ZERO-SHOT MEDICAL ANOMALY SEGMENTATION UNDER DOMAIN SHIFT: A PROTOTYPE-LEVEL ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot anomaly segmentation localizes abnormalities in unseen domains without target-domain supervision. Yet existing evaluations reveal little about how source-domain evidence manifests in unseen domains or why transfer fails. We introduce ProtoZSAD, which decomposes normal and abnormal representations into learned prototype queries that provide intermediate units for analyzing source-to-target transfer while supporting abnormal-region localization. Trained on MVTec and evaluated without target-domain adaptation on four medical datasets, ProtoZSAD achieves competitive abnormal-region localization across domains. Prototype-level interventions show that individual prototypes have distinct, domain-dependent effects on final predictions, while greater prototype response to normal anatomy relative to lesions is associated with lower Dice and increased false positives. A targeted Brain-domain intervention further reduces the implicated prototype responses and false positives while improving segmentation. Together, these results show how prototype-level analysis can identify when source-associated prototype responses remain aligned with true abnormalities and when they instead become aligned with normal target anatomy.

1 INTRODUCTION

Zero-shot anomaly detection and segmentation aim to localize abnormal regions in unseen target domains without requiring target-domain supervision. Existing methods such as Zhou et al. (2024) and Cao et al. (2024) are typically evaluated through final segmentation metrics such as AUC or Dice, together with a generic abnormality map. While these evaluations measure final performance, they provide limited means to understand what source-domain visual evidence appears in unseen target domains, whether it contributes to target predictions, or why the resulting predictions become unreliable under domain shift.

Prototype-based representations (Ayoobi et al., 2025) offer a way to expose local visual properties of an image as intermediate representations for interpreting model decisions. Instead of explaining a prediction only through the final output, prototype responses can reveal recurring local visual patterns associated with a class. This suggests a complementary role for prototypes in zero-shot anomaly segmentation: not only as prediction representations, but as intermediate units for examining cross-domain behavior. The key question is then not simply which visual patterns a model detects, but how source-associated prototype responses manifest in unseen targets, whether those responses affect the final prediction, and when they become aligned with normal target anatomy rather than true abnormalities.

In this work, we introduce ProtoZSAD, a prototype-based framework that decomposes normal and abnormal representations into multiple learned prototype queries. The learned queries are shared across images, while their patch assignments and response maps are image-dependent. ProtoZSAD preserves these prototype-specific responses between local image patches and the final segmentation output, enabling us to examine source-associated evidence, its manifestation in unseen targets, and its contribution to predictions. A dual-condition abnormal gate (DCAG) aggregates abnormal prototype responses for localization by retaining strong abnormal evidence while suppressing responses that do not exceed competing normal evidence.

We train ProtoZSAD on the industrial-domain MVTec dataset (Bergmann et al., 2019) and evaluate it without target-domain adaptation on four medical datasets: Retina-RESC, Breast, Brain, and

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.

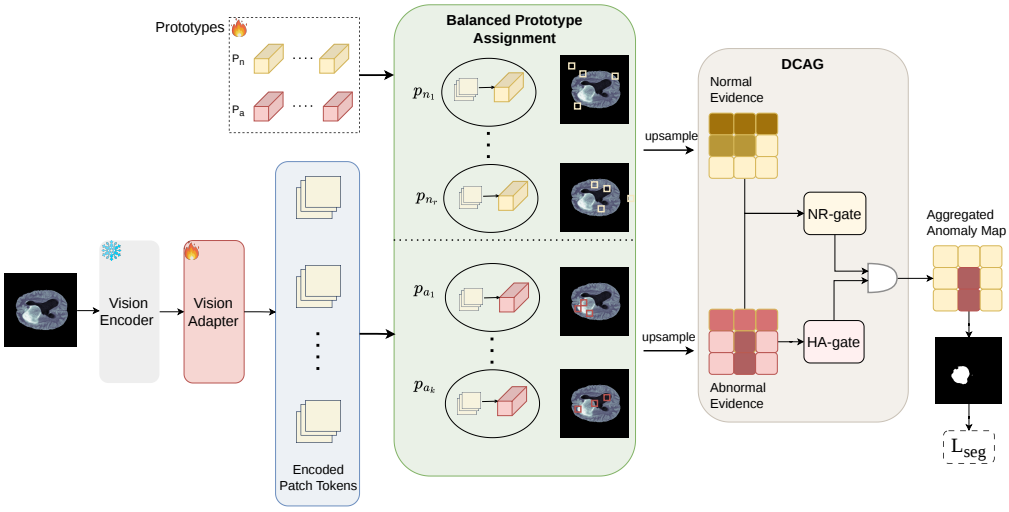


Figure 1: Overview of the ProtoZSAD architecture

Lung. ProtoZSAD achieves competitive Dice-1 across the four target domains, providing a viable localization model for the subsequent prototype-level transfer analysis. We then examine the resulting prototype behavior at three levels: how source-associated prototype responses manifest across target domains, whether prototype responses contribute to final predictions, and which prototype-level behaviors are associated with transfer failure. Prototype-removal interventions reveal distinct, domain-dependent effects on final predictions, while greater prototype response to normal anatomy relative to lesions is generally associated with lower Dice and increased false positives. In a focused Brain failure case, modifying the implicated ventricular structures reduces the corresponding prototype responses and false positives while improving segmentation. Together, these results show how prototype-level analysis can identify when source-associated evidence remains aligned with true abnormalities and when it is redirected toward normal target anatomy.

