ICASSP 2026

AD-DINOv3: Enhancing DINOv3 for Zero-Shot Anomaly Detection with Anomaly-Aware Calibration

Jingyi Yuan, Jianxiong Ye, Wenkang Chen, Chenqiang Gao*
School of Intelligent Systems Engineering, Sun Yat-Sen University
{yuanjy36, yejx55, chenwk25} @mail2.sysu.edu.cn
*gaochq6@mail.sysu.edu.cn

Abstract

Zero-Shot Anomaly Detection (ZSAD) seeks to identify anomalies from arbitrary novel categories, offering a scalable and annotation-efficient solution. Traditionally, most ZSAD works have been based on the CLIP model, which performs anomaly detection by calculating the similarity between visual and text embeddings. Recently, vision foundation models such as DINOv3 have demonstrated strong transferable representation capabilities. In this work, we are the first to adapt DINOv3 for ZSAD. However, this adaptation presents two key challenges: (i) the domain bias between large-scale pretraining data and anomaly detection tasks leads to feature misalignment; and (ii) the inherent bias toward global semantics in pretrained representations often leads to subtle anomalies being misinterpreted as part of the normal foreground objects, rather than being distinguished as abnormal regions. To overcome these challenges, we introduce AD-DINOv3, a novel vision-language multimodal framework designed for ZSAD. Specifically, we formulate anomaly detection as a multimodal contrastive learning problem, where DINOv3 is employed as the visual backbone to extract patch tokens and a CLS token, and the CLIP text encoder provides embeddings for both normal and abnormal prompts. To bridge the domain gap, lightweight adapters are introduced in both modalities, enabling their representations to be recalibrated for the anomaly detection task. Beyond this baseline alignment, we further design an Anomaly-Aware Calibration Module (AACM), which explicitly guides the CLS token to attend to anomalous regions rather than generic foreground semantics, thereby enhancing discriminability. Finally, anomaly localization are achieved by measuring the similarity between adapted visual features and prompt embeddings. Extensive experiments on eight industrial and medical benchmarks demonstrate that AD-DINOv3 consistently matches or surpasses state-of-the-art methods, verifying its effectiveness and broad applicability as a general zero-shot anomaly detection framework. The code will be available at https://github.com/Kaisor-Yuan/AD-DINOv3

1 Introduction

dinov3（Meta）：
·强大的迁移能力：在大规模自然图像进行自监督训练
·更具判别性特征：相比于CLIP模型，dinov3视觉特征在判别正常异常区域更具有潜力

Anomaly detection plays a vital role in various real-world applications, such as industrial quality inspection, medical image analysis, and security monitoring [27]. Traditional supervised anomaly detection methods typically rely on large amounts of labeled abnormal samples for each category, which is often impractical in large-scale or dynamic environments [4, 1, 3]. In contrast, Unsupervised Anomaly Detection (UAD) aims to identify anomalies in images with only normal samples for training without requiring any anomalous samples [30, 11, 13]. Nevertheless, it remains challenging to obtain plenty of normal images for training.

*Corresponding Author

·领域偏差：预训练自然图像与工业异常检测任务存在很大差异
·全局语义偏差：dinov3倾向于关注物体整体语义而非局部缺陷

Traditional unsupervised anomaly detection methods rely on modeling the distribution of normal samples, which makes them sensitive to domain shifts and requires retraining whenever a new product category is introduced. This limitation has motivated a growing interest in Zero Shot Anomaly Detection (ZSAD), which aims to identify anomalies in previously unseen objects without using any target-dataset samples during training. In practice, ZSAD leverages the transferability of large-scale pretrained models to generalize across categories and domains. Among them, pre-trained vision–language models (VLMs) have become particularly popular due to their strong cross-modal alignment and generalization capability. In particular, CLIP [28], with its powerful image–text alignment, has inspired a series of ZSAD approaches [7, 44, 9, 29, 42]. These methods typically construct anomaly maps by computing the cosine similarity between image patch features and textual prompts, thereby enabling pixel-level localization of anomalous regions.

