Beyond Static Anchors: Bounded Prototype Conditioning for Language-Free Medical Anomaly Detection

Yibo Wan[1], **Jinyu Cai**[2*], **See-Kiong Ng**[2]

[1]School of Computing, National University of Singapore, Singapore
[2]Institute of Data Science, National University of Singapore, Singapore
yibo.wan19@outlook.com, {jinyucai,seekiong}@nus.edu.sg

arXiv:2608.00442v1 [cs.CV] 1 Aug 2026

Abstract

Medical anomaly detection identifies abnormal images and localizes lesions under scarce supervision while generalizing across organs and modalities. Existing CLIP-based methods reduce annotation requirements through vision–language alignment, but their normal and abnormal references, whether text prompts or learned visual tokens, remain fixed across test images. Such static references may not transfer reliably to unseen targets in a cross-domain medical imaging scenario. To address this, we propose ReCAP, a language-free framework that replaces static anchors with input-conditioned visual prototypes. ReCAP re-centers separated normal and abnormal prototypes for each image through a bounded gated modulation, enabling query-adaptive anomaly scoring while constraining context-induced prototype drift. For the few-shot setting, we introduce a non-parametric normal-reference memory to preserve instance-level target-domain variation and complement the conditional prototype branch. Across six medical benchmarks, ReCAP achieves the best image-level AUROC on all zero-shot and 23 of 24 few-shot settings, and the best zero-shot pixel-level AUROC on all three segmentation datasets. Particularly, it reduces inference latency by over 70% compared to the fastest baseline, without text prompts or test-time gradient updates.

1 Introduction

Medical anomaly detection (AD) aims to identify abnormal images and localize lesion regions under limited supervision (Cai et al. 2025; Wolleb et al. 2022). This problem is especially important in medical imaging, where lesions are rare, diverse, and expensive to annotate. Therefore, a practical medical detector must operate under scarce supervision while remaining transferable across organs and modalities such as brain MRI, liver CT, and chest X-ray. This cross-domain setting is particularly challenging: normal appearance changes substantially across anatomical structures and imaging protocols, so an anomaly reference learned from source domains may be poorly calibrated for an unseen target domain.

Recent CLIP-based anomaly detectors reduce annotation cost by exploiting transferable vision–language representations (Radford et al. 2021). Prompt-based methods define normal and abnormal semantics with hand-crafted or learnable text prompts, while visual adaptation methods use multi-level CLIP image features for anomaly classification and localization (Huang et al. 2024; Shiri, Beyan, and Murino 2025; Ma et al. 2025). More recent language-free methods, such as VisualAD (Hou et al. 2026) and UniADet (Gao and Wang 2026), remove the text encoder and learn normal/abnormal visual tokens directly in the image feature space. Although these methods differ in how they construct anomaly anchors, most of them share a common assumption: the normal/abnormal reference is fixed after training and reused for all test images. Such a static scoring reference can be brittle in cross-domain medical AD, where the feature-space location of normality may vary sharply across target domains.

This static-anchor formulation leads to three limitations. First, text prompts such as "normal" and "abnormal" provide coarse semantic supervision and are sensitive to wording, whereas medical lesions are often subtle, localized, and difficult to describe with generic language. Second, removing text does not by itself solve cross-domain miscalibration, because learned visual anchors can still define a fixed anomaly boundary that is misplaced for an unseen target. Third, compact global anchors cannot preserve the diverse instance-level normal patterns needed for fine-grained localization, especially when only a few target-domain support images are available. These limitations suggest that robust medical AD requires not only language-free anomaly anchors, but also a mechanism that adapts the anomaly boundary to each input while preserving normal/abnormal separability.

To address these challenges, we propose **Re-C**entered **A**nomaly **P**rototypes (**ReCAP**), a language-free framework for zero- and few-shot medical anomaly detection, as illustrated in Figure 1. The central idea is to replace a static anomaly reference with an input-conditioned visual prototype boundary. ReCAP re-centers a learned and separated prototype pair for each image according to its global visual context, producing an image-specific boundary without text prompts or test-time gradient updates. This re-centering is deliberately bounded to mitigate excessive prototype drift, because the conditioning signal is pooled from the input image itself and may be contaminated by the very lesion we aim to detect. Therefore, the boundary could shift toward anomalous content and explain it away as normal if left unconstrained. Together with a boundary-separation regularizer, the bounded design allows the prototype pair to adapt toward the current domain while preserving normal/abnormal

*Corresponding author.

