Robust Zero-shot Anomaly Detection under Limited Auxiliary Anomaly Priors

Guanyu Lu, Fang Zhou(✉), and Cheqing Jin

East China Normal University, Shanghai, China
gylu@stu.ecnu.edu.cn, {fzhou, cqjin}@dase.ecnu.edu.cn

Abstract. Zero-shot anomaly detection aims to identify defects in arbitrary novel domains; however, existing models assume that the auxiliary data contains a rich diversity of anomalies, neglecting the far more complex and unpredictable variations in real-world target domains. This study introduces DIVE, the first approach to investigate the scenario of limited auxiliary anomaly priors and resolve the resulting substantial performance degradation. Through a shallow-and-deep text embedding injection strategy during visual encoding, DIVE learns to abstract generic anomaly concepts shared across the auxiliary training domain and diverse target domains. Moreover, we propose a disentanglement mechanism to tackle the suboptimal alignment between visual embeddings entangled with object semantics and object-agnostic textual prompts. Experiments demonstrate that, under the setting of limited anomaly patterns in auxiliary data, DIVE outperforms SOTA baselines by up to 16.2% and 28.5% on two classification metrics, and 23.4%, 24.1%, and 47.0% on three segmentation metrics, in terms of average performance across twelve datasets. Furthermore, it maintains highly competitive performance when auxiliary data exhibits sufficient anomaly diversity.

Keywords: Anomaly Detection · Zero-shot Learning · Prompt Engineering

1 Introduction

Anomaly detection (AD) [5,26] plays a crucial role in identifying instances that deviate from normal patterns, forming the foundation of a wide range of applications, such as industrial defect detection [19,50], medical diagnosis [4,44], and financial fraud identification [28]. Traditional AD approaches predominantly rely on the assumption that an adequate amount of training data can be acquired beforehand. However, compiling extensive datasets for every newly encountered scenario is often infeasible. This limitation is prominent in two scenarios: 1) accessing or centralizing training data may violate strict privacy policies, such as handling sensitive patient scans in collaborative diagnostics; 2) the target domain may intrinsically lack historical data, as seen in customized manufacturing where novel products demand immediate inspection.

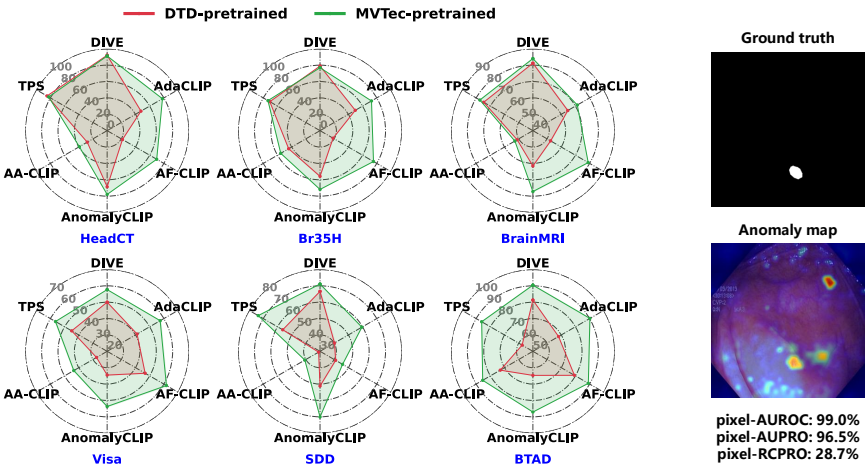

Fig. 1: Left: AP results of five SOTA baseline models and DIVE (our model), evaluated on six target datasets under different auxiliary pre-training settings. The green and red regions denote models pre-trained on MVTec and DTD, respectively. Right: The visualization result of AnomalyCLIP on a single test image from the ColonDB dataset, with performance measured by AUROC, AUPRO, and RCPRO (our proposed metric).

To overcome the aforementioned limitations, Zero-shot Anomaly Detection (ZSAD) has been proposed to identify anomalies in entirely unseen target domains without any prior domain-specific data collection. Existing ZSAD methods predominantly rely on large Vision-Language Models (VLMs), such as CLIP [34], which are pre-trained on massive image-text pairs and provide rich open-world visual priors together with strong cross-modal alignment capabilities. Early studies primarily adopt hand-crafted prompt engineering [9, 10, 20], aligning visual features with manually designed text prompts. This paradigm heavily depends on human expertise and task-specific prompt design, which restricts model scalability and adaptability to diverse anomaly patterns. To alleviate this issue, recent works have shifted towards learnable prompt tuning [11,31,47,49], enabling the model to automatically optimize text embeddings for anomaly detection. Despite the enhanced flexibility brought by learnable prompt tuning, existing ZSAD paradigms still encounter two challenges:

Challenge 1. Degraded generalization under limited auxiliary anomaly priors. Existing ZSAD methods primarily leverage learnable prompts or adapter networks to fine-tune VLMs, operating under the optimistic assumption that they can fully exploit rich anomaly information in the auxiliary data. When such auxiliary anomaly priors are scarce, these methods are highly susceptible to overfitting on specific anomaly distributions. This inevitably leads to a failure when dealing with target domains that harbor more complex defects alongside a potentially continuous influx of novel object categories. For instance, as illustrated in Fig. 1 (left), when the auxiliary data is switched from MVTec (comprising diverse industrial defects such as holes, stains, and cracks) to DTD (containing

only texture anomalies), the average AP performance of these methods across target datasets drops by 9.9% to 23.3%.

Challenge 2. Suboptimal alignment due to visual embeddings entangled with object semantics. The feature space of pre-trained vision encoders is dominated by object semantics. When the visual embeddings are aligned with text prompts designed exclusively for anomaly evaluation, the dominant object semantics act as disruptive noise. Consequently, the model's inability to disentangle subtle defect variations from salient object identities critically hinders the learning of anomaly-discriminative representations.

To this end, we propose a novel model called **DIVE**[1] (**D**isentangled and text-**I**njected **V**isual **E**mbeddings) for ZSAD. To overcome the generalization bottleneck caused by insufficient abnormality coverage (Challenge 1), DIVE introduces a shallow-and-deep text embedding injection strategy. At the shallow level, the cross-modal projection of learnable semantic text prompts into visual prompts enables synergistic optimization for effective generalization across unseen object semantics. Subsequently, by selectively fusing each patch token extracted at deep layers with the text embeddings of LLM-generated descriptions about normality and abnormality through a cross-attention formulation, the model is empowered to abstract generic anomaly concepts from specific auxiliary visual patterns. For instance, a texture anomaly might share similarities in shape and contour with an industrial fabric stain or a medical tumor shadow; thus, its corresponding patch can match with enriched text prompts like `"a photo of an object with a stain"` or `"a photo of an object with a shadow"`. As verified in Fig. 1 (left), benefiting from this strategy, DIVE restricts the average performance penalty to a mere 5.3%, effectively reducing the degradation magnitude by approximately 46.5% to 77.3% compared to the drastic drop observed in baselines under the DTD-pretrained setting. Furthermore, to eliminate the interference of object semantics on AD tasks (Challenge 2), a disentanglement mechanism is applied to decouple the global visual embeddings into independent anomaly-state and object-semantic subspaces, where each is optimized for cross-modal alignment under the guidance of its respective text prompt learner. By imposing orthogonal constraints, the state branch focuses exclusively on visual cues essential for anomaly identification.

This paper additionally introduces a segmentation evaluation metric to resolve the inherent flaws of the standard AUPRO metric [2, 3] widely adopted in prior works [11, 13, 18, 31, 47, 49], which ignores false positives in normal regions and over-segmentation of anomaly regions. As depicted in Fig. 1 (right), while AnomalyCLIP highlights three primary anomalous regions on a given test instance, two of them are stark false positives (including the region with the highest anomaly score). Since the AUPRO metric solely focuses on the coverage of ground-truth anomalous regions, it yields a misleadingly inflated score of 96.5%. By integrating a region-wise prediction quality penalty, the proposed metric RCPRO circumvents this issue, restricting the score of the aforementioned test case to a realistic 28.7%.