2 Related Work

Existing zero-shot anomaly detection methods commonly represent abnormality through a generic normal–abnormal semantic space. AnomalyCLIP (Zhou et al., 2024), AdaCLIP (Cao et al., 2024), FECLIP (Gong et al., 2025), AA-CLIP (Ma et al., 2025), and WinCLIP (Jeong et al., 2023) compute anomaly scores by comparing local visual features with normal and abnormal language representations used as semantic anchors. More recent approaches reduce or remove this dependence on language cues. VisualAD (Hou et al., 2026) replaces language prompts with learnable visual tokens representing normal and abnormal patterns, while MRAD-CLIP (Xu et al., 2026) uses an abnormal memory bank and similarity-based retrieval to compute anomaly scores. Although these approaches differ architecturally, they are designed primarily for anomaly discrimination and localization, rather than for analyzing how source-associated visual evidence manifests in unseen target domains, whether such evidence contributes to final predictions, and when it becomes unreliable under domain shift.

Interpretability is particularly important in medical imaging, where understanding the visual evidence underlying model predictions is important for assessing model reliability (Chaddad et al., 2023; Zhang et al., 2021; Kazmierczak et al., 2025). Prototype-based models provide a natural intermediate representation for this purpose by relating local visual features to representative patterns. Prior work introduces class-specific prototypes to support interpretable prediction (Chen et al., 2019; Nauta et al., 2023; Wang et al., 2023; Ma et al., 2023) and extends prototype-based reason-

ing to Vision Transformer architectures (Ma et al., 2024; Gurung et al., 2026; Xue et al., 2022). These works motivate the use of prototypes as structured intermediate representations, but do not center on tracing source-to-target behavior under zero-shot domain shift. Recently, Malenia (Jiang et al., 2025) represents disease-agnostic medical attributes through structured textual descriptions, while ReCAP (Wan et al., 2026) uses prototypes as latent conditional anchors formed from aggregated patch tokens for normal/abnormal segmentation. Our focus is complementary: rather than treating prototype use itself as the contribution, we use shared learned prototype queries as analysis units to examine how source-associated responses manifest in unseen target domains, whether those responses contribute to final predictions, and when they become aligned with normal anatomy and contribute to transfer failure. Sinkhorn-based patch-to-prototype assignment (Cuturi, 2013) provides the intermediate representation needed for this analysis.

3 PROTOZSAD ARCHITECTURE

Overview and notation. Given an RGB image batch $\mathbf{I} \in \mathbb{R}^{B\times 3\times H\times W}$, the segmentation model produces a dense abnormality probability $\mathbf{P}^a \in [0,1]^{B\times H\times W}$. Its active path consists of a frozen ViT-L/14 visual encoder (Radford et al., 2021), multi-level segmentation adapters, K learned prototype queries, balanced patch-to-prototype assignment, and a normal-gated prototype localization rule. Let $H \times W$ denote the image resolution and $G \times G$ the patch-grid resolution, yielding $P = G^2$ patch tokens. In the reference configuration, the visual-transformer width is 1024, and the adapter and prototype width is $D = 768$. We use batch index $b \in \{0,\dots,B-1\}$, patch indices $i, j \in \{0,\dots,P-1\}$, prototype indices $k, r \in \{0,\dots,K-1\}$, and pixel coordinate $u \in \{0,\dots,H-1\}\times\{0,\dots,W-1\}$. The normal and abnormal prototype counts are denoted by K_n and K_a, with $K = K_n + K_a$. Accordingly, $\mathbb{K}_n = \{0,\dots,K_n-1\}$ and $\mathbb{K}_a = \{K_n,\dots,K-1\}$. The default configuration uses $K_n = K_a = 4$, hence $K = 8$, with normal indices $\{0,1,2,3\}$ and abnormal indices $\{4,5,6,7\}$. Experimental labels $p_{04},\dots,p_{07}$ refer to $\boldsymbol{\rho}_4,\dots,\boldsymbol{\rho}_7$, respectively. Images are converted to tensors and resized to $H \times W$ and Binary masks $\mathbf{M} \in \{0,1\}^{B\times 1\times H\times W}$. These masks are used only as training targets. We next describe the model in computational order: multi-level visual encoding, balanced prototype assignment, dual-condition abnormal gate, and optimization losses.