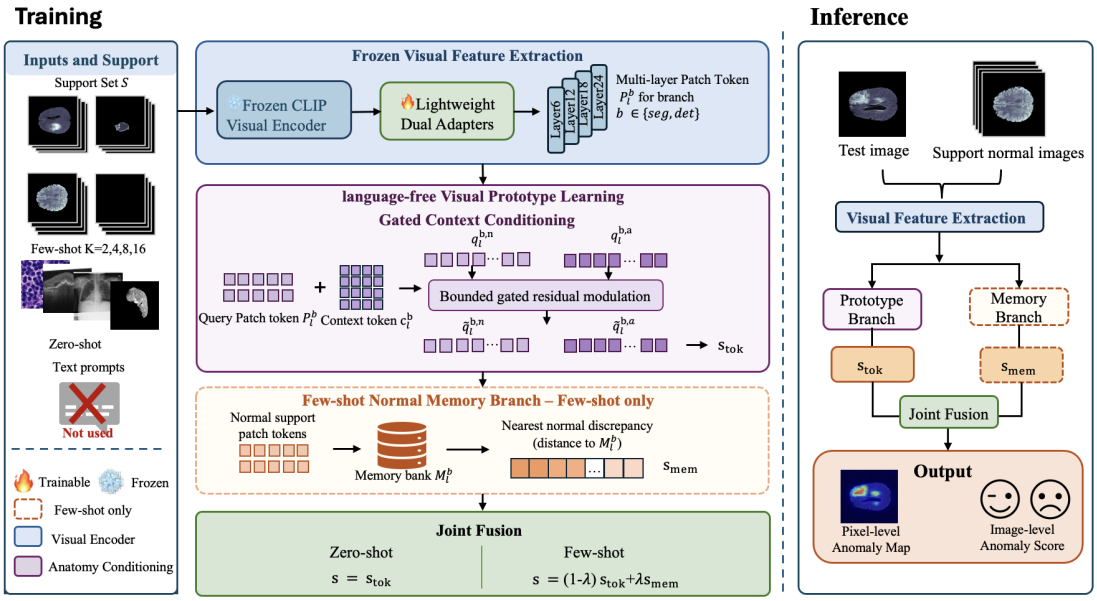

Figure 1: Overview of ReCAP. Instead of aligning visual features with hand-crafted text prompts, our framework learns normal and abnormal visual prototypes directly in the CLIP visual space. The prototypes are further modulated by query-specific anatomical context to produce adaptive visual anchors for anomaly scoring. In the few-shot setting, an auxiliary normal memory bank provides instance-level normal references, while the zero-shot setting relies solely on the learned visual prototypes.

discrimination, providing a controlled, gradient-free adaptation mechanism that supports zero-shot transfer. For few-shot detection, ReCAP further uses target-domain support images to initialize target-specific prototypes and builds a non-parametric normal-reference memory from normal support samples, preserving instance-level normal variation that complements the conditional prototype branch. Multi-layer fusion combines evidence from different CLIP layers and integrates prototype-based discrimination with memory-based deviation, producing both image-level anomaly scores and pixel-level anomaly maps. Built on a frozen CLIP visual encoder with lightweight adapters, ReCAP remains language-free, adaptive, and efficient. Our main contributions are summarized as follows:

- We identify static anomaly anchors as a key limitation of CLIP-based medical AD under cross-domain shifts, and propose input-conditioned visual prototypes for language-free image-specific boundary prediction.
- We introduce bounded gated prototype modulation to re-centers normal/abnormal visual prototypes according to input context while mitigating uncontrolled prototype drift.
- We incorporate normal-reference memory and multi-layer fusion for few-shot detection and localization, achieving state-of-the-art image- and pixel-level performance with improved inference efficiency across six benchmarks.