[1] Code is available at https://github.com/ZhouF-ECNU/DIVE.

In summary, this paper makes the following key contributions:

- To the best of our knowledge, this work pioneers the study of ZSAD under the setting of limited auxiliary anomaly priors and analyzes the severe generalization bottlenecks induced by this constraint. This realistic scenario reflects the unpredictable variations and the continual emergence of novel anomaly patterns in real-world target domains.
- We propose DIVE, a tailored model for this challenging scenario, with two core components for visual embedding learning: text embedding injection to enable generalization to unseen object categories and abstraction of generic anomaly concepts, and a disentanglement mechanism to alleviate semantic interference in global embeddings.
- We formulate RCPRO, a novel metric that rectifies AUPRO's leniency towards false positives in normal regions and over-segmentation of anomalous regions. Experiments on twelve datasets verify the superiority and robustness of DIVE under both limited and sufficient anomaly diversity settings.

2 Related Work

2.1 Anomaly Detection

Conventional AD methods are typically built on the assumption of *i.i.d.* training and testing data, meaning they are sampled from the same domain. Based on the availability of labeled anomalies, these methods can be categorized into unsupervised [15,40,41,48] and semi-supervised [22,29,36,39] settings. Such a closed-set paradigm leads to limited generalization across diverse domains, necessitating costly retraining when deployed in new scenarios.

To achieve broader applicability, Zero-shot Anomaly Detection has emerged as a promising direction. Generally, current ZSAD methods can be categorized into two paradigms based on their utilization of Large Language Models (LLMs): (i) prompting-based detection [14, 43], which directly queries LLMs to generate descriptive responses indicating anomalies, yet suffers from the imprecise articulation of fine-grained anomaly patterns; (ii) contrasting-based detection [24, 50] offers a more effective alternative by leveraging pre-trained VLMs (e.g., CLIP) to compute similarity between image features and textual prompt embeddings for anomaly identification. As the pioneering work applying CLIP to ZSAD, WinCLIP [20] aligns window-level visual features with hand-crafted text prompts, whose manual construction restricts adaptability and generalization, a limitation also shared by VAND [9] and AnoVL [10]. To solve this issue, AnomalyCLIP [47] introduces an object-agnostic prompt design, where specific portions of the text prompts are substituted with learnable tokens for dynamic optimization. Furthermore, some recent works like TPS [31] and FAPrompt [49] focus on designing fine-grained text prompts to capture nuanced abnormal patterns and enhance local grounding. While these approaches have somewhat contributed to improvements in ZSAD, their generalization heavily relies on the assumption of sufficient anomaly diversity in the auxiliary data, causing severe performance degradation when target domains exhibit significant shifts.

2.2 Prompt Engineering

Vision-Language Models have exhibited impressive zero-shot capabilities, which are largely contingent upon effective prompt engineering. Early attempts relied on manually designed static text templates [34], which are not only labor-intensive but also highly sensitive to specific wording choices. To overcome the limitations of manual design, CoOp [46] pioneered the use of continuous learnable vectors in the text encoder. Subsequent research further advanced textual prompt tuning by learning prompt distributions (ProDA [30]), incorporating knowledge guidance (KgCoOp [42]), or leveraging optimal transport (PLOT [7]). In response to the limited generalization of static learned prompts to unseen classes, CoCoOp [45] introduced instance-conditional dynamic prompting, where learnable tokens are conditioned on individual image features. Recognizing that single-modality adaptation is suboptimal for complex cross-modal alignment, MaPLe [23] explores multi-modal prompt learning, jointly optimizing learnable prompts in both vision and language branches. In this paper, we construct visual embeddings through text-injected synergistic learning and disentanglement, making CLIP more suitable for downstream AD tasks and enabling it to abstract anomaly concepts that are common across diverse domains.

3 Problem Formulation

Consider an auxiliary training dataset $\mathcal{D}_{train} = \{(x_i, y_i, \mathbf{G}_i)\}_{i=1}^{l}$ comprising l samples from a source domain, where $x_i \in \mathbb{R}^{H\times W\times 3}$ is the input image, $y_i \in \{0, 1\}$ denotes the image-level label (with 0 representing normal and 1 representing anomalous), and $\mathbf{G}_i \in \{0, 1\}^{H\times W}$ represents the ground-truth pixel-level anomaly mask. The objective of Zero-Shot AD is to learn a generalized model on $\mathcal{D}_{train}$ capable of handling multiple test datasets $\{\mathcal{D}_{test}^{k}\}_{k=1}^{K}$ derived from K distinct unseen target domains that are disjoint from the source domain. During inference, for any query image $x \in \mathcal{D}_{test}$, the model is expected to simultaneously predict a scalar anomaly score $s \in [0, 1]$ indicating the probability of the image being anomalous, and a pixel-wise anomaly map $\mathbf{M} \in [0, 1]^{H\times W}$ that localizes the anomalous regions. For both the image-level score s and the pixel values in $\mathbf{M}$, a higher value indicates a higher degree of abnormality.

4 Methodology

The overall workflow of DIVE is illustrated in Fig. 2. Following the CLIP-based AD paradigm [6, 31, 32, 47, 49], DIVE aligns global and local visual embeddings with designated "normal" and "abnormal" text prompts. Notably, a text embedding injection strategy (Vision-Language (V-L) coupling function $\mathcal{F}$ and cross-attention networks $\mathcal{CA}$ in Fig. 2) is applied to enhance the model's generalization to unseen object categories and complex anomaly patterns. Furthermore, to prevent visual embeddings from entangling with object semantics, we introduce a disentanglement mechanism (branch networks $\mathcal{G}_{state}$ and $\mathcal{G}_{semantic}$ in Fig. 2) to capture global visual information related to anomaly identification.

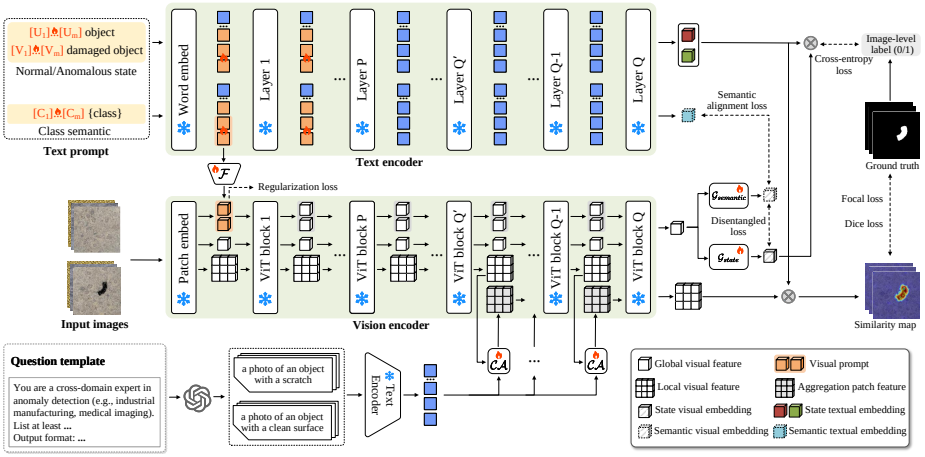

Fig. 2: The workflow of DIVE.

