This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DLVP-CLIP: Enhancing Fine-Grained Zero-Shot Anomaly Detection via Dynamic Local Visual Prompting

Gaowei Zhang, Lihe Zhang*
Dalian University of Technology
zhanggaowei@mail.dlut.edu.cn, zhanglihe@dlut.edu.cn

Abstract

Zero-shot anomaly detection (ZSAD) aims to utilize auxiliary data to train models for generalized learning of unseen categories, which has important application value in fields such as industrial quality inspection and medical diagnosis. Although methods based on CLIP show potential, their pre-training objective of focusing on overall semantic alignment between images and text makes the model insensitive to local details, which is inherently contradictory to the need for fine-grained local features in anomaly detection. Existing improvement methods rely on predefined text prompt frameworks to perceive local information, but struggle to effectively address the issue of insufficient local perception. To address this, this paper proposes a dynamic local visual prompting method based on CLIP (DLVP-CLIP). DLVP dynamically identifies and extracts local visual features from key regions in images as prompt tokens using the Semantic-Aware Local Feature Selector (SLFS) module, and utilizes the multi-modal local prompt (MLoP) module to jointly optimize representations in both visual and textual spaces, achieving more precise cross-modal alignment. Additionally, the high-low frequency decomposition module (HFD) is introduced to separate and process global structural and local textural information via wavelet transformation, thereby enhancing detail perception. Extensive experiments on 13 anomaly detection datasets demonstrate that DLVP-CLIP achieves outstanding ZSAD performance on datasets from the industrial and medical domains.

1. Introduction

The core objective of anomaly detection (AD) is to model data distributions to identify anomalous samples, such as industrial product defects [2] or medical image lesions [9]. While traditional methods [11, 23, 24] perform well when labeled data is abundant, their strong reliance on labeled data severely limits the model's generalization ability on scarce samples or new categories. This limitation is particularly pronounced in industrial quality inspection [18, 29] and medical diagnosis domains [4, 27]. To address this, researchers have proposed Zero-Shot Anomaly Detection (ZSAD), which aims to directly generalize the model to unseen categories after training with auxiliary data [13, 16, 34]. However, differences in background features across products/organs, the diversity of anomaly morphologies, and variations in visual appearance make it challenging for existing methods to achieve robust cross-domain generalization.

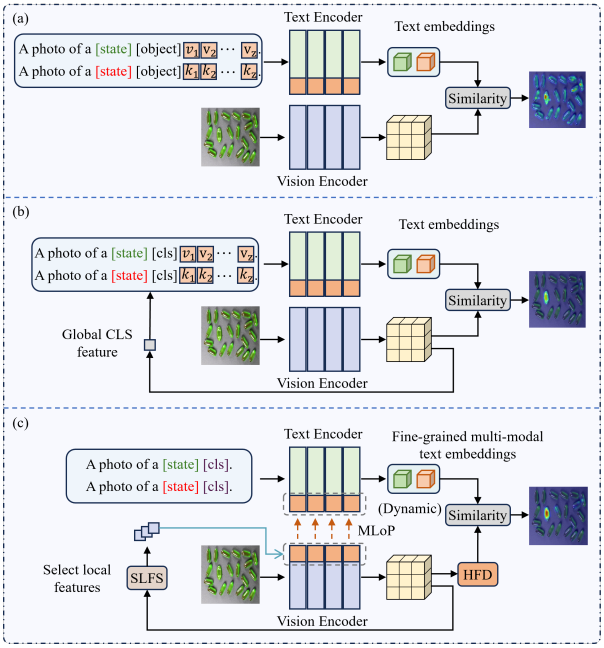


Figure 1. Comparison between existing CLIP-based methods and our DLVP-CLIP. (a) Generate general text prompts. (b) Adapt the global visual features of CLIP to the text prompts. (c) Our DLVP-CLIP extracts local visual features as dynamic prompts and achieves more accurate cross-modal alignment through visual-textual joint encoding.

In recent years, CLIP [32] has provided new insights for

*Corresponding author.

few/zero-shot anomaly detection due to its strong zero-shot generalization capabilities [18]. These methods leverage the alignment between visual and textual representations to describe anomalies through language prompts, eliminating the need for category-specific training. As shown in Figure 1(a), AnomalyCLIP [43] introduces object-agnostic prompts by replacing product names with a generic "object" token, enabling domain transfer. AA-CLIP [26] employs residual adapters to construct anomaly-aware text representations and guide visual attention toward potential defect regions. TPS [25] enhances multi-level alignment by introducing dual-level text prompts for global anomaly classification and pixel-level segmentation. Additionally, Figure 1(b) illustrates that AdaCLIP [4] and VCP-CLIP [30] improve the encoding of category semantics in the text space by incorporating global image features.

Although existing methods have improved CLIP's performance in anomaly detection, they have failed to address its inherent weakness of being insensitive to local features. Anomaly detection relies heavily on the accurate capture of local details in images. However, CLIP's pre-training objective focuses on the overall semantic alignment between images and text. This causes the model to tend to extract object-level global semantic features, while paying insufficient attention to the fine visual features of local areas. This characteristic creates an inherent contradiction with the requirements of anomaly detection.

Constrained by this contradiction, current improvement methods face challenges. Text-driven methods (AnomalyCLIP, AA-CLIP, TPS) rely on pre-designed text prompts to perceive local details, which makes it difficult to dynamically adapt to complex and variable anomaly patterns. On the other hand, methods (AdaCLIP, VCP-CLIP) that introduce visual global CLS features can enhance the overall description at the instance level. However, by further focusing on global representations, they inevitably weaken the ability to represent subtle local anomalies.