2 Related Work

Traditional Medical Anomaly Detection. Medical anomaly detection has been studied through reconstruction-based, feature embedding-based, and supervised paradigms (Deng and Li 2022; He et al. 2024; Sabokrou et al. 2018; Wang et al. 2025). Reconstruction methods learn normal-image distributions and detect anomalies by reconstruction discrepancy (Schlegl et al. 2017; Zong et al. 2018; Gong et al. 2019; Gudovskiy, Ishizaka, and Kozuka 2022; Li et al. 2025), while embedding-based methods model normality in a representation space using one-class objectives, memory banks, or distributional estimates (Roth et al. 2022; Liu et al. 2023; Koshil et al. 2024; Wang et al. 2025). Supervised methods can provide stronger discriminative signals but require costly anomaly annotations and often generalize poorly to unseen anomaly types (Wang et al. 2025; Rolih, Fučka, and Skočaj 2024). These limitations are especially pronounced in medical imaging, where abnormal samples are scarce and normal anatomy varies substantially across organs and modalities.

CLIP-based Anomaly Detection. CLIP-based anomaly detectors reduce annotation cost by transferring vision–language representations to zero-/few-shot anomaly detection (Radford et al. 2021; Huang et al. 2024; Zhang et al. 2024). Most methods define anomaly semantics through textual prompts or learnable text-side representations, and score images or patches by contrasting normal and abnormal prompt similarities (Wang et al. 2022; Zhou et al. 2025; Ma et al. 2025; Cao et al. 2024; Shiri, Beyan, and Murino 2025; Huang et al. 2025; Zhang et al. 2024). In medical imaging, this paradigm has been extended with medical vision–language pretraining, multi-level visual adaptation, and stronger normal/abnormal semantic separation (Huang et al. 2024; Ma et al. 2025; Zhou et al. 2023; Shiri, Beyan, and

Murino 2025). However, prompt-based methods remain sensitive to textual descriptions, whereas many medical anomalies are subtle structural deviations that are difficult to capture with generic words such as *normal* and *abnormal*.

Language-free CLIP-based Anomaly Detection. Recent language-free methods remove the dependence on textual prompts by learning anomaly semantics directly in the CLIP visual space (Hou et al. 2026; Gao and Wang 2026). VisualAD (Hou et al. 2026) replaces text prompts with learnable normal/abnormal visual tokens, and UniADet (Gao and Wang 2026) learns task- and level-specific visual decision weights for universal anomaly classification and segmentation. Although these methods demonstrate the promise of prompt-free anomaly modeling, their anomaly references are still largely shared across images or tasks. Such static visual anchors can be poorly calibrated under cross-domain medical shifts, where normal appearance and abnormal evidence depend strongly on anatomy, modality, and local context. Different from these methods, ReCAP learns language-free visual prototypes that are conditionally re-centered for each input and further complemented by support-derived normal references for few-shot medical anomaly detection.

3 Methodology

3.1 Problem Formulation

We study language-free medical anomaly detection under zero-shot and few-shot settings. Given a query image $x \in \mathbb{R}^{H\times W\times 3}$, the goal is to predict an image-level anomaly score $s(x)$ and, when pixel-level annotations are available, an anomaly map $\mathbf{A}(x) \in \mathbb{R}^{H\times W}$. The benchmark contains multiple medical domains $\{\mathcal{D}^{(j)}\}_{j=1}^{J}$. In the leave-one-out zero-shot setting, one domain is held out for testing, while the remaining domains form $\mathcal{D}_{\text{src}}$ for source-domain training. In the few-shot setting, a target-domain support set $\mathcal{S} = \{(x_i, y_i, m_i)\}_{i=1}^{N}$ is provided, where $y_i \in \{0, 1\}$ denotes the image-level label and m_i is the lesion mask when available.

ReCAP performs anomaly scoring entirely in the visual feature space using two complementary branches (Figure 1). The conditional prototype branch produces an input-specific normal/abnormal boundary, while the few-shot memory branch measures deviations from normal support references. The complete procedure is provided in Appendix A.

