.1	88.4	94.7	99.6	98.3	93.1	96.6	99.7	98.6
Tile	100.0	100.0	98.4	96.1	100.0	100.0	98.4	95.9	100.0	100.0	98.3	95.8
Toothbrush	100.0	100.0	99.6	97.8	100.0	100.0	99.6	97.7	100.0	100.0	99.6	97.8
Transistor	99.9	99.9	88.5	85.0	100.0	100.0	91.7	87.2	100.0	100.0	94.1	88.2
Wood	97.6	99.2	96.1	95.8	97.8	99.3	96.3	96.0	97.8	99.3	96.4	96.1
Zipper	99.8	99.9	97.9	95.2	100.0	100.0	98.2	95.9	99.9	100.0	98.4	96.3
Mean	**98.4**	**99.3**	**97.2**	**95.9**	**98.7**	**99.5**	**97.5**	**96.3**	**99.1**	**99.7**	**97.8**	**96.5**

Table A2: Category-wise performance on VisA under different few-shot settings.

Category	1-shot				2-shot				4-shot			
	Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO
Candle	95.2	94.9	99.6	97.7	95.3	95.1	99.6	97.8	95.5	95.2	99.7	97.8
Capsules	98.0	98.6	99.5	98.5	98.1	98.7	99.5	98.4	98.3	98.9	99.5	98.6
Cashew	95.9	98.0	99.4	98.3	96.3	98.4	99.5	98.3	96.4	98.1	99.6	98.3
Chewing Gum	99.5	99.6	99.7	96.6	99.4	99.7	99.7	96.7	99.4	99.7	99.7	96.4
Fryum	97.6	98.4	95.9	94.6	96.7	98.5	96.1	94.8	96.8	98.5	96.5	94.9
Macaroni1	92.1	90.2	99.8	95.8	92.3	90.7	99.8	95.6	92.1	90.6	99.8	96.4
Macaroni2	87.1	85.4	99.5	95.1	87.9	85.8	99.7	97.1	88.4	86.5	99.8	97.9
PCB1	94.1	91.3	99.6	95.9	94.4	91.4	99.7	96.6	95.2	92.3	99.7	96.7
PCB2	92.0	87.4	98.1	94.1	92.5	87.9	98.4	94.9	92.7	87.9	98.7	95.6
PCB3	92.5	92.1	98.6	93.1	96.2	95.7	99.0	94.7	97.2	96.6	99.2	95.6
PCB4	98.8	99.1	98.0	93.4	99.0	99.1	98.1	92.9	99.1	99.1	98.3	93.5
Pipe Fryum	95.3	97.4	99.1	98.0	95.9	97.6	99.2	98.1	96.3	97.6	99.3	98.2
Mean	**94.8**	**94.4**	**98.9**	**95.9**	**95.3**	**94.9**	**99.0**	**96.3**	**95.6**	**95.1**	**99.1**	**96.6**

A.3 Hyperparameter Analysis

SAE sparsity and dictionary capacity. Tables A4 and A5 investigate the sparsity level and dictionary capacity of the BatchTopK SAE. Increasing Top-K from 8 to 32 consistently improves both image-level and pixel-level performance. The results saturate at $K = 32$, while increasing K to 64 produces no additional improvement at the reported precision. This indicates that a moderate

Table A3: Category-wise performance on MPDD under different few-shot settings.

Category	1-shot				2-shot				4-shot			
	Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO
Bracket Black	62.1	63.8	98.7	95.9	67.5	71.1	98.4	96.4	63.3	75.2	98.2	96.8
Bracket Brown	60.1	75.1	94.9	89.9	64.3	74.1	95.5	91.4	68.1	80.8	96.4	92.3
Bracket White	74.8	72.5	99.6	98.4	71.2	68.0	99.4	97.0	85.1	79.9	99.5	97.1
Connector	75.5	59.2	97.2	90.8	93.6	89.3	98.2	94.1	96.2	92.7	98.5	95.1
Metal Plate	99.7	99.9	98.1	94.8	99.7	99.9	98.4	95.5	99.2	99.7	98.6	96.2
Tubes	99.9	99.9	99.5	98.2	98.6	99.5	99.5	98.2	98.9	99.5	99.5	98.2
Mean	**78.7**	**78.4**	**98.0**	**94.7**	**82.5**	**83.6**	**98.2**	**95.4**	**85.1**	**88.0**	**98.5**	**95.9**

number of active concepts is sufficient to represent normal features while avoiding unnecessarily dense reconstructions.

