VarProtoAD: Variational Prototype-Conditioned Prompting for Zero-Shot Anomaly Detection

Mengyang Zhao[1], Zhuolin He[1], Haiyang Yu[1], Teng Fu[1], Ke Niu[1], and Xiangyang Xue[1](✉)

College of Computer Science and Artificial Intelligence, Fudan University, 200438 Shanghai, China {myzhao20, hyyu20, xyxue}@fudan.edu.cn {zlhe22, tfu23, kniu22}@m.fudan.edu.cn

Abstract. Zero-shot anomaly detection (ZSAD) requires detecting and localizing diverse defects in unseen categories. Recent vision-language model (VLM) based approaches improve ZSAD via prompt learning, yet their prompts are often weakly grounded to transferable visual evidence, making them prone to semantic drift under domain shifts and limiting the reuse of normal/abnormal primitives. We propose a Variational Prototype-conditioned prompting framework (**VarProtoAD**) that explicitly conditions prompt learning on visual prototypes discovered in the feature space. Specifically, multi-layer patch tokens are ℓ_2-normalized and modeled with a variational von Mises–Fisher mixture to learn layer-wise normal and abnormal prototype banks. These prototypes form multi-cluster semantic anchors on the hypersphere and are naturally aligned with cosine-based VLM matching, where anomalies correspond to directional deviations. Conditioned on these anchors, learnable prompt context tokens interact with the prototype banks via confidence-gated cross-attention, producing explicit prototype-conditioned prompts as well as implicit deviation prompts derived from normal prototypes to better cover unseen anomaly semantics. Finally, multi-layer image–text similarities are fused within a unified framework to jointly produce image-level anomaly scores and pixel-level anomaly maps. Experiments on 15 industrial and medical datasets demonstrate strong cross-domain performance, indicating that VarProtoAD yields more stable and robust prompt learning.

Keywords: Anomaly Detection · Zero-shot Learning · Prompt Learning

1 Introduction

Zero-shot anomaly detection (ZSAD) aims to detect and localize anomalies in unseen categories, without requiring any image of that category for training. This task is central to industrial inspection [13,14,26,29,42] and medical imaging [30,41] due to pervasive cold-start scenarios [6] and expensive defect annotation when encountering novel categories. Robust generalization remains challenging under domain shifts, where anomalies are subtle and open-ended, and normal patterns vary across categories and datasets.

Recently, vision-language models (VLMs), such as CLIP [33] and ALIGN [19], have been widely adopted for ZSAD because of their powerful generalization capability. Consequently, a common paradigm [17,41] performs cross-modal matching by comparing visual embeddings with text embeddings derived from "normal" and "abnormal" prompts. However, ZSAD focuses on subtle local defects, but standard VLMs mainly capture coarse, global semantics. As a result, it often misses small local anomalies, leading to weak localization and unstable performance under domain shifts.

Motivated by this mismatch, recent VLM-based ZSAD methods have increasingly focused on prompt learning. AnomalyCLIP [41] learns object-agnostic normal/abnormal prompts and optimizes them with both image- and pixel-level supervision (Fig. 1a). AdaCLIP [6] improves adaptability by combining image-agnostic static prompts with instance-specific dynamic prompts (Fig. 1b). Concretely, its dynamic prompts are generated from global image features (e.g., the class token) via a lightweight generator, echoing instance-conditioned prompt learning as in CoCoOp [39]. More recently, Bayes-PFL [32] models prompts as a learnable distribution and samples diverse prompts to better cover the prompt space. Despite these advances, a key limitation remains: prompt updates are only weakly grounded in transferable local evidence. Optimizing prompts mainly in an unconstrained text space, or conditioning them on coarse image-level features, provides little fine-grained regularization across domains, leading to semantic drift under domain shifts and hindering the discovery of reusable normal/anomalous primitives.

Building on this observation, we propose a Variational Prototype-conditioned prompting framework for ZSAD (**VarProtoAD**), which grounds prompt learning on fine-grained visual prototypes discovered in the frozen VLM feature space. Unlike prior conditioning schemes that rely on global image-level features, we instead extract transferable visual evidence from multi-layer patch embeddings to provide fine-grained anchors for prompting (Fig. 1c). Since cross-modal matching in VLMs is often performed with cosine similarity in an ℓ_2-normalized embedding space, these patch tokens can be treated as directional data on a unit hypersphere. Building on this geometry, we introduce a prototype learning module that fits a variational von Mises–Fisher (vMF) mixture [2, 11] to the normalized patch tokens, yielding stable layer-wise normal and abnormal prototype banks as transferable semantic anchors. To inject these anchors into prompt learning, we propose a confidence-gated cross-attention mechanism that enables learnable context tokens to selectively retrieve reliable prototype cues. Ultimately, we obtain explicit normal/abnormal prompts conditioned on normal/abnormal prototypes, together with implicit deviation abnormal prompts transformed from normal prototypes to broaden coverage of unseen anomaly semantics.

Overall, our main contributions are as follows:

- We propose a prototype-anchored prompting paradigm for ZSAD that conditions prompt learning on fine-grained, patch-level visual evidence via layer-wise prototype banks, rather than coarse global feature conditioning.

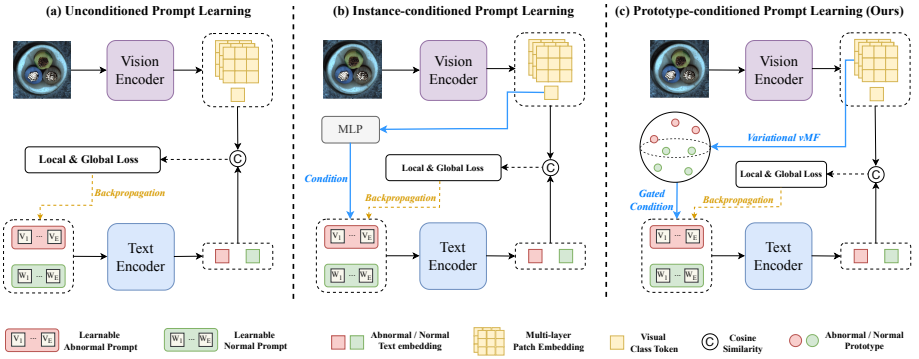

Fig. 1: Comparison of VLM-based prompting paradigms for ZSAD. (a) Tuning optimizes context tokens strictly in the text space, lacking explicit visual grounding. (b) Instance-conditioned prompt learning using global visual features, providing only coarse cues. (c) Guiding prompt learning with fine-grained visual prototypes, which are learned via a vMF mixture from dense patch embeddings.