3.2 Multi-level Features and Visual Prototypes

Multi-level Patch Tokens. We use a frozen CLIP visual encoder and insert lightweight segmentation and detection adapters into selected transformer layers Ω ($\{6, 12, 18, 24\}$ for ViT-L/14). The segmentation adapter preserves spatially faithful patch features for dense anomaly localization, while the detection adapter produces discriminative patch responses for image-level anomaly scoring. We extract ℓ_2-normalized multi-level patch tokens from both adapters and use them for prototype-based scoring. Detailed multi-level patch token extraction is provided in Appendix B.1.

Visual Prototypes. In place of text embeddings, each branch and layer maintains a normal prototype $q_l^{b,n} \in \mathbb{R}^C$ and abnormal prototype $q_l^{b,a} \in \mathbb{R}^C$. These prototypes define a base normal/abnormal boundary that is re-centered for each input by the conditional modulation module. In the few-shot setting, the prototypes are initialized from the corresponding normal and abnormal support patch statistics; in the zero-shot setting, source-trained prototypes serve as the base anchors for the unseen target. In both settings, the prototypes are optimized together with the adapters and conditioning modules. Detailed initialization is provided in Appendix B.2.

3.3 Conditional Prototype Modulation

A static prototype is a single point in feature space, yet the optimal normal/abnormal boundary depends on the location of the input. For example, the normal-feature distribution of brain MRI sits far from that of liver CT or retinal OCT. We therefore make the boundary input-conditional and predict it for each image based on its own appearance.

Input Context Descriptor. For each branch and layer, we summarize the query into a context descriptor by average-pooling its patch tokens followed by normalization:

$$c_l^b(x) = \text{Norm}\left(\frac{1}{R_l}\sum_{r=1}^{R_l} p_{l,r}^b(x)\right) \in \mathbb{R}^C. \quad (1)$$

This descriptor captures the global appearance of the image, such as its modality and dominant structure, and serves as the conditioning signal.

Bounded Gated Modulation. Each prototype is re-centered by the context through a bounded gated residual:

$$\tilde{q}_l^{b,z} = \text{Norm}\left(q_l^{b,z} + \sigma(\beta_l^{b,z})\,\eta\,\tanh\left(W_l^{b,z}\,c_l^b(x)\right)\right), \quad (2)$$

where $W_l^{b,z} \in \mathbb{R}^{C\times C}$ and $\beta_l^{b,z} \in \mathbb{R}$ are learnable, $\sigma(\cdot)$ is the sigmoid, and $\eta > 0$ is a fixed modulation strength. The bound on this residual is deliberate and specific to anomaly detection. Because $c_l^b(x)$ is pooled over the entire image, on an abnormal input, it is *contaminated* by the lesion itself. An unconstrained modulation could exploit this by shifting the normal anchor toward the anomalous content, thereby lowering the training loss while making genuine anomalies appear normal at test time. To mitigate this, the learnable gate $\sigma(\beta_l^{b,z}) \in (0, 1)$ and the bounded $\tanh$ keep the residual small, so that the conditioned anchor $\tilde{q}_l^{b,z}$ stays close to the learned base prototype. The boundary, therefore, re-centers toward the current domain rather than collapsing onto arbitrary content. As a result, a single set of source-trained prototypes can adapt to an unseen domain in a single forward pass, providing an amortized, gradient-free alternative to test-time adaptation that requires no target labels.

Patch-level and Image-level Scoring. For a query patch $\hat{p}_{l,r}^b(x)$, we form scaled-cosine logits against the conditioned anchors:

$$g_{l,r}^b(x) = \alpha\big[\,\langle \hat{p}_{l,r}^b(x), \tilde{q}_l^{b,n}\rangle, \langle \hat{p}_{l,r}^b(x), \tilde{q}_l^{b,a}\rangle\,\big] \in \mathbb{R}^2, \quad (3)$$

with a fixed logit scale α. The patch-level abnormal probability is the abnormal entry of the softmax, $a_{l,r}^b(x) =$