To this end, this paper proposes Dynamic Local Visual Prompting based on CLIP (DLVP-CLIP), which explicitly extracts key local visual features from images as dynamic prompts and achieves more precise cross-modal alignment through visual-text joint encoding. As shown in Figure 1(c), DLVP consists of a semantic-aware local feature selector (SLFS) and a multi-modal local prompt module (MLoP). SLFS identifies semantic key regions through the attention map of the CLS token and extracts features from high-attention image blocks as dynamic prompt tokens. Subsequently, the MLoP module utilizes these fine-grained visual features to jointly optimize the feature representations of the text and visual spaces. Embedded visual tokens provide rich, fine-grained semantic context, significantly enhancing the model's ability to capture image details. This approach mitigates the limitations of predefined text prompts and reduces overemphasis on global features. Additionally, ViT exhibits a low-pass filtering effect during global feature aggregation that tends to smooth fine textures [5]. By decoupling and separately modeling low-frequency global structures and high-frequency local textures through wavelet transformation, High-Low Frequency Decomposition (HFD) effectively preserves subtle details, enhancing the model's fine-grained perception capabilities.

Our contributions are summarized as follows:

- We propose DLVP-CLIP, which achieves more flexible and context-aware cross-modal alignment by co-encoding dynamically extracted visual cues with text prompts, significantly improving fine-grained anomaly detection accuracy while maintaining zero-shot generalization capabilities.
- We propose DLVP, which consists of SLFS and MLoP modules. SLFS is used to extract high-attention image patches as dynamic prompt tokens, and the MLoP utilizes the extracted prompt tokens to jointly optimize feature representations in both text and visual spaces through joint visual-text encoding.
- To further enhance the model's ability to perceive details, we propose a HFD method based on wavelet transformation to separately process global structure and local texture.
- Comprehensive experiments on 13 real-world industrial and medical datasets demonstrate that DLVP-CLIP achieves outstanding ZSAD performance.

2. Related Work

2.1. Zero-Shot Anomaly Detection (ZSAD)

ZSAD has achieved significant performance gains by leveraging large visual language models (e.g., CLIP). WinCLIP [18] was the first to apply CLIP to the ZSAD domain, aggregating classification results by aligning text with sub-images at multiple scales. However, due to CLIP being pre-trained on general object semantics rather than abnormal patterns, its anomaly detection capabilities remain limited. APRIL-GAN [6] and CLIP-AD [7] introduce adapters to optimise visual features and enhance image block-level visual representations. AnomalyCLIP [43] aligns patch-level markers by learning category-independent prompt embeddings and adding learnable markers to the text intermediate layer to optimise the base CLIP representation. AdaCLIP [4] and VCP-CLIP [30] adopt a similar approach and further integrate visual knowledge into the text prompt embedding. Bayes-PFL [31] learns the distribution of the text prompt and uses sampling to cover the prompt space. FE-CLIP [10] injects frequency information through two complementary adapters, thereby better capturing anomaly patterns in the cross-domain.