- We introduce a variational von Mises–Fisher mixture for learning reusable hyperspherical normal/abnormal prototypes from patch embeddings.
- We design a confidence-gated cross-attention module that injects prototype evidence into prompt contexts in a selective manner.
- VarProtoAD achieves superior performance on 15 industrial and medical benchmarks, indicating the stable and generalization of our prompt learning strategy.

2 Related Work

Zero-shot anomaly detection with VLMs. Zero-shot anomaly detection (ZSAD) targets anomaly classification and localization on unseen categories, thus requiring strong cross-category generalization, especially under cross-dataset domain shifts. With vision-language models such as CLIP [33] and ALIGN [19], a prevalent paradigm performs cosine-based matching between visual embeddings (image-level and patch-level) and text embeddings derived from prompts describing normality/abnormality. WinCLIP [17] is a representative VLM-based baseline that achieves zero-/few-shot anomaly classification and segmentation by prompt engineering and patch-level similarity aggregation. APRIL-GAN [7] ensembles normal/abnormal prompt templates to form text embeddings, whereas CLIP-AD [8] samples multiple prompts and selects representative text embeddings for matching. Beyond handcrafted prompts, AnomalyCLIP [41] learns object-agnostic normal/abnormal context tokens with image- and pixel-level supervision, AdaCLIP [6] further combines shared static prompts with instance-specific dynamic prompts conditioned on global image features. Bayes-PFL [32] models prompts as a learnable distribution and samples diverse prompts to enlarge semantic coverage. While effective, these methods mainly adapt prompts in

the text space or condition them on coarse global cues, which provides limited explicit grounding to transferable fine-grained normal/abnormal evidence and may lead to semantic drift under domain shifts. This leaves an open question of how to ground prompt adaptation in reusable, fine-grained visual primitives that remain stable across datasets.

Prompt Learning. Early VLM-based adaptation often relies on manual prompt engineering, where hand-crafted templates are used to construct textual descriptions for downstream tasks [12, 17, 38]. Prompt learning replaces such fixed templates with trainable context tokens. CoOp [40] is a canonical formulation that learns continuous context tokens in text space. Inspired by CoOp, AnomalyCLIP [41] learns context tokens $[V_1]\cdots[V_z]$, which are concatenated with fixed template words $[state][object]$ to form prompts and optimized with image- and pixel-level supervision for ZSAD. CoCoOp [39] further introduces instance-conditioned prompting to improve generalization by adapting prompts based on image features. Following the idea of CoCoOp in the ZSAD setting, AdaCLIP [6] conditions text-prompt learning on global image features to generate instance-specific dynamic prompts, aiming to improve cross-category transfer under domain shifts. Our method also learns context tokens, but conditions them on fine-grained patch-level evidence via prototype representations.

3 Preliminaries

3.1 Zero-shot Anomaly Detection

Zero-shot anomaly detection aims to detect and localize anomalies in previously unseen target categories. Following recent VLM-based protocols [6, 32, 41], we assume access to an auxiliary training set $\mathcal{D}_{\text{aux}} = \{(I, y, \mathbf{M})\}$, where $y \in \{0, 1\}$ indicates whether image I is normal or anomalous, and $\mathbf{M} \in \{0, 1\}^{H\times W}$ is the pixel-level anomaly mask available during training. At test time, the model is evaluated on target datasets with different unseen categories. The output is an image-level anomaly score $s(I)$ and a pixel-level anomaly map $A(I) \in [0, 1]^{H\times W}$.

3.2 VLM-based ZSAD

Existing VLM-based ZSAD pipelines [13,17] usually leverage pre-trained VLMs to compare visual representations (global and dense) against text embeddings derived from normality/abnormality text prompts. VLMs provide an image encoder E_v and a text encoder E_t. Given an input image I, the vision encoder can extract a set of visual embeddings consisting of a global image embedding $\mathbf{v} \in \mathbb{R}^D$ and intermediate dense patch embeddings $\mathbf{f}_i^{(l)} \in \mathbb{R}^D$:

$$\mathbf{v},\ \{\mathbf{F}^{(l)}\}_{l\in\mathcal{L}} = E_v(I), \qquad \mathbf{F}^{(l)} = [\mathbf{f}_1^{(l)}; \ldots; \mathbf{f}_N^{(l)}], \tag{1}$$

where $\mathcal{L}$ denotes a set of selected intermediate layers. On the text side, a prompt is encoded by E_t to obtain a text embedding $\mathbf{t} \in \mathbb{R}^D$. CLIP compares a visual

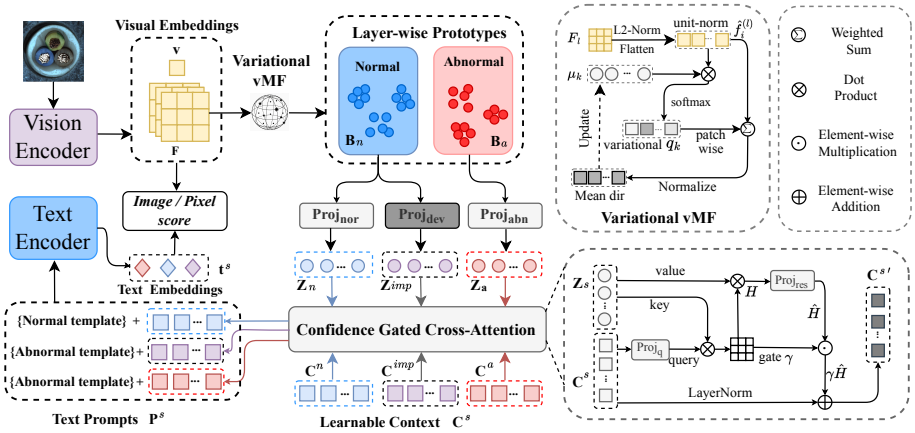

Fig. 2: VarProtoAD overall pipeline: layer-wise prototypes anchor prompt conditioning, and fused multi-layer logit margins yield image-level scores and pixel-level anomaly maps.

token $\mathbf{u} \in \mathbb{R}^D$ with a text embedding by cosine similarity with temperature τ:

$$\text{sim}(\mathbf{u}, \mathbf{t}) = \frac{\langle \hat{\mathbf{u}}, \hat{\mathbf{t}} \rangle}{\tau}, \qquad \hat{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|_2}, \ \hat{\mathbf{t}} = \frac{\mathbf{t}}{\|\mathbf{t}\|_2}. \tag{2}$$

By contrasting similarities between visual embeddings and normal/ abnormal text prompts, VLM-based ZSAD methods can derive the image- and pixel-level anomaly score.