$\mathrm{softmax}\big(g^b_{l,r}(x)\big)_a$. Equivalently, $a^b_{l,r}(x) = \sigma\big(\phi^b_{l,r}(x)\big)$ with the patch anomaly logit:

$$\phi^b_{l,r}(x) = \alpha\big(\langle \hat{p}^b_{l,r}(x), \tilde{q}^{b,a}_l \rangle - \langle \hat{p}^b_{l,r}(x), \tilde{q}^{b,n}_l \rangle\big), \quad (4)$$

which is the normal-versus-abnormal margin under the conditioned boundary. For localization, the segmentation adapter reshapes the two-class patch logits $\{g^{\mathrm{seg}}_{l,r}(x)\}$ into 2D logit maps and upsamples them to the input resolution. A channel-wise softmax is then applied, and the abnormal channel forms the layer-wise anomaly map. For image-level detection, the patch-level abnormal probabilities $\{a^{\mathrm{det}}_{l,r}(x)\}$ are aggregated using the multi-layer fusion described next.

3.4 Normal-Reference Memory and Layer Fusion

Normal-Reference Memory. The conditional prototype branch models normality using a compact re-centered anchor. To retain instance-level normal variation in the few-shot setting, we build a non-parametric memory $\mathcal{M}^b_l$ from the patch tokens of all normal support images ($y_i = 0$). Each query patch is scored by its nearest-neighbor cosine distance:

$$d^b_{l,r}(x) = \min_{u \in \mathcal{M}^b_l} \left(1 - \langle \hat{p}^b_{l,r}(x), \hat{u} \rangle\right), \quad (5)$$

where $\hat{u}$ is an ℓ_2-normalized memory feature. Since the cosine distance between normalized features lies in $[0, 2]$, we map it to the prototype probability range as

$$\bar{d}^b_{l,r}(x) = \mathrm{clip}\left(\frac{d^b_{l,r}(x)}{2},\, 0,\, 1\right), \quad (6)$$

where $\bar{d}^b_{l,r}(x)$ denotes the normalized memory anomaly response used for subsequent layer and branch fusion. Large normalized distances indicate deviations from observed target-domain normal patterns. The memory branch therefore complements the compact prototype boundary with instance-level normal references.

Learnable Layer Fusion. Both branches produce anomaly evidence at each selected layer in Ω. For image-level detection, the patch-level abnormal probabilities are aggregated by top-k pooling and fused with learnable layer weights:

$$\begin{aligned} s_{\mathrm{tok}}(x) &= \sum_{l \in \Omega} \omega_l \,\mathrm{TopK}_r\left(a^{\mathrm{det}}_{l,r}(x)\right), \\ \{\omega_l\}_{l \in \Omega} &= \mathrm{softmax}(\theta), \end{aligned} \quad (7)$$

where $\mathrm{TopK}_r(\cdot)$ averages the largest k abnormal probabilities within each layer. The memory score is aggregated from the normalized memory distances using the same layer weights:

$$s_{\mathrm{mem}}(x) = \sum_{l \in \Omega} \omega_l \,\mathrm{TopK}_r\left(\bar{d}^{\mathrm{det}}_{l,r}(x)\right). \quad (8)$$

For localization, the patch-level abnormal probabilities and normalized memory distances from the segmentation branch are reshaped into 2D maps, upsampled to the input resolution, and fused with segmentation-specific layer weights to obtain $A_{\mathrm{tok}}(x)$ and $A_{\mathrm{mem}}(x)$.

Branch Fusion and Inference. In the few-shot setting, the prototype and memory predictions are fused as

$$s(x) = (1 - \lambda) s_{\mathrm{tok}}(x) + \lambda s_{\mathrm{mem}}(x), \quad (9)$$

where $\lambda \in [0, 1]$ balances the two branches and is set to 0.5 by default. In the zero-shot setting, no memory is constructed, and prediction reduces to the prototype branch with $\lambda = 0$. The pixel-level anomaly map $A(x)$ is fused analogously from $A_{\mathrm{tok}}(x)$ and $A_{\mathrm{mem}}(x)$.

