applied sciences

MDPI

Article

CLIP-Based Adaptive Weighted Anomaly Detection with Cross-Layer Semantic Decoupling

Fudong Li [1,*], Yong Wang [1], Xian Tao [2], Mintao Wei [1], Qiang Yang [1], Jianquan Li [3], Zhichao Yuan [3] and Yuequan Yang [1]

[1] School of Information and Artificial Intelligence, Yangzhou University, Yangzhou 225009, China; mx120240562@stu.yzu.edu.cn (Y.W.); mz120250992@stu.yzu.edu.cn (M.W.); mz120260993@stu.yzu.edu.cn (Q.Y.); yang@yzu.edu.cn (Y.Y.)
[2] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; taoxian2013@ia.ac.cn
[3] CASisense Technology Co., Ltd., Beijing 100192, China; lijianquan@isensecn.com (J.L.); yzc@isensecn.com (Z.Y.)
* Correspondence: fdli@yzu.edu.cn

Abstract

Image anomaly detection is an important task in industrial quality inspection and has also been investigated on medical imaging benchmarks. However, many existing methods require normal samples from the target domain, which may be difficult to obtain in practical applications. Recent zero-shot anomaly detection methods based on vision–language models have shown promising performance. However, cross-layer semantic redundancy in hierarchical visual–text alignment and the aggregation of sparse anomaly information for image-level prediction remain challenges. To address these challenges, we propose CLIP-based adaptive weighted anomaly detection (AWAD), a cross-domain zero-shot anomaly detection framework. AWAD introduces a cross-layer semantic decoupling (CLSD) module to construct stage-specific textual representations and reduce excessive cross-layer semantic similarity. An adaptive anomaly guidance (AAG) module is further developed to aggregate spatial features according to cross-modal anomaly responses for image-level discrimination. Experiments on 12 public industrial and medical imaging benchmarks and a supplementary Toothbrush Bristle Defect Dataset (TBDD) evaluate the proposed framework under cross-domain zero-shot settings. AWAD achieves a pixel-level AUROC of 95.8% and a PRO of 89.6% on VisA, together with a pixel-level AUROC of 91.6% on MVTec AD. The results show promising pixel-level anomaly localization performance on several benchmarks, while image-level detection performance varies across datasets. The medical imaging experiments serve as cross-domain transfer benchmarks and do not constitute clinical validation.

Keywords: anomaly detection; zero-shot anomaly detection; vision–language model; prompt learning

Check for updates

Academic Editor: Piotr A. Kowalski

Received: 30 September 2026
Revised: 17 October 2026
Accepted: 21 October 2026
Published: 23 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Visual anomaly detection (VAD) aims to identify and localize regions that deviate from normal patterns and plays an important role in industrial quality inspection. Related anomaly detection methods have also been investigated on medical imaging benchmarks to evaluate their applicability across different visual domains [1,2]. Unsupervised anomaly detection (UAD) has been a dominant paradigm in this field, but its performance often relies on sufficient and carefully curated normal samples from the target domain [3]. In practical applications, particularly for rapidly evolving product lines, collecting such data can be

costly and time consuming. This limitation has motivated growing interest in zero-shot anomaly detection (ZSAD), which aims to eliminate the requirement for target-domain normal training samples [4].

Recent vision–language models (VLMs) have established generalizable semantic spaces through large-scale cross-modal pre-training, providing a promising foundation for ZSAD [5–7]. To adapt these models to downstream tasks, prompt learning has been widely explored [8–10]. However, most existing prompting strategies are designed for object-level or coarse-grained semantics, whereas anomaly detection requires sensitivity to subtle local texture defects and structural distortions. This mismatch calls for more task-specific adaptation mechanisms.

Despite recent progress, VLM-based ZSAD still faces two key limitations. First, in multi-scale visual-text alignment, some methods rely on shared prompts or weakly discriminative anomaly semantics [11–14], which may produce excessively similar textual representations across stages. Here, cross-layer semantic redundancy refers to excessive similarity among stage-specific textual representations associated with visual features at different depths. Second, image-level anomaly classification often relies on a nearly uniform aggregation of patch features [15], which may cause sparse anomaly cues to be overwhelmed by dominant normal background regions. These issues reduce the reliability of both pixel-level localization and image-level prediction.

Inspired by AnomalyCLIP [11], we develop AWAD, a CLIP-based zero-shot anomaly detection framework that integrates hierarchical visual–text alignment with anomaly-guided image-level feature aggregation. We also construct a supplementary real-world Toothbrush Bristle Defect Dataset (TBDD) to assess the method in an industrial inspection setting.

AWAD consists of two main components: a hierarchical visual–text alignment mechanism with a cross-layer semantic decoupling constraint, and an AAG module. The former associates visual features at different depths with stage-specific normal and anomalous textual representations to enhance layer-wise specificity and reduce semantic redundancy. The latter uses high-level cross-modal anomaly cues to reweight local features and construct anomaly-sensitive image-level representations.

The main contributions of this paper are summarized as follows:

1. We propose AWAD, a unified CLIP-based zero-shot anomaly detection framework with cross-layer semantic decoupling and adaptive anomaly guidance, which jointly supports image-level anomaly classification and pixel-level anomaly localization.
2. We introduce a CLSD module to learn layer-specific textual representations for hierarchical visual–text alignment. By extracting stage-specific textual representations from a shared object-agnostic prompt bank and suppressing cross-layer semantic similarity, CLSD improves layer-wise semantic specificity and reduces excessive similarity between stage-specific textual representations.
3. We design an AAG module for image-level anomaly discrimination. By leveraging high-level cross-modal anomaly cues to reweight spatial features, AAG constructs anomaly-sensitive global representations and reduces the interference of dominant normal background regions.
4. We evaluate AWAD on public industrial and medical imaging benchmarks, together with the supplementary Toothbrush Bristle Defect Dataset (TBDD), to assess its performance under cross-domain zero-shot settings.

2. Related Work

2.1. Conventional Anomaly Detection

Early unsupervised anomaly detection (UAD) methods mainly modeled normality through reconstruction. Autoencoders and related variants detected anomalous regions by

measuring reconstruction errors [16,17]. Subsequent studies introduced anomaly synthesis and self-supervised learning strategies, such as CutPaste [18] and SimpleNet [19], shifting the focus from pure reconstruction to more discriminative feature learning. Nevertheless, these methods still depend on sufficient normal samples from the target domain.

With the development of stronger visual representations, feature-space modeling has become a dominant paradigm for anomaly detection. Methods such as PatchCore [20], PaDiM [21], and SPADE [22] improve anomaly discrimination through nearest-neighbor retrieval or feature distribution modeling. Recent industrial anomaly detection methods have further explored dictionary-based sparse representation for surface defect detection [23] and mask-based training with teacher–student joint memory for multimodal anomaly detection [24]. In parallel, teacher–student distillation frameworks [25–27] and multi-directional feature aggregation methods [28] have been investigated for unsupervised anomaly detection. Despite their effectiveness, these methods generally rely on target-domain normal data during training, limiting their direct applicability to the cross-domain zero-shot anomaly detection setting.

2.2. Prompt Learning and CLIP Adaptation

Vision–language models (VLMs), represented by CLIP [5], provide a unified semantic space for cross-modal transfer. To adapt pretrained VLMs to downstream tasks, prompt learning has been widely explored. Representative methods include CoOp [8] and CoCoOp [9], which learn continuous prompts and instance-aware prompts, respectively. RegionCLIP [10] extends language supervision to region-level representation learning, while multimodal adaptation methods such as MaPLe [29] and VPT [30] further improve task-specific transfer. Adapter-based tuning [31,32] provides another lightweight solution for efficient adaptation.

Although these adaptation strategies have demonstrated effectiveness in various downstream vision–language tasks, including recognition, retrieval, and relational reasoning [33], they are mainly designed to preserve object-level semantic consistency. In anomaly detection, however, the model must be sensitive to subtle local deviations and semantic variations across feature levels. As a result, general-purpose prompt or adapter tuning does not directly address the multi-scale alignment challenges required by zero-shot anomaly detection.

2.3. Zero-Shot Anomaly Detection

A central issue in zero-shot anomaly detection (ZSAD) is how to represent anomaly semantics without target-domain training data. Early methods such as WinCLIP [15] relied on handcrafted textual templates for zero-shot classification and segmentation. AnomalyCLIP [11] introduced object-agnostic prompt learning and significantly improved cross-category generalization. Bayes-PFL [34] further models the text prompt space as a learnable probability distribution from a Bayesian perspective, enabling diverse prompt sampling and improving fine-grained visual–text alignment through residual cross-attention. VAND [12] enhances visual–text matching through an additional mapping layer, while FiLo [35] and AnomalyGPT [36] enrich anomaly representation using fine-grained textual descriptions and large-model priors, respectively. SimCLIP [13] and AnoVL [14] also improve ZSAD by refining visual–text alignment and unified anomaly localization. However, these methods mainly focus on improving anomaly semantic representation and visual–text matching, while the semantic roles of hierarchical visual features are generally not explicitly differentiated during multi-scale alignment.

Another line of work aims to improve visual adaptation to local defects. AdaCLIP [37], AF-CLIP [38], PromptAD [39], and CLIP-AD [40] enhance anomaly sensitivity through learnable prompts, lightweight adapters, or dual-path designs. More recent studies, in-

cluding DLVP-CLIP [41] and FB-CLIP [42], have investigated dynamic local visual prompting and foreground–background disentanglement, respectively, to improve fine-grained anomaly perception. These studies provide complementary approaches to visual–text adaptation, while AWAD focuses on explicit cross-stage textual semantic regularization and anomaly-guided spatial aggregation.

Image-level anomaly decision making remains relatively underexplored in existing ZSAD methods. Many approaches directly use the global feature from the visual backbone or aggregate local features through global average pooling [39,40]. Although region-aware aggregation has been explored with the help of foundation segmentation models [43], sparse anomaly cues may still be suppressed by dominant normal background regions. Consequently, image-level prediction can remain insensitive to weak but critical local anomalies.

In summary, existing studies have substantially advanced ZSAD through prompt learning, visual adaptation, and cross-modal alignment. Unlike methods primarily focusing on prompt optimization or visual adaptation, AWAD addresses two specific issues: cross-stage semantic redundancy and the dilution of sparse anomaly information during image-level aggregation. Specifically, CLSD constructs stage-specific textual representations from grouped intermediate text features and introduces a margin-based constraint to reduce excessive same-state cross-stage similarity. Meanwhile, AAG adaptively weights spatial features according to high-level cross-modal anomaly responses, rather than relying on uniform aggregation. These two mechanisms jointly support hierarchical anomaly localization and image-level prediction under cross-domain zero-shot settings.

