THE CLASS PRIOR AS A DOMAIN-DEPENDENT BIAS: LAYER-WISE VISUAL PROMPT SYNTHESIS FOR FROZEN-CLIP ZERO-SHOT ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP-based zero-shot anomaly detection treats the class-semantic prior as a resource the text branch must supply, whether through class names, handcrafted templates, or per-class descriptions. This paper argues that under cross-domain transfer the prior is instead a domain-dependent bias term. Replacing the class-name slot with a per-layer, image-conditioned prototype read-out stays within 0.5 points of the strongest conditional-prompt baseline, while the class-name prompt itself trails that read-out by 3.7 points, so the prior is not a requirement. Its removal yields 5.1/6.6 points at pixel level in the medical domain against 0.3/0.1 in the industrial domain, an order-of-magnitude difference between two domains evaluated under one protocol, which is the pattern predicted by a bias term and not by a method that is uniformly stronger. Within the prompt the learnable component is what carries the adaptation: discarding it costs 8.8/17.0 points, so what proves indispensable is the prompt's optimizable position in CLIP text space rather than what it says. Because each tapped visual layer synthesizes its own prompt, the construction is structurally distinct from prior methods, which match multi-layer evidence against a single prompt. On 14 industrial and medical benchmarks under a unified protocol, the method matches or exceeds the strongest baselines in accuracy while using 1284 GFLOPs against AdaCLIP's 5570, a factor of 4.3, and running at 12.0 FPS against 5.5, with neither a class name nor a handcrafted state word at inference.

1 INTRODUCTION

Anomaly detection (AD) aims to identify anomalies from established normal conditions in visual data (Pang et al., 2021), making important contributions to industrial inspection (Roth et al., 2022; Wan et al., 2022; Zhao et al., 2023), medical image analysis (Antonelli et al., 2022; Zhao et al., 2021), autonomous driving (Bogdoll et al., 2022), and other safety-critical scenarios. AD can be broadly divided into supervised and unsupervised settings (Zhang et al., 2023; Gudovskiy et al., 2022). The latter includes zero-shot anomaly detection (ZSAD) (Chen et al., 2023), which aims to detect anomalies in unseen categories without relying on target-domain anomaly samples during training. Recently, ZSAD has attracted increasing attention because it reduces annotation cost while improving deployment flexibility across heterogeneous domains.

Most recent ZSAD methods adopt a matching-based paradigm and leverage pre-trained vision-language models such as CLIP (Radford et al., 2021b) to establish consistency between visual evidence and textual prompts (Jeong et al., 2023; Cao et al., 2025b; Zhou et al., 2024). These CLIP-based methods can be roughly categorized into three types, namely *prompt tuning*, *partial tuning*, and *adapter tuning*, as summarized in Fig. 1. Although these strategies have improved anomaly localization and recognition, their empirical gains remain strongly dependent on prompt construction and adaptation topology.

Prompt tuning. Inspired by CoOp and related prompt-learning methods (Zhou, Kaiyang and Yang, Jingkang and Loy, Chen Change and Liu, Ziwei, 2022; Zhou et al., 2022), this line of work improves textual representation through learnable context optimization. In ZSAD, prompt tuning is usually combined with paired positive-negative templates and category-specific descriptors, as shown in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

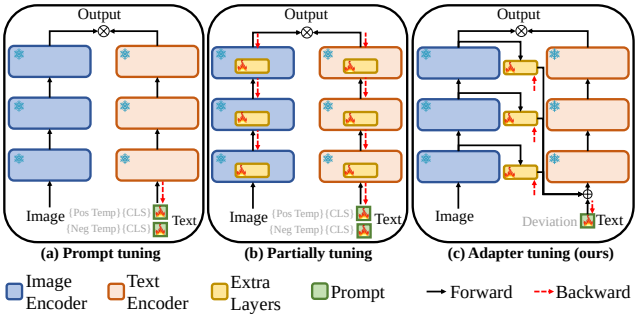


Figure 1: Overview of recent ZSAD methods, categorized as Prompt Tuning, Partial Tuning, and Adapter Tuning. {Pos Temp} represents positive template, {Neg Temp} represents negative template, {CLS} indicates category-specific information.

Fig. 1(a). This design can improve discriminability, but it also increases the prompt count and makes the prediction sensitive to the quality of the manually specified text.

Partial tuning. A second line of work inserts trainable layers into the image encoder, text encoder, or both to strengthen task-specific cross-modal alignment (Cao et al., 2025b; Zhou et al., 2024). As shown in Fig. 1(b), this strategy can improve representation learning capacity, but it changes the encoder pathway and introduces non-negligible parameter and latency overhead.

Adapter tuning. Adapter-style methods modify visual or textual features through lightweight feature adaptation modules (Yang et al., 2024). Compared with partial tuning, they are often more parameter-efficient, but their effectiveness still depends on how prompt semantics and feature adaptation are coordinated.

However, these methods still suffer from two unresolved issues. 1) *Category-specific prompt brittleness*. Most CLIP-based ZSAD methods rely on class names, handcrafted prompt templates, or manually designed semantic priors. In practical scenarios, especially under open-vocabulary industrial inspection and medical decision-support settings, category descriptions may be incomplete, inaccurate, or overly generic. When the textual prior is unreliable, prompt semantics may become a source of confusion rather than a source of supervision. 2) *Sequential adaptation overhead*. Many strong baselines enhance cross-modal interaction by stacking paired templates, extra layers, or sequentially inserted adaptation modules. Although these designs may improve accuracy, they also increase latency, parameter count, and layer coupling, weakening their efficiency and maintainability in real deployment pipelines. More importantly, sequential encoder-side adaptation changes the frozen CLIP pathway and accumulates latency across layers, which makes efficiency gains harder to interpret under a deployment-oriented setting. This concern is especially relevant in practical inspection systems where multimodal fusion, inference efficiency, and incomplete sensory inputs must be balanced simultaneously.