3.5 Training Objectives

ReCAP is optimized on the source domains $\mathcal{D}_{\mathrm{src}}$ in the zero-shot setting or on the target-domain support set $\mathcal{S}$ in the few-shot setting. The training objective contains three terms: a pixel-level localization loss when lesion masks are available, an image-level detection loss, and a prototype separation regularizer. The first two losses supervise the task-specific scoring functions, while the separation regularizer keeps the base normal/abnormal prototypes discriminative before input-specific re-centering.

Pixel-level Localization and Image-level Detection. For a training sample (x_i, y_i, m_i), when the lesion mask m_i is available, the predicted anomaly map $\mathbf{A}(x_i)$ is supervised by a focal-plus-Dice loss:

$$\mathcal{L}_{\mathrm{seg}} = \mathcal{L}_{\mathrm{focal}}\big(\mathbf{A}(x_i), m_i\big) + \mathcal{L}_{\mathrm{dice}}\big(\mathbf{A}(x_i), m_i\big). \quad (10)$$

This loss encourages high anomaly responses on lesion regions and suppresses activations on normal tissues. For image-level detection, the anomaly score $s(x_i)$ is supervised by binary cross-entropy:

$$\mathcal{L}_{\mathrm{det}} = \mathrm{BCEWithLogits}\left(\mathrm{logit}(s(x_i)), y_i\right). \quad (11)$$

This objective trains the detection branch to separate normal and abnormal images under the same scoring function used at inference.

Prototype Separation Regularization. Since conditional modulation re-centers an existing normal/abnormal boundary, the base prototypes should remain discriminative before adaptation. We therefore enforce a margin between the normal and abnormal prototypes in each branch and layer:

$$\mathcal{L}_{\mathrm{sep}} = \frac{1}{2|\Omega|} \sum_{b \in \{\mathrm{seg}, \mathrm{det}\}} \sum_{l \in \Omega} \max\left(0, \cos(q^{b,n}_l, q^{b,a}_l) - \delta\right), \quad (12)$$

where δ is the separation margin. This regularizer is coupled with bounded modulation: the margin preserves a discriminative base boundary, while the modulation bound allows the boundary to shift toward the current input without collapsing onto lesion-contaminated context.

Overall Loss Function. The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}} + \mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{sep}} \mathcal{L}_{\mathrm{sep}}. \quad (13)$$

For datasets without pixel-level annotations, $\mathcal{L}_{\mathrm{seg}}$ is omitted, and ReCAP is trained with image-level supervision and prototype separation.