Frozen visual encoder and multi-level segmentation representation. The image is processed by a frozen ViT-L/14 visual backbone augmented with trainable visual adapters. We retain the segmentation-adapter outputs from blocks $\mathbb{S} = \{6,12,18,24\}$ following MVFA (Huang et al., 2024), $\mathbf{H}_{\text{seg}}^{(\ell)} \in \mathbb{R}^{(P+1)\times B\times D}$, where each sequence contains one class token at sequence index 0 followed by P encoded patch tokens at sequence indices $1,\dots,P$. Thus, patch index i corresponds to sequence index $i+1$. The internal patch embedding, transformer blocks, and adapter reinjection are subsumed into this encoder–adapter mapping. We discard the class token, average the encoded patch tokens across the four levels, and normalize every resulting feature:

$$\boldsymbol{z}_{bi} = \operatorname{norm}_2\left(\frac{1}{4}\sum_{\ell\in\mathbb{S}}(\mathbf{H}_{\text{seg}}^{(\ell)})_{i+1,b,:}\right), \qquad \mathbf{Z} = [\boldsymbol{z}_{bi}] \in \mathbb{R}^{B\times P\times D}. \tag{1}$$

Learned prototypes and balanced patch assignment. The model maintains K trainable global prototype queries,

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{\rho}_0^\top \\ \vdots \\ \boldsymbol{\rho}_{K-1}^\top \end{bmatrix} \in \mathbb{R}^{K\times D}. \tag{2}$$

Membership in the normal or abnormal group is determined only by these index sets. A shared bias-free projection $\boldsymbol{W}_A \in \mathbb{R}^{D\times D}$ is applied to both patch features and queries before normalization:

$$\widehat{\boldsymbol{z}}_{bi} = \frac{\boldsymbol{W}_A\boldsymbol{z}_{bi}}{\|\boldsymbol{W}_A\boldsymbol{z}_{bi}\|_2}, \qquad \widehat{\boldsymbol{\rho}}_k = \frac{\boldsymbol{W}_A\boldsymbol{\rho}_k}{\|\boldsymbol{W}_A\boldsymbol{\rho}_k\|_2}. \tag{3}$$

Their cosine score is

$$\mathcal{S}_{bik} = \widehat{\boldsymbol{z}}_{bi}^\top\widehat{\boldsymbol{\rho}}_k, \qquad \mathbf{S} \in \mathbb{R}^{B\times P\times K}. \tag{4}$$

The implementation converts these scores into soft assignments using exactly three finite Sinkhorn normalization iterations. It first forms

$$R^{(0)}_{bki} = \frac{\exp(S_{bik}/\epsilon - m_b)}{\sum_{r=0}^{K-1}\sum_{j=0}^{P-1}\exp(S_{bjr}/\epsilon - m_b)}, \qquad m_b = \max_{j,r}\frac{S_{bjr}}{\epsilon}, \qquad \epsilon = 0.05, \tag{5}$$

where $\mathbf{R}^{(0)} \in \mathbb{R}^{B\times K\times P}$. For $\nu \in \{0,1,2\}$, prototype-row and patch-column normalizations are applied successively:

$$\bar{R}^{(\nu+1)}_{bki} = \frac{R^{(\nu)}_{bki}}{K\,\max\left(\sum_{j=0}^{P-1} R^{(\nu)}_{bkj}, 10^{-12}\right)}, \qquad R^{(\nu+1)}_{bki} = \frac{\bar{R}^{(\nu+1)}_{bki}}{P\,\max\left(\sum_{r=0}^{K-1} \bar{R}^{(\nu+1)}_{bri}, 10^{-12}\right)}. \tag{6}$$

The returned patch-to-prototype assignment is

$$A_{bik} = P\,R^{(3)}_{bki}, \qquad \mathbf{A} \in \mathbb{R}^{B\times P\times K}. \tag{7}$$

After three finite iterations, $\sum_{k=0}^{K-1} A_{bik} \approx 1$ and $\sum_{i=0}^{P-1} A_{bik} \approx P/K$. All K queries participate jointly in this normalization.

Dual-Condition Abnormal Gate (DCAG) The abnormal assignment channels $k \in \mathbb{K}_a$ define K_a candidate maps, while the normal channels $r \in \mathbb{K}_n$ provide a spatial competitor. For each abnormal channel, define

$$a^{\min}_{bk} = \min_i A_{bik}, \qquad \Delta_{bk} = \max_i A_{bik} - a^{\min}_{bk}. \tag{8}$$

The spatial min–max normalization preserves the explicit constant-channel case:

$$\widetilde{A}_{bik} = \begin{cases} \dfrac{A_{bik} - a^{\min}_{bk}}{\max(\Delta_{bk}, 10^{-8})}, & \Delta_{bk} > 10^{-8}, \\ 0, & \Delta_{bk} \le 10^{-8}, \end{cases} \qquad k \in \mathbb{K}_a. \tag{9}$$

Each normalized channel is reshaped to the $G \times G$ patch grid and bilinearly interpolated to image resolution:

$$U_{bk} = f^{\text{int}}_{G\times G\to H\times W}\left(\text{Grid}_{G\times G}(\widetilde{\mathbf{A}}_{b,:,k})\right) \in \mathbb{R}^{H\times W}. \tag{10}$$

The same interpolation is applied to each raw abnormal assignment, while the normal competitor is obtained by taking the patchwise maximum over normal prototype channels before interpolation:

$$U^{\text{raw}}_{bk} = f^{\text{int}}_{G\times G\to H\times W}\left(\text{Grid}_{G\times G}(\mathbf{A}_{b,:,k})\right), \qquad k \in \mathbb{K}_a, \tag{11}$$

$$U^{n}_{b} = f^{\text{int}}_{G\times G\to H\times W}\left(\text{Grid}_{G\times G}\left(\left[\max_{r\in\mathbb{K}_n} A_{bir}\right]_{i=0}^{P-1}\right)\right). \tag{12}$$