4.1 Independent Parallel Text Prompting

In traditional VLMs, text prompt templates typically entangle anomaly descriptions with specific object identities (e.g., `"a photo of a damaged {class}"`) [20]. To decouple these semantics, AnomalyCLIP [47] discards the token `{class}` from the prompts. Alternatively, it introduces a set of learnable context vectors to construct two object-agnostic text prompt templates:

$$t_n = [U_1][U_2]\dots[U_m][object], \quad t_a = [V_1][V_2]\dots[V_m][damaged][object].$$

However, directly aligning visual embeddings extracted by a pre-trained vision encoder with object-agnostic text prompts is suboptimal, as it forces semantic-dominated visual embeddings into a textual space exclusively designed for anomaly detection. To address this issue, we further design an independent semantic-aware prompt learner on the text side. Operating in parallel with the aforementioned state-aware prompt learner, this module is specifically dedicated to capturing the semantic category information of target objects. This parallel design provides textual anchors for the subsequent precise disentanglement of state-related visual embeddings and semantic-related visual embeddings at the visual end. Specifically, for an auxiliary training dataset comprising R distinct object categories, we formulate the semantic-aware text prompt template using a separate set of learnable context vectors as follows:

$$t_c = [C_1][C_2]\dots[C_m][\text{class}_r], \quad r \in \{1, 2, \dots, R\},$$

where $[\text{class}_r]$ represents the word token for the r-th specific object class.

To encourage the textual context prompts in the two prompt learners to better generalize to both anomaly detection and object recognition tasks, a new set of learnable tokens is introduced into each transformer block $\mathtt{T}_j$ of the text

encoder up to depth P:

$$\begin{gathered}[\_, U_j^i|_{h+1}^m, N_j] = \mathtt{T}_j([U_{j-1}^i|_1^h, U_{j-1}^i|_{h+1}^m, N_{j-1}]), \quad [\_, V_j^i|_{h+1}^m, A_j] = \mathtt{T}_j([V_{j-1}^i|_1^h, V_{j-1}^i|_{h+1}^m, A_{j-1}]), \\ [\_, C_j^i|_{h+1}^m, S_j] = \mathtt{T}_j([C_{j-1}^i|_1^h, C_{j-1}^i|_{h+1}^m, S_{j-1}]), \quad j = 1, 2, \ldots, P,\end{gathered} \tag{1}$$

where N, A, and S denote the fixed input tokens; $U^i|_1^h$, $V^i|_1^h$, and $C^i|_1^h$ represent the newly introduced learnable tokens with a sequence length of h; and $[\cdot, \cdot, \cdot]$ denotes the concatenation operation. Beyond the P^{th} Transformer layer, each subsequent layer takes the prompts from its preceding layer as input:

$$\begin{gathered}[U_{j'}, N_{j'}] = \mathtt{T}_{j'}([U_{j'-1}, N_{j'-1}]), \quad [V_{j'}, A_{j'}] = \mathtt{T}_{j'}([V_{j'-1}, A_{j'-1}]), \\ [C_{j'}, S_{j'}] = \mathtt{T}_{j'}([C_{j'-1}, S_{j'-1}]), \quad j' = P+1, \ldots, Q.\end{gathered} \tag{2}$$

Ultimately, by feeding the final tokens N_Q^m, A_Q^m, and C_Q^m into the text projection head, the two parallel prompt learners generate the state textual embeddings z_n, z_a and the semantic textual embedding z_c, respectively.

4.2 Textual Embedding Injection

Lacking prior knowledge regarding unseen categories and normal/abnormal patterns in the target domain, the pre-trained visual encoder suffers from restricted generalization capabilities in novel scenarios, which incurs a substantial performance degradation. Inspired by the multi-modal prompting mechanism of MaPLe [23], learnable text prompts are introduced to guide visual feature extraction, thereby preserving the cross-modal alignment capability for novel object categories. Given that the shallow layers of Vision Transformer (ViT) are primarily responsible for capturing coarse-grained local structures such as edges and textures, we map the first h learnable tokens of the semantic text prompt into visual prompts via a V-L coupling function $\mathcal{F}$ before the first layer, injecting them into the visual feature extraction process, i.e., $[c_1, E_1, \tilde{C}_1^i|_1^h] = \mathtt{ViT}_1([c_0, E_0, \mathcal{F}(C_0^i|_1^h)])$, where c denotes the global visual feature, and E represents the local visual feature comprising $H \times W$ patch tokens.

While the shallow injection of semantic text prompts improves generalization to unseen object categories, the model remains highly susceptible to overfitting the known anomaly patterns in the auxiliary dataset. When encountering novel variants, this overfitting directly leads to the misclassification of unseen anomalies and false alarms on normal samples. Given that the deep layers of ViT possess strong conceptual abstraction capabilities to represent object-agnostic state features, we directly fuse the rich LLM-generated typological descriptions of normality and abnormality with the patch tokens via cross-attention in these deep layers. This process compels the model to abstract the concrete visual patterns from the auxiliary training data and map them onto more generalizable, universal anomaly concepts. Formally, we execute this cross-modal fusion after the Q'^{th} ViT block:

$$[c_d, E_d, \tilde{C}_d] = \mathtt{ViT}_d([c_{d-1}, E_{d-1}, \tilde{C}_{d-1}]), \quad d = 2, 3, \ldots, Q', \tag{3}$$

$$[c_{d'}, E_{d'}, \tilde{C}_{d'}] = \texttt{ViT}_{d'}([c_{d'-1}, E_{d'-1} + \gamma \cdot \mathcal{CA}_{d'-1}(E_{d'-1}, Z_g W_{proj}^{\top}, Z_g W_{proj}^{\top}), \tilde{C}_{d'-1}]), \quad d' = Q'+1, \ldots, Q, \tag{4}$$

where Z_g denotes the text embeddings obtained by feeding diverse LLM-generated descriptions of normality and abnormality (the question template is detailed in the supplementary material) into a pre-trained CLIP text encoder, and $W_{proj}^{\top}$ is a learnable linear projection matrix that maps these textual embeddings into the visual feature space; $\mathcal{CA}(\cdot,\cdot,\cdot)$ denotes the cross-attention operation, utilizing the visual patch tokens as the query, and the projected text features as the key and value. Notably, these cross-modal state textual priors are injected via a residual connection, where γ is a hyperparameter designed to balance the original visual features and the injected textual information.

To extract the final visual embeddings, the global visual feature z_x and local patch features $z_e^{(i,j)} (i \in [1, H], j \in [1, W])$ are derived by processing the output tokens c_Q and E_Q via the image projection head.

4.3 Visual Embedding Disentanglement

The global visual feature z_x extracted by the vision encoder is a highly entangled embedding, inherently coupling the object identity semantics with the anomaly state information. Directly aligning this entangled embedding with text prompts would inevitably blur the boundary between normal and anomalous instances across different categories. Therefore, we design two parallel learnable Multi-Layer Perceptron (MLP) networks, denoted as $\mathcal{G}_{state}(\cdot)$ and $\mathcal{G}_{semantic}(\cdot)$, to explicitly disentangle z_x into two independent subspaces. This process yields the state visual embedding z_{st} and the semantic visual embedding z_{se}, i.e.,

$$z_{st} = z_x + \Delta z_{st} = z_x + \mathcal{G}_{state}(z_x), \quad z_{se} = z_x + \Delta z_{se} = z_x + \mathcal{G}_{semantic}(z_x). \tag{5}$$

Here, instead of learning an absolute transformation, we formulate the disentanglement through a residual learning paradigm, where Δz_{st} and Δz_{se} represent the learned residual offsets for the state and semantic subspaces, respectively. This not only mitigates the risk of catastrophic forgetting of pre-trained visual concepts but also stabilizes the optimization process.

4.4 Training and Inference

During the training phase, the backbone parameters of the pre-trained vision and text encoders are strictly frozen to retain their open-world generalization capabilities. Only the state-aware and semantic-aware prompt learners, the V-L coupling function $\mathcal{F}$, the introduced cross-attention layers, and the visual MLP projection heads ($\mathcal{G}_{state}$ and $\mathcal{G}_{semantic}$) are trainable.