4 Method

4.1 Overview: Prototype-Conditioned Prompting

Recent VLM-based approaches improve ZSAD via prompt learning, yet prompts are often weakly grounded in transferable visual evidence, making them prone to semantic drift under domain shifts [6, 32, 41]. To address this, we propose VarProtoAD, which anchors prompt learning with explicit prototypes discovered in the frozen visual feature space.

As shown in Fig 2, VarProtoAD consists of three components: (i) layer-wise normal/abnormal prototype banks are learned from intermediate patch embeddings via a variational von Mises–Fisher mixture to provide reusable visual anchors (§4.2); (ii) prompt context tokens are conditioned on these anchors through confidence-gated cross-attention to obtain prototype-conditioned prompts with improved stability under domain shifts (§4.3); and (iii) multi-layer image-text similarities are fused to produce image- and pixel-level anomaly scores (§4.5).

4.2 Layer-wise Prototype Banks via vMF Mixtures

To reduce semantic drift in prompt learning due to weak visual evidence, we explicitly extract a compact set of reusable local visual primitives from frozen vision features and use them as anchors for subsequent prompt conditioning. Unlike global representations, dense patch embeddings capture recurring local textures, parts, and defect-related patterns across categories and domains [41]. Moreover, patch embeddings from different depths provide hierarchical cues, ranging from low-level textures to higher-level semantic patterns. Such multi-layer features are widely used in anomaly detection to enable robust localization [10, 24], motivating our layer-wise prototype banks.

Most modern VLMs perform cross-modal matching via cosine similarity (cf. Eq. (2)), where the similarity is mainly determined by the direction of ℓ_2-normalized embeddings. Consequently, the decision boundary is governed by angular distances, and patch tokens can be viewed as directional observations on the unit hypersphere. This makes hyperspherical clustering a geometry-consistent way to summarize recurring local patterns. Building on this insight, we introduce a variational vMF prototype learner that discovers layer-wise normal/abnormal prototype banks directly in the frozen feature space. Unlike hard-assignment clustering, our variational formulation produces temperature-controlled soft responsibilities and includes an explicit entropy term, which stabilizes prototype discovery and yields a compact set of reusable anchors.

Patch sets. For each selected layer $l \in \mathcal{L}$, patch tokens are collected from auxiliary data and form two sets, $\mathcal{S}_n^{(l)}$ and $\mathcal{S}_a^{(l)}$, corresponding to normal and abnormal regions, respectively. Each patch token is ℓ_2-normalized:

$$\hat{\mathbf{f}}_i^{(l)} = \frac{\mathbf{f}_i^{(l)}}{\|\mathbf{f}_i^{(l)}\|_2} \in \mathbb{R}^D, \qquad \|\hat{\mathbf{f}}_i^{(l)}\|_2 = 1. \tag{3}$$

Thus $\hat{\mathbf{f}}_i^{(l)}$ lies on the unit hypersphere $\mathbb{S}^{D-1} = \{\mathbf{z} \in \mathbb{R}^D : \|\mathbf{z}\|_2 = 1\}$, where $\mathbb{S}^{D-1}$ denotes the unit sphere in $\mathbb{R}^D$.

Prototype banks. We maintain two banks per layer: $\{\boldsymbol{\mu}_{n,k}^{(l)}\}_{k=1}^{K_n}$ for normal prototypes and $\{\boldsymbol{\mu}_{a,k}^{(l)}\}_{k=1}^{K_a}$ for abnormal prototypes, with unit-norm constraints $\|\boldsymbol{\mu}\|_2 = 1$.

Each prototype $\boldsymbol{\mu}$ is a representative direction on the hypersphere capturing a dominant cluster of patch embeddings, providing reusable reference points to condition prompt updates and mitigate drift under domain shifts.

Variational vMF mixture. A vMF component on $\mathbb{S}^{D-1}$ has density $p(\hat{\mathbf{f}} \mid z{=}k) = C_D(\kappa)\exp(\kappa \boldsymbol{\mu}_k^\top \hat{\mathbf{f}})$, where $\boldsymbol{\mu}_k \in \mathbb{S}^{D-1}$ is the mean direction and $\kappa > 0$ is the concentration. Given mixture weights $\boldsymbol{\pi}$ with $\pi_k \geq 0$ and $\sum_k \pi_k = 1$, the unnormalized log-responsibility (logit) for component k is

$$e_k(\hat{\mathbf{f}}) = \kappa\, \boldsymbol{\mu}_k^\top \hat{\mathbf{f}} + \log \pi_k, \tag{4}$$

which equals $\log p(z{=}k, \hat{\mathbf{f}})$ up to the additive constant $\log C_D(\kappa)$ (shared across k when κ is fixed). We fix κ during training for stability and to avoid optimizing the κ-dependent normalization term. We then adopt a tempered variational

assignment as follows:

$$q_k(\hat{\mathbf{f}}) = \frac{\exp(e_k(\hat{\mathbf{f}})/\tau_q)}{\sum_{j=1}^{K} \exp(e_j(\hat{\mathbf{f}})/\tau_q)}, \tag{5}$$

We use a tempered responsibility $q = \mathrm{softmax}(e/\tau_q)$ to stabilize training, where τ_q controls the sharpness of the assignments; when $\tau_q = 1$, it reduces to the standard entropy-regularized soft assignment consistent with Eq. (6).

Objective and optimization. For a patch set $\mathcal{S}$ (normal or abnormal), we minimize the following entropy-regularized variational objective:

$$\mathcal{L}_{\mathrm{vmf}}(\mathcal{S}) = -\frac{1}{|\mathcal{S}|} \sum_{\hat{\mathbf{f}} \in \mathcal{S}} \sum_{k} q_k(\hat{\mathbf{f}}) \Big(e_k(\hat{\mathbf{f}}) - \log q_k(\hat{\mathbf{f}}) \Big). \tag{6}$$

We optimize Eq. (6) with mini-batch stochastic gradient optimization, where q serves as a soft assignment that guides prototype updates.

Finally, to learn more discriminative and diverse prototypes, we additionally introduce a simple separation loss $\mathcal{L}_{\mathrm{sep}}$ to encourage normal and abnormal prototype banks to be well separated, and an orthogonality loss $\mathcal{L}_{\mathrm{orth}}$ to promote diversity within each bank. The overall prototype loss is

$$\mathcal{L}_{\mathrm{proto}} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \Big(\mathcal{L}_{\mathrm{vmf}}(\mathcal{S}_n^{(l)}) + \mathcal{L}_{\mathrm{vmf}}(\mathcal{S}_a^{(l)}) \Big) + \lambda_{\mathrm{sep}} \mathcal{L}_{\mathrm{sep}} + \lambda_{\mathrm{orth}} \mathcal{L}_{\mathrm{orth}}. \tag{7}$$

We defer the detailed definitions of $\mathcal{L}_{\mathrm{sep}}$ and $\mathcal{L}_{\mathrm{orth}}$ to the supplementary material.

