The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

PromptMoE: Generalizable Zero-Shot Anomaly Detection via Visually-Guided Prompt Mixtures

Yuheng Shao[1,2], **Lizhang Wang**[1], **Changhao Li**[1], **Peixian Chen**[2], **Qinyuan Liu**[1, 3*]

[1]School of Computer Science and Technology, Tongji University
[2]Ant Group
[3]Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University
{yuhengshao, lizhangwang, chli, liuqy}@tongji.edu.cn, peixian.cpx@antgroup.com

Abstract

Zero-Shot Anomaly Detection (ZSAD) aims to identify and localize anomalous regions in images of unseen object classes. While recent methods based on vision-language models like CLIP show promise, their performance is constrained by existing prompt engineering strategies. Current approaches, whether relying on single fixed, learnable, or dense dynamic prompts, suffer from a representational bottleneck and are prone to overfitting on auxiliary data, failing to generalize to the complexity and diversity of unseen anomalies. To overcome these limitations, we propose `PromptMoE`. Our core insight is that robust ZSAD requires a compositional approach to prompt learning. Instead of learning monolithic prompts, `PromptMoE` learns a pool of expert prompts, which serve as a basis set of composable semantic primitives, and a visually-guided Mixture-of-Experts (MoE) mechanism to dynamically combine them for each instance. Our framework materializes this concept through a Visually-Guided Mixture of Prompt (VGMoP) that employs an image-gated sparse MoE to aggregate diverse normal and abnormal expert state prompts, generating semantically rich textual representations with strong generalization. Extensive experiments across 15 datasets in industrial and medical domains demonstrate the effectiveness and state-of-the-art performance of `PromptMoE`.

Introduction

Anomaly detection is a crucial technology with widespread applications, such as ensuring quality in industrial manufacturing (Bergmann et al. 2019; Jezek et al. 2021; Zou et al. 2022; Wang et al. 2024) and assisting in medical diagnosis (Buda, Saha, and Mazurowski 2019; Huang et al. 2024). For example, in automated production lines, models must accurately detect subtle defects like scratches, blemishes, or deformations. In practice, however, new product categories and unforeseen defect types frequently emerge, making it infeasible to collect extensive labeled data or retrain models for every new class. This has highlighted the importance of Zero-Shot Anomaly Detection (ZSAD), which aims to detect anomalies in object categories that are not seen during training. ZSAD is both essential and inherently challenging.

Recently, the emergence of large-scale vision-language models, particularly CLIP (Radford et al. 2021), offers new possibilities for ZSAD due to their strong generalization and zero-shot recognition capabilities. CLIP, trained on massive image-text pairs, learns a shared semantic space that aligns visual and textual information. Motivated by this, recent studies have explored adapting CLIP for ZSAD. Some methods (Jeong et al. 2023; Chen, Han, and Zhang 2023a) reformulate ZSAD as an image-text matching task using hand-crafted prompts to describe "normal" and "abnormal" states. To reduce the need for manual design, later methods like AnomalyCLIP (Zhou et al. 2024) introduced learnable prompts that are optimized on auxiliary datasets.

Although prompt-based methods have shown promising results, they face notable limitations in representational power and generalization. Firstly, the single-prompt strategy, whether manually designed or learned as a single normal/abnormal pair, suffers from a clear representational bottleneck,

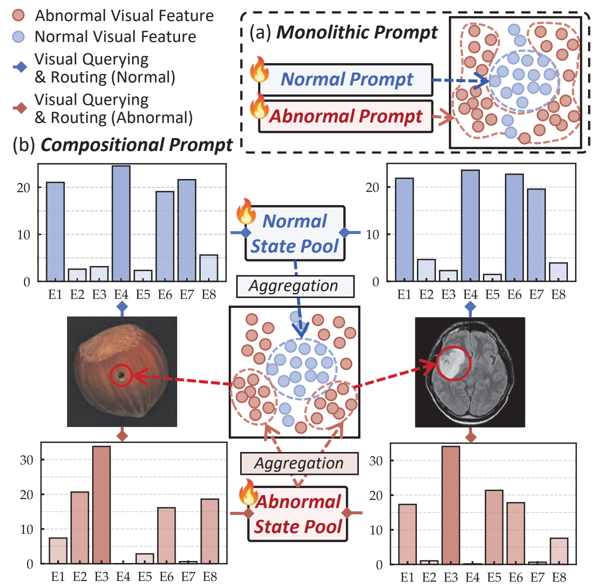

Figure 1: (a) Monolithic Prompt, (b) Compositional Prompt. Unlike monolithic prompts with fixed representations, our compositional prompts employs visual querying to dynamically aggregate expert prompts, treated as semantic primitives, into an instance-specific textual representation.

*Corresponding author.

as a fixed prompt vector struggles to capture the diverse patterns of normality and abnormality in unseen classes. Secondly, simply increasing the number of learnable static prompts is not a viable solution, as it significantly increases the risk of overfitting to the auxiliary data; the model tends to memorize specific pattern-prompt combinations from the training set instead of learning generalizable abstract concepts. Furthermore, approaches (Cao et al. 2024; Qu et al. 2024) inspired by CoCoOp (Zhou et al. 2022) that dynamically generate prompts from visual instances often rely on a single mapping network to handle all visual variations, making it difficult to generate specialized prompts for specific, fine-grained anomaly patterns. Finally, the overfitting issue inherent in static designs persists even in deep prompting approaches. As demonstrated by previous work (Zhou et al. 2024), when static learnable prompts are inserted into the text encoder's intermediate layers, they are effective only in shallow stages. Applying them to deeper layers harms zero-shot detection on unseen classes, revealing a critical generalization failure of static prompt designs.