Unlike these methods, DLVP-CLIP employs a locally

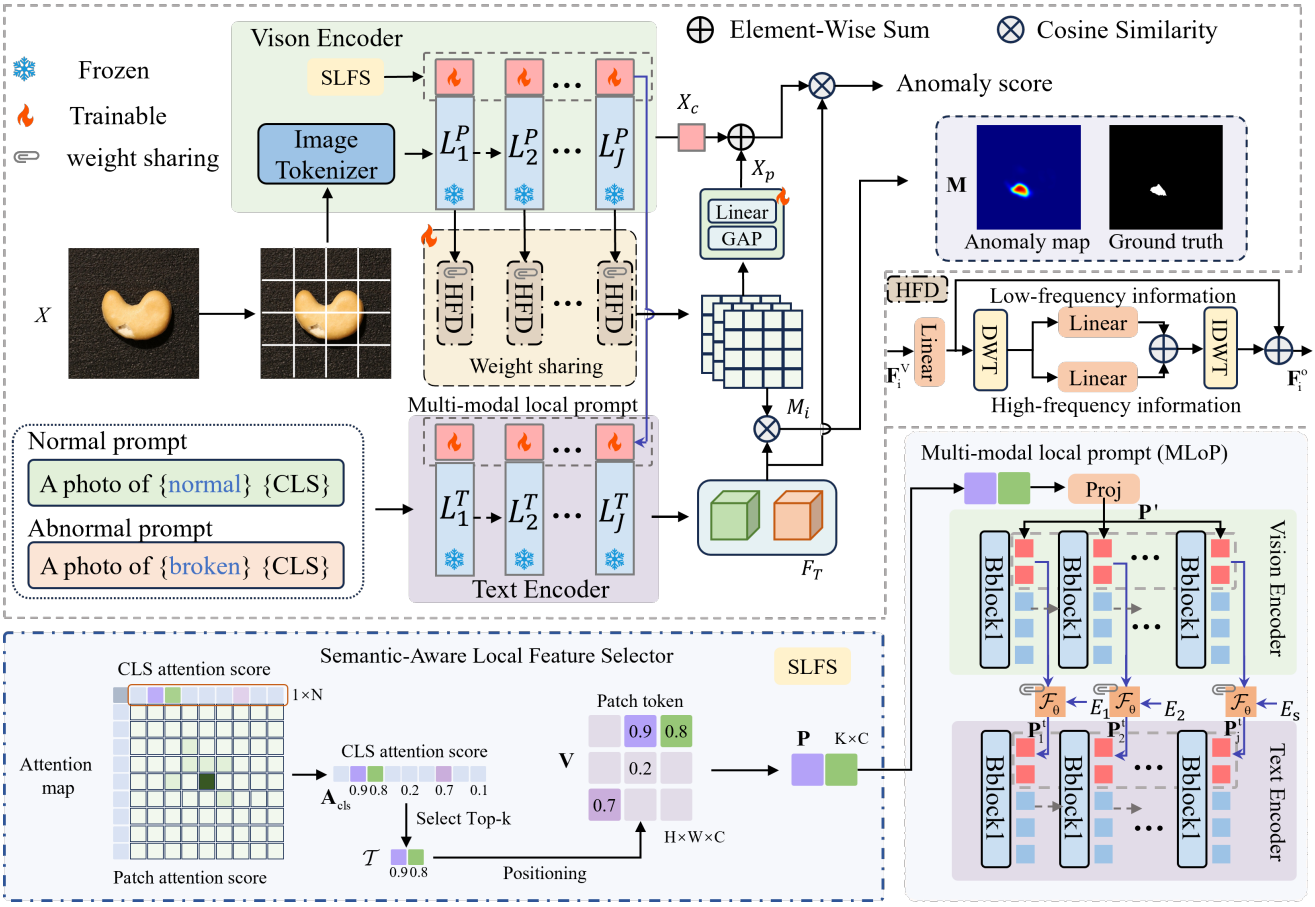


Figure 2. Architecture of DLVP-CLIP. First, the SLFS module generates dynamic visual prompts from CLIP's CLS attention maps. Then, MLoP embeds these prompts into Transformer layers of visual/text encoders, guiding local detail focus and semantic anchoring. Finally, HFD uses wavelet transforms to separately process frequency components, jointly supports image-level anomaly detection and pixel-level anomaly localization.

visual-driven dynamic prompt generation mechanism. It extracts key local features from images to build context-aware prompt semantics and enhances the alignment between text and visual spaces through multi-modal interaction.

2.2. Prompt Learning

Prompt learning methods [33, 35, 38, 41] have become the core paradigm for adapting visual-language models (VLMs) to downstream tasks. Early studies such as CoOp [42] pioneered the use of learnable text prompts to replace traditionally manually designed prompts, significantly enhancing the model's adaptability to specific tasks. However, such methods still face challenges in terms of insufficient transferability when handling out-of-distribution data. To address this, CoCoOp [41] guides the pre-trained backbone network to adapt to the target domain using visual context prompts. Furthermore, MaPLe [21] and PromptSRC [22] extend prompt learning to the bimodal space, achieving deeper cross-modal alignment by jointly regulating image and text features.

3. Method

3.1. Problem Formulation

ZSAD aims to identify anomalies from unseen object categories that are disjoint from the training categories. Formally, a model is trained on a dataset $\mathcal{D}_{\text{train}}$ containing normal or anomalous samples to learn cross-category shared patterns, enabling generalization to n distinct test sets $\{\mathcal{D}^1_{\text{test}}, \mathcal{D}^2_{\text{test}}, \ldots, \mathcal{D}^n_{\text{test}}\}$ where $\mathcal{D}^i_{\text{test}} \cap \mathcal{D}_{\text{train}} = \emptyset$ $(\forall i)$. At the image level, AD is defined as binary classification: for an input x, the model predicts $y = 0$ (normal) or $y = 1$ (anomaly). For pixel-level anomaly segmentation, the model additionally generates a mask $\mathcal{S} \in \{0,1\}^{H\times W}$ where $s_{h,w} = 1$ indicates an anomalous pixel at location (h, w).

3.2. Overview

Figure 2 illustrates the main structure of DLVP-CLIP. First, SLFS extracts the CLS token attention map from the original CLIP visual encoder. Based on this attention map, it extracts key local details and forms visual prompt tokens. Subsequently, MLoP embeds these visual prompt tokens into the Transformer layers of CLIP's visual and text encoders. In the visual encoder, the local prompts guide the model to focus on key local details during feature extraction. Meanwhile, in the text encoder, they anchor text feature representations to specific image details, thereby enhancing semantic precision and mitigating the limitations of generic text prompts in capturing fine-grained associations. Additionally, to further enhance feature representations, we introduce HFD. This module utilizes wavelet transforms to decompose visual features into high-frequency and low-frequency components, processing them separately to extract local details and global structural information collaboratively.

3.3. Dynamic Local Visual Prompt

To effectively capture and understand critical local details in images and their fine-grained semantic associations with text descriptions, we propose the DLVP. DLVP consists of the SLFS and MLoP.

Semantic-Aware Local Feature Selector. Existing prompt-based visual-language models (such as CLIP) are typically optimized at the global image-text alignment level, making it difficult to effectively capture key local details and their fine-grained semantic associations with text. This limitation significantly restricts the model's performance in tasks requiring a precise understanding of image content. Unlike AdaCLIP [4], which uses the global features of the CLS token as prompts, this paper proposes explicitly injecting local visual semantics to enhance the fine-grained alignment of the text space through key local information.

