Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084208

ARTICLE

LaRP-CLIP: Layer-Aware Refinement with Prototype Guidance for Zero-Shot Anomaly Detection

Xing Fang[1], **Yuanfang Chen**[1,2,*], **Qiang Lin**[3], **Kun Yang**[2,4] **and Gyu Myoung Lee**[5]

[1]School of Cyberspace, Hangzhou Dianzi University, Hangzhou, China
[2]The State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China
[3]School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
[4]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[5]School of Computer Science and Mathematics, Liverpool John Moores University, Liverpool, UK

*Corresponding Author: Yuanfang Chen. Email: yuanfang.chen.tina@gmail.com

Received: 18 April 2026; Accepted: 22 May 2026; Published: 23 July 2026

ABSTRACT: The deployment of supervised anomaly detection is typically limited by the high cost of annotation, privacy constraints, and the scarcity of anomalous samples. These constraints have motivated the use of vision-language pre-trained models for zero-shot anomaly detection. However, existing CLIP-based methods still face three limitations: a shared set of prompts is applied across feature layers, anomaly maps are fused by fixed strategies, and image-level anomaly scores are determined solely by global image-text similarity. These limitations reduce the accuracy of pixel-level localization and weaken the reliability of image-level anomaly prediction. To overcome these limitations, LaRP-CLIP is proposed. It introduces layer-aware prompt decoupling to better match feature layers with different semantic characteristics, adaptive fusion with error-prior-guided local refinement to produce cleaner and more precise anomaly maps, and a prototype branch to improve image-level scoring. Experiments on four industrial datasets and seven medical datasets show that LaRP-CLIP achieves strong performance in both image-level detection and pixel-level localization.

KEYWORDS: Zero-shot anomaly detection; vision-language models; layer-aware prompts; local refinement; prototype branch

1 Introduction

Anomaly detection aims to identify samples that deviate from normal patterns and plays an important role in industrial quality inspection, medical image analysis, and intelligent manufacturing [1–5]. Traditional supervised anomaly detection methods usually require a large number of normal and anomalous samples for each new product category during training [6–8]. In real industrial scenarios, however, data collection is costly, privacy constraints are common, and anomalous samples are often extremely scarce [9,10]. Therefore, this paper focuses on zero-shot anomaly detection, where the goal is to detect anomalies in unseen target domains without requiring target-domain training samples [11–15].

To reduce dependence on target-domain annotation, zero-shot anomaly detection (ZSAD) has attracted increasing attention in recent years [16,17]. This paradigm relies on pre-trained vision-language models, such as Contrastive Language-Image Pre-training (CLIP) [18], and performs anomaly detection without requiring anomalous samples from the target domain. Among existing ZSAD approaches, CLIP-based methods have

drawn particular interest because of their strong cross-modal representation capability [19–24]. As shown in Fig. 1a, these methods usually extract multi-layer features from a ViT backbone, compute similarity scores with a shared set of text prompts to obtain layer-wise anomaly maps, fuse these maps for pixel-level prediction, and use global image features for image-level anomaly scoring.

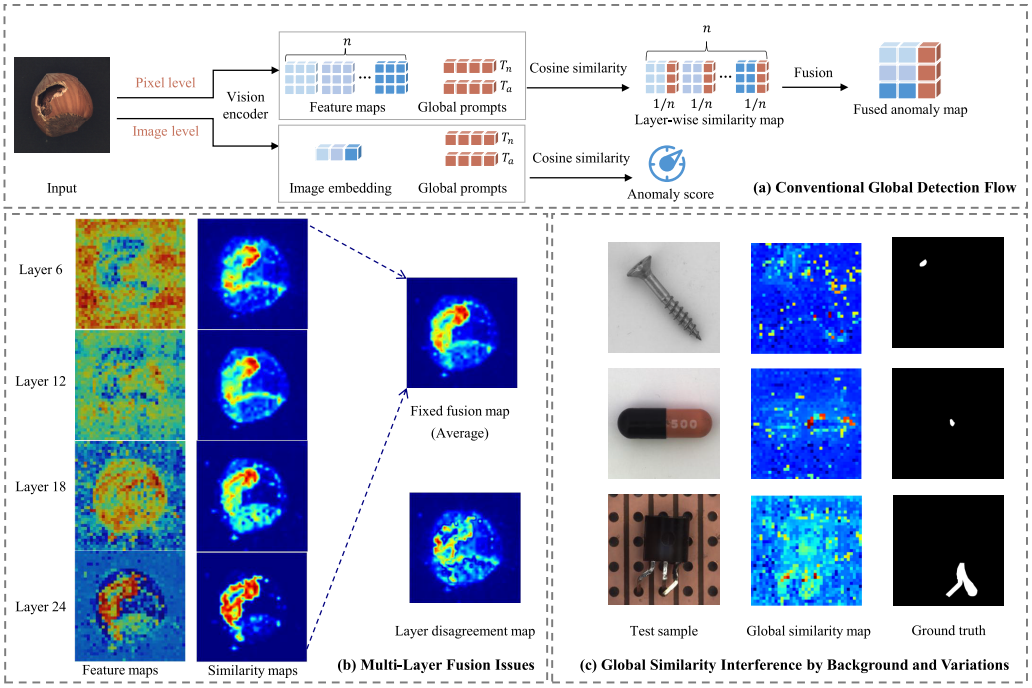

Figure 1: Typical pipeline and three major limitations of existing CLIP-based zero-shot anomaly detection methods. (**a**) Overall workflow. (**b**) Pixel-level limitations: all layers share the same text prompts and use fixed fusion. (**c**) Image-level limitation: reliance on global image-text similarity.

Despite recent progress, existing CLIP-based ZSAD methods still exhibit three major limitations. At the pixel level, all feature layers typically share the same set of text prompts, as illustrated in Fig. 1b. This design is suboptimal because shallow layers mainly capture texture details, whereas deep layers encode higher-level structure and semantics. Using the same prompts across all layers therefore introduces a mismatch between prompt semantics and feature characteristics, which degrades the quality of the resulting anomaly maps. Moreover, layer-wise anomaly maps are usually combined by a fixed fusion strategy that cannot adapt to the content of each input image. Responses from different layers may also conflict on the same defect, resulting in blurred boundaries and increased noise in the fused maps.

At the image level, existing methods usually rely only on global image–text similarity, as shown in Fig. 1c. When the input image contains a complex background or normal samples exhibit slight appearance variations, the global similarity score can be easily affected by irrelevant regions or benign changes. This weakens the robustness of image-level anomaly scoring.

To address these limitations, LaRP-CLIP is proposed. The main contributions of this paper are summarized as follows:

- Layer-Aware Prompt Decoupling: The learnable prompts are divided into three independent groups, namely Global Prompts, Shallow Prompts, and Deep Prompts. This design allows different feature layers to interact with prompts that better match their semantic characteristics, thereby reducing the mismatch between shallow-layer texture information and deep-layer semantic representations.

- Adaptive Weighted Fusion and Local Refinement: Learnable layer-wise weights are introduced to adaptively fuse the anomaly maps from multiple layers. In addition, an Error-Prior Guided Local Relation Refiner is designed to perform local correction by modeling inter-layer response discrepancies, producing anomaly maps with clearer boundaries and less noise.
- Prototype Branch: During inference, high-confidence normal patch features are selected from the current test image to construct an image-specific normal prototype. This prototype is then combined with the global similarity score, which helps reduce the influence of complex backgrounds and normal appearance variations and improves the robustness of image-level anomaly scoring.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed method. Section 4 reports the experimental results and analysis. Section 5 concludes the paper.

2 Related Work

2.1 Traditional Anomaly Detection Methods

Traditional anomaly detection methods are generally divided into supervised and unsupervised paradigms. Supervised methods rely on a sufficient number of normal and anomalous samples for each category during training [3,25]. In industrial practice, however, anomalous samples are often scarce, annotation is expensive, and data sharing may be restricted by privacy or production constraints. These factors greatly hinder the rapid deployment of anomaly detection models on new product lines.

Unsupervised methods are usually trained with normal samples only and identify anomalies through reconstruction error [26], autoencoding mechanisms [27], or one-class modeling [28]. Although such methods avoid the need for anomalous training data, they are still tied to the distribution of the training domain. Once the product type, surface texture, or imaging condition changes, their performance often degrades noticeably. This weak cross-domain generalization makes them less suitable for settings that require rapid deployment on unseen categories.

2.2 Zero-Shot Anomaly Detection

To reduce dependence on target-domain annotation, zero-shot anomaly detection (ZSAD) has received increasing attention in recent years [29,30]. The goal of ZSAD is to detect anomalies in a target domain without using anomalous samples from that domain during training [16]. Earlier studies mainly relied on reconstruction-based schemes or distribution modeling. In complex industrial scenes, however, reconstruction quality is easily affected by normal appearance variation, which weakens the distinction between normal samples and true anomalies [17,18].