To address the aforementioned challenges, we propose `PromptMoE`, a novel framework that shifts prompt learning from a monolithic to a compositional paradigm, as conceptually illustrated in Fig. 1. Our key insight is that to effectively represent the open-ended and diverse nature of normal and abnormal patterns in ZSAD, a compositional representation capability for prompts is key, rather than learning a fixed or single dynamic prompt. Accordingly, `PromptMoE` shifts the paradigm from learning a monolithic state description to learning a basis set of composable semantic primitives (i.e., expert prompts) and a router to dynamically combine them based on the visual instance. This Mixture-of-Experts (MoE) approach allows for the creation of a rich variety of instance-specific prompts from a finite set of parameters, while the sparse activation inherent to MoE helps mitigate overfitting. Our framework implements this strategy through a novel **Visually-Guided Mixture of Prompt (VGMoP)** module, which employs an image-gated sparse MoE for specialized selection from distinct expert pools to construct fine-grained, instance-adaptive textual prompts. This process is further regularized by expert load balancing and decoupling losses to enhance the diversity of the expert pool, thereby improving generalization to unseen classes.

Our main contributions are summarized as follows:

- We propose `PromptMoE`, a framework for ZSAD that replaces monolithic prompts with a compositional learning paradigm. It employs a MoE mechanism to dynamically combine a learned basis of semantic primitives, enhancing generalization and mitigating overfitting.
- We design a VGMoP module to realize this paradigm. It utilizes an image-gated sparse router and distinct expert pools to dynamically compose specialized normal and abnormal state prompts for each visual instance.
- Extensive experiments 15 industrial and medical datasets demonstrate that our method achieves SOTA performance, validating its effectiveness on the ZSAD task.

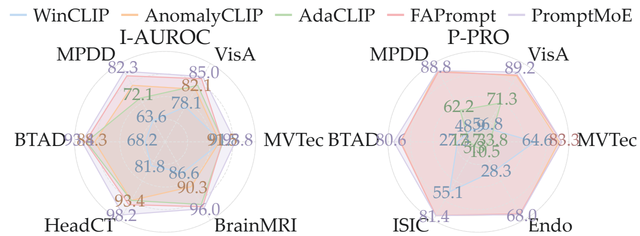

Figure 2: ZSAD performance of `PromptMoE` compared to state-of-the-art methods. Left: I-AUROC. Right: P-PRO.

Related Work

Zero-Shot Anomaly Detection

Recent research in ZSAD has been significantly advanced by Vision-Language Models (VLM) like CLIP (Radford et al. 2021), which typically reframe the task as image-text matching. Early pioneering works, such as WinCLIP (Jeong et al. 2023), introduce a prompt ensembling strategy, generating rich textual descriptions by combining numerous handcrafted templates with state words. To overcome the reliance on manual design, AnomalyCLIP (Zhou et al. 2024) proposes prompt learning, which optimizes learnable vectors to replace fixed textual prompts, thereby automatically learning general anomaly concepts.

However, static learnable prompts show limited adaptability to different visual instances, leading subsequent research toward dynamic, visually-guided prompt generation. For instance, AdaCLIP (Cao et al. 2024) combines static and dynamic prompts, while VCP-CLIP (Qu et al. 2024) injects visual context at multiple stages of the text encoder. Another approach, taken by Bayes-PFL (Qu et al. 2025), models the prompt space as a learnable probability distribution and generates diverse prompts via sampling. Similarly, FAPrompt (Zhu et al. 2025) learns a set of fine-grained prompts and adapts them using a visual prior extracted from the test image. While these methods enhance instance-adaptability, they often rely on complex generative or probabilistic modeling. In contrast, our work explores using a MoE mechanism to achieve visually-guided prompt generation in a compositional manner.

Mixture of Experts

Mixture-of-Experts is a classic conditional computation paradigm that effectively scales model capacity without a proportional increase in computational cost by activating only a sparse subset of "expert" subnetworks for each input (Shazeer et al. 2017). This strategy has recently been widely and successfully applied to scale large-scale language models to trillions of parameters (Fedus, Zoph, and Shazeer 2022) and has also been adapted for vision models (Riquelme et al. 2021). In these mainstream applications, MoE typically functions as a dynamic router that selects one of several expert MLP layers to process each input token. In contrast, our `PromptMoE` framework explores a novel application of MoE in prompt engineering: we repurpose the MoE mechanism as a strategy for compositional

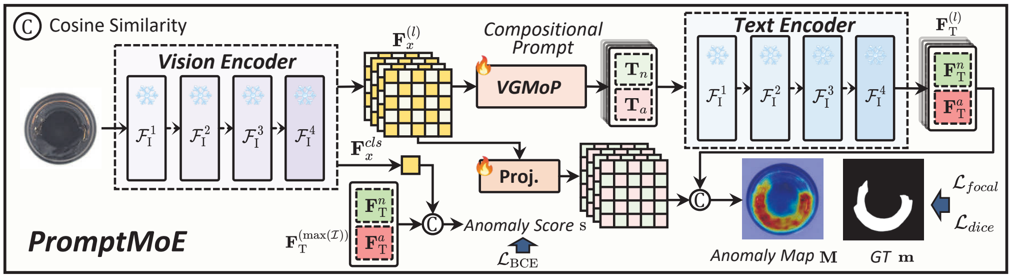

Figure 3: Framework of PromptMoE.

prompt generation. Under this framework, the "experts" are learnable prompt embedding segments, and an image-gated router learns to select a specialized subset of these experts to dynamically construct a highly adaptive textual representation for each visual instance.

Method

Problem Formulation

The goal of ZSAD is to detect and segment anomalous regions in images, especially for object classes that are unseen during training. Formally, given an input image $\mathbf{x} \in \mathbb{R}^{h\times w\times 3}$, the model aims to predict an image-level anomaly score s as well as a pixel-level anomaly map $\mathbf{M} \in \mathbb{R}^{h\times w}$. We utilize an auxiliary training dataset $\mathcal{D}_{train} = \{(\mathbf{x}_i, c_i, \mathbf{m}_i)\}_{i=1}^{N_{train}}$, where each sample consists of an image $\mathbf{x}_i$, an image-level label c_i, and a pixel-level anomaly mask $\mathbf{m}_i$. The label $c_i \in \{0, 1\}$ indicates whether the image is normal ($c_i = 0$) or abnormal ($c_i = 1$), and the mask $\mathbf{m}_i \in \{0, 1\}^{h\times w}$ provides pixel-wise annotations for anomalous regions. The test set $\mathcal{D}_{test} = \{\mathbf{x}_j\}_{j=1}^{N_{test}}$ contains images from unseen classes that are disjoint from those in the training set, i.e., $\mathcal{C}_{train} \cap \mathcal{C}_{test} = \emptyset$. The model is trained to learn a generalizable patterns of normality and abnormality from $\mathcal{D}_{train}$, and to achieve the detection of anomalies in unseen classes within $\mathcal{D}_{test}$.