In the visual Transformer, the CLS token generates image-level semantic representations by aggregating global contextual information, with its attention weights essentially encoding the contribution of each image region to the overall semantics. Given the attention matrix $A \in \mathbb{R}^{(N+1)\times(N+1)}$ from the final transformer layer, where N is the number of image patches. We extract the attention vector $\mathbf{A}_{cls} \in \mathbb{R}^N$ for the CLS token, where each element corresponds to a semantic importance score for an image patch. Based on this, we select the top K token indices with the highest attention weights: $\mathcal{T} = \{t_i \mid \mathbf{A}_{cls}[t_i] \in Topk(A_{cls}, K)\}_{i=1}^{K}$. Then, local features are extracted from the visual features $\mathbf{V}$ based on $\mathbf{P} = \mathbf{V}[\mathcal{T}] \in R^{K\times C}$, where $\mathbf{P}$ aggregates K semantically critical local features, serving as input for subsequent prompt learning.

Multi-modal Local Prompt. To promote collaborative optimization between text and visual modalities, we inject the projected local features as dynamic prompts into both visual and text encoders. This design establishes fine-grained semantic associations within the cross-modal feature space, thereby enhancing the model's sensitivity to subtle local variations.

Specifically, the local visual feature matrix $\mathbf{P}$ is transformed into prompt tokens $\mathbf{P}' \in \mathbb{R}^{M\times C}$. In the visual encoder, $\mathbf{P}'$ is concatenated with the regular input tokens $\mathbf{T} \in \mathbb{R}^{N\times C}$ along the sequence dimension, with the constraint $M \ll N$. Let L_j^P denote the j-th prompt-enhanced Transformer layer, whose forward propagation is processed layer-wise as:

$$[\mathbf{T}_{j+1}, _] = L_j^P([\mathbf{T}_j, \mathbf{P}']), \quad j \le J, \tag{1}$$

$$[\mathbf{T}_{j+1}, \mathbf{P}'_{j+1}] = L_j^P([\mathbf{T}_j, \mathbf{P}']), \quad j > J, \tag{2}$$

where J denotes the maximum layer depth of learnable prompts, and $[\cdot, \cdot]$ denotes concatenation along the sequence dimension.

On the text side, the projected features $\mathbf{P}'$ are passed through a projection layer $\mathcal{F}\theta$ and combined with a static learnable vector $\mathbf{E}j \in \mathbb{R}^{M\times D}$ to generate text prompt tokens:

$$\mathbf{P}_j^t = \mathcal{F}_\theta(\mathbf{P}') + \mathbf{E}_j, \quad \mathbf{P}_j^t \in \mathbb{R}^{M\times D}. \tag{3}$$

These tokens are injected into the text encoder via the same mechanism, concatenated with text tokens $\mathbf{S}_j \in \mathbb{R}^{L\times D}$ to form $[\mathbf{S}_j, \mathbf{P}_j^t] \in \mathbb{R}^{(L+M)\times D}$, where D is the text embedding dimension and L is the text token length:

$$[\mathbf{S}_{j+1}, _] = L_j^T([\mathbf{S}_j, \mathbf{P}_j^t]), \quad j \le J, \tag{4}$$

$$[\mathbf{S}_{j+1}, \mathbf{P}_{tj+1}] = L_j^T([\mathbf{S}_j, \mathbf{P}_j^t]), \quad j > J, \tag{5}$$

where L_j^T denotes the j-th prompt-enhanced text Transformer layer.

3.4. High-low Frequency Decomposition

Anomalies often appear as subtle local texture variations. However, the self-attention mechanism in ViT aggregates features globally, effectively acting as a low-pass filter. It computes similarities between each patch and all others, fusing them into a global representation. This process smooths out high-frequency signals (e.g., fine textures) while emphasizing low-frequency, semantically consistent patterns (e.g., global structures). To address this limitation, HFD explicitly decomposes features into low- and high-frequency components via DWT. By modeling the high-frequency branch independently, HFD enhances fine-grained anomaly cues that would otherwise be suppressed.

Given the input feature $\mathbf{F}_i^V \in \mathbb{R}^{N\times C}$ at layer i, we first adjust the channel dimension through a linear projection layer to generate spatially-aware features:

$$\tilde{\mathbf{F}}_i^V = \mathcal{F}_v\left(\mathbf{F}_i^V\right). \tag{6}$$

Subsequently, $\tilde{\mathbf{F}}_i^V$ is decomposed into low-frequency components $\mathbf{F}_i^L$ and high-frequency components $\mathbf{F}_i^H$:

$$\mathbf{F}_i^H = [DWT_{lh}(\tilde{\mathbf{F}}_i^V), DWT_{hl}(\tilde{\mathbf{F}}_i^V), DWT_{hh}(\tilde{\mathbf{F}}_i^V)],$$
$$\mathbf{F}_i^L = DWT_{ll}\left(\tilde{\mathbf{F}}_i^V\right), \tag{7}$$

DWT_{xy} denotes the wavelet coefficients of different frequency bands, $x, y \in \{l, h\}$ represent low-frequency and high-frequency components, respectively. The low-frequency component $\mathbf{F}_i^L$ mainly encodes global structural information, while the high-frequency component $\mathbf{F}_i^H$ contains detailed information such as edge textures. We process these two features using linear projection to obtain $\{\mathbf{F}_i^L, \mathbf{F}_i^H\}$. The processed features are spatially reconstructed via inverse discrete wavelet transform:

$$\mathbf{F}_i^D = \text{IDWT}\left(\mathbf{F}_i^L + \mathbf{F}_i^H\right). \tag{8}$$