With the development of large-scale foundation models, recent ZSAD research has shifted toward transfer-based solutions. Instead of learning anomaly-specific representations from target-domain data, these methods attempt to transfer general visual or cross-modal knowledge acquired during large-scale pre-training to downstream anomaly detection tasks. This shift has made vision-language models an important direction for zero-shot anomaly detection.

ZSAD is related to zero-shot learning (ZSL), where semantic knowledge is transferred to recognize unseen classes. Recent ZSL studies, such as SVIP [31], further improve visual-semantic alignment by modeling semantically contextualized visual patches. However, unlike ZSL that focuses on unseen category recognition, ZSAD aims to detect category-agnostic abnormal patterns in unseen domains. LaRP-CLIP adapts this semantic transfer principle to anomaly detection through normal/abnormal cross-modal alignment, layer-aware prompt decoupling, and image-specific prototype construction.

2.3 CLIP-Based Zero-Shot Anomaly Detection Methods

The emergence of CLIP [18] has accelerated the development of CLIP-based ZSAD methods because of its strong image–text alignment capability. WinCLIP [19] is an early representative method that performs zero-shot anomaly classification and segmentation through a window-based inference scheme and prompt-template ensemble. AnomalyCLIP [20] further introduces object-agnostic learnable prompts and optimizes them with both image-level and pixel-level supervision, showing that CLIP can be adapted more effectively to zero-shot anomaly detection.

Subsequent studies have mainly focused on improving prompt design, feature utilization, and anomaly discrimination. AA-CLIP [24] strengthens anomaly awareness in both the textual and visual spaces, thereby improving the discrimination between normal and abnormal patterns. GenCLIP [21] further explores prompt generalization and multi-layer feature usage to improve transferability across datasets. Despite this progress, most existing methods still rely on a single prompt form across different layers and use fixed fusion strategies, which limits their ability to handle layer-wise semantic differences and image-specific normal variations.

Table 1 summarizes representative CLIP-based ZSAD methods and highlights the differences from LaRP-CLIP.

Table 1: Comparison of representative CLIP-based zero-shot anomaly detection methods. ✗ indicates not met; ✓ indicates met.

Method	Venue	Prompt Design	Layer-Specific Prompting	Adaptive Fusion	Local Refinement	Prototype Guidance
AnomalyCLIP [20]	ICLR'24	Learnable object-agnostic	✗	✗	✗	✗
WinCLIP [19]	CVPR'23	Hand-crafted templates	✗	✗	✗	✗
AA-CLIP [24]	CVPR'25	Anomaly-aware prompts	✗	✗	✗	✗
GenCLIP [21]	PR'26	Generalized prompts	✗	✗	✗	✗
LaRP-CLIP	–	Layer-aware prompts	✓	✓	✓	✓

Decoupled prompts and local refinement have been explored in prompt learning and segmentation literature but are not claimed as isolated contributions. The novelty of LaRP-CLIP lies in adapting these ideas for CLIP-based ZSAD: prompt decoupling is driven by the semantic hierarchy of ViT layers, local refinement is guided by inter-layer anomaly-response disagreement, and image-level scoring is enhanced by an inference-time normal prototype. This approach directly addresses cross-layer semantic mismatch, fixed map fusion, and unreliable global-only scoring in existing CLIP-based ZSAD methods.

3 Method

3.1 Overall Architecture

LaRP-CLIP is designed for both pixel-level anomaly localization and image-level anomaly classification in the zero-shot setting. As shown in Fig. 2, given an input image $x \in \mathbb{R}^{3\times H\times W}$, the visual encoder extracts a global image feature together with multi-layer local patch features. Meanwhile, the prompt learner generates text prompts for different semantic levels, which are then projected into a unified semantic space by the text encoder. The method contains two complementary branches: a pixel-level branch for anomaly localization and an image-level branch for robust anomaly scoring.

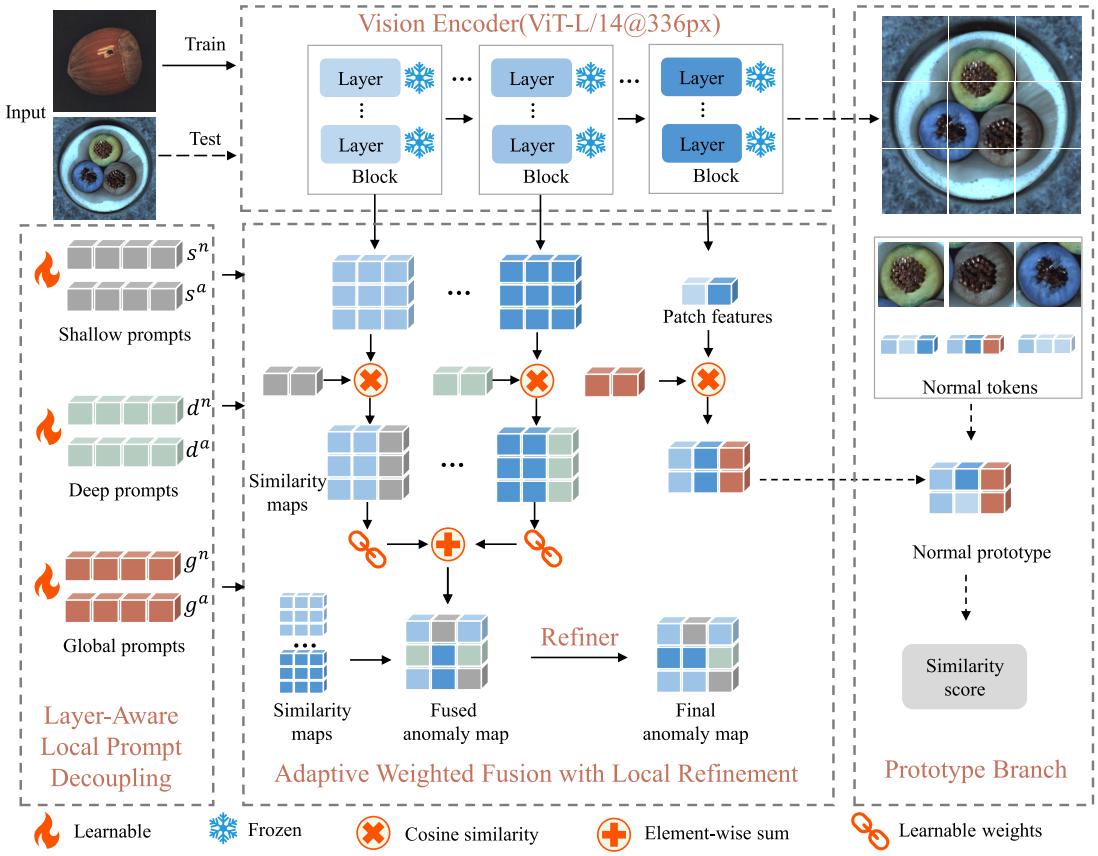

Figure 2: LaRP-CLIP consists of three main components: layer-aware prompt decoupling (**left**), adaptive weighted fusion with local refinement (**middle**), and the prototype branch for robust image-level scoring (**right**).

Let the global feature extracted by the visual encoder be denoted by $f_g \in \mathbb{R}^D$, and let the selected multi-layer patch features be

$$F = \{F^1, F^2, \ldots, F^L\}, \quad F^l \in \mathbb{R}^{N_l \times D}, \tag{1}$$

where N_l is the number of patches at the l-th layer and D is the feature dimension.

The pixel-level branch is built upon three steps: layer-aware semantic alignment, adaptive cross-layer fusion, and local refinement. To better match the heterogeneous semantics of different ViT layers, the local prompts are decoupled into shallow prompts and deep prompts, allowing patch features from different layers to be aligned with anomaly semantics at suitable granularities. Based on the resulting layer-wise anomaly responses, learnable layer weights are introduced to perform adaptive fusion across layers. A lightweight Error-Prior Guided Local Relation Refiner is then applied to the fused anomaly map to suppress noisy responses and improve local boundary quality.

The image-level branch is introduced to improve the robustness of anomaly scoring. During inference, a Prototype Branch constructs an image-specific normal prototype from high-confidence normal patches selected from the current test image. This prototype serves as an intra-image normal reference and is combined with the global similarity score, so that image-level anomaly prediction becomes less sensitive to background clutter and benign appearance variation.