In recent, vision foundation models pretrained on billions of natural images have recently shown excellent cross-task transferability. To date, most ZSAD methods still rely on vision–language models such as CLIP [28], exploiting their image–text alignment to detect anomalies. In comparison, self-supervised visual encoders like DINOv3 [35] remain largely unexplored for this task. However, the direct application of DINOv3 to ZSAD faces two key challenges: (i) the domain bias between large-scale pretraining data and anomaly detection tasks leads to feature misalignment, limiting the direct applicability of pretrained representations to anomaly detection; and (ii) the learned representations tend to emphasize global object semantics while overlooking subtle defect cues, causing anomalies to be treated as background noise rather than recognized as abnormal regions. Fig. 1 illustrates the differences between original DINOv3 and our proposed AD-DINOv3. For DINOv3, the similarity maps show two major issues. First, when attending to normal regions (upper row), the responses incorrectly extend to anomalous areas, indicating a semantic misalignment where anomalies are regarded as part of normal regions. Second, when attending to anomalous regions (lower row), the responses not only include spurious activations in normal areas but also exhibit insufficient contrast, preventing anomalies from standing out as coherent and distinctive clusters.

In this work, we propose AD-DINOv3, a novel vision–language multimodal framework and the first to adapt DINOv3 to the ZSAD task. Our approach leverages DINOv3 as the visual backbone to extract both patch tokens and CLS token enabling unified anomaly localization. To fully exploit the hierarchical representations of DINOv3, we introduce a multi-level feature adaptation strategy with lightweight adapters, preserving complementary low- and high-level cues. Moreover, we design a novel Anomaly-Aware Calibration Module (AACM), which explicitly guides the CLS token to attend to anomalous regions rather than generic foreground semantics. In parallel, the text branch employs the CLIP text encoder to represent both normal and abnormal prompts, which are further aligned to the target domain via adapters, performing cross modal contrastive learning with visual branch. Finally, anomaly detection is realized by comparing visual and textual representations: patch tokens against prompt embeddings for pixel-level anomaly maps. As shown in Fig. 1, compared with the original DINOv3, our proposed AD-DINOv3 produces cleaner responses: normal regions are more compact and free from spurious activations, while anomalous regions appear more prominent and coherent. Overall, the separation between normal and abnormal regions is significantly improved.

Our main contributions include the following:

1. We present AD-DINOv3, the first framework that adapts DINOv3 to zero-shot anomaly detection, bridging the gap between self-supervised visual encoders and anomaly detection tasks.

2. We introduce a Cross Modal Contrastive Learning (CMCL) strategy with lightweight adapters to fully exploit hierarchical representations of DINOv3 for zero shot anomaly detection.

3. We design a novel Anomaly-Aware Calibration Module (AACM), which explicitly guides the CLS token to focus on anomalous regions, alleviating its bias toward generic object semantics.

4. Our method surpasses or matches state-of-the-art methods on eight industrial and medical benchmarks, underscoring its effectiveness as a general ZSAD framework.

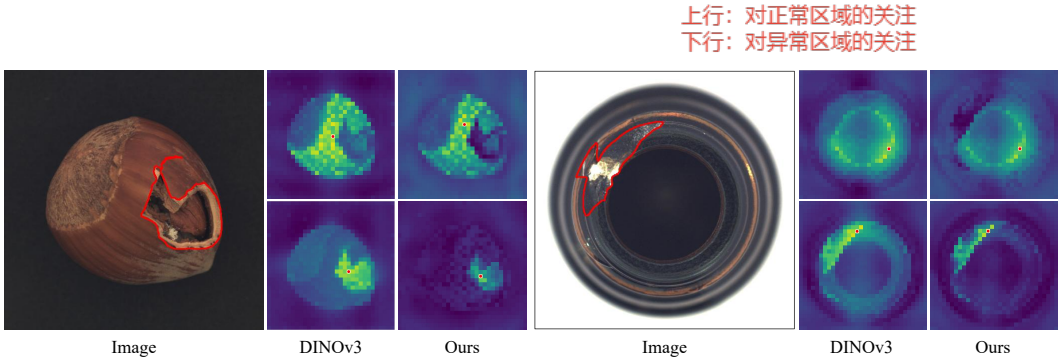

Figure 1: Cosine similarity maps between the red-dotted patch and all other patches. The left shows the ground truth, the center corresponds to original DINOv3, and the right to our proposed AD-DINOv3. The upper and lower rows illustrate attention to normal regions and anomalous regions, respectively. Compared with original DINOv3, AD-DINOv3 reduces spurious responses in normal regions, highlights anomalous areas more prominently and coherently, and enhances the separation between normal and abnormal regions.

2 Related Work

2.1 Traditional Anomaly Detection

Traditional anomaly detection methods in computer vision typically model the distribution of normal data and treat deviations as potential anomalies. Reconstruction-based methods, such as Autoencoders and Variational Autoencoders [17, 22], attempt to reconstruct input images and use reconstruction errors to identify anomalous regions. GAN-based models further extend this idea by leveraging adversarial training to synthesize normal patterns and detect deviations [1]. Another line of work focuses on feature embedding. One-Class SVM and its deep extensions aim to learn compact decision boundaries around normal samples [33, 32]. Memory-augmented approaches such as MemAE introduced external memory modules to enhance the reconstruction ability of normal data [14]. Industrial anomaly detection has also benefited from representation learning on large datasets. PatchCore [31] proposed a nearest-neighbor search strategy on features extracted from pretrained networks, achieving state-of-the-art performance on MVTec AD [4]. Despite their success, these methods rely heavily on abundant and clean normal data; in scenarios where data is noisy or distributions shift, their generalization ability is significantly limited.

2.2 CLIP-based Anomaly Detection

With the introduction of vision-language models such as CLIP, anomaly detection has been extended to a zero-shot setting. CLIP learns transferable representations by aligning image and text embeddings, enabling models to generalize to unseen categories without task-specific training. Early attempts such as WinCLIP [18] explored the use of CLIP for anomaly detection by aggregating multi-scale features and aligning them with text prompts describing normal classes. Other works refined this idea by incorporating prompt learning. For instance, CoOp and its variants [43] introduced learnable textual prompts, which were later adapted for anomaly detection to improve feature alignment. Similarly, AnomalyCLIP [39] integrated multi-prompt ensembles to better capture diverse anomaly types. More recent works introduced rectification mechanisms to address CLIP's limitations in distinguishing subtle abnormal cues. For example, AFR-CLIP [40] proposed a stateless-to-stateful anomaly feature rectification strategy, embedding defect-related information into textual prompts to improve anomaly localization. Other methods such as PromptAD [16] employed domain-specific prompt adaptation to reduce the semantic gap between textual and visual features.

While these advances highlight the potential of CLIP for anomaly detection, challenges remain in capturing fine-grained anomalies under complex backgrounds and ensuring robust performance across unseen domains. This motivates the exploration of stronger visual backbones DINOv3 and adaptive prompt-learning mechanisms tailored for anomaly detection.

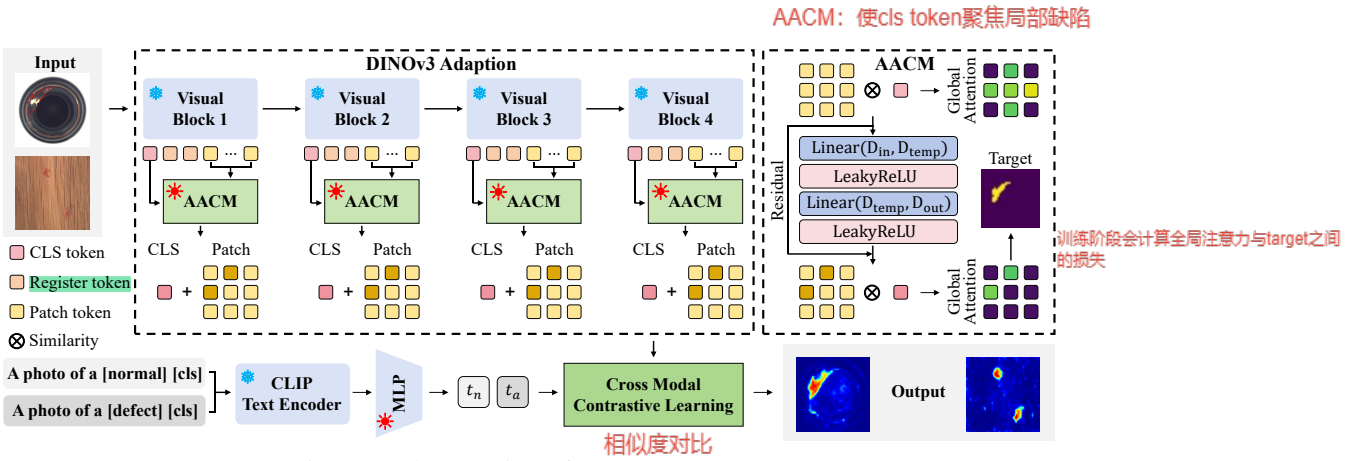

Figure 2: The overview of our AD-DINOv3.

