CRAG-CLIP: CROSS-LAYER RELATIONAL ABNORMALITY GEOMETRY FOR ZERO-SHOT IMAGE ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP-based zero-shot image anomaly detection has achieved encouraging progress by leveraging transferable vision–language representations to identify anomalies in previously unseen datasets. Recent source-only approaches further improve cross-dataset transfer by learning anomaly-aware text prompts from auxiliary source data. However, existing prompt-based approaches typically represent the abnormal concept with a single normal-abnormal opposition and fuse multi-layer visual evidence through fixed heuristics. This paradigm inevitably under-utilizes the multi-faceted nature of real-world anomalies and the complementary semantic representations across different transformer layers. To address these limitations, we propose CRAG-CLIP, a source-only zero-shot anomaly detection framework that models abnormality as a cross-layer relational geometry in the CLIP text space. The framework represents abnormality through multiple object-agnostic normal–anomaly semantic relations, constrains the learned directions with frozen language priors to mitigate source-specific semantic drift, and promotes complementary semantic coverage through diversity regularization. It further estimates the reliability of layer-wise patch evidence from posterior uncertainty and cross-layer disagreement, enabling data-dependent fusion across visual layers. For image-level prediction, global semantic evidence and risk-sensitive local anomaly evidence are jointly aggregated to capture both holistic context and localized defects. Extensive evaluations on five industrial and nine medical datasets establish the effectiveness and competitive performance of CRAG-CLIP.

1 INTRODUCTION

Image anomaly detection, the task of identifying and localizing visual patterns that deviate from expected normal appearances, is a fundamental problem in computer vision with critical applications in industrial inspection (Liu et al., 2024; Li et al., 2025; Tao et al., 2022) and medical screening (Fernando et al., 2021; Tschuchnig & Gadermayr, 2021). Given the prohibitive cost of collecting exhaustive anomaly examples for each emerging domain, zero-shot anomaly detection has attracted considerable interest. This setting encompasses both training-free methods that directly deploy pre-trained models on unseen targets (Jeong et al., 2023; Li et al., 2024; Zhang et al., 2026) and methods that learn transferable anomaly representations from auxiliary source data (Zhou et al., 2024). This work adopts the latter source-only cross-domain setting. The success of large-scale vision-language pre-training, notably CLIP (Radford et al., 2021), has stimulated CLIP-based zero-shot anomaly detection. Early methods typically paired class names with handcrafted normal and anomaly state descriptions, allowing them to exploit known object semantics but limiting transfer when target identities are unavailable or differ substantially across domains (Jeong et al., 2023). This limitation motivated the shift toward object-agnostic learnable prompts, which decouple anomaly semantics from category names and support more transferable visual-text alignment across heterogeneous targets (Zhou et al., 2024).

Despite these encouraging advancements, two major limitations persist. First, the majority of prompt-based detectors collapse the definition of abnormality into a single normal-anomaly opposition. However, real-world anomalies are intrinsically multi-faceted: a defect might manifest as

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.

Algorithm 1 Training and testing of CRAG-CLIP

Input: Source dataset $\mathcal{D}_s$; frozen CLIP E_v, E_t; descriptor pairs $\{(c_k^n, c_k^a)\}_{k=1}^K$.
Output: Pixel-level anomaly map $\mathbf{M}$ and image-level score s_{img}.

Initialize: learnable contexts $\{\mathbf{C}_k^n, \mathbf{C}_k^a\}$, deep prompts, and frozen language anchors $\{\mathbf{d}_k^{\text{lang}}\}_{k=1}^K$.
for each epoch **do**
  for each batch $(\mathbf{x}, y, \mathbf{Y}) \in \mathcal{D}_s$ **do**
    $\mathbf{g}, \{\mathbf{p}^{(l)}\}_{l\in\mathcal{L}} \leftarrow E_v(\mathbf{x})$; $\{\mathbf{T}^n, \mathbf{T}^a\} \leftarrow E_t(\text{prompts})$
    $\mathbf{d}_k \leftarrow \text{norm}(\mathbf{t}_k^a - \mathbf{t}_k^n),\ \forall k$; for each layer $l \in \mathcal{L}$: compute $s^{(l)}$ and $\boldsymbol{\pi}^{(l)}$ via Eqs. 1–2
    $q_n^{(l)} \leftarrow 1 - \frac{H(\boldsymbol{\pi}_n^{(l)})}{\log K} - \lambda_\delta \frac{D_{\text{KL}}(\boldsymbol{\pi}_n^{(l)} \,\|\, \bar{\boldsymbol{\pi}}_n)}{\log K}$ {Quality score}
    $\hat{\mathbf{s}} \leftarrow \sum_{l=1}^L \text{softmax}(q^{(l)}/\tau_l) \cdot s^{(l)}$; $s_{\text{patch}} \leftarrow \text{CVaR}_q(\hat{\mathbf{s}})$; $s_{\text{global}} \leftarrow s(\mathbf{g})$
    $s_{\text{img}} \leftarrow \text{SmoothMax}(s_{\text{global}}, s_{\text{patch}})$ {Aggregate evidence}
    Compute loss $\mathcal{L}$ (Eq. 10) and update learnable prompts.
  end for
end for
Inference: for a target image $\mathbf{x}_{\text{target}}$, repeat Steps 4–8 without gradient updates; $\mathbf{M} \leftarrow \text{Upsample}(\hat{\mathbf{s}})$
return $\mathbf{M}$ and s_{img}

a structural crack, surface contamination, missing components, or color discoloration, and each type corresponds to a distinct semantic direction within the CLIP text space.

Consolidating these variations into a singular prompt discards the rich relational structure distinguishing different anomaly types, thereby impairing zero-shot transferability to targets whose dominant anomaly type differs substantially from the source. Second, the aggregation of visual evidence across network layers often relies on fixed heuristics that disregard the inherent reliability of individual layers: previous approaches average feature layers uniformly or combine them with static weights, while extracting the image-level score from a single global representation. These choices fail to acknowledge that different transformer layers encode anomaly cues at varying granularities (e.g., texture, structure, and high-level semantics), and that their agreement is critically informative for determining the genuine anomalousness of a local patch.

To address these challenges, we propose CRAG-CLIP (cross-layer relational abnormality geometry for CLIP), an innovative framework that formulates anomaly detection as reasoning over a relational geometry in the text space, harmonized across multiple visual layers through posterior consensus. Specifically, CRAG-CLIP contributes three tightly coupled modules:

- **Language-Anchored Paired Axes (LAPA).** We conceptualize abnormality as K paired, object-agnostic normal-anomaly descriptor relations (e.g., intact-broken, clean-contaminated), with each relation instantiated by independent learnable prompt contexts. To ensure cross-domain generalization, we anchor the learned directions to frozen CLIP text directions derived from the descriptor sentences, preventing source-domain semantic drift. Furthermore, a Gram-based diversity regularizer is introduced to maintain semantic separation among the axes.
- **Cross-Layer Posterior Consensus (CLPC).** We extract multi-level patch features from various CLIP transformer layers and compute a posterior distribution over the K anomaly axes for each patch. Subsequently, patches are fused across layers using a per-patch quality score that explicitly rewards high confidence (low posterior entropy) and penalizes the Kullback-Leibler divergence between the individual layer's posterior and the cross-layer consensus. This yields a dynamic, data-dependent layer weighting strategy.
- **Global-Local Evidence Aggregation (GLEA).** For image-level anomaly scoring, we aggregate evidence via a smooth-max function applied to the global CLIP logit and a top-q conditional value-at-risk (CVaR) of the fused local patch evidence. This aggregation ensures that the final image score captures both the presence of salient local anomalies and the holistic global semantic context, maintaining strict consistency between training and inference phases.

Crucially, all three modules adhere to a unified design philosophy: anomaly evidence is fundamentally defined by relations between normal and abnormal text directions and by the agreement

across multiple layers and scales. CRAG-CLIP necessitates zero target-domain adaptation, test-time memory banks, or source normal prototypes.

2 Related Work

2.1 Vision-Language Models

Vision–language models, represented by CLIP (Radford et al., 2021), learn a shared visual–textual embedding space through contrastive pre-training on large-scale image–text pairs. By comparing globally pooled image features with text embeddings used as semantic prototypes, CLIP enables open-vocabulary recognition and transfer without updating the pretrained encoders. Prompt learning further adapts this interface efficiently: CoOp (Zhou et al., 2022b) optimizes continuous context tokens, CoCoOp (Zhou et al., 2022a) introduces image-conditioned prompts, and MaPLe (Khattak et al., 2023) performs hierarchical prompting in both visual and textual branches. However, the original CLIP objective supervises only image-level alignment. RegionCLIP (Zhong et al., 2021) learns region–text alignment, DenseCLIP (Rao et al., 2022) extends image–text matching to pixel-level prediction, and MaskCLIP (Dong et al., 2023) strengthens local patch representations through masked self-distillation, indicating that effective local modeling still requires dedicated adaptation mechanisms.

2.2 Zero-Shot Image Anomaly Detection

Zero-shot image anomaly detection aims to identify and localize anomalies in unseen categories or domains without target-specific training data. Early CLIP-based methods established the feasibility of this setting. WinCLIP (Jeong et al., 2023) combines handcrafted state-word prompt ensembles with window-level visual features for zero-shot anomaly classification and segmentation. AnomalyCLIP (Zhou et al., 2024) replaces category-dependent descriptions with object-agnostic learnable prompts and introduces dual-path attention modules to improve patch-level localization. More recent methods further strengthen anomaly-aware adaptation, AdaCLIP (Cao et al., 2024) combines static and image-conditioned prompts, APRIL-GAN (Chen et al., 2023) adapts CLIP with lightweight linear projections, and FAPrompt (Zhu et al., 2025) models decomposed fine-grained abnormality semantics with a data-dependent prior.

Despite these advances, strict source-only transfer remains challenging because anomalies manifest through multiple semantic relations and multi-layer patch evidence may be uncertain under substantial domain shifts. CRAG-CLIP addresses this by modeling abnormality as a cross-layer relational geometry.

3 Method

3.1 Overview and Notation

Figure 1 presents the overall architecture of CRAG-CLIP, which is trained on a labeled source dataset and directly evaluated on previously unseen industrial and medical datasets. Given a frozen CLIP image encoder E_v and text encoder E_t, CRAG-CLIP performs anomaly reasoning using K learnable object-agnostic normal–anomaly prompt pairs and multi-level visual evidence. The CLIP backbone, including the diagonally prominent attention map (DPAM) mechanism (Zhou et al., 2024), remains frozen throughout training, while only the shallow and deep continuous prompt parameters are optimized.