3.2 Layer-Aware Local Prompt Learning

Local anomaly representations exhibit clear hierarchical differences across ViT layers. Shallow features mainly encode edges, textures, and other high-frequency local patterns, and are therefore more suitable for describing fine-grained anomalies such as scratches, cracks, and stains. By contrast, deep features place greater emphasis on structure, shape, and semantic consistency, making them more suitable for structural anomalies such as missing parts, misalignment, and deformation. If all layers share a single set of local prompts, features from different semantic levels are forced into the same text space, which leads to a clear semantic granularity mismatch.

To alleviate this issue, LaRP-CLIP divides the learnable prompts into three groups, namely global prompts, shallow prompts, and deep prompts. After the text encoder, the corresponding text embeddings are written as

$$[(\hat{g}^n, \hat{g}^a), (\hat{s}^n, \hat{s}^a), (\hat{d}^n, \hat{d}^a)] \in \mathbb{R}^{2\times D}, \tag{2}$$

where the superscripts n and a denote the normal and abnormal classes, respectively.

The layers used for local anomaly modeling are divided into a shallow set S and a deep set D, satisfying

$$S \cup D = \{1, 2, \ldots, L\}, \quad S \cap D = \varnothing. \tag{3}$$

For the patch features F^l at the l-th layer, each patch feature is first normalized by

$$\hat{f}_i^l = \frac{f_i^l}{\|f_i^l\|_2}, \tag{4}$$

where f_i^l denotes the i-th patch feature at layer l. The text embeddings are normalized in the same way. If $l \in S$, the shallow prompt embeddings $(\hat{s}^n, \hat{s}^a)$ are used; if $l \in D$, the deep prompt embeddings $(\hat{d}^n, \hat{d}^a)$ are used.

The similarity between the i-th patch and the corresponding normal or abnormal text embedding is defined as

$$z_{i,c}^l = \tau\,(\hat{f}_i^l)^\top \hat{e}_c^l, \quad c \in \{n, a\}, \tag{5}$$

where τ is the temperature coefficient and $\hat{e}_c^l$ denotes the text embedding assigned to layer l. The anomaly response of the i-th patch is then obtained by binary softmax:

$$p_{i,a}^l = \frac{\exp(z_{i,a}^l)}{\exp(z_{i,n}^l) + \exp(z_{i,a}^l)}. \tag{6}$$

The anomaly map of the l-th layer is written as

$$M^l = \mathrm{Reshape}\left(\{p_{i,a}^l\}_{i=1}^{N_l}\right). \tag{7}$$

For image-level scoring, the global feature interacts only with the text embeddings corresponding to the global prompts. The resulting text-driven anomaly score is defined as

$$s_{\text{text}} = \frac{\exp(\tau_g \hat{f}_g^\top \hat{g}^a)}{\exp(\tau_g \hat{f}_g^\top \hat{g}^n) + \exp(\tau_g \hat{f}_g^\top \hat{g}^a)}. \tag{8}$$

With this design, shallow patch features are aligned with shallow anomaly semantics, deep patch features are aligned with deep anomaly semantics, and the global feature is used for image-level anomaly judgment. In this way, the cross-modal alignment process becomes layer-aware, which helps reduce the semantic mismatch introduced by using a shared prompt across all feature layers.

3.3 Layer-Adaptive Fusion and Local Relation Refinement

The layer-wise anomaly maps $\{M^l\}_{l=1}^{L}$ differ in detail sensitivity, structural stability, and noise level. Fixed fusion or equal-weight averaging cannot fully exploit the complementary information carried by different layers. To address this issue, LaRP-CLIP introduces a set of learnable fusion parameters $\{\alpha^l\}_{l=1}^{L}$, which are normalized by softmax to obtain the layer weights:

$$\omega_l = \frac{\exp(\alpha_l)}{\sum_{k=1}^{L} \exp(\alpha_k)}, \quad l = 1, \ldots, L. \tag{9}$$

The fused anomaly map is then computed as

$$M_{\text{fused}} = \sum_{l=1}^{L} \omega_l M^l. \tag{10}$$

This design allows the contribution of each layer to be learned from data, rather than being fixed in advance.

Although adaptive fusion improves the integration of multi-layer responses, the fused map may still contain local noise or imprecise boundaries. This issue becomes more evident when different layers produce inconsistent responses in the same region. To capture such disagreement, an error prior is introduced to measure the deviation of each layer from the fused result. For the l-th layer, the deviation map is defined as

$$E^l = |M^l - M_{\text{fused}}|. \tag{11}$$

The aggregated error prior is then written as

$$E_{\text{prior}} = \sum_{l=1}^{L} \omega_l E^l. \tag{12}$$

The error prior provides a local measure of cross-layer inconsistency. Regions with similar responses across layers tend to have small values, whereas regions with conflicting responses yield larger values. Based on this cue, LaRP-CLIP concatenates the fused anomaly map with the error prior and feeds them into a lightweight local relation refinement module. This module adopts a 3×3 convolutional structure to predict a residual correction:

$$\Delta M = \psi([M_{\text{fused}}, E_{\text{prior}}]). \tag{13}$$

The final pixel-level anomaly map is given by

$$M_{\text{final}} = M_{\text{fused}} + \Delta M. \tag{14}$$

The refinement module is used to correct local errors in the fused anomaly map rather than to regenerate the entire prediction. It mainly adjusts regions with blurred boundaries, discontinuous responses, or noticeable local noise, while preserving the main structure provided by the preceding layer-wise fusion stage.

3.4 Prototype Branch

Relying only on global text matching is often insufficient for stable image-level anomaly scoring, because it does not fully reflect the local distribution of anomaly evidence within an image. This limitation becomes more apparent when anomalous regions are small, the background is complex, or normal samples exhibit noticeable appearance variation. To improve the robustness of image-level scoring, a Prototype Branch is introduced during inference. This branch constructs a normal prototype from high-confidence normal regions in the current test image and uses it as an additional reference for anomaly judgment.

The layer-wise anomaly maps are first accumulated to obtain a text-driven anomaly map:

$$M_{\text{text}} = \sum_{l=1}^{L} M^{l}. \tag{15}$$

The patch features from the last layer, denoted by $\{p_i\}_{i=1}^{N}$, are then used as local semantic representations. The map M_{text} is downsampled to the patch grid to produce an anomaly score q_i for each patch. Based on these scores, a set of high-confidence normal patches is selected, and its index set is denoted by $\mathcal{N}$. The normal prototype is computed as

$$p_{\text{norm}} = \frac{\sum_{i\in\mathcal{N}} \omega_i p_i}{\sum_{i\in\mathcal{N}} \omega_i}, \tag{16}$$

where ω_i is derived from q_i, so that patches with lower anomaly scores receive larger weights.

The deviation of each patch from the normal prototype is then measured by

$$\mu_i = 1 - \cos(p_i, p_{\text{norm}}). \tag{17}$$

These deviations are reshaped into a two-dimensional map to form the prototype-guided anomaly map, which is further combined with the text-driven anomaly map to obtain an auxiliary anomaly map for image-level scoring. Since image-level prediction is more influenced by the most salient anomalous regions than by the average response over the whole image, top-k pooling is applied to the fused map to obtain the auxiliary score s_{map}. The final image-level anomaly score is defined as

$$s_{\text{final}} = \frac{s_{\text{text}} + \lambda s_{\text{map}}}{1 + \lambda}, \tag{18}$$

where λ is a balancing coefficient.

Unlike a fixed normal template shared across all images, the Prototype Branch derives a normal reference directly from the current test image. This makes the image-level score less sensitive to background clutter and benign appearance variation. The branch is used only for image-level scoring and does not affect the generation of pixel-level anomaly maps.

The Prototype Branch is excluded from training because it is designed as a non-parametric test-time adaptation module. It constructs an image-specific normal prototype from high-confidence normal patches in each test image, rather than learning a fixed prototype from auxiliary training data.

3.5 Training Objective

To jointly optimize image-level anomaly discrimination and pixel-level anomaly localization, LaRP-CLIP uses a combined training objective. The overall loss is defined as

$$\mathcal{L} = \mathcal{L}_{\text{img}} + \mathcal{L}_{\text{pixel}}, \tag{19}$$

where $\mathcal{L}_{\text{img}}$ is the image-level loss and $\mathcal{L}_{\text{pixel}}$ is the pixel-level loss.

For image-level supervision, let $y \in \{0,1\}$ denote the image label and let s_{text} denote the global text-driven anomaly score. The image-level loss is defined by binary cross-entropy:

$$\mathcal{L}_{\text{img}} = -y \log s_{\text{text}} - (1-y)\log(1-s_{\text{text}}). \tag{20}$$

For pixel-level supervision, the intermediate outputs are not treated as independent prediction targets. Instead, the loss is divided according to the two stages of the pixel-level branch, namely alignment and refinement:

$$\mathcal{L}_{\text{pixel}} = \mathcal{L}_{\text{align}} + \lambda_r \mathcal{L}_{\text{ref}}, \tag{21}$$

where $\mathcal{L}_{\text{align}}$ supervises the front-end layer-aware semantic alignment and adaptive fusion, $\mathcal{L}_{\text{ref}}$ supervises the refinement stage, and λ_r is the weight assigned to the refinement term.