Overview

We propose PromptMoE, a framework that utilizes visually-guided prompt mixtures to improve generalization to unseen classes, thereby enhancing ZSAD performance. As illustrated in Fig. 3, given an input image $\mathbf{x}$, a vision encoder extracts comprehensive patch-level features $\mathbf{F}_x^{(l)} \in \mathbb{R}^{(HW+1)\times D_x}$ and a global feature $\mathbf{F}_x^{cls} \in \mathbb{R}^{D_x}$, where H and W denote the number of patches, and D_x is the feature dimension. The core innovation of PromptMoE is the VGMoP module, detailed in Fig. 4. It takes the visual features $\mathbf{F}_x^{(l)}$ as input and dynamically generates fine-grained, instance-specific normal and abnormal textual prompts, $\mathbf{T}_n^{(l)}$ and $\mathbf{T}_a^{(l)}$, for each visual instance. Finally, layer-wise similarities between textual embeddings $\mathbf{F}_\mathrm{T}^{(l)}$ and patch features $\mathbf{F}_x^{(l)}$ are aggregated to produce the anomaly map $\mathbf{M}$.

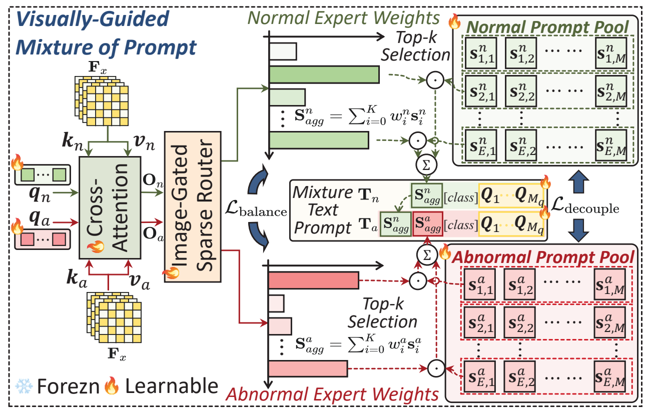

Figure 4: Architecture of our proposed Visually-Guided Mixture of Prompt (VGMoP) module. A sparse router, guided by cross-attention between learnable queries and image features, dynamically selects and aggregates top-k experts from respective prompt pools to construct instance-specific normal and abnormal text prompts.

Visually-Guided Mixture of Prompt

Previous ZSAD method (Zhou et al. 2024) introduces learnable pairs of normal and abnormal text prompts to guide pre-trained CLIP models in distinguishing these two concepts, thereby avoiding tedious manual template design. However, such single, fixed prompts possess limited expressive power, struggling to adequately represent the complex and diverse anomaly patterns found in unseen classes. Although some subsequent studies (Cao et al. 2024; Qu et al. 2024) attempt to dynamically generate prompts using visual features to enhance instance adaptability, their effectiveness is often constrained by the inherent representational bottlenecks of visual encoders. Furthermore, simply increasing the number of prompts or their complexity can easily lead to model overfitting on seen classes, consequently impairing its generalization ability to unseen classes.

To address these limitations, we propose VGMoP, a visually-guided prompt mixture that effectively utilizes richer prompt information to enhance generalization to unseen classes. Unlike fixed or single dynamically-generated prompts, VGMoP employs an image-gated sparse MoE mechanism. This adaptively selects and aggregates a sparse subset of relevant expert prompts from a pool for the current visual query instance, thereby constructing distinct textual representations for normal and abnormal states.

Mixture Text Prompt Structure. Under the VGMoP strategy, we construct two mixture text prompts for each visual instance $\mathbf{x}$: a normal-state prompt $\mathbf{T}_n \in \mathbb{R}^{N_n \times D}$ and an abnormal-state prompt $\mathbf{T}_a \in \mathbb{R}^{N_a \times D}$, where D denotes the token embedding dimension. The prompt are defined as:

$$\begin{aligned} \mathbf{T}_n &= [\mathbf{S}^n_{\text{agg}}][\texttt{cls}][\mathbf{Q}_{\text{ctx}}], \\ \mathbf{T}_a &= [\mathbf{S}^n_{\text{agg}}][\mathbf{S}^a_{\text{agg}}][\texttt{cls}][\mathbf{Q}_{\text{ctx}}], \end{aligned} \tag{1}$$

where $\mathbf{S}^n_{\text{agg}} \in \mathbb{R}^{M_n \times D}$ represents the aggregated normal state. Inspired by Li et al. (2024), we append an aggregated abnormal state suffix $\mathbf{S}^a_{\text{agg}} \in \mathbb{R}^{M_a \times D}$ to $\mathbf{S}^n_{\text{agg}}$ to construct the abnormal-state prompt. Both $\mathbf{S}^n_{\text{agg}}$ and $\mathbf{S}^a_{\text{agg}}$ are generated by the visually-guided MoE mechanism. $\mathbf{Q}_{\text{ctx}} \in \mathbb{R}^{M_q \times D}$ denotes a set of learnable context tokens shared by both prompts, and $[\texttt{cls}]$ denotes the embedding of a given category name or a generic placeholder such as "object". The resulting $\mathbf{T}_n$ and $\mathbf{T}_a$ provide the text encoder with dynamically adapted, instance-specific semantic guidance.