4.3 Prototype-Conditioned Prompt Ensemble

Given the prototype banks learned in Sec. 4.2, we define a prototype-conditioned prompt ensemble for ZSAD. The ensemble updates learnable context tokens by querying the prototype banks and injecting prototype-derived residuals, such that prompt representations are anchored to reusable visual primitives rather than being optimized in a weakly constrained text space.

Learnable prompt contexts. An object-agnostic two-state prompting scheme is employed same as AnomalyCLIP [41], consisting of normal and abnormal states, with an ensemble of M prompts maintained for each state.

For each state s and prompt index m, L_c learnable context tokens $\mathbf{C}_m^s \in \mathbb{R}^{L_c \times D}$ are introduced. A prompt is assembled by inserting these context tokens between a fixed start token and a fixed suffix:

$$\mathbf{P}_m^s = [\mathbf{p}_{\mathrm{sot}};\ \mathbf{C}_m^{s\prime};\ \mathbf{p}_{\mathrm{suffix}}], \qquad \mathbf{t}_m^s = E_t(\mathbf{P}_m^s). \tag{8}$$

Here $\mathbf{C}_m^{s\prime}$ denotes the prototype-conditioned context tokens described below.

Prototype-to-prompt conditioning. Prototypes are concatenated into two banks, $\mathbf{B}_n$ and $\mathbf{B}_a$, and projected via MLPs into conditioning sets $\mathbf{Z}_n$ and $\mathbf{Z}_a$.

Since explicit anomalies ($\mathbf{Z}_a$) may not fully cover open-ended defects, we also derive an implicit abnormal set $\mathbf{Z}_{\text{imp}}$ by transforming normal prototypes:

$$\mathbf{Z}_n = \text{Proj}_{\text{nor}}(\mathbf{B}_n), \quad \mathbf{Z}_a = \text{Proj}_{\text{abn}}(\mathbf{B}_a), \quad \mathbf{Z}_{\text{imp}} = \text{Proj}_{\text{dev}}(\mathbf{B}_n). \tag{9}$$

Consequently, normal prompts are conditioned on $\mathbf{Z}_n$, whereas abnormal prompts are conditioned on either $\mathbf{Z}_a$ or $\mathbf{Z}_{\text{imp}}$, improving semantic coverage.

Confidence-gated cross-attention. Given a conditioning set $\mathbf{Z} \in \{\mathbf{Z}_n, \mathbf{Z}_a, \mathbf{Z}_{\text{imp}}\}$, the context tokens $\mathbf{C}_m^s$ for the corresponding state $s \in \{n, a, \text{imp}\}$ are projected into queries and updated via cross-attention. To improve stability, queries and keys are ℓ_2-normalized before the scaled dot-product:

$$\mathbf{Q}_m^s = \text{Proj}_q(\mathbf{C}_m^s), \qquad \mathbf{A}_m^s = \text{softmax}\Big(\tau_{\text{attn}} \tilde{\mathbf{Q}}_m^s \tilde{\mathbf{Z}}^\top\Big), \qquad \mathbf{H}_m^s = \mathbf{A}_m^s \mathbf{Z}, \tag{10}$$

where $\text{Proj}_q(\cdot)$ denotes a projection mapping, $\tilde{(\cdot)}$ denotes ℓ_2-normalization, τ_{attn} is a learnable temperature scale, and $\mathbf{Z}$ provides keys and values. The attention output $\mathbf{H}_m^s$ aggregates prototype evidence for each context token.

To suppress noisy updates under diffuse attention, a token-wise confidence gate $\gamma_{m,j}^s \in [0, 1]$ is derived from the maximum attention probability. Let $\mathbf{a}_{m,j}^s \in \mathbb{R}^K$ denote the j-th row of $\mathbf{A}_m^s$, representing the attention distribution of the j-th context token. The confidence score is defined as $c_{m,j}^s = \max_k(\mathbf{a}_{m,j}^s)_k$, and the gate is formulated as:

$$\gamma_{m,j}^s = \sqrt{\frac{\max(0, c_{m,j}^s - 1/K)}{1 - 1/K + \epsilon}}, \tag{11}$$

where $1/K$ represents the expected confidence of a uniform distribution and ϵ ensures numerical stability. Residual injection is then applied token-wise:

$$\mathbf{C}_m^{s\prime} = \text{LN}(\mathbf{C}_m^s) + \boldsymbol{\gamma}_m^s \odot \text{Proj}_{\text{res}}(\mathbf{H}_m^s), \tag{12}$$

where $\text{LN}(\cdot)$ denotes Layer Normalization, $\boldsymbol{\gamma}_m^s$ groups the token-wise gates, and $\odot$ is broadcasted element-wise multiplication. To ensure training stability, the attention output $\mathbf{H}_m^s$ is refined by a MLP $\text{Proj}_{\text{res}}(\cdot)$, which comprises residual projections and layer normalizations. This gating strategy promotes strong prototype conditioning under confident matching while preserving the original context representation when attention remains uncertain.

Explicit and implicit prompts. By feeding the prototype-conditioned prompts into the text encoder E_t, we obtain the corresponding text embeddings. Specifically, conditioning the normal state on $\mathbf{Z}_n$ yields $\mathbf{t}_m^n$. For the abnormal state, conditioning on $\mathbf{Z}_a$ and $\mathbf{Z}_{\text{imp}}$ yields the explicit text embedding $\mathbf{t}_m^{a,\text{exp}}$ and implicit text embedding $\mathbf{t}_m^{a,\text{imp}}$, respectively. Thus, three prompt sets are constructed:

$$\mathcal{T}_n = \{\mathbf{t}_m^n\}_{m=1}^M, \quad \mathcal{T}_{\text{exp}} = \{\mathbf{t}_m^{a,\text{exp}}\}_{m=1}^M, \quad \mathcal{T}_{\text{imp}} = \{\mathbf{t}_m^{a,\text{imp}}\}_{m=1}^M, \tag{13}$$

and the complete abnormal prompt set is defined as $\mathcal{T}_a = \mathcal{T}_{\text{exp}} \cup \mathcal{T}_{\text{imp}}$. The abnormal prompt space integrates prototype-grounded abnormal semantics from $\mathcal{T}_{\text{exp}}$ with deviation-based representations derived from $\mathcal{T}_{\text{imp}}$, resulting in more stable and semantically enriched text embeddings.

4.4 Training Objectives

We train VarProtoAD on the auxiliary dataset $\mathcal{D}_{\text{aux}} = \{(I, y, \mathbf{M})\}$ with image-level labels $y \in \{0, 1\}$ and pixel-level masks $\mathbf{M}$.

Prompt-ensemble logits. For any visual embedding $\mathbf{u}$, we define a state logit by softly aggregating similarities over the prompt set:

$$z_s(\mathbf{u}) = \log\left(\frac{1}{|\mathcal{T}_s|}\sum_{\mathbf{t}\in\mathcal{T}_s}\exp\big(\text{sim}(\mathbf{u},\mathbf{t})\big)\right), \qquad s \in \{n, a\}, \tag{14}$$

where $\text{sim}(\mathbf{u}, \mathbf{t})$ is the similarity defined in Eq. (2). We then use the logit margin

$$\Delta(\mathbf{u}) = z_a(\mathbf{u}) - z_n(\mathbf{u}) \tag{15}$$

as a unified anomaly evidence signal.

Patch-level loss. For each selected layer $l \in \mathcal{L}$ and patch token $\mathbf{f}_i^{(l)}$, we compute the logit margin $\Delta(\mathbf{f}_i^{(l)})$ (Eq. (15)). During training, we fuse multi-layer patch evidence by simple averaging:

$$\Delta_i^{\text{patch}} = \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}}\Delta(\mathbf{f}_i^{(l)}). \tag{16}$$

We supervise the patch logits Δ_i^{patch} with a combination of weighted binary cross-entropy loss [34] and a Dice loss [25].

Image-level loss. We form image-level logits from global embedding $\mathbf{v}$ as $\big[z_n(\mathbf{v}),\, z_a(\mathbf{v})\big]$ and optimize a standard two-class cross-entropy loss. We further include an MIL loss derived from patch logits to encourage consistency between detection and localization.

The full training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{img}} + \lambda_{\text{pix}}\mathcal{L}_{\text{pix}} + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}}. \tag{17}$$

More details about training objective and optimization strategy refer to supplementary material.

4.5 Inference Stage

Pixel-level anomaly map. Different layers may emphasize different defect patterns. We therefore aggregate multi-layer patch logit margins with a temperature-controlled smooth maximum. Define a generic smooth aggregation operator

$$\phi_{\tau_a}\big(\{x_j\}_{j=1}^{J}\big) = \tau_a \log\left(\frac{1}{J}\sum_{j=1}^{J}\exp(x_j/\tau_a)\right). \tag{18}$$

For each patch index i, we fuse layer-wise margins by

$$\Delta_i^{\text{fuse}} = \phi_{\tau_a}\Big(\{\Delta(\mathbf{f}_i^{(l)})\}_{l\in\mathcal{L}}\Big), \tag{19}$$

and obtain the pixel-level anomaly map by converting fused logits to probabilities and resizing:

$$A(I) = \mathrm{Upsample}\Big(\mathrm{Reshape}\big(\{\sigma(\Delta_i^{\mathrm{fuse}})\}_{i=1}^{N}\big)\Big) \in [0,1]^{H\times W}, \tag{20}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

Image-level anomaly score. Using the same operator, we pool fused patch logits into an image score

$$s_{\mathrm{patch}}(I) = \phi_{\tau_a}\left(\{\Delta_i^{\mathrm{fuse}}\}_{i=1}^{N}\right), \qquad s(I) = s_{\mathrm{patch}}(I) + \lambda\,\Delta(\mathbf{v}), \tag{21}$$

where τ_a is shared for both layer fusion and patch pooling, and λ balances localized and global evidence.

5 Experiments

5.1 Datasets

We evaluate ZSAD on 15 real-world datasets from industrial inspection and medical imaging. The industrial domain includes seven benchmarks: MVTec-AD [3], VisA [23], BTAD [31], KSDD [35], KSDD2 [5], DAGM [37], and DTD-Synthetic [1]. The medical domain includes eight datasets: HeadCT [22], Brain-MRI [20], Br35H [15] (brain tumor classification), ColonDB [36], ClinicDB [4], Endo [16], Kvasir [18] (colon polyp segmentation), and ISIC [9] (skin lesion segmentation).

Following [6, 41], we perform auxiliary training on labeled dataset and directly evaluate on the remaining datasets without additional training. To avoid category leakage, the auxiliary dataset should not share categories with the target test set. Although ClinicDB, ColonDB, Endo, and Kvasir are contain colon polyp images, their visual distributions differ substantially. Thus, we follow AdaCLIP [6] and use MVTec-AD (industrial) and ClinicDB (medical) as auxiliary datasets. When evaluating on MVTec-AD and ClinicDB themselves, we instead use VisA and ColonDB for auxiliary training, respectively.

5.2 Evaluation Metrics

We follow AnomalyCLIP [41] for a fair comparison. For image-level detection, we report AUROC and average precision (AP), which are threshold-free and suitable for imbalanced anomaly detection [27, 28]. For pixel-level localization, we report pixel-level AUROC and Per-Region Overlap (PRO), measuring pixel-wise separability and region-level overlap, respectively.

5.3 Implementation Details

We use the public OpenAI CLIP model ViT-L/14@336 [33] as the backbone and resize all images to 518×518 following prior work [6, 32, 41]. We use $M = 12$

Table 1: Comparison with existing state-of-the-art methods for **image-level** anomaly detection. Metrics are reported as AUROC / AP in (%). Bold numbers indicate the best results, while underlined numbers denote the second-best results

Dataset	WinCLIP	APRIL-GAN	AnomalyCLIP	AdaCLIP	Bayes-PFL	Ours
Industrial						
MVTec-AD	91.8 / 95.1	86.1 / 93.5	91.5 / 96.2	92.0 / 96.4	**92.3** / **96.7**	91.0 / 95.8
VisA	78.1 / 77.5	78.0 / 81.4	82.1 / 85.4	83.0 / 84.9	87.0 / 89.2	**90.0** / **91.6**
BTAD	83.3 / 84.1	74.2 / 71.7	89.1 / 91.1	91.6 / 92.4	93.2 / **96.5**	**94.0** / 96.4
KSDD	68.2 / -	96.8 / 92.5	97.8 / **94.2**	97.1 / 89.1	94.1 / 73.3	**97.9** / 94.1
KSDD2	93.5 / 77.9	90.3 / 74.4	92.1 / 77.8	95.9 / 95.9	**97.3** / **97.9**	96.7 / 97.6
DAGM	89.6 / 90.4	90.4 / 90.1	95.6 / 94.6	96.5 / 95.7	**97.7** / 97.0	**97.7** / **97.1**
DTD-Syn.	95.0 / 97.9	83.9 / 93.6	94.5 / 97.7	92.8 / 97.0	95.1 / 98.4	**96.6** / **98.5**
Indust. Avg.	85.6 / -	85.7 / 85.3	91.8 / 91.0	92.7 / 93.1	93.8 / 92.7	**94.8** / **95.9**
Medical						
HeadCT	83.7 / 81.6	89.3 / 89.6	95.3 / 95.2	93.4 / 92.2	**96.5** / 95.5	**96.5** / **96.8**
BrainMRI	92.0 / 90.7	89.6 / 84.5	96.1 / 92.3	94.9 / 94.2	96.2 / 92.4	**96.6** / **96.5**
Br35H	80.5 / 82.2	93.1 / 92.9	97.3 / 96.1	95.7 / 95.7	97.8 / 96.2	**99.0** / **99.0**
Medical Avg.	85.4 / 84.8	90.7 / 89.0	96.2 / 94.5	94.7 / 94.0	96.8 / 94.7	**97.4** / **97.4**
Overall Avg.	85.6 / -	87.2 / 86.4	93.1 / 92.1	93.3 / 93.4	94.7 / 93.3	**95.6** / **96.3**

prompts with context length $L_c = 12$, and use a fixed value $\lambda = 1$ for all benchmarks. For each selected layer, we set the numbers of normal and abnormal prototypes to 96 and 24, respectively. We train the VarProtoAD 50 epochs using Adam [21] with batch size of 48. All experiments are run on a single Tesla H100 GPU with 5 random seeds, results are averaged over categories in each target dataset. More details of the implementation are provided in the supplementary material.

5.4 Comparison to State of the Art

Baselines. We compare our method against representative VLM-based ZSAD methods, including WinCLIP [17], APRIL-GAN [7], AnomalyCLIP [41], AdaCLIP [6], and Bayes-PFL [32].

Quantitative results. Tab. 1 and Tab. 2 summarize the quantitative comparison with representative VLM-based ZSAD methods. In terms of image-level anomaly detection, VarProtoAD achieves the best overall average of 95.6% AUROC and 96.3% AP, outperforming the strongest competitor (Bayes-PFL) by $+$**0.9%** AUROC and $+$**3.0%** AP, with consistent gains on both industrial (94.8%/95.9%) and medical (97.4%/97.4%) benchmarks. In terms of pixel-level anomaly localization, VarProtoAD further delivers the best average localization performance (94.8% pixel AUROC / 87.2% PRO), improving over Bayes-PFL by $+$**2.0%** AUROC and $+$**2.6%** PRO; notably, the medical-domain average rises from 87.7%/76.2% to 92.2%/81.3%, indicating sharper and more reliable anomaly maps under cross-dataset shifts.

Qualitative comparison. Fig. 3 presents a qualitative comparison of anomaly localization by visualizing the predicted maps of our method and contrasting

Table 2: Comparison with existing state-of-the-art methods for **pixel-level** anomaly localization. Metrics are reported as AUROC / PRO in (%). Bold numbers indicate the best results, while underlined numbers denote the second-best results

Dataset	WinCLIP	APRIL-GAN	AnomalyCLIP	AdaCLIP	Bayes-PFL	Ours
Industrial						
MVTec-AD	85.1 / 64.6	87.6 / 44.0	91.1 / 81.4	86.8 / 33.8	**91.8** / **87.4**	91.1 / 85.9
VisA	79.6 / 56.8	94.2 / 86.8	95.5 / 87.0	95.1 / 71.3	95.6 / 88.9	**96.3** / **90.1**
BTAD	71.4 / 32.8	91.3 / 23.0	93.3 / 69.3	87.7 / 17.1	93.9 / 76.6	**94.2** / **78.3**
KSDD	95.8 / -	92.9 / 84.3	98.1 / **94.9**	97.7 / 36.1	96.5 / 91.1	**98.5** / 94.8
KSDD2	97.9 / 91.2	97.9 / 51.1	99.4 / 92.7	96.1 / 70.8	**99.6** / 97.6	**99.6** / **98.6**
DAGM	83.2 / 55.4	99.2 / 44.7	99.1 / 93.6	97.0 / 40.9	**99.3** / **98.0**	97.7 / 97.1
DTD-Syn.	82.5 / 55.4	96.6 / 41.6	97.6 / 88.3	94.1 / 24.9	97.8 / 94.3	**98.7** / **95.1**
Indust. Avg.	85.1 / -	94.2 / 53.6	96.3 / 86.7	93.5 / 42.1	96.4 / 90.6	**96.6** / **91.4**
Medical						
ISIC	83.3 / 55.1	85.8 / 13.7	88.4 / 78.1	85.4 / 5.3	**92.2** / **87.6**	**92.2** / 86.6
ColonDB	64.8 / 28.4	78.4 / 28.0	81.9 / 71.4	79.3 / 6.5	82.1 / 76.1	**88.3** / **84.7**
ClinicDB	70.7 / 32.5	86.0 / 41.2	85.9 / 69.6	84.3 / 14.6	89.6 / 78.4	**93.0** / **85.7**
Endo	68.2 / 28.3	84.1 / 32.3	86.3 / 67.3	84.0 / 10.5	89.2 / 74.8	**94.6** / **88.9**
Kvasir	69.8 / 31.0	80.2 / 27.1	81.8 / 53.8	79.4 / 12.3	85.4 / **63.9**	**93.1** / 60.5
Medical Avg.	71.4 / 35.1	82.9 / 28.5	84.9 / 68.0	82.5 / 9.8	87.7 / 76.2	**92.2** / **81.3**
Overall Avg.	79.4 / -	89.5 / 43.2	91.5 / 79.0	88.9 / 28.7	92.8/ 84.6	**94.8** / **87.2**

them with AnomalyCLIP [41] and Bayes-PFL [32]. Across the shown examples, VarProtoAD tends to yield more concentrated anomaly responses that align better with the annotated defect regions, while producing fewer diffuse activations on normal structures. In addition, compared with the prompt-based baselines, the predicted maps often show reduced spurious responses on visually salient yet non-defective areas, and exhibit relatively clearer boundaries in several cases. These observations suggest that the proposed prototype-conditioned prompting can provide more stable localization results under cross-dataset evaluation.

5.5 Ablation Studies

In this section, ablation experiments are conducted on the VisA [23] and ColonDB [36] datasets to further investigate the impact of the proposed VarProtoAD.

Ablation on prototype extraction. To verify the effectiveness of our method, we evaluate different prototype extraction strategies: (i) No Prototype (prompt-only, without a prototype bank), (ii) K-means Prototype (K-means cluster centroid used as prototypes), and (iii) Variational vMF Prototype (ours). All variants share the same backbone, prompt settings, and training schedule; only the prototype extraction module differs. As shown in Tab. 3, introducing visual prototypes yields consistent improvements in both anomaly detection and localization on VisA and ColonDB.

On VisA, replacing prompt-only with K-means prototypes improves the image-level score from 85.4%/87.0% to 89.1%/90.3%, and slightly improves pixel-level performance from 95.2%/86.8% to 95.5/87.6. Using variational vMF prototypes

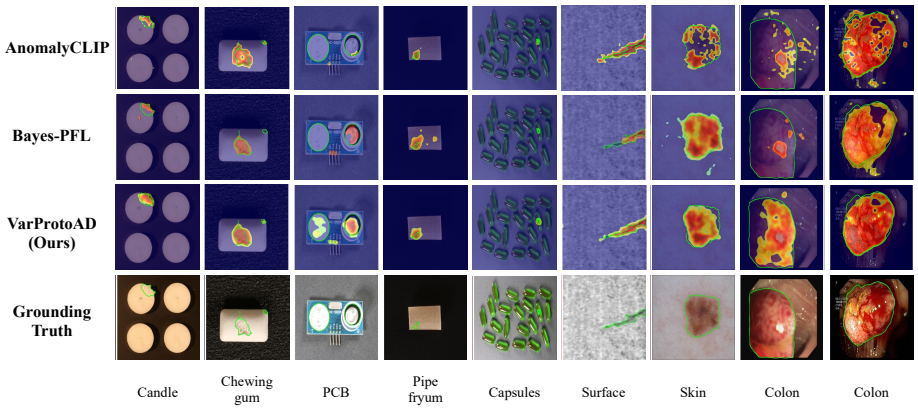

Fig. 3: Qualitative comparison of anomaly localization under cross-dataset evaluation. The first five columns are from VisA [23], whereas the last four columns are from KSDD [35], ISIC [9], ClinicDB [4], and Endo [16], respectively. The green curve is used to mark the boundary of abnormal regions from the ground truth. In the heatmaps, brighter colors indicate higher predicted anomaly values.

Table 3: The results with different prototype extraction methods.

Prototype	VisA		ColonDB
	Image-level	Pixel-level	Pixel-level
No Prototype (Prompt-only)	85.4 / 87.0	95.2 / 86.8	84.5 / 76.2
K-means Prototype	89.1 / 90.3	95.5 / 87.6	87.2 / 82.2
vMF Prototype (Ours)	90.0 / 91.6	96.3 / 90.1	88.3 / 84.7

further increases the image-level score to 90.0%/91.6% and provides a clearer gain on localization, improving pixel-level AUROC/PRO to 96.3%/90.1%. A similar trend is observed on ColonDB: vMF improves pixel-level AUROC/PRO from 84.5%/76.2% (prompt-only) to 88.3%/84.7%, outperforming K-means prototypes significantly. Overall, across both datasets, the vMF prototype variant achieves the best performance among the compared extraction strategies.

Ablation on prompt conditioning. The purpose of this experiment is to evaluate the effectiveness of the confidence-gated cross-attention for prototype-to-prompt conditioning. With the vMF prototype bank fixed, we compare different conditioning modes in Tab. 4. Prompt-only (no conditioning) serves as the baseline and is identical to "No Prototype" in Tab. 3. We then enable prototype conditioning via cross-attention under three variants: Cross-Attn (No Gate) (gate $\equiv 1$), Cross-Attn (Scalar-Gated) using a learnable global coefficient α, and Cross-Attn (Confidence-Gated, Ours) using confidence-dependent gating.

As shown in Tab. 4, cross-attention conditioning improves performance over the prompt-only baseline on both datasets (e.g., VisA image-level 85.4%/87.0% $\rightarrow$ 88.2%/89.3%; ColonDB pixel-level 84.5/76.2 $\rightarrow$ 86.5%/82.0%). Adding gat-

Table 4: Ablation results with different prompt conditioning modes.

Conditioning	VisA		ColonDB
	Image-level	Pixel-level	Pixel-level
Prompt-only (no conditioning)	85.4 / 87.0	95.2 / 86.8	84.5 / 76.2
Cross-Attn (No Gate)	88.2 / 89.3	95.6 / 87.9	86.5 / 82.0
Cross-Attn (Scalar-Gated)	88.7 / 90.0	95.8 / 88.6	87.2 / 83.1
Cross-Attn (Confidence-Gated, Ours)	90.0 / 91.6	96.3 / 90.1	88.3 / 84.7

Table 5: Ablation on explicit/implicit abnormal prompt composition.

Abnormal prompt setting	VisA		ColonDB
	Image-level	Pixel-level	Pixel-level
Explicit-only ($\mathcal{T}_a = \mathcal{T}_{\text{exp}}$)	89.2 / 90.7	96.1 / 89.8	87.6 / 84.1
Implicit-only ($\mathcal{T}_a = \mathcal{T}_{\text{imp}}$)	89.1 / 90.5	96.1 / 89.9	87.5 / 84.2
Explicit+Implicit (Ours)	90.0 / 91.6	96.3 / 90.1	88.3 / 84.7

ing yields further gains, with the confidence-gated variant achieving the best results overall (VisA 90.0%/91.6% at image level and 96.3%/90.1% at pixel level; ColonDB 88.3%/84.7% at pixel level). These results suggest that confidence-aware gating provides a more effective conditioning strategy than ungated or globally gated cross-attention under the same setting.

Ablation on explicit/implicit abnormal prompts. This experiment aims to examine the contribution of the proposed explicit/implicit abnormal prompts. All settings are kept identical (backbone, vMF prototype banks, and confidence-gated conditioning), and only the abnormal prompt composition is varied. Specifically, we compare three settings: (i) Explicit-only, where abnormal prompts are conditioned only on abnormal prototypes $\mathbf{Z}_a$; (ii) Implicit-only, where abnormal prompts are constructed only from transformed normal prototypes $\mathbf{Z}_{\text{imp}}$; and (iii) Explicit+Implicit (ours), which combines both prompt sets.

As shown in Tab. 5, the Explicit-only and Implicit-only variants achieve similar performance on both datasets. For example, on VisA they obtain 89.2%/90.7% and 89.1%/90.5% at the image level, and comparable pixel-level results. Combining the two prompt types yields the best overall results, reaching 90.0/91.6 at the image level and 96.3%/90.1% at the pixel level on VisA, and improving ColonDB pixel-level performance to 88.3%/84.7%. These observations suggest that combining explicit abnormal prototypes with implicit deviation prompts provides complementary abnormal cues, which may contribute to more stable prompt representations under cross-domain settings.