The alignment loss is applied to the shallow-layer anomaly maps, the deep-layer anomaly maps, and the fused anomaly map:

$$\begin{aligned}\mathcal{L}_{\text{align}} = {} & \frac{1}{|S|}\sum_{l\in S}\mathcal{L}_{\text{seg}}(M^l, Y) \\ & + \frac{1}{|D|}\sum_{l\in D}\mathcal{L}_{\text{seg}}(M^l, Y) \\ & + \lambda_f \mathcal{L}_{\text{seg}}(M_{\text{fused}}, Y),\end{aligned} \tag{22}$$

where Y is the pixel-level ground-truth mask, M^l is the anomaly map at layer l, M_{fused} is the adaptively fused anomaly map, and S and D denote the shallow-layer set and deep-layer set, respectively. The function $\mathcal{L}_{\text{seg}}(\cdot)$ denotes the segmentation loss, implemented as a combination of focal loss and dual-channel Dice loss for the normal and abnormal classes. The coefficient λ_f controls the relative contribution of the fused-map supervision.

This design serves two purposes. First, separate supervision on shallow and deep anomaly maps encourages the decoupled prompts to learn anomaly semantics at different granularities. Second, direct supervision on M_{fused} ensures that the fusion stage itself produces a discriminative anomaly map before refinement.

The refinement loss is applied only to the final anomaly map:

$$\mathcal{L}_{\text{ref}} = \mathcal{L}_{\text{seg}}(M_{\text{final}}, Y) + \lambda_\Delta \| M_{\text{final}} - M_{\text{fused}} \|_1, \tag{23}$$

where λ_Δ is the weight of the residual regularization term. This term limits the difference between the refined anomaly map and the fused anomaly map, thereby constraining the magnitude of the correction introduced by the refinement module.

The Prototype Branch is used only during inference and is therefore not included in the training objective.

3.6 Training and Inference Procedure

To improve the readability and reproducibility of the proposed method, the pseudo code of LaRP-CLIP is provided in Algorithm 1. The algorithm summarizes the complete training and inference procedure, including layer-aware prompt matching, adaptive fusion with local refinement, and the Prototype Branch for image-level scoring. During training, the loss in Eq. (19) is used to optimize the learnable parameters. During inference, all learnable parameters are fixed, and only the forward pass is performed to obtain the final anomaly scores.

Algorithm 1: The pseudo code of LaRP-CLIP for zero-shot anomaly detection

Require: Input image $x \in \mathbb{R}^{3\times H\times W}$, pre-trained visual and text encoders, learnable prompts $\{\alpha^l, g^n, g^a, s^n, s^a, d^n, d^a\}$, refinement module ψ, balancing coefficient λ

Ensure: Pixel-level anomaly map M_{final}, image-level anomaly score s_{final}

Extract global feature f_g and multi-layer patch features $F = \{F^l\}_{l=1}^{L}$ {Eq. (1)}

Generate and normalize global, shallow, and deep text embeddings: $(\hat{g}^n, \hat{g}^a), (\hat{s}^n, \hat{s}^a), (\hat{d}^n, \hat{d}^a)$ {Eq. (2)}

Partition the selected layers into shallow set S and deep set D {Eq. (3)}

for each layer $l = 1, \ldots, L$ **do**

  Normalize patch features: $\hat{f}_i^l = f_i^l / \|f_i^l\|_2$ {Eq. (4)}

  Select the corresponding prompt embedding according to $l \in S$ or $l \in D$

  Compute similarity scores: $z_{i,c}^l = \tau(\hat{f}_i^l)^\top \hat{e}_c^l$ {Eq. (5)}

  Compute the anomaly response $p_{i,a}^l$ and form the anomaly map M^l {Eqs. (6) and (7)}

end for

Compute normalized layer weights ω_l by softmax {Eq. (9)}

Fuse the layer-wise anomaly maps: $M_{\text{fused}} = \sum_{l=1}^{L} \omega_l M^l$ {Eq. (10)}

Compute deviation maps and error prior: $E^l = |M^l - M_{\text{fused}}|$, $E_{\text{prior}} = \sum_{l=1}^{L} \omega_l E^l$ {Eqs. (11) and (12)}

Predict residual correction and obtain the final anomaly map: $\Delta M = \psi([M_{\text{fused}}, E_{\text{prior}}])$, $M_{\text{final}} = M_{\text{fused}} + \Delta M$ {Eqs. (13) and (14)}

Compute the global text-driven anomaly score s_{text} {Eq. (8)}

Accumulate the layer-wise anomaly maps: $M_{\text{text}} = \sum_{l=1}^{L} M^l$ {Eq. (15)}

Select high-confidence normal patches and compute the normal prototype p_{norm} {Eq. (16)}

Compute the prototype-guided auxiliary score s_{map} by top-k pooling

Compute the final image-level score: $s_{\text{final}} = (1-\lambda)s_{\text{text}} + \lambda s_{\text{map}}$ {Eq. (18)}

if training **then**

  Compute the total loss $\mathcal{L} = \mathcal{L}_{\text{img}} + \mathcal{L}_{\text{pixel}}$ {Eqs. (19)–(23)}

end if

return $M_{\text{final}}, s_{\text{final}}$

4 Experiments

4.1 Experimental Settings

The experiments are evaluated on the MVTec AD [32], VisA [23], MPDD [33], and BTAD [34] industrial defect datasets, as well as the HeadCT [35], BrainMRI [35], Br35H [36], ISIC [37], CVC-ColonDB [38], CVC-ClinicDB [38], and TN3K [39] medical imaging datasets. For evaluation metrics, pixel-level performance is measured by pixel-level AUROC and PRO, while image-level performance is measured by image-level AUROC and AP.

All experiments were conducted on a single NVIDIA RTX 4090 GPU (24 GB VRAM) using Python 3.12.7 and the PyTorch 2.0.0 framework. The Adam optimizer was employed with an initial learning rate of

1×10^{-3}, a batch size of 8, and a total of 15 training epochs. To ensure reproducibility, the random seed was fixed at 111. The visual backbone network adopts the pre-trained CLIP model (ViT-L/14@336px), with the input resolution uniformly set to 518×518. During training, only the parameters of the learnable prompt learner are optimized, while the CLIP visual encoder remains frozen.

The prompt learner uses a learnable prompt length of 12, a text embedding depth of 9, and feature layers selected as $[6, 12, 18, 24]$. The temperature coefficient is $\tau = 0.07$, the balancing coefficient is $\lambda = 1.0$, $\lambda_f = 1$, $\lambda_r = 0.5$, and $\lambda_\Delta = 0.2$. The Prototype Branch is enabled during the testing phase with top-k set to 8.

4.2 Main Results

Tables 2 and 3 summarize the quantitative comparison between LaRP-CLIP and state-of-the-art zero-shot anomaly detection methods on industrial defect datasets and medical imaging datasets, respectively. The highest value of each metric (AUROC or AP/PRO) is highlighted in **red** and the second-highest value is highlighted in blue.

Table 2: Quantitative comparison on industrial defect datasets (MVTec AD, VisA, MPDD, BTAD). ↑ indicates that a higher value is better.