Visually-Guided MoE for State Aggregation. To achieve effective compositional representations and enhance generalization to unseen classes, $\mathbf{S}^n_{\text{agg}}$ and $\mathbf{S}^a_{\text{agg}}$ are generated using a visually-guided MoE procedure applied independently for each layer $l \in \mathcal{I}$. For each layer, the process begins by constructing a comprehensive visual representation $\mathbf{F}^{(l)}_x \in \mathbb{R}^{(HW+1) \times D_x}$, obtained by concatenating the patch features of the layer with the global feature $\mathbf{F}^{cls}_x$. A set of learnable, layer-specific state queries $\mathbf{q}^{(l)} \in \mathbb{R}^{N_q \times D}$ then attend to $\mathbf{F}^{(l)}_x$ via a cross-attention layer to extract fine-grained, state-relevant visual contextual representations $\mathbf{O}^{(l)} \in \mathbb{R}^{N_q \times D}$, computed as:

$$\begin{aligned} &Q^{(l)} = \mathbf{q}^{(l)}, K^{(l)} = \mathbf{F}^{(l)}_x \mathbf{W}^{(l)}_K, V^{(l)} = \mathbf{F}^{(l)}_x \mathbf{W}^{(l)}_V, \\ &\mathbf{O}^{(l)} = \text{Softmax}\left(\frac{Q^{(l)} {K^{(l)}}^\top}{\sqrt{D}}\right) V^{(l)}, \end{aligned} \tag{2}$$

where $\mathbf{W}^{(l)}_K, \mathbf{W}^{(l)}_V \in \mathbb{R}^{D_x \times D}$ are linear projection matrices mapping $\mathbf{F}^{(l)}_x$ to dimension D. The routing representation is then obtained by average pooling: $\mathbf{r}^{(l)} = \text{mean}(\mathbf{O}^{(l)}) \in \mathbb{R}^D$.

This representation $\mathbf{r}^{(l)}$ is fed into a layer-specific image-gated sparse router $G^{(l)}(\cdot)$, composed of a two-layer MLP (i.e., Linear-ReLU-Linear). The router produces routing logits $\mathbf{z}^{(l)}$ for the corresponding layer-specific expert prompt pool $\mathcal{E}^{(l)} = \{\mathbf{s}^{(l)}_j \in \mathbb{R}^{M \times D}\}^E_{j=1}$. From these logits, we select the top-k values, denoted as $\mathbf{z}^{(l)}_{\text{top}}$, and retrieve their associated expert prompts $\{\mathbf{s}^{(l)}_{\text{top},i}\}^k_{i=1}$. Normalized gating weights are then computed as $\mathbf{w}^{(l)} = \text{Softmax}(\mathbf{z}^{(l)}_{\text{top}})$, which are used to aggregate the layer-specific state prompt $\mathbf{S}^{(l)}_{\text{agg}}$:

$$\mathbf{S}^{(l)}_{\text{agg}} = \sum_{i=1}^{k} \mathbf{w}^{(l)}_i \mathbf{s}^{(l)}_{\text{top},i}. \tag{3}$$

Note that we utilize independent expert prompt pools for normal and abnormal states, $\mathcal{E}_n$ and $\mathcal{E}_a$ respectively, and apply this Visually-Guided MoE procedure to each. This separated design is intended to provide specialized representational spaces for normality and abnormality, thereby enhancing generalization to complex patterns in unseen classes.

Auxiliary Loss. For each visual instance, VGMoP selects k expert state prompts from the expert pool. However, during training, prompts that perform better in the early stages are more likely to be repeatedly selected by the sparse router, leading to imbalanced prompt utilization. This imbalance may cause the prompt pool to degenerate into a fixed combination of prompt states, undermining the dynamics and sparsity of selection. Drawing inspiration from prior research (Shazeer et al. 2017; Fedus, Zoph, and Shazeer 2022), this loss encourages a uniform average contribution from all experts across the batch, defined as:

$$\begin{aligned} \mathcal{L}_{\text{balance}} &= \alpha \sum_{l \in \mathcal{I}} \left(E \sum_{j=1}^{E} \left(\frac{1}{B} \sum_{i=1}^{B} \mathbf{p}^{(l)}_{i,j} \right)^2 \right), \\ \mathbf{p}^{(l)} &= \text{Softmax}(\mathbf{z}^{(l)}), \end{aligned} \tag{4}$$

where B is the batch size, E is the number of experts, $\mathbf{p}^{(l)}$ is the router's output probability distribution over experts at layer l, and α is the loss weight. In practice, we find this loss is most effective when combined with a learning rate warmup schedule during the initial training epochs.

To further enhance the generalization ability of state prompts, we introduce an expert decoupling loss, $\mathcal{L}_{\text{decouple}}$, aimming to promote representational diversity within the expert prompt pool, ensuring that individual state prompt learn discriminative features. Let $\bar{\mathbf{s}}^{(l)}_j = \text{mean}(\mathbf{s}^{(l)}_j) \in \mathbb{R}^D$ denote the mean embedding representation of each state prompt $\mathbf{s}^{(l)}_j$ from the prompt pool $\mathcal{E}^{(l)}$. These are then normalized to yield $\hat{\mathbf{s}}_j = \bar{\mathbf{s}}_j / \|\bar{\mathbf{s}}_j\|_2$. The normalized mean representations $\hat{\mathbf{s}}^{(l)}_j$ form the rows of a matrix $\hat{\mathbf{S}}^{(l)} \in \mathbb{R}^{E \times D}$. The expert decoupling loss $\mathcal{L}_{\text{decouple}}$ is defined as:

$$\mathcal{L}_{\text{decouple}} = \beta \sum_{l \in \mathcal{I}} \|\hat{\mathbf{S}}^{(l)} (\hat{\mathbf{S}}^{(l)})^T - \mathbf{I}_E\|^2_F, \tag{5}$$

where β is a hyperparameter weight, $\mathbf{I}_E$ is the identity matrix, and $\|\cdot\|^2_F$ denotes the squared Frobenius norm. This loss promotes orthogonality among the mean representations of different expert prompts within each layer's pool, enhancing the overall model's diversity.