Few-shot	Metric	Dataset	DRA	APRIL-GAN	MediCLIP	MVFA	MadCLIP	VisualAD	ReCAP
			N/A	✗	✗	✗	✗	✓	✓
$K=2$	Image-level (AUROC, F_1-max, AP)	HIS	(72.91, 71.52, 73.20)	(69.57, 69.24, 70.52)	(64.49, 66.81, 65.20)	(79.75, 76.96, 71.72)	(83.62, 77.60, 84.22)	(53.80, 67.13, 53.14)	(**85.89**, **78.44**, **86.10**)
		Chest	(72.22, 96.82, 97.82)	(69.84, 96.64, 97.48)	(61.69, 96.39, 96.12)	(78.38, 97.68, 98.48)	(84.56, 97.60, 98.90)	(53.70, **97.70**, 95.62)	(**84.71**, 97.68, **99.00**)
		OCT17	(98.08, 96.41, 98.70)	(99.21, 92.80, 99.40)	(93.40, 92.73, 96.10)	(96.53, 96.83, 98.56)	(99.06, 97.50, 99.30)	(92.40, 91.02, 97.64)	(**99.40**, **98.26**, **99.70**)
		BrainMRI	(71.78, 84.64, 89.20)	(78.45, 88.34, 92.40)	(85.13, 90.61, 94.80)	(91.95, 92.88, 98.23)	(93.93, **94.20**, 98.20)	(83.90, 92.04, 95.33)	(**94.35**, 94.08, **98.48**)
		LiverCT	(57.17, 62.34, 65.89)	(57.80, 63.11, 66.23)	(68.48, 70.20, 74.51)	(87.23, 77.79, 88.74)	(84.48, 78.40, 87.30)	(84.42, 74.23, 84.91)	(**90.45**, **80.20**, **90.00**)
		RESC	(85.69, 78.82, 85.60)	(89.44, 81.24, 88.23)	(83.96, 78.11, 84.40)	(94.70, 84.87, 93.58)	(95.09, 86.70, 93.40)	(91.20, 82.60, 87.42)	(**95.47**, **87.29**, **94.33**)
	Pixel-level (AUROC, F_1-max, PRO)	BrainMRI	(72.09, 18.44, 52.64)	(94.02, 36.82, 78.40)	(97.39, 42.60, 86.74)	(97.41, 35.84, 77.85)	(**97.92**, 45.21, 88.64)	(95.00, 45.71, 80.21)	(97.52, **50.24**, **89.03**)
		LiverCT	(63.13, 20.10, 48.54)	(95.87, 44.62, 78.22)	(97.09, 51.83, 82.44)	(96.67, 62.17, 82.79)	(99.39, 59.40, 87.60)	(99.10, 34.90, 76.67)	(**99.64**, **62.89**, **84.61**)
		RESC	(65.59, 28.62, 46.81)	(96.39, 63.50, 75.64)	(96.01, 61.84, 74.20)	(97.10, 62.34, **87.55**)	(97.18, 67.80, 78.40)	(97.10, 64.90, 78.16)	(**98.37**, **70.85**, 78.91)
$K=4$	Image-level (AUROC, F_1-max, AP)	HIS	(68.73, 69.84, 68.71)	(76.11, 73.20, 76.20)	(70.85, 70.64, 72.83)	(80.05, 77.50, 75.67)	(80.05, 76.40, 78.26)	(67.30, 70.00, 64.40)	(**82.29**, **79.69**, **79.85**)
		Chest	(75.81, 96.30, 98.10)	(77.43, 97.14, 98.20)	(56.83, 96.12, 95.40)	(82.51, 97.69, 97.84)	(88.14, 95.70, **99.20**)	(65.20, 97.30, 97.50)	(**88.92**, **98.68**, 98.84)
		OCT17	(99.06, 97.34, 99.30)	(99.41, 98.00, 99.45)	(89.07, 91.50, 95.40)	(99.38, 98.64, 99.77)	(99.37, 98.15, 99.63)	(98.30, 96.42, 99.43)	(**99.86**, **99.24**, **99.95**)
		BrainMRI	(80.62, 88.22, 93.44)	(89.18, 92.30, 96.80)	(83.82, 90.14, 94.12)	(88.89, 91.62, 97.57)	(**95.25**, **95.40**, 98.28)	(77.50, 91.23, 93.62)	(95.14, 95.01, **98.90**)
		LiverCT	(59.64, 64.12, 68.55)	(53.05, 60.70, 62.82)	(81.53, 75.64, 84.74)	(82.10, 73.49, 83.95)	(82.97, 76.80, 86.20)	(59.40, 61.92, 56.63)	(**87.56**, **77.30**, **87.47**)
		RESC	(90.90, 82.32, 89.44)	(94.70, 85.40, 93.22)	(87.52, 80.25, 87.10)	(94.67, 87.04, 91.61)	(96.62, 88.30, 95.60)	(91.20, 81.40, 87.60)	(**96.94**, **89.29**, **96.62**)
	Pixel-level (AUROC, F_1-max, PRO)	BrainMRI	(74.77, 20.62, 55.30)	(94.67, 38.54, 79.82)	(96.86, 41.80, 84.70)	(95.05, 36.64, 76.91)	(**97.90**, 45.20, 88.20)	(84.90, 19.80, 82.10)	(97.56, **56.76**, **89.19**)
		LiverCT	(71.79, 25.32, 54.72)	(96.24, 47.90, 79.44)	(98.61, 55.44, 84.61)	(99.65, 60.10, **91.05**)	(99.29, 58.81, 87.23)	(98.40, 22.20, 79.22)	(**99.77**, **61.08**, 89.42)
		RESC	(77.28, 34.22, 55.82)	(97.98, 70.40, 82.60)	(96.65, 64.54, 76.84)	(98.73, 77.86, **93.42**)	(98.90, 75.80, 89.80)	(97.50, 65.00, 90.40)	(**99.11**, **78.05**, 92.37)