The final output result is obtained via residual connection:

$$\mathbf{F}_i^o = \mathbf{F}_i^V + \mathbf{F}_i^D. \tag{9}$$

3.5. Anomaly Map and Anomaly Score

Pixel-level anomaly map. Given an input image $X \in \mathbb{R}^{h\times w\times 3}$, the visual encoder extracts multiple patch-level features $F_i^V \in \mathbb{R}^{H\times W\times C}, i = 1, 2, \cdots L$ from L intermediate layers, where $H = h/patchsize$, $W = w/patchsize$. After optimization by the HFD module, $\mathbf{F}_i^o \in \mathbb{R}^{H\times W\times C}$ is obtained. The text encoder generates normal/abnormal text embeddings $\mathbf{F}_T = [\mathbf{F}_T^N, \mathbf{F}_T^A]$ based on predefined prompts. The anomaly map from the l-th layer is computed as:

$$M_i = Softmax(Up(\widetilde{\mathbf{F}}_i^o \widetilde{\mathbf{F}}_T)), i = 1, 2, \cdots L, \tag{10}$$

where $M_i = [M_i^a, M_i^n], M_i \in \mathbb{R}^{h\times w\times 2}$, $\widetilde{(\cdot)}$ denotes L_2 normalization along the embedding dimension, and $Up(\cdot)$ denotes upsampling.

Image-level anomaly score. For anomaly classification, we combine global and local features for similarity alignment with the text features. Specifically, the global feature X_c of the category token is first extracted. Meanwhile, the multi-scale local patch features $\mathbf{F}_i^o \in \mathbb{R}^{H\times W\times C}$ are concatenated along the channel dimension, then aggregated into a representation X_p via spatial compression and linear projection:

$$\mathbf{F}_{\text{cat}} = \text{Concat}\left(\mathbf{F}_1^o, \cdots, \mathbf{F}_L^o\right), \tag{11}$$
$$X_p = \text{Linear}\left(\text{GAP}(\mathbf{F}_{\text{cat}})\right), \tag{12}$$

where Concat denotes concatenate, GAP denotes global average pooling over the spatial dimensions. The local representation X_p is then fused with the global feature X_c to obtain the final image representation $X = X_p + X_c$. The image-level anomaly score is computed by aligning X with the text features:

$$S_p = Softmax(\widetilde{X}\widetilde{\mathbf{F}}_T). \tag{13}$$

3.6. Loss Function

During training, we calculate the classification loss L_{cls} and segmentation loss L_{seg} to optimize the parameters:

$$L_{total} = L_{cls}(S_p, y) + \sum\nolimits_{l=1}^{L} L_{seg}^l. \tag{14}$$

The classification loss uses binary cross-entropy to compare the classification prediction S_p with the true label y. The segmentation loss L_{seg} combines the Dice loss and Focal loss, which are applied to the segmentation prediction M_i and the abnormal segmentation mask S_{gt}, respectively:

$$L_{seg} = Dice(M_l^a, S_{gt}) + Dice(M_l^n, I - S_{gt}) \tag{15}$$
$$+ Focal(M_l^a, S_{gt}). \tag{16}$$

In the inference process, the final result $\mathbf{M} \in \mathbb{R}^{h\times w}$ is obtained by calculating the average value of the anomaly map of the l layer patch features:

$$\mathbf{M} = \frac{1}{L}\sum\nolimits_{l=1}^{L}(M_l^a + I - M_l^n)/2, \tag{17}$$

where I represents the full-one matrix.

4. Experiments

4.1. Experiment Setup

Dataset. To evaluate the model's ZSAD performance, we conducted experiments on 13 real-world datasets from industrial and medical domains. Specifically, we utilized 6 commonly used industrial anomaly detection datasets, including MVTec-AD [2], VisA [44], BTAD [28], SDD2 [3], DAGM [39], and DTD-Synthetic [1]. In the medical field, we utilized 7 different datasets: HeadCT [36], BrainMRI [20], and Br35H [12] for brain tumor classification; Endo [14] and Kvasir [19] for colon polyp segmentation; ISIC [8] for skin cancer detection; and RESC [15] for retinal edema segmentation. More details about the datasets can be found in Appendix C.

Evaluation Metrics. Following [17, 26, 37, 40], we adopted the Area Under the Receiver Operating Characteristic Curve (AUROC) as the evaluation metric. By calculating the AUROC values at both the image and pixel levels, we comprehensively assessed the model's effectiveness in anomaly detection and localization.

Implementation Details. Following [4, 30, 43], we use the OpenCLIP and ViT-L/14 architectures as the backbone network, adjusting the input image to a size of 518×518. We selected layers 6, 12, 18, and 24 from the 24-layer image encoder to extract patch embeddings. The number of local prompts was set to 4, the prompt depth to 9. We trained our model using the Adam optimizer with a learning rate of $1.5e^{-4}$. Following AnomalyCLIP [43], we fine-tuned our model on the MVTec-AD dataset, whose object categories

Table 1. Image-level AUROC comparison of the ZSAD method. Best performance is marked in bold, and the second-best is shown with an underline. The results using the $*$ method are copied from the paper or inferred from the official weights.

Datasets	Category	$\|C\|$	WinCLIP*	AnomalyCLIP*	AdaCLIP	TPS	BayesPFL*	Ours
			CVPR 23	ICLR 24	ECCV 24	AAAI 25	CVPR 25	-
MVTec-AD	Obj & texture	15	91.8	91.5	91.1	90.1	<u>92.3</u>	**94.2**
VisA	Obj	12	78.1	82.1	84.5	83.3	87.0	**87.9**
BTAD	Obj	3	68.2	88.3	<u>90.5</u>	88.1	<u>90.5</u>	**92.6**
SDD2	Obj	1	94.3	94.1	96.0	94.7	<u>97.4</u>	**97.9**
DAGM	Texture	10	91.8	97.5	97.1	97.0	<u>98.9</u>	**99.3**
DTD-Synthetic	Texture	12	93.2	93.5	94.7	93.4	<u>95.1</u>	**96.8**
HeadCT	Brain	1	81.8	93.4	<u>94.0</u>	92.3	88.2	**95.0**
BrainMRI	Brain	1	86.6	90.3	<u>96.0</u>	92.4	91.2	**96.2**
Br35H	Brain	1	80.5	94.6	95.6	<u>96.2</u>	94.0	**98.1**

Table 2. Pixel-level AUROC comparison of the ZSAD method. Best performance is marked in bold, and the second-best is shown with an underline. The results using the $*$ method are copied from the paper or inferred from the official weights.

Datasets	Category	$\|C\|$	WinCLIP*	AnomalyCLIP*	AdaCLIP	TPS	BayesPFL*	Ours
			CVPR 23	ICLR 24	ECCV 24	AAAI 25	CVPR 25	-
MVTec-AD	Obj & texture	15	85.1	<u>91.1</u>	89.4	88.5	**91.8**	90.4
VisA	Obj	12	79.6	95.5	<u>95.6</u>	95.5	<u>95.6</u>	**95.7**
BTAD	Obj	3	72.7	**94.2**	91.5	**94.2**	92.3	<u>93.4</u>
SDD2	Obj	1	94.5	<u>98.5</u>	97.4	96.7	97.6	**98.8**
DAGM	Texture	10	87.6	<u>95.6</u>	95.1	93.8	95.1	**96.8**
DTD-Synthetic	Texture	12	83.9	97.9	<u>98.5</u>	97.9	98.3	**98.9**
ISIC	Skin	1	83.3	89.7	89.3	<u>91.2</u>	89.5	**92.1**
Kvasir	Colon	1	69.7	78.9	<u>81.6</u>	78.8	<u>81.6</u>	**81.8**
Endo	Colon	1	68.2	84.1	<u>85.4</u>	84.7	85.0	**86.0**
RESC	Retina	1	90.9	<u>94.3</u>	94.0	94.2	93.9	**95.7**

do not overlap with those of the remaining datasets. For the results on MVTec-AD, we follow prior work and fine-tune the model on the VisA dataset. All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU. More implementation details can be found in Appendix A.2.

4.2. Comparison with SOTA Methods

This paper compares our DLVP-CLIP with five state-of-the-art methods, including WinCLIP [18], AnomalyCLIP [43], AdaCLIP [4], TPS [25], and Bayes-PFL [31]. To ensure fair comparison, we retrained models originally trained on different datasets to match the dataset configuration of other methods. For more details on these comparison methods, see Appendix A.1.

Zero-shot anomaly classification. As shown in Table 1, our method achieves significant and consistent advantages in ZSAD across both industrial and medical domains. Specifically, on the texture-based datasets DAGM and DTD-Synthetic, it achieves AUROC scores of 99.3% and 96.8%, respectively, significantly outperforming other methods and showcasing its robust detection capability for complex texture anomalies. In the medical field, our method demonstrates significant advantages on three brain datasets, with an AUROC of 98.1% on the Br35H dataset, indicating its high sensitivity in detecting subtle structural anomalies in medical images.

For WinCLIP, it can maintain basic performance in simple scenarios, but in complex backgrounds or low-contrast medical images (such as HeadCT 81.8%), it reveals the deficiency of its semantic alignment capability. AnomalyCLIP introduced object-agnostic prompts, simplifying the prompt generation process. However, its static prompt design is difficult to adapt to cross-domain scenarios, resulting in slightly inferior performance in medical datasets. AdaCLIP dynamically provides hints based on global information, while ignoring local perception. Our approach achieves this by using dynamic local cues to guide the coarse-grained global features of CLIP towards local abnormal patterns, thereby enabling the unified modeling of global semantics and fine-grained anomalies. As a result, it has achieved excellent performance in both industrial quality inspection and medical diagnosis.

Zero-shot anomaly segmentation. As shown in Table 2, we compared the pixel-level anomaly segmentation performance on seven industrial datasets and four medical datasets. The experimental results indicate that our method

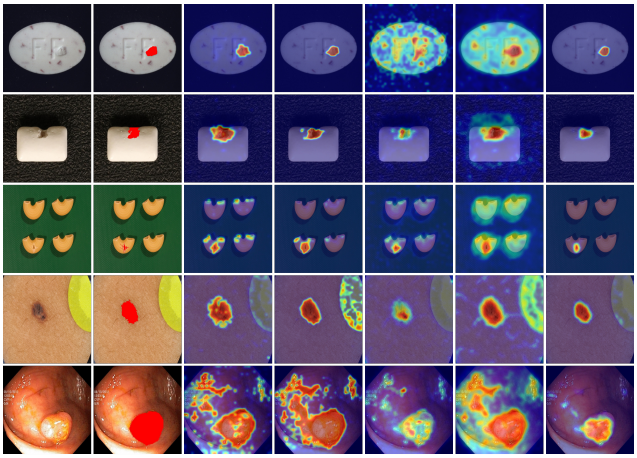


Figure 3. Visualization of anomaly localization results of AnomalyCLIP, AdaCLIP, TPS, Bayes-PFL and our method.

Table 3. Ablation Results of DLVP. S and M respectively represent SLFS and MLoP. I-AUROC and P-AUROC represent Image-level and Pixel-level AUROC, respectively.

S	M	MVTec-AD		VisA	
		I-AUROC	P-AUROC	I-AUROC	P-AUROC
✗	✗	89.5	89.0	84.2	95.0
✔	✗	91.0	89.3	86.3	95.2
✗	✔	90.4	89.6	85.1	95.1
✔	✔	**94.2**	**90.4**	**87.9**	**95.7**

demonstrates comprehensive competitiveness in both industrial and medical fields, particularly in scenarios involving texture defects and medical lesions. Specifically, while TPS performs well on some tasks, its performance declines in texture detection tasks (e.g., DAGM, DTD-Synthetic) and medical dataset tasks. AdaCLIP and AnomalyCLIP perform adequately in industrial texture detection but show a significant decline in performance for medical image anomaly segmentation tasks. In contrast, our method performs outstandingly in industrial texture and medical lesion tasks, verifying its generalization advantage. Figure 3 shows the anomaly maps for industrial and medical datasets. Compared to other methods, our DLVP-CLIP model demonstrates superior performance in segmentation accuracy and the completeness of anomaly region localization.

4.3. Ablation

Influence of DLVP. Table 3 shows the impact of DLVP on model performance. When both the SLFS and MLoP modules are removed, the model achieves an I-AUROC of 89.5% and a P-AUROC of 89.0% on the MVTec-AD dataset, while obtaining an I-AUROC of 84.2% and a P-AUROC of 95.0% on the VisA dataset. These results indicate that, without DLVP, the model has limited ability to capture fine-grained anomalous patterns and distinguish subtle local deviations. Adding either module improves performance. Notably, the combination of both modules achieves optimal performance across all metrics. This is because the local semantic information provided by SLFS lays the foundation for cross-modal alignment in MLoP, while MLoP dynamically couples local features with the text space through its prompting mechanism, enabling the model to more accurately distinguish local differences between normal and abnormal samples. As shown in Figure 4, the visualization results demonstrate that the model tends to overlook subtle anomaly regions when DLVP is removed. In contrast, the anomaly boundaries are clearly captured when they are used.

Table 4. Ablation Results of HFD. I-AUROC and P-AUROC represent Image-level and Pixel-level AUROC, respectively.

HFD	MVTec-AD		VisA	
	I-AUROC	P-AUROC	I-AUROC	P-AUROC
✗	92.8	89.5	85.8	95.3
✔	**94.2**	**90.4**	**87.9**	**95.7**

Table 5. Validity ablation for SLFS and HFD on the VisA dataset.

(a) Effect of SLFS.

Setting	I/P-AUROC
Global	85.6, 95.6
SLFS	**87.9**, **95.7**

(b) Effect of HFD.

Setting	I/P-AUROC
Sobel	86.7, 95.2
HFD	**87.9**, **95.7**

Table 6. Ablation results of different Top-K values. I-AUROC and P-AUROC represent image-level and pixel-level AUROC, respectively.

Num.	MVTec-AD		VisA	
	I-AUROC	P-AUROC	I-AUROC	P-AUROC
1	92.1	88.5	84.5	95.3
2	92.4	88.9	85.7	95.3
3	92.9	90.1	86.7	95.5
4	**94.2**	**90.4**	**87.9**	**95.7**
5	92.7	89.7	85.7	**95.7**

Influence of HFD. We conducted an ablation experiment on HFD, and the results are shown in Table 4. The experimental results indicate that introducing HFD improves ZSAD performance on MVTec-AD and VisA. This

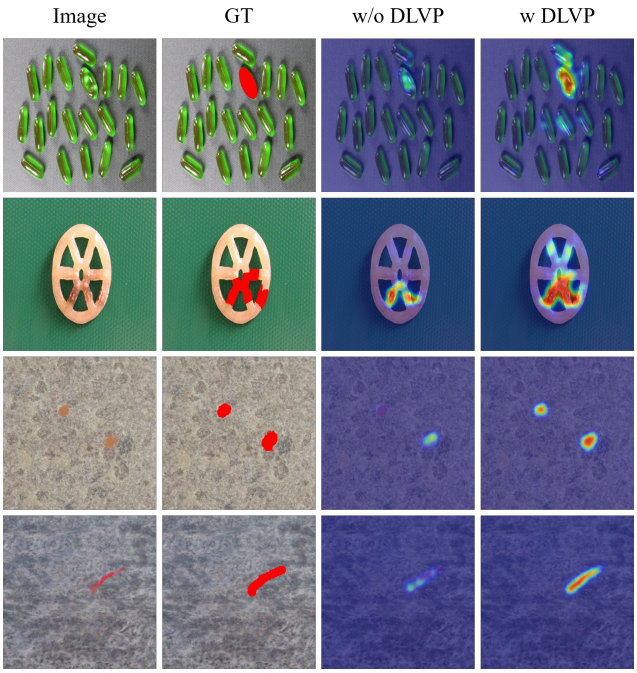


Figure 4. Visualization comparison of anomaly detection results with and without DLVP. After introducing DLVP, the model produces more precise anomaly localization results.

module decomposes features into low-frequency and high-frequency components via DWT, where low-frequency components suppress noise and high-frequency components enhance the perception of minor anomalies, effectively balancing global and local features.

As shown in Table 5, we further analyzed the effectiveness of SLFS and HFD. Compared to using a global token, SLFS focuses on semantically important local regions, capturing fine-grained anomaly clues that global aggregation often overlooks. Furthermore, although the Sobel filter extracts high-frequency gradients, it captures only limited edge-based responses. In contrast, HFD performs structured multi-band frequency decomposition via DWT, enabling richer and more diverse high-frequency representations for subtle anomaly patterns.