For the anomaly classification task, we compute the cosine similarity between the state visual embedding z_{st} and the normal/abnormal state textual embeddings $\{z_n, z_a\}$. The state prediction probability p_{st} is formulated as: $p_{st} =$

softmax $\left([z_{st}^{\top} z_n, z_{st}^{\top} z_a]/\tau\right)$, where τ is a temperature hyperparameter, set to 0.07 following the standard setting in SimCLR [8]. The state-aware alignment is optimized using the cross-entropy loss: $\mathcal{L}_{state} = \mathrm{CE}(p_{st}, y_i)$.

Similarly, to ensure that the semantic branch accurately captures the object identity, we align the semantic visual embedding z_{se} with the semantic textual embeddings $\{z_c^r\}_{r=1}^R$ via a contrastive learning objective. Specifically, z_{se} is pulled closer to the textual embedding of its corresponding ground-truth category y_{cls}, while being pushed apart from those of other mismatched categories: $\mathcal{L}_{sem} = -\log \frac{\exp(z_{se}^{\top} z_c^{y_{cls}}/\tau)}{\sum_{r=1}^{R} \exp(z_{se}^{\top} z_c^r/\tau)}$, where $z_c^{y_{cls}}$ represents the positive textual embedding anchor associated with the actual class of the input image.

To enforce the disentanglement between the object semantics and anomaly states, we apply an orthogonal constraint on the learned residual offsets. The disentanglement loss is calculated using the squared cosine similarity between the two offset vectors: $\mathcal{L}_{dis} = \left(\frac{\Delta z_{st}^{\top} \Delta z_{se}}{\|\Delta z_{st}\|_2 \|\Delta z_{se}\|_2}\right)^2$.

For the pixel-level segmentation task, we compute the similarity map $\mathbf{S}$ by calculating the similarity scores between the local patch features $z_e^{(i,j)}$ and the state textual embeddings $\{z_n, z_a\}$. Focal loss [27] and Dice loss [25] are jointly applied to optimize the anomaly segmentation map against the ground-truth mask $\mathbf{G}$: $\mathcal{L}_{seg} = \mathrm{Focal}(Up(\mathbf{S}), \mathbf{G}) + \mathrm{Dice}(Up(\mathbf{S}_a), \mathbf{G}) + \mathrm{Dice}(Up(\mathbf{S}_n), \mathbf{I} - \mathbf{G})$, where $\mathbf{S}_n$ and $\mathbf{S}_a$ denote the normal and abnormal probability channels of $\mathbf{S}$, respectively, and $\mathbf{I}$ is the full-one matrix. The operator $Up(\cdot)$ represents the upsampling operation used to restore the feature resolution to the original image size. Furthermore, to prevent the visual prompt tokens from overpowering the original visual features and causing instability, we impose a regularization constraint $\mathcal{L}_{reg} = \frac{1}{|\tilde{C}|}||\tilde{C}||_2^2$ on the generated visual prompts $\tilde{C}$, where $|\tilde{C}|$ denotes the total number of elements in the prompt tensor.

The overall objective function is optimized as follows: $\mathcal{L} = \mathcal{L}_{state} + \lambda \mathcal{L}_{sem} + \lambda' \mathcal{L}_{dis} + \mathcal{L}_{seg} + \lambda'' \mathcal{L}_{reg}$, where λ, λ' and λ'' are trade-off weighting coefficients.

In the detection phase, inference is performed using the frozen model weights optimized on the auxiliary dataset. Given a query test image x, we first calculate its image-level anomaly score $s \in [0, 1]$. We directly apply the abnormal channel of the state prediction probability as the anomaly score, which tends toward 1 when the abnormal state textual embedding z_a is highly aligned with the state visual embedding z_{st}:

$$s = \frac{\exp(z_{st}^{\top} z_a/\tau)}{\exp(z_{st}^{\top} z_n/\tau) + \exp(z_{st}^{\top} z_a/\tau)}. \tag{6}$$

For pixel-wise predictions, the final-layer patch features $z_e^{(i,j)}$ are used to compute a similarity map. Its abnormal channel $\mathbf{S}_a$ is then upsampled and spatially smoothed to generate the final anomaly map: $\mathbf{M} = G_\sigma\left(Up(\mathbf{S}_a)\right)$, where G_σ represents a Gaussian filter with a standard deviation σ designed to smooth the final segmentation boundaries and eliminate noise.

5 Experiments

5.1 Experimental Setup

Datasets and Baselines. To comprehensively evaluate the proposed DIVE, we selected twelve publicly available datasets, spanning both industrial defect detection (MVTec [2], DTD [1], Visa [51], MPDD [21], SDD [37], and BTAD [33]) and clinical medical imaging (HeadCT [35], Br35H [17], BrainMRI [35], ISIC [16], ColonDB [38], and TN3K [12]). The proposed DIVE is compared against several state-of-the-art baselines representing the latest advancements in ZSAD: AnomalyCLIP [47], AdaCLIP [6], AF-CLIP [11], AA-CLIP [32], and TPS [31]. Recognizing that anomalous events are scarce in real-world applications, these datasets were resampled to simulate an imbalanced setting, where anomalies account for only a minor proportion of the entire distribution. The exceptions are ISIC, ColonDB, and TN3K. Given that they are exclusively utilized for segmentation evaluation and contain no normal instances, we follow their original testing setups. Further statistical details for the datasets, along with descriptions of the baseline models, are provided in the supplementary material.

Evaluation Metrics. Following standard evaluation procedures in ZSAD, we assess performance at both the image and pixel levels. For image-level anomaly classification, the Area Under the Receiver Operating Characteristic curve (AUROC) and Average Precision (AP) are reported. For pixel-level anomaly segmentation, AUROC and the Area Under the Per-Region Overlap (AUPRO) [2,3] are used to evaluate localization accuracy.

We introduce a new segmentation metric, Region-Calibrated PRO (RCPRO), designed to overcome key limitations of the AUPRO metric (as analyzed in Section 1). First, a series of binary anomaly maps is generated by varying the threshold from the minimum to the maximum anomaly score. For each threshold, two core curve coordinates are then computed: the X-axis represents the standard PRO score, and the Y-axis represents the Region-Calibrated score, which quantifies the average prediction quality across all predicted anomalous regions. Specifically, if the maximum overlap between a predicted region and any ground-truth anomaly falls below a tolerance threshold (set to 0.05 of the matched ground-truth area in our setting), the predicted region receives a score of zero, thereby penalizing false positives in normal regions. On the other hand, if a correctly located prediction expands excessively beyond the ground-truth boundary, its score decays inversely proportional to the excess area ratio to penalize over-segmentation. Ultimately, the RCPRO value is defined as the area under the curve formed by these coordinates. Detailed formulas are provided in the supplementary material.

Implementation Details. The pre-trained CLIP model (ViT-L/14@336px)[2] was adopted as our backbone, with all input images uniformly resized to a resolution of 518×518. The learnable parameters detailed in Section 4.4 were updated

[2] https://github.com/mlfoundations/open_clip

Table 1: Performance comparison of DIVE with baseline models utilizing DTD as auxiliary data for pre-training, evaluated across two tasks: image-level classification (AUROC, AP) and pixel-level segmentation (AUROC, AUPRO, RCPRO). The best results are highlighted in **bold** and the second-best results are underlined.