Specifically, the Q–K pathway provides the ℓ_2-normalized global CLS feature $\mathbf{g} \in \mathbb{R}^d$, whereas the frozen DPAM branch produces ℓ_2-normalized local patch features $\mathbf{p}^{(l)} \in \mathbb{R}^{N\times d}$ from a predefined set of layers $\mathcal{L} = \{l_1, \ldots, l_L\}$, where N and d denote the number of patches and feature dimension, respectively. For the k-th semantic axis, the frozen text encoder outputs normalized normal and anomaly text embeddings $\mathbf{t}_k^n, \mathbf{t}_k^a \in \mathbb{R}^d$. LAPA converts the paired visual–text similarities into complementary semantic evidence, CLPC reliability-weights the layer-wise local responses into a fused anomaly map, and GLEA combines the global semantic evidence with the local cue to produce

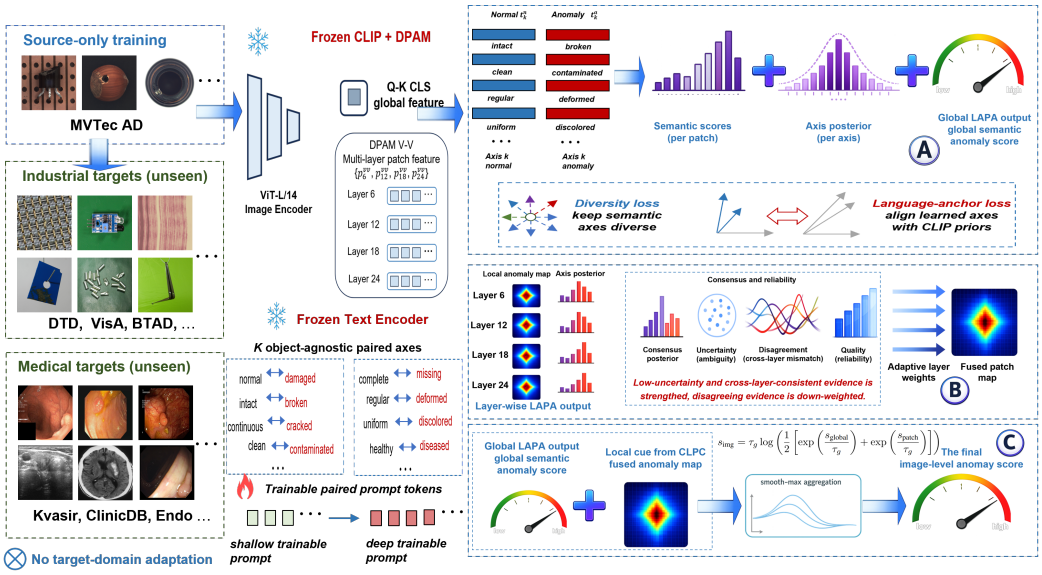


Figure 1: Overview of CRAG-CLIP under the source-only zero-shot evaluation protocol. The model is trained on a single source dataset and directly evaluated on previously unseen datasets without target-data adaptation. The frozen CLIP image encoder with the frozen DPAM branch extracts a Q–K global CLS feature and V–V multi-layer patch features from layers $\{6, 12, 18, 24\}$, while the frozen text encoder embeds eight learnable object-agnostic normal–anomaly prompt pairs. LAPA converts paired semantic margins into axis posteriors and a global anomaly score, under diversity and language-anchor regularization. CLPC estimates the reliability of layer-wise patch evidence from posterior uncertainty and cross-layer disagreement, and fuses the resulting local anomaly maps. GLEA combines the global semantic score with the fused local cue through normalized smooth-max aggregation to produce the final image-level anomaly score.

the image-level anomaly score; all margin computations use the same paired semantic axes for both global and local features.

3.2 Language-Anchored Paired Axes (LAPA)

CRAG-CLIP abandons the conventional single normal-anomaly prompt strategy in favor of K paired axes. Each axis $k \in \{1, \ldots, K\}$ is explicitly associated with a pair of object-agnostic semantic descriptors (c_k^n, c_k^a) (e.g., intact–broken). For each axis, we introduce independent learnable context tokens $\mathbf{C}_k^n, \mathbf{C}_k^a \in \mathbb{R}^{M \times d}$ of length M. These are concatenated with a fixed prefix and a suffix to formulate the complete prompt, which is subsequently processed by the frozen text encoder E_t.

The anomaly direction vector for axis k is defined as $\mathbf{d}_k = \mathrm{norm}(\mathbf{t}_k^a - \mathbf{t}_k^n)$, where $\mathrm{norm}(\cdot)$ denotes ℓ_2-normalization. For a given visual feature $\mathbf{v} \in \mathbb{R}^d$, the anomaly margin with respect to axis k quantifies the extent to which the feature aligns more closely with the abnormal concept:

$$m_k(\mathbf{v}) = \frac{\mathbf{v}^\top \mathbf{t}_k^a - \mathbf{v}^\top \mathbf{t}_k^n}{\tau_s}, \tag{1}$$

where τ_s is a scalar temperature hyper-parameter controlling the sharpness of the margin.

To consolidate the evidence from all K axes into a robust scalar anomaly score while preserving invariance to the total number of axes, we employ a log-mean-exp (LME) pooling strategy:

$$s(\mathbf{v}) = \tau_m \log \left(\frac{1}{K} \sum_{k=1}^{K} \exp \left(\frac{m_k(\mathbf{v})}{\tau_m} \right) \right), \tag{2}$$

where $s(\mathbf{v}) \in \mathbb{R}$ denotes the aggregated evidence for feature $\mathbf{v}$, and τ_m is the mixture temperature. The posterior probability distribution over the K axes is computed via the softmax function:

$$\boldsymbol{\pi}(\mathbf{v}) = \mathrm{softmax}\left(\frac{\mathbf{m}(\mathbf{v})}{\tau_p}\right), \tag{3}$$

where $\boldsymbol{\pi}(\mathbf{v}) \in \mathbb{R}^K$ is the posterior vector, $\mathbf{m}(\mathbf{v}) = [m_1(\mathbf{v}), \ldots, m_K(\mathbf{v})]^\top$, and τ_p is the posterior temperature.

To prevent the learnable directions from overfitting to source-domain shortcuts, each learned direction is anchored to a frozen text direction $\mathbf{d}_k^{\text{lang}}$ derived purely from the original descriptor sentences. The language anchor loss is formulated as the average cosine distance:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{K}\sum_{k=1}^{K}\big(1 - \mathbf{d}_k^\top \mathbf{d}_k^{\text{lang}}\big). \tag{4}$$

Furthermore, a Gram matrix-based diversity loss is applied to prevent semantic collapse among the K learned directions:

$$\mathcal{L}_{\text{div}} = \frac{2}{K(K-1)}\sum_{i<j}\big(\mathbf{d}_i^\top \mathbf{d}_j\big)^2. \tag{5}$$

3.3 Cross-Layer Posterior Consensus (CLPC)

Different CLIP transformer layers capture visual features at varying semantic levels. Let $s_n^{(l)}$ and $\boldsymbol{\pi}_n^{(l)}$ denote the LME evidence and the axis posterior for the n-th patch at the l-th layer, respectively. We compute the cross-layer consensus posterior as $\bar{\boldsymbol{\pi}}_n = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\pi}_n^{(l)}$. The per-patch quality score $q_n^{(l)}$ is designed to reward high posterior confidence and penalize divergence from the multi-layer consensus:

$$q_n^{(l)} = 1 - \frac{H(\boldsymbol{\pi}_n^{(l)})}{\log K} - \lambda_\delta \frac{D_{\mathrm{KL}}\big(\boldsymbol{\pi}_n^{(l)} \,\|\, \bar{\boldsymbol{\pi}}_n\big)}{\log K}, \tag{6}$$

where $H(\cdot)$ represents the Shannon entropy, $D_{\mathrm{KL}}(\cdot\|\cdot)$ is the Kullback-Leibler divergence, and λ_δ controls the weight of the disagreement penalty. The layer-specific weights $w_n^{(l)}$ are then calculated using a temperature-scaled softmax, producing the final fused anomaly evidence $\hat{s}_n$:

$$w_n^{(l)} = \frac{\exp(q_n^{(l)}/\tau_l)}{\sum_{j=1}^{L}\exp(q_n^{(j)}/\tau_l)}, \qquad \hat{s}_n = \sum_{l=1}^{L} w_n^{(l)}\, s_n^{(l)}, \tag{7}$$

where τ_l is the layer fusion temperature.

3.4 Global-Local Evidence Aggregation (GLEA)

To summarize the local patch evidence robustly, we utilize a top-q conditional value-at-risk (CVaR) formulation over the fused patch evidence vector $\hat{\mathbf{s}} \in \mathbb{R}^N$:

$$s_{\text{patch}} = \frac{1}{\lceil qN \rceil}\sum_{i=1}^{\lceil qN \rceil}\hat{s}_{(i)}, \tag{8}$$

where $q \in (0, 1]$ is the specified fraction of top anomalous patches, and $\hat{s}_{(i)}$ denotes the i-th largest fused patch score.

The global CLS token feature $\mathbf{g}$ is evaluated using the LAPA function (Eq. 2) to yield s_{global}. The final image-level anomaly score s_{img} is synthesized using a smooth-max (log-sum-exp) fusion of both branches:

$$s_{\text{img}} = \tau_g \log\left(\frac{1}{2}\left[\exp\left(\frac{s_{\text{global}}}{\tau_g}\right) + \exp\left(\frac{s_{\text{patch}}}{\tau_g}\right)\right]\right), \tag{9}$$

where τ_g is the image fusion temperature. This differentiable aggregation mechanism is applied symmetrically during both training and inference.

3.5 Training and Inference Pipeline

During training, CRAG-CLIP is optimized end-to-end on the source domain using a composite objective function. The pixel-level fused patch evidence $\hat{\mathbf{s}}$ is supervised by the ground truth mask $\mathbf{Y}$ via Dice and Focal losses, while the image-level score s_{img}, global score s_{global}, and patch score s_{patch} are jointly optimized with binary cross-entropy (BCE) loss against the image label y. Incorporating the relational constraints from the LAPA module, the overall loss $\mathcal{L}_{\text{total}}$ is formally defined as:

$$\begin{aligned}\mathcal{L}_{\text{total}} = &\underbrace{\text{BCE}(s_{\text{img}}, y)}_{\text{Classification}} + \underbrace{\lambda_{\text{aux}}\mathcal{L}_{\text{comp}}}_{\text{Auxiliary Cls}} \\ &+ \underbrace{\lambda_{\text{seg}}\big(\text{Dice}(\hat{\mathbf{s}}, \mathbf{Y}) + \text{Focal}(\hat{\mathbf{s}}, \mathbf{Y})\big)}_{\text{Segmentation}} \\ &+ \underbrace{\lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{anchor}}\mathcal{L}_{\text{anchor}}}_{\text{Relational Constraints}},\end{aligned} \tag{10}$$

where $\mathcal{L}_{\text{comp}} = \frac{1}{2}\big(\text{BCE}(s_{\text{global}}, y) + \text{BCE}(s_{\text{patch}}, y)\big)$ denotes the auxiliary classification constraints, and $\lambda_{\text{seg}}, \lambda_{\text{aux}}, \lambda_{\text{div}}, \lambda_{\text{anchor}}$ are balancing hyper-parameters.

During inference, CRAG-CLIP operates in a strictly zero-shot manner without any target-domain adaptation. For an unseen target image $\mathbf{x}$, the pixel-level anomaly map $\mathbf{M}$ is obtained by aggregating and bilinearly up-sampling the cross-layer fused patch evidence $\hat{\mathbf{s}} = \sum_{l\in\mathcal{L}} w^{(l)}\, s^{(l)}$ to the original spatial resolution (H, W), and the image-level anomaly score is computed by GLEA via the smooth-max fusion in Eq. 9. A higher anomaly score s indicates a stronger probability that the input image contains anomalous patterns. Algorithm 1 summarizes the complete training and inference procedure of CRAG-CLIP.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We evaluate CRAG-CLIP on five industrial and nine medical anomaly detection datasets. The industrial benchmarks include MVTec AD (Bergmann et al., 2019), VisA (Zou et al., 2022), MPDD (Jezek et al., 2021), BTAD (Mishra et al., 2021), and DTD-Synthetic (Aota et al., 2023). For medical anomaly detection, HeadCT (Salehi et al., 2021), BrainMRI (Salehi et al., 2021), and Br35H (Hamada, 2020) are used for image-level evaluation, whereas ISIC (Codella et al., 2018), ClinicDB (Bernal et al., 2015), ColonDB (Tajbakhsh et al., 2015), Kvasir (Jha et al., 2019), Endo (Hicks et al., 2021), and TN3K (Gong et al., 2021) are used for pixel-level evaluation. Under a strict source-only zero-shot protocol, CRAG-CLIP is trained exclusively on MVTec AD and directly evaluated on all remaining datasets. To further assess transferability in the reverse direction, we additionally train CRAG-CLIP on VisA and directly evaluate it on MVTec AD under the same protocol. Performance is assessed using AUROC and AP at the image level, and AUROC and PRO at the pixel level (Bergmann et al., 2019).

Implementation Details. We adopt the pretrained CLIP ViT-L/14@336px (Radford et al., 2021) as a frozen backbone, incorporating DPAM at each of the selected feature-extraction layers, $\{6, 12, 18, 24\}$, whose outputs are used as multi-level representations. We employ eight normal-abnormal semantic pairs, with the prompt length, prompt depth, and deep-context length set to 12, 9, and 4, respectively, the semantic, mixture, and layer-fusion temperatures are set to 0.07, 1.0, and 0.25, respectively, while the disagreement weight and CVaR ratio are set to 1.0 and 5%. Training is performed for up to 15 epochs using AdamW with a batch size of 4, a learning rate of 10^{-3}, and a weight decay of 10^{-4}, based on a $90/10$ class- and label-stratified source-domain train/validation split and early stopping with a patience of three epochs. The auxiliary segmentation and image-level loss terms are each weighted by 0.25, whereas the diversity and language-anchor regularizers are each weighted by 0.05. All experiments are conducted using PyTorch 2.9.1 on an NVIDIA RTX PRO 6000 Blackwell Server Edition GPU.

Table 1: Zero-shot anomaly detection performance comparison on industrial and medical domains. The models are trained on MVTec AD unless specified otherwise. Best results are in **bold**, and second-best are underlined.

	Metric	Datasets	CLIP	WinCLIP	APRIL-GAN	AnomalyCLIP	AdaCLIP	FAPrompt	**CRAG-CLIP**
Industrial	Image-level (AUROC, AP)	MVTec AD	(74.1, 87.6)	(91.8, **96.5**)	(86.1, 93.5)	(91.5, 96.2)	(**92.0**, 96.4)	(91.9, 95.7)	(90.4, 95.3)
		VisA	(66.4, 71.5)	(78.1, 81.2)	(78.0, 81.4)	(81.5, 84.3)	(83.0, 84.9)	(**84.5**, **86.8**)	(83.5, 86.1)
		MPDD	(54.3, 65.4)	(63.6, 69.9)	(73.0, 80.2)	(77.0, 82.0)	(72.1, 77.6)	(**80.6**, **83.3**)	(76.1, 80.0)
		BTAD	(34.5, 52.5)	(68.2, 70.9)	(73.6, 68.6)	(88.3, 87.3)	(91.6, 92.4)	(92.0, 92.2)	(**92.4**, **94.5**)
		DTD-Synthetic	(71.6, 85.7)	(93.2, 92.6)	(86.4, 95.0)	(93.5, 97.0)	(92.8, 97.0)	(**95.9**, **98.3**)	(95.6, 97.8)
		Average	(60.2, 72.5)	(79.0, 82.2)	(79.4, 83.7)	(86.4, 89.4)	(86.3, 89.7)	(**89.0**, **91.3**)	(87.6, 90.7)
	Pixel-level (AUROC, PRO)	MVTec AD	(38.4, 11.3)	(85.1, 64.6)	(87.6, 44.0)	(**91.0**, 80.4)	(86.8, 33.8)	(90.6, **83.3**)	(87.9, 81.0)
		VisA	(46.6, 14.8)	(79.6, 56.8)	(94.2, 86.8)	(89.2, 77.1)	(95.1, 71.3)	(**95.9**, **87.5**)	(95.1, 86.7)
		MPDD	(62.1, 33.0)	(76.4, 48.9)	(94.1, 83.2)	(**96.5**, 88.7)	(96.4, 62.2)	(**96.5**, 87.9)	(96.2, **89.2**)
		BTAD	(30.6, 4.4)	(72.7, 27.3)	(60.8, 25.0)	(94.2, 74.8)	(87.7, 17.7)	(**95.6**, 75.2)	(93.3, **76.3**)
		DTD-Synthetic	(33.9, 12.5)	(83.9, 57.8)	(95.3, 86.9)	(97.9, 92.3)	(94.1, 24.9)	(**98.3**, **93.1**)	(96.9, 91.6)
		Average	(42.3, 15.2)	(79.5, 51.1)	(86.4, 65.2)	(93.8, 82.7)	(92.0, 42.0)	(**95.4**, **85.4**)	(93.9, 85.0)
Medical	Image-level (AUROC, AP)	HeadCT	(56.5, 58.4)	(81.8, 80.2)	(89.1, 89.4)	(93.4, 91.6)	(93.4, 92.2)	(94.8, 93.5)	(**94.9**, **95.2**)
		BrainMRI	(73.9, 81.7)	(86.6, 91.5)	(89.3, 90.9)	(90.3, 92.2)	(94.9, 94.2)	(95.5, 95.6)	(**96.9**, **97.9**)
		Br35H	(78.4, 78.8)	(80.5, 82.2)	(93.1, 92.9)	(94.6, 94.7)	(95.7, 95.7)	(97.8, 97.5)	(**98.2**, **98.2**)
		Average	(69.6, 73.0)	(83.0, 84.6)	(90.5, 91.1)	(92.8, 92.8)	(94.7, 94.0)	(96.0, 95.5)	(**96.7**, **97.1**)
	Pixel-level (AUROC, PRO)	ISIC	(33.1, 5.8)	(83.3, 55.1)	(89.4, 77.2)	(89.7, 78.4)	(85.4, 5.3)	(**90.9**, **81.2**)	(83.3, 69.9)
		ClinicDB	(47.5, 18.9)	(51.2, 13.8)	(80.5, 60.7)	(82.9, 67.8)	(**85.3**, 14.6)	(84.7, 70.1)	(84.9, **71.2**)
		ColonDB	(49.5, 15.8)	(70.3, 32.5)	(78.4, 64.6)	(81.9, 71.3)	(79.3, 6.5)	(**84.6**, **74.7**)	(81.8, 73.0)
		Kvasir	(44.6, 17.7)	(69.7, 24.5)	(75.0, 36.2)	(78.9, 45.6)	(79.4, 12.3)	(**81.2**, 47.8)	(79.8, **48.4**)
		Endo	(45.2, 15.9)	(68.2, 28.3)	(81.9, 54.9)	(84.1, 63.6)	(84.0, 10.5)	(**86.4**, **67.2**)	(83.9, 55.3)
		TN3K	(70.7, 39.8)	(73.6, 37.8)	(81.5, 50.4)	(77.2, 46.8)	(80.5, 45.8)	(**84.7**, **54.6**)	(74.9, 47.4)
		Average	(48.4, 20.3)	(71.1, 35.8)	(81.5, 58.8)	(81.8, 62.3)	(83.2, 43.7)	(**85.5**, **65.9**)	(81.4, 60.9)

4.2 Comparisons with State-of-the-Art Methods

Quantitative Comparisons. Table 1 presents quantitative comparisons with state-of-the-art zero-shot baselines under the source-only protocol. For industrial image-level detection, CRAG-CLIP attains the second-best average AUROC of 87.6 and AP of 90.7, which are 1.4 and 0.6 points below the corresponding results of FAPrompt (Zhu et al., 2025). It is nevertheless the best method on BTAD (92.4 AUROC, 94.5 AP). For industrial pixel-level localization, CRAG-CLIP again ranks second overall, with an average AUROC of 93.9 and PRO of 85.0. In particular, it obtains the highest PRO on MPDD at 89.2 and on BTAD at 76.3. The advantage of CRAG-CLIP is most pronounced for medical image-level anomaly detection. It achieves the best average results, with an AUROC of 96.7 and an AP of 97.1, improving upon FAPrompt by 0.7 AUROC points and 1.6 AP points. CRAG-CLIP also ranks first on HeadCT, BrainMRI, and Br35H, where it achieves AUROC and AP scores of 94.9 and 95.2, 96.9 and 97.9, and 98.2 and 98.2, respectively. Medical pixel-level localization exhibits greater variation across datasets. CRAG-CLIP attains an average AUROC of 81.4 and PRO of 60.9, while achieving the highest PRO on ClinicDB at 71.2 and on Kvasir at 48.4. Overall, the results indicate that CRAG-CLIP provides particularly strong cross-domain generalization for image-level anomaly detection, especially in medical domains with substantial semantic shifts.

Qualitative Comparisons. Figure 2 compares the pixel-level anomaly maps of AnomalyCLIP and CRAG-CLIP on unseen industrial and medical targets. Across the heterogeneous appearances of VisA, MPDD, BTAD, DTD-Synthetic, and four medical datasets, CRAG-CLIP produces anomaly responses that are more spatially coherent and better aligned with the annotated defect regions. In particular, it more reliably captures subtle local defects (e.g., VisA and DTD-Synthetic) and extended irregular regions with diverse morphologies (e.g., MPDD, ClinicDB, ColonDB, and Kvasir), while preserving discriminative responses on ultrasound data (TN3K); AnomalyCLIP often yields weaker, fragmented, or incomplete responses under substantial domain shifts. These results qualitatively support that the complementary semantic axes, reliability-aware cross-layer fusion, and global–local aggregation improve the transferability and localization robustness of CRAG-CLIP.

Analysis of Reliability-Aware Layer Fusion. Figure 3 contrasts AnomalyCLIP, which assigns identical normalized contributions (0.25) to all selected layers irrespective of their reliability, with CRAG-CLIP, which allocates fusion weights according to posterior uncertainty and cross-layer dis-

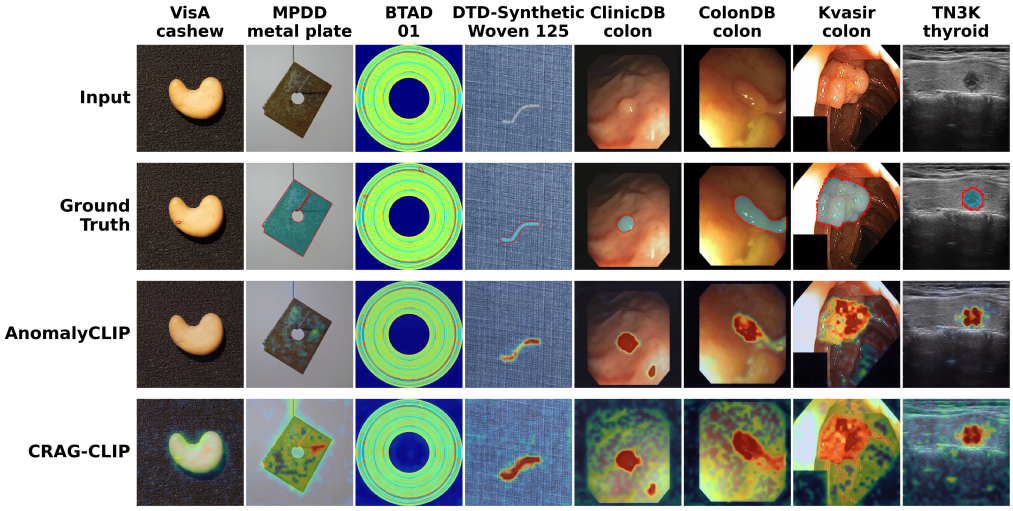


Figure 2: Qualitative comparison of anomaly localization results on representative industrial and medical datasets. From left to right: VisA, MPDD, BTAD, DTD-Synthetic, ClinicDB, ColonDB, Kvasir, and TN3K.

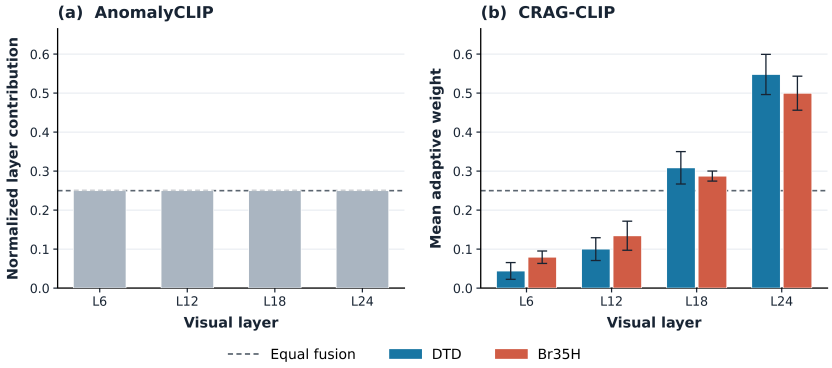


Figure 3: Comparison between fixed and reliability-aware multi-layer fusion. (a) AnomalyCLIP equally sums the anomaly maps from layers $\{6, 12, 18, 24\}$, fixing each layer's normalized contribution at 0.25. (b) CRAG-CLIP derives patch-wise layer weights from the uncertainty and cross-layer disagreement of the eight-axis semantic posteriors. Bars report the mean image-wise weights over 12 deterministically selected anomalous test images from DTD and Br35H; error bars denote the standard deviation across images.

agreement, as defined by the quality score in Eq. 6. On anomalous samples from DTD and Br35H, the learned weights differ substantially across layers and consistently emphasize deeper layers (especially 18 and 24), suppressing less reliable shallow-layer evidence. This verifies that CLPC performs reliability-aware, input-dependent fusion rather than fixed equal-weight aggregation, explaining the localization gains in Table 1.

4.3 Ablation Study

We ablate LAPA, CLPC, and GLEA, summarizing performance with evaluation metrics averaged over all target domains.

Language-Anchored Paired Axes (LAPA). Table 2 shows that full LAPA achieves an image-level AUROC of 91.1 and a pixel-level AUROC of 87.0. Collapsing the formulation to a single semantic axis produces the largest degradation (1.8 points in pixel-level AUROC and 3.0 points in PRO), demonstrating that cross-domain abnormalities cannot be adequately represented by a single

Table 2: Ablation study on components of the language-anchored paired axes (LAPA) module.

LAPA Variants	Image-level		Pixel-level	
	AUROC	AP	AUROC	PRO
Single Axis ($K = 1$)	90.1	91.9	85.2	67.9
Unpaired Axes	90.9	92.4	86.7	70.6
w/o Language Anchor	90.9	92.6	86.7	**71.5**
w/o Diversity Loss	90.8	92.2	86.7	70.8
Full LAPA (Ours)	**91.1**	**92.8**	**87.0**	70.9

Table 3: Ablation study on fusion strategies in cross-layer posterior consensus (CLPC).

CLPC Variants	Image-level		Pixel-level	
	AUROC	AP	AUROC	PRO
Uniform Layer Weights	90.7	92.2	86.7	70.8
w/o Disagreement Penalty	**91.1**	92.8	86.9	70.8
Full CLPC (Ours)	**91.1**	**92.8**	**87.0**	**70.9**

Table 4: Ablation study on image-level scoring strategies in GLEA.

GLEA Variants	Image-level		Pixel-level	
	AUROC	AP	AUROC	PRO
Global Evidence Only	89.7	91.3	87.3	71.5
Patch Evidence Only	90.3	91.8	**87.6**	**71.7**
Arithmetic Mean Fusion	91.0	**92.8**	87.1	70.7
Full GLEA (Ours)	**91.1**	**92.8**	87.0	70.9

relation between normality and abnormality. Removing paired relations or diversity regularization also consistently reduces all four metrics, confirming the importance of non-redundant axes. The language anchor improves the remaining three metrics, although removing it increases PRO from 70.9 to 71.5. Mechanistically, the language anchor keeps each learned direction aligned with its frozen descriptor-derived CLIP direction, mitigating source-specific semantic drift, while the diversity loss penalizes redundant directions, preserving complementary coverage.

Cross-Layer Posterior Consensus (CLPC). Table 3 isolates the contribution of our data-dependent fusion mechanism. Substituting CLPC with uniform, fixed layer weights directly diminishes pixel-level PRO, verifying that heuristic averaging fails to capture the varying semantic reliability of layers, while removing the disagreement penalty also yields suboptimal fusion. Inter-layer agreement thus acts as an effective self-regularizing signal for precise localization.

Global-Local Evidence Aggregation (GLEA). Table 4 validates the image-level aggregation design. Relying solely on global CLIP features ignores small local defects, while using only patch-level CVaR discards holistic scene semantics. Replacing the smooth-max function with an arithmetic mean likewise degrades performance, showing that our soft-maximum aggregation captures salient anomalies without suppressing global context.

5 Conclusion

We presented CRAG-CLIP, a zero-shot anomaly detection framework that reformulates anomaly reasoning as a cross-layer relational abnormality geometry in the text space, overcoming single-prompt heuristics via LAPA, CLPC, and GLEA. Extensive evaluations across industrial and medical datasets demonstrate competitive zero-shot transferability, especially for medical image-level detection, without target-data adaptation.

REPRODUCIBILITY STATEMENT

This work is fully empirical and does not involve human subjects or private data collection. To support reproducibility, Algorithm 1 in the main text summarizes the complete training and inference procedure of CRAG-CLIP, and the Implementation Details paragraph in the Experiments section reports all architectural and hyper-parameter settings. Appendix A further provides the checkpoint-resolved configuration, including the frozen backbone and DPAM setup, the prompt parameterization, the eight paired semantic axes, the source-domain training and checkpoint-selection protocol, and the frozen target-domain evaluation protocol, so that the entire pipeline can be reproduced without any target-domain information. Appendix B reports the hyperparameter sensitivity analysis, Appendix C reports detailed per-category quantitative results for every evaluated dataset, and Appendix D presents category-wise qualitative examples generated with the exact checkpoint used for the reported quantitative results. All fourteen datasets used in this work are publicly available and cited accordingly, and we will release the complete source code used for conducting and analyzing the experiments upon publication.

ETHICS STATEMENT

This work uses only publicly available academic benchmarks for industrial inspection and medical imaging, all of which are properly cited and used for their intended research purposes. No new human-subject data are collected, and no personally identifiable information is processed, since the medical datasets are de-identified public research benchmarks. The proposed method is intended to assist anomaly screening in industrial and medical scenarios and is not a substitute for professional diagnosis; any deployment in safety-critical applications would require further validation, calibration, and regulatory compliance. Moreover, because CRAG-CLIP follows a strict source-only protocol and never adapts to target-domain data, it reduces the risk of transferring sensitive information across institutions or domains. We do not foresee harmful uses of this work beyond those common to anomaly detection research.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the language, improving the clarity of presentation, and assisting with the LaTeX formatting of the manuscript. We have not used generative AI tools for generating research ideas, designing the methodology, implementing the model, conducting experiments, or analyzing results, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for checking citation and notation consistency. We have reviewed all AI-assisted work: every AI-polished passage was checked by the authors against the original technical content to ensure factual accuracy, and no AI-generated text was used without verification. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

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, 2023.

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.

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, 2024.

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, 2023.

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, 2018.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10995–11005, June 2023.

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

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, 2021.

A. Hamada. Br35H: Brain tumor detection. `https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection`, 2020. Kaggle dataset.

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

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, 2021.

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

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

Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *The Twelfth International Conference on Learning Representations*, pp. 2172–2197, 2024.

Zhuo Li, Yuhao Yan, Xiangheng Wang, Yifei Ge, and Lin Meng. A survey of deep learning for industrial visual anomaly detection. *Artificial Intelligence Review*, 58(9):279, 2025.

Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1): 104–135, 2024.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. *arXiv preprint arXiv:2104.10036*, 2021.

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*, volume 139, pp. 8748–8763, 2021.

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

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

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.

Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 71:1–21, 2022.

Maximilian E Tschuchnig and Michael Gadermayr. Anomaly detection in medical imaging-a mini review. In *International Data Science Conference*, pp. 33–38, 2021.

Yi Zhang, Jiawen Zhu, Lele Fu, and Guansong Pang. Anomalyagent: Training-free agentic models for zero-/few-shot anomaly detection, 2026.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining, 2021.

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*, pp. 16816–16825, 2022a.

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

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*, pp. 49705–49737, 2024.

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*, pp. 22241–22251, 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, 2022.

A Reproducible Experimental Configuration

All values in this section are recovered from the exact CRAG-CLIP checkpoint used for the reported quantitative and qualitative results. The configuration is stated explicitly to make the experimental protocol reproducible without relying on target-domain information.

A.1 Architecture and Prompt Configuration

CRAG-CLIP uses a frozen CLIP visual–language backbone and optimizes only the prompt parameters. The image encoder, including the DPAM branch, remains frozen throughout training and target-domain evaluation. Table 5 lists the checkpoint-resolved architecture, prompt, and inference settings.

Table 5: Checkpoint-resolved architecture and prompt configuration. All CLIP backbone and DPAM parameters are frozen; only the continuous prompt parameters are optimized.

Component	Checkpoint-resolved setting
Visual–language backbone	Frozen CLIP ViT-L/14@336px; DPAM is applied across the final 19 transformer blocks (layers 6–24), using a DPAM boundary index of 20.
Input and visual layers	Input resolution 518×518; patch features from transformer layers $\{6, 12, 18, 24\}$.
Prompt parameterization	12 shallow context tokens, prompt depth 9, and 4 deep context tokens for each semantic axis.
Paired semantic axes	normal–damaged; intact–broken; continuous–cracked; clean–contaminated; complete–missing; regular–deformed; uniform–discolored; healthy–diseased.
LAPA temperatures	Semantic-margin temperature $\tau_s = 0.07$ and log-mean-exp mixture temperature $\tau_m = 1.0$.
CLPC settings	Posterior temperature $\tau_p = 1.0$, disagreement weight $\lambda_\delta = 1.0$, and layer-fusion temperature $\tau_l = 0.25$.
GLEA settings	CVaR/top-patch fraction $q = 0.05$ and global–local smooth-max temperature $\tau_g = 1.0$.

A.2 Source-Domain Training and Checkpoint Selection

Training is performed exclusively on MVTec AD. The configured source metadata split is test; it is partitioned internally into source-domain training and validation subsets, stratified jointly by category and image label. No target dataset is provided to the training procedure. The selected best.pth checkpoint corresponds to epoch 4 and is determined solely by the source validation objective.

A.3 Frozen Target-Domain Evaluation Protocol

At evaluation time, the selected checkpoint is placed in evaluation mode and applied directly to each target test category with batch size 1. No parameters are updated, and no target-domain statistics are accumulated. The anomaly map is Gaussian-smoothed with standard deviation 4.0 before pixel-level metric computation. Table 7 specifies the exact reporting protocol.

B Hyperparameter Sensitivity Analysis

Figure 4 examines the sensitivity of three key design choices on the Br35H and DTD target domains. We vary the semantic-axis number K, the log-mean-exp mixture temperature τ_m, and the CLPC layer-fusion temperature τ_l, while fixing all other settings. The default configuration ($K = 8$, $\tau_m = 1.0$, and $\tau_l = 0.25$) achieves the highest Image AUROC and Image AP among the tested values on both domains. Increasing K from 4 to 8 improves semantic coverage, whereas further increasing it to 12 introduces redundant or competing relations. Likewise, $\tau_m = 1.0$ balances dominant-axis selection at low temperature against excessive multi-axis averaging at high temperature. For cross-layer fusion, $\tau_l = 0.25$ avoids both unstable near-hard layer selection and overly uniform layer

Table 6: Source-domain optimization and checkpoint-selection protocol. The validation objective intentionally excludes target data, target statistics, and target performance metrics.

Component	Checkpoint-resolved setting
Source-only split	MVTec AD configured test split, using a fixed 90%/10% source-domain training/validation partition stratified by category and image label.
Optimization	AdamW with batch size 4, learning rate 10^{-3}, weight decay 10^{-4}, and $(\beta_1, \beta_2) = (0.9, 0.999)$. Gradients are clipped to an ℓ_2 norm of 5.0.
Training schedule	At most 15 epochs; patience of 3 validation epochs; checkpoint saved every epoch.
Segmentation objective	Fused-map Dice plus binary focal loss, together with the mean layer-wise auxiliary segmentation loss weighted by 0.25.
Image-level objective	Binary cross-entropy on the final image logit, plus the mean of global and patch binary cross-entropy terms weighted by 0.25.
Regularization weights	Total training objective weights: segmentation 1.0, image 1.0, diversity 0.05, and language anchor 0.05.
Checkpoint selection	best.pth is epoch 4, selected by the lowest source validation objective: fused-map segmentation loss plus binary cross-entropy of the final image logit. No target-domain metric is used for model selection.

Table 7: Frozen target-domain evaluation and metric-reporting protocol. All reported scores are multiplied by 100 and shown as percentages.

Component	Evaluation protocol
Inference mode	Frozen checkpoint; direct target-domain test inference; no gradient computation, parameter update, target adaptation, or target-statistic accumulation.
Image-level metrics	AUROC and average precision (AP) computed from image-level labels and final image anomaly scores.
Pixel-level metrics	Pixel AUROC and AUPRO computed from ground-truth masks and Gaussian-smoothed anomaly maps.
AUPRO implementation	200 linearly spaced thresholds; region overlap is averaged over connected ground-truth regions; only operating points with false-positive rate strictly below 0.3 are retained; the accepted false-positive-rate span is normalized to $[0, 1]$ before area computation.
Category aggregation	Each metric is computed separately for every category. Reported dataset means are arithmetic means over finite category values; undefined metrics are retained as N/A and excluded from the corresponding mean.

averaging. These results support the selected defaults as balanced operating points for semantic aggregation and cross-layer consensus fusion.

C Detailed Per-Category Results

This appendix reports detailed zero-shot results of CRAG-CLIP for every evaluated category across the target datasets. The model is trained only on MVTec AD and evaluated directly on each target dataset without target-domain adaptation. All values are percentages, and higher values indicate better performance. Image-level evaluation uses AUROC and average precision (AP); pixel-level evaluation uses AUROC and area under the per-region-overlap curve (PRO). Metrics that are not defined or reported by a benchmark are omitted rather than imputed.

Table 8: Detailed per-category zero-shot anomaly detection results on VisA. CRAG-CLIP is trained on MVTec AD and evaluated on VisA without target-domain adaptation. All values are percentages; higher is better.

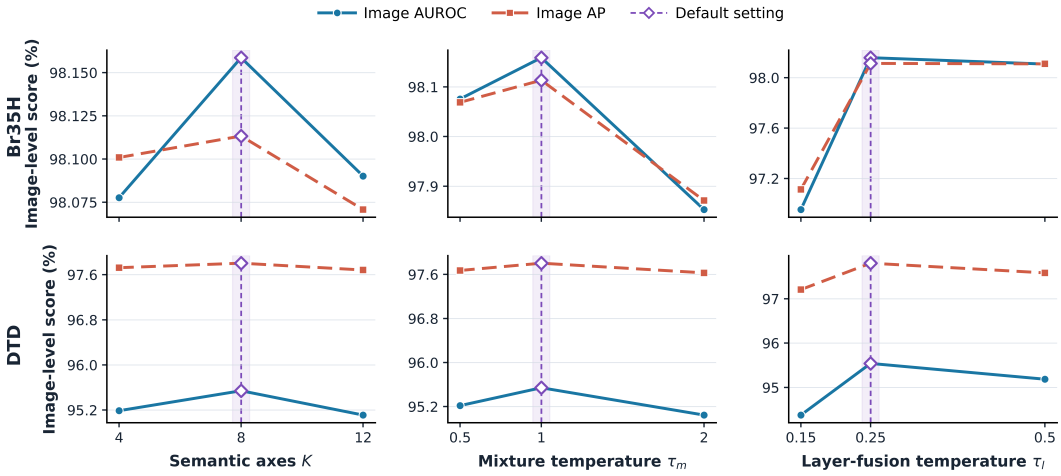


Figure 4: Sensitivity to the semantic-axis number K, axis-mixture temperature τ_m, and layer-fusion temperature τ_l on Br35H and DTD. Shaded regions and diamond markers denote the default settings: $K = 8$, $\tau_m = 1.0$, and $\tau_l = 0.25$.

Category	Image AUROC	Image AP	Pixel AUROC	Pixel PRO
candle	78.51	81.41	98.88	97.12
capsules	90.25	94.26	95.37	84.49
cashew	86.02	93.85	93.69	91.48
chewinggum	97.86	99.08	99.38	92.45
fryum	91.10	96.08	94.06	91.00
macaroni1	83.10	81.99	98.64	88.45
macaroni2	71.51	70.85	97.53	79.81
pcb1	76.82	77.36	92.70	80.82
pcb2	69.99	71.75	90.94	74.95
pcb3	66.48	72.51	88.02	75.18
pcb4	96.80	96.63	95.08	88.23
pipe_fryum	93.84	97.06	96.75	96.72
Mean	**83.52**	**86.07**	**95.09**	**86.72**

Table 9: Detailed per-category zero-shot anomaly detection results on MPDD. CRAG-CLIP is trained on MVTec AD and evaluated on MPDD without target-domain adaptation. All values are percentages; higher is better.

Category	Image AUROC	Image AP	Pixel AUROC	Pixel PRO
bracket_black	60.57	69.27	94.99	86.56
bracket_brown	63.88	79.61	94.74	84.02
bracket_white	68.11	64.89	99.81	98.81
connector	81.43	72.47	96.37	87.73
metal_plate	87.92	95.84	92.97	84.35
tubes	94.52	97.62	98.19	93.80
Mean	**76.07**	**79.95**	**96.18**	**89.21**

Table 10: Detailed per-category zero-shot anomaly detection results on BTAD. CRAG-CLIP is trained on MVTec AD and evaluated on BTAD without target-domain adaptation. All values are percentages; higher is better.

Category	Image AUROC	Image AP	Pixel AUROC	Pixel PRO
01	97.38	99.04	90.99	70.11
02	83.22	97.27	93.39	68.22
03	96.60	87.26	95.60	90.49
Mean	**92.40**	**94.52**	**93.33**	**76.27**

Table 11: Detailed per-category zero-shot anomaly detection results on DTD-Synthetic. CRAG-CLIP is trained on MVTec AD and evaluated on DTD-Synthetic without target-domain adaptation. All values are percentages; higher is better.

Category	Image AUROC	Image AP	Pixel AUROC	Pixel PRO
Blotchy_099	99.50	99.87	99.29	95.90
Fibrous_183	98.38	99.61	99.17	97.85
Marbled_078	99.44	99.87	98.64	96.42
Matted_069	91.52	97.88	99.27	84.71
Mesh_114	88.89	95.77	92.68	76.39
Perforated_037	92.94	98.26	92.72	84.09
Stratified_154	98.56	99.67	99.14	98.11
Woven_001	100.00	100.00	99.59	98.41
Woven_068	95.18	97.12	98.00	93.94
Woven_104	98.19	99.60	93.02	88.45
Woven_125	100.00	100.00	99.23	97.49
Woven_127	84.41	86.16	92.36	87.92
Mean	**95.58**	**97.82**	**96.93**	**91.64**

Table 12: Detailed zero-shot image-level anomaly detection result on HeadCT. Pixel-level metrics are not reported because this benchmark provides image-level evaluation only. All values are percentages; higher is better.

Category	Image AUROC	Image AP
brain	94.89	95.15

Table 13: Detailed zero-shot image-level anomaly detection result on BrainMRI. Pixel-level metrics are not reported because this benchmark provides image-level evaluation only. All values are percentages; higher is better.

Category	Image AUROC	Image AP
brain	96.91	97.94

Table 14: Detailed zero-shot image-level anomaly detection result on Br35H. Pixel-level metrics are not reported because this benchmark provides image-level evaluation only. All values are percentages; higher is better.

Category	Image AUROC	Image AP
brain	98.23	98.18

Table 15: Detailed zero-shot pixel-level anomaly localization result on ISIC. Image-level metrics are not reported for this benchmark. All values are percentages; higher is better.

Category	Pixel AUROC	Pixel PRO
skin	83.30	69.88

Table 16: Detailed zero-shot pixel-level anomaly localization result on ClinicDB. Image-level metrics are not reported for this benchmark. All values are percentages; higher is better.

Category	Pixel AUROC	Pixel PRO
colon	84.89	71.18

Table 17: Detailed zero-shot pixel-level anomaly localization result on ColonDB. Image-level metrics are not reported for this benchmark. All values are percentages; higher is better.

Category	Pixel AUROC	Pixel PRO
colon	81.83	72.96

Table 18: Detailed zero-shot pixel-level anomaly localization result on Kvasir. Image-level metrics are not reported for this benchmark. All values are percentages; higher is better.

Category	Pixel AUROC	Pixel PRO
colon	79.82	48.37

Table 19: Detailed zero-shot pixel-level anomaly localization result on Endo. Image-level metrics are not reported for this benchmark. All values are percentages; higher is better.

Category	Pixel AUROC	Pixel PRO
colon	83.93	55.30

Table 20: Detailed zero-shot pixel-level anomaly localization result on TN3K. Image-level metrics are not reported for this benchmark. All values are percentages; higher is better.

Category	Pixel AUROC	Pixel PRO
thyroid	74.94	47.35

D Qualitative Results by Category

We present one anomalous test image for each of the 42 target-dataset categories. All panels are generated with the MVTec AD-trained best.pth checkpoint (epoch 4), without target-domain adaptation. Each displayed instance is fixed before inference by a deterministic, prediction-independent selection protocol; the protocol does not use anomaly scores, model outputs, or ground-truth-mask geometry. For benchmarks with pixel-level annotations, the candidate pool is limited to anomalous test images with an available ground-truth mask. Each triplet comprises the input image, ground-truth annotation, and CRAG-CLIP anomaly-map overlay. Ground-truth masks are rendered with cyan filling and a red boundary. HeadCT, BrainMRI, and Br35H provide only image-level labels, which are indicated by a red frame rather than an artificial pixel mask. All overlays share one global 1–99 percentile score range and use the turbo colormap with a maximum opacity of 0.72; warmer colors therefore consistently indicate stronger anomaly evidence.

VisA / candle



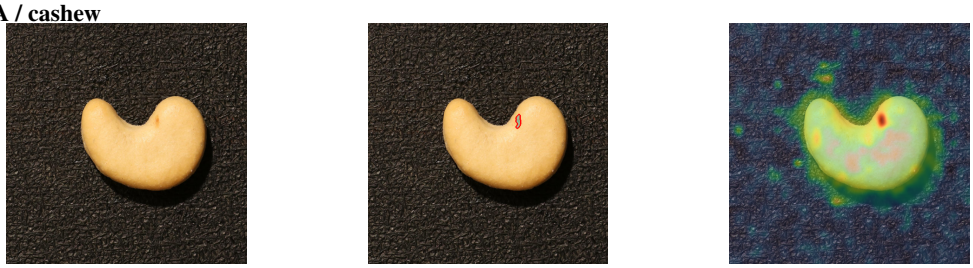
Input | Ground truth | CRAG-CLIP overlay

VisA / capsules



Input | Ground truth | CRAG-CLIP overlay

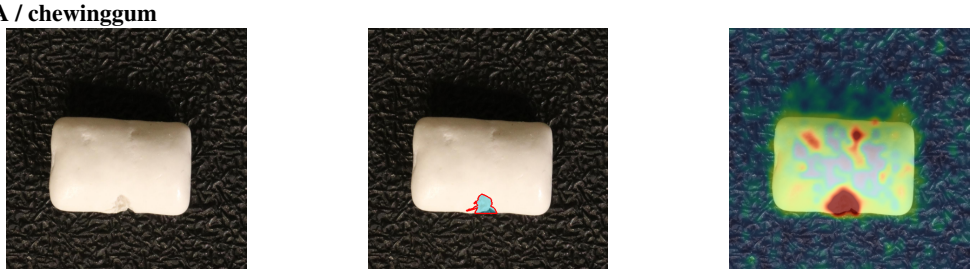
VisA / cashew



Input | Ground truth | CRAG-CLIP overlay

Figure 5: Category-wise qualitative zero-shot localization on VisA (part 1 of 4).

VisA / chewinggum



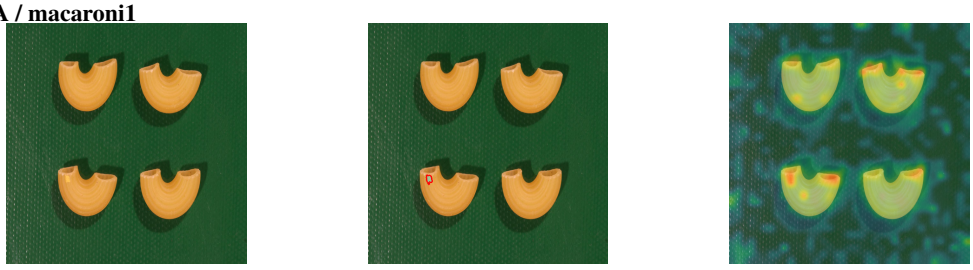
Input | Ground truth | CRAG-CLIP overlay

VisA / fryum



Input | Ground truth | CRAG-CLIP overlay

VisA / macaroni1



Input | Ground truth | CRAG-CLIP overlay

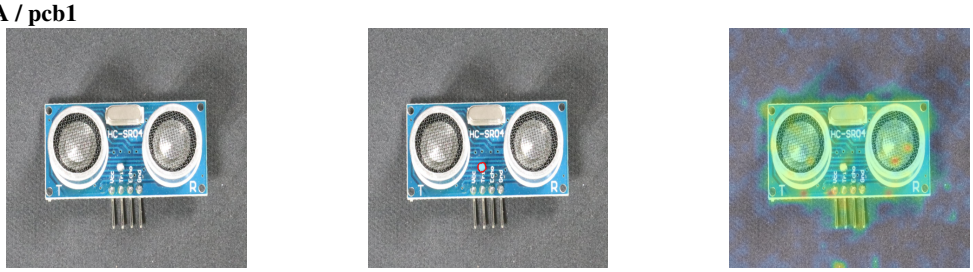
Figure 6: Category-wise qualitative zero-shot localization on VisA (part 2 of 4).

VisA / macaroni2



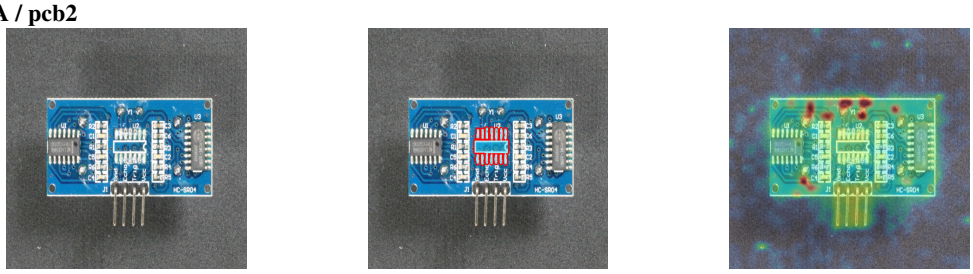
Input | Ground truth | CRAG-CLIP overlay

VisA / pcb1



Input | Ground truth | CRAG-CLIP overlay

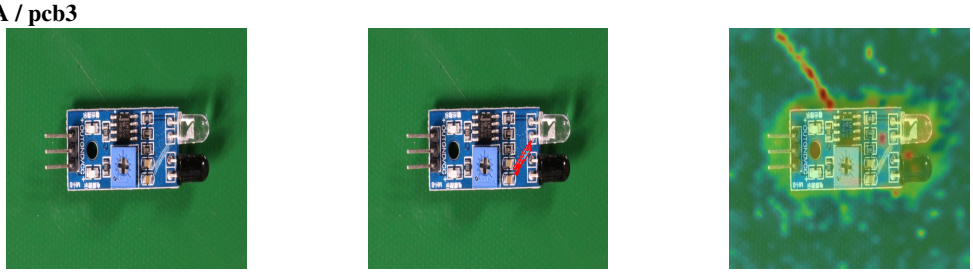
VisA / pcb2



Input | Ground truth | CRAG-CLIP overlay

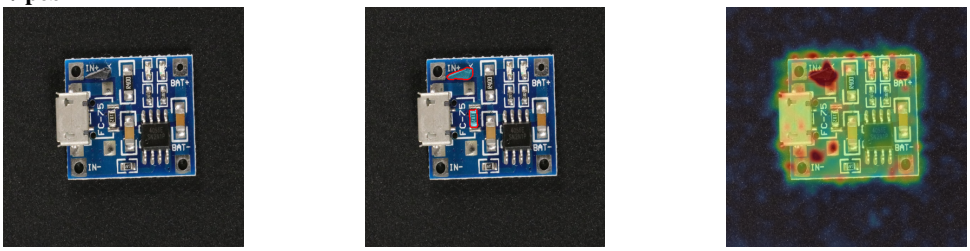
Figure 7: Category-wise qualitative zero-shot localization on VisA (part 3 of 4).

VisA / pcb3



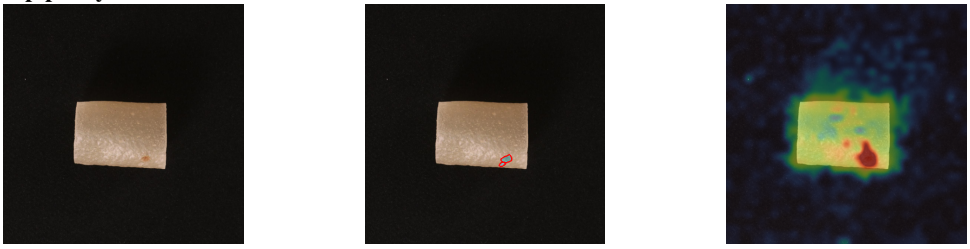
Input | Ground truth | CRAG-CLIP overlay

VisA / pcb4



Input | Ground truth | CRAG-CLIP overlay

VisA / pipe_fryum



Input | Ground truth | CRAG-CLIP overlay

Figure 8: Category-wise qualitative zero-shot localization on VisA (part 4 of 4).

MPDD / bracket_black



Input | Ground truth | CRAG-CLIP overlay

MPDD / bracket_brown



Input | Ground truth | CRAG-CLIP overlay

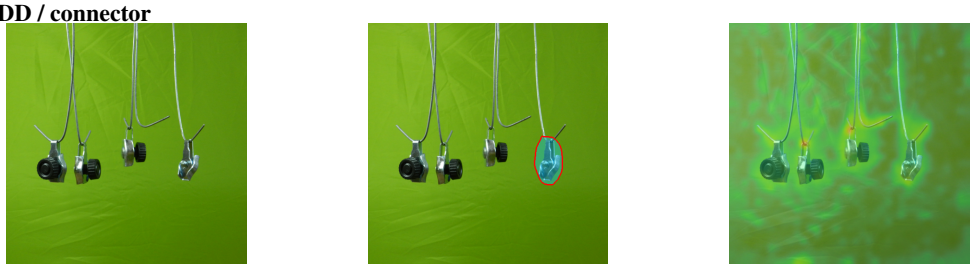
MPDD / bracket_white



Input | Ground truth | CRAG-CLIP overlay

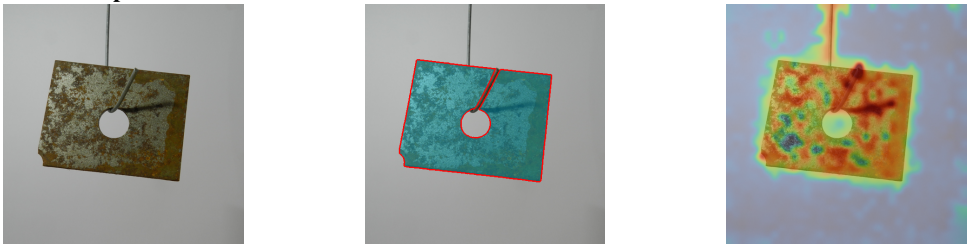
Figure 9: Category-wise qualitative zero-shot localization on MPDD (part 1 of 2).

MPDD / connector



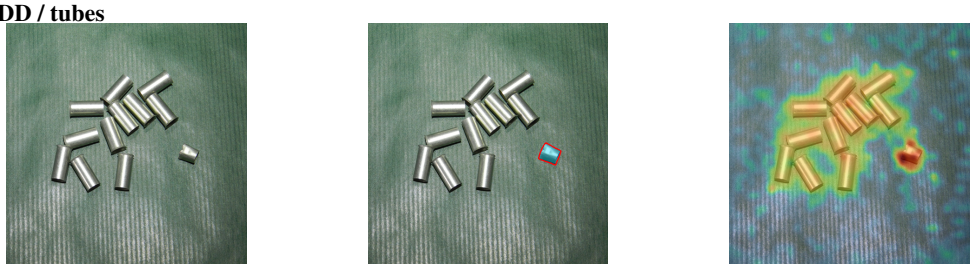
Input | Ground truth | CRAG-CLIP overlay

MPDD / metal_plate



Input | Ground truth | CRAG-CLIP overlay

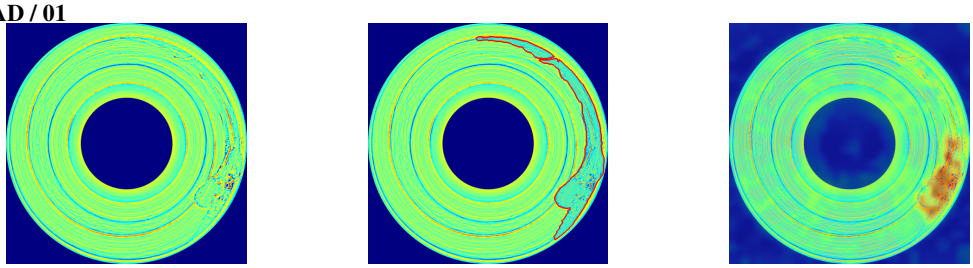
MPDD / tubes



Input | Ground truth | CRAG-CLIP overlay

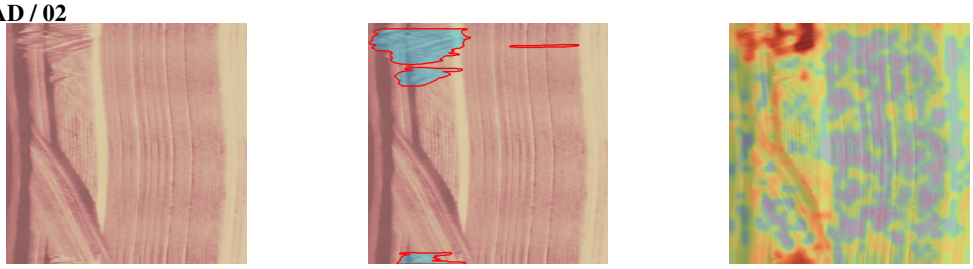
Figure 10: Category-wise qualitative zero-shot localization on MPDD (part 2 of 2).

BTAD / 01



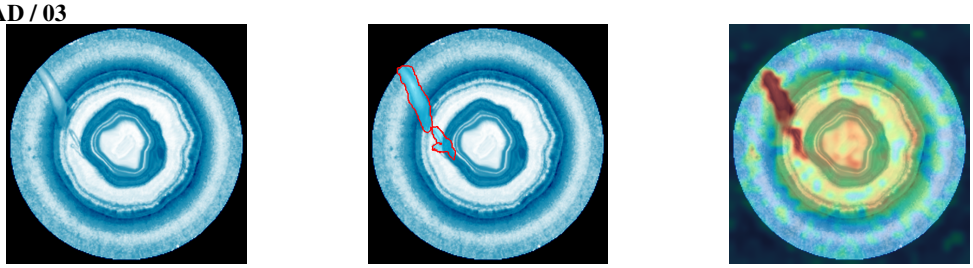
Input | Ground truth | CRAG-CLIP overlay

BTAD / 02



Input | Ground truth | CRAG-CLIP overlay

BTAD / 03



Input | Ground truth | CRAG-CLIP overlay

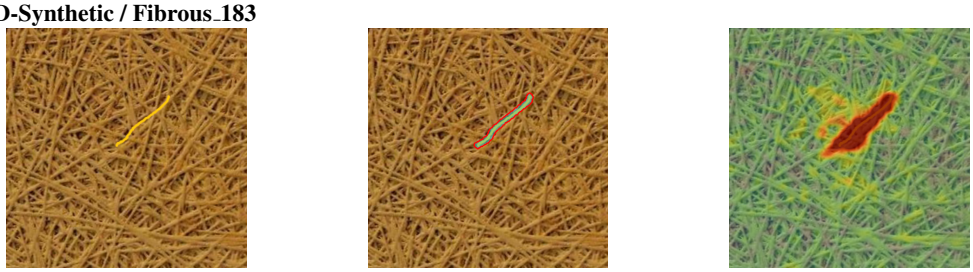
Figure 11: Category-wise qualitative zero-shot localization on BTAD.

DTD-Synthetic / Blotchy_099



Input | Ground truth | CRAG-CLIP overlay

DTD-Synthetic / Fibrous_183



Input | Ground truth | CRAG-CLIP overlay

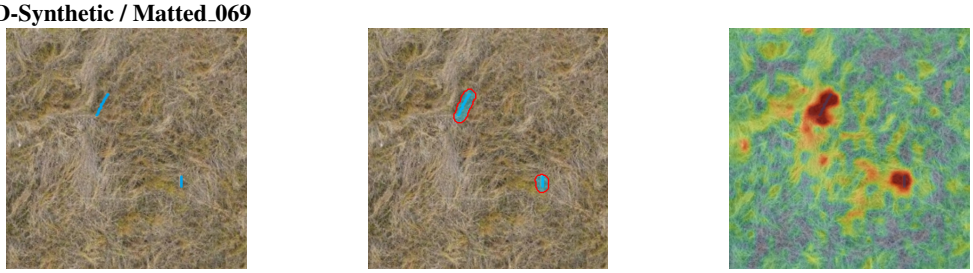
DTD-Synthetic / Marbled_078



Input | Ground truth | CRAG-CLIP overlay

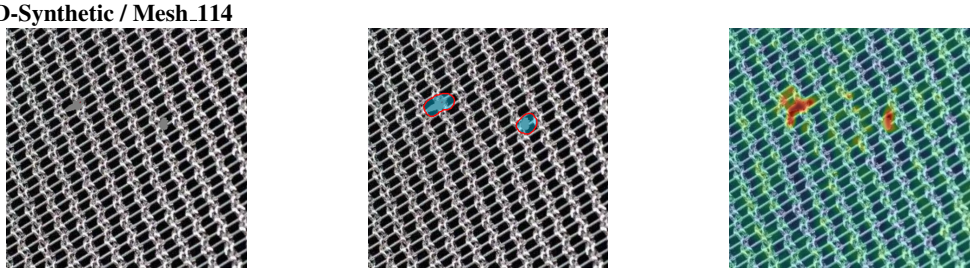
Figure 12: Category-wise qualitative zero-shot localization on DTD-Synthetic (part 1 of 4).

DTD-Synthetic / Matted_069



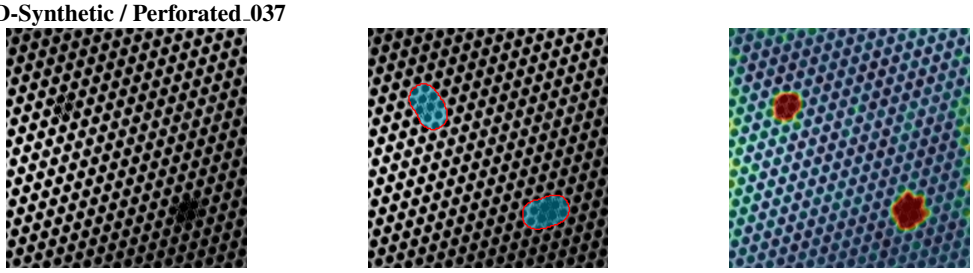
Input | Ground truth | CRAG-CLIP overlay

DTD-Synthetic / Mesh_114



Input | Ground truth | CRAG-CLIP overlay

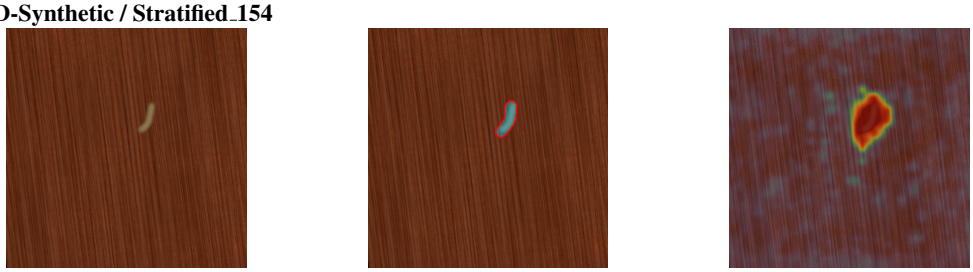
DTD-Synthetic / Perforated_037



Input | Ground truth | CRAG-CLIP overlay

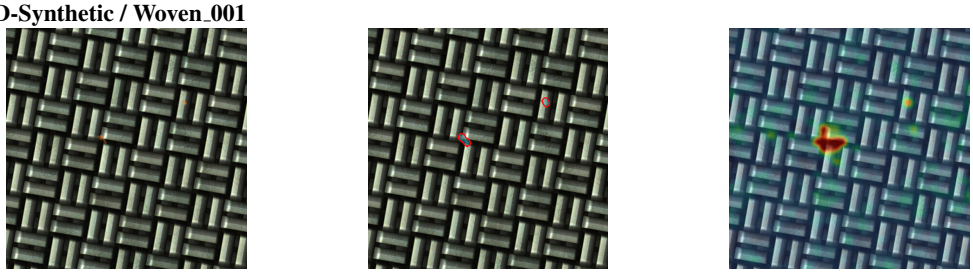
Figure 13: Category-wise qualitative zero-shot localization on DTD-Synthetic (part 2 of 4).

DTD-Synthetic / Stratified_154



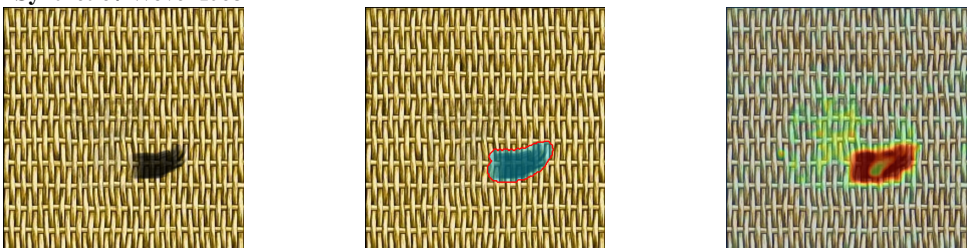
Input | Ground truth | CRAG-CLIP overlay

DTD-Synthetic / Woven_001



Input | Ground truth | CRAG-CLIP overlay

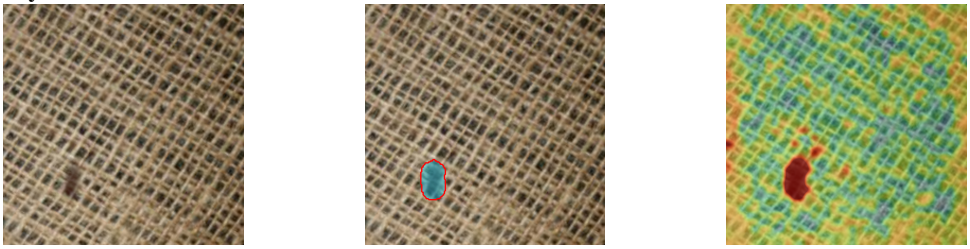
DTD-Synthetic / Woven_068



Input | Ground truth | CRAG-CLIP overlay

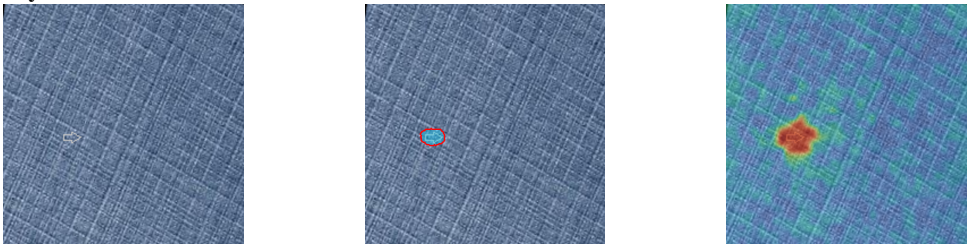
Figure 14: Category-wise qualitative zero-shot localization on DTD-Synthetic (part 3 of 4).

DTD-Synthetic / Woven_104



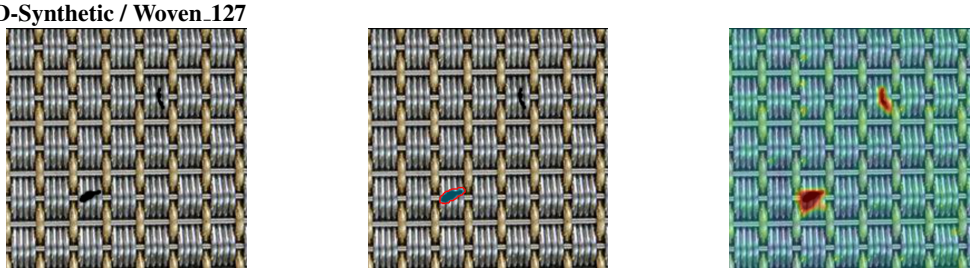
Input | Ground truth | CRAG-CLIP overlay

DTD-Synthetic / Woven_125



Input | Ground truth | CRAG-CLIP overlay

DTD-Synthetic / Woven_127



Input | Ground truth | CRAG-CLIP overlay

Figure 15: Category-wise qualitative zero-shot localization on DTD-Synthetic (part 4 of 4).

HeadCT / brain



Input | Ground truth | CRAG-CLIP overlay

BrainMRI / brain



Input | Ground truth | CRAG-CLIP overlay

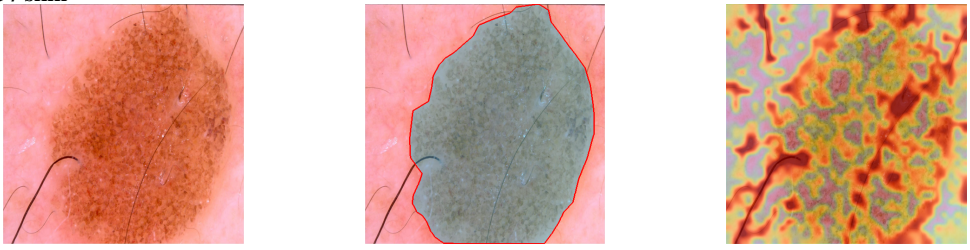
Br35H / brain



Input | Ground truth | CRAG-CLIP overlay

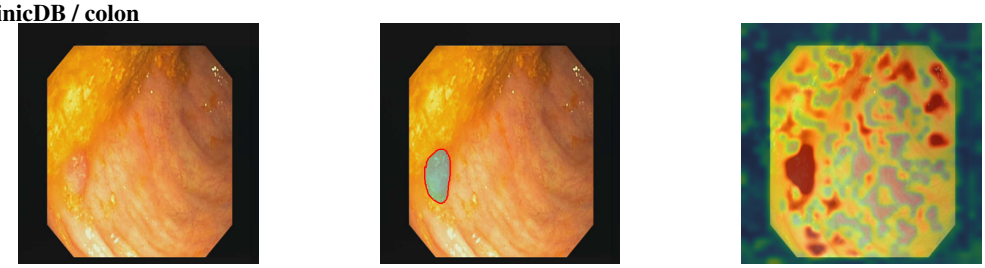
Figure 16: Category-wise qualitative zero-shot image-level anomaly detection on HeadCT, BrainMRI, and Br35H. Red frames denote anomalous image-level labels.

ISIC / skin



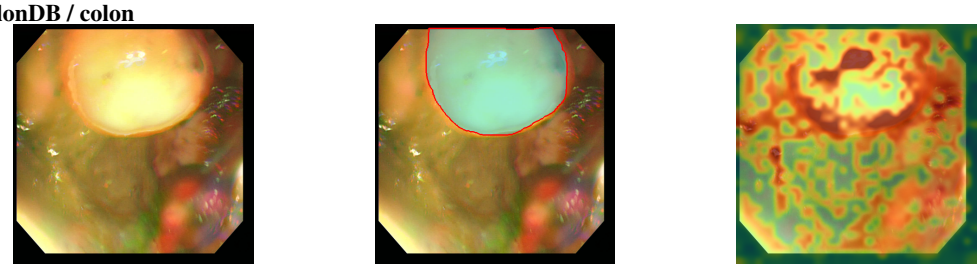
Input | Ground truth | CRAG-CLIP overlay

ClinicDB / colon



Input | Ground truth | CRAG-CLIP overlay

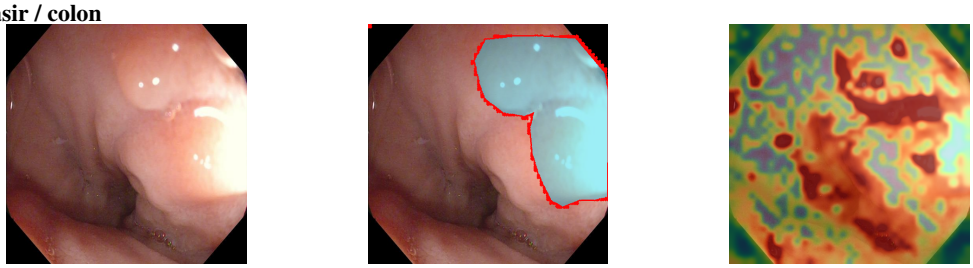
ColonDB / colon



Input | Ground truth | CRAG-CLIP overlay

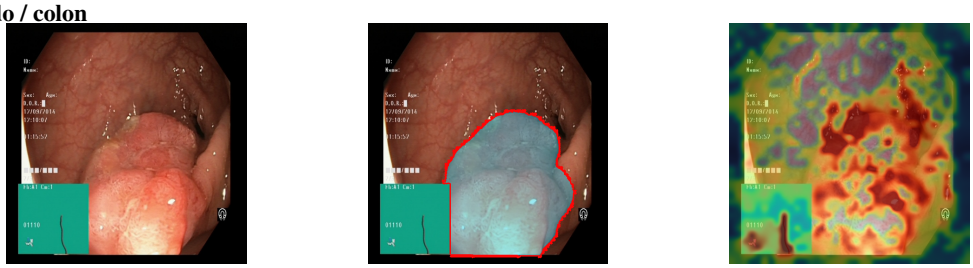
Figure 17: Category-wise qualitative zero-shot localization on ISIC, ClinicDB, and ColonDB.

Kvasir / colon



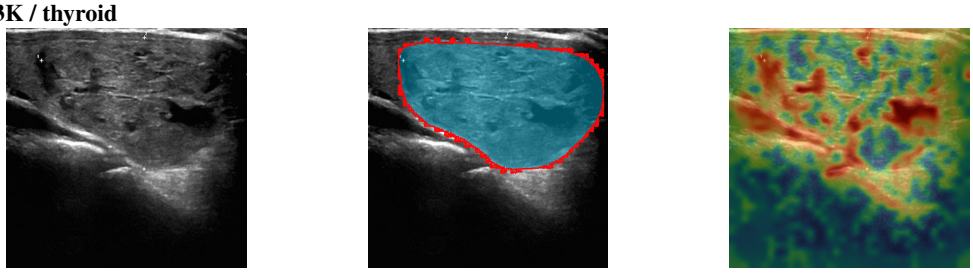
Input | Ground truth | CRAG-CLIP overlay

Endo / colon



Input | Ground truth | CRAG-CLIP overlay

TN3K / thyroid



Input | Ground truth | CRAG-CLIP overlay

Figure 18: Category-wise qualitative zero-shot localization on Kvasir, Endo, and TN3K.