The dictionary expansion factor exhibits a similar saturation trend. Increasing the expansion factor from $8\times$ to $16\times$ improves the image AUROC from 99.0% to 99.1%, while the pixel-level metrics remain unchanged. Further increasing the capacity to $32\times$ does not provide additional benefits. We therefore select $K = 32$ and a $16\times$ expansion factor as a balanced configuration that provides sufficient representational capacity without increasing the optimization difficulty under limited support data.

Table A4: SAE Top-K ablation on MVTec-AD under the 2-shot setting.

	Image-level		Pixel-level	
Top-K	AUROC	AUPR	AUROC	AUPRO
8	98.0	99.2	97.1	95.9
16	98.4	99.4	97.4	96.2
32	**98.7**	**99.5**	**97.5**	**96.2**
64	**98.7**	**99.5**	**97.5**	**96.2**

Table A5: SAE dictionary expansion ablation on MVTec-AD under the 4-shot setting.

	Image-level		Pixel-level	
Factor	AUROC	AUPR	AUROC	AUPRO
$8\times$	99.0	**99.6**	**97.8**	**96.4**
$16\times$	**99.1**	**99.6**	**97.8**	**96.4**
$32\times$	99.0	**99.6**	**97.8**	**96.4**

Temperature parameter. Table A6 evaluates the temperature parameter τ used in the normality weighting function. Very small values of τ lead to more selective weighting and can restrict adaptation to only a small subset of patches. In contrast, overly large values distribute substantial weights to less reliable regions and may allow abnormal patterns to influence the adaptation process. The results show a relatively stable performance plateau between $\tau = 1.0$ and $\tau = 2.0$. In particular, $\tau = 1.5$ provides a strong compromise between image-level detection and pixel-level localization across the different shot settings. We therefore use $\tau = 1.5$ in all experiments.

Table A6: Ablation study of the temperature parameter τ on MVTec-AD. Best results are determined using the original unrounded values.

τ	1-shot				2-shot				4-shot			
	Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO
0.5	97.6	98.9	**97.4**	95.1	98.1	99.3	**97.8**	95.8	99.0	99.6	**98.1**	96.3
1.0	**98.4**	**99.4**	97.1	95.6	98.6	99.4	97.4	95.9	**99.2**	**99.7**	97.7	96.2
1.5	**98.4**	**99.4**	97.2	**95.9**	**98.7**	**99.5**	97.5	**96.3**	99.1	**99.7**	97.8	**96.5**
2.0	98.0	99.2	97.1	**95.9**	98.4	99.4	97.5	**96.3**	98.9	99.6	97.8	**96.5**
2.5	97.8	99.1	97.0	95.6	98.3	99.4	97.4	96.1	98.9	99.6	97.8	96.3
3.0	97.5	99.0	96.9	95.4	98.2	99.3	97.4	95.9	98.8	99.5	97.7	96.2

Input resolution. Table A7 analyzes the effect of input resolution. Increasing the resolution generally improves localization, since more spatial details are preserved in the extracted patch features. On MVTec-AD, the pixel AUROC and PRO reach their best values at 672px in most shot settings, whereas the image-level metrics largely plateau around 512px. A similar trend is observed on VisA: image-level classification improves rapidly at lower resolutions, while pixel-level localization continues to benefit from the higher spatial resolution. We therefore adopt 672px as the default resolution because localization is the more resolution-sensitive component of the task.

Table A7: Ablation study of the input resolution on MVTec-AD and VisA. The best displayed results are highlighted in bold.

Dataset	Resolution	1-shot				2-shot				4-shot			
		Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO
MVTec-AD	256px	97.1	98.9	96.4	93.3	97.5	99.1	96.8	93.8	98.3	99.4	97.3	94.4
	336px	98.2	99.3	96.6	94.6	98.5	99.4	97.1	95.1	99.0	99.6	97.6	95.6
	448px	98.7	99.4	96.8	95.2	99.0	99.6	97.3	95.8	99.3	**99.7**	97.7	96.2
	512px	**98.8**	**99.5**	96.9	95.1	**99.1**	**99.7**	97.3	95.6	**99.4**	**99.7**	97.7	95.9
	672px	98.4	99.3	**97.2**	**95.9**	98.7	99.5	**97.5**	**96.3**	99.1	**99.7**	**97.8**	**96.5**
VisA	256px	94.1	93.9	98.1	88.0	94.3	94.6	98.5	90.4	95.7	95.2	98.8	91.7
	336px	94.6	**94.7**	98.6	92.3	94.9	95.0	98.9	93.8	95.7	95.2	99.0	94.4
	448px	94.7	94.1	98.8	94.8	95.0	**95.1**	**99.0**	95.5	96.0	**95.4**	**99.1**	95.9
	512px	**95.1**	94.3	98.7	93.9	**95.4**	**95.1**	98.9	94.8	**96.1**	**95.4**	99.0	95.2
	672px	94.8	94.4	**98.9**	**95.9**	95.3	94.9	**99.0**	**96.3**	95.6	95.1	**99.1**	**96.6**

Aggregation degree $\mathcal{R}$. Table A8 evaluates different combinations of spatial aggregation degrees. Adding $r = 3$ to the single-scale configuration $\{1\}$ generally improves localization while preserving image-level performance. For example, under the 2-shot setting, the pixel AUROC on MVTec-AD improves from 97.0% to 97.5%, and the PRO score improves from 95.8% to 96.3%. Similar improvements are observed on VisA and MPDD.

Table A8: Ablation study of the Aggregation degree set $\mathcal{R}$ on MVTec-AD, VisA, and MPDD. The best displayed results are highlighted in bold.

Dataset	$\mathcal{R}$	1-shot				2-shot				4-shot			
		Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO
MVTec-AD	$\{1\}$	**98.4**	**99.3**	96.6	95.4	**98.8**	**99.5**	97.0	95.8	**99.2**	**99.7**	97.5	96.2
	$\{1,3\}$	**98.4**	**99.3**	97.2	95.9	98.7	**99.5**	97.5	**96.3**	99.1	**99.7**	97.8	**96.5**
	$\{1,3,5\}$	98.1	99.2	97.4	**96.0**	98.4	99.4	97.7	**96.3**	99.0	99.6	97.9	**96.5**
	$\{1,3,5,7\}$	97.6	99.0	**97.5**	**96.0**	97.9	99.1	**97.8**	96.2	98.7	99.5	**98.0**	96.4
VisA	$\{1\}$	**95.1**	94.2	98.6	**95.9**	95.1	94.3	98.8	**96.4**	**96.9**	**96.5**	98.9	**96.7**
	$\{1,3\}$	94.8	**94.4**	98.9	**95.9**	**95.3**	**94.9**	99.0	96.3	95.6	95.1	99.1	96.6
	$\{1,3,5\}$	94.6	94.3	**99.0**	95.5	95.0	94.5	**99.1**	96.1	94.3	93.6	**99.2**	96.3
	$\{1,3,5,7\}$	94.1	93.6	**99.0**	95.1	94.3	93.9	**99.1**	95.6	93.7	93.0	**99.2**	96.0
MPDD	$\{1\}$	77.4	75.9	97.8	94.0	**82.5**	**83.8**	98.1	95.2	**85.3**	87.2	**98.5**	**96.1**
	$\{1,3\}$	**78.7**	**78.4**	**98.0**	**94.7**	**82.5**	83.6	**98.2**	**95.4**	85.1	**88.0**	**98.5**	95.9
	$\{1,3,5\}$	77.4	76.8	**98.0**	94.5	80.6	81.2	98.0	95.0	83.9	85.3	98.4	95.6
	$\{1,3,5,7\}$	75.6	74.8	97.9	94.3	78.7	76.5	98.0	95.0	81.2	80.7	98.2	95.2

Normality-weighting exponent. A9 study the exponent λ in the normality weighting function. When $\lambda = 0$, all patches receive equal weights during test-time adaptation. This setting causes a substantial performance drop, particularly on MPDD, where the 2-shot image AUROC and AUPR decrease to 65.5% and 69.6%, respectively. This result confirms that treating all patches equally can cause anomalous regions to contaminate the learned normality representation.

As λ increases, the influence of potentially abnormal patches is reduced and the performance improves. On MPDD, the image AUROC increases from 65.5% at $\lambda = 0$ to 83.0% at $\lambda = 6$. However, excessively large values do not consistently improve pixel-level localization and may slightly reduce the reconstruction quality of difficult normal regions. On MVTec-AD and VisA, the performance remains relatively stable within a moderate range. We therefore select $\lambda = 4$, which provides a robust balance between image-level detection and pixel-level localization.

A.4 Feature and Backbone Analysis

Feature aggregation strategy. Table **??** compares mean pooling over intermediate layers, mean pooling over final layers, and using only the last layer. Using only the last layer consistently produces inferior results, especially on VisA, where the image AUROC/AUPR decreases to 89.0%/91.3%. This confirms that the last layer alone may lose fine-grained structural information that is important for localizing small defects.