Task	Dataset	AnomalyCLIP (ICLR' 2024)	AdaCLIP (ECCV' 2024)	AF-CLIP (MM' 2025)	AA-CLIP (CVPR' 2025)	TPS (AAAI' 2025)	DIVE
Classification (image-level)	HeadCT	(95.1, 68.6)	(84.1, 47.7)	(70.9, 21.6)	(75.4, 28.1)	(97.9, 84.8)	(**99.1**, **92.1**)
	Br35H	(91.6, 55.8)	(92.2, 49.7)	(59.6, 18.4)	(80.5, 44.2)	(94.8, 70.8)	(**97.3**, **78.9**)
	BrainMRI	(89.9, 61.6)	(92.8, 64.6)	(81.8, 52.5)	(83.1, 51.1)	(93.6, 74.7)	(**96.4**, **81.2**)
	MVTec	(81.7, 55.1)	(84.8, 63.9)	(88.2, 70.0)	(78.4, 56.1)	(85.5, 69.5)	(**89.3**, **73.9**)
	Visa	(66.8, 34.5)	(71.2, 41.1)	(76.3, 46.8)	(60.8, 27.5)	(75.0, 45.0)	(**77.8**, **50.0**)
	MPDD	(58.0, 21.9)	(63.0, 21.2)	(68.2, 21.2)	(43.5, 15.3)	(65.6, 25.0)	(**71.8**, **30.1**)
	SDD	(79.1, 51.2)	(72.0, 40.2)	(75.2, 41.0)	(72.1, 31.0)	(78.8, 56.3)	(**81.7**, **66.6**)
	BTAD	(72.0, 64.7)	(80.4, 68.1)	(**86.2**, 79.3)	(74.8, 73.0)	(61.6, 57.6)	(85.2, **81.3**)
	Average	(79.3, 51.7)	(80.1, 49.6)	(75.8, 43.9)	(71.1, 40.8)	(81.6, 60.5)	(**87.3**, **69.3**)
Segmentation (pixel-level)	ISIC	(70.7, 45.5, 41.2)	(83.5, 15.5, 66.5)	(74.1, 64.0, 64.1)	(80.2, 71.8, **89.3**)	(70.2, 45.2, 61.8)	(**86.4**, **76.5**, 74.2)
	ColonDB	(62.9, 26.2, 15.5)	(76.3, 17.3, 30.1)	(72.7, 58.8, 44.4)	(78.7, 63.2, **59.6**)	(57.8, 18.6, 26.2)	(**81.8**, **68.2**, 56.2)
	TN3K	(69.6, 21.5, 21.1)	(72.6, 8.1, 54.0)	(65.6, 44.9, 36.4)	(79.0, 52.7, **68.6**)	(64.9, 26.2, 43.8)	(**81.3**, **55.8**, 56.8)
	MVTec	(80.9, 48.2, 23.1)	(89.4, 37.8, 24.3)	(90.0, 80.2, 36.4)	(89.6, 76.9, 21.9)	(66.0, 20.2, 0.3)	(**90.5**, **80.9**, **37.0**)
	Visa	(82.1, 45.2, 3.6)	(90.6, 48.9, 14.1)	(91.2, 73.9, 10.9)	(90.0, 71.6, 8.7)	(68.8, 21.2, 0.2)	(**92.0**, **80.2**, **23.2**)
	MPDD	(75.0, 22.0, 0.4)	(91.7, 57.5, 15.7)	(91.9, 80.9, 9.0)	(91.8, 78.2, 5.2)	(81.9, 42.4, 0.9)	(**93.7**, **81.7**, **24.3**)
	SDD	(78.5, 38.3, 20.4)	(84.5, 0.8, 4.2)	(86.2, 53.8, 8.3)	(86.3, 55.3, 8.9)	(49.2, 9.5, 0.3)	(**86.9**, **65.0**, **24.4**)
	BTAD	(63.3, 13.1, 5.4)	(84.4, 7.2, 15.9)	(**87.0**, **67.6**, 25.4)	(79.9, 49.7, 11.9)	(52.3, 13.4, 4.3)	(85.7, 60.6, **27.2**)
	Average	(72.9, 32.5, 16.3)	(84.1, 24.1, 28.1)	(82.3, 65.5, 29.4)	(84.4, 64.9, 34.3)	(63.9, 24.6, 17.2)	(**87.3**, **71.1**, **40.4**)

using the Adam optimizer. Training was conducted with a batch size of 8, applying a learning rate of 1×10^{-3} to the prompt learners and coupling function $\mathcal{F}$, and 5×10^{-5} to the cross-attention network $\mathcal{CA}$. Within the text encoder, the learnable prompt length was set to $m = 12$ and the prompting depth to $P = 9$. The length of the text prompts used for projection was configured to $h = 4$. For the vision encoder, cross-attention modules were introduced into the final four ViT blocks (i.e., from $Q' = 21$ to $Q = 24$) to enable the injection of the LLM-generated text embeddings. The trade-off parameters λ, λ', and λ'' were set to 0.5, 0.1, and 0.01, respectively. For all baseline models, we used their official implementations released by the authors. All experiments were implemented in PyTorch 2.0.0 and conducted on a server equipped with a single NVIDIA L40S GPU, an Intel(R) Xeon(R) Silver 4110 CPU, and 160 GB of RAM.

5.2 Experimental Results and Analysis

Overall Performance. Comparing Table 1 and Table 2 reveals a significant performance degradation when utilizing DTD as auxiliary data instead of MVTec. In the image-level classification task, baseline models suffer an average drop of 7.1%-11.6% in AUROC and 9.9%-23.3% in AP, with a similar downward trend consistently observed in the segmentation task. This discrepancy arises because MVTec provides diverse anomaly classes that naturally share closer patterns with the target domains. Since real-world applications frequently encounter unpredictable domain shifts, the DTD pre-training setting presents a more realistic and challenging scenario. In contrast, DIVE successfully mitigates this degradation under this challenging setting, restricting the performance drop in the classification task to a mere 2.8% for AUROC and 5.3% for AP. DIVE yields the best results on nearly all testing datasets and ranks second-best in the few exceptions. For image-level classification, DIVE achieves average improvements of

Table 2: Performance comparison of DIVE with baseline models utilizing MVTec as auxiliary data for pre-training.