3 Method

3.1 Problem Definition

Given a test sample $I \in \mathbb{R}^{H\times W\times 3}$, the objective of zero-shot anomaly detection (ZSAD) is to generate a pixel-wise anomaly map $M \in [0,1]^{H\times W}$. In line with existing ZSAD frameworks, we utilize an auxiliary dataset $D_a = \{(I_1, G_1), \ldots, (I_n, G_n) \mid G_i \in [0,1]^{H\times W}\}$, which contains both normal and anomalous samples $\{I_i\}_{i=1}^n$ along with their ground-truth masks $\{G_i\}_{i=1}^n$. The model is subsequently evaluated on a disjoint test dataset $D_t = \{I_{t_1}, \ldots, I_{t_m}\}$, comprising unseen images from a different benchmark or domain. Importantly, to guarantee the zero-shot setting, the auxiliary and test datasets must be non-overlapping, i.e., $D_a \cap D_t = \emptyset$.

3.2 Overview

We employ DINOv3 as the visual backbone of AD-DINOv3. As illustrated in Fig. 2, the image branch extracts patch tokens and a CLS token, which are refined by lightweight adapters together with an Anomaly-Aware Calibration Module (AACM). The AACM explicitly guides the CLS token to focus on anomalous patch regions under mask supervision, redirecting its global attention from generic foreground objects, as favored in natural image pretraining, toward abnormalities. In parallel, the text branch encodes both normal and abnormal prompts (e.g., "a photo of a [state] [class]") and is likewise refined by adapters to better align with the target domain. For anomaly localization, patch tokens are compared with the prompt embeddings to generate pixel-level anomaly maps.

3.3 Cross Modal Contrastive Learning (CMCL)

Vision-language models pretrained on large-scale natural images primarily capture high-level semantics, such as object categories or foreground–background separation. While such priors are beneficial for many recognition tasks, they are not ideal for anomaly detection, which instead demands fine-grained discrimination between normal and abnormal regions that may differ only in subtle texture or local structure. Directly applying pretrained features to this task often results in weak separability: patches covering anomalous defects may still appear highly similar to normal counterparts, and textual embeddings are not explicitly aligned with anomaly semantics.

To mitigate this issue, we introduce lightweight adapters for both the visual and textual branches. Each adapter is implemented as a bottleneck MLP with two linear layers and LeakyReLU activations, which recalibrates representations toward the anomaly detection domain while keeping the powerful pretrained backbones frozen. We denote this transformation as $\mathrm{Ada}(\cdot)$.

Formally, given an image, the visual encoder yields patch tokens $\{p_i\}_{i=1}^N$ and a CLS token c, while the text encoder provides embeddings for both normal and abnormal prompts $t \in \mathbb{R}^{d\times 2}$. The adapted representations are written as

$$\{\tilde{p}_i, \tilde{c}, \tilde{t}\} = \mathrm{Ada}(\{p_i, c, t\}). \quad (1)$$

We then formulate pixel-level localization as a cross modal contrastive alignment task. For the adapted visual patch tokens $\tilde{p}_i$, its similarity to text embeddings is computed as

$$s = \frac{\tilde{p}_i^\top \tilde{t}}{\|\tilde{p}_i\|\|\tilde{t}\|}, \quad p = \sigma(s), \tag{2}$$

where $p \in \mathbb{R}^{N\times 2}$ provides normal/abnormal probabilities and σ means softmax function over the class dimension. At the patch level, the abnormal probabilities are rearranged into a coarse map and bilinearly upsampled to the original image resolution, yielding the pixel-level anomaly map $\hat{S} \in \mathbb{R}^{H\times W}$.

3.4 Anomaly-Aware Calibration Module (AACM)

Although the CLS token aggregates global context by design, it is heavily biased toward generic foreground objects due to natural-image pretraining. This bias causes the model to confuse anomalies with salient object regions, leading to poor perception to subtle defects. To address this limitation, we propose the Anomaly-Aware Calibration Module (AACM), which introduces a mask-guided training objective that explicitly encourages the CLS token and its interaction with patch tokens to emphasize anomalous regions.

Specifically, for each patch i, the similarity to the adapted CLS token is given by

$$s_i = \frac{\tilde{c}^\top \tilde{p}_i}{\|\tilde{c}\|\|\tilde{p}_i\|}. \tag{3}$$

The similarity distribution $\sigma(\{s_i\})$ is then optimized against the anomaly mask M using a combination of focal and Dice losses:

$$\mathcal{L}_{\text{AACM}} = \mathcal{L}_{\text{focal}}(\sigma(\{s_i\}), M) + \mathcal{L}_{\text{dice}}(\sigma(\{s_i\}), M). \tag{4}$$

Here, the focal loss [24] emphasizes hard-to-classify anomalous patches, while the Dice loss [25] improves spatial consistency and alleviates the severe imbalance between normal and anomalous pixels.

Through this objective, the CLS token is gradually guided to allocate more of its global attention to anomalous regions, which in turn reshapes the feature space of patch tokens associated with those regions. In effect, AACM enhances anomaly awareness across both global and local representations, while remaining lightweight.

3.5 Training and Inference

Training. The cross-modal alignment loss is computed by comparing the cross-modal anomaly map P against the ground-truth mask M. Formally,

$$\mathcal{L}_{\text{CM}} = \mathcal{L}_{\text{focal}}(P, M) + \mathcal{L}_{\text{dice}}(P, M). \tag{5}$$

The complete training objective combines the cross-modal alignment loss with the AACM regularization:

$$\mathcal{L} = \lambda_{\text{CM}}\mathcal{L}_{\text{CM}} + \lambda_{\text{AACM}}\mathcal{L}_{\text{AACM}}, \tag{6}$$

where λ_{CM} and λ_{AACM} are balancing weights.

Testing. At inference time, anomaly masks are unavailable. Detection therefore relies solely on cross-modal similarity: patch–text similarities yield pixel-level anomaly maps. Since AACM is used only during training, our framework maintains the same efficiency as the baseline model at test time.

4 Experiments

4.1 Experiments Setup

Datasets. To comprehensively assess the effectiveness of AD-DINOv3, we perform experiments in both industrial and medical scenarios. For the industrial setting, we adopt the MVTec AD [5], VisA [45], BTAD [26], and MPDD [20] benchmarks, while in the medical setting we evaluate on

ISIC [10], ColonDB [36], ClinicDB [6], and TN3K [15]. Since the object categories in VisA are distinct from those in the other datasets, we follow a cross-dataset evaluation strategy: model trained on VisA is tested on the remaining datasets, and conversely, we utilize the model trained on MVTec AD for VisA testing.

Evaluating Metrics. Consistent with prior works in zero-shot anomaly detection, we adopt several evaluation metrics for anomaly localization. Specifically, pixel-level anomaly localization is assessed by the Area Under the Receiver Operating Characteristic Curve (AUROC), which measures the discriminative capability and the balance between precision and recall. In addition, to ensure fair comparison with existing ZSAD approaches, we include the maximum F1 score (F1). Finally, to provide a holistic assessment, we also summarize the mean performance across all domains.

Implementation Details. In our experiments, we employ the pretrained DINOv3 with a ViT-L/16 backbone released by Meta AI as the default image encoder, while the text encoder from pretrained **CLIP** (OpenAI) is used to generate textual embeddings. All input images are uniformly resized to 512×512 for both training and inference. The DINOv3 backbone contains 24 transformer layers, which we partition into four stages, extracting patch embeddings from the 6th, 12th, 18th, and 24th layers, respectively. Model training is performed for 10 epochs with a batch size of 64, optimized using Adam. The learning rate is initialized at 1×10^{-4}. All experiments are conducted on a single NVIDIA RTX A6000 GPU (48 GB).

Comparison Methods. For a comprehensive comparison, we benchmark AD-DINOv3 against representative approaches from two distinct paradigms: those that operate in a training-free manner and those that rely on auxiliary datasets. Within the first paradigm, we consider WinCLIP [19], a pioneering CLIP-based framework that dispenses with any additional training data. Within the second paradigm, we examine APRIL-GAN [8], AdaCLIP [7], and AnomalyCLIP [44], which are trained on external auxiliary data and subsequently evaluated on unseen object categories.

4.2 Comparison with State-of-the-Art Methods

Comparison on industrial image benchmarks.

As shown in Tab. 1, AD-DINOv3 achieves state-of-the-art performance on industrial benchmarks. On VisA, our method reaches an AUROC of 95.6% and an F1 score of 37.7%, surpassing previous methods including WinCLIP (79.6% / 14.8%) and AnomalyCLIP (95.4% / 28.3%). Similarly, on BTAD and MPDD, our approach yields AUROC scores of 93.5% and 96.2%, respectively, consistently outperforming AdaCLIP and APRILGAN. On the widely-used MVTec AD, AD-DINOv3 achieves 91.6% AUROC and 50.1% F1, setting a new benchmark.

When averaged across all industrial datasets, AD-DINOv3 attains an AUROC of 94.2% and an F1 score of 44.6%, clearly exceeding all competing methods. This demonstrates the advantage of

Table 1: Pixel-level performance comparisons of the ZSAD methods on the industrial and medical datasets. The best performance is in **bold** and the second-best is underlined.

Domain	Dataset	WinCLIP		APRILGAN		AnomalyCLIP		AdaCLIP		**Ours**	
		AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1
Industrial	VisA	(79.6 ,	14.8)	(95.2 ,	32.3)	(95.4 ,	28.3)	(95.5 ,	37.7)	(**95.6** ,	**37.7**)
	BTAD	(72.6 ,	18.5)	(89.5 ,	38.4)	(**94.0** ,	49.7)	(92.1 ,	51.7)	(93.5 ,	**51.7**)
	MPDD	(71.2 ,	15.4)	(95.1 ,	30.6)	(**96.4** ,	34.2)	(96.1 ,	34.9)	(96.2 ,	**38.8**)
	MVTec AD	(85.1 ,	31.6)	(83.7 ,	39.8)	(90.9 ,	39.1)	(88.7 ,	43.4)	(**91.6** ,	**50.1**)
	Average	(77.1 ,	20.1)	(90.9 ,	35.3)	(94.2 ,	37.5)	(93.1 ,	41.9)	(**94.2** ,	**44.6**)
Medical	ISIC	(83.3 ,	48.5)	(85.8 ,	67.3)	(89.7 ,	70.6)	(**90.3** ,	**72.6**)	(89.0 ,	72.1)
	ColonDB	(70.3 ,	19.6)	(78.4 ,	33.2)	(81.9 ,	37.3)	(**82.6** ,	36.1)	(80.8 ,	**38.4**)
	ClinicDB	(51.2 ,	24.4)	(83.2 ,	42.3)	(82.9 ,	42.1)	(82.8 ,	40.9)	(**90.4** ,	**54.3**)
	TN3K	(70.7 ,	30.0)	(74.4 ,	39.7)	(**81.5** ,	**47.9**)	(76.8 ,	40.7)	(77.8 ,	41.1)
	Average	(68.9 ,	30.6)	(80.5 ,	45.6)	(84.0 ,	49.5)	(83.1 ,	47.6)	(**84.5** ,	**51.5**)

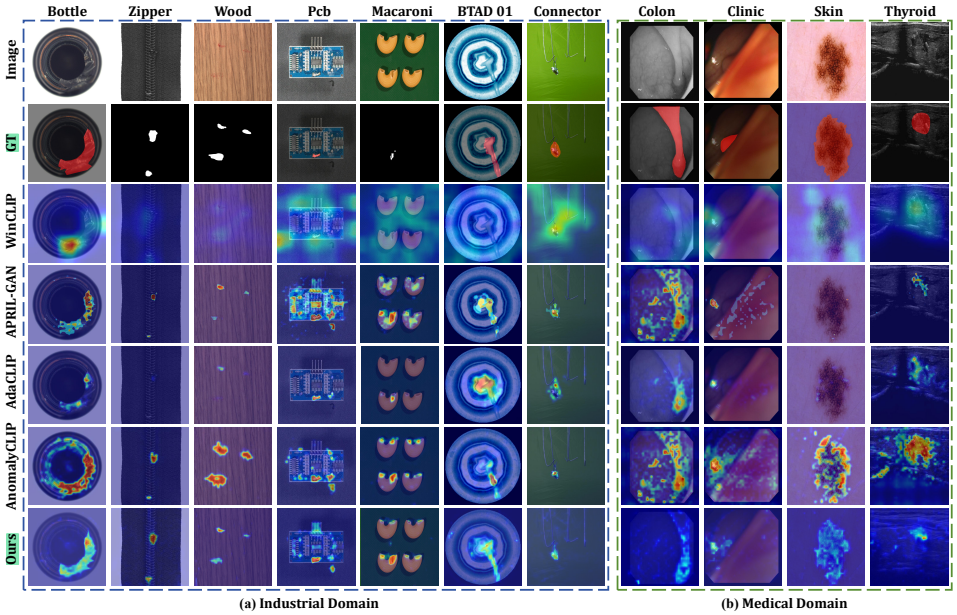

Figure 3: Visualization of ZSAD methods. (a) and (b) present the localization results across industrial and medical domains, respectively.

incorporating DINOv3's discriminative representations, which enhance the separability between normal and anomalous features.

Qualitative comparisons in Fig. 3(a) further confirm these findings. Competing approaches, particularly WinCLIP and APRILGAN, often produce blurry or noisy heatmaps, failing to accurately delineate defect regions. In contrast, AD-DINOv3 generates sharp and precise anomaly maps, effectively highlighting defective components such as scratches on metal surfaces and structural defects in PCB boards, while suppressing irrelevant background noise.

Comparison on medical image benchmarks.

We further evaluate AD-DINOv3 on medical datasets to assess its ability to handle fine-grained and visually subtle anomalies. As summarized in Tab. 1, our method achieves consistent improvements over prior works. On ClinicDB, AD-DINOv3 delivers an AUROC of 90.4% and an F1 score of 54.3%, significantly outperforming AnomalyCLIP (82.9% / 42.1%) and AdaCLIP (82.8% / 40.9%). On ISIC, our method achieves an AUROC of 89.0% with an F1 score of 72.1%, competitive with AdaCLIP while maintaining balanced localization accuracy. Similar gains are observed on ColonDB and TN3K, where our approach consistently improves upon existing CLIP-based baselines.

On average, across the four medical datasets, AD-DINOv3 achieves an AUROC of 84.5% and an F1 score of 51.5%, establishing new state-of-the-art performance. This highlights the robustness of our framework in capturing anomalies in medical images, which are typically small in size and highly ambiguous.

The qualitative results in Fig. 3(b) further illustrate this advantage. Competing approaches often misclassify normal tissues as abnormal or fail to localize subtle lesions, especially in dermoscopic and endoscopic images. Conversely, AD-DINOv3 generates clearer and more discriminative heatmaps, successfully identifying lesions in skin and colon images and localizing irregular patterns in thyroid scans. These results underscore the potential of our method for real-world medical applications where high precision and sensitivity are critical.

4.3 Ablation Study

As shown in Tab. 2 we dismantle AD-DINOv3 step-by-step on MVTec AD to verify the contribution of each component. For the baseline, We simply $\ell 2$-normalize the raw DINOv3 patch tokens, average over channel dimension, and upsample to the image size as the anomaly map. This yields 76.20%

Table 2: Ablation Results of AD-DINOv3's Components on MVTec AD.

Num.	CMCL	AACM	AUROC	F1
1.	✗	✗	76.20	20.49
2.	✔	✗	90.98 (+14.78)	47.00 (+26.51)
3.	✔	✔	91.60 (+15.40)	50.13 (+29.64)

Table 3: Ablation Results of multi-layer features on MVTec AD.

Num.	Last Layer	Multi Layer	AUROC	F1
1.	✔		90.38	48.24
2.		✔	91.60 (+1.22)	50.13 (+1.89)

AUROC and 20.49% F1, confirming that generic self-supervised features alone cannot reliably separate normal and anomalous regions.

Ablation study on the Cross-Modal Contrastive Learning (CMCL). As shown in row 2 of Tab. 2, introducing CMCL, which explicitly aligns visual and textual features, leads to a large performance gain: AUROC rises to 90.98% (+14.78%) and F1 to 47.00% (+26.51%). This substantial improvement indicates that cross-modal alignment effectively narrows the gap between image features and semantic prompts. By enforcing consistency across modalities, CMCL calibrates the embedding space so that normal and abnormal regions become more separable, thereby refining pixel-level decision boundaries and enhancing overall discriminability. These results confirm that leveraging semantic cues from the text branch is crucial for guiding the visual representations toward anomaly-aware feature learning.

Ablation study on the Anomaly-Aware Calibration Module (AACM). As shown in row 3 of Tab. 2, adding the proposed AACM further increases AUROC to 91.60% and F1 to 50.13%. This gain demonstrates that explicitly guiding the global CLS token to attend to defective regions enhances. By redirecting attention from generic foreground semantics to anomalies, AACM calibrates the representation space and makes the model focus on subtle yet critical cues, thereby yielding more accurate anomaly maps and more reliable image-level decisions.

Ablation study on multi-level features. To investigate the effect of utilizing features from different layers of the visual backbone, we conduct an ablation study comparing single-layer and multi-layer settings. In the single-layer variant, only the last-layer output is used to compute similarity with text embeddings for anomaly detection. In contrast, the multi-layer variant aggregates features from the 6th, 12th, 18th, and 24th layers, and the anomaly maps from these four stages are averaged to obtain the final prediction.

As shown in Tab. 3, incorporating multi-level features leads to consistent improvements across all evaluation metrics. Compared with the single-layer baseline, the multi-layer strategy achieves an increase of +1.22% in AUROC and +1.89% in F1 score, demonstrating that intermediate features contain complementary information that helps capture both low-level appearance cues and high-level semantic context. This indicates that leveraging hierarchical representations enhances the model's ability to distinguish anomalies from normal regions.

5 Conclusion

In this work, we introduced AD-DINOv3, the first framework that adapts DINOv3 to the ZSAD task. Our approach integrates lightweight adapters and a novel Anomaly-Aware Calibration Module (AACM) to refine patch and CLS tokens, explicitly guiding the model to focus on anomalous regions rather than generic foreground semantics. Across 8 industrial and medical benchmarks, AD-DINOv3 matches or surpasses state-of-the-art approaches, underscoring its robustness and broad applicability to ZSAD. We further observe that DINOv3 representations fragment normal regions into multiple clusters due to intra-class variations, while anomalies fail to form a distinctive cluster. Overall, our study shows that DINOv3 yields more discriminative visual features than CLIP, offering a powerful and generalizable solution to ZSAD.

References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. *Asian Conference on Computer Vision (ACCV)*, pages 622–637, 2018.

[2] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[3] Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *International MICCAI Brainlesion Workshop*, pages 161–169. Springer, 2018.

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9592–9600, 2019.

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec AD - A comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9592–9600. Computer Vision Foundation / IEEE, 2019.

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

[7] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024.

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

[9] Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, and Yong Liu. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. In *International Joint Conference on Artificial Intelligence*, pages 17–33. Springer, 2024.

[10] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[11] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *ArXiv*, abs/2005.02357, 2020.

[12] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid features. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1001–1010, 2020.

[13] Thomas Defard, Aleksandr Setkov, Angélique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *ICPR Workshops*, 2020.

[14] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1705–1714, 2019.

[15] Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 257–261. IEEE, 2021.

[16] Lei Han, Yixuan Xu, Ming Li, Liang Gao, and Yi Yang. Promptad: Learning prompts with anomaly-aware feature rectification for zero-shot anomaly detection. *arXiv preprint arXiv:2403.01234*, 2024.

[17] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[18] Jongheon Jeong, Hwanjun Song, Namgi Kim, and Seunghoon Yoon. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19606–19616, 2023.

[19] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19606–19616. IEEE, 2023.

[20] 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)*, pages 66–71. IEEE, 2021.

[21] Pranita Balaji Kanade and PP Gumaste. Brain tumor detection using mri images. *Brain*, 3(2):146–150, 2015.

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[23] Fumio C. Kitamura. Head ct - hemorrhage. https://www.kaggle.com/datasets/fourthbrain/head-ct-hemorrhage, 2018. https://doi.org/10.34740/KAGGLE/DSV/152137.

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[25] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

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

[27] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, 2021.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763. PMLR, 2021.

[29] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18061–18070. IEEE, 2022.

[30] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter V. Gehler. Towards total recall in industrial anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14298–14308. IEEE, 2022.

[31] Karsten Roth, Luca Pemula, Joaquín Zepeda, Bernhard Scholkopf, 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)*, pages 14318–14328, 2022.

[32] Lukas Ruff, Robert Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib A Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4393–4402. PMLR, 2018.

[33] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[34] Vikash Sehwag, Ping-yeh Chiang, Prateek Mittal, and Soheil Feizi. Zero-shot anomaly detection with pre-trained vision-language models. In *Proceedings of the NeurIPS Workshop on Distribution Shifts*, 2022.

[35] Oriane Simoni and et al. Dinov3: Self-supervised learning for vision at unprecedented scale. *arXiv preprint arXiv:2508.10104*, 2025.

[36] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

[37] Yoad Tewel, Gal Shalev, Elad Ben-Zaken, Uri Alon, Shai Shalev-Shwartz, and Amir Globerson. Zero-shot anomaly detection by learning to leverage vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 1–14, 2022.

[38] Mohamed Wageh, Khalid Amin, Abeer D Algarni, Ahmed M Hamad, and Mina Ibrahim. Brain tumor detection based on deep features concatenation and machine learning classifiers with genetic selection. *IEEE Access*, 2024.

[39] Kaifeng You, Yutong Shi, Yixuan Yan, Jizheng Zhao, Zhizhong Zhang, Yuesheng Chen, and Yaqian Wang. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pages 5280–5288, 2023.

[40] Jingyi Yuan, Chenqiang Gao, Pengyu Jie, Xuan Xia, Shangri Huang, and Wanquan Liu. Afr-clip: Enhancing zero-shot industrial anomaly detection with stateless-to-stateful anomaly feature rectification. *arXiv preprint arXiv:2503.12910*, 2025.

[41] Kaiyang Zhou, Chen Change Loy, and Bo Dai. Can transformers be strong anomaly detectors? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9988–9997, 2022.

[42] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16795–16804. IEEE, 2022.

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7266–7276, 2022.

[44] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2023.

[45] 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.