For each abnormal prototype, the **high-activation (HA) gate** favors locations with high spatially normalized activation, while the **normal-relative (NR) gate** suppresses locations where raw abnormal assignment does not exceed the normal competitor. Their outputs are combined by an elementwise minimum, ensuring that the final prototype contribution is high only when both conditions are satisfied.

$$G_{bk}(u) = \min\left\{\sigma\left(\frac{U_{bk}(u) - \theta_h}{\tau_h}\right), \sigma\left(\frac{U^{\text{raw}}_{bk}(u) - U^{n}_{b}(u)}{\tau_h}\right)\right\}, \qquad \theta_h = 0.7, \qquad \tau_h = 0.05. \tag{13}$$

The abnormal probability is the hard maximum across the K_a normal-gated prototype contributions, followed by numerical clipping:

$$p^{a}_{b}(u) = \text{clip}_{[10^{-6},\,1-10^{-6}]}\left(\max_{k\in\mathbb{K}_a} G_{bk}(u)\right). \tag{14}$$

The returned two-class probability representation is

$$\mathbf{P}^{\text{seg}}_{b}(u) = \begin{bmatrix} 1 - p^{a}_{b}(u) \\ p^{a}_{b}(u) \end{bmatrix}, \qquad \mathbf{P}^{\text{seg}} \in [0,1]^{B\times 2\times H\times W}. \tag{15}$$

This continuous probability, rather than a thresholded mask, is used by the pixel losses. The finite Sinkhorn normalization, interpolation, sigmoid gates, elementwise minimum, and abnormal-prototype maximum remain in the training graph. Consequently, normal queries receive a direct segmentation gradient path through U^{n}_{b} in addition to their joint Sinkhorn coupling with abnormal queries.

Loss function. Pixel supervision combines the focal loss (Lin et al., 2017) on the two-class probability map and the Dice loss (Milletari et al., 2016) on its abnormal channel:

$$\mathcal{L}_{\text{focal}} = \text{Focal}(\mathbf{P}^{\text{seg}}, \mathbf{M}), \qquad \mathcal{L}_{\text{dice}} = \text{Dice}(\mathbf{P}^{a}, \mathbf{M}). \tag{16}$$

The continuous probabilities are supplied directly to both losses without an additional $\mathrm{softmax}$ or binary threshold.

For prototype compactness, the ground-truth mask is resized by nearest-neighbor interpolation to the patch grid:

$$m_{bi} = \text{Flatten}\big(\text{Resize}^{\text{nearest}}_{G\times G}(\mathbf{M}_b)\big)_i \in \{0, 1\}. \tag{17}$$

The discrete same-class target query is

$$t_{bi} = \begin{cases} \arg\max\limits_{k\in\mathbb{K}_n} \boldsymbol{A}_{bik}, & m_{bi} = 0, \\ \arg\max\limits_{k\in\mathbb{K}_a} \boldsymbol{A}_{bik}, & m_{bi} = 1. \end{cases} \tag{18}$$

This target-selection $\arg\max$ is non-differentiable. Let $\mathbb{V}$ be the set of patches whose resized label is exactly zero or one. The compactness loss (Zhou et al., 2022) is

$$\mathcal{L}_{\text{PPC}} = \frac{1}{|\mathbb{V}|} \sum_{(b,i)\in\mathbb{V}} \|\boldsymbol{z}_{bi} - \boldsymbol{\rho}_{t_{bi}}\|_2^2 . \tag{19}$$

Here $\boldsymbol{z}_{bi}$ is normalized, whereas the selected learned query is not normalized in this loss.

For prototype diversity, a trainable projection $\boldsymbol{W}_P : \mathbb{R}^{768} \to \mathbb{R}^{256}$ produces

$$\bar{\boldsymbol{\rho}}_k = \frac{\boldsymbol{W}_P \boldsymbol{\rho}_k}{\|\boldsymbol{W}_P \boldsymbol{\rho}_k\|_2}, \qquad \mathcal{L}_{\text{PPD}} = \sum_{\substack{0\le k,r\le K-1 \\ k\neq r}} \big(\bar{\boldsymbol{\rho}}_k^{\top} \bar{\boldsymbol{\rho}}_r\big)^2 . \tag{20}$$

The sum includes all ordered off-diagonal pairs.

The overall segmentation loss is

$$\boxed{\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{dice}} + 0.1\,\mathcal{L}_{\text{PPC}} + 0.1\,\mathcal{L}_{\text{PPD}}}. \tag{21}$$

4 Experiments

We train our model on the general industrial anomaly dataset MVTec AD (Bergmann et al., 2019) and evaluate zero-shot transfer on four medical anomaly segmentation datasets: Retina-RESC, Breast, Brain, and Lung. Detailed dataset preparation, training settings, and evaluation protocols are provided in the Appendix.