Discussion

In summary, the design of our `PromptMoE` framework addresses several key considerations. First, to effectively utilize the rich yet redundant patch features from the visual encoder, we employ a query-driven cross-attention mechanism. This mechanism learns to actively distill state-relevant visual signals from the numerous patches, avoiding the loss of critical local information that can occur with simple average pooling. Second, this query-and-distill mechanism is extended layer-wise, enabling the generation of fine-grained textual prompts that correspond to visual features at diverse semantic depths. Finally, dedicated auxiliary losses are introduced to ensure the MoE system trains stably and generalizes effectively: a load balancing loss maintains the router's dynamism, while a decoupling loss enhances generalization to unseen classes by promoting expert diversity.

Training and Inference

`PromptMoE` computes the pixel-level anomaly map $\mathbf{M}$ by aggregating similarity maps over a predefined set of visual encoder layers $\mathcal{I}$. For each layer $l \in \mathcal{I}$, the map is derived from the similarity between its patch features $\mathbf{F}_x^{(l)}$ and the corresponding textual embedding $\mathbf{F}_\mathrm{T}^{(l)}$:

$$\mathbf{M} = \sum_{l \in \mathcal{I}} \mathrm{Softmax}(\mathrm{Up}(\mathbf{F}_x^{(l)} {\mathbf{F}_\mathrm{T}^{(l)}}^\top)/\tau), \tag{6}$$

where $\mathrm{Up}(\cdot)$ upsamples the similarity map to (h, w) and τ is a temperature parameter. The image-level anomaly score s combines the peak value of the anomaly map $\mathbf{M}$ with the similarity between the global image feature $\mathbf{F}_x^{cls}$ and the final layer's textual embedding $\mathbf{F}_\mathrm{T}^{(\max(\mathcal{I}))}$:

$$s = \frac{1}{2}(\max(\mathbf{M}) + \mathrm{Softmax}(\mathbf{F}_x^{cls} {\mathbf{F}_\mathrm{T}^{(\max(\mathcal{I}))}}^\top / \tau')). \tag{7}$$

Similar to previous works (Zhou et al. 2024; Gu et al. 2024; Cao et al. 2024), `PromptMoE` supervises the anomaly map $\mathbf{M}$ using Dice (Milletari, Navab, and Ahmadi 2016) and Focal (Lin et al. 2017) losses against the ground truth mask $\mathbf{m}$. In addition, the anomaly score s is optimized using BCE loss with the ground truth label c. The overall loss $\mathcal{L}_\text{total}$ is:

$$\begin{aligned}\mathcal{L}_\text{total} = \underbrace{\mathrm{BCE}(s, c)}_{\text{Classify}} + \underbrace{\mathcal{L}_\text{balance} + \mathcal{L}_\text{decouple}}_{\text{Auxiliary}} \\ + \underbrace{\mathrm{Dice}(\mathbf{M}, \mathbf{m}) + \mathrm{Focal}(\mathbf{M}, \mathbf{m})}_{\text{Segment}}.\end{aligned} \tag{8}$$

Experiments

Experiment Settings

Datasets. To comprehensively evaluate the generalization capability of our proposed `PromptMoE` on the ZSAD task, we conduct extensive experiments on 15 real-world datasets spanning both industrial manufacturing and medical diagnosis domains. We utilize seven widely-used industrial benchmarks, including MVTec AD (Bergmann et al. 2019), VisA (Zou et al. 2022), MPDD (Jezek et al. 2021), BTAD (Mishra et al. 2021), SDD (Tabernik et al. 2020), DAGM (Wieler and Hahn 2007), and DTD-Synthetic (Aota, Tong, and Okatani 2023); as well as eight medical datasets from diverse modalities and application scenarios, including HeadCT (Salehi et al. 2021), BrainMRI (Kanade and Gumaste 2015), Br35H (Hamada 2020), ISIC (Codella et al. 2018), CVC-ColonDB (Tajbakhsh, Gurudu, and Liang 2015), CVC-ClinicDB (Bernal et al. 2015), Kvasir (Jha et al. 2019), and Endo (Hicks et al. 2021). Following the standard ZSAD setting, we primarily train our model on the MVTec AD dataset and then perform zero-shot inference and evaluation on the other 14 datasets. To evaluate performance on MVTec AD itself, we train our model on VisA, a dataset with a disjoint set of object categories.

Metrics. To evaluate the performance of our model, we follow the standard protocols used in previous works (Zhou et al. 2024). For the image-level anomaly detection, we report the AUROC and AP. For the pixel-level anomaly localization, we use pixel-wise AUROC and the area under the Per-Region Overlap curve (PRO) (Bergmann et al. 2019).

Implementation details. We utilize the pre-trained CLIP (Radford et al. 2021) (ViT-L/14@336px) released by OpenAI as our backbone, with all of its parameters frozen during training. All input images are resized to 518×518 pixels. We extract patch features from the $\{6, 12, 18, 24\}$-th layers of the 24-layer visual encoder. The model is trained for 15 epochs using the Adam optimizer with a batch size of 16 and a learning rate of 0.001, employing a warm-up strategy during the first 3 epochs.

For our VGMoP module, we set the number of queries $N_q = 8$ and experts $E = 8$, from which we sparsely select the top $k = 4$ for aggregation; its cross-attention utilizes 8 heads and its router has a hidden dimension of 256. The resulting normal and abnormal state sequences have lengths $M_n = 5$ and $M_a = 6$, respectively, while the shared context $\mathbf{Q}_{ctx}$ has a length of $M_q = 8$. The auxiliary loss coefficients α and β are set to 0.01 and 0.005, respectively. All experiments are conducted on a RTX3090 using PyTorch 1.13.1.

Comparisons with State-of-the-Art Methods

We compare `PromptMoE` against recent SOTA methods, including WinCLIP (Jeong et al. 2023), APRIL-GAN (Chen, Han, and Zhang 2023b), AnomalyCLIP (Zhou et al. 2024), AdaCLIP (Cao et al. 2024), and FAPrompt (Zhu et al. 2025). For a fair comparison, all results are cited from their original papers, while any missing values are reproduced using the official code under a unified setting.