Task	Dataset	AnomalyCLIP (ICLR' 2024)	AdaCLIP (ECCV' 2024)	AF-CLIP (MM' 2025)	AA-CLIP (CVPR' 2025)	TPS (AAAI' 2025)	DIVE
Classification (image-level)	HeadCT	(98.3, 77.9)	(97.4, 78.0)	(92.9, 69.9)	(87.5, 39.2)	(98.1, 82.4)	(**99.1**, **91.2**)
	Br35H	(95.1, 71.8)	(95.4, 72.6)	(96.0, 75.4)	(82.5, 55.4)	(95.2, 73.4)	(**96.1**, **76.9**)
	BrainMRI	(95.5, 77.2)	(91.2, 71.3)	(96.5, 79.0)	(84.9, 52.7)	(94.1, 77.4)	(**97.3**, **84.0**)
	DTD	(91.5, 78.3)	(**97.6**, **92.1**)	(89.6, 75.4)	(86.6, 77.2)	(91.3, 81.4)	(92.8, 84.0)
	Visa	(81.7, 53.7)	(83.4, 57.4)	(**85.0**, **61.8**)	(73.9, 43.5)	(83.3, 56.4)	(84.0, 57.7)
	MPDD	(72.2, 31.4)	(75.1, 38.8)	(74.0, 40.8)	(59.7, 23.3)	(73.1, 31.9)	(**75.6**, **41.1**)
	SDD	(83.0, 70.1)	(81.6, 59.5)	(73.7, 45.9)	(71.6, 40.7)	(83.8, **73.6**)	(**83.9**, 71.1)
	BTAD	(91.5, 87.1)	(90.9, 90.2)	(91.8, 89.1)	(82.2, 85.4)	(90.7, 86.1)	(**92.0**, **90.4**)
	Average	(88.6, 68.4)	(89.1, 70.0)	(87.4, 67.2)	(78.6, 52.2)	(88.7, 70.3)	(**90.1**, **74.6**)
Segmentation (pixel-level)	ISIC	(84.9, 69.4, 69.6)	(81.4, 13.9, 69.4)	(**90.1**, **78.1**, **85.7**)	(77.6, 68.4, 85.4)	(66.4, 40.0, 59.4)	(88.5, 76.4, 73.9)
	ColonDB	(82.1, 69.7, 53.4)	(80.8, 21.2, 36.7)	(80.7, 66.6, 51.2)	(73.4, 63.0, 54.3)	(60.1, 25.5, 28.1)	(**83.2**, **71.4**, **54.9**)
	TN3K	(82.0, 49.3, 51.2)	(78.4, 5.8, 52.8)	(81.4, **62.2**, **62.7**)	(67.1, 35.5, 52.4)	(65.5, 28.9, 45.1)	(**84.1**, 54.5, 55.7)
	DTD	(98.6, 92.8, 48.1)	(99.0, 65.9, 27.9)	(99.5, 92.0, 37.0)	(92.9, 81.5, 33.3)	(86.9, 65.5, 11.1)	(**99.7**, **94.7**, **49.9**)
	Visa	(94.9, 81.8, 24.5)	(92.6, 33.9, 16.2)	(**96.2**, 81.1, 23.9)	(92.7, 76.8, 10.9)	(79.4, 44.7, 0.6)	(95.0, **85.1**, **25.4**)
	MPDD	(92.3, 77.8, 17.8)	(94.5, 24.3, 17.7)	(95.3, 82.4, **20.4**)	(94.5, 82.5, 7.6)	(87.5, 58.6, 0.2)	(**95.5**, **84.4**, 18.5)
	SDD	(91.3, 59.9, 32.6)	(86.0, 35.2, 26.8)	(**90.5**, 55.5, 26.0)	(77.0, 44.5, 9.9)	(72.6, 33.8, 6.2)	(88.2, **60.0**, **33.5**)
	BTAD	(93.8, 72.9, 34.0)	(93.6, 47.9, 29.4)	(**95.0**, 53.8, 34.6)	(91.3, 71.7, 20.3)	(66.4, 29.9, 6.1)	(94.3, **73.6**, **37.6**)
	Average	(90.0, 71.7, 41.4)	(88.3, 31.0, 34.6)	(**91.1**, 71.5, 42.7)	(83.3, 65.5, 34.3)	(73.1, 40.9, 19.6)	(**91.1**, **75.0**, **43.7**)

5.7%-16.2% in AUROC and 8.8%-28.5% in AP. Similarly, for pixel-level segmentation, it delivers substantial average gains of 2.9%-23.4% in AUROC, 5.6%-47.0% in AUPRO, and 6.1%-24.1% in our proposed RCPRO metric. Furthermore, Table 2 also demonstrates that given sufficient anomaly diversity in the auxiliary data, DIVE maintains highly competitive performance, yielding the best average results across all evaluated datasets.

To delve into several intriguing phenomena observed in Table 1, Fig. 3 visualizes the anomaly maps generated by DIVE and three relevant baselines for further analysis. *Observation 1.* Although TPS ranks second or third in image-level classification across most datasets, it underperforms in segmentation. As illustrated by its anomaly maps, although TPS possesses the capacity to recognize anomalous images globally, it generates numerous small-scale false positives scattered across normal regions. *Observation 2.* AdaCLIP outperforms AA-CLIP in our proposed RCPRO metric on several industrial datasets (e.g., MVTec, Visa, and MPDD), despite falling significantly behind in AUPRO. The visualization results show that AA-CLIP tends to predict excessively large connected components, leading to severe over-segmentation and prominent false positives in normal regions. In contrast, AdaCLIP can localize the core anomalies, albeit with potentially incomplete boundaries. As the AUPRO metric primarily emphasizes the recall of ground-truth regions without imposing penalties for extraneous predictions, AA-CLIP receives an inflated AUPRO score. This phenomenon compellingly demonstrates the necessity and rationality of our RCPRO metric. *Observation 3.* AA-CLIP yields higher RCPRO scores than DIVE on several medical imaging datasets, most notably ISIC. This exception stems from the specific characteristics of the ISIC dataset, where anomalous regions occupy a dominant portion of the entire image. Consequently, AA-CLIP's tendency to predict excessively expansive connected regions, which causes over-segmentation elsewhere, inadvertently becomes an advantage here. Nevertheless, DIVE still precisely segments the core pathological areas of these anomalies.

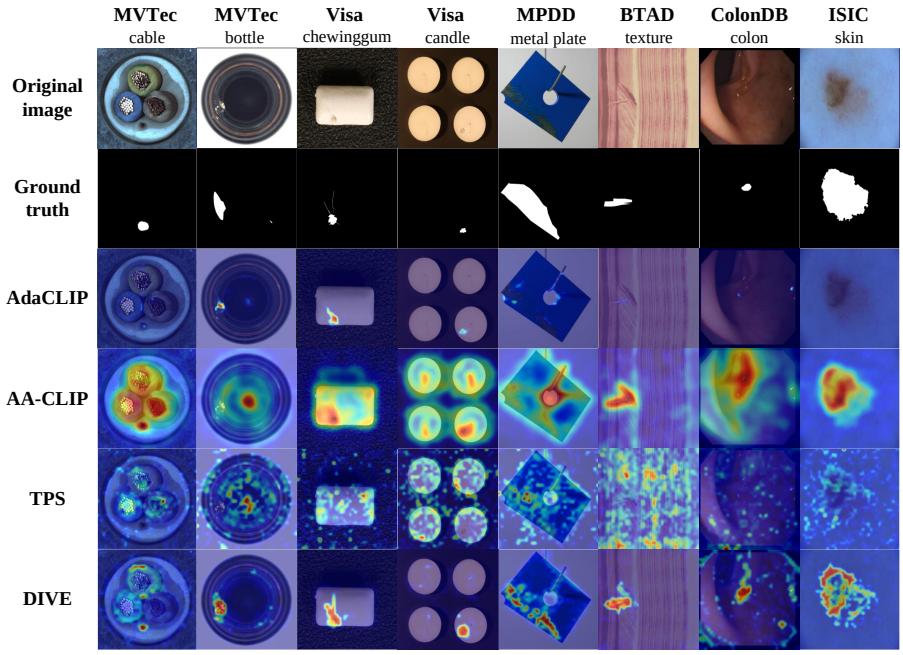

Fig. 3: Visualization of anomaly maps generated by DIVE and baseline models across diverse datasets, under the setting where DTD is utilized as auxiliary pre-training data.

Ablation Study. To evaluate the contribution of each proposed component, we perform ablation studies using several DIVE variants: 1) DIVE$_{\text{-dis}}$ ablates the disentanglement mechanism; 2) DIVE$_{\text{-text}}$ entirely discards the text embedding injection. DIVE$_{\text{-text(s)}}$ and DIVE$_{\text{-text(d)}}$ are further introduced to ablate the shallow-level and deep-level injections, respectively. Table 3 presents the results of the aforementioned ablated models under the setting of using DTD as auxiliary data. Comparing DIVE$_{\text{-dis}}$ with the full DIVE model reveals that the performance degradation predominantly occurs in the two classification metrics. This is attributed to the fact that the disentanglement operates on the global visual embeddings, which are aligned with state textual embeddings to derive the image-level anomaly scores. Furthermore, we observe that the performance degradation of DIVE$_{\text{-text}}$ is more pronounced than that of DIVE$_{\text{-dis}}$. In scenarios where the auxiliary data lacks sufficient anomaly diversity, the text embedding injection strategy is critical to capture generic anomaly concepts, thereby safeguarding the model against a decline in generalization.