4.1 RQ1: Does prototype decomposition preserve competitive zero-shot abnormal-region localization?

Table 1(a) compares our multi-prototype formulation with baselines that rely on generic normal/abnormal representations, where image patches are matched against class-level representations instantiated by text prompts, aggregated visual features, or visual memory features. Our method achieves the best Dice-1 on Retina-RESC, Breast, and Brain, although the margins over the second-best methods are modest. Lung remains the most challenging target domain for our model and for most of the compared methods, and we analyze these domain-dependent differences in a later section. These results suggest that introducing prototype decomposition does not generally compromise zero-shot localization performance and can provide modest gains over generic representations. Table 1(b) further isolates the effect of prototype decomposition by keeping the model structure unchanged and comparing the default 4-normal/4-abnormal prototype setting with a single normal and single abnormal prototype. The multi-prototype setting improves Dice-1 on all four target datasets, although the magnitude of improvement is domain-dependent. Together with the prototype-count analysis in Appendix A.7, these results show that prototype decomposition can support competitive abnormal-region localization while providing intermediate units for subsequent transfer analysis. Results for ProtoZSAD are averaged over five seeds.

(a) Zero-shot segmentation performance

Method	Dice-1 (%) ↑				
	Retina	Breast	Brain	Lung	Avg.
AnomalyCLIP	0.74	20.71	12.31	7.58	10.33
AdaCLIP	8.45	34.63	17.13	8.77	17.25
AA-CLIP	<u>57.77</u>	26.20	24.51	**18.59**	31.77
MRAD-CLIP	52.20	36.60	28.90	11.60	<u>32.33</u>
VisualAD	29.72	<u>45.33</u>	<u>33.95</u>	15.19	31.05
Ours	**60.01**	**47.81**	**34.91**	<u>18.24</u>	**40.24**

(b) Prototype multiplicity

Dataset	Setting	Dice-1 ↑
Retina-RESC	Multiple	60.01
	Single	47.46
Breast	Multiple	47.81
	Single	40.86
Brain	Multiple	34.91
	Single	27.77
Lung	Multiple	18.24
	Single	17.51

Table 1: (a) Zero-shot segmentation performance on unseen medical datasets. Dice-1 (%) denotes the mean Dice score over abnormal images. Best results are **bold** and second-best results are <u>underlined</u>. (b) Effect of prototype multiplicity, comparing the multi-prototype and single-prototype settings.

4.2 RQ2: What visual evidence do individual prototypes capture, and how does it affect target predictions?

4.2.1 How do source-associated prototype responses manifest in target domains?

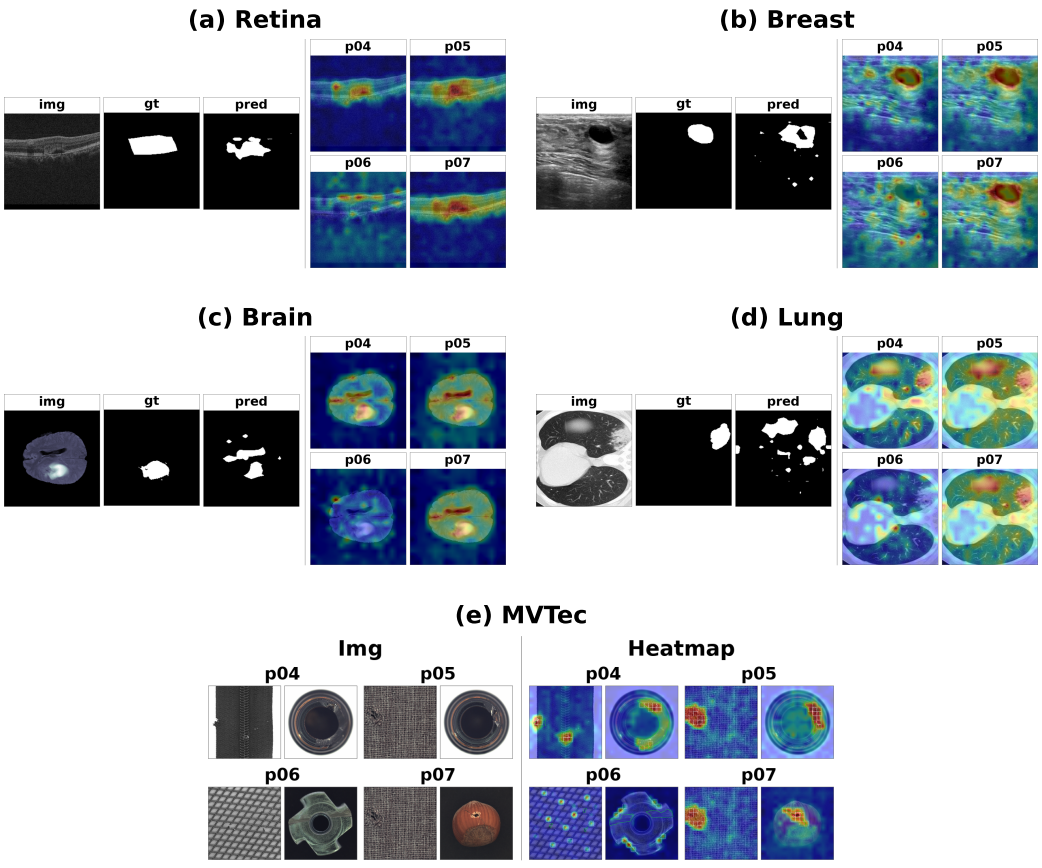


Figure 2: Spatial activation patterns of abnormal prototypes across datasets. The visualization illustrates how individual prototypes respond to different local visual patterns and how their activation behavior changes under zero-shot transfer. gt is ground-truth mask while pred is model's prediction