6 Conclusion

We proposed a variational prototype-conditioned prompting framework for zero-shot anomaly detection. By learning layer-wise normal/abnormal prototype banks

from ℓ_2-normalized patch embeddings via a variational von Mises–Fisher mixture, our method provides reusable hyperspherical anchors that align naturally with cosine-based VLM matching. We further introduced confidence-gated cross-attention to inject reliable prototype evidence into learnable context tokens, together with implicit deviation prompts derived from normal prototypes to broaden abnormal semantics. Extensive experiments on 15 datasets demonstrate strong cross-dataset performance, indicating that grounding prompt learning in visual prototypes improves the stability and transferability of prompt representations.

Acknowledgements

This work was supported by the NSFC General Program (Grant No. 62576110).

References

1. Aota, T., Tong, L.T.T., Okatani, T.: Zero-shot versus many-shot: Unsupervised texture anomaly detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5564–5572 (2023)
2. Banerjee, A., Dhillon, I.S., Ghosh, J., Sra, S., Ridgeway, G.: Clustering on the unit hypersphere using von mises-fisher distributions. Journal of Machine Learning Research **6**(9) (2005)
3. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., Steger, C.: The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. Int. J. Comput. Vis. **129**(4), 1038–1059 (2021)
4. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015)
5. Bozic, J., Tabernik, D., Skocaj, D.: Mixed supervision for surface-defect detection: From weakly to fully supervised learning. Comput. Ind. **129**, 103459 (2021). https://doi.org/10.1016/J.COMPIND.2021.103459, https://doi.org/10.1016/j.compind.2021.103459
6. Cao, Y., Zhang, J., Frittoli, L., Cheng, Y., Shen, W., Boracchi, G.: Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In: European conference on computer vision. pp. 55–72. Springer (2024)
7. Chen, X., Han, Y., Zhang, J.: 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)
8. Chen, X., Zhang, J., Tian, G., He, H., Zhang, W., Wang, Y., Wang, C., Liu, Y.: Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. In: International joint conference on artificial intelligence. pp. 17–33. Springer (2024)
9. 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.: 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). pp. 168–172. IEEE (2018)
10. Damm, S., Laszkiewicz, M., Lederer, J., Fischer, A.: Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1319–1329. IEEE (2025)
11. Fisher, R.A.: Dispersion on a sphere. Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences **217**(1130), 295–305 (05 1953). https://doi.org/10.1098/rspa.1953.0064, https://doi.org/10.1098/rspa.1953.0064
12. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
13. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: Anomalygpt: Detecting industrial anomalies using large vision-language models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 1932–1940 (2024)
14. Guo, J., Lu, S., Zhang, W., Chen, F., Li, H., Liao, H.: Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 20405–20415. Computer Vision Foundation / IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.01900, https://openaccess.thecvf.com/content/CVPR2025/html/Guo_Dinomaly_The_Less_Is_More_Philosophy_in_Multi-Class_Unsupervised_Anomaly_CVPR_2025_paper.html
15. Hamada, A.: Br35h: Brain tumor detection 2020. Kaggle Dataset (2020), https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection
16. Hicks, S.A., Jha, D., Thambawita, V., Halvorsen, P., Hammer, H.L., Riegler, M.A.: The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In: International Conference on Pattern Recognition. pp. 263–274. Springer (2021)
17. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 19606–19616. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01878, https://doi.org/10.1109/CVPR52729.2023.01878
18. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International conference on multimedia modeling. pp. 451–462. Springer (2019)
19. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
20. Kanade, P.B., Gumaste, P.: Brain tumor detection using mri images. Brain **3**(2), 146–150 (2015)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Kitamura, F.C.: Head ct - hemorrhage. Kaggle Dataset (2018). https://doi.org/10.34740/KAGGLE/DSV/152137, https://www.kaggle.com/dsv/152137
23. Li, N., Jiang, K., Ma, Z., Wei, X., Hong, X., Gong, Y.: Anomaly detection via self-organizing map. In: 2021 IEEE International Conference on Image Processing, ICIP 2021, Anchorage, AK, USA, September 19-22, 2021. pp. 974–978. IEEE (2021)

24. Li, X., Zhang, Z., Tan, X., Chen, C., Qu, Y., Xie, Y., Ma, L.: 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. pp. 16838–16848 (2024)
25. Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced nlp tasks. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 465–476 (2020)
26. Li, Y., Cao, Y., Liu, C., Xiong, Y., Dong, X., Huang, C.: IAD-R1: reinforcing consistent reasoning in industrial anomaly detection. CoRR **abs/2508.09178** (2025). https://doi.org/10.48550/ARXIV.2508.09178, https://doi.org/10.48550/arXiv.2508.09178
27. Liu, Y., Xia, Z., Zhao, M., Wei, D., Wang, Y., Liu, S., Ju, B., Fang, G., Liu, J., Song, L.: Learning causality-inspired representation consistency for video anomaly detection. In: Proceedings of the 31st ACM international conference on multimedia. pp. 203–212 (2023)
28. Liu, Y., Yang, D., Wang, Y., Liu, J., Liu, J., Boukerche, A., Sun, P., Song, L.: Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. ACM Computing Surveys **56**(7), 1–38 (2024)
29. Luo, W., Cao, Y., Yao, H., Zhang, X., Lou, J., Cheng, Y., Shen, W., Yu, W.: 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 (CVPR). pp. 9974–9983 (June 2025)
30. Ma, W., Zhang, X., Yao, Q., Tang, F., Wu, C., Li, Y., Yan, R., Jiang, Z., Zhou, S.: Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4744–4754 (June 2025)
31. Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., Foresti, G.L.: VT-ADL: A vision transformer network for image anomaly detection and localization. In: 30th IEEE International Symposium on Industrial Electronics, ISIE 2021, Kyoto, Japan, June 20-23, 2021. pp. 1–6. IEEE (2021)
32. Qu, Z., Tao, X., Gong, X., Qu, S., Chen, Q., Zhang, Z., Wang, X., Ding, G.: Bayesian prompt flow learning for zero-shot anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30398–30408 (2025)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
34. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
35. Tabernik, D., Šela, S., Skvarč, J., Skočaj, D.: Segmentation-based deep-learning approach for surface-defect detection. Journal of Intelligent Manufacturing **31**(3), 759–776 (2020)
36. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging **35**(2), 630–644 (2015)
37. Wieler, M., Hahn, T.: Weakly supervised learning for industrial optical inspection. In: DAGM symposium in. vol. 6, p. 11 (2007)

38. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18134–18144 (2022)
39. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
40. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International journal of computer vision **130**(9), 2337–2348 (2022)
41. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In: The Twelfth International Conference on Learning Representations
42. Zhu, J., Pang, G.: Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. CoRR **abs/2403.06495** (2024). https://doi.org/10.48550/ARXIV.2403.06495, https://doi.org/10.48550/arXiv.2403.06495