To address these, we propose an efficient ZSAD framework, termed EAD, which consists of two coordinated designs, namely category-agnostic prompt synthesis (CAP) and hierarchical parallel adaptation (HiP). Specifically, the CAP module generates prompt tokens directly from learnable prototypes and image features, reducing reliance on manually specified class-aware text. The HiP framework then organizes multiple CAP modules in parallel across selected visual layers, avoiding encoder surgery and mitigating the cumulative overhead of sequential adaptation. In this sense, EAD is not positioned as a new pre-training objective, but as an evidence-driven design for improving the trade-off between category-agnostic prompting and computational efficiency in ZSAD.

In summary, the main contributions are:

- We propose an efficient ZSAD framework EAD, to address two practical issues in CLIP-based anomaly detection, namely category-specific prompt brittleness and sequential adaptation overhead.

- We develop two coordinated designs in EAD: the category-agnostic prompt synthesis module, CAP, which generates prompt tokens from visual prototypes and image features with

a shared semantic anchor, and the hierarchical parallel adaptation framework, HiP, which performs multi-level adaptation without encoder surgery.

- Extensive experiments on 14 industrial and medical benchmarks, together with qualitative analysis, efficiency comparison, and ablation studies, show that EAD provides a favorable accuracy-efficiency trade-off for practical deployment.

2 Related work

2.1 Zero-shot anomaly detection

Zero-shot anomaly detection (ZSAD) aims to identify anomalies without relying on target-domain anomaly examples during training (Aota et al., 2023a). Compared with supervised or many-shot settings, ZSAD is more flexible, but it is also more difficult because the model must transfer anomaly-related knowledge from external priors rather than from target-domain anomaly annotations (Zhang et al., 2024). Existing ZSAD methods can be broadly divided into two groups: visual-only methods and vision-language-model (VLM)-based methods (Jeong et al., 2023). Visual-only methods usually assume relatively homogeneous textures and detect anomalous regions by measuring deviations from normal feature patterns (Aota et al., 2023b). Although these methods are simple and efficient, their effectiveness is often limited when semantic variation is large. In contrast, VLM-based methods leverage models such as CLIP to exploit rich semantic priors and cross-modal alignment, which makes prompt construction and feature adaptation central to the final performance. From this perspective, category-specific text and adaptation topology become two key design variables in CLIP-based ZSAD.

2.2 VLM transfer learning

Although pre-trained VLMs provide strong generalization ability, they still face non-trivial gaps when transferred to downstream tasks. These gaps arise from differences in image style, text format, and task objective between pre-training and target deployment (Bian et al., 2024; Qu et al., 2024). Recent studies also show that multimodal fusion, prompt automation, and incomplete-modality handling remain important when adapting anomaly detectors to practical industrial settings (Chen et al., 2026a; Zhou et al., 2026). Existing transfer strategies fall into prompt tuning, partial tuning, and adapter tuning. Prompt tuning leaves the backbone fixed and optimizes the text-side input, and its two canonical forms differ in a way that matters for the present work. CoOp (Zhou, Kaiyang and Yang, Jingkang and Loy, Chen Change and Liu, Ziwei, 2022) learns a context that is constant across inputs, so the prompt is a set of free parameters. CoCoOp (Zhou et al., 2022) adds a single image-conditioned offset produced from the final-layer global feature and added uniformly to every context vector, which makes the prompt a function of the image while keeping the class token in place. Partial tuning and adapter tuning instead place trainable layers or feature adapters inside the image encoder, the text encoder, or both (Houlsby et al., 2019; Cao et al., 2025b; Yang et al., 2024). Across all three families the category prior is assumed to be a resource that the text branch supplies, and stronger adaptation is bought with either a more specific category prior or a more intrusive adaptation module.

3 Methodology

3.1 Overview

We address ZSAD on $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{S}_i, \delta_i)\}_{i=1}^{I_N}$, images with masks and image-level labels. A frozen CLIP encoder $\mathcal{G}$ provides the class token and multi-level patch features, from which D CAP modules read out prototype tokens. With a shared learnable anchor $\mathbf{Q}$, they form layer-wise prompts $\mathbf{W}_d$ that the text encoder maps to the image-level score and pixel-level mask. All branches read from the encoder without writing back, making cost additive in D. Fig. 2 summarizes the framework and the per-module workflow.

3.2 Category-Agnostic Prompt Synthesis

CAP here denotes the module below, not the compound abnormality prompt of FAPrompt (Zhu et al., 2025). A frozen vision encoder $\mathcal{G}$ (Jeong et al., 2023; Cao et al., 2025b) supplies the class

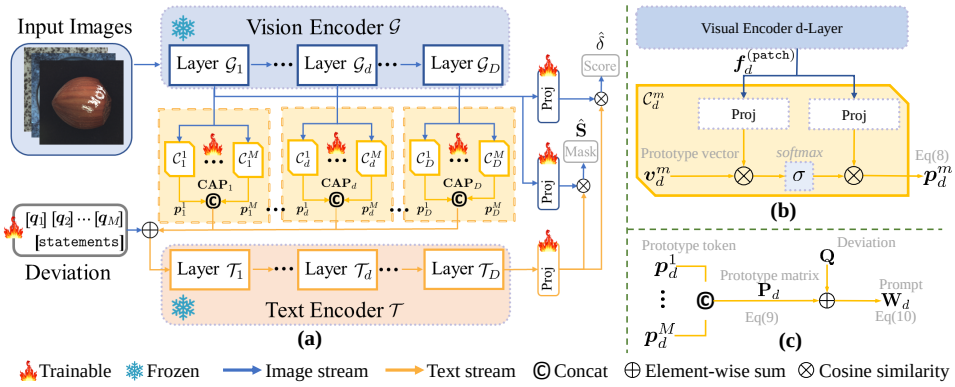


Figure 2: (a) Overview of EAD. The input to EAD is an image together with the shared semantic anchor, and the outputs are the image-level anomaly score and the pixel-level anomaly mask. (b) Workflow of the d-th CAP module, detailed in 3.2. (c) Calculation of Prompt, detailed in 3.2.

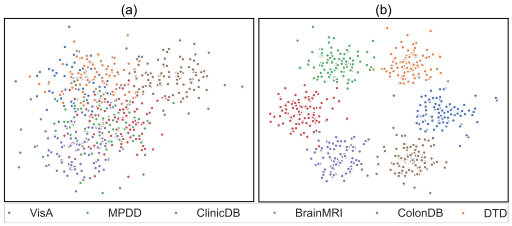


Figure 3: The visualization of features and prototype tokens using t-SNE (Van der Maaten & Hinton, 2008). (a) visualizes 1000 patch features $\mathbf{F}_d^{(\texttt{patch})}$ from six categories at the 18th CLIP layer. (b) visualizes the corresponding prototype tokens $\boldsymbol{p}_d$ generated from these features.

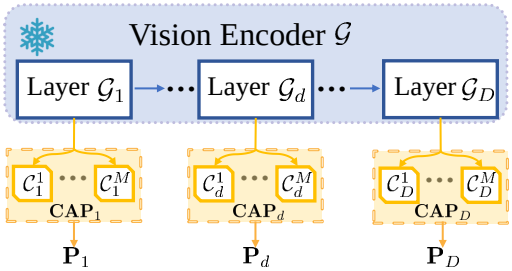


Figure 4: Overview of HiP framework. The HiP is designed with an emphasis on hierarchical parallelism, integrating CAP modules across multiple levels in a seamless manner.

token and the patch features of D selected layers,

$$\left(\boldsymbol{f}^{(\texttt{cls})}, \{\mathbf{F}_d^{(\texttt{patch})}\}_{d=1}^{D}\right) = \mathcal{G}(\mathbf{X}). \tag{1}$$

where L is the total number of blocks. Layers 6, 12, 18 and 24 are tapped by convention, not by design. Each selected layer owns one module, in which the m-th prototype token attends over the patch matrix against a learnable query $\boldsymbol{v}_d^m$,

$$\boldsymbol{a}_d^m = \sigma\left(\frac{\mathcal{H}_1^m(\mathbf{F}_d^{(\texttt{patch})})(\boldsymbol{v}_d^m)^\top}{\sqrt{C_T}}\right) \in \mathbb{R}^N, \tag{2}$$

$$\boldsymbol{p}_d^m = (\boldsymbol{a}_d^m)^\top \mathcal{H}_2^m(\mathbf{F}_d^{(\texttt{patch})}) \in \mathbb{R}^{C_T}. \tag{3}$$

with learnable projections $\mathcal{H}_1^m, \mathcal{H}_2^m$ and a scaling function (softmax) $\sigma(\cdot)$. The M tokens fill a prompt matrix of fixed CLIP-compatible length K,

$$\mathbf{P}_d = [\,\boldsymbol{e}_{\text{SOS}}; \boldsymbol{p}_d^1; \cdots; \boldsymbol{p}_d^M; \boldsymbol{e}_{\text{EOS}}; \underbrace{\phi; \cdots; \phi}_{K-M-2 \text{ empty slots}}\,] \in \mathbb{R}^{K \times C_T}, \tag{4}$$

where $\boldsymbol{e}_{\text{SOS}}$, $\boldsymbol{e}_{\text{EOS}}$ and ϕ are the special and empty token embeddings, so $\mathbf{P}_d$ depends only on layer-d evidence. Fig. 3 shows these tokens separating categories more cleanly than their source patches.

3.3 Hierarchical Parallel Prompt Encoding

The HiP framework runs one module per selected layer as parallel branches in the shared vision–language space (Fig. 4). Each branch computes $\mathbf{P}_d = \text{CAP}_d(\mathbf{F}_d^{(\texttt{patch})})$ and returns nothing to the

encoder, so block $\ell{+}1$ observes $\boldsymbol{h}_{\ell+1} = B_\ell(\boldsymbol{h}_\ell)$ as in the unmodified model; sequential adaptation makes that state depend on adaptation parameters, recomputing the remaining $L - \ell$ blocks. Branches are thus independent and cost is additive in D. The bypass arrangement is prior work (He et al., 2021; Houlsby et al., 2019; Yang et al., 2024; Chen et al., 2024), which here lets every branch own its prompt. Prompts combine the prototype matrix with a shared anchor $\mathbf{Q}$,

$$\mathbf{Q} = [\boldsymbol{e}_{\mathrm{SOS}}; \boldsymbol{q}^1; \cdots; \boldsymbol{q}^M; \boldsymbol{e}_{\mathrm{EOS}}; \underbrace{\phi; \cdots; \phi}_{K-M-2 \text{ slots}}] \in \mathbb{R}^{K \times C_T}, \tag{5}$$

$$\mathbf{W}_d = \alpha \cdot \mathbf{P}_d + \beta \cdot \mathbf{Q}. \tag{6}$$

where $\boldsymbol{q}^m$ are learnable anchor tokens and (α, β) balance visual evidence against the shared prior. Both terms share the shape $K \times C_T$, so Eq. (6) is an element-wise weighted sum. The frozen text encoder then yields

$$\boldsymbol{f}_d^{(\texttt{text})} = \mathcal{H}_4^d(\mathcal{T}(\mathbf{W}_d)) \in \mathbb{R}^{C_T}. \tag{7}$$

with $\mathcal{T}(\cdot)$ the text encoder and $\mathcal{H}_4^d$ a per-layer projection.

3.4 Relation to context optimization

With $[\mathrm{CLS}]$ denoting the class-name embedding, the three constructions read

$$\text{CoOp:} \quad \mathbf{t} = [\boldsymbol{v}^1; \cdots; \boldsymbol{v}^M; [\mathrm{CLS}]], \tag{8}$$

$$\text{CoCoOp:} \quad \mathbf{t}(\mathbf{X}) = [\boldsymbol{v}^1{+}\boldsymbol{\pi}; \cdots; \boldsymbol{v}^M{+}\boldsymbol{\pi}; [\mathrm{CLS}]], \tag{9}$$

$$\text{ours:} \quad \mathbf{W}_d = \alpha \cdot \mathbf{P}_d + \beta \cdot \mathbf{Q}, \tag{10}$$

where $\boldsymbol{\pi} = h_\theta(\boldsymbol{f}^{(\texttt{cls})})$ is one offset shared by all M contexts. Two properties are shared rather than novel: $\mathbf{Q}$ is learnable freely through a frozen text encoder, i.e. Eq. (8) without the class token, and image conditioning comes from Eq. (9). The difference is $\mathbf{P}_d$, a function rather than a parameter: each $\boldsymbol{p}_d^m$ attends over the patch matrix, whereas CoOp's $\boldsymbol{v}^m$ is input-independent and CoCoOp's offset is position-shared. Because $\boldsymbol{v}_d^m$ is indexed by layer, so are $\mathbf{P}_d$, $\mathbf{W}_d$ and $\hat{\mathbf{S}}_i^d$ – an index neither baseline admits. The weighted sum keeps $\mathbf{P}_d$ and $\mathbf{Q}$ separately ablatable.

3.5 Anomaly Scoring and Optimization Objectives

EAD predicts an image-level score $\hat{\delta}_i$ and a pixel-level map $\hat{\mathbf{S}}_i$, optimizing only the CAP modules and projections while both encoders stay frozen. The image score matches the class token against the summed layer-wise text embeddings,

$$\hat{\delta}_i = \mathcal{H}_3(\boldsymbol{f}^{(\texttt{cls})}) \cdot \sum_{d=1}^{D} \left(\mu_d \cdot \boldsymbol{f}_d^{(\texttt{text})}\right). \tag{11}$$

with projection $\mathcal{H}_3$ and branch weights μ_d. Pixel-wise, each layer first scores patches and reshapes them to the image grid,

$$\hat{\boldsymbol{s}}_i^d = \mathcal{H}_5(\mathbf{F}_d^{(\texttt{patch})}) \cdot \boldsymbol{f}_d^{(\texttt{text})} \in \mathbb{R}^N, \tag{12}$$

$$\hat{\mathbf{S}}_i^d = \mathcal{U}\Big(\mathrm{reshape}\Big(\hat{\boldsymbol{s}}_i^d, \sqrt{N}, \sqrt{N}\Big)\Big) \in \mathbb{R}^{H \times W \times 1}, \tag{13}$$

where $\mathcal{H}_5$ maps patches into text space and $\mathcal{U}(\cdot)$ is bilinear upsampling; intermediates are then aggregated with learnable λ_d,

$$\hat{\mathbf{S}}_i = \sum_{d=1}^{D} \lambda_d \cdot \hat{\mathbf{S}}_i^d. \tag{14}$$

Supervision combines an image-level and a dense localization term,

$$\mathcal{L} = \mathcal{L}_{\mathrm{CLS}} + \mathcal{L}_{\mathrm{LOC}}. \tag{15}$$

the first an L_1 distance (Wu et al., 2023), avoiding the calibration cross-entropy would need over a similarity,

$$\mathcal{L}_{\mathrm{CLS}} = \frac{1}{I_N} \sum_{i=1}^{I_N} |\hat{\delta}_i - \delta_i|. \tag{16}$$

Table 1: **Comparisons of ZSAD methods in the industrial domain.** The best performance is in **bold**, and the second-best is underlined. † indicates that the prior textual information is **not** incorporated into the method.

Metric	Dataset	Prompt tuning		Partial tuning			Adapter tuning	
		ACLIP (Zhou et al., 2024)	WinCLIP (Jeong et al., 2023)	DINOv2† (Oquab et al., 2024)	SAM† (Kirillov et al., 2023)	INP-Former (Luo et al., 2025)	AdaCLIP (Cao et al., 2025b)	**Ours**†
Image-level AUROC, max-F1	MVTec AD	91.5, 90.7	**91.8**, **92.9**	74.4, 87.4	70.8, 86.0	80.8, 89.1	89.2, 90.6	90.0, 91.5
	VisA	82.1, **85.1**	78.1, 80.7	75.2, 78.5	61.9, 73.9	67.5, 75.0	**85.8**, 83.1	**85.8**, 82.9
	MPDD	77.0, 78.9	61.4, 77.5	62.4, 74.9	63.0, 77.0	75.5, 80.9	76.0, 82.5	**80.1**, **82.9**
	BTAD	88.3, 86.0	68.2, 67.6	79.3, 69.3	89.4, 85.7	90.3, 87.5	88.6, 88.2	**93.6**, **90.8**
	KSDD	84.7, 79.6	93.3, 79.0	94.9, 77.5	65.8, 37.9	95.0, 90.6	97.1, 90.7	**98.2**, **93.5**
	DAGM	97.5, 96.1	91.7, 87.6	90.7, 89.2	82.7, 83.6	93.5, 92.1	**99.1**, **97.5**	98.6, 96.3
	DTD	93.5, 86.0	95.1, 94.1	85.8, 93.5	81.9, 91.1	90.9, 89.7	95.5, 94.7	**96.3**, **95.5**
	Average	87.8, 86.1	82.8, 82.8	80.4, 81.5	73.6, 76.4	84.8, 86.4	90.2, 89.6	**91.8**, **90.5**
Pixel-level AUROC, max-F1	MVTec AD	**91.1**, **44.4**	85.1, 31.6	85.9, 39.6	85.4, 29.4	88.0, 39.5	88.7, 43.4	88.1, 43.0
	VisA	95.5, 27.7	79.6, 14.8	95.0, 30.3	92.6, 18.2	88.7, 11.8	95.5, **37.7**	**95.7**, 33.7
	MPDD	**96.5**, 33.1	71.2, 15.4	95.6, 31.1	94.8, 22.1	94.2, 20.6	96.1, **34.9**	95.9, 33.7
	BTAD	**94.2**, 47.7	72.6, 18.5	91.9, 43.4	93.8, 46.9	92.1, 47.8	92.1, 51.7	92.7, **52.8**
	KSDD	90.6, 50.9	95.8, 21.3	**99.3**, 50.6	91.2, 18.4	90.9, 27.7	97.7, 54.5	98.9, **55.2**
	DAGM	**95.6**, **60.6**	81.3, 13.9	90.9, 52.0	88.6, 40.7	89.1, 46.4	91.5, 57.5	92.1, 59.7
	DTD	97.9, 47.7	79.5, 16.1	97.0, 63.4	95.0, 56.7	92.9, 50.3	97.9, 71.6	**98.1**, **74.3**
	Average	**94.5**, 44.6	80.7, 18.8	93.7, 44.3	91.7, 33.2	90.8, 34.9	94.2, 50.2	**94.5**, **50.3**

Table 2: **Comparisons of ZSAD methods in the medical domain.** The best performance is in **bold**, and the second-best is underlined. † indicates that the prior textual information is **not** incorporated into the method.

Metric	Dataset	Prompt tuning		Partial tuning		Adapter tuning		
		ACLIP (Zhou et al., 2024)	WinCLIP (Jeong et al., 2023)	DINOv2† (Oquab et al., 2024)	SAM† (Kirillov et al., 2023)	AA-CLIP (Ma et al., 2025)	AdaCLIP (Cao et al., 2025b)	**Ours**†
Image-level AUROC, max-F1	HeadCT	93.4, 87.8	84.1, 79.8	71.4, 72.4	78.4, 76.4	81.2, 77.9	91.4, 85.2	**93.8**, **88.1**
	BrainMRI	90.3, 92.9	89.8, 86.3	78.3, 82.7	71.5, 78.9	80.2, 83.4	94.8, 91.2	**95.2**, **93.0**
	Br35H	94.6, 92.3	81.6, 74.4	69.1, 70.5	59.0, 67.2	75.9, 73.5	**97.7**, 92.4	97.6, **92.5**
	Average	92.8, 91.0	85.2, 80.2	72.9, 75.2	69.7, 74.1	79.1, 78.3	94.6, 89.6	**95.5**, **91.2**
Pixel-level AUROC, max-F1	ISIC	89.7, 70.7	67.1, 48.5	**94.2**, 79.6	**94.2**, **81.0**	93.2, 80.0	89.3, 71.4	93.5, 80.0
	ColonDB	81.9, 39.8	61.1, 19.6	87.3, 56.5	86.1, 45.7	84.0, 47.8	90.4, 58.2	**92.7**, **65.4**
	ClinicDB	82.9, 42.4	67.1, 24.4	83.3, 56.2	83.5, 43.0	89.9, 48.2	84.4, 58.2	**90.0**, **59.6**
	TN3K	81.5, 50.1	67.2, 30.0	73.3, 35.7	70.1, 32.5	78.4, 49.2	77.2, 41.9	**85.4**, **50.8**
	Average	84.0, 50.8	65.6, 30.6	84.5, 57.0	83.5, 50.5	86.4, 56.3	85.3, 57.4	**90.4**, **64.0**

the second a Frobenius loss applied to every intermediate map and to the final mask,

$$\mathcal{L}_{\text{LOC}} = \sum_{d=1}^{D} \mathcal{L}_{\text{LOC}}^{d} + \mathcal{L}_{\text{LOC}}^{\text{output}}, \quad \mathcal{L}_{\text{LOC}}^{d} = \frac{1}{I_N} \sum_{i=1}^{I_N} \|\hat{\mathbf{S}}_i^d - \mathbf{S}_i\|_F^2, \quad \mathcal{L}_{\text{LOC}}^{\text{output}} = \frac{1}{I_N} \sum_{i=1}^{I_N} \|\hat{\mathbf{S}}_i - \mathbf{S}_i\|_F^2. \tag{17}$$

so each branch is supervised on its own and λ_d calibrates rather than suppresses it.

4 Experiment

4.1 Experimental Setup

Metrics. We report image-level and pixel-level AUROC together with max-F1 at the optimal threshold (Qu et al., 2024). In addition to per-dataset performance, we report industrial and medical averages to make the cross-domain behavior of EAD explicit.

4.2 Main Results

Table 3: Efficiency comparison. Acc indicates the average AUROC on image-level and pixel-level tasks. GFLOPs are computed per image at the common input size 518×518, and FPS is measured with batch size 1 on the same RTX 4080 under the unified inference pipeline.

Name	GFLOPs ↓	MParam	FPS ↑	Acc ↑
SAM	2735	641	0.5	63.3, 59.4
WinCLIP	26359	**78**	0.6	78.6, 53.1
ACLIP	2570	498	7.2	89.8, 68.1
AdaCLIP	5570	528	5.5	91.1, 71.7
Ours	**1284**	542	**12.0**	**93.1**, **74.0**

Name	Parameter breakdown (MParam)			
	Text	Image	Other	Total
CLIP	304	124	–	428
AdaCLIP	351	171	2	528
Ours	304	124	114	542

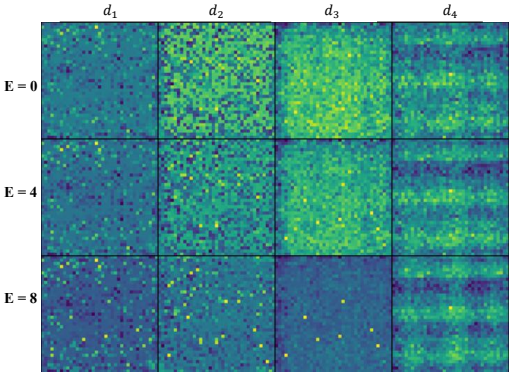


Figure 5: Distance visualization of visual and textual embeddings across training epochs.

Table 4: Ablation results of paired templates and extra layers.

Name	GFLOPs	Mparam	FPS
Ours	**1284**	**542**	**12.0**
w/ paired templates	2460	**542**	6.9
w/ extra layers	4164	550	7.1
w/ both	5372	550	5.3

Datasets. We evaluate EAD on 14 anomaly detection benchmarks from industrial and medical domains. Dataset descriptions and statistics are provided in Appendix A.4 (Table A3).

Details. The full implementation settings—backbone and tapped layers, input resolution, auxiliary domains, optimizer and learning rate, prompt configuration, and the evaluation protocol—are given in Appendix A.1.

Accuracy. Table 1 and Table 2 compare EAD with representative baselines under the protocol of Sec. 4.1. Against AdaCLIP, the strongest prompt-based baseline in both tables (Cao et al., 2025b), pixel-level averages improve by 5.1/6.6 points in AUROC/max-F1 in the medical domain against 0.3/0.1 in the industrial one. This differential is what the third prediction of Sec. 1 anticipates: a method that were simply stronger would lead by comparable margins in both domains, whereas these differ by more than an order of magnitude under one protocol. No method dominates every benchmark: EAD trails WinCLIP on MVTec AD at image level and AdaCLIP on DAGM, the two benchmarks whose class names carry the most usable semantics. Two readings survive, since the medical baselines may be weaker for reasons unrelated to prompt construction and category-semantic strength is never measured independently of the outcome.

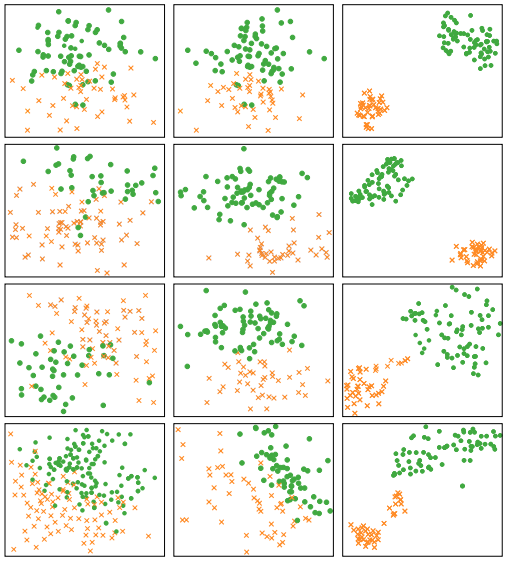
Category-agnostic baselines. SAM and DINOv2 also avoid class names, but without prompt-based cross-modal adaptation; SAM remains strong on ISIC, DINOv2 on texture-heavy KSDD.

Prompt-based baselines. WinCLIP gains from explicit descriptors yet transfers less stably when semantics are ambiguous, while AA-CLIP and AdaCLIP reach their gains through category- or template-dependent prompts. EAD matches or exceeds them while remaining category-agnostic.

Efficiency. Table 3 reports computation, parameters and throughput together, measured per image on one RTX 4080 in a single pipeline, which makes the entries comparable to each other in a way cross-paper numbers are not. EAD uses 1284 GFLOPs against AdaCLIP's 5570, a factor of 4.3, while carrying *more* total parameters, 542M against 528M; the saving therefore cannot be attributed to a smaller model, and it follows from placing the trainable components in a branch that reads from the frozen encoder, so their cost is added once rather than propagated through the remaining blocks. Two caveats: total parameters include the frozen backbone, and AdaCLIP's reported decomposition sums to 524M against a stated 528M.

Table 4 locates that cost in the two mechanisms this design omits. Paired templates raise computation to 2460 GFLOPs and halve throughput to 6.9 FPS at unchanged parameters, extra encoder layers reach 4164 GFLOPs and 550M at 7.1 FPS, and both together reach 5372 GFLOPs at 5.3 FPS, close to the measured cost of AdaCLIP. That table carries no accuracy column, so the saving is not shown to be free. Per-class results and their aggregation conventions are in Appendix A.5.

ACLIP and AdaCLIP, which suggests that prompt synthesis and layer aggregation sharpen the discriminative structure of the anomaly features.



(a) DINOv2 (b) AdaCLIP (c) Ours

Figure 6: t-SNE visualization of predicted pixel-level features on bottle (MVTec AD), 01 (BTAD), ISIC, and Br35H. Normal samples are shown as green and abnormal samples are shown as orange.

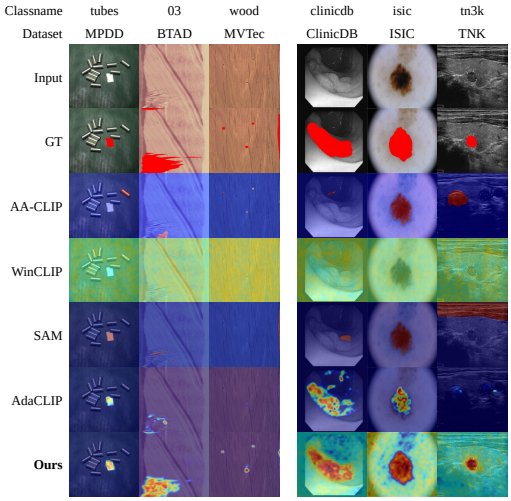


Figure 7: Segmentation visualization of different ZSAD methods.

Table 5: Localization-loss comparison under the same model and training budget. Formula lists the core form of each loss, and Acc reports the overall average AUROC and max-F1 across the industrial and medical domains.

Localization loss	Acc
Frobenius(ours) Formula: $\|\hat{\mathbf{S}} - \mathbf{S}\|_F^2$	**93.1, 74.0**
Focal Formula: $-(1-p_t)^\gamma \log p_t$	92.4, 72.9
Dice Formula: $1 - \frac{2\sum \hat{\mathbf{S}}\mathbf{S}+\epsilon}{\sum \hat{\mathbf{S}}+\sum \mathbf{S}+\epsilon}$	92.0, 72.1
BCE-style Formula: $-\mathbf{S}\log\hat{\mathbf{S}} - (1-\mathbf{S})\log(1-\hat{\mathbf{S}})$	92.7, 73.0

Table 6: Prompt comparison among class-name prompts, CoCoOp-style prompts, and CAP. Avg denotes the overall average AUROC and max-F1 across the industrial and medical domains.

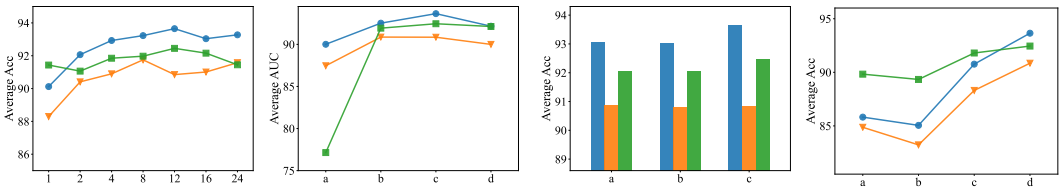
Prompt design	Class-name prior	Conditional prompt	Acc
Class-name	✓	✗	89.4, 69.1
CoCoOp	✓	✓	92.6, 73.2
CAP (ours)	✗	✓	**93.1, 74.0**

Qualitative analysis. Fig. 5 tracks whether optimization improves cross-modal consistency: as prototypes are updated the dark regions concentrate, indicating that visual and textual embeddings become progressively better aligned. Fig. 6 examines whether the learned representations separate normal from abnormal samples across datasets, where EAD separates more clearly in t-SNE space than DINOv2, Fig. 7 inspects final localization quality, and the segmentation maps show more accurate defect boundaries and fewer irrelevant activations than AA-CLIP, WinCLIP and SAM, consistent with the pixel-level gains rather than with image-level scoring alone. None of the three figures is scored, so they are read as consistent with, not as independent evidence for, the reported metrics.

4.3 Ablation Studies

Table 5 compares four localization objectives whose forms appear there: Frobenius performs best, consistent with the dense target formed by aggregating intermediate maps, whereas focal, dice and BCE-style losses emphasise hard pixels, set overlap or independent pixels. Table 6 varies only where category information enters: a class-name prompt, a CoCoOp-style prompt retaining the class token, and the synthesis of Eq. (10). The class-name prompt is weakest by 3.7/4.9 points, so the prior is not necessary; the 0.5/0.8 margin over the CoCoOp-style prompt is within run-to-run variation, supporting only non-inferiority. Table 7 varies topology alone: bypass adaptation reaches 1284 GFLOPs and 12.0 FPS against 4830 and 3.2, factors of 3.8 and 3.75, at accuracy level within resolution. The bypass arrangement is prior work (Sec. A.2); what is measured is its cost.

Table 8 removes $\mathbf{P}$ and $\mathbf{Q}$ in turn: $\mathbf{Q}$ costs 8.8/17.0 points against 0.5/0.8 for $\mathbf{P}$, so substituting for the class name is close to free, making discarding it viable. The same row collapses per-layer



(a) Prompt length M. (b) Coupling weights (α, β). (c) Layer weights (μ, λ). (d) Backbone comparison.

Figure 8: Sensitivity of prompt length M (a), coupling weights (α, β) (b), layer weights (μ, λ) (c), and CLIP backbones (d). The blue, green, and orange curves denote average pixel-level AUROC, image-level AUROC, and image-level max-F1, respectively.

Table 7: Structure comparison between sequential adaptation and HiP. Avg denotes the overall average AUROC and max-F1 across the industrial and medical domains.

Structure	GFLOPs	FPS	Acc
Sequential adaptation	4830	3.2	92.7, 72.4
Parallel adaptation (HiP)	**1284**	**12.0**	**93.1, 74.0**

Table 8: Ablation results of prototype matrix $\mathbf{P}$ and deviation $\mathbf{Q}$ in CAP.

Name	Industrial		Medical	
	AUROC	max-F1	AUROC	max-F1
Ours	**93.2**	**70.4**	**93.0**	**77.6**
w/o $\mathbf{Q}$	84.4	53.4	82.0	63.8
w/o $\mathbf{P}$	92.7	69.6	91.1	75.9

differentiation, since $\mathbf{W}_d = \beta \cdot \mathbf{Q}$ is then identical across layers, and that costs 0.5 points: per-layer prompting is thus a structural property, not a demonstrated source of accuracy.

Table 9 shows the localization loss, image-level loss, σ and $\mathcal{H}_5$ each contributing. Fig. 8 locates the capacity optimum near $M = 12$, confirms the dependence on $\mathbf{Q}$, shows learned (μ, λ) beating uniform weighting, and gains with capacity.

Table 9: Ablation results of $\mathcal{L}_{\text{LOC}}$, $\mathcal{L}_{\text{CLS}}$, σ and $\mathcal{H}_5$.

$\mathcal{L}_{\text{LOC}}$	$\mathcal{L}_{\text{CLS}}$	σ	$\mathcal{H}_5$	Industrial		Medical	
				AUROC	max-F1	AUROC	max-F1
✗	✓	✓	✓	81.2	52.1	78.7	60.9
✓	✗	✓	✓	75.4	59.3	84.8	70.7
✓	✓	✗	✓	89.9	65.6	88.8	72.5
✓	✓	✓	✗	91.7	68.1	89.4	76.6
✓	✓	✓	✓	**93.2**	**70.4**	**93.0**	**77.6**

5 Conclusion

The class-semantic prior on the text side of a frozen-CLIP detector is treated across this literature as a resource to be supplied more carefully. The evidence here supports a different account: replacing the class name with a per-layer read-out of visual evidence costs nothing measurable and gains most where category names carry least, which is the pattern a bias term produces and an informative prior does not. Within the prompt, what proves indispensable is not the class-name prior it replaces but the freely optimizable coordinates the learnable component occupies in CLIP text space: discarding that component costs 8.8/17.0 points. Engineering what the prompt says is therefore unlikely to be repaid, whereas the position it occupies and its freedom to vary with depth are what carry the adaptation. The per-layer construction is structurally distinct from prior methods, and isolating its contribution to accuracy with a broadcast-prompt control is the immediate next step.

AI use statement

In this work, we used generative AI tools to assist with translation, to modify scientific figures, and to propose keywords.

Ethics statement

All authors declare that this work was conducted in accordance with relevant ethical guidelines and standards.

Reproducibility statement

To ensure reproducibility, we will release the complete source code upon acceptance of this paper. All datasets used in the experiments are publicly available, and the detailed data preprocessing steps are described in the supplementary materials. Additional implementation details and experimental configurations are provided in the appendix.

References

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5564–5572, January 2023a.

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

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

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

Jintang Bian, Xiaohua Xie, Jian-Huang Lai, and Feiping Nie. Multi-view contrastive clustering via integrating graph aggregation and confidence enhancement. *Information Fusion*, 108:102393, 2024.

Daniel Bogdoll, Maximilian Nitsche, and J. Marius Zöllner. Anomaly detection in autonomous driving: A survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4488–4499, June 2022.

Yunkang Cao, Xiaohao Xu, Yuqi Cheng, Chen Sun, Zongwei Du, Liang Gao, and Weiming Shen. Personalizing vision-language models with hybrid prompts for zero-shot anomaly detection. *IEEE Transactions on Cybernetics*, 2025a.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pp. 55–72. Springer, 2025b.

Peter D Chang, Edward Kuoy, Jack Grinband, Brent D Weinberg, Matthew Thompson, Richelle Homo, Jefferson Chen, Hermelinda Abcede, Mohammad Shafie, Leo Sugrue, et al. Hybrid 3d/2d convolutional neural network for hemorrhage evaluation on head ct. *American Journal of Neuroradiology*, 39(9):1609–1616, 2018.

Hanzhi Chen, Jingbin Que, Kexin Zhu, Zhide Chen, Fei Zhu, Wencheng Yang, Xu Yang, and Xuechao Yang. Malm-clip: A generative multi-agent framework for multimodal fusion in few-shot industrial anomaly detection. *Information Fusion*, 127:103765, 2026a. ISSN 1566-2535.

Qiyu Chen, Zhen Qu, Wei Luo, Haiming Yao, Yunkang Cao, Yuxin Jiang, Yinan Duan, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. Cops: Conditional prompt synthesis for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8554–8563, 2026b.

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

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

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 257–261. IEEE, 2021.

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

Ahmed Hamada. Br35H : Brain Tumor Detection 2020. https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection, 2020.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

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

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. 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)*, pp. 66–71. IEEE, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Yiting Li, Adam Goodge, Fayao Liu, and Chuan-Sheng Foo. Promptad: Zero-shot anomaly detection using text prompts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1093–1102, January 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. 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.

Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S.Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4744–4754, June 2025.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024, 2024.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

Jun Yeong Park, JunYoung Seo, Minji Kang, and Yu Rang Park. Moeclip: patch-specialized experts for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35534–35544, 2026.

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

Zhen Qu, Xian Tao, Xinyi Gong, ShiChen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30398–30408, June 2025.

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

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

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14318–14328, 2022.

Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020.

Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Qian Wan, Liang Gao, Xinyu Li, and Long Wen. Unsupervised image anomaly detection and segmentation based on pretrained feature mapping. *IEEE Transactions on Industrial Informatics*, 19 (3):2330–2339, 2022.

Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM Symposium In*, volume 6, pp. 11, 2007.

Minghu Wu, Shufan Zhang, Fan Zhang, Rui Sun, Jing Tang, and Sheng Hu. Anomaly detection method for lithium-ion battery cells based on time series decomposition and improved manhattan distance algorithm. *ACS omega*, 9(2):2409–2421, 2023.

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23826–23837, 2024.

Hui Zhang, Zuxuan Wu, Zheng Wang, Zhineng Chen, and Yu-Gang Jiang. Prototypical residual networks for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16281–16291, June 2023.

Zhaoxiang Zhang, Hanqiu Deng, Jinan Bao, and Xingyu Li. Dual-image enhanced clip for zero-shot anomaly detection. *arXiv preprint arXiv:2405.04782*, 2024.

He Zhao, Yuexiang Li, Nanjun He, Kai Ma, Leyuan Fang, Huiqi Li, and Yefeng Zheng. Anomaly detection for medical images using self-supervised and translation-consistent features. *IEEE Transactions on Medical Imaging*, 40(12):3641–3651, 2021. doi: 10.1109/TMI.2021.3093883.

Qing Zhao, Yan Wang, Boyang Wang, Junxiong Lin, Shaoqi Yan, Wei Song, Antonio Liotta, Jiawen Yu, Shuyong Gao, and Wenqiang Zhang. Msc-ad: A multiscene unsupervised anomaly detection dataset for small defect detection of casting surface. *IEEE Transactions on Industrial Informatics*, 20(4):6041–6052, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, June 2022.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Xunkuai Zhou, Xi Chen, Jie Chen, and Ben M. Chen. An accurate and resource-efficient network for surface anomaly detection via enhanced downsampling and activation representation. *Advanced Engineering Informatics*, 69:103891, 2026. doi: 10.1016/j.aei.2025.103891.

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

Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22241–22251, October 2025.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pp. 392–408. Springer, 2022.