Table 1: Comparisons with state-of-the-art few-shot anomaly detection methods under different few-shot settings ($K=2, 4$). Image-level results are reported as (AUROC, F_1-max, AP), and pixel-level results as (AUROC, F_1-max, PRO). ✓denotes language-free methods, while ✗denotes methods using text prompts, and N/A denotes non-CLIP methods. All values are reported in percentage.

4 Experiments

4.1 Experimental Setup

Datasets. Following prior medical anomaly detection benchmarks (Huang et al. 2024), we evaluate ReCAP on six datasets spanning five imaging domains. These include BrainMRI for brain MRI (Baid et al. 2021; Bakas et al. 2017; Menze et al. 2014), LiverCT for liver CT (Bilic et al. 2023; Landman et al. 2015), RESC and OCT17 for retinal OCT (Kermany et al. 2018; Hu, Chen, and Yi 2019), ChestXray for chest X-ray (Wang et al. 2017), and HIS for digital histopathology (Ehteshami Bejnordi et al. 2017). BrainMRI, LiverCT, and RESC contain pixel-level annotations and are evaluated for both anomaly classification (AC) and anomaly segmentation (AS); the remaining datasets are evaluated for AC only. Dataset statistics and preprocessing details are provided in Appendix C.

Competing Methods and Baselines. We compare ReCAP with representative medical anomaly detection methods, including DRA (Ding, Pang, and Shen 2022), APRIL-GAN (Chen, Han, and Zhang 2023), MediCLIP (Zhang et al. 2024), MVFA (Huang et al. 2024), MadCLIP (Shiri, Beyan, and Murino 2025), and VisualAD (Hou et al. 2026). For fair comparison, all methods are evaluated under the same dataset splits, support-shot settings, and image-level or pixel-level evaluation protocols whenever applicable. For all few-shot experiments, every method uses the same predefined support indices for each dataset and shot setting.

Evaluation Metrics. For anomaly classification (AC), we report image-level AUROC, the maximum F_1 score across decision thresholds (F_1-max), and average precision (AP). For anomaly segmentation (AS), we use pixel-level AUROC, pixel-level F_1-max, and per-region overlap (PRO). All metrics are reported as percentages.

Implementation Details. ReCAP uses a frozen CLIP ViT-L/14@336px visual encoder with lightweight segmentation and detection adapters. Images are resized to 240×240, and patch tokens from layers $\{6, 12, 18, 24\}$ are used for multi-level anomaly scoring. Only the adapters, visual prototypes, conditioning modules, and layer-fusion weights are trainable. Training combines image-level detection, pixel-level localization when masks are available, and prototype separation. Complete optimization settings, support protocols, augmentation, and computing resources are listed in Appendix D.

4.2 Comparison with State-of-the-Art Methods

Quantitative Comparison. Table 1 compares ReCAP with representative few-shot anomaly detection methods under low-shot settings ($K = 2, 4$). Under $K = 2$, ReCAP achieves the best image-level AUROC on all six datasets. In particular, it improves the LiverCT AUROC from 84.48% with MadCLIP to 90.45%. For pixel-level segmentation, ReCAP achieves the best AUROC on LiverCT and RESC, reaching 99.64% and 98.37%, respectively, while obtaining the best F_1-max on all three segmentation datasets. Under $K = 4$, ReCAP maintains the best image-level AUROC on five of six datasets and the best pixel-level AUROC on LiverCT and RESC. It also achieves the best pixel-level F_1-max on all three datasets. These results demonstrate the effectiveness of ReCAP for both anomaly classification and localization with limited target-domain supervision.

Table 2 reports zero-shot results on six medical datasets. Without using target-domain support samples, ReCAP achieves the best image-level AUROC on all six datasets, outperforming both prompt-based CLIