Exploring High-order-aware Prompt Learning for Zero-shot Anomaly Detection

Shun Wei[1], **Jielin Jiang**[1,2,3*], **Xiaolong Xu**[1,2,3]

[1]School of Software, Nanjing University of Information Science and Technology, Nanjing, China
[2]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[3]Jiangsu Province Engineering Research Center of Advanced Computing and Intelligent Services, Nanjing, China
pangdatangtt@nuist.edu.cn, jiangjielin2008@nuist.edu.cn, xlxu@nuist.edu.cn

Abstract

Many methods have demonstrated promising results in zero-shot anomaly detection (ZSAD) by incorporating prompt learning (PL) to fine-tune Vision-Language Models. However, the prompt learners proposed in recent studies remain relatively simple, such as learnable textual and visual prompts. Relying solely on the current PL paradigm restricts the ability to generate more precise prompts, thereby hindering improved ZSAD performance. To mitigate this issue, this paper proposes a high-order-aware prompt learning framework, termed HiPL, which facilitates the detection of unseen anomalies through generating prompts fortified by hypergraphs. Specifically, HiPL models high-order correlations among patches through a dynamically constructed hypergraph structure. Then we leverage a hypergraph semantic convolution to capture potential collaborative information by propagating high-order correlations by hyperedges. Meanwhile, HiPL introduces a Mixture-of-Experts prompt learner (MoEPLer), where the experts within MoEPLer can generate multiple distinct prompts based on the modeled high-order correlations. Then, the final high-order-aware textual prompts can be formed by synthetically considering each expert's prompt by gating weights. This enables a comprehensive understanding of potential anomalous patterns, thereby facilitating ZSAD performance. Large-scale experiments conducted on 12 datasets, spanning natural, industrial, and medical domains, demonstrate the validity of proposed HiPL.

Introduction

Anomaly detection (AD), as a critical link of deep learning, has been extensively applied in various fields, including industrial defect inspection (Wang et al. 2023; Rudolph et al. 2023; Jiang et al. 2024; Wang, Peng, and Fu 2024) and medical imaging analysis (Guo et al. 2023; Chai et al. 2024; Rahman, Munir, and Marculescu 2024; Wei, Jiang, and Xu 2025). Due to the rarity of anomalous samples, conventional studies have a tendency to follow a one-model-one-category paradigm and train the models in an unsupervised manner. However, maintaining this tendency has become increasingly challenging as factors such as the emergence of new products or the need to protect commercial secrets, which restrict the availability of sufficient normal samples for training. Fortunately, recent advancements have introduced the zero-shot learning paradigm as a promising solution to this challenge.

With the rise of prompt learning, Vision-Language Models (VLMs) like CLIP (Radford et al. 2021) can be fine-tuned to adapt to different downstream tasks (Radford et al. 2021; Rao et al. 2022; Zhou et al. 2022b,a; Zhu et al. 2023), demonstrating exceptional generalization capacity. As a result, numerous CLIP-based methods have been proposed for zero-shot anomaly detection (ZSAD), achieving impressive performance across various domains and even surpassing several existing full-shot AD methods. Following earlier CLIP-based image classification methods (Radford et al. 2021), WinCLIP (Jeong et al. 2023) initially adopts predefined prompt templates for few/zero-shot AD, as illustrated in Fig. 1(a). However, subsequent works have highlighted that such rigid templates make it challenging for models to capture fine-grained details within images, such as shapes and colors. To address this issue, recent works (Gu et al. 2024; Li et al. 2024; Zhou et al. 2024b) adopt learnable word embeddings to acquire textual prompts in a static way (see Fig. 1(b)), while others (Zhou et al. 2022a; Zha et al. 2023; Cao et al. 2024) explore dynamic prompt learning via visual prompts (see Fig. 1(c)). Despite these advancements, both prevailing paradigms remain suboptimal. A key reason is that most CLIP-based frameworks rely on Transformer architectures, which, due to their self-attention mechanism, capture pairwise correlations among all patches (see Fig. 2(a)). This often introduces redundant information (Chai et al. 2024), indirectly affecting prompt learning, so the quality of prompts may degrade.

Recently, hypergraphs have demonstrated significant potential in visual tasks. Unlike Transformers, hypergraphs can capture complex high-order correlations, as shown in Fig. 2(b). Motivated by hypergraphs, this paper investigates *how to generate prompts with high grade for accurate ZSAD by integrating high-order correlations to reduce redundancy.*

To fulfill the goal, this paper proposes a novel prompt learning framework, high-order-aware prompt learning (HiPL), as shown in Fig. 1(d). Specifically, there are two crucial steps: *dynamic hypergraph construction* and *high-order-aware prompt generation*. 1) Unlike conventional hypergraph-based methods that assign the fixed number of vertices (*e.g.*, vertices=3 with each vertex seen as a patch)

*Corresponding Author.

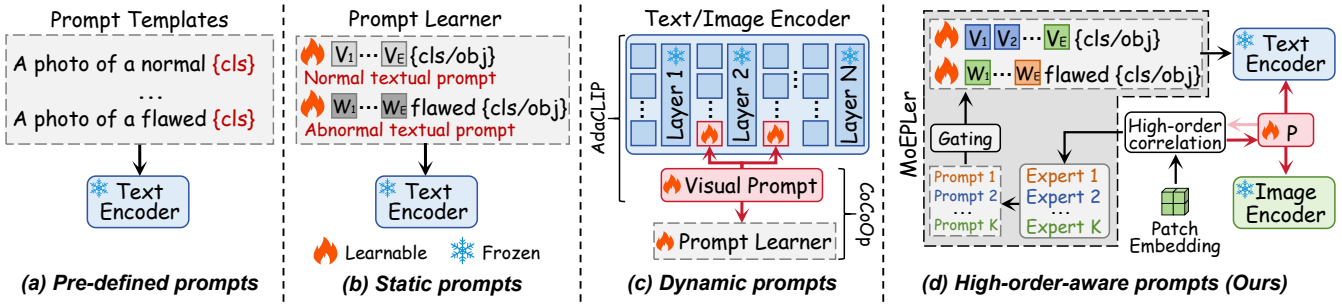

Figure 1: Comparison of different prompt learning paradigms. Existing three common prompt learning methods: (a), (b), and (c). Our high-order-aware prompt learning framework: (d).

for a hypergraph construction (see Fig. 2(b)), our aim is to dynamically construct the hypergraph structure based on the importance of different patches (see Fig. 2(c)). Because some patches include significant semantic information, while others do not, *e.g.*, backgrounds. Thus, we first measure the semantic similarity of different patches. Based on this, we dynamically construct a hypergraph structure by generating different number of hyperedges for each vertex. After this, a hypergraph semantic convolution is applied on hypergraph to capture potential collaborative information and facilitate message passing. 2) HiPL further introduces Mixture-of-Experts prompt learner (MoEPLer), which consists of multiple experts. With modeled high-order correlations, MoEPLer empowers experts to generate different prompts, as visualized in Fig. 1(d). At the same time, MoEPLer also integrates a cross-modal gating network, which determines the final high-order-aware prompt generation by comprehensively considering each expert's initial prompt through vision-text interaction weights. This helps more effectively understand generic normal and abnormal patterns. Besides textual prompts, HiPL also introduces high-order-aware visual prompts generated by a projection layer to text/image encoder for more effective modeling capabilities. In summary, the main contributions are as follows:

- To our best knowledge, we are the first to extend VLMs with hypergraphs to explore high-order-aware prompts, proposing a novel prompt learning framework, dubbed HiPL, for enhanced ZSAD performance.
- Proposing a novel dynamic hypergraph construction method ensures that high-order correlations among crucial patches within images can be modeled. Based on this, high-order-aware textual and visual prompts are formed via MoEPLer and a projection layer, respectively.
- Large-scale experiments on multiple datasets, covering industrial, medical, and natural domains, demonstrate the effectiveness of proposed HiPL.

Related Work

Zero-shot Anomaly Detection. Unlike traditional AD methods (Guo et al. 2023; Rudolph et al. 2023; Jiang et al. 2025), ZSAD methods only train the model on one auxiliary dataset and then can test it on any other dataset without fine-tuning. WinCLIP (Jeong et al. 2023), for instance, which is the pioneering ZSAD approach, employs off-the-shelf VLMs with manual textual prompts to detect anomalies. Followed by this, AnomalyCLIP (Zhou et al. 2024a) proposes using learnable object-agnostic textual prompts to capture generic anomaly semantics, while AnomalyGPT (Gu et al. 2024) focuses on learnable class-specific textual prompts but for unsupervised AD. In addition to textual prompts, AdaCLIP (Cao et al. 2024) further achieves enhanced ZSAD performance by introducing visual prompts to encoders. More recently, without relying on VLMs, some methods like INP-Former (Luo et al. 2025) also achieve impressive results, proving non-VLMs methods' potential in ZSAD field.

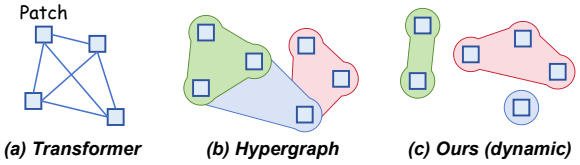

Figure 2: Comparison of the ability of different methods to capture correlations among patches. (a) Pairwise correlations; (b) High-order correlation modeling; (c) Dynamic high-order correlation modeling.

Prompt Learning. Compared to fine-tune the entire framework for different tasks, prompt learning focuses on optimizing the model to adapt to specific tasks based on input prompts. CoOp (Zhou et al. 2022b) is the first to apply prompt learning to the visual domain. PromptAD (Li et al. 2024) combines normal prompts with anomalous suffixes to generate abnormal prompts for few-shot AD. AnomalyCLIP (Zhou et al. 2024a) introduces object-agnostic prompt learning to understand the general concepts of normality and abnormality. However, these methods typically fall under static prompt learning paradigm, which may cause distribution shifts (Zhou et al. 2022a; Cao et al. 2024). In contrast, CoCoOp (Zhou et al. 2022a) and AdaCLIP (Cao et al. 2024) introduce dynamic prompts to enhance modeling ability.

Hypergraph Learning. Unlike traditional graph structures, hypergraphs can capture not only binary relationships

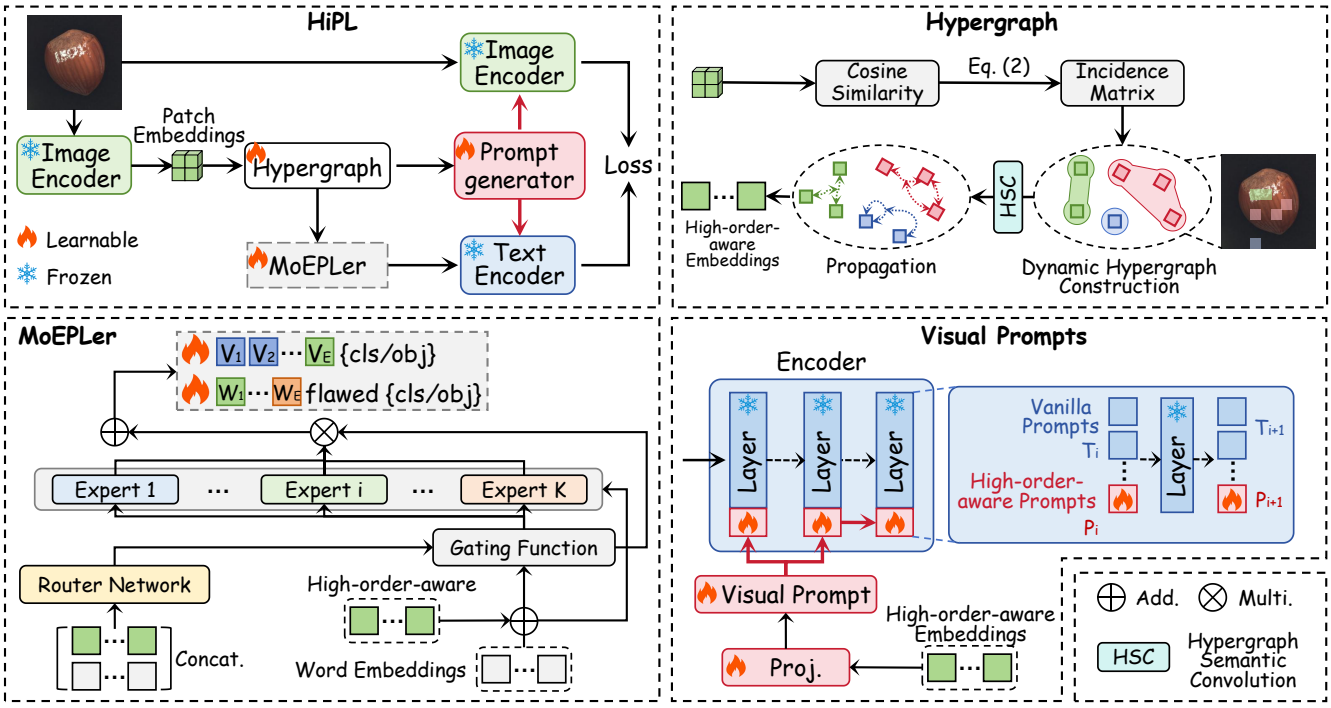

Figure 3: Framework of proposed HiPL. There are two key steps: *dynamic hypergraph construction* and *high-order-aware prompt generation*. 1) HiPL models high-order correlations among patches by a dynamically constructed hypergraph. 2) With high-order correlations, HiPL acquires textual prompts by MoEPLer and obtains visual prompts via a projection layer (Proj.).

between vertices but also high-order correlations among multiple vertices through their hyperedges. This capability makes them valuable for applications across different fields, including recommendation systems (La Gatta et al. 2022; Khan et al. 2025), social networks (Yang et al. 2020; Meng and Motevalli 2024), and data mining (Jin, Wang, and Zhang 2019). Recently, hypergraph learning methods have gradually been used for visual tasks. Hyper-YOLO (Feng et al. 2025) utilizes high-order information to improve visual backbones for object detection, while Chai et al. (Chai et al. 2024) proposed an adaptive hypergraph neural network for medical image segmentation, showing their effectiveness in capturing complex relationships within images. For the first time, through combining hypergraphs with VLMs, this paper explores high-order-aware prompt learning for ZSAD.

Preliminary: Hypergraphs

Unlike ordinary graphs, a hypergraph comprises hyperedges, each connecting multiple vertices, which helps establish high-order correlations. Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{W}\}$ be a hypergraph, where the vertex set and the hyperedge set are respectively marked as $\mathcal{V}$ and $\mathcal{E}$, and $\mathcal{W} \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{E}|}$ represents a diagonal matrix, with $\mathcal{W}_{i,i}$ representing the weight of the i^{th} hyperedge. For convenience, the hypergraph can be denoted by an incidence matrix $\mathbf{H} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{E}|}$, where $\mathbf{H}(v,e) = 1$ if the hyperedge e includes the vertex $v \in \mathcal{V}$, and $\mathbf{H}(v,e) = 0$ otherwise. Let vertex and hyperedge matrices respectively be $\mathbf{D}_v$ and $\mathbf{D}_e$, where the corresponding i^{th} diagonal element in these two matrices is $\sum_{i=1}^{|\mathcal{V}|} \mathbf{H}_{ie}$ and $\sum_{e=1}^{|\mathcal{E}|} \mathbf{H}_{ie}$. In this paper, we consider patches as a set of vertices and construct the hypergraph structure based on them.

Methodology

Overall Framework

Fig. 3 illustrates the proposed prompt learning framework, HiPL, which contains two key innovations: *dynamic hypergraph construction* and *high-order-aware prompt generation*. In the first step, a hypergraph structure is first dynamically constructed through generating varying hyperedges for each vertex based on the semantic similarity among patches. Then, a hypergraph semantic convolution (HSC) is introduced to further propagate high-order correlations among patches with homogeneous semantics. In the second step, with such high-order correlations, HiPL introduces MoEPLer, which enables multiple experts to generate textual prompts. The final high-order-aware textual prompts are then generated by comprehensively aggregating the prompt from each expert via vision-text interaction weights. Besides, visual prompts for the encoders are obtained by projecting the high-order-aware embeddings.

Dynamic Hypergraph Construction

Conventional hypergraph-based methods (Feng et al. 2019; Peng et al. 2022) construct a hypergraph structure by using a fixed number of vertices. In contrast, we propose to construct the hypergraph based on the significance of patches in

a dynamic way, as larger regions containing homogeneous semantic information need to be assigned more neighbors.

Formally, given the patch embeddings $F_P \in \mathbb{R}^{N\times D}$ after transforming and removing $[cls]$ token, which can be viewed as a graph with N vertices with D features at each vertex, where N and D are respectively the number of patches and feature dimension. Notably, the $[cls]$ token is not involved in hypergraph construction, as it represents global information. Here, we consider a vertex as a patch. The detailed steps of dynamic hypergraph construction are as follows:

- To reduce computational overhead caused by cluster-based algorithms and generate different number of hyperedges for each vertex based on its importance, we first measure the semantic similarities among patches through cosine similarity:

$$S = \frac{F_P \cdot (F_P)^\top}{\|F_P\| \cdot \|F_P\|} \in \mathbb{R}^{N\times N}, \tag{1}$$

  where S is a similarity matrix.
- Inspired by (Chai et al. 2024) that determines the number of hyperedges by introducing a scale factor that controls vertex degree matrix $\mathbf{D}_v$, we propose to dynamically assign heperedges by an incidence matrix generated by a sigmoid function:

$$I = Sigmoid(\alpha(S + 10^{-4})), \tag{2}$$

  where I represents the incidence matrix and α is a large number that avoids values approaching 0. Then, the patch can include a hyperedge to connect another patch with high similarity based on $I' = [I > 0.5]$.

Therefore, our dynamic hypergraph is defined as $\mathcal{G}_d = \{\mathcal{V}_d, \mathcal{E}_d\}$. The vertex and hyperedge sets are represented as $\mathcal{V}_d = \{F_P^i \in \mathbb{R}^D\}_{i=1}^N$ and $\mathcal{E}_d = \{\mathcal{E}_k(v)|v \in \mathcal{V}_d\}$, where $\mathcal{E}_k(v)$ denotes the set of the k highest-similarity neighbors of vertex v.

After dynamic hypergraph construction, the HSC is designed to capture potential collaborative information and facilitate high-order semantic correlations. Motivated by (Feng et al. 2019), the HSC is mathematically defined as:

$$\tilde{F}_P = a(F_P + \sigma(\mathbf{D}_v^{-\frac{1}{2}}\mathbf{H}\mathcal{W}\mathbf{D}_e^{-1}\mathbf{H}^\intercal\mathbf{D}_v^{-\frac{1}{2}}F_P\Theta)) + b, \tag{3}$$

where a and b are a learnable weight matrix and a bias vector, respectively. $\Theta \in \mathbb{R}^{d\times d}$ is the learnable weight for HSC layer. σ is the activation function. After HSC, the high-order-aware patch embeddings $\tilde{F}_P$ can be generated. Then, we obtain the global high-order-aware embeddings for prompt learning, which can be expressed as follow:

$$F_P^g = Linear(\frac{1}{N}\sum_{i=1}^{N}\tilde{F}_P(i,:)), \tag{4}$$

where $Linear(\cdot)$ represents a linear layer.

High-order-aware Prompt Generation

Textual prompt generation. High-order-aware patch embeddings are generated through the constructed hypergraph, which aids in learning generic semantic information (*i.e.*, semantics associated with seen categories may also be similar to those of unseen categories (Zhou et al. 2024a; Cao et al. 2024)). Then, the next goal is to generate high-quality textual prompts for improved ZSAD performance. To enhance flexibility, most of existing prompt learning methods (Zhou et al. 2024a; Li et al. 2024; Zhou et al. 2024b) optimize textual prompts in a static manner:

$$t_n = [V_1]\cdots[V_L][class/object],$$
$$t_a = [W_1]\cdots[W_L][flawed][class/object],$$

where t_n and t_a respectively represent learnable textual templates for normality and abnormality, V and W denote learnable word embeddings, and L is the length of word embeddings. $class$ and $object$ are used in class-specific (Jeong et al. 2023; Cao et al. 2024) and object-agnostic textual templates (Zhou et al. 2024a,b), respectively. However, building upon prior studies, it is found that the key to ZSAD lies not only in capturing generic normal and abnormal semantic representations (Zhou et al. 2024a) but also in generating more precise prompts based on input visual patterns to avoid static prompt learning (Li et al. 2024; Zhou et al. 2024a). Thus, we expect that elaborate and dynamic textual prompts are generated based on input high-order-aware embeddings.

To achieve it, inspired by the strong capability of Mixture-of-Experts (MoE) (Shazeer et al. 2017; Meng et al. 2024) to handling multiple tasks, we employ MoEPLer to generate prompts with high grade. Specifically, MoEPLer consists of K experts and a cross-modal gate function which determines the final prompt generation. Based on input visual patterns, our dynamic textual prompts can be obtained by modifying prior templates:

$$t_n = \sum_{i=1}^{K} G(concat(F_P^g, [V_1]\cdots[V_L]))$$
$$E_i(F_P^g + ([V_1]\cdots[V_L]))[class/object],$$
$$t_a = \sum_{i=1}^{K} G(concat(F_P^g, [W_1]\cdots[W_L]))$$
$$E_i(F_P^g + ([W_1]\cdots[W_L]))[flawed][class/object].$$

Here, E_i represents the i^{th} expert, where the architecture for each expert is different, unlike the uniform architectures used in prior MoE-based methods. To maintain efficiency, the architecture for each expert is simple, *e.g.*, convolutional layers, MLPs, and attention mechanisms. $G(\cdot)$ is the output of cross-modal gating function G, which is defined as:

$$G(x) = \frac{exp(R(x))}{\sum_{i=1}^{K}\exp(R_i(x))}, \tag{5}$$

where $R(x)$, the matching score, is the output of our router network R, which ensures that learnable word embeddings interact with high-order-aware embeddings. R is a nonlinear transformation:

$$R(x) = max(0, a_1x + b_1)a_2 + b_2, \tag{6}$$

where a_1 and a_2 are learnable weight matrices, and b_1 and b_2 are bias vectors. Based on the cross-modal weights generated by G, we comprehensively consider each expert's initial prompt to form the final high-order-aware prompts.

2D Task	Datasets	w/o HiPL				w/ HiPL	
		WinCLIP CVPR'23	AnomalyCLIP ICLR'24	AdaCLIP ECCV'24	MultiADS† ICCV'25	AnomalyCLIP ICLR'24	AdaCLIP ECCV'24
(I-ROC, AP)	MVTec	(91.5, 96.5)	(90.9, 96.0)	(87.5, 91.0)	(–, –)	(**93.5**, **97.0**)	(90.5, 93.3)
	VisA	(78.6, 81.3)	(81.5, 85.0)	(80.8, 84.0)	(82.5, 86.5)	(84.5, 87.0)	(**85.0**, **87.1**)
	BTAD	(68.8, 70.9)	(87.9, 87.3)	(88.0, 89.5)	(–. –)	(88.9, **91.8**)	(**89.0**, 91.1)
	MPDD	(63.3, 69.5)	(77.0, 82.0)	(76.0, 76.5)	(78.3, 78.4)	(**79.5**, **82.9**)	(78.4, 80.3)
(P-ROC, PRO)	MVTec	(85.1, 64.6)	(91.0, 81.4)	(88.0, 79.2)	(89.1, –)	(**91.8**,**83.0**)	(89.9, 81.1)
	VisA	(79.6, 56.8)	(95.0, 86.7)	(95.0, 85.3)	(95.0, **89.7**)	(**95.9**,88.0)	(95.5, 86.4)
	BTAD	(73.0, 30.5)	(94.2, 74.8)	(92.1, 72.9)	(–, –)	(**94.8**, **79.5**)	(93.2, 74.0)
	MPDD	(75.9, 48.0)	(96.0, 87.5)	(96.3, 88.0)	(95.8, **89.7**)	(96.5, 88.9)	(**96.7**, 89.5)
3D Task	Datasets	CLIP ICML'21	PointCLIP ICCV'23	AnomalyCLIP ICLR'24	PointAD NIPS'24	AnomalyCLIP ICLR'24	PointAD NIPS'24
(I-ROC, AP)	MVT 3D	(60.1, 84.0)	(78.1, 53.4)	(52.4, 81.4)	(77.3, 93.0)	(55.3, 85.1)	(**82.0**, **95.3**)
	Eyecandies	(65.6, 67.2)	(48.0, 50.5)	(58.0, 59.5)	(68.0, 72.1)	(59.1, 61.2)	(**72.1**, **75.3**)
	Real3D	(65.7, 70.0)	(57.0, 59.2)	(54.1, 56.3)	(74.3, 76.4)	(57.2, 59.9)	(**77.0**, **78.9**)
(P-ROC, PRO)	MVT 3D	(79.5, 53.0)	(87.5, 53.8)	(88.5, 61.2)	(95.2, 84.4)	(89.5, 63.9)	(**96.0**, **85.5**)
	Eyecandies	(80.1, 35.6)	(46.4, –)	(77.7, 46.2)	(91.6, 70.8)	(78.5, 50.3)	(**92.0**, **72.8**)
	Real3D	(43.2, –)	(50.9, –)	(49.5, –)	(72.9, –)	(55.5, –)	(**74.1**, –)

Table 1: Performance comparison of ZSAD without and with HiPL across different industrial datasets. The best result is highlighted. I-ROC: image-level AUROC and P-ROC: pixel-level AUROC. We reproduced these results using their publicly available code. †: Results are taken from original paper (Sadikaj et al. 2025).

Task	Datasets	w/o HiPL				w/ HiPL	
		WinCLIP CVPR'23	MVFA-AD CVPR'24	AnomalyCLIP ICLR'24	AdaCLIP ECCV'24	AnomalyCLIP ICLR'24	AdaCLIP ECCV'24
(P-ROC, PRO)	Kvasir	(69.7, 24.5)	(**81.9**, –)	(78.5, 45.0)	(77.7, 43.7)	(80.0, **47.4**)	(79.0, 45.2)
	ClinicDB	(51.2, 13.8)	(83.9, –)	(82.0, 67.1)	(84.4, 68.1)	(83.1, 68.9)	(**85.0**, **69.5**)
	ColonDB	(70.3, 32.5)	(78.4, –)	(80.9, 70.5)	(85.4, 69.3)	(81.6, **72.1**)	(**86.2**, 71.3)

Table 2: Performance comparison of ZSAD without and with HiPL across different medical datasets.

Visual prompt generation. High-order-aware visual prompts can be achieved in the similar way. Following (Cao et al. 2024), we introduce prompting layers to replace the layers in encoders of CLIP. Prompting layers concatenate dynamic visual prompt tokens to the original tokens $\mathbf{T}$.

Formally, the high-order-aware embeddings are projected into prompts $\mathbf{P}$ through a learnable linear layer. Let the i^{th} prompting layer be L_i^P, the process is:

$$[\mathbf{T}_{i+1}, \_] = L_i^P([\mathbf{T}_i, \mathbf{P}_i]), \tag{7}$$

$$[\mathbf{T}_{i+1}, \mathbf{P}_{i+1}] = L_i^P([\mathbf{T}_i, \mathbf{P}_i]), \tag{8}$$

Loss functions. Following prior works (Gu et al. 2024; Li et al. 2024), the commonly used dice loss and focal loss are used to optimize pixel-level anomaly maps, and the focal loss is used to optimize the image-level anomaly scores.

Experiments

Experimental Setup

Datasets and evaluation metrics. Extensive experiments were conducted to evaluate HiPL on 12 commonly available datasets, covering diverse industrial, medical, and natural domains. For industrial scenarios, MVTec AD (Bergmann et al. 2019), VisA, BTAD (Mishra et al. 2021), MVTec 3D (Bergmann et al. 2021), Eyecandies (Bonfiglioli et al. 2022), and Real3D-AD (Liu et al. 2023) were considered. In medical imaging analysis, experiments were assessed on Kvasir (Jha et al. 2020), CVC-ClinicDB (Bernal et al. 2015), and CVC-ColonDB (Tajbakhsh, Gurudu, and Liang 2015). Besides, two natural datasets (CIFAR10 and CIFAR100) were used for zero-shot image classification.

Following prior works (Jeong et al. 2023; Zhou et al. 2024a), two detection metrics (image-level Area Under the Receiver Operating Characteristic Curve (AUROC) and average precision (AP)) and two segmentation metrics (pixel-level AUROC and PRO (Bergmann et al. 2020)) were reported. For image classification, accuracy and Top-5 accuracy were considered.

Implementation details & baselines. The experiments were conducted on a system equipped with an NVIDIA GeForce RTX3090. The Adam and SGD optimizers were respectively used for MoEPLer and hypergraph structure, with a uniform learning rate of 0.001. The length of learnable word embeddings L was set to 12. Hyper-parameter K was set to 4. Following prior works (Zhou et al. 2024a; Cao et al. 2024), we trained all methods on MVTec AD dataset and then tested them on any other dataset. For MVTec AD

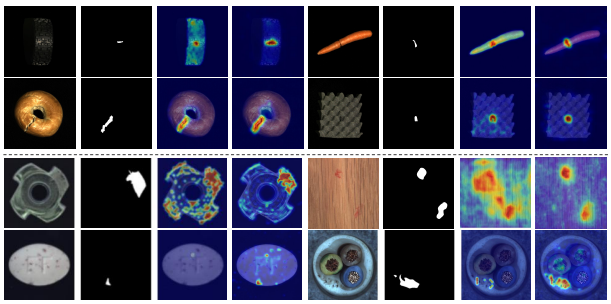

Figure 4: Qualitative results on the MVTec 3D-AD and MVTec AD datesets. Each group, from left to right, consists of image, ground-truth, anomaly map generated by methods w/o HiPL, and anomaly map generated by methods w/ HiPL. Method: PointAD (3D) and AnomalyCLIP (2D).

dataset, we fine-tuned methods on VisA dataset. For zero-shot 3D AD, following (Zhou et al. 2024b), we understood 3D information from 2D perspectives by rendering point clouds into multi-view images. We trained methods on MVTec 3D-AD and then assessed them on remaining two datasets. As for MVTec 3D-AD, methods were trained on Real3D-AD dataset. For zero-shot image classification, we trained methods on certain categories of each dataset and then tested them on the remaining categories. We run experiments under three different seeds.

For ZSAD in 2D, the top competing methods including MultiADS (Sadikaj et al. 2025), AnomalyCLIP (Zhou et al. 2024a), AdaCLIP (Cao et al. 2024), MVFA-AD (Huang et al. 2024), and WinCLIP (Jeong et al. 2023). For ZSAD in 3D, a recent leading method (Zhou et al. 2024b) was included. However, due to few ZSAD methods in 3D, following (Zhou et al. 2024b), we also included a leading point cloud classification method, PointCLIP V2 (Zhu et al. 2023), and fine-tuned AnomalyCLIP and CLIP for zero-shot 3D AD. For classification, three representative methods were considered: CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a), and ProText (Khattak et al. 2025). We replaced prompt learning in existing methods with HiPL to show its effectiveness.

Main Results

ZSAD performance on industrial datasets. As reported in Table 1, in zero-shot 2D AD, without HiPL, WinCLIP (Jeong et al. 2023) achieves superior results on the MVTec AD dataset, while MultiADS (Sadikaj et al. 2025) outperforms all competing methods across VisA and MPDD datasets. However, after introducing HiPL, both AnomalyCLIP (Zhou et al. 2024a) and AdaCLIP (Cao et al. 2024) achieve superior ZSAD performance across all datasets. Hypergraph structure aids in modeling high-order correlations among patches and learning generic semantic information, which helps generalize to unseen objects. Then, both methods can dynamically generate high-order-aware prompts based on the high-order-aware embeddings.

Besides zero-shot 2D AD, HiPL was also adopted for zero-shot 3D AD. Table 1 reports the results of several competing methods over three 3D datasets. For the MVTec 3D-AD dataset, without HiPL, PointCLIP V2 achieves optimal I-AUROC results, while PointAD shows good performance on P-AUROC metric. For the remaining two datasets, the recent leading method PointAD achieves promising results across multiple metrics. However, after introducing HiPL, PointAD improves I-AUROC by 4.7%, AP by 2.3%, PRO by 1.1%, and showing a slight gain of 0.8% in P-AUROC on the MVTec 3D-AD. Additionally, for Eyecandies and Real3D-AD datasets, it also achieves superior results.

To provide more intuitive results, Fig. 4 shows the anomaly segmentation results on two datasets. For MVTec 3D-AD dataset, without HiPL, PointAD can segment anomalous regions but it may also wrongly identify normal areas as anomalies. But, introducing HiPL to PointAD ensures that PointAD can more accurately segment anomalies. Similarly, For MVTec AD, AnomalyCLIP also segments anomalous regions effectively when using HiPL.

ZSAD performance on medical datasets. Following previous methods (Zhou et al. 2024a; Cao et al. 2024), we further assessed the validity of HiPL after integrating it into methods on medical image datasets. Table 2 lists the detailed results. Without HiPL, AdaCLIP shows superior ZSAD performance on medical datasets after being trained on the industrial dataset. However, after using HiPL, both AnomalyCLIP and AdaCLIP show comparable ZSAD performance on these three datasets. In particular, "AnomalyCLIP+HiPL" achieves the best results on the Kvasir dataset, while "AdaCLIP+HiPL" shows the best results on the ClinicDB and ColonDB datasets. With HiPL, both methods have significantly improved metrics across all datasets, which further demonstrates the effectiveness of HiPL.

Performance on natural datasets. We also evaluated the effectiveness of HiPL on two natural datasets, CIFAR10 and CIFAR100, as shown in Table 3. Three classic zero-shot image classification methods were compared: CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a), and ProText (Khattak et al. 2025). Without HiPL, CoCoOp outperforms all competing methods, including recent ProText. With HiPL, both CoOp and CoCoOP markedly improves these two metrics across two natural datasets. "CoOp+HiPL" improves accuracy by 1.2% and Top-5 accuracy by 2.1% on CIFAR10 dataset. With HiPL, CoCoOp also improves accuracy by 1.4% and Top-5 accuracy by 1.1% on CIFAR10 dataset. For the CIFAR100 dataset, after using HiPL, they still outperform their original prompt learning methods.

Methods	CIFAR10	CIFAR100
CoOp	(55.2, 70.3)	(60.1, 77.9)
CoCoOp	(58.4, 75.7)	(68.9, 80.1)
ProText	(57.9, 69.8)	(66.5, 78.3)
CoOp‡	(56.0, 72.4)	(61.5, 79.0)
CoCoOp‡	(**59.9**, **77.7**)	(**70.3**, **82.6**)

Table 3: Performance comparison on two natural datasets, with (accuracy, Top-5 accuracy) listed. The best result is highlighted. ‡: w/ HiPL.

$\mathbf{P_T^H}$	I-ROC	AP	P-ROC	PRO
–	92.5	96.5	91.4	82.5
✓	**93.5**	**97.0**	**91.8**	**83.0**

Table 4: Study on the impact of high-order-aware textual prompts ($\mathbf{P_T^H}$) on performance. Method: AnomalyCLIP + HiPL.

$(\mathbf{P_V^H}) \rightarrow \mathbf{E_T}$	$(\mathbf{P_V^H}) \rightarrow \mathbf{E_I}$	I-ROC	AP	P-ROC	PRO
–	–	87.5	91.0	88.0	79.2
✓	–	88.5	91.9	88.9	80.7
–	✓	89.1	92.4	89.5	80.3
✓	✓	**90.5**	**93.3**	**89.9**	**81.1**

Table 5: Study on the impact of adding high-order-aware visual prompts ($\mathbf{P_V^H}$) to text encoder ($\mathbf{E_T}$) and image encoder ($\mathbf{E_I}$) on performance. Method: AdaCLIP + HiPL.

L	I-ROC	AP	P-ROC	PRO
8	80.0	92.9	95.4	84.0
10	80.7	93.5	**96.1**	84.5
12	**82.0**	**95.3**	96.0	**85.5**
14	79.1	92.8	95.3	83.9

Table 6: Study on the number of learnable word embeddings L. Method: PointAD + HiPL.

K	I-ROC	AP	P-ROC	PRO
3	80.4	93.0	95.7	85.0
4	**82.0**	**95.3**	**96.0**	**85.5**
5	80.7	93.1	95.6	84.6
6	80.6	93.1	95.6	84.2

Table 7: Study on the number of experts K. Method: PointAD + HiPL.

Combination	I-ROC / AP / P-ROC / PRO
$\{MLP, MLP, MLP, MLP\}$	82.6 / 85.9 / 95.4 / 87.0
$\{Conv, Conv, MLP, MLP\}$	83.5 / 86.2 / 95.4 / 87.3
$\{Conv, Conv, Attn, Attn\}$	83.9 / 86.6 / 95.6 / 87.6
$\{Conv, Attn, MLP, Mamba\}$	**84.5 / 87.0 / 95.9 / 88.0**

Table 8: Study on the impact of expert architecture. Method: AnomalyCLIP + HiPL.

n	I-ROC	AP	P-ROC	PRO
Manual setting				
4	91.2	95.0	90.8	81.2
7	92.3	96.0	91.0	**83.0**
9	92.0	95.5	90.7	82.1
Dynamic generation				
ours	**93.5**	**97.0**	**91.8**	**83.0**

Table 9: Study on the number of neighbors n. Method: AnomalyCLIP + HiPL.

Ablation Study

Study on prompts. For high-order-aware textual prompts $\mathbf{P_T^H}$, we assessed the effect of them on the performance of AnomalyCLIP on the MVTec AD dataset. Notably, high-order-aware visual prompts in its text encoder were not removed. Table 4 shows that, without $\mathbf{P_T^H}$, the performance of AnomalyCLIP declines, particularly with a significant drop in the I-ROC evaluation metric. For high-order-aware visual prompts $\mathbf{P_V^H}$, we considered AdaCLIP because of its two encoders (text and image) used originally. Table 5 lists the detailed results. It is clear that adding $\mathbf{P_V^H}$ to one of encoders can increase the performance of AdaCLIP compared to its original prompts in encoders. Also, text encoder with $\mathbf{P_V^H}$ shows better results in PRO, while image encoder with $\mathbf{P_V^H}$ achieves better image-level metrics. But, after adding $\mathbf{P_V^H}$ to both encoders shows the best results, improving I-AUROC by 3.0%, AP by 2.3%, and P-AUROC and PRO by 1.9%.

Study on word embedding length. The length of learnable word embeddings plays a critical role, as it similarly influences the quality of generated textual prompts. Table 6 investigates its impact on performance on the MVTec 3D-AD dataset. With the increase of the length of word embeddings, overall performance improves gradually. However, it is important to note that the excessively long embeddings can lead to a decline in performance, with the best performance achieved when L is set to 12.

Study on expert number and architecture. Another factor affecting the quality of textual prompts is the number of experts, as different experts are responsible for distinct tasks. Table 7 reports the results on the MVTec 3D-AD dataset, The optimal number of experts is found to be 4. Particularly, fewer experts may be suitable for anomaly segmentation tasks, while an excessive number of experts tends to degrade performance. Besides, Table 8 shows the effect of different combinations of architectures on performance on the VisA dataset. It is found that the diversity of expert architectures helps improve performance.

Study on neighbor number. Conventional hypergraph-based methods construct hyperedges by allocating a fixed number of neighbor to each vertex. However, this approach is not ideal for accurately capturing the attributes of different objects. Table 9 illustrates the impact of the number of neighbors on performance on the MVTec AD dataset. It is observed that manually setting neighbors n causes the results across different metrics to fluctuate. When $n = 7$, it yields better performance. Differently, our method dynamically generates neighbors, achieving optimal results.

Conclusion

In this paper, we propose a novel prompt learning framework HiPL for enhanced ZSAD. HiPL introduces two key innovations: dynamic hypergraph construction and high-order-aware prompt generation. HiPL first models high-order correlations among patches within images by dynamically constructing a hypergraph structure and implementing hypergraph semantics convolution. Then, with high-order correlations, HiPL further introduces MoEPLer to generate final high-order-aware textual prompts by comprehensively considering initial prompts generated by experts. Extensive experiments demonstrate the effectiveness of proposed HiPL.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62372242 and Postgraduate Research & Practice Innovation Program of Jiangsu Province (Project No. KYCX25_ 1680).

References

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9592–9600.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2020. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4183–4192.

Bergmann, P.; Jin, X.; Sattlegger, D.; and Steger, C. 2021. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. *arXiv preprint arXiv:2112.09045*.

Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; and Vilariño, F. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43: 99–111.

Bonfiglioli, L.; Toschi, M.; Silvestri, D.; Fioraio, N.; and De Gregorio, D. 2022. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *Proceedings of the Asian Conference on Computer Vision*, 3586–3602.

Cao, Y.; Zhang, J.; Frittoli, L.; Cheng, Y.; Shen, W.; and Boracchi, G. 2024. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, 55–72.

Chai, S.; Jain, R. K.; Mo, S.; Liu, J.; Yang, Y.; Li, Y.; Tateyama, T.; Lin, L.; and Chen, Y.-W. 2024. A novel adaptive hypergraph neural network for enhancing medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 23–33.

Feng, Y.; Huang, J.; Du, S.; Ying, S.; Yong, J.-H.; Li, Y.; Ding, G.; Ji, R.; and Gao, Y. 2025. Hyper-yolo: When visual object detection meets hypergraph computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 2388–2401.

Feng, Y.; You, H.; Zhang, Z.; Ji, R.; and Gao, Y. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3558–3565.

Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Tang, M.; and Wang, J. 2024. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 1932–1940.

Guo, J.; Lu, S.; Jia, L.; Zhang, W.; and Li, H. 2023. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. *Advances in Neural Information Processing Systems*, 36: 10721–10740.

Huang, C.; Jiang, A.; Feng, J.; Zhang, Y.; Wang, X.; and Wang, Y. 2024. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11375–11385.

Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; and Dabeer, O. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19606–19616.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; De Lange, T.; Johansen, D.; and Johansen, H. D. 2020. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, 451–462.

Jiang, J.; Liu, X.; Yan, P.; Wei, S.; and Cui, Y. 2025. Localize-diffusion based dual-branch anomaly detection. *Neural Networks*, 107439.

Jiang, J.; Wei, S.; Xu, X.; Cui, Y.; and Liu, X. 2024. Unsupervised anomaly detection and localization based on two-hierarchy normalizing flow. *IEEE Transactions on Instrumentation and Measurement*, 73: 1–11.

Jin, M.; Wang, H.; and Zhang, Q. 2019. Association rules redundancy processing algorithm based on hypergraph in data mining. *Cluster Computing*, 22(4): 8089–8098.

Khan, B.; Wu, J.; Yang, J.; and Ma, X. 2025. Heterogeneous hypergraph neural network for social recommendation using attention network. *ACM Transactions on Recommender Systems*, 3(3).

Khattak, M. U.; Naeem, M. F.; Naseer, M.; Van Gool, L.; and Tombari, F. 2025. Learning to prompt with text only supervision for vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4230–4238.

La Gatta, V.; Moscato, V.; Pennone, M.; Postiglione, M.; and Sperlí, G. 2022. Music recommendation via hypergraph embedding. *IEEE transactions on neural networks and learning systems*, 34(10): 7887–7899.

Li, X.; Zhang, Z.; Tan, X.; Chen, C.; Qu, Y.; Xie, Y.; and Ma, L. 2024. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16838–16848.

Liu, J.; Xie, G.; Chen, R.; Li, X.; Wang, J.; Liu, Y.; Wang, C.; and Zheng, F. 2023. Real3d-ad: A dataset of point cloud anomaly detection. *Advances in Neural Information Processing Systems*, 36: 30402–30415.

Luo, W.; Cao, Y.; Yao, H.; Zhang, X.; Lou, J.; Cheng, Y.; Shen, W.; and Yu, W. 2025. Exploring Intrinsic Normal Prototypes within a Single Image for Universal Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9974–9983.

Meng, C.; and Motevalli, H. 2024. Link prediction in social networks using hyper-motif representation on hypergraph. *Multimedia Systems*, 30(3): 123.

Meng, S.; Meng, W.; Zhou, Q.; Li, S.; Hou, W.; and He, S. 2024. MoEAD: A parameter-efficient model for multi-class anomaly detection. In *European Conference on Computer Vision*, 345–361.

Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; and Foresti, G. L. 2021. VT-ADL: A vision transformer network for image anomaly detection and localization. In *ISIE*, 01–06.

Peng, J.; Yang, J.; Xia, C.; Li, X.; Guo, Y.; Fu, Y.; Chen, X.; and Cui, Z. 2022. Make U-Net Greater: An Easy-to-Embed Approach to Improve Segmentation Performance Using Hypergraph. *Computer Systems Science & Engineering*, 42(1).

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Rahman, M. M.; Munir, M.; and Marculescu, R. 2024. EMCAD: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11769–11779.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18082–18091.

Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; and Wandt, B. 2023. Asymmetric student-teacher networks for industrial anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2592–2602.

Sadikaj, Y.; Zhou, H.; Halilaj, L.; Schmid, S.; Staab, S.; and Plant, C. 2025. MultiADS: Defect-aware Supervision for Multi-type Anomaly Detection and Segmentation in Zero-Shot Learning. *arXiv preprint arXiv:2504.06740*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Tajbakhsh, N.; Gurudu, S. R.; and Liang, J. 2015. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions medical imaging*, 35(2): 630–644.

Wang, Y.; Peng, J.; Zhang, J.; Yi, R.; Wang, Y.; and Wang, C. 2023. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8032–8041.

Wang, Y.; Peng, K.-C.; and Fu, Y. 2024. Towards zero-shot 3D anomaly localization. *arXiv preprint arXiv:2412.04304*.

Wei, S.; Jiang, J.; and Xu, X. 2025. UniNet: A Contrastive Learning-guided Unified Framework with Feature Selection for Anomaly Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9994–10003.

Yang, D.; Qu, B.; Yang, J.; and Cudré-Mauroux, P. 2020. Lbsn2vec++: Heterogeneous hypergraph embedding for location-based social networks. *IEEE Transactions on Knowledge and Data Engineering*, 34(4): 1843–1855.

Zha, Y.; Wang, J.; Dai, T.; Chen, B.; Wang, Z.; and Xia, S.-T. 2023. Instance-aware dynamic prompt tuning for pre-trained point cloud models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14161–14170.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, Q.; Pang, G.; Tian, Y.; He, S.; and Chen, J. 2024a. AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection. In *The Twelfth International Conference on Learning Representations*.

Zhou, Q.; Yan, J.; He, S.; Meng, W.; and Chen, J. 2024b. PointAD: Comprehending 3D anomalies from points and pixels for zero-shot 3D anomaly detection. *arXiv preprint arXiv:2410.00320*.

Zhu, X.; Zhang, R.; He, B.; Guo, Z.; Zeng, Z.; Qin, Z.; Zhang, S.; and Gao, P. 2023. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2639–2650.