Methods	CLIP [18]	CLIP-AC [18]	WinCLIP [19]	VAND [22]	CoOp [40]	AnomalyCLIP [20]	AACLIP [24]	GenCLIP [21]	LaRP-CLIP
Image-level (AUROC, AP)↑									
MVTec AD	(74.1, 87.6)	(71.5, 86.4)	(91.8, **96.5**)	(86.1, 93.5)	(88.8, 94.8)	(91.5, 96.2)	(90.5, 94.9)	(90.9, 96.1)	**(92.8, 96.5)**
VisA	(66.4, 71.5)	(65.0, 70.1)	(78.1, 81.2)	(78.0, 81.4)	(62.8, 68.1)	(82.1, 85.4)	(79.0, 81.9)	(83.3, 87.5)	**(86.0, 88.0)**
MPDD	(54.3, 65.4)	(56.2, 66.0)	(63.6, 69.9)	(73.0, 80.2)	(55.1, 64.2)	(77.0, **82.0**)	(56.0, 66.6)	(73.7, 79.6)	(**77.7**,80.7)
BTAD	(34.5, 52.5)	(51.0, 62.1)	(68.2, 70.9)	(73.6, 68.6)	(66.8, 77.4)	(88.3, 87.3)	(**92.9**,96.5)	(90.0, **96.9**)	(91.6, 94.3)
Pixel-level (AUROC, PRO)↑									
MVTec AD	(38.4, 11.3)	(38.2, 11.6)	(85.2, 64.6)	(87.6, 44.0)	(33.3, 6.7)	(91.1, 81.4)	(92.3, 85.7)	(**92.7, 88.1**)	(91.5, 85.8)
VisA	(46.6, 14.8)	(47.8, 17.3)	(79.6, 56.8)	(94.2, 86.8)	(24.2, 3.8)	(95.5, 87.0)	(94.8, 83.7)	(92.7, 88.1)	**(96.7, 91.7)**
MPDD	(62.1, 33.0)	(58.7, 29.1)	(76.4, 48.9)	(94.1, 83.2)	(15.4, 2.3)	(96.5, 88.7)	(95.7, 84.8)	(96.2, 89.3)	**(97.6, 90.1)**
BTAD	(30.6, 4.4)	(32.8, 8.3)	(72.7, 27.3)	(60.8, 25.0)	(28.6, 3.8)	(94.2, 74.8)	(94.5, **79.1**)	(93.6, 75.6)	(**94.8**,77.5)

Table 3: Quantitative comparison on medical imaging datasets (HeadCT, BrainMRI, Br35H, ISIC, CVC-ColonDB, CVC-ClinicDB, TN3K). ↑ indicates that a higher value is better.

Methods	CLIP [18]	CLIP-AC [18]	WinCLIP [19]	VAND [22]	CoOp [40]	AnomalyCLIP [20]	AACLIP [24]	GenCLIP [21]	LaRP-CLIP
Image-level (AUROC, AP)↑									
HeadCT	(56.5, 58.4)	(60.0, 60.7)	(81.8, 80.2)	(89.1, 89.4)	(78.4, 78.8)	(93.4, 91.6)	(84.8, 87.4)	(96.1, 96.2)	**(96.7, 97.3)**
BrainMRI	(73.9, 81.7)	(80.6, 86.4)	(86.6, 91.5)	(89.3, 90.9)	(61.3, 44.9)	(90.3, 92.2)	(90.3, 93.3)	(92.4, 94.4)	**(95.6, 96.4)**
Br35H	(78.4, 78.8)	(82.7, 81.3)	(80.5, 82.2)	(93.1, 92.9)	(86.0, 87.5)	(94.6, 94.7)	(86.9, 88.6)	(95.0, 95.0)	**(97.7, 97.6)**
Pixel-level (AUROC, PRO)↑									
ISIC	(33.1, 5.8)	(36.0, 7.7)	(83.3, 55.1)	(89.4, 77.2)	(51.7, 15.9)	(89.7, 78.4)	(**94.0**, 79.7)	(93.1,**88.1**)	(92.5, 83.0)
CVC-ColonDB	(49.5, 15.8)	(49.5, 11.5)	(70.3, 32.5)	(78.4, 64.6)	(40.5, 2.6)	(**81.9**,71.3)	(80.2, 56.6)	(80.3, **87.8**)	(81.8, 71.0)
CVC-ClinicDB	(47.5, 18.9)	(48.5, 12.6)	(51.2, 13.8)	(80.5, 60.7)	(34.8, 2.4)	(82.9, 67.8)	(81.3, 53.2)	(81.3, **82.8**)	(**85.2**, 67.1)
TN3K	(42.3, 7.3)	(35.6, 5.2)	(70.7, 39.8)	(73.6, 37.8)	(34.0, 9.5)	(**81.5**, 50.4)	(75.9, 44.7)	(73.8, 80.6)	(**82.5**,51.9)

On the four industrial defect datasets, LaRP-CLIP achieves the best image-level performance on MVTec AD and VisA, with AUROC/AP scores of (92.8, 96.5) and (86.0, 88.0), respectively. On MPDD and BTAD, it also remains among the strongest competitors, yielding the second-best image-level results. For pixel-level evaluation, LaRP-CLIP consistently ranks first or second across all four datasets and surpasses previous CLIP-based methods in both AUROC and PRO in most cases. In particular, compared with AnomalyCLIP, the pixel-level performance is improved from (95.5, 87.0) to (96.7, 91.7) on VisA and from (96.5, 88.7) to (97.6, 90.1) on MPDD. The results show that LaRP-CLIP performs consistently well across industrial datasets with diverse defect patterns and visual characteristics.

On the medical imaging benchmarks, LaRP-CLIP also shows strong transferability. It achieves the highest image-level AUROC and AP on HeadCT, BrainMRI, and Br35H, with scores of (96.7, 97.3), (95.6, 96.4), and (97.7, 97.6), respectively. For pixel-level anomaly localization, LaRP-CLIP attains the best AUROC on CVC-ClinicDB and TN3K, while remaining highly competitive on ISIC and CVC-ColonDB. It also achieves strong PRO performance, particularly on datasets where lesion boundaries are difficult to delineate. This cross-domain performance indicates that LaRP-CLIP remains reliable even when the target anomalies differ markedly from the training domain in both appearance and imaging style.

It is also observed that LaRP-CLIP does not achieve the best PRO on all medical datasets. On ISIC and CVC-ColonDB, lower PRO scores indicate incomplete localization for lesions with irregular, low-contrast, or ambiguous boundaries. This suggests that although LaRP-CLIP shows competitive cross-domain transferability, stronger boundary-aware modeling is still needed for challenging medical lesions.

Fig. 3 presents a qualitative comparison of anomaly maps produced by different prompt strategies. The three prompts show notably different response patterns across samples. The Shallow Prompt is more sensitive to local appearance variations and can better preserve fine-grained details, making it effective for subtle texture anomalies. The Deep Prompt generates more compact and semantically aligned activation on abnormal regions, which is beneficial for anomalies with relatively clear structural semantics. In contrast, the Global Prompt frequently produces diffuse or less discriminative responses, leading to incomplete localization or increased background interference. This comparison indicates that different ViT layers capture anomaly cues at different semantic granularities, while the proposed layer-aware prompt decoupling offers a more suitable mechanism for aligning prompt representations with layer-specific features.

Fig. 4 presents qualitative anomaly localization results of LaRP-CLIP on representative samples from both industrial and medical domains. The third row shows the fused anomaly maps obtained after multi-layer aggregation, while the fourth row shows the refined anomaly maps after applying the Error-Prior Guided Local Relation Refiner. As observed, the fused maps can already roughly highlight the abnormal regions, but they still contain scattered responses and imprecise boundaries in several cases. After refinement, the anomaly regions become more compact and better aligned with the ground-truth masks, with reduced background interference and clearer structural outlines. This improvement is consistently visible in both industrial samples with thin or irregular defects and medical samples with relatively compact lesion regions, indicating that the refinement stage is beneficial for producing more precise pixel-level localization.

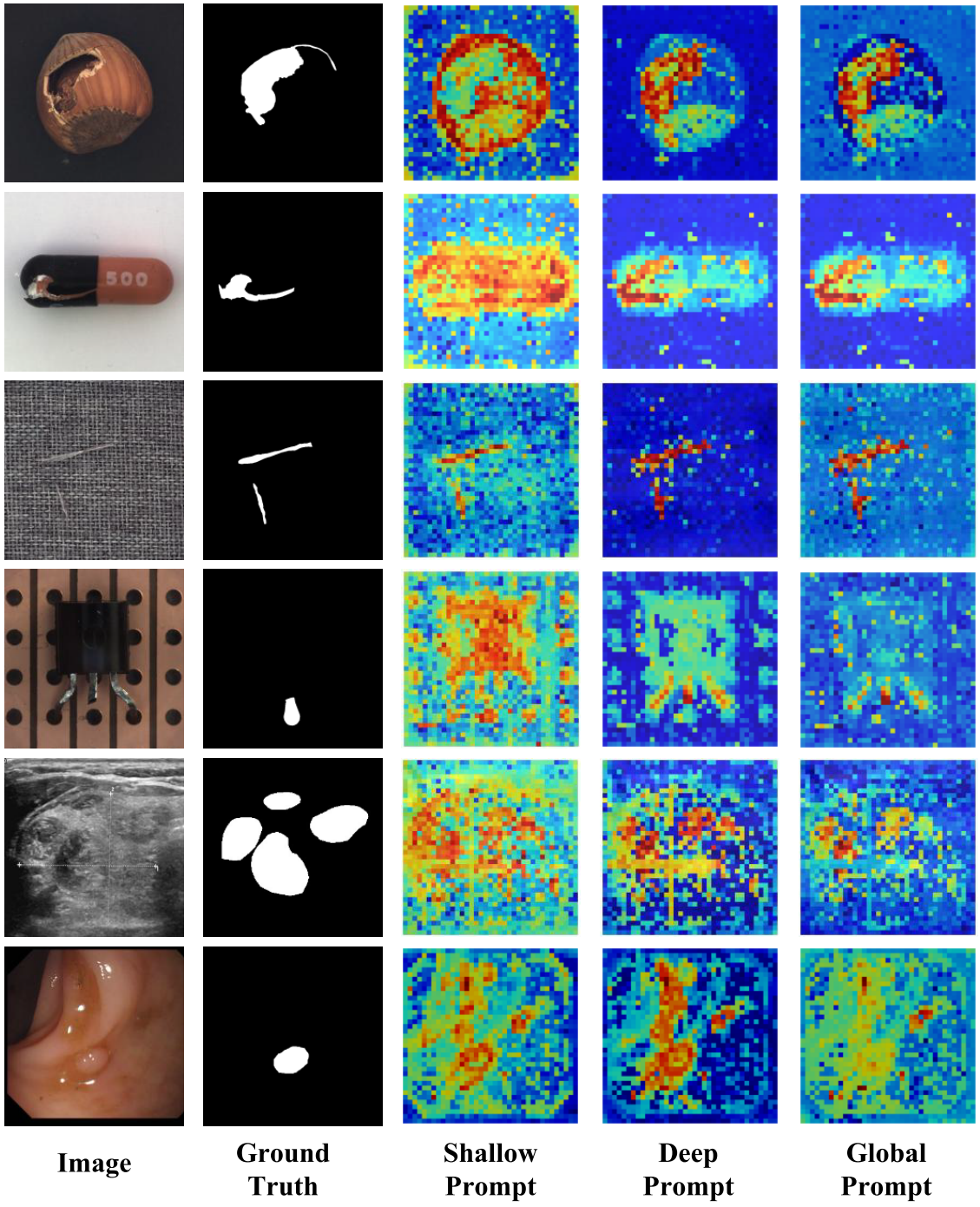

Figure 3: Qualitative comparison of anomaly maps generated using different prompts. From left to right: input image, ground-truth mask, anomaly map using shallow prompt, deep prompt, and global prompt.

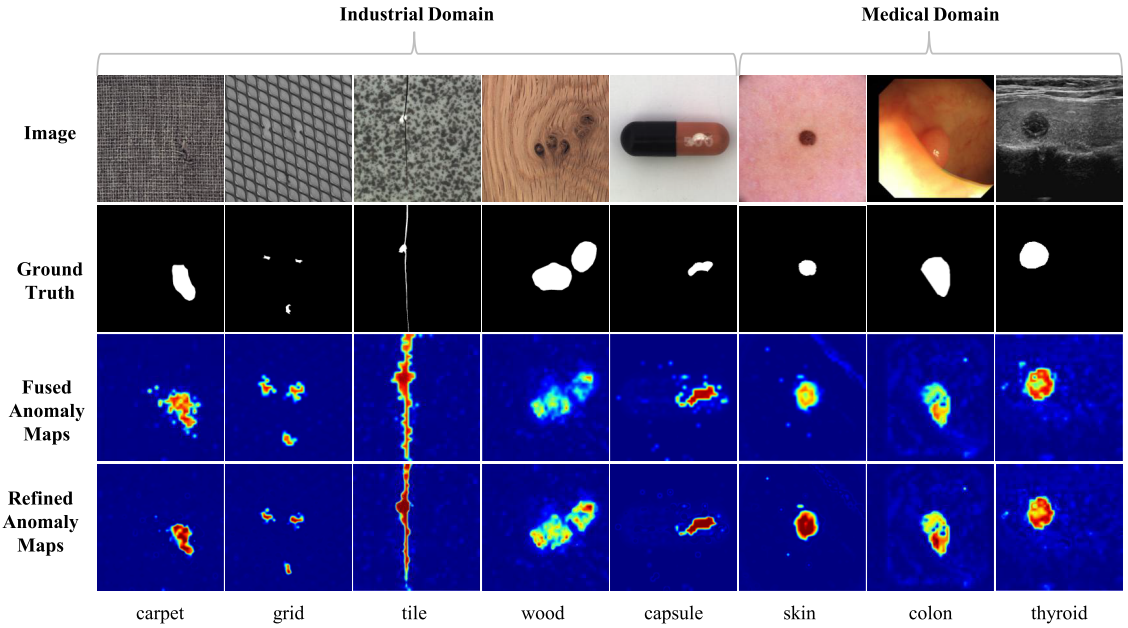

Figure 4: Qualitative anomaly localization results of LaRP-CLIP on representative samples from industrial and medical domains. From top to bottom: input image, ground-truth mask, fused anomaly map, and refined anomaly map. The refined maps exhibit clearer boundaries and reduced background interference compared with the fused maps.

Fig. 5 presents a qualitative comparison between the global similarity map and the prototype-guided similarity map. The selected high-confidence normal prototype regions in the second column are used to build an image-specific normal reference for the current test sample. Compared with the global similarity map, which tends to exhibit scattered or distracting responses on normal structures, the prototype-guided similarity map yields cleaner activations and better highlights the actual abnormal regions. The improvement is visible across samples from both industrial and medical domains, indicating that the Prototype Branch helps reduce the influence of normal appearance variation and complex background content during anomaly localization.

4.3 Ablation Study

To assess the contribution of each component in LaRP-CLIP, an ablation study is conducted on MVTec AD. The results are reported in Table 4, where image-level performance is evaluated by AUROC and AP, and pixel-level performance is evaluated by AUROC and PRO.

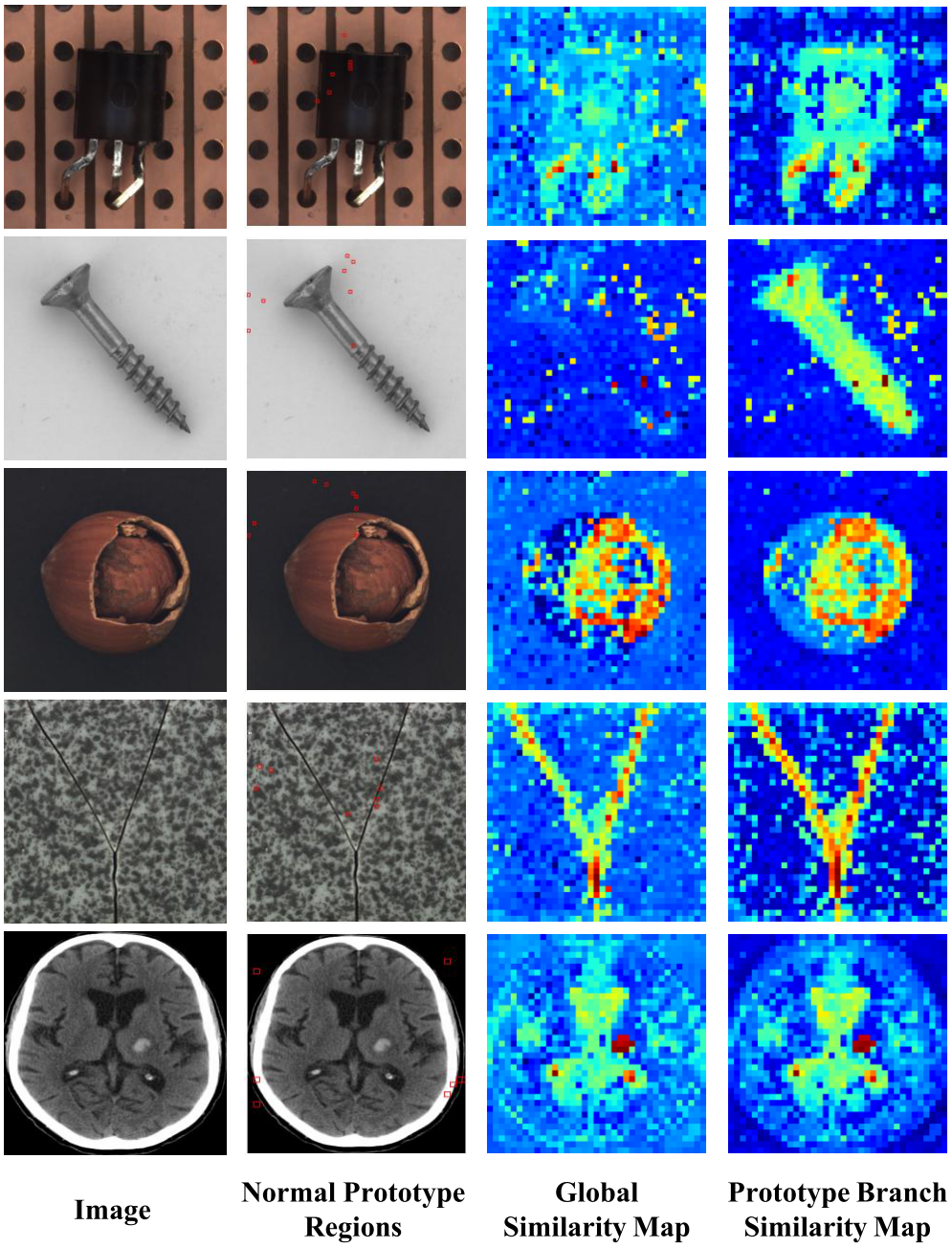

Figure 5: Qualitative results showing the effect of the prototype branch. From left to right: input image, selected high-confidence normal prototype regions (red dots), global similarity map, and prototype-guided similarity map.

Table 4: Ablation study on MVTec AD. "w/o" denotes the removal of the corresponding component.

Variant	Image-Level	Pixel-Level
Full LaRP-CLIP	(92.8, 96.5)	(91.5, 85.8)
w/o Layer-Aware Prompt Decoupling	(88.4, 93.7)	(87.9, 80.2)
w/o Adaptive Weighted Fusion	(89.7, 94.9)	(89.1, 82.6)
w/o Local Relation Refiner	(90.2, 95.1)	(88.4, 81.9)
w/o Prototype Branch	(90.5, 95.4)	(91.0, 85.1)

As shown in Table 4, removing any component leads to a decline in performance, indicating that each module contributes to the overall capability of the framework. Among all variants, removing the Layer-Aware Prompt Decoupling causes the largest performance drop, with image-level results decreasing from (92.8, 96.5) to (88.4, 93.7) and pixel-level results decreasing from (91.5, 85.8) to (87.9, 80.2). This result suggests that aligning prompt representations with the semantic granularity of different ViT layers is critical for both anomaly classification and localization, which is also consistent with the qualitative comparison in Fig. 3.

When the Adaptive Weighted Fusion is removed, the performance decreases to (89.7, 94.9) at the image level and (89.1, 82.6) at the pixel level, suggesting that fixed fusion is less effective for aggregating anomaly cues from multiple layers. Removing the Local Relation Refiner also results in a noticeable decline, particularly in pixel-level localization, where the PRO score drops from 85.8 to 81.9. This observation is consistent with the qualitative comparison in Fig. 4, where the refinement stage produces cleaner anomaly regions and clearer boundaries.

By comparison, removing the Prototype Branch leads to a relatively smaller drop, from (92.8, 96.5) to (90.5, 95.4) at the image level and from (91.5, 85.8) to (91.0, 85.1) at the pixel level. Although its effect is less pronounced than that of the other modules, it still brings consistent gains, indicating that image-specific normal prototypes help stabilize anomaly scoring under appearance variation. This trend is also supported by the visual comparison in Fig. 5, where the prototype-guided similarity maps show more concentrated responses on abnormal regions and less interference from normal structures.

The results reveal clear performance degradation when any component is removed. Removing layer-aware prompt decoupling leads to the largest drop, decreasing image-level AUROC by 4.4% and pixel-level AUROC by 3.6%, confirming its critical role in resolving semantic granularity mismatch across layers. Removing the Error-Prior Guided Local Relation Refiner causes noticeable declines in pixel-level metrics (AUROC drops by 3.1% and PRO by 4.0%), highlighting its importance in sharpening boundaries and suppressing noise. Eliminating the Prototype Branch primarily affects image-level robustness, while equal-weight fusion and the absence of error prior both result in suboptimal fusion quality.

To further evaluate the influence of key hyperparameters, sensitivity analyses are conducted on MVTec AD. Three factors are considered: the balancing coefficient λ, the top-k value used in prototype-guided pooling, and the shallow/deep layer partition. Since λ and top-k directly affect the final image-level anomaly score, their influence is evaluated using image-level AUROC and AP. The shallow/deep layer partition affects layer-aware prompt matching and anomaly map generation, and is therefore evaluated using pixel-level AUROC and PRO.

As shown in Table 5, LaRP-CLIP remains stable with different λ and top-k values. When $\lambda = 0$, the score relies only on global text similarity, leading to weaker detection. Increasing λ incorporates local anomaly evidence, improving performance, but too large a λ overemphasizes local responses, slightly degrading performance. The best result is achieved at $\lambda = 1.0$, where global and local evidence complement each other.

Table 5: Sensitivity analysis of λ and top-k on MVTec AD. Image-level results are reported as (AUROC, AP).

Parameter	Setting	Image-Level
λ	0.0	(91.6, 96.1)
λ	0.5	(92.4, 96.4)
λ	1.0	(92.8, 96.5)
λ	2.0	(92.5, 96.3)
λ	4.0	(92.1, 96.0)
top-k	1	(92.0, 96.1)
top-k	4	(92.5, 96.4)
top-k	8	(92.8, 96.5)
top-k	16	(92.6, 96.4)
top-k	32	(92.2, 96.1)

For top-k, a very small value is more sensitive to noisy local activations, while a very large value may include normal background regions and dilute anomaly evidence. The setting top-k = 8 achieves the best performance and is therefore adopted as the default configuration.

Table 6 shows that assigning only layer 6 to the shallow group lacks sufficient texture supervision, while assigning layer 18 weakens the distinction between shallow and deep prompts. The best pixel-level performance is achieved with the partition $[6, 12]/[18, 24]$, aligning with the ViT feature hierarchy: shallow layers capture texture, while deep layers encode structure and semantics.

Table 6: Sensitivity analysis of shallow/deep layer partition on MVTec AD. Pixel-level results are reported as (AUROC, PRO).

Layer Partition (S/D)	Pixel-Level
$[6]/[12, 18, 24]$	(90.7, 84.6)
$[6, 12]/[18, 24]$	(91.5, 85.8)
$[6, 12, 18]/[24]$	(91.0, 85.1)

4.4 Time Complexity Analysis

Let N denote the number of image patches and L denote the number of selected feature layers. In LaRP-CLIP, the extra computation beyond the frozen CLIP backbone comes from layer-wise prompt matching, adaptive fusion, refinement, and prototype construction.

Layer-aware prompt matching and patch–text similarity computation across L layers require $O(LN)$. Adaptive weighted fusion and error-prior estimation are also performed on layer-wise anomaly responses, with complexity $O(LN)$. The Local Relation Refiner is implemented using lightweight 3×3 convolutions and has complexity $O(N)$. The Prototype Branch, including high-confidence patch selection and prototype construction, also scales as $O(N)$.

The total additional complexity is therefore $O(LN)$, excluding the frozen backbone. This is on the same order as existing CLIP-based methods that rely on multi-layer inference. On a single NVIDIA RTX 4090 GPU, with an input resolution of 518×518, LaRP-CLIP takes about 280 ms to process one image during inference.

5 Conclusion

This paper presents LaRP-CLIP, a zero-shot anomaly detection framework for industrial and medical images. The framework introduces layer-aware prompt decoupling, adaptive weighted fusion with an Error-Prior Guided Local Relation Refiner, and a Prototype Branch for image-specific normal prototype construction. These designs are intended to alleviate three common issues in existing CLIP-based anomaly detection methods, namely semantic granularity mismatch across layers, inflexible fusion across feature levels, and interference from background or normal appearance variations.

Experimental results on multiple industrial defect and medical imaging datasets show that LaRP-CLIP achieves strong performance in both image-level detection and pixel-level localization. The qualitative results further show that the proposed framework produces cleaner anomaly responses and more precise boundaries in challenging cases. These results suggest that LaRP-CLIP is a practical zero-shot solution for anomaly detection in scenarios where anomalous samples are limited or unavailable. Future work will consider extending the framework to multi-class anomaly settings and to more complex real-world environments.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the Key Research and Development Program of Zhejiang Province No. 2023C01141, the Science and Technology Innovation Community Project of Yangtze River Delta No. 23002410100.

Author Contributions: Xing Fang: Responsible for software implementation, experimental validation, and preparation of the original manuscript draft. Yuanfang Chen: Led the conceptualization and methodological design, provided project administration and supervision, contributed to manuscript revision and editing, and secured funding. Qiang Lin: Conducted data collection. Kun Yang: Offered methodological guidance, technical support, and constructive feedback during manuscript review. Gyu Myoung Lee: Provided conceptual guidance and supervision. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All datasets used in this study are publicly accessible, and links to the original data sources are included in the manuscript for reference.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wu P, Pan C, Yan Y, Pang G, Yan Q, Wang P, et al. Deep learning for video anomaly detection: a review. IEEE Trans Neural Netw Learn Syst. 2026. doi:10.1109/TNNLS.2025.3647892.
2. Huang H, Wang P, Pei J, Wang J, Alexanian S, Niyato D. Deep learning advancements in anomaly detection: a comprehensive survey. IEEE Internet Things J. 2025;12(21):44318–42. doi:10.1109/JIOT.2025.3585884.
3. Li Z, Yan Y, Wang X, Ge Y, Meng L. A survey of deep learning for industrial visual anomaly detection. Artif Intell Rev. 2025;58(9):279. doi:10.1007/s10462-025-11287-7.
4. Ammar MB, Mendoza A, Belkhir N, Manzanera A, Franchi G. Foundation models and Transformers for anomaly detection: a survey. Inf Fusion. 2025;126(7):103517. doi:10.1016/j.inffus.2025.103517.
5. Laghari AA, Khan AA, Ksibi A, Hajjej F, Kryvinska N, Almadhor A, et al. A novel and secure artificial intelligence enabled zero trust intrusion detection in industrial internet of things architecture. Sci Rep. 2025;15(1):26843. doi:10.1038/s41598-025-11738-9.
6. Xie S, Wu X, Wang MY. Semi-patchcore: a novel two-staged method for semi-supervised anomaly detection and localization. IEEE Trans Instrum Meas. 2025;74:3506012. doi:10.1109/TIM.2025.3527588.

7. Li R, Ma H, Wang R, Song H, Zhou X, Wang L, et al. Application of unsupervised learning methods based on video data for real-time anomaly detection in wire arc additive manufacturing. J Manuf Process. 2025;143:37–55. doi:10.1016/j.jmapro.2025.03.113.
8. Koren O, Koren M, Peretz O. A procedure for anomaly detection and analysis. Eng Appl Artif Intell. 2023;117(2):105503. doi:10.1016/j.engappai.2022.105503.
9. Nizam H, Zafar S, Lv Z, Wang F, Hu X. Real-time deep anomaly detection framework for multivariate time-series data in industrial IoT. IEEE Sens J. 2022;22(23):22836–49. doi:10.1109/JSEN.2022.3211874.
10. Liu L, Li J, Lv J, Wang J, Zhao S, Lu Q. Privacy-preserving and secure industrial big data analytics: a survey and the research framework. IEEE Internet Things J. 2024;11(11):18976–99. doi:10.1109/JIOT.2024.3353727.
11. Aich A, Peng KC, Roy-Chowdhury AK. Cross-domain video anomaly detection without target domain adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2023 Jan 2–7; Waikoloa, HI, USA. p. 2579–91. doi:10.1109/WACV56688.2023.00261.
12. Hashemi MJ, Keller E, Tizpaz-Niari S. Detecting unseen anomalies in network systems by leveraging neural networks. IEEE Trans Netw Serv Manag. 2022;20(3):2515–28. doi:10.1109/TNSM.2022.3220775.
13. Yang Z, Soltani I, Darve E. Anomaly detection with domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. p. 2958–67. doi:10.1109/CVPRW59228.2023.00297.
14. Cho M, Kim T, Shim M, Wee D, Lee S. Towards multi-domain learning for generalizable video anomaly detection. Adv Neural Inf Process Syst. 2024;37:50256–84. doi:10.52202/079017-1591.
15. Mao W, Wang G, Kou L, Liang X. Deep domain-adversarial anomaly detection with one-class transfer learning. IEEE/CAA J Autom Sin. 2023;10(2):524–46. doi:10.1109/JAS.2023.123228.
16. Li A, Qiu C, Kloft M, Smyth P, Rudolph M, Mandt S. Zero-shot anomaly detection via batch normalization. Adv Neural Inf Process Syst. 2023;36:40963–93. doi:10.52202/075280-1784.
17. Peng Y, Lin X, Ma N, Du J, Liu C, Liu C, et al. SAM-LAD: segment anything model meets zero-shot logic anomaly detection. Knowl-Based Syst. 2025;314(4):113176. doi:10.1016/j.knosys.2025.113176.
18. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning; 2021 Jul 18–24; Virtually. p. 8748–63. doi:10.48550/arXiv.2103.00020.
19. Jeong J, Zou Y, Kim T, Zhang D, Ravichandran A, Dabeer O. WinCLIP: zero-/few-shot anomaly classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. p. 19606–16. doi:10.1109/CVPR52729.2023.01878.
20. Zhou Q, Pang G, Tian Y, He S, Chen J. AnomalyCLIP: object-agnostic prompt learning for zero-shot anomaly detection. In: Proceedings of the International Conference on Learning Representation; 2023 May 1–5; Kigali, Rwanda. doi:10.48550/arXiv.2310.18961.
21. Kim D, Park C, Cho S, Lim H, Kang M, Lee J, et al. Generalizing CLIP prompts for zero-shot anomaly detection. Pattern Recognit. 2026;178(9):113406. doi:10.1016/j.patcog.2026.113406.
22. 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:2305.17382. 2023. doi:10.48550/arXiv.2305.17382.
23. 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; 2022 Oct 23–27; Tel Aviv, Israel. Cham, Switzerland: Springer; 2022. p. 392–408. doi:10.1007/978-3-031-20056-4_23.
24. Ma W, Zhang X, Yao Q, Tang F, Wu C, Li Y, et al. AA-CLIP: enhancing zero-shot anomaly detection via anomaly-aware CLIP. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025 Jun 10–17; Nashville, TN, USA. p. 4744–54. doi:10.1109/CVPR52734.2025.00447.
25. Ruff L, Kauffmann JR, Vandermeulen RA, Montavon G, Samek W, Kloft M, et al. A unifying review of deep and shallow anomaly detection. Proc IEEE. 2021;109(5):756–95. doi:10.1109/JPROC.2021.3052449.

26. Gong D, Liu L, Le V, Saha B, Mansour MR, Venkatesh S, et al. Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 1705–14. doi:10.1109/ICCV.2019.00179.
27. Schlegl T, Seeböck P, Waldstein SM, Langs G, Schmidt-Erfurth U. f-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks. Med Image Anal. 2019;54(3):30–44. doi:10.1016/j.media.2019.01.010.
28. Ruff L, Vandermeulen R, Goernitz N, Deecke L, Siddiqui SA, Binder A, et al. Deep one-class classification. In: Proceedings of the International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden. p. 4393–402.
29. Ren J, Tang T, Jia H, Xu Z, Fayek H, Li X, et al. Foundation models for anomaly detection: vision and challenges. AI Mag. 2025;46(4):e70045. doi:10.1002/aaai.70045.
30. Li Y, Goodge A, Liu F, Foo CS. PromptAD: zero-shot anomaly detection using text prompts. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2024 Jan 3–8; Waikoloa, HI, USA. p. 1093–102. doi:10.1109/WACV57701.2024.00113.
31. Chen Z, Zhao Z, Guo J, Li J, Huang Z. SVIP: semantically contextualized visual patches for zero-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2025 Oct 19–20; Honolulu, HI, USA. p. 3346–56. doi:10.1109/ICCV51701.2025.00320.
32. Bergmann P, Fauser M, Sattlegger D, Steger C. MVTec AD—a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 16–20; Long Beach, CA, USA. p. 9592–600. doi:10.1109/CVPR.2019.00982.
33. 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 2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT); 2021 Oct 25–27; Brno, Czech Republic. New York, NY, USA: IEEE. p. 66–71. doi:10.1109/ICUMT54235.2021.9631567.
34. Mishra P, Verk R, Fornasier D, Piciarelli C, Foresti GL. VT-ADL: a vision transformer network for image anomaly detection and localization. arXiv:2104.10036. 2021. doi:10.48550/arxiv.2104.10036.
35. Salehi M, Sadjadi N, Baselizadeh S, Rohban MH, Rabiee HR. Multiresolution knowledge distillation for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 14902–12. doi:10.1109/CVPR46437.2021.01466.
36. Hamada A. Br35H:: brain tumor detection 2020. IEEE Dataport. 2025. doi:10.21227/t5k6-5r54.
37. Codella NC, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 2018 Apr 4–7; Washington, DC, USA. New York, NY, USA: IEEE; 2018. p. 168–72. doi:10.1109/ISBI.2018.8363547.
38. Tajbakhsh N, Gurudu SR, Liang J. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imaging. 2015;35(2):630–44. doi:10.1109/TMI.2015.2487997.
39. Gong H, Chen G, Wang R, Xie X, Mao M, Yu Y, et al. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In: Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI); 2021 Apr 20–21; Nice, France. New York, NY, USA: IEEE; 2021. p. 257–61. doi:10.1109/ISBI48211.2021.9434087.
40. Zhou K, Yang J, Loy CC, Liu Z. Learning to prompt for vision-language models. Int J Comput Vis. 2022;130(9):2337–48. doi:10.1007/s11263-022-01653-1.