(a) Prototype behavior across domains

Prototype	Source Pattern	Successful Transfer	Failure Mode
p_{05}/p_{07}^{*}	Compact high contrast defects with positive cosine similarity, indicating direct prototype-patch matching.	Breast and Retina lesions.	Brain ventricles and Lung vessel/airway structures.
p_{04}	Boundary and intensity-transition structures highlighted by prototype assignments, despite negative mean cosine similarity.	Some lesion boundaries.	Normal anatomical boundaries.
p_{06}	Fine-scale local variations highlighted by prototype assignments, despite negative mean cosine similarity.	Limited transfer	Predominantly normal non-lesion anatomy.

(b) Mean prototype evidence

Proto.	Cosine	GT Ov.
p_{04}	-0.0423	0.4025
p_{05}	$+0.3381$	0.4383
p_{06}	-0.1800	0.1868
p_{07}	$+0.3324$	0.4030

Table 2: Summary of abnormal prototype behavior and evidence across source and target domains. * p_{05} and p_{07} exhibit substantial spatial redundancy but are not identical.

Figure 2 shows a case study providing representative examples of how abnormal prototypes behave across the source and target domains. The red regions indicate the spatial areas on which each prototype focuses most strongly. Table 2b summarizes the domain-averaged evidence statistics for each prototype across the four medical target domains. Cosine measures the average pre-assignment prototype-patch similarity, while GT Ov. quantifies the average spatial alignment between prototype responses and the annotated abnormal regions. Prototype p_{04} can be interpreted as a boundary- or transition-sensitive unit. Although its mean cosine similarity remains negative, this does not preclude high patch-to-prototype assignment weights, since Sinkhorn assignment depends on relative similarities across patches and prototypes under balancing constraints rather than on whether individual cosine scores are positive. The visualizations suggest that regions emphasized by p_{04} consistently exhibit boundary or intensity-transition structures. This behavior aligns with true abnormalities when lesions appear against relatively homogeneous backgrounds, as in MVTec, Breast, and Retina, but is redirected toward normal anatomical boundaries in Brain and Lung. In contrast, p_{05} and p_{07} behave as local high-contrast blob/edge detectors, capturing localized dark or bright regions with sharp boundaries and showing positive cosine similarity, which is consistent with more direct prototype–patch matching. They transfer well to Breast and Retina, but become less specific in Brain and Lung, where normal anatomy contains repeated high-contrast cavities, vessels, and airway-related structures that resemble the source-domain defect patterns. Prototype p_{06} also exhibits negative mean cosine similarity, but its assigned regions show recurring fine-scale local variations rather than a clear lesion-specific pattern. These responses are generally weakly aligned with annotated abnormalities, with Retina being the main exception where this behavior overlaps more substantially with lesion regions. A summary of these behaviors is provided in Table 2a. Overall, a shared prototype query provides a common analysis index across source and target images, while its assigned patches and response maps remain image-dependent. Recurring visual characteristics associated with a query can therefore be compared across domains without assuming that the query represents an identical semantic concept in every image. Successful transfer is characterized by prototype responses that remain aligned with true abnormalities, whereas failures arise when similar responses instead align with normal target anatomy.

(a) Prototype–GT IoU

Proto.	MVTec	Retina	Breast	Brain	Lung
p_4	33.11	13.94	20.58	15.13	11.82
p_5	39.92	33.21	26.42	18.14	10.44
p_6	0.55	7.64	2.91	1.58	0.41
p_7	39.48	32.09	23.59	18.48	10.44

(b) Pairwise Prototype IoU

Pair	MVTec	Retina	Breast	Brain	Lung
p_4–p_5	60.89	33.12	31.10	30.91	30.52
p_4–p_6	0.02	2.39	1.54	0.46	0.07
p_4–p_7	62.40	29.14	29.27	25.78	31.81
p_5–p_6	0.22	4.55	2.63	1.12	1.11
p_5–p_7	80.79	61.88	60.26	59.71	61.56
p_6–p_7	0.32	4.29	2.65	1.36	1.01

(c) Prediction relevance of individual abnormal prototypes. Affected $> 1\%$ denotes the percentage of abnormal images for which prototype removal changes more than 1% of spatial predictions. Metric changes are defined as $\Delta m = m_{\text{remove}} - m_{\text{original}}$.

Dataset	Removed	Affected $> 1\%$	ΔDice-1	ΔFPO
MVTec	p_{04}	1.11	+0.01	-0.06
	p_{05}	5.48	-0.21	-0.15
	p_{06}	1.11	+4.08	-0.25
	p_{07}	11.37	-0.78	-0.22
Retina	p_{04}	4.29	-0.26	-0.07
	p_{05}	22.86	-2.17	-0.32
	p_{06}	7.14	-0.28	-0.22
	p_{07}	67.14	-5.89	-0.97
Breast	p_{04}	1.54	-0.16	-0.07
	p_{05}	46.15	+0.87	-1.11
	p_{06}	7.69	+2.05	-0.47
	p_{07}	49.23	+0.77	-1.30
Brain	p_{04}	0.00	-0.01	-0.04
	p_{05}	34.09	-0.07	-1.03
	p_{06}	4.55	+0.77	-0.36
	p_{07}	54.55	+2.22	-1.39
Lung	p_{04}	0.00	-0.05	-0.08
	p_{05}	53.52	+0.55	-1.92
	p_{06}	7.04	+0.33	-0.42
	p_{07}	46.48	-0.55	-1.81

Table 3: Prototype-level complementarity and prediction relevance across source and target domains. Prototype–GT IoU measures lesion alignment, pairwise prototype IoU measures spatial redundancy, and prototype removal evaluates the contribution of each prototype to the final prediction.

4.2.2 Do prototypes provide distinct or redundant evidence?

Tables 3a and 3b show that the abnormal prototypes exhibit a mixture of spatial redundancy and complementarity. In particular, p_{05} and p_{07} consistently show substantial pairwise overlap, while p_{04} also partially overlaps with these prototypes. In contrast, p_{06} remains largely spatially distinct from the others across all domains. However, this spatial distinctiveness does not necessarily correspond to useful abnormal evidence: p_{04}, p_{05}, and p_{07} maintain substantially higher overlap with the ground-truth abnormal regions, whereas p_{06} shows consistently weak lesion alignment. Overall, prototype decomposition yields a mixture of shared and distinct representations, but spatial complementarity alone does not guarantee lesion-relevant evidence.

4.2.3 Are prototype responses prediction-relevant?

Section 4.2.1 establishes how source-domain visual patterns are represented by individual prototypes and where these prototype-specific patterns reappear in the target domain. We next examine whether these target-domain prototype responses contribute to the final segmentation prediction, completing the chain from source visual evidence, through prototype representation and target activation, to model output. For this analysis, we remove each abnormal prototype individually from the final union at inference time while keeping all upstream assignments fixed, and measure how often the segmentation prediction changes together with the resulting changes in Dice-1 and abnormal false-positive-outside rate.

Table 3c shows that removing p_{05} or p_{07} generally changes predictions in a larger fraction of images than removing p_{04} or p_{06} under the 1% change criterion. However, spatial involvement does not imply uniformly positive utility: the effect of prototype removal on Dice varies across both prototypes and domains. On Retina-RESC, removing any abnormal prototype degrades segmentation, whereas in other domains certain prototypes can improve Dice when removed; notably, p_{06} shows beneficial removal effects across several datasets, including MVTec. Since removal from the fixed max aggregation can only decrease abnormality scores, the accompanying FPO reductions are expected,

(a) Image-level association between prototype confusion and segmentation performance. Each cell reports Spearman ρ for Dice / abnormal FPO. Bold values indicate that the 95% bootstrap CI excludes zero in the hypothesized direction (negative for Dice, positive for FPO).

Dataset	p_{04}	p_{05}	p_{06}	p_{07}
Retina	**-0.68 / +0.42**	**-0.70 / +0.47**	-0.22 / **+0.26**	**-0.65 / +0.50**
Breast	**-0.26 / +0.45**	**-0.39 / +0.56**	-0.07 / -0.15	**-0.38 / +0.58**
Brain	**-0.48 / +0.43**	**-0.56** / +0.20	-0.06 / +0.08	**-0.53** / +0.25
Lung	**-0.62 / +0.26**	**-0.67 / +0.29**	**-0.50** / +0.08	**-0.68 / +0.27**
Significant		Dice: 13/16	FPO: 11/16	

(b) Prototype contributions and false-positive behavior in the Brain ablation study. **Vent.** denotes its mean gated contribution over non-lesion ventricular pixels within each image. All values are reported in percentage (%).

Condition	p_{05} **Vent.**	p_{07} **Vent.**	**Abn. FPO**	**Dice-1**
Brain	58.59	56.81	10.81	34.91
BrainVent	41.38	38.39	7.89	37.06
BrainSmooth	59.56	58.61	9.59	34.41

Table 4: Prototype-level analysis of transfer behavior and the Brain failure case.

while the Dice changes reflect the trade-off between false-positive suppression and true-positive loss. Overall, these results establish prediction relevance while showing that prototype utility is domain-dependent and distinct from the extent of spatial involvement.

4.3 RQ3: When does prototype evidence become confused with normal target anatomy, and how is this associated with transfer failure?

From Section 4.2.1, p_{04}, p_{05}, and p_{07} capture source-domain defect patterns characterized by visually distinctive regions relative to their surroundings. These patterns transfer more successfully to Retina and Breast, where abnormalities remain visually distinguishable from surrounding tissue. In contrast, normal anatomical structures in Brain and Lung, such as ventricles, vessels, and airways, can exhibit similar visual cues and attract abnormal prototype responses. This suggests that transfer is more successful when prototype-associated abnormal patterns remain visually distinctive from normal target-domain structures. To quantify this behavior, for each abnormal target image and prototype, we define a prototype confusion score as the difference between its mean activation on non-lesion region and on the lesion. Higher scores indicate greater relative activation on non-lesion region. We compute Spearman correlations between this score and per-image Dice and abnormal FPO, with 95% bootstrap confidence intervals. As shown in Table 4a, greater prototype confusion is generally associated with lower Dice and higher false-positive behavior, particularly for p_{04}, p_{05}, and p_{07}, while p_{06} shows a weaker and less consistent association.