Influence of Top-K. We conducted ablation analysis on the number of Top-K in the SLFS module, as shown in Table 6. When K = 1, only the single highest attention region is selected, which may overlook critical local details. As K increases to 4, the model gradually covers more local regions strongly associated with anomalies, effectively supplementing the missing details in global features, achieving optimal performance. However, when K = 5, although more regions are covered, some regions with low attention weights may contain background noise or normal texture variations unrelated to anomalies, resulting in a performance decline. More ablation studies on the hyperparameters can be found in Appendix B.

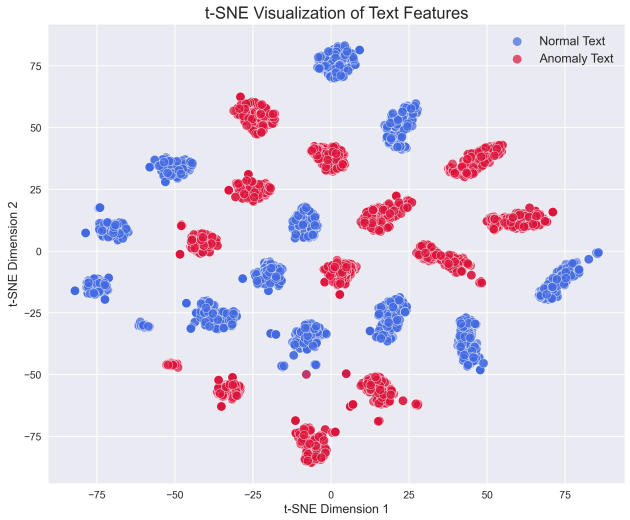


Figure 5. t-SNE visualization results of text features on the VisA dataset.

We visualize the text-prompt feature embeddings using t-SNE to assess the effectiveness of DLVP. As shown in Figure 5, the normal (blue) and abnormal (red) samples form clearly separable clusters, indicating that DLVP injects discriminative local visual cues into the prompt space, enabling the model to capture fine-grained semantic differences between normal and anomalous samples.

5. Conclusion

This paper proposes DLVP-CLIP, a ZSAD method based on dynamic local visual prompts, to address the inherent shortcomings of CLIP in fine-grained anomaly localization. This method utilizes Semantic-Aware Local Feature Selector (SLFS) to extract visual features from key regions as dynamic prompt markers and employs multi-modal local prompt (MLoP) to achieve visual-text joint encoding, integrating fine-grained visual clues into the cross-modal alignment process. Furthermore, the high-low frequency decomposition (HFD) module based on wavelet transformation is introduced to distinguish high-frequency textures from low-frequency structural components, enabling the collaborative representation of local details and global semantics. Extensive experimental results on 13 diverse public datasets, including industrial and medical images, demonstrate that our method achieves superior performance in the ZSAD task.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 62431004.

References

[1] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5564–5572, 2023.

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[3] Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129: 103459, 2021.

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

[5] Linwei Chen, Lin Gu, and Ying Fu. Frequency-dynamic attention modulation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22620–22632, 2025.

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

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

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

[9] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37, 2021.

[10] Tao Gong, Qi Chu, Bin Liu, Wei Zhou, and Nenghai Yu. Fe-clip: Frequency enhanced clip model for zero-shot anomaly detection and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21220–21230, 2025.

[11] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 98–107, 2022.

[12] A. Hamada. Br35h: Brain tumor detection 2020. *Online. Available: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection*, 2020.

[13] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8472–8480, 2024.

[14] Steven A Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo L Hammer, and Michael A Riegler. The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In *International Conference on Pattern Recognition*, pages 263–274. Springer, 2021.

[15] Junjie Hu, Yuanyuan Chen, and Zhang Yi. Automated segmentation of macular edema in oct using deep neural networks. *Medical image analysis*, 55:216–227, 2019.

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

[17] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11375–11385, 2024.

[18] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023.

[19] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019.

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

[21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023.

[22] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023.

[23] Daehyun Kim, Sungyong Baik, and Tae Hyun Kim. Sanflow: Semantic-aware normalizing flow for anomaly detection. *Advances in neural information processing systems*, 36: 75434–75454, 2023.

[24] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection

and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20402–20411, 2023.

[25] Jitao Ma, Weiying Xie, Hangyu Ye, Daixun Li, and Leyuan Fang. Aligning and prompting anything for zero-shot generalized anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5964–5972, 2025.

[26] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4744–4754, 2025.

[27] Dwarikanath Mahapatra, Behzad Bozorgtabar, and Zongyuan Ge. Medical image classification using generalized zero shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3344–3353, 2021.

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

[29] Zhen Qu, Xian Tao, Fei Shen, Zhengtao Zhang, and Tao Li. Investigating shift equivalence of convolutional neural networks in industrial defect segmentation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–17, 2023.

[30] Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In *European Conference on Computer Vision*, pages 301–317. Springer, 2024.

[31] Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30398–30408, 2025.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[33] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022.

[34] Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2155–2162, 2023.

[35] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. *arXiv preprint arXiv:2306.01195*, 2023.

[36] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021.

[37] Fenfang Tao, Guo-Sen Xie, Fang Zhao, and Xiangbo Shu. Kernel-aware graph prompt learning for few-shot anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7347–7355, 2025.

[38] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5749–5757, 2024.

[39] Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, page 11, 2007.

[40] Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15204–15213, 2025.

[41] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[42] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[43] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2024.

[44] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation. In *European conference on computer vision*, pages 392–408. Springer, 2022.