3. Method

3.1. Problem Formulation

We consider a cross-domain zero-shot anomaly detection setting, in which the model is trained on an auxiliary source domain and directly evaluated on an unseen target domain without accessing any target-domain samples during training. Let the source-domain dataset be denoted by

$$\mathcal{D}_s = \{(x_i^s, y_i^s, M_i^s)\}_{i=1}^{N_s}. \quad (1)$$

where x_i^s denotes a source-domain image, $y_i^s \in \{0, 1\}$ denotes the corresponding image-level binary anomaly label, and M_i^s represents the pixel-level anomaly mask for local supervision.

The target-domain dataset is denoted by

$$\mathcal{D}_t = \left\{x_j^t\right\}_{j=1}^{N_t}. \quad (2)$$

No target-domain samples are used for model optimization. The objective is to learn a model from $\mathcal{D}_s$ that can directly perform both image-level anomaly classification and pixel-level anomaly localization on the unseen target domain $\mathcal{D}_t$.

It should be emphasized that AWAD is not a training-free method. The learnable prompt-related parameters and task-specific modules are optimized exclusively on the auxiliary source domain, whereas inference on the unseen target domain is performed without using any target-domain samples for training or adaptation.

3.2. Overall Architecture

To jointly address pixel-level anomaly localization and image-level anomaly classification, we propose AWAD, as illustrated in Figure 1. AWAD is built upon a frozen pre-trained CLIP backbone. Given an input image, the CLIP visual encoder extracts patch-level features from multiple selected Transformer layers, providing hierarchical visual representations with different semantic granularities. In parallel, a shared object-agnostic prompt bank is processed by the frozen CLIP text encoder to generate stage-wise textual representations

of normal and anomalous states. This design enables hierarchical visual–text alignment without relying on category names or defect descriptions from the target domain.

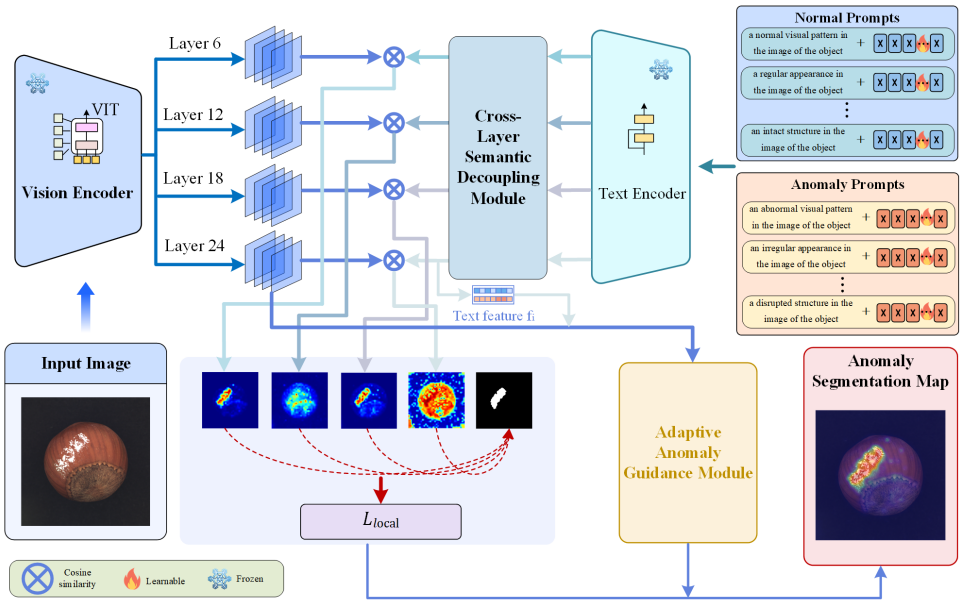


Figure 1. Overall architecture of the proposed AWAD framework, consisting of the cross-layer semantic decoupling (CLSD) and adaptive anomaly guidance (AAG) modules.

AWAD contains two main components. The CLSD module constructs stage-specific textual representations and aligns them with the corresponding visual feature levels for pixel-level anomaly localization. A margin-based semantic decoupling loss is applied to reduce excessive same-state cross-stage similarity, thereby promoting more differentiated stage-specific representations. The AAG module exploits high-level cross-modal anomaly responses to reweight spatial features and construct an anomaly-sensitive global representation for image-level classification. These components are jointly optimized under a unified objective, enabling AWAD to generate anomaly heatmaps and image-level predictions within a single forward pass while generalizing to unseen target categories without requiring target-domain samples for training or adaptation.

3.3. Cross-Layer Semantic Decoupling Module

The CLSD module constructs stage-specific normal and anomalous textual representations using a shared object-agnostic prompt bank [11]. Its implementation involves text-layer grouping, intermediate feature projection, and cross-stage semantic decoupling, as detailed below.

Specifically, we select the 6th, 12th, 18th, and 24th Transformer layers from the CLIP visual encoder. Correspondingly, the 12 layers of the CLIP text encoder are divided into four consecutive groups, namely, $G_1 = \{1, 2, 3\}$, $G_2 = \{4, 5, 6\}$, $G_3 = \{7, 8, 9\}$, and $G_4 = \{10, 11, 12\}$. These four text-stage groups are sequentially paired with the selected visual layers, forming a one-to-one hierarchical visual–text alignment scheme. In this work, a "layer" refers to an individual Transformer layer, whereas a "stage" refers to one of the four hierarchical visual–text alignment levels, each comprising a group of text Transformer layers paired with a selected visual Transformer layer.

To map intermediate textual features into the shared CLIP embedding space, we extract the end-of-text (EOT) token representation from each text Transformer layer. Each intermediate representation is processed by the pre-trained final layer normalization and text projection matrix, followed by L2 normalization. The same projection parameters

are shared across all text layers, ensuring that the resulting features have a consistent embedding dimension and can be directly compared with visual representations. The projected textual features are then aggregated within each predefined text-stage group to construct stage-specific semantic representations.

Let P_k^n and P_k^a denote the k-th normal-state prompt and anomalous-state prompt, respectively. The numbers of normal and anomalous prompts are denoted by K_n and K_a, respectively. For the l-th text-stage group, the pooled textual representation of a prompt P is computed as

$$\bar{E}_{\text{text}}^{(l)}(P) = \frac{1}{|G_l|} \sum_{r \in G_l} E_{\text{text}}^{(r)}(P), \tag{3}$$

where $|G_l|$ denotes the number of layers in G_l and $E_{\text{text}}^{(r)}(\cdot)$ denotes the textual feature extracted from the r-th layer of the CLIP text encoder.

The normal-state and anomalous-state textual representations of the l-th stage are then obtained as

$$\mathbf{t}_l^n = \text{Norm}\left(\frac{1}{K_n} \sum_{k=1}^{K_n} \text{Norm}\left(\bar{E}_{\text{text}}^{(l)}(P_k^n) \right) \right), \tag{4}$$

and

$$\mathbf{t}_l^a = \text{Norm}\left(\frac{1}{K_a} \sum_{k=1}^{K_a} \text{Norm}\left(\bar{E}_{\text{text}}^{(l)}(P_k^a) \right) \right), \tag{5}$$

where $\text{Norm}(\cdot)$ denotes ℓ_2 normalization. The two representations are grouped as

$$T_l = \begin{bmatrix} (\mathbf{t}_l^n)^{\mathrm{T}} \\ (\mathbf{t}_l^a)^{\mathrm{T}} \end{bmatrix}, \tag{6}$$

For the visual branch, intermediate patch features extracted from the 6th, 12th, 18th, and 24th Transformer layers are mapped into the shared CLIP embedding space using the pre-trained visual post-layer normalization and projection matrix. The projected features are further L2-normalized before cross-modal similarity computation. Since all selected features originate from the same ViT backbone and share an identical patch-token grid, their spatial correspondence is inherently preserved across different layers. Specifically, with an input resolution of 518×518 and a patch size of 14, each selected layer produces a 37×37 patch grid. For pixel-level anomaly localization, the CLS token is excluded from the patch-level similarity maps, and the spatial responses are reshaped and resized to the input resolution before multi-layer aggregation. Therefore, no additional spatial alignment module is required.

During visual–text matching, the patch features extracted from the l-th selected visual stage are aligned with the corresponding textual representation T_l. The resulting cross-modal similarity maps serve as local anomaly responses for pixel-level anomaly localization.

Although stage-wise textual representations are extracted from different text-layer groups, representations associated with different stages may still become excessively similar during optimization. To explicitly reduce this cross-stage redundancy, CLSD introduces a margin-based semantic decoupling loss between same-state textual representations. The overall decoupling loss is defined as

$$\mathcal{L}_{\text{dec}} = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \Big[\max\left(0, \cos(\mathbf{t}_i^n, \mathbf{t}_j^n) - m_{\text{dec}}\right) + \max\left(0, \cos(\mathbf{t}_i^a, \mathbf{t}_j^a) - m_{\text{dec}}\right) \Big]. \tag{7}$$

where m_{dec} denotes the decoupling margin, L is the total number of text stages, and $\cos(\cdot, \cdot)$ denotes cosine similarity. The loss is activated only when the cosine similarity between same-state representations from different stages exceeds m_{dec}. Therefore, it suppresses excessive cross-stage semantic redundancy while imposing no decoupling penalty on stage

pairs whose similarity falls below the margin, thereby encouraging less redundant stage-wise textual representations for hierarchical visual–text alignment. The decoupling loss constrains only same-state cross-stage similarities, while within-stage normal-anomaly discrimination is encouraged by the classification and localization objectives. Importantly, the decoupling constraint is imposed on the aggregated stage-specific textual representations rather than on individual prompt embeddings.

3.4. Adaptive Anomaly Guidance Module

While CLSD provides stage-wise local anomaly responses for pixel-level localization, image-level anomaly discrimination requires a global representation that is sensitive to sparse anomalous regions. To this end, we introduce the AAG module, which adaptively reweights patch features according to high-level cross-modal anomaly responses. Conventional aggregation strategies, such as global average pooling, assign nearly uniform importance to different spatial regions and may therefore dilute weak anomaly cues within dominant normal backgrounds. In contrast, AAG assigns larger contributions to spatial locations with stronger anomaly responses when constructing the global representation. The structure of the AAG module is illustrated in Figure 2.

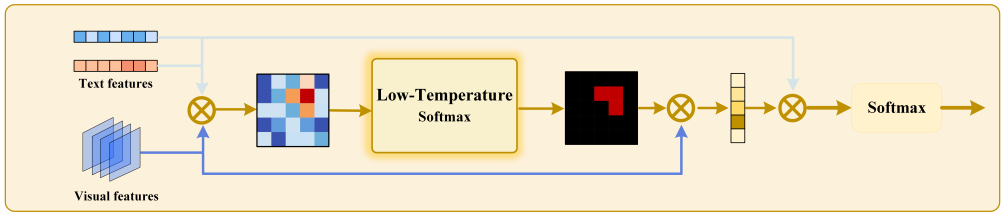


Figure 2. Structure of the AAG module.

Since AAG is designed to construct image-level discriminative representations rather than refine fine-grained anomaly boundaries, a selected high-level visual stage is used as the guidance source. By default, the 24th visual layer is adopted because it provides high-level semantic information suitable for image-level anomaly discrimination. The corresponding textual representations are taken from the fourth CLSD text stage, $G_4 = \{10, 11, 12\}$, which is paired with the 24th visual Transformer layer.

Let the patch features extracted from the selected guidance layer be denoted by $\mathcal{F} = \{\mathbf{f}_i\}_{i=1}^{N}$, where N is the number of patch tokens and $\mathbf{f}_i \in \mathbb{R}^d$ denotes the feature of the i-th patch token. Let the corresponding normal-state and anomalous-state textual representations generated by CLSD be denoted by $\mathbf{t}^n \in \mathbb{R}^d$ and $\mathbf{t}^a \in \mathbb{R}^d$, respectively. To obtain spatial anomaly responses, we compute the cosine similarity between each patch feature and the anomalous-state textual representation as

$$s_i = \cos(\mathbf{f}_i, \mathbf{t}^a), \quad i = 1, 2, \ldots, N, \tag{8}$$

where s_i denotes the anomaly response of the i-th spatial position.

A temperature-controlled Softmax operation is then applied to normalize the spatial anomaly responses:

$$\beta_i = \frac{\exp(s_i/\tau)}{\sum_{j=1}^{N} \exp(s_j/\tau)}, \tag{9}$$

where β_i denotes the normalized spatial weight of the i-th patch and τ is the temperature parameter controlling the sharpness of the spatial weight distribution. A smaller value of τ produces a more concentrated weight distribution, allowing high-response regions to contribute more strongly to the subsequent global aggregation.

Based on these spatial weights, AAG constructs an anomaly-sensitive global visual representation through weighted aggregation:

$$\mathbf{g} = \sum_{i=1}^{N} \beta_i \mathbf{f}_i, \tag{10}$$

where $\mathbf{g} \in \mathbb{R}^d$ denotes the reconstructed global representation.

Finally, $\mathbf{g}$ is matched with the corresponding normal-state and anomalous-state textual representations to obtain the image-level prediction logits. A fixed temperature-based logit scaling operation is applied to the cosine similarity scores:

$$z^n = \frac{\cos(\mathbf{g}, \mathbf{t}^n)}{T_{\text{cls}}}, \qquad z^a = \frac{\cos(\mathbf{g}, \mathbf{t}^a)}{T_{\text{cls}}}, \tag{11}$$

and

$$\mathbf{z} = \begin{bmatrix} z^n \\ z^a \end{bmatrix}, \tag{12}$$

where z^n and z^a denote the scaled logits corresponding to the normal and anomalous states, respectively. T_{cls} is a fixed classification temperature, set to 0.07 in our implementation. This operation is equivalent to applying a fixed logit scale of $1/T_{\text{cls}}$ and is independent of the spatial weighting temperature τ used in Equation (9). The resulting logits are subsequently used for image-level anomaly discrimination.

3.5. Training Objective and Inference

With the outputs of the CLSD and AAG modules, AWAD is optimized using a unified objective that jointly considers image-level anomaly discrimination, pixel-level anomaly localization, and cross-layer semantic decoupling:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{global}} + \lambda_{\text{loc}} \mathcal{L}_{\text{local}} + \lambda_{\text{dec}} \mathcal{L}_{\text{dec}}, \tag{13}$$

where $\mathcal{L}_{\text{global}}$ denotes the image-level classification loss, $\mathcal{L}_{\text{local}}$ denotes the multi-scale local supervision loss, and $\mathcal{L}_{\text{dec}}$ denotes the cross-layer semantic decoupling loss introduced in the CLSD module. The weight of the global classification loss $\mathcal{L}_{\text{global}}$ is fixed to 1, while the coefficients λ_{loc} and λ_{dec} control the relative contributions of the local supervision loss and the semantic decoupling loss, respectively.

For pixel-level anomaly localization, each selected visual layer produces paired prediction maps corresponding to the normal and anomalous states. The ground-truth anomaly mask is resized to the spatial resolution of the l-th layer and denoted by $Y_l \in \{0,1\}^{H_l \times W_l}$, where 1 and 0 indicate anomalous and normal pixels, respectively. Let $S_l = [S_l^n, S_l^a] \in [0,1]^{2 \times H_l \times W_l}$ denote the normalized prediction maps of the l-th layer, where S_l^n and S_l^a represent the normal-state and anomalous-state prediction maps, respectively. The local supervision loss for the l-th layer is defined as

$$\mathcal{L}_{\text{loc}}^l = \mathcal{L}_{\text{focal}}([S_l^n, S_l^a], Y_l) + \mathcal{L}_{\text{dice}}(S_l^a, Y_l) + \mathcal{L}_{\text{dice}}(S_l^n, 1 - Y_l). \tag{14}$$

The focal loss reduces the contribution of well-classified pixels and emphasizes difficult samples during local anomaly localization:

$$\mathcal{L}_{\text{focal}}([S_l^n, S_l^a], Y_l) = -\frac{1}{N} \sum_{i=1}^{N} \Big[Y_{l,i} (1 - S_{l,i}^a)^\gamma \log(S_{l,i}^a + \epsilon_1) + (1 - Y_{l,i})(1 - S_{l,i}^n)^\gamma \log(S_{l,i}^n + \epsilon_1) \Big]. \tag{15}$$

The Dice losses are applied to both anomalous-state and normal-state prediction maps to improve region-level consistency:

$$\mathcal{L}_{\text{dice}}(S_l^a, Y_l) = 1 - \frac{2\sum_{i=1}^{N} S_{l,i}^a Y_{l,i}}{\sum_{i=1}^{N} S_{l,i}^a + \sum_{i=1}^{N} Y_{l,i} + \epsilon_2}, \tag{16}$$

and

$$\mathcal{L}_{\text{dice}}(S_l^n, 1 - Y_l) = 1 - \frac{2\sum_{i=1}^{N} S_{l,i}^n (1 - Y_{l,i})}{\sum_{i=1}^{N} S_{l,i}^n + \sum_{i=1}^{N} (1 - Y_{l,i}) + \epsilon_2}. \tag{17}$$

Here, $N = H_l \times W_l$ denotes the number of pixels in the prediction map of the l-th layer, i denotes the pixel index, and $S_{l,i}^a$ and $S_{l,i}^n$ denote the anomalous-state and normal-state prediction scores of the i-th pixel, respectively. The parameter γ is the focusing factor, while ϵ_1 and ϵ_2 are small constants introduced for numerical stability. The multi-scale local supervision loss is obtained by summing the local losses over all selected visual layers:

$$\mathcal{L}_{\text{local}} = \sum_{l=1}^{L} \mathcal{L}_{\text{loc}}^l. \tag{18}$$

At the image level, the AAG module produces prediction logits $\mathbf{z}$ by matching the anomaly-sensitive global representation with the corresponding normal-state and anomalous-state textual representations. The image-level classification loss is defined as

$$\mathcal{L}_{\text{global}} = \mathcal{L}_{\text{ce}}(\mathbf{z}, y), \tag{19}$$

where $L_{\text{ce}}(\cdot)$ denotes the cross-entropy loss and $y \in \{0, 1\}$ denotes the image-level anomaly label.

During inference, AWAD simultaneously outputs an image-level anomaly score and a pixel-level anomaly heatmap in a single forward pass. The image-level anomaly probability is obtained by applying Softmax to the AAG prediction logits:

$$p^a = \frac{\exp(z^a)}{\exp(z^n) + \exp(z^a)}, \tag{20}$$

where p^a denotes the predicted image-level anomaly probability.

For pixel-level anomaly localization, the anomaly response maps generated from all selected visual layers are fused using a parameter-free summation strategy:

$$A = \sum_{l=1}^{L} A_l, \tag{21}$$

where A_l denotes the anomaly response map obtained from visual–text matching at the l-th selected visual layer and resized to the input image resolution, and A denotes the fused anomaly map. Spatial smoothing is subsequently applied to A to suppress local prediction noise and obtain the final pixel-level anomaly heatmap.

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate the zero-shot anomaly detection performance of AWAD on 12 public benchmark datasets and one newly constructed real-world dataset. To assess cross-domain generalization, these datasets are grouped into two domains, namely, industrial and medical anomaly detection.

As summarized in Table 1, the industrial datasets include MVTec AD [3], VisA [44], MPDD [45], BTAD [46], SDD [47], our self-constructed TBDD, and DTD-Synthetic [48], covering diverse object categories, complex textures, and defect patterns. The medical datasets include ISIC [49], CVC-ClinicDB [50], ENDO [51], HeadCT [52], BrainMRI [52],

and Br35H [53], spanning skin lesion analysis, colonoscopy imaging, and brain tumor detection. These datasets exhibit substantial variations in appearance, imaging modality, and anomaly characteristics, providing a challenging benchmark for evaluating cross-domain generalization.

Table 1. Summary of the industrial and medical anomaly detection datasets evaluated in our experiments.

Dataset	Category	Modality	$\|C\|$	Normal	Anomalous
MVTec AD	Object and texture	Photography	15	467	1258
VisA	Object	Photography	12	962	1200
MPDD		Photography	6	176	282
BTAD		Photography	3	451	290
SDD		Photography	1	347	52
TBDD		Photography	1	40	249
DTD-Synthetic	Texture	Photography	12	357	947
ISIC	Skin	Photography	1	0	379
CVC-ClinicDB	Colon	Endoscopy	1	0	612
ENDO		Endoscopy	1	0	200
HeadCT	Brain	Radiology (CT)	1	100	100
BrainMRI		Radiology (MRI)	1	98	155
Br35H		Radiology (MRI)	1	1500	1500

The licensing conditions and intended research uses of the evaluated datasets are summarized in Table 2. All datasets are used in this study for academic research and benchmark evaluation. The industrial datasets are employed for source-domain training or industrial anomaly detection evaluation, while the medical imaging datasets are used exclusively as cross-domain transfer benchmarks. No medical target-domain samples are used for model training or adaptation. For datasets with specific licensing or usage restrictions, the original providers' terms remain applicable. In particular, the medical imaging experiments are intended to assess cross-domain generalization rather than clinical diagnostic performance.

In addition to the public benchmarks, we construct a real-world industrial test set for toothbrush defect inspection, termed TBDD, to evaluate the proposed method under practical conditions. Sample images from TBDD are shown in Figure 3. TBDD contains 40 normal samples and 249 anomalous samples covering five defect types: Oil Stain, Long Bristle, Whorled Bristles, Missing Tuft, and Short-filled Tuft. The images were collected under a fixed industrial inspection setup and manually annotated at the image level. Since pixel-level annotations are not available, TBDD is used as a supplementary image-level evaluation dataset rather than as primary evidence for the main conclusions. The dataset and annotations will be released upon publication, subject to institutional approval. Given the imbalanced class distribution of TBDD, I-AUROC is adopted as the primary image-level evaluation metric, with I-AP reported as a supplementary metric. All reported TBDD baselines are evaluated on the same test images under the cross-domain zero-shot protocol.

To evaluate AWAD under a focused comparison setting, we select representative CLIP-based zero-shot anomaly detection baselines, with particular emphasis on AnomalyCLIP as the underlying framework of the proposed method. The comparison is not intended to exhaustively cover all recent ZSAD approaches, and the conclusions are restricted to the evaluated baselines and experimental settings. WinCLIP [15] represents vision–language inference based on handcrafted prompts. AnomalyCLIP [11] represents object-agnostic

prompt learning. CLIP-AC [15], introduced in the WinCLIP study, serves as a CLIP-based anomaly classification baseline with two-class prompts. VAND [12] represents joint vision–language feature mapping. Bayes-PFL [34] represents Bayesian prompt distribution learning for zero-shot anomaly detection. CoOp [8] is included as a basic learnable prompt optimization baseline.

Table 2. Licensing conditions and intended research uses of the datasets evaluated in this study.

Dataset	License or Usage Conditions	Intended Use in This Study
Industrial datasets		
MVTec AD	CC BY-NC-SA 4.0	Source-domain training or industrial evaluation
VisA	CC BY 4.0	Source-domain training or industrial evaluation
MPDD	CC BY-NC-SA 4.0	Industrial defect localization benchmark
BTAD	CC BY-SA	Industrial defect detection benchmark
SDD	CC BY-NC-SA 4.0	Surface defect localization benchmark
DTD-Synthetic	Source-specific terms; separate license not verified	Synthetic texture anomaly benchmark
TBDD	Release subject to institutional approval	Supplementary image-level industrial evaluation
Medical imaging datasets		
ISIC	CC0 (2016 Challenge)	Skin lesion localization transfer benchmark
CVC-ClinicDB	Research and educational use only; non-commercial	Endoscopic polyp localization transfer benchmark
ENDO	Provider-specific terms; separate license not verified	Endoscopic polyp localization transfer benchmark
HeadCT	CC0	Head CT image-level transfer benchmark
BrainMRI	Data files copyright original authors	Brain MRI image-level transfer benchmark
Br35H	CC BY 4.0 (reported in prior literature)	Brain MRI image-level transfer benchmark

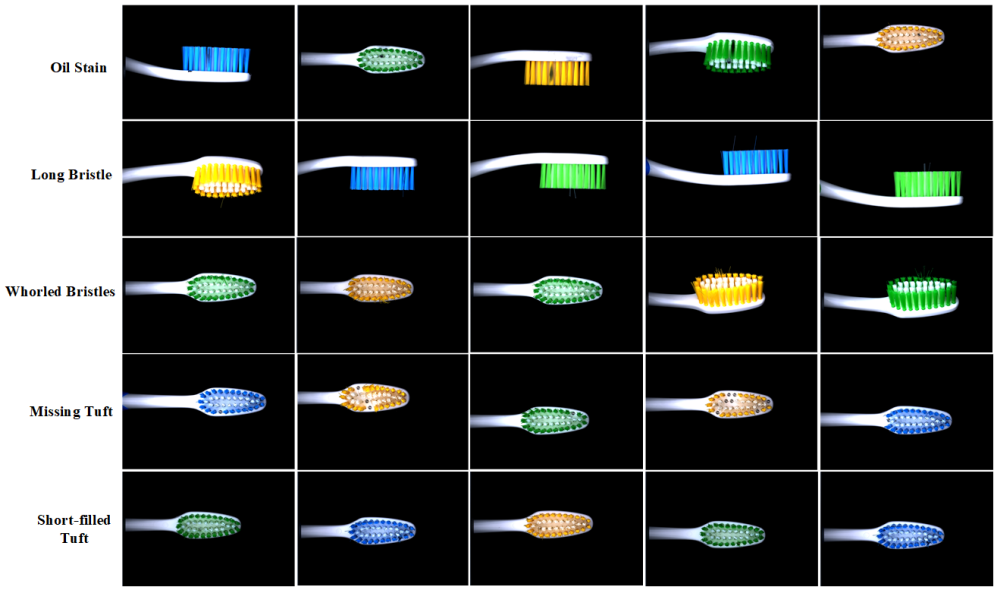


Figure 3. Sample images from our self-constructed Toothbrush Bristle Defect Dataset (TBDD).

For quantitative comparison, the results of CoOp and CLIP-AC, as well as those of WinCLIP and VAND on public benchmarks, are obtained from published references. AnomalyCLIP and Bayes-PFL are independently reproduced under our experimental environment, while WinCLIP and VAND are additionally evaluated on the newly constructed TBDD dataset. The independently reproduced AnomalyCLIP and Bayes-PFL results serve as our main experimental reference points, while results adopted from published studies provide broader contextual comparisons. Since the latter were not re-evaluated under a fully unified experimental setup, numerical differences involving these methods should be interpreted cautiously.

The symbol "–" in Table 3 denotes results not reported in the comparison. Specifically, the CoOp results on DTD-Synthetic are not provided in the adopted published comparison data. For TBDD, quantitative results are reported for WinCLIP, VAND, AnomalyCLIP, and Bayes-PFL, while those for CoOp and CLIP-AC are not included.

Table 3. ZSAD performance comparison on industrial anomaly detection datasets. The best results are shown in bold, and the second-best results are underlined. The superscript † indicates results independently reproduced in our experiments; unmarked baseline results are taken from published references. The symbol "–" denotes results not reported in the comparison.

Task	Category	Dataset	$\|C\|$	CoOp	CLIP-AC	WinCLIP	VAND	AnomalyCLIP †	Bayes-PFL †	AWAD (Ours)
Image-level (AUROC, AP)	Obj. and Texture	MVTec AD	15	(88.8, 94.8)	(71.5, 86.4)	(91.8, **96.5**)	(86.1, 93.5)	(91.5, 96.2)	(**92.3**, 95.8)	(89.6, 96.3)
	Object	VisA	12	(62.8, 68.1)	(65.0, 70.1)	(78.1, 81.2)	(78.0, 81.4)	(81.5, 84.8)	(**86.4**, 85.6)	(82.9, **85.7**)
		MPDD	6	(55.1, 64.2)	(56.2, 66.0)	(63.6, 69.6)	(73.0, 80.2)	(76.7, 81.5)	(74.9, 80.4)	(**77.3**, **82.6**)
		BTAD	3	(66.8, 77.4)	(51.0, 62.1)	(68.2, 70.9)	(73.6, 68.6)	(86.3, 88.2)	(**92.5**, 90.2)	(91.1, **92.3**)
		SDD	1	(74.9, 65.1)	(65.2, 45.7)	(84.3, 77.4)	(79.8, 71.4)	(89.2, 83.9)	(97.0, 84.5)	(**98.4**, **91.3**)
	Texture	DTD-Synthetic	12	(–, –)	(66.8, 83.2)	(93.2, 92.6)	(86.4, 95.0)	(93.4, 97.0)	(**94.7**, 96.8)	(93.9, **97.3**)
	–	TBDD	1	(–, –)	(–, –)	(42.2, 85.6) †	(89.0, 98.1) †	(74.9, 95.8)	(90.2, 98.3)	(**92.5**, **98.6**)
Pixel-level (AUROC, PRO)	Obj. and Texture	MVTec AD	15	(33.3, 6.7)	(38.2, 11.6)	(85.1, 64.6)	(87.6, 44.0)	(91.0, 81.7)	(91.3, **87.7**)	(**91.6**, 85.4)
	Object	VisA	12	(24.2, 3.8)	(47.8, 17.3)	(79.6, 56.8)	(94.2, 86.8)	(95.4, 86.4)	(95.4, 89.0)	(**95.8**, **89.6**)
		MPDD	6	(15.4, 2.3)	(58.7, 29.1)	(76.4, 48.9)	(94.1, 83.2)	(96.4, 88.4)	(97.0, 90.0)	(**97.2**, **91.7**)
		BTAD	3	(28.6, 3.8)	(32.8, 8.3)	(72.7, 27.3)	(60.8, 25.0)	(94.1, 74.6)	(91.2, 76.0)	(**95.1**, **79.6**)
		SDD	1	(28.9, 7.1)	(32.5, 5.8)	(68.8, 24.2)	(79.8, 65.1)	(90.5, 75.4)	(90.4, 77.6)	(**93.9**, **85.9**)
	Texture	DTD-Synthetic	12	(–, –)	(23.7, 5.5)	(83.9, 57.8)	(95.3, 86.9)	(97.8, 91.5)	(98.2, 94.5)	(**98.4**, **94.7**)

Performance is evaluated at both the image level and the pixel level. For image-level classification, I-AUROC and I-AP are adopted. For pixel-level localization, P-AUROC and PRO are used. P-AUROC measures pixel-wise anomaly discrimination, while PRO evaluates the overlap of individual connected anomaly regions and assigns equal importance to each region, reducing the dominance of larger anomalous regions. Therefore, P-AUROC and PRO are considered complementary metrics for evaluating pixel-level anomaly localization performance.

4.2. Implementation Details

We adopt the pre-trained CLIP ViT-L/14@336px (OpenAI, San Francisco, CA, USA) as the backbone vision–language model. Its 24-layer visual Transformer enables hierarchical feature extraction from layers 6, 12, 18, and 24 while preserving pretrained vision–language alignment. Both encoders remain frozen during training and inference, with only prompt-related parameters optimized during training. We follow a cross-domain zero-shot protocol: MVTec AD serves as the source domain for all other datasets, while VisA is used when evaluating MVTec AD, ensuring no target-domain samples are used for training. All images are resized to 518 × 518, with visual tokens generated from the corresponding patch grid and positional embeddings bilinearly interpolated to match the resolution. Following AnomalyCLIP, we retain DPAM in the visual encoder, while CLSD and AAG constitute the main architectural modifications of AWAD.

During training, we use the Adam optimizer with an initial learning rate of 0.001, a batch size of 8, and 15 training epochs. In the joint objective, the local losses from the four selected visual layers are summed, and the corresponding weight is set to $\lambda_{\text{loc}} = 4$. The weight of the cross-layer semantic decoupling loss is set to $\lambda_{\text{dec}} = 0.1$. In the AAG module, the temperature parameter controlling the sharpness of the spatial weight distribution is fixed at $\tau = 0.05$.

The prompt context length is set to 12, with a compound prompt depth of 9 and a compound text prompt length of 4. The shared prompt bank contains $K_n = 8$ normal-state prompts and $K_a = 8$ anomalous-state prompts. The decoupling margin is set to $m_{\text{dec}} = 0.3$, while the focusing factor of the focal loss is set to $\gamma = 2$. The random seed is fixed to 111. The Adam optimizer is configured with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The complete set of 16 object-agnostic prompt templates is provided in Appendix A to facilitate reproducibility.

During inference, the image-level anomaly score is obtained directly through the AAG pathway. For pixel-level anomaly localization, anomaly maps are first generated from the four selected visual layers, then fused using a parameter-free summation strategy, and finally smoothed by Gaussian filtering with $\sigma = 4$. All experiments are conducted on a single NVIDIA RTX 4090D GPU with 24 GB of memory. The software environment is based on PyTorch 2.0.0, Python 3.8, and CUDA 11.8.

To evaluate the computational efficiency of AWAD, we measure the model complexity and inference performance with an input resolution of 518 × 518 and a batch size of 1. The complete model contains 434.311 million parameters, of which 5.555 million (approximately 1.28%) are trainable, with a computational cost of 1186.15 GFLOPs. The average inference latency is 78.28 ± 0.22 ms per image, corresponding to a throughput of 12.77 images per second. The peak GPU memory consumption during inference is 2071.0 MiB. These results characterize the computational requirements of AWAD for joint image-level anomaly classification and pixel-level anomaly localization.

4.3. Quantitative Evaluations

We evaluate AWAD on benchmarks covering industrial defects, complex textures, and medical imaging. AWAD is compared with representative anomaly detection methods based on vision–language models. The quantitative results are reported in Tables 3 and 4, while the qualitative results are presented in Figures 4–6. Overall, AWAD shows dataset-dependent performance in image-level anomaly discrimination and pixel-level anomaly localization across the evaluated benchmarks.

In industrial defect detection and complex texture scenarios, AWAD achieves competitive performance in both image-level anomaly discrimination and pixel-level anomaly localization. At the image level, AWAD obtains the highest reported I-AUROC and I-AP on SDD and TBDD among the compared methods, while its I-AP advantage on VisA is marginal. However, on MVTec AD, AWAD achieves an I-AUROC of 89.6%, which is lower than those of WinCLIP (91.8%), AnomalyCLIP (91.5%), and Bayes-PFL (92.3%). On VisA, although AWAD achieves an I-AP of 85.7%, the improvement over Bayes-PFL (85.6%) is only 0.1 percentage points, while its I-AUROC remains lower (82.9% versus 86.4%). These results indicate that the image-level performance advantages of AWAD are dataset-dependent and that marginal numerical improvements should be interpreted cautiously. At the pixel level, AWAD obtains the highest reported P-AUROC on the six evaluated industrial datasets and the highest reported PRO on VisA, MPDD, BTAD, SDD, and DTD-Synthetic, although some numerical differences are small. On VisA, AWAD achieves a P-AUROC of 95.8%, compared with 95.4% for both the independently reproduced AnomalyCLIP and Bayes-PFL baselines. Its PRO reaches 89.6%, compared with 86.4% and 89.0% for these two methods, respectively. On MVTec AD, AWAD achieves the highest P-AUROC of 91.6, slightly out-

performing Bayes-PFL (91.3). However, its PRO of 85.4 is lower than that of Bayes-PFL (87.7), indicating that improved pixel-wise discrimination does not necessarily translate into better region-level overlap. Nevertheless, AWAD maintains competitive overall localization performance across diverse industrial objects, textures, and defect patterns.

Table 4. ZSAD performance comparison on medical anomaly detection datasets. The best results are shown in bold, and the second-best results are underlined. The superscript † indicates results independently reproduced in our experiments; unmarked baseline results are taken from published references.

Task	Category	Dataset	|C|	CoOp	CLIP-AC	WinCLIP	VAND	AnomalyCLIP †	Bayes-PFL †	AWAD (Ours)
Image-level (AUROC, AP)	Brain	HeadCT	1	(78.4, 78.8)	(60.0, 60.7)	(81.8, 80.2)	(89.1, 89.4)	(92.4, 89.9)	(94.7, 94.3)	(**94.9**, **94.6**)
		BrainMRI	1	(61.3, 44.9)	(80.6, 86.4)	(86.6, 91.5)	(89.3, 90.9)	(90.8, 92.3)	(94.3, 90.0)	(**94.7**, **94.0**)
		Br35H	1	(86.0, 87.5)	(82.7, 81.3)	(80.5, 82.2)	(93.1, 92.9)	(94.8, 95.1)	(96.3, **96.2**)	(**96.5**, 92.9)
Pixel-level (AUROC, PRO)	Skin	ISIC	1	(51.7, 15.9)	(36.0, 7.7)	(83.3, 55.1)	(89.4, 77.2)	(89.3, 78.0)	(**92.2**, **83.8**)	(90.0, 79.7)
	Colon	CVC-ClinicDB	1	(34.8, 2.4)	(48.5, 12.6)	(51.2, 13.8)	(80.5, 60.7)	(81.3, 61.9)	(**84.7**, 67.8)	(83.0, **68.0**)
		ENDO	1	(40.6, 3.9)	(46.6, 12.6)	(68.2, 28.3)	(81.9, 54.9)	(83.5, **63.3**)	(84.4, 63.0)	(**84.5**, 62.9)

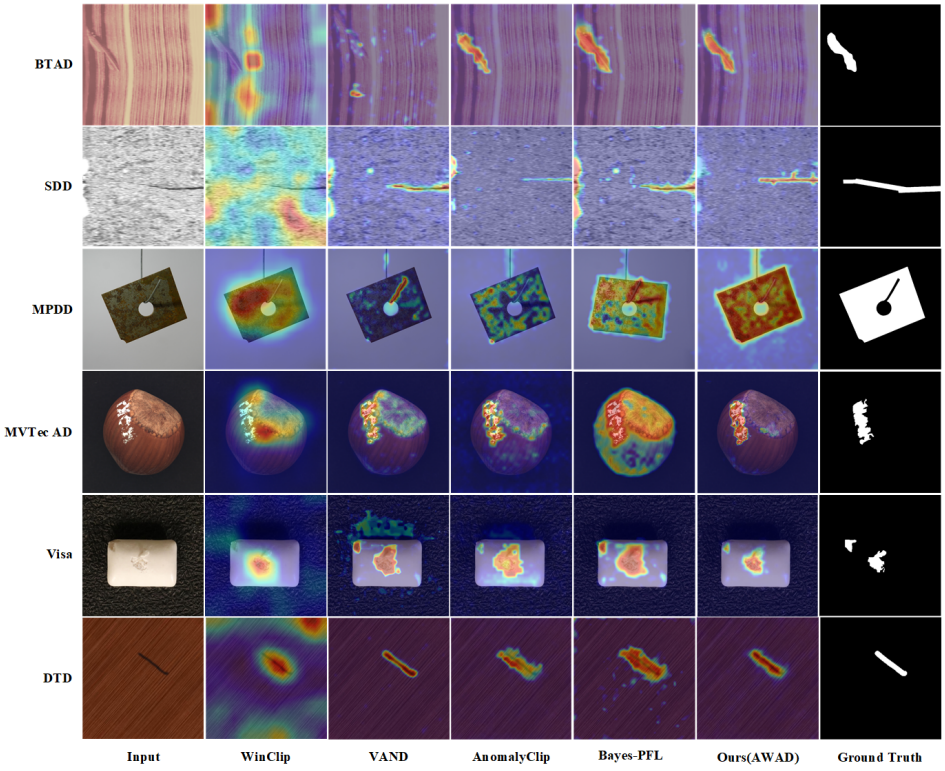


Figure 4. Qualitative comparison of zero-shot anomaly localization results on industrial anomaly detection datasets.

AWAD also maintains competitive cross-domain performance on medical imaging datasets. At the image level, AWAD achieves the best AUROC on HeadCT, BrainMRI, and Br35H, as well as the best AP on HeadCT and BrainMRI. However, on Br35H, AWAD achieves an I-AP of 92.9%, which is lower than the 96.2% obtained by Bayes-PFL, despite its slightly higher I-AUROC (96.5% versus 96.3%). This further indicates that improvements in one evaluation metric do not necessarily translate into consistent gains across other metrics. For pixel-level anomaly localization, AWAD achieves the best PRO on CVC-ClinicDB and the highest P-AUROC on ENDO. However, on ISIC, AWAD achieves a P-AUROC of 90.0 and a PRO of 79.7, compared with 92.2 and 83.8 for Bayes-PFL, respectively. This indicates that AWAD still has room for improvement in pixel-level localization on certain medical imaging datasets. Overall, these results demonstrate the cross-domain applicability of

AWAD, while also highlighting that its performance advantages vary across datasets and evaluation metrics.

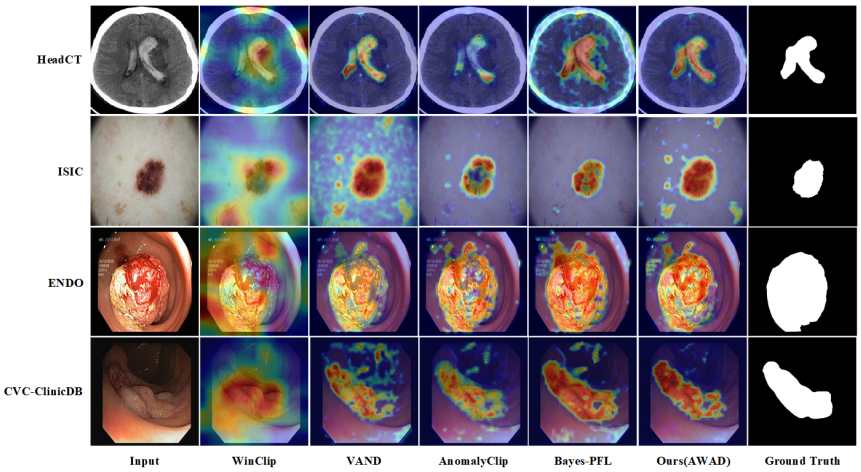


Figure 5. Qualitative comparison of zero-shot anomaly localization results on medical anomaly detection datasets.

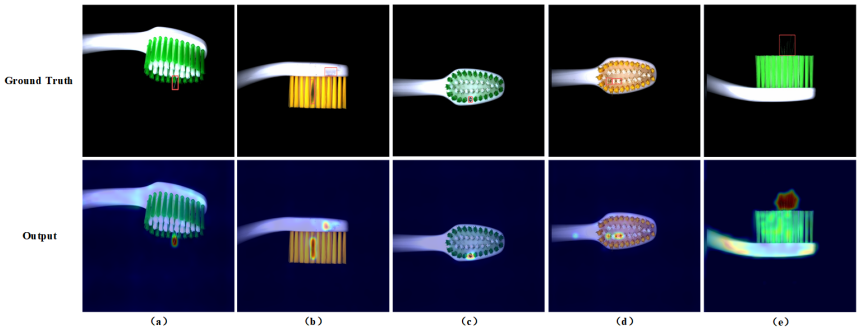


Figure 6. Qualitative anomaly localization results of AWAD on TBDD: (**a**) Long Bristle; (**b**) Oil Stain; (**c**) Missing Tuft; (**d**) Short-filled Tuft; (**e**) Whorled Bristles. The first and second rows show reference boxes and localization outputs, respectively.

The qualitative visualizations complement the quantitative results. As shown in Figures 4 and 5, some baseline methods exhibit spurious background activations, fragmented anomaly responses, or inaccurate localization in the displayed examples involving complex backgrounds and periodic textures. For the displayed VisA and DTD-Synthetic samples, AWAD produces localized anomaly responses with comparatively limited activation in normal background regions. Figure 6 further presents qualitative results on TBDD, illustrating the anomaly responses produced by AWAD on independently collected industrial inspection images. Since the current version of TBDD does not provide pixel-level annotations, coarse bounding-box annotations are manually provided for a small subset of representative anomalous samples for visualization only. These annotations are not involved in model training or quantitative evaluation. In Figure 6, (a)–(e) correspond to Long Bristle, Oil Stain, Missing Tuft, Short-filled Tuft, and Whorled Bristles, respectively.

4.4. Ablation Studies

We conduct incremental ablation studies to assess the contributions of the Prompt Bank, CLSD, and AAG modules, followed by a sensitivity analysis of the AAG temperature coefficient τ.

4.4.1. Component-Wise Ablation Study

To evaluate the contribution of each component in the proposed AWAD framework, we conduct an incremental ablation study by progressively incorporating the Prompt Bank, CLSD, and AAG modules. As shown in Table 5, the baseline corresponds to our independently reproduced AnomalyCLIP results. The subsequent configurations are evaluated to examine the effects of the proposed components on image-level anomaly discrimination and pixel-level anomaly localization.

Table 5. Incremental ablation study of different components in AWAD. "Img" and "Pix" denote image-level (AUROC, AP) and pixel-level (AUROC, PRO) metrics, respectively. HeadCT reports only image-level metrics, whereas ISIC reports only pixel-level metrics. The baseline results are obtained from our independent reproduction of AnomalyCLIP. The best results are shown in bold, and the second-best results are underlined.

Module	MVTec AD		VisA		BTAD		MPDD		HeadCT	ISIC
	Img	Pix	Img	Pix	Img	Pix	Img	Pix	Img	Pix
Baseline	(**91.5**, 96.2)	(91.0, 81.7)	(81.5, 84.8)	(95.4, 86.4)	(86.3, 88.2)	(94.1, 74.6)	(76.7, 81.5)	(96.4, 88.4)	(92.4, 89.9)	(89.3, 78.0)
+ Prompt Bank	(88.1, 95.9)	(91.1, 82.0)	(81.9, 85.1)	(95.2, 87.3)	(84.8, 88.4)	(94.1, 75.3)	(76.3, 81.3)	(96.5, 90.4)	(94.0, 92.1)	(89.6, 78.6)
+ Prompt Bank + CLSD	(91.2, 96.1)	(91.3, 84.1)	(82.3, **85.7**)	(95.6, 88.6)	(85.3, 90.1)	(94.7, 77.3)	(76.9, 81.7)	(96.7, 91.1)	(94.2, 93.0)	(**90.0**, **79.9**)
+ Prompt Bank + CLSD + AAG	(89.6, **96.3**)	(**91.6**, **85.4**)	(**82.9**, **85.7**)	(**95.8**, **89.6**)	(**91.1**, **92.3**)	(**95.1**, **79.6**)	(**77.3**, **82.6**)	(**97.2**, **91.7**)	(**94.9**, **94.6**)	(**90.0**, 79.7)

Introducing the Prompt Bank alone produces dataset-dependent effects. For example, the image-level AUROC on MVTec AD decreases from 91.5% to 88.1%. One possible explanation is that additional learnable prompts may alter visual–text alignment without stage-specific semantic constraints. After incorporating CLSD, the I-AUROC recovers to 91.2%, although it remains below the reproduced baseline, suggesting that stage-specific modeling may partially mitigate this degradation. After incorporating CLSD, pixel-level localization performance improves across the evaluated datasets. Specifically, the PRO increases from 82.0% to 84.1% on MVTec AD, from 87.3% to 88.6% on VisA, and from 75.3% to 77.3% on BTAD. These results suggest that stage-wise textual representations and cross-layer semantic constraints can provide useful guidance for hierarchical anomaly localization.

The introduction of AAG leads to dataset-dependent changes in image-level anomaly detection performance. Compared with the CLSD-based configuration, the image-level AUROC on BTAD increases from 85.3% to 91.1%, whereas that on MVTec AD decreases from 91.2% to 89.6%. Interestingly, despite the decrease in image-level performance on MVTec AD, the pixel-level PRO improves from 84.1% to 85.4%. These observations suggest a potential trade-off between emphasizing localized anomaly responses and preserving global discriminative information. In particular, anomaly-guided spatial weighting may overemphasize high-response regions while reducing the contribution of contextual features useful for image-level discrimination. Although AAG does not directly participate in pixel-level anomaly map generation during inference, the image-level and pixel-level pathways share learnable prompt representations during training. Gradients from the image-level classification loss can therefore update these shared representations through the AAG pathway, indirectly affecting pixel-level anomaly localization. Importantly, both the CLSD-only and CLSD+AAG configurations employ the same image-level classification supervision; thus, AAG does not introduce an additional supervision objective but changes the feature aggregation and gradient propagation within the existing joint optimization framework. Overall, the effectiveness of AAG varies with dataset characteristics and the distribution of anomaly responses.

Compared with the independently reproduced baseline, the complete AWAD framework improves the PRO from 81.7% to 85.4% on MVTec AD, from 86.4% to 89.6% on VisA,

and from 74.6% to 79.6% on BTAD. At the image level, the AUROC on BTAD increases from 86.3% to 91.1%, while that on MVTec AD decreases from 91.5% to 89.6%. Overall, these findings indicate that the proposed framework achieves favorable improvements in pixel-level anomaly localization on several datasets, whereas its image-level discrimination performance remains dependent on dataset characteristics and anomaly distributions.

To isolate the contribution of L_{dec}, we compare two configurations that differ only in the decoupling loss while keeping the remaining components unchanged. As shown in Table 6, introducing L_{dec} improves I-AUROC from 89.6% to 91.2% and PRO from 83.0% to 84.1% on MVTec AD, while BTAD achieves improvements of 1.8 and 1.3 percentage points in I-AP and PRO, respectively. These results indicate metric-dependent benefits of the decoupling constraint.

Table 6. Controlled ablation study of the decoupling loss. Both configurations use identical prompt banks, stage grouping, and layer pooling without AAG. The best results are shown in bold.

Configuration	MVTec AD				BTAD			
	I-AUROC	**I-AP**	**P-AUROC**	**PRO**	**I-AUROC**	**I-AP**	**P-AUROC**	**PRO**
w/o L_{dec}	89.6	95.6	**91.4**	83.0	**85.3**	88.3	**94.7**	76.0
w/ L_{dec}	**91.2**	**96.1**	91.3	**84.1**	**85.3**	**90.1**	**94.7**	**77.3**

4.4.2. Analysis of Cross-Layer Semantic Decoupling

As shown in Figure 7, the mean inter-stage cosine similarities of normal and anomalous textual representations decrease from 0.975 and 0.976 at initialization to 0.267 and 0.236 after joint training with L_{dec}, respectively. The overall post-training mean similarity is 0.252, indicating substantially reduced cross-stage similarity during joint optimization. The margin of 0.3 serves as the activation threshold of the hinge loss rather than a target similarity value. Once the similarity of a stage pair falls below the margin, its corresponding decoupling penalty becomes inactive. However, the stage-specific textual representations can still be updated by the image-level classification and pixel-level localization losses, allowing their pairwise similarities to change further. The controlled ablation in Table 6 further examines the independent contribution of the decoupling loss to detection performance.

4.4.3. Sensitivity Analysis of the Temperature Coefficient

To analyze the effect of the temperature coefficient τ in the AAG module, we conduct sensitivity experiments under $\tau \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$. Except for the temperature coefficient, all other training and inference settings remain consistent with those of the full model. As reported in Table 7, the influence of τ varies across datasets, with noticeable fluctuations in image-level detection performance on BTAD and HeadCT. A smaller τ produces a sharper spatial weight distribution, emphasizing high-response regions but potentially reducing contextual information, whereas a larger τ results in smoother spatial weighting and weaker anomaly concentration. The default temperature of 0.05 was selected based on cross-domain evaluations between MVTec AD and VisA and then consistently applied to all subsequent experiments without dataset-specific adjustment. Nevertheless, this empirical selection does not establish the optimality of the parameter under an independent validation protocol.

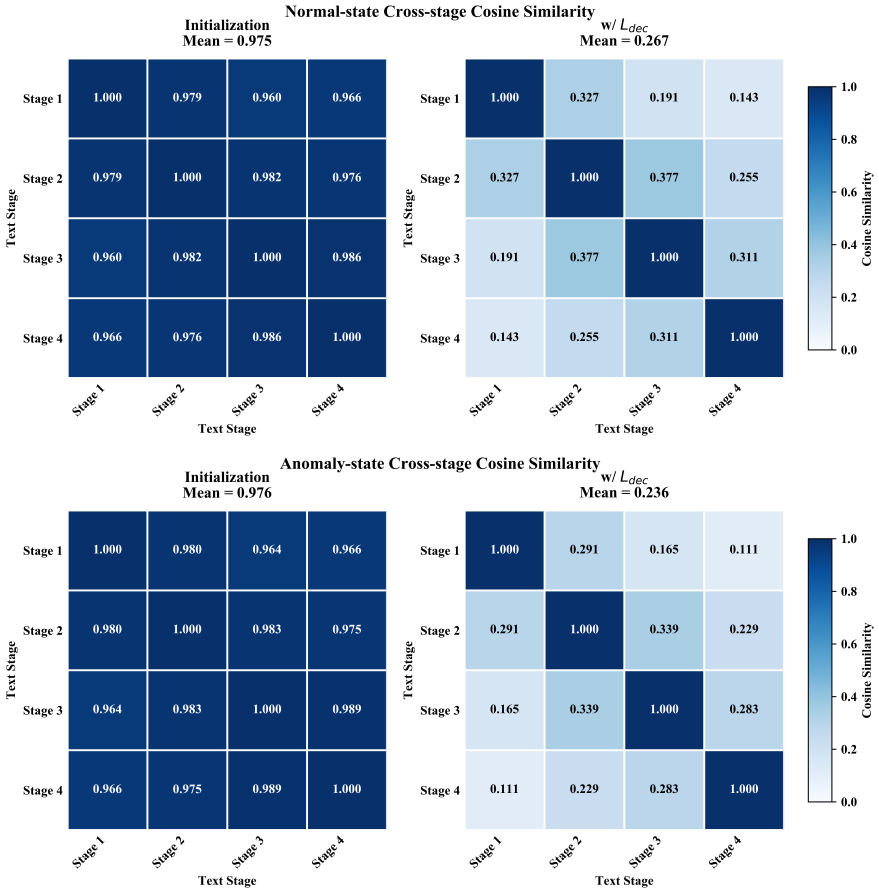


Figure 7. Cross-stage cosine similarity of textual representations at initialization and after training with L_{dec}. The upper and lower panels show normal-state and anomalous-state representations, respectively. The reported means are calculated over the six unique off-diagonal stage pairs. The model is trained on VisA for evaluation on MVTec AD.

Table 7. Sensitivity analysis of the temperature coefficient τ. "Img" and "Pix" denote image-level (AUROC, AP) and pixel-level (AUROC, PRO) metrics, respectively. The best results are shown in bold, and the second-best results are underlined.

Setting	MVTec AD		VisA		BTAD		MPDD		HeadCT	ISIC
	Img	Pix	Img	Pix	Img	Pix	Img	Pix	Img	Pix
$\tau = 0.01$	(87.9, 93.6)	(91.4, 84.4)	(80.5, 83.9)	(95.5, 88.7)	(87.7, 81.6)	(94.7, 77.0)	(**77.9**, 81.6)	(96.5, 90.3)	(90.3, 89.5)	(89.4, 79.4)
$\tau = 0.03$	(88.9, 94.4)	(91.4, 85.1)	(81.9, 84.9)	(95.5, 89.1)	(90.4, 88.0)	(94.8, 79.0)	(77.2, 81.6)	(96.7, 90.9)	(91.5, 90.1)	(89.8, 79.5)
$\tau = 0.05$	(**89.6**, **96.3**)	(**91.6**, **85.4**)	(**82.9**, **85.7**)	(**95.8**, **89.6**)	(**91.1**, **92.3**)	(95.1, **79.6**)	(77.3, **82.6**)	(**97.2**, **91.7**)	(**94.9**, **94.6**)	(**90.0**, **79.7**)
$\tau = 0.07$	(89.5, 94.9)	(90.4, 84.5)	(82.5, 85.4)	(95.3, 89.4)	(90.9, 92.0)	(94.9, 79.3)	(76.9, 82.5)	(96.5, 90.8)	(94.7, 94.2)	(89.5, 79.6)
$\tau = 0.1$	(89.3, 94.8)	(89.6, 83.0)	(81.4, 84.4)	(95.0, 88.2)	(89.8, 91.8)	(**95.3**, **79.6**)	(77.1, 81.6)	(95.9, 89.4)	(94.5, 94.5)	(89.5, 79.0)

4.4.4. Comparison of Aggregation Strategies and Defect-Area Analysis

To evaluate the effectiveness of AAG, we compare it with GAP, max pooling, top-k pooling, and fused-map scoring. All aggregation strategies are evaluated using the same trained AWAD checkpoints and test samples without additional training, ensuring a controlled comparison of inference-time aggregation and scoring strategies. As shown in Table 8, AAG achieves the highest I-AP on all three datasets and the highest I-AUROC on VisA and BTAD, while fused-map scoring performs better on MVTec AD in terms of I-AUROC. These results suggest that anomaly-guided aggregation can improve image-level discrimination compared with uniform pooling, although its advantages vary across datasets and metrics.

Table 8. Comparison of different aggregation strategies for image-level anomaly detection. The best results are shown in bold, and the second-best results are underlined.

Method	MVTec AD		VisA		BTAD	
	I-AUROC	I-AP	I-AUROC	I-AP	I-AUROC	I-AP
GAP	79.9	90.2	66.5	72.5	84.8	83.6
Max Pooling	86.4	93.5	82.2	85.4	88.4	82.4
Top-k (5%)	89.2	95.1	82.0	84.7	90.9	91.3
Fused Map	**91.2**	95.8	80.2	83.3	89.8	89.4
AAG (Ours)	89.6	**96.3**	**82.9**	**85.7**	**91.1**	**92.3**

To examine the effect of anomaly sparsity, we further evaluate different aggregation strategies across four defect-area groups on MVTec AD. As shown in Table 9, AAG outperforms GAP substantially for defects smaller than 1% (80.4% versus 39.6% in I-AUROC) and achieves the highest numerical performance for defect areas below 10%. However, top-k pooling performs better for larger defects. These findings support the motivation for anomaly-guided aggregation, particularly for sparse anomalies.

Table 9. Image-level detection performance on MVTec AD grouped by ground-truth defect-area ratio. Each entry reports I-AUROC/I-AP. The best results are shown in bold, and the second-best results are underlined.

Defect Area	Anomalous Images	GAP	Max	Top-k (5%)	Fused Map	AAG
<1%	499	39.6/42.8	80.2/81.7	67.7/66.1	61.9/60.1	**80.4**/**81.9**
1–5%	516	64.8/63.2	82.1/85.1	77.1/80.2	75.1/79.2	**82.9**/**85.8**
5–10%	121	75.8/46.0	83.8/65.5	85.0/67.8	84.9/67.8	**86.4**/**70.2**
≥10%	122	82.3/53.0	81.9/57.9	**89.2**/**71.7**	85.8/67.9	88.5/67.6

Note: Anomalous images from all MVTec AD categories are pooled within each defect-area group and evaluated against all normal test images. Unlike the category-averaged results in the main comparison tables, these values use cross-category pooled evaluation and cannot be directly aggregated to reproduce the overall benchmark scores.

4.5. Limitations and Discussion

Although the proposed AWAD framework achieves competitive performance, it occasionally produces false positives in highly complex industrial scenarios, as illustrated in Figure 8. In particular, structurally salient normal regions and repetitive background textures may trigger spurious anomaly responses. Future work will explore foreground-aware prompt learning and explicit background suppression to improve foreground-background disentanglement.

Despite these limitations, CLSD plays an important role in the hierarchical representation learning of AWAD. Unlike conventional zero-shot methods that rely on monolithic prompts, CLSD extracts stage-wise textual representations from a shared prompt bank, thereby decomposing the concepts of normal and anomalous into hierarchical semantic granularities. Furthermore, the margin-based decoupling loss discourages excessive same-state cross-stage similarity, promoting more differentiated stage-specific textual representations. However, reduced cosine similarity alone does not establish the semantic specialization of individual stages. Further stage-wise analysis is required to verify their specific roles in texture-level and structural anomaly detection.

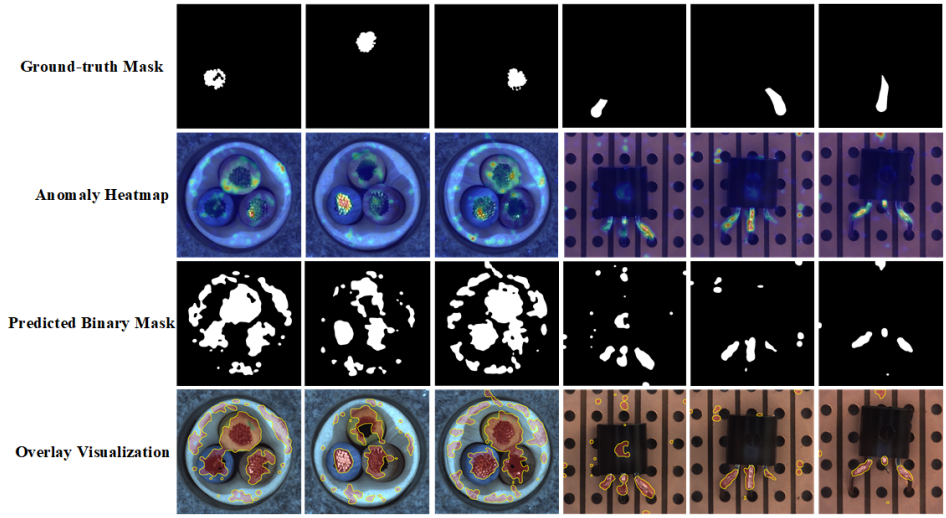


Figure 8. Typical failure cases of AWAD in complex industrial scenarios. The first three columns show over-response to normal structural regions in cable cross-section samples, while the last three columns illustrate interference from repetitive background patterns in electronic component samples. From top to bottom, each case presents the ground-truth mask, anomaly heatmap, predicted binary mask, and overlay visualization, respectively.

5. Conclusions

This study investigated cross-layer semantic decoupling and anomaly-guided spatial aggregation for CLIP-based zero-shot anomaly detection. The proposed CLSD module reduces excessive similarity among stage-specific textual representations, while AAG incorporates spatial anomaly responses into image-level feature aggregation. AWAD achieves a P-AUROC of 95.8% and a PRO of 89.6% on VisA, together with a P-AUROC of 91.6% on MVTec AD. Controlled ablation experiments further show that the decoupling loss improves PRO from 83.0% to 84.1% on MVTec AD, while introducing AAG increases I-AUROC from 85.3% to 91.1% on BTAD. These results support the effectiveness of cross-stage semantic regularization and anomaly-guided feature aggregation, although the improvements remain dataset dependent. False positives in structurally complex backgrounds remain a limitation, motivating further research on foreground-background disentanglement.

Author Contributions: F.L.: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Visualization, Writing—Original Draft, and Writing—Review and Editing. Y.W.: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Visualization, and Writing—Original Draft. X.T.: Conceptualization, Methodology, Formal Analysis, Validation, Supervision, and Writing—Review and Editing. M.W.: Data Curation, Investigation, and Validation. Q.Y.: Investigation and Validation. J.L.: Resources and Investigation. Z.Y.: Investigation and Validation. Y.Y.: Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Validation, and Writing—Review and Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by the National Natural Science Foundation of China (Grant No. 62073322) and the opening project of the State Key Laboratory of Autonomous Intelligent Unmanned Systems of China (No. ZZKF2025-3-1).

Informed Consent Statement: Not applicable.

Data Availability Statement: The public benchmark datasets used in this study, including MVTec AD, VisA, MPDD, BTAD, SDD, DTD-Synthetic, ISIC, CVC-ClinicDB, ENDO, HeadCT, BrainMRI, and Br35H, are available from their respective original sources, subject to the applicable licensing and usage conditions summarized in Table 2. The medical imaging datasets are used solely for cross-domain research benchmarking and do not constitute clinical validation. The Toothbrush

Bristle Defect Dataset (TBDD) constructed in this study will be released upon publication, subject to institutional approval.

Conflicts of Interest: Jianquan Li and Zhichao Yuan are employed by CASisense Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Object-Agnostic Prompt Bank Templates

The shared prompt bank used in AWAD consists of eight normal-state and eight anomalous-state textual templates. As shown in Table A1, these templates describe normality and abnormality from complementary visual perspectives, including appearance regularity, local consistency, and structural integrity. All templates are object-agnostic and do not contain target-specific category names, defect types, or medical terminology.

Table A1. The 16 object-agnostic textual templates used in the shared prompt bank.

No.	Normal-State Prompt	Anomalous-State Prompt
1	a normal visual pattern in the image	an abnormal visual pattern in the image
2	a regular appearance in the image	an irregular appearance in the image
3	a consistent local region in the image	an inconsistent local region in the image
4	an intact structure in the image	a disrupted structure in the image
5	a typical image region	an atypical image region
6	a normal image region	an anomalous image region
7	a visually consistent sample	a visually inconsistent sample
8	a normal sample	an abnormal sample

References

1. Lin, Y.; Chang, Y.; Tong, X.; Yu, J.; Liotta, A.; Huang, G.; Song, W.; Zeng, D.; Wu, Z.; Wang, Y.; et al. A survey on RGB, 3D, and multimodal approaches for unsupervised industrial image anomaly detection. *Inf. Fusion* **2025**, *121*, 103139. [CrossRef]
2. Tschuchnig, M.E.; Gadermayr, M. Anomaly detection in medical imaging—A mini review. In *Proceedings of the International Data Science Conference*; Springer: Wiesbaden, Germany, 2021; pp. 33–38.
3. Bergmann, P.; Batzner, K.; Fauser, M.; Sattlegger, D.; Steger, C. The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *Int. J. Comput. Vis.* **2021**, *129*, 1038–1059. [CrossRef]
4. Xu, J.; Lo, S.Y.; Safaei, B.; Patel, V.M.; Dwivedi, I. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 20370–20382.
5. 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 Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; Volume 139, pp. 8748–8763.
6. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.V.; Sung, Y.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; Volume 139, pp. 4904–4916.
7. Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; Kiela, D. FLAVA: A foundational language and vision alignment model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15638–15650.
8. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. [CrossRef]
9. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16816–16825.
10. Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L.H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. RegionCLIP: Region-based language-image pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16793–16803.
11. Zhou, Q.; Pang, G.; Tian, Y.; He, S.; Chen, J. AnomalyCLIP: Object-Agnostic Prompt Learning for Zero-Shot Anomaly Detection. In Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.

12. Chen, X.; Han, Y.; Zhang, J. A zero-/few-shot anomaly classification and segmentation method for CVPR 2023 VAND workshop challenge tracks 1 & 2. *arXiv* **2023**, arXiv:2305.17382.
13. Deng, C.; Xu, H.; Chen, X.; Xu, H.; Tu, X.; Ding, X.; Huang, Y. SimCLIP: Refining image-text alignment with simple prompts for zero-/few-shot anomaly detection. In Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM), Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 1761–1770.
14. Deng, H.; Zhang, Z.; Bao, J.; Li, X. AnoVL: Adapting vision-language models for unified zero-shot anomaly localization. *arXiv* **2023**, arXiv:2308.15939.
15. Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravikumar, A.; Dabeer, O. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 19606–19616.
16. Liu, W.; Wang, C.; Zhang, Y. Industrial surface defect detection by multi-scale Inpainting-GAN. *Vis. Comput.* **2025**, *41*, 5643–5660. [CrossRef]
17. Zavrtanik, V.; Kristan, M.; Skočaj, D. DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 8330–8339.
18. Li, C.L.; Sohn, K.; Yoon, J.; Pfister, T. CutPaste: Self-supervised learning for anomaly detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9664–9674.
19. Liu, Z.; Zhou, Y.; Xu, Y.; Wang, Z. SimpleNet: A simple network for image anomaly detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 20402–20411.
20. Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; Gehler, P. Towards total recall in industrial anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14318–14328.
21. Defard, T.; Setkov, A.; Loesch, A.; Audigier, R. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *ICPR International Workshops and Challenges*; Springer: Cham, Switzerland, 2021; pp. 475–489.
22. Cohen, N.; Hoshen, Y. Sub-image anomaly detection with deep pyramid correspondences. *arXiv* **2020**, arXiv:2005.02357.
23. Meng, F.; Gong, T.; Wu, D.; Li, Y. Unsupervised surface defect detection using dictionary-based sparse representation. *Eng. Appl. Artif. Intell.* **2025**, *143*, 110020. [CrossRef]
24. Liu, Y.; Zhang, C.; Dong, X.; Wang, H. A multimodal industrial anomaly detection method based on mask training and teacher–student joint memory. *Eng. Appl. Artif. Intell.* **2025**, *161*, 112299. [CrossRef]
25. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4183–4192.
26. Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; Wandt, B. Asymmetric student-teacher networks for industrial anomaly detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–7 January 2023; pp. 2592–2602.
27. Deng, H.; Li, X. Anomaly detection via reverse distillation from one-class embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 9737–9746.
28. Wang, X.; Fan, J.; Yan, F.; Liu, Y. Unsupervised fur anomaly detection with B-spline noise-guided Multi-directional Feature Aggregation. *Vis. Comput.* **2025**, *41*, 6169–6185. [CrossRef]
29. Khattak, M.U.; Rasheed, H.; Maaz, M.; Khan, S.; Khan, F.S. MaPLe: Multi-modal prompt learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 19113–19122.
30. Jia, M.; Tang, L.; Chen, B.; Cardie, C.; Belongie, S.; Hariharan, B.; Lim, S.-N. Visual prompt tuning. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 709–727.
31. Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H. CLIP-Adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.* **2024**, *132*, 581–595. [CrossRef]
32. Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Hwang, K.; Li, H. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 493–510.
33. Yang, Y.; Zhang, H.; Cao, Z.; Liu, X. CORE-CLIP: Smart collaborative reasoning driven by CLIP for human-object interaction detection. *Pattern Recognit.* **2025**, *171*, 112242.. [CrossRef]
34. Qu, Z.; Tao, X.; Gong, X.; Zhang, Y. Bayesian prompt flow learning for zero-shot anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 30398–30408.

35. Gu, Z.; Zhu, B.; Zhu, G.; Li, Y. FiLo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM), Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 2041–2049.
36. Gu, Z.; Zhu, B.; Zhu, G.; Li, Y. AnomalyGPT: Detecting industrial anomalies using large vision-language models. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 1932–1940. [CrossRef]
37. Cao, Y.; Zhang, J.; Frittoli, L.; Boracchi, G. AdaCLIP: Adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September 2024; pp. 55–72.
38. Fang, Q.; Lv, W.; Su, Q. AF-CLIP: Zero-shot anomaly detection via anomaly-focused CLIP adaptation. In Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM), Dublin, Ireland, 27–31 October 2025; pp. 4846–4855.
39. Li, Y.; Goodge, A.; Liu, F.; Li, C. PromptAD: Zero-shot anomaly detection using text prompts. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024; pp. 1093–1102.
40. Chen, X.; Zhang, J.; Tian, G.; Wang, Y. CLIP-AD: A language-guided staged dual-path model for zero-shot anomaly detection. In *Human Activity Recognition and Anomaly Detection*; Springer: Singapore, 2024; pp. 17–33.
41. Zhang, G.; Zhang, L. DLVP-CLIP: Enhancing Fine-Grained Zero-Shot Anomaly Detection via Dynamic Local Visual Prompting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA, 3–7 June 2026; pp. 35524–35533.
42. Hu, M.; Huo, Y.; Dou, M.; Yin, J.; Zhao, P.; Wang, Y.; Hu, C.; Hu, B.; Wang, Q. FB-CLIP: Fine-Grained Zero-Shot Anomaly Detection with Foreground-Background Disentanglement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA, 3–7 June 2026; pp. 35659–35669.
43. Cao, Y.; Xu, X.; Sun, C.; Wang, Y. Segment any anomaly without training via hybrid prompt regularization. *arXiv* **2023**, arXiv:2305.10724.
44. Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; Dabeer, O. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 392–408.
45. 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 *Proceedings of the 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*; IEEE: New York, NY, USA, 2021; pp. 66–71.
46. Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; Foresti, G.L. VT-ADL: A vision transformer network for image anomaly detection and localization. *arXiv* **2021**, arXiv:2104.10036.
47. Tabernik, D.; Šela, S.; Skvarč, J.; Skočaj, D. Segmentation-based deep-learning approach for surface-defect detection. *J. Intell. Manuf.* **2020**, *31*, 759–776. [CrossRef]
48. Aota, T.; Tong, L.T.T.; Okatani, T. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–7 January 2023; pp. 5564–5572.
49. Gutman, D.; Codella, N.C.F.; 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* **2016**, arXiv:1605.01397.
50. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; Vilariño, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **2015**, *43*, 99–111. [CrossRef] [PubMed]
51. Hicks, S.A.; Jha, D.; Thambawita, V.; Halvorsen, P.; Riegler, M.A. The EndoTect 2020 challenge: Evaluation and comparison of classification, segmentation and inference time for endoscopy. In *Pattern Recognition. ICPR International Workshops and Challenges*; Springer: Cham, Switzerland, 2021; pp. 263–274.
52. Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M.H.; Rabiee, H.R. Multiresolution knowledge distillation for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 14902–14912.
53. Hamada, A. Br35h: Brain Tumor Detection 2020, 2020. Available online: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection (accessed on 20 October 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.