Quantitative Comparisons. We quantitatively compare `PromptMoE` against recent SOTA methods across 15 industrial and medical datasets. The detailed results in Table 1,

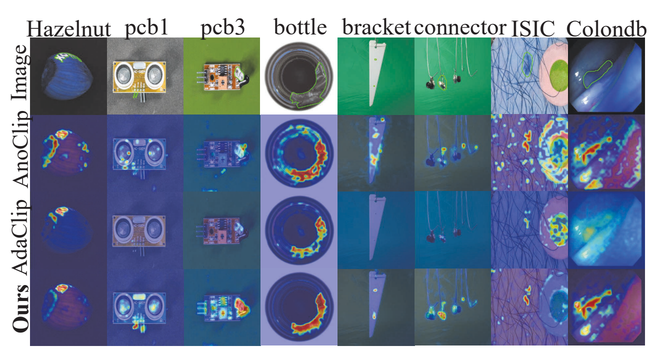

Figure 5: Qualitative comparison of anomaly localization results across different ZSAD methods.

Domain	Metric	Datasets	CLIP	WinCLIP	APRIL-GAN	AnomalyCLIP	AdaCLIP	FAPrompt	PromptMoE
Industrial	Image-level (AUROC, AP)	MVTec AD	(74.1, 87.6)	(91.8, 96.5)	(86.1, 93.5)	(91.5, 96.2)	(92.0, 96.4)	(91.9, 95.7)	(**93.8**, **97.2**)
		VisA	(66.4, 71.5)	(78.1, 81.2)	(78.0, 81.4)	(82.1, 85.4)	(83.0, 84.9)	(84.5, 86.8)	(**85.0**, **88.2**)
		MPDD	(54.3, 65.4)	(63.6, 69.9)	(73.0, 80.2)	(77.0, 82.0)	(72.1, 77.6)	(80.6, 83.3)	(**82.3**, **84.5**)
		BTAD	(34.5, 52.5)	(68.2, 70.9)	(73.6, 68.6)	(88.3, 87.3)	(91.6, 92.4)	(92.0, 92.2)	(**93.4**, **95.7**)
		SDD	(65.7, 45.2)	(84.3, 77.4)	(79.8, 71.4)	(84.7, 80.0)	(81.2, 72.6)	(**98.6**, **95.9**)	(97.4, 94.0)
		DAGM	(79.6, 59.0)	(91.8, 79.5)	(94.4, 83.8)	(97.5, 92.3)	(96.5, 95.7)	(98.9, 95.7)	(**98.9**, **96.4**)
		DTD-Synthetic	(71.6, 85.7)	(93.2, 92.6)	(86.4, 95.0)	(93.5, 97.0)	(92.8, 97.0)	(**95.9**, **98.3**)	(**95.9**, 98.1)
		Average	(63.7, 66.7)	(81.6, 81.1)	(81.6, 82.0)	(87.8, 88.6)	(87.0, 88.0)	(91.7, 92.5)	(**92.4**, **93.4**)
	Pixel-level (AUROC, PRO)	MVTec AD	(38.4, 11.3)	(85.1, 64.6)	(87.6, 44.0)	(91.1, 81.4)	(86.8, 33.8)	(90.6, **83.3**)	(**91.8**, 83.2)
		VisA	(46.6, 14.8)	(79.6, 56.8)	(94.2, 86.8)	(95.5, 87.0)	(95.1, 71.3)	(**95.9**, 87.5)	(95.6, **89.2**)
		MPDD	(62.1, 33.0)	(76.4, 48.9)	(94.1, 83.2)	(96.5, 88.7)	(96.4, 62.2)	(96.5, 87.9)	(**96.8**, **88.8**)
		BTAD	(30.6, 4.4)	(72.7, 27.3)	(60.8, 25.0)	(94.2, 74.8)	(87.7, 17.7)	(**95.6**, 75.2)	(94.9, **80.6**)
		SDD	(39.0, 8.9)	(68.8, 24.2)	(79.8, 65.1)	(90.6, 67.8)	(71.7, 17.6)	(**98.3**, 93.6)	(98.1, **95.6**)
		DAGM	(28.2, 2.9)	(87.6, 65.7)	(82.4, 66.2)	(95.6, 91.0)	(97.0, 40.9)	(**98.3**, **95.4**)	(97.8, 93.9)
		DTD-Synthetic	(33.9, 12.5)	(83.9, 57.8)	(95.3, 86.9)	(97.9, 92.3)	(94.1, 24.9)	(**98.3**, 93.1)	(**98.3**, **93.2**)
		Average	(39.8, 12.5)	(79.1, 49.3)	(84.8, 65.3)	(94.5, 83.3)	(89.8, 38.4)	(**96.2**, 88.0)	(**96.2**, **89.2**)
Medical	Image-level (AUROC, AP)	HeadCT	(56.5, 58.4)	(81.8, 80.2)	(89.1, 89.4)	(93.4, 91.6)	(93.4, 92.2)	(94.8, 93.5)	(**98.2**, **98.2**)
		BrainMRI	(73.9, 81.7)	(86.6, 91.5)	(89.3, 90.9)	(90.3, 92.2)	(94.9, 94.2)	(95.5, 95.6)	(**96.0**, **96.6**)
		Br35H	(78.4, 78.8)	(80.5, 82.2)	(93.1, 92.9)	(94.6, 94.7)	(95.7, 95.7)	(97.8, 97.5)	(**97.9**, **97.7**)
		Average	(69.6, 73.0)	(83.0, 84.6)	(90.5, 91.1)	(92.8, 92.8)	(94.7, 94.0)	(96.0, 95.5)	(**97.4**, **97.5**)
	Pixel-level (AUROC, PRO)	ISIC	(33.1, 5.8)	(83.3, 55.1)	(89.4, 77.2)	(89.7, 78.4)	(85.4, 5.3)	(90.9, 81.2)	(**91.1**, **81.4**)
		CVC-ColonDB	(49.5, 15.8)	(70.3, 32.5)	(78.4, 64.6)	(81.9, 71.3)	(79.3, 6.5)	(**84.6**, 74.7)	(84.3, **74.9**)
		CVC-ClinicDB	(47.5, 18.9)	(51.2, 13.8)	(80.5, 60.7)	(82.9, 67.8)	(**85.3**, 14.6)	(84.7, 70.1)	(84.5, **70.9**)
		Kvasir	(44.6, 17.7)	(69.7, 24.5)	(75.0, 36.2)	(78.9, 45.6)	(79.4, 12.3)	(81.2, 47.8)	(**81.6**, **48.8**)
		Endo	(45.2, 15.9)	(68.2, 28.3)	(81.9, 54.9)	(84.1, 63.6)	(84.0, 10.5)	(**86.4**, 67.2)	(86.3, **68.0**)
		Average	(44.0, 14.8)	(68.5, 30.8)	(81.0, 58.2)	(83.5, 65.3)	(82.7, 9.8)	(**85.6**, 68.2)	(**85.6**, **68.8**)

Table 1: ZSAD performance comparison on industrial and medical domains. Best results are in **bold**, second-best are underlined.

visually summarized by the radar chart in Fig. 2, clearly indicate that our PromptMoE demonstrates comprehensive superiority over competing methods in both image-level and pixel-level metrics. For instance, it achieves a 93.8% Image-level AUROC on the MVTec AD benchmark (+1.8% over the runner-up). This strong generalization is further validated on medical datasets, notably on HeadCT, where our model surpasses the next-best method by a significant 3.4% in AUROC and 4.7% in AP. The exceptional performance on the unseen medical domain, despite being trained solely on industrial data, strongly validates the success of our visually-guided compositional prompts in learning generalizable concepts of normality and abnormality.

Qualitative Comparisons. To provide a visual intuition of the superiority of PromptMoE, we present a series of anomaly localization results in Fig. 5, comparing qualitatively against AnomalyCLIP and AdaCLIP. As can be observed, the results from competing methods are often imprecise, suffering from issues such as false positives on normal regions or incomplete coverage of defective areas. In contrast, the anomaly maps generated by our PromptMoE not only focus more accurately on the true defects but also exhibit highlighting that aligns more closely with the actual anomaly contours. These results suggest that our visually-guided compositional prompt mechanism, by generating more discriminative textual representations for each instance, enables a finer-grained image-text alignment that leads to superior localization performance.

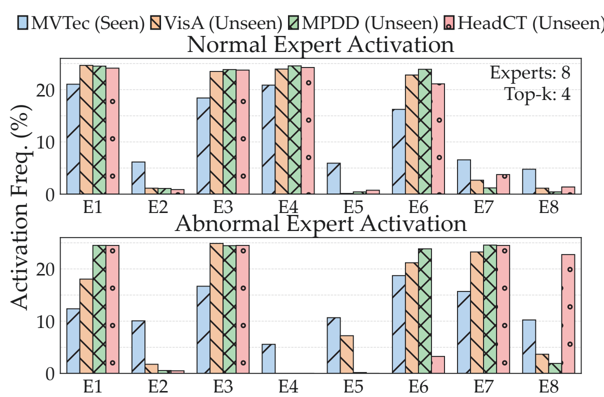

Figure 6: Activation frequencies of normal and abnormal experts across training and test datasets.

Analysis of Expert Activation. To gain deeper insight into the dynamic routing behavior of the VGMoP module, we analyze its expert activation patterns. As illustrated in Fig. 6, we compute the cumulative activation frequency for each expert in both the normal and abnormal prompt pools on the seen dataset (MVTec AD) and across several unseen datasets.

The results clearly reveal that the normal and abnormal branches learn two distinct and highly task-adaptive routing strategies. For the normal state, the router's selections consistently converge to a small, fixed set of core experts across all test datasets. This indicates that our model successfully distills a core set of highly generalizable semantic primitives for the concept of "normality" from the diverse training data.

In stark contrast, the expert activation for the abnormal state remains highly dynamic and sparse across different unseen datasets, demonstrating that the model flexibly composes different expert primitives for varied anomaly patterns. For instance, we observe that a general-purpose expert (e.g., E3) is combined with a dataset-specific subset of other experts (e.g., higher activation of E7 and E8 on VisA vs. E1 and E6 on MPDD). Furthermore, the specific activation weights for the instances in Fig. 1 reveal that the router assigns non-uniform weights to the selected experts, forming a meaningful weighted combination rather than a simple average. This provides further evidence of the router's nuanced, instance-specific decision-making.

Ablation Studies

Analysis of the Compositional Prompting. To validate the effectiveness of our proposed compositional prompting strategy, we conduct an ablation study progressively building from a simple static prompt baseline to our full `PromptMoE` model. As shown in Table 2, the "Static Prompt" baseline (using a single learnable prompt) yields limited performance. While "Static Ensemble", which uses a set of static prompts and averages their outputs, provides some improvement, its effectiveness remains constrained. Notably, replacing the static ensemble with "VG-MoP" (single-layer, without auxiliary losses) leads to a significant performance boost (a +0.9% gain in I-AUC on MVTec AD). This strongly demonstrates that the core advantage of our method stems from the dynamic, visually-guided composition of experts, rather than merely increasing the number of static prompts. Finally, our `PromptMoE`, benefiting from its multi-layer prompt design and regularization from auxiliary losses, achieves the best performance, fully validating the superiority of our compositional framework.

Configuration	MVTec AD		VisA	
	I-AUC	PRO	I-AUC	PRO
Static Prompt	91.7	82.0	82.4	88.0
+*Static Ensemble*	92.2	82.9	83.3	88.3
+*VGMoP*	93.1	**83.3**	84.1	89.0
`PromptMoE`	**93.8**	83.2	**85.0**	**89.2**

Table 2: Ablation on the effectiveness of compositional prompting strategy.

Synergy of Expert Diversity and Load Balancing. To validate the synergy between our two auxiliary losses, we visualize the evolution of key metrics during training with and without $\mathcal{L}_{\text{decouple}}$ in Fig. 7. The results clearly show

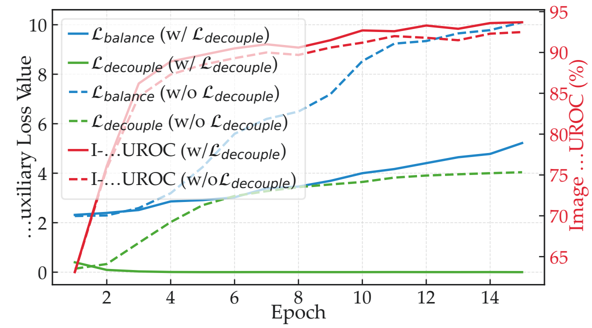

Figure 7: Effect of $\mathcal{L}_{decouple}$ loss on MoE training dynamics.

that in the absence of $\mathcal{L}_{\text{decouple}}$ (dashed lines), expert representations become redundant (reflected by the increasing value of $\mathcal{L}_{\text{decouple}}$), which in turn leads to a sharp degradation of $\mathcal{L}_{\text{balance}}$ and ultimately harms the model's I-AUC performance. Conversely, when both losses are used (solid lines), both auxiliary losses are effectively suppressed to healthy levels, and the model achieves superior final performance. This strongly demonstrates that the expert diversity enforced by $\mathcal{L}_{\text{decouple}}$ is a crucial prerequisite for achieving effective load balancing and optimal performance.

Analysis on the Load Balancing Loss. To further validate the necessity of $\mathcal{L}_{\text{balance}}$, we conduct an ablation study on its coefficient α. As shown in Table 3, completely removing the load balancing loss ($\alpha = 0$) leads to a significant performance drop. A moderate weight ($\alpha = 0.01$) effectively promotes expert balancing and yields the best performance. However, an excessively high weight ($\alpha = 0.1$) might over-penalize the router's specialization, which also negatively impacts the main task's performance.

Weight α	MVTec AD		VisA	
	I-AUC	PRO	I-AUC	PRO
$\alpha = 0$ (w/o $\mathcal{L}_{\text{balance}}$)	92.1	82.5	83.4	88.2
$\alpha = 0.01$	**93.8**	**83.2**	**85.0**	**89.2**
$\alpha = 0.1$	92.9	83.1	84.2	**89.2**

Table 3: Ablation study on the weight α of $\mathcal{L}_{\text{balance}}$.

Conclusion

In this paper, we have demonstrated that reframing prompt learning as a compositional paradigm is an effective approach to tackling the generalization challenge in ZSAD. Our proposed `PromptMoE`, through its visually-guided MoE mechanism, successfully learns to dynamically construct tailored textual prompts for each instance from a basis set of learnable semantic primitives. This approach overcomes the representational bottlenecks and overfitting risks associated with monolithic prompts. The SOTA performance achieved across 15 industrial and medical datasets strongly supports our central conclusion: learning how to compose, rather than learning a single complete prompt, is key to enhancing a model's ability to handle the diversity of unseen anomalies.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (2022YFB4501704); the National Science Foundation of China (62222312, 62473285); the Shanghai Science and Technology Innovation Action Plan Project of China (22511100700); and the Fundamental Research Funds for the Central Universities.

References

Aota, T.; Tong, L. T. T.; and Okatani, T. 2023. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5564–5572.

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.

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.

Buda, M.; Saha, A.; and Mazurowski, M. A. 2019. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology and Medicine*, 109: 218–225.

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. Springer.

Chen, X.; Han, Y.; and Zhang, J. 2023a. 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*.

Chen, X.; Han, Y.; and Zhang, J. 2023b. 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*, 2305: 17382.

Codella, N. C.; Gutman, D.; Celebi, M. E.; Helba, B.; Marchetti, M. A.; Dusza, S. W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. 2018. 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)*, 168–172. IEEE.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

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*, 1932–1940.

Hamada, A. 2020. Br35H :: Brain Tumor Detection 2020. Kaggle.

Hicks, S. A.; Jha, D.; Thambawita, V.; Halvorsen, P.; Hammer, H. L.; and Riegler, M. A. 2021. The EndoTect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In *International Conference on Pattern Recognition*, 263–274. Springer.

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.

Jezek, S.; Jonak, M.; Burget, R.; Dvorak, P.; and Skotak, M. 2021. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 66–71. IEEE.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; De Lange, T.; Johansen, D.; and Johansen, H. D. 2019. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, 451–462. Springer.

Kanade, P. B.; and Gumaste, P. 2015. Brain tumor detection using MRI images. *Brain*, 3(2): 146–150.

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.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571. IEEE.

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 *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, 01–06. IEEE.

Qu, Z.; Tao, X.; Gong, X.; Qu, S.; Chen, Q.; Zhang, Z.; Wang, X.; and Ding, G. 2025. Bayesian Prompt Flow Learning for Zero-Shot Anomaly Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 30398–30408.

Qu, Z.; Tao, X.; Prasad, M.; Shen, F.; Zhang, Z.; Gong, X.; and Ding, G. 2024. VCP-CLIP: A visual context prompting model for zero-shot anomaly segmentation. In *European Conference on Computer Vision*, 301–317. Springer.

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. PMLR.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595.

Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M. H.; and Rabiee, H. R. 2021. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14902–14912.

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*.

Tabernik, D.; Šela, S.; Skvarč, J.; and Skočaj, D. 2020. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3): 759–776.

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

Wang, C.; Zhu, W.; Gao, B.-B.; Gan, Z.; Zhang, J.; Gu, Z.; Qian, S.; Chen, M.; and Ma, L. 2024. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22883–22892.

Wieler, M.; and Hahn, T. 2007. Weakly supervised learning for industrial optical inspection. In *DAGM Symposium in*, volume 6, 11.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

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

Zhu, J.; Ong, Y.-S.; Shen, C.; and Pang, G. 2025. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22241–22251.

Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; and Dabeer, O. 2022. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, 392–408. Springer.