We further examine Brain as a focused failure case, where p_{05} and p_{07} frequently respond to ventricular structures. Using image intensity and spatial location, we identify candidate ventricular regions without lesion masks and replace them using Telea inpainting (Telea, 2004), while leaving the model unchanged. We compare against BrainSmooth, a control applied to a non-ventricular parenchymal region outside the annotated lesion. Table 4b shows that modifying the ventricular region substantially reduces the contributions of p_{05} and p_{07}, lowers false positives, and improves segmentation, whereas BrainSmooth does not reproduce this improvement. Together, these results show that prototype-level analysis can identify normal target structures associated with transfer failures.

5 Conclusion

We introduced ProtoZSAD, a prototype-based framework that supports zero-shot abnormal-region localization while exposing intermediate units for analyzing source-to-target transfer. Across four unseen medical domains, prototype-level interventions reveal distinct, domain-dependent contributions to final predictions, while greater prototype response to normal anatomy relative to lesions is associated with lower Dice and increased false positives. A targeted Brain-domain intervention further reduces the implicated prototype responses and false positives while improving segmentation. Together, these results show how prototype-level analysis can reveal when source-associated evidence remains aligned with true abnormalities and when it is redirected toward normal target anatomy under domain shift.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, including correcting grammatical and spelling errors and improving the coherence and readability of the manuscript. We also used generative AI to assist in converting experimental results from CSV files into LaTeX tables. We manually reviewed all AI-assisted content, with particular attention to terminology, scientific claims, and descriptions of experimental results, to ensure that the generated wording accurately reflected the underlying methods, observations, and interpretations.

REFERENCES

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Hamed Ayoobi, Nico Potyka, and Francesca Toni. Protoargnet: Interpretable image classification with super-prototypes and argumentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 1791–1799, 2025.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592. IEEE, 2019.

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, 2024.

Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable ai techniques in healthcare. *Sensors*, 23(2):634, 2023.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE transactions on medical imaging*, 39(8):2626–2637, 2020.

Tao Gong, Qi Chu, Bin Liu, Wei Zhou, and Nenghai Yu. Fe-clip: Frequency enhanced clip model for zero-shot anomaly detection and segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21220–21230. IEEE, 2025.

Rose Gurung, Ronilo Ragodos, Chiyu Ma, Tong Wang, and Chaofan Chen. Protopairnet: Interpretable regression through prototypical pair reasoning. *Advances in Neural Information Processing Systems*, 38:110144–110181, 2026.

Yanning Hou, Peiyuan Li, Zirui Liu, Yitong Wang, Yanran Ruan, Jianfeng Qiu, and Ke Xu. Visualad: Language-free zero-shot anomaly detection via vision transformer. *arXiv preprint arXiv:2603.07952*, 2026.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11375–11385. IEEE, 2024.

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.

Yankai Jiang, Wenhui Lei, Xiaofan Zhang, and Shaoting Zhang. Unleashing the potential of vision-language pre-training for 3d zero-shot lesion segmentation via mask-attribute alignment. In *International Conference on Learning Representations*, volume 2025, pp. 38926–38951, 2025.

Rémi Kazmierczak, Eloïse Berthier, Goran Frehse, and Gianni Franchi. Explainability and vision foundation models: A survey. *Information Fusion*, 122:103184, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Chiyu Ma, Brandon Zhao, Chaofan Chen, and Cynthia Rudin. This looks like those: Illuminating prototypical concepts using multiple visualizations. *Advances in Neural Information Processing Systems*, 36:39212–39235, 2023.

Chiyu Ma, Jon Donnelly, Wenjun Liu, Soroush Vosoughi, Cynthia Rudin, and Chaofan Chen. Interpretable image classification with adaptive prototype-based vision transformers. *Advances in Neural Information Processing Systems*, 37:41447–41493, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4744–4754. IEEE, 2025.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Meike Nauta, J¨org Schl¨otterer, Maurice Van Keulen, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2744–2753, 2023.

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, 2021.

Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004.

Yibo Wan, Jinyu Cai, et al. Beyond static anchors: Bounded prototype conditioning for language-free medical anomaly detection. *arXiv preprint arXiv:2608.00442*, 2026.

Chong Wang, Yuyuan Liu, Yuanhong Chen, Fengbei Liu, Yu Tian, Davis McCarthy, Helen Frazer, and Gustavo Carneiro. Learning support and trivial prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2062–2072, 2023.

Chaoran Xu, Chengkan Lv, Qiyu Chen, Feng Zhang, and Zhengtao Zhang. Mrad: Zero-shot anomaly detection with memory-driven retrieval. *arXiv preprint arXiv:2602.00522*, 2026.

Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. *arXiv preprint arXiv:2208.10431*, 2022.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE transactions on emerging topics in computational intelligence*, 5(5):726–742, 2021.

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*, volume 2024, pp. 49705–49737, 2024.

Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2582–2593, 2022.