A further analysis of the DIVE$_{\text{-text}}$ variants shows that DIVE$_{\text{-text(d)}}$ exhibits a more pronounced performance degradation than DIVE$_{\text{-text(s)}}$, underscoring the importance of incorporating LLM-generated descriptions of normality and abnormality into the deep-level patch tokens. Within the DIVE framework, the cross-attention networks are tasked with integrating these text embeddings, and

Table 3: Performance of DIVE and its ablated variants on the MVTec and Visa datasets. The evaluation metrics for the classification task are (AUROC, AP), and for the segmentation task are (AUROC, AUPRO, RCPRO).

Models	MVTec		Visa	
	Classification	Segmentation	Classification	Segmentation
DIVE$_{\text{-dis}}$	(88.7$_{\downarrow 0.6}$, 72.1$_{\downarrow 1.8}$)	(89.9$_{\downarrow 0.6}$, 80.5$_{\downarrow 0.4}$, 36.8$_{\downarrow 0.2}$)	(76.2$_{\downarrow 1.6}$, 48.4$_{\downarrow 1.6}$)	(91.8$_{\downarrow 0.2}$, 79.5$_{\downarrow 0.7}$, 22.9$_{\downarrow 0.3}$)
DIVE$_{\text{-text(s)}}$	(88.8$_{\downarrow 0.5}$, 73.2$_{\downarrow 0.7}$)	(90.2$_{\downarrow 0.3}$, 80.5$_{\downarrow 0.4}$, 36.2$_{\downarrow 0.8}$)	(76.9$_{\downarrow 0.9}$, 49.1$_{\downarrow 0.9}$)	(91.6$_{\downarrow 0.4}$, 79.4$_{\downarrow 0.8}$, 23.1$_{\downarrow 0.1}$)
DIVE$_{\text{-text(d)}}$	(87.3$_{\downarrow 2.0}$, 70.8$_{\downarrow 3.1}$)	(89.7$_{\downarrow 0.8}$, 77.3$_{\downarrow 3.6}$, 35.6$_{\downarrow 1.4}$)	(75.3$_{\downarrow 2.5}$, 46.0$_{\downarrow 4.0}$)	(91.1$_{\downarrow 0.9}$, 77.9$_{\downarrow 2.3}$, 21.6$_{\downarrow 1.6}$)
DIVE$_{\text{-text}}$	(86.9$_{\downarrow 2.4}$, 69.5$_{\downarrow 4.4}$)	(88.3$_{\downarrow 2.2}$, 76.6$_{\downarrow 4.3}$, 35.5$_{\downarrow 1.5}$)	(74.4$_{\downarrow 3.4}$, 43.9$_{\downarrow 6.1}$)	(90.4$_{\downarrow 1.6}$, 74.4$_{\downarrow 5.8}$, 20.5$_{\downarrow 2.7}$)
DIVE	(89.3, 73.9)	(90.5, 80.9, 37.0)	(77.8, 50.0)	(92.0, 80.2, 23.2)

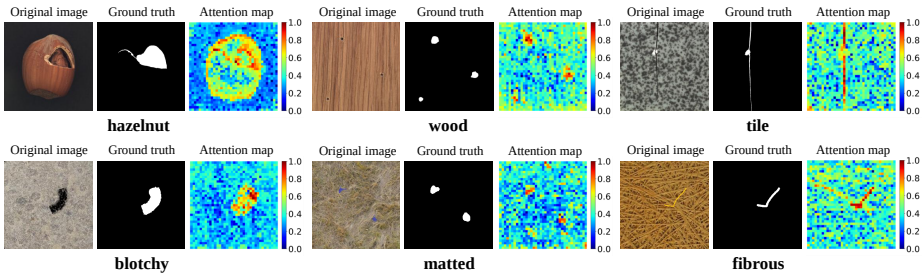

Fig. 4: Visualization of attention maps between image patches of representative categories from the auxiliary data (first row: MVTec, second row: DTD) and the embeddings of LLM-generated descriptions. The attention score for each patch is calculated by summing the normalized similarity probabilities of all anomaly descriptions.

the resulting patch-wise attention maps are visualized in Fig. 4. It can be observed that the cross-attention module in DIVE excels at focusing on anomalous regions, though it also exhibits a slight tendency to attend to object boundaries. For instance, regarding the hazelnut category in Fig. 4, high attention scores are assigned not only to the damaged area but also to the contour of the hazelnut against the background. Consequently, as shown in Fig. 3, DIVE may occasionally introduce minor false positive predictions along object edges (e.g., the anomaly maps for the cable in MVTec and the colon in ColonDB).

Hyperparameter Sensitivity. In this section, we conduct a sensitivity analysis of DIVE on the MVTec and Visa datasets concerning three hyperparameters: the prompting depth P, the learnable prompt length m, and the length of the text prompts used for projection h. First, to evaluate the model's sensitivity to P, we keep the other two parameters fixed and vary the value of P within the set $\{3, 6, 9, 12\}$. As illustrated in Fig. 5, DIVE is relatively sensitive to the parameter P. With the increase of P, the model's performance exhibits an initial upward trend followed by a subsequent decline. This phenomenon can be attributed to the fact that deeper prompt layers facilitate a more refined construction of the textual embedding space regarding anomalous states and object semantics; however, when the prompting layers become excessively deep, it may degrade the original representation capacity of the pre-trained VLMs.

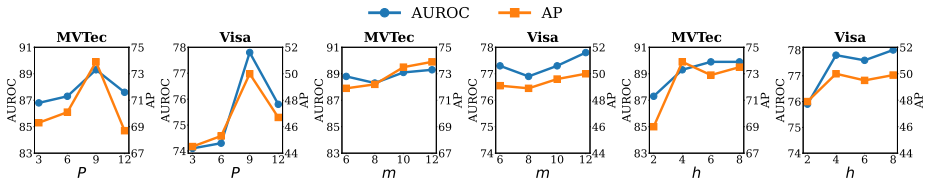

Fig. 5: AUROC (left Y-axis) and AP (right Y-axis) results under varying values of the hyperparameters P, m, and h.

Next, we evaluate the sensitivity of DIVE to the hyperparameters m and h, with their variation ranges set to $\{6, 8, 10, 12\}$ and $\{2, 4, 6, 8\}$, respectively. The hyperparameter m denotes the length of the learnable context initialized at the first layer of the text encoder. As observed in Fig. 5, DIVE demonstrates relatively low sensitivity to m. This can be interpreted as indicating that a small number of learnable context tokens is already sufficient to adequately capture the essential information required to distinguish between normal and anomalous states. Regarding the hyperparameter h, we select a relatively small value (e.g., 4). This choice is motivated by the observation that DIVE's performance tends to plateau beyond this value. Furthermore, maintaining a smaller value prevents the visual encoding process from incurring excessive overhead or introducing redundant noise.

6 Conclusion

In this paper, we investigate a more realistic and challenging scenario for ZSAD characterized by limited auxiliary anomaly priors. To address this, we propose DIVE, a novel model that integrates and fine-tunes learnable networks and prompts to construct robust visual embeddings. By leveraging text embedding injection and a disentanglement mechanism, DIVE is able to capture anomaly patterns shared between the auxiliary training domain and unknown target domains. Extensive experiments demonstrate that DIVE exhibits superior generalization on testing data that heavily deviates from the auxiliary training distribution. Future work will focus on mitigating the small false-positive noise patches introduced by DIVE along object boundaries to further enhance its fine-grained anomaly localization capability.

Acknowledgments

This work was supported by Shanghai "Science and Technology Innovation Action Plan" Project (No. 23511100700).

References

1. Aota, T., Tong, L.T.T., Okatani, T.: Zero-shot versus many-shot: Unsupervised texture anomaly detection. In: WACV. pp. 5564–5572 (2023)

2. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: MVTec AD–a comprehensive real-world dataset for unsupervised anomaly detection. In: CVPR. pp. 9592–9600 (2019)
3. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: CVPR. pp. 4183–4192 (2020)
4. Cai, Y., Chen, H., Cheng, K.T.: Rethinking autoencoders for medical anomaly detection from a theoretical perspective. In: MICCAI. pp. 544–554 (2024)
5. Cao, Y., Xu, X., Zhang, J., Cheng, Y., Huang, X., Pang, G., Shen, W.: A survey on visual anomaly detection: Challenge, approach, and prospect. arXiv preprint arXiv:2401.16402 (2024)
6. Cao, Y., Zhang, J., Frittoli, L., Cheng, Y., Shen, W., Boracchi, G.: AdaCLIP: Adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. In: ECCV. pp. 55–72 (2024)
7. Chen, G., Yao, W., Song, X., Li, X., Rao, Y., Zhang, K.: PLOT: Prompt learning with optimal transport for vision-language models. In: ICLR (2023)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)
9. Chen, X., Han, Y., Zhang, J.: A zero-/few-shot anomaly classification and segmentation method for CVPR 2023 VAND workshop challenge tracks 1&2: 1st place on zero-shot AD and 4th place on few-shot AD. arXiv preprint arXiv:2305.17382 (2023)
10. Deng, H., Zhang, Z., Bao, J., Li, X.: AnoVL: Adapting vision-language models for unified zero-shot anomaly localization. arXiv preprint arXiv:2308.15939 (2023)
11. Fang, Q., Lv, W., Su, Q.: AF-CLIP: Zero-shot anomaly detection via anomaly-focused CLIP adaptation. In: ACM MM. pp. 4846–4855 (2025)
12. Gong, H., Chen, G., Wang, R., Xie, X., Mao, M., Yu, Y., Chen, F., Li, G.: Multi-task learning for thyroid nodule segmentation with thyroid region prior. In: ISBI. pp. 257–261 (2021)
13. Gong, T., Chu, Q., Liu, B., Zhou, W., Yu, N.: FE-CLIP: Frequency enhanced CLIP model for zero-shot anomaly detection and segmentation. In: ICCV. pp. 21220–21230 (2025)
14. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: AnomalyGPT: Detecting industrial anomalies using large vision-language models. In: AAAI. vol. 38, pp. 1932–1940 (2024)
15. Guo, J., Lu, S., Zhang, W., Chen, F., Li, H., Liao, H.: Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In: CVPR. pp. 20405–20415 (2025)
16. Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1605.01397 (2016)
17. Hamada, A.: Br35h: brain tumor detection 2020. `https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection` (2020), Accessed: 28 June 2026
18. He, J., Cao, M., Peng, S., Xie, Q.: RareCLIP: Rarity-aware online zero-shot industrial anomaly detection. In: ICCV. pp. 24478–24487 (2025)
19. Ho, C.H., Peng, K.C., Vasconcelos, N.: Long-tailed anomaly detection with learnable class names. In: CVPR. pp. 12435–12446 (2024)
20. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: WinCLIP: Zero-/few-shot anomaly classification and segmentation. In: CVPR. pp. 19606–19616 (2023)

21. Jezek, S., Jonak, M., Burget, R., Dvorak, P., Skotak, M.: Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In: ICUMT. pp. 66–71 (2021)
22. Jiang, M., Han, S., Huang, H.: Anomaly detection with score distribution discrimination. In: SIGKDD. pp. 984–996 (2023)
23. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: MaPLe: Multi-modal prompt learning. In: CVPR. pp. 19113–19122 (2023)
24. Li, X., Zhang, Z., Tan, X., Chen, C., Qu, Y., Xie, Y., Ma, L.: PromptAD: Learning prompts with only normal samples for few-shot anomaly detection. In: CVPR. pp. 16838–16848 (2024)
25. Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced nlp tasks. In: ACL. pp. 465–476 (2020)
26. Li, Z., Yan, Y., Wang, X., Ge, Y., Meng, L.: A survey of deep learning for industrial visual anomaly detection. Artif. Intell. Rev. **58**(9), 279 (2025)
27. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
28. Lu, G., Lin, X., Pavlovski, M., Zhang, X., Zhou, F.: Targeted detection of anomalous merchants on integrated payment platforms via multifaceted transaction representation learning. In: IEEE BigData. pp. 2170–2178 (2024)
29. Lu, G., Zhou, F., Pavlovski, M., Zhou, C., Jin, C.: A robust prioritized anomaly detection when not all anomalies are of primary interest. In: ICDE. pp. 775–788 (2024)
30. Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X.: Prompt distribution learning. In: CVPR. pp. 5206–5215 (2022)
31. Ma, J., Xie, W., Ye, H., Li, D., Fang, L.: Aligning and prompting anything for zero-shot generalized anomaly detection. In: AAAI. vol. 39, pp. 5964–5972 (2025)
32. Ma, W., Zhang, X., Yao, Q., Tang, F., Wu, C., Li, Y., Yan, R., Jiang, Z., Zhou, S.K.: AA-CLIP: Enhancing zero-shot anomaly detection via anomaly-aware CLIP. In: CVPR. pp. 4744–4754 (2025)
33. Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., Foresti, G.L.: VT-ADL: A vision transformer network for image anomaly detection and localization. In: ISIE. pp. 01–06 (2021)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
35. Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M.H., Rabiee, H.R.: Multiresolution knowledge distillation for anomaly detection. In: CVPR. pp. 14902–14912 (2021)
36. Shou, H., Lu, G., Pavlovski, M., Zhou, F.: READ: Robust and efficient anomaly detection under data contamination and limited supervision. In: SIGKDD. pp. 2586–2596 (2025)
37. Tabernik, D., Šela, S., Skvarč, J., Skočaj, D.: Segmentation-based deep-learning approach for surface-defect detection. J. Intell. Manuf. **31**(3), 759–776 (2020)
38. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imaging **35**(2), 630–644 (2015)
39. Wei, R., He, Z., Pavlovski, M., Zhou, F.: GAD: A generalized framework for anomaly detection at different risk levels. In: CIKM. pp. 2513–2522 (2024)
40. Xu, H., Pang, G., Wang, Y., Wang, Y.: Deep isolation forest for anomaly detection. IEEE Trans. Knowl. Data Eng. **35**(12), 12591–12604 (2023)

41. Xu, H., Wang, Y., Wei, J., Jian, S., Li, Y., Liu, N.: Fascinating supervisory signals and where to find them: Deep anomaly detection with scale learning. In: ICML. pp. 38655–38673 (2023)
42. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: CVPR. pp. 6757–6767 (2023)
43. Zhang, J., He, H., Chen, X., Xue, Z., Wang, Y., Wang, C., Xie, L., Liu, Y.: GPT-4V-AD: Exploring grounding potential of VQA-oriented GPT-4V for zero-shot anomaly detection. In: IJCAI. pp. 3–16 (2024)
44. Zhang, X., Xu, M., Qiu, D., Yan, R., Lang, N., Zhou, X.: MediCLIP: Adapting CLIP for few-shot medical image anomaly detection. In: MICCAI. pp. 458–468 (2024)
45. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022)
46. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vis. **130**(9), 2337–2348 (2022)
47. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In: ICLR (2024)
48. Zhou, Y., Xu, X., Song, J., Shen, F., Shen, H.T.: MSFlow: Multiscale flow-based framework for unsupervised anomaly detection. IEEE Trans. Neural Netw. Learn. Syst. **36**(2), 2437–2450 (2025)
49. Zhu, J., Ong, Y.S., Shen, C., Pang, G.: Fine-grained abnormality prompt learning for zero-shot anomaly detection. In: ICCV. pp. 22241–22251 (2025)
50. Zhu, J., Pang, G.: Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In: CVPR. pp. 17826–17836 (2024)
51. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: ECCV. pp. 392–408 (2022)