Aggregating multiple intermediate layers provides a better balance between semantic abstraction and spatial detail. The middle-layer aggregation achieves the best overall performance on MVTec-AD and VisA, reaching 99.1%/99.6% image-level performance and 97.8%/96.4% pixel-level performance on MVTec-AD. On MPDD, the final-layer aggregation performs slightly better, suggesting

Table A9: Ablation study of the exponent λ on MVTec-AD and VisA. The best displayed results are highlighted in bold.

Dataset	λ	1-shot				2-shot				4-shot			
		Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO
MVTec-AD	0	95.8	97.9	96.8	94.6	97.0	98.5	97.3	95.3	98.2	99.2	97.6	95.6
	1	96.8	98.6	96.9	95.1	97.7	99.0	97.4	95.6	98.6	99.4	97.7	95.9
	2	98.0	99.2	97.1	**95.9**	98.4	99.4	**97.5**	**96.3**	98.9	99.6	**97.8**	**96.5**
	4	**98.4**	99.3	**97.2**	**95.9**	98.7	**99.5**	**97.5**	**96.3**	99.1	**99.7**	**97.8**	**96.5**
	6	**98.4**	**99.4**	97.1	95.8	**98.8**	**99.5**	**97.5**	96.2	**99.2**	**99.7**	**97.8**	96.4
	8	**98.4**	**99.4**	97.1	95.8	**98.8**	**99.5**	**97.5**	96.1	**99.2**	**99.7**	97.7	96.3
VisA	0	94.3	94.2	98.7	94.9	95.1	**94.9**	98.9	95.5	**95.9**	**95.6**	99.0	95.8
	1	94.7	94.3	98.8	95.2	95.2	94.1	98.9	95.7	95.7	95.3	99.0	96.0
	2	**94.8**	94.2	**98.9**	95.8	95.3	94.8	**99.0**	96.2	**95.9**	95.5	**99.1**	96.5
	4	**94.8**	**94.4**	**98.9**	**95.9**	95.3	**94.9**	**99.0**	**96.3**	95.6	95.1	**99.1**	**96.6**
	6	94.6	95.7	**98.9**	95.8	**95.6**	94.5	**99.0**	**96.3**	95.3	94.8	**99.1**	**96.6**
	8	94.2	95.4	**98.9**	95.8	95.4	94.3	**99.0**	96.2	95.2	94.6	**99.1**	**96.6**

that the optimal feature depth can depend on the complexity and visual characteristics of the dataset. Nevertheless, intermediate-layer aggregation provides the most consistent performance across the evaluated benchmarks and is used as the default strategy.

DINOv2 backbone capacity. Table A10 evaluates four DINOv2 backbones with different capacities. Overall, larger backbones provide stronger representations and improve both image-level detection and pixel-level localization. DINOv2-G achieves the best image-level results across all settings on both MVTec-AD and VisA. For example, on VisA, it improves the image AUROC from 90.2% with DINOv2-S to 94.8% with DINOv2-G in the 1-shot setting.

The improvement is not perfectly monotonic for every individual metric. For example, DINOv2-L slightly outperforms DINOv2-G on some MVTec-AD pixel-level metrics in the 4-shot setting. Nevertheless, DINOv2-G provides the strongest and most consistent image-level performance, while maintaining competitive localization accuracy. We therefore use DINOv2-G as the default backbone.

A.5 Interpretability and Efficiency

Interpretability of learned concepts. Figure A1 visualizes representative concepts learned by the EvoSAE. The purpose of this visualization is to examine whether the learned concepts correspond to recurring normal patterns rather than isolated noise. The concept-level activations provide a qualitative explanation of how the sparse dictionary represents normal structures at different spatial scales. These visualizations should be interpreted as qualitative evidence: they support the interpretability of the learned representation, but do not constitute an independent quantitative evaluation.

Runtime analysis. Table A11 compares the per-image runtime and detection performance on MVTec-AD. The frozen SAE requires 471 ms per image, which is slightly slower than AnomalyDINO and SubspaceAD, at 375 ms and 440 ms, respectively. Incorporating the test-time adaptation procedure without counting its additional optimization time increases the reported runtime to 494 ms, while improving the image-level AUROC/AUPR from 98.9%/99.6% to 99.1%/99.7%.

When the actual test-time adaptation cost is included, the runtime increases to 1,009 ms per image, approximately twice the cost of the corresponding non-adaptive setting. At the reported precision, the image-level and pixel-level metrics remain unchanged after including this additional time. This result highlights an explicit accuracy–efficiency trade-off: the adaptation mechanism improves the reconstruction-based representation, but its computational cost should be considered in latency-sensitive applications.

Table A10: Ablation study of different DINOv2 backbones on MVTec-AD and VisA. The best displayed results are highlighted in bold.

Dataset	Backbone	1-shot				2-shot				4-shot			
		Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO	Img AUROC	Img AUPR	Pxl AUROC	PRO
MVTec-AD	DINOv2-S	95.6	97.8	96.4	94.4	96.9	98.4	97.1	95.3	97.5	98.9	97.6	95.9
	DINOv2-B	96.1	97.8	96.2	93.4	97.2	98.3	96.8	94.1	98.3	99.0	97.2	94.5
	DINOv2-L	96.5	97.5	97.0	95.8	97.3	98.1	**97.5**	**96.3**	98.4	99.2	**97.9**	**96.6**
	DINOv2-G	**98.4**	**99.3**	**97.2**	**95.9**	**98.7**	**99.5**	**97.5**	**96.3**	**99.1**	**99.7**	97.8	96.5
VisA	DINOv2-S	90.2	89.0	97.9	92.8	90.5	89.6	98.2	93.5	90.9	89.1	98.4	94.1
	DINOv2-B	91.4	90.6	97.9	93.1	91.7	91.0	98.1	93.6	91.8	91.1	98.3	94.0
	DINOv2-L	93.7	92.9	98.6	95.6	94.3	93.5	98.8	96.1	94.4	93.4	98.9	96.5
	DINOv2-G	**94.8**	**94.4**	**98.9**	**95.9**	**95.3**	**94.9**	**99.0**	**96.3**	**95.6**	**95.1**	**99.1**	**96.6**

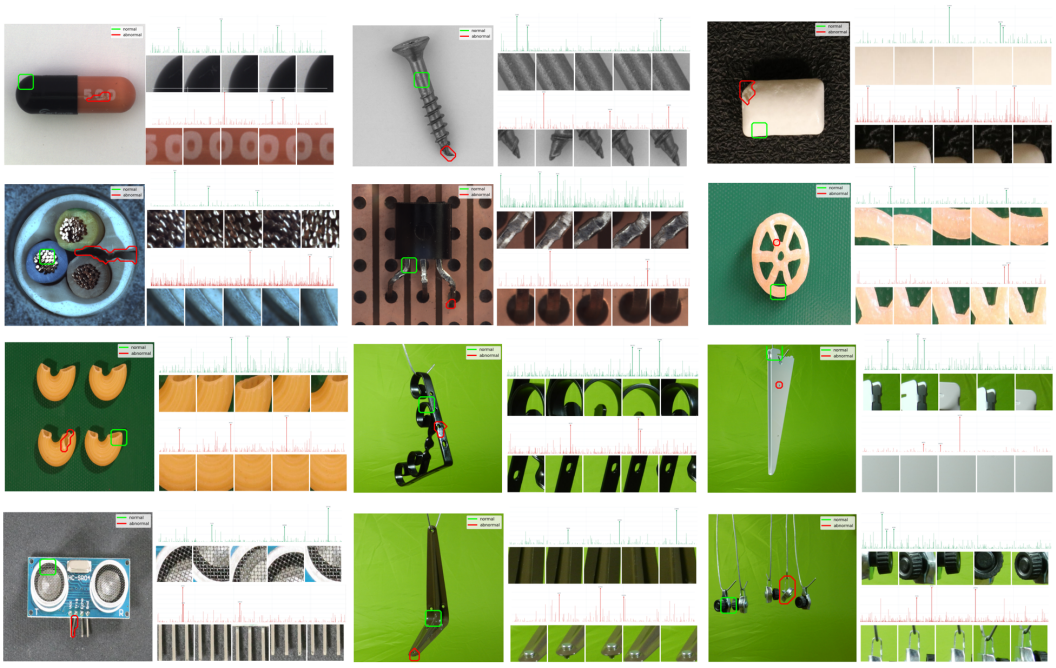


Figure A1: Interpretability of EvoSAE concept.

Table A11: Per-image runtime and 4-shot anomaly detection performance on MVTec-AD using DINOv2-G with an input resolution of 672px.

Method	Time (ms)	Image AUROC/AUPR	Pixel AUROC/PRO
AnomalyDINO	375	97.7/98.7	97.2/93.4
SubspaceAD	440	98.4/99.3	98.0/94.2
Ours (Frozen SAE)	471	98.9/99.6	**98.1**/96.2
Ours (w/o TTT time)	494	**99.1**/**99.7**	97.8/**96.5**
Ours (w/ TTT time)	1,009	**99.1**/**99.7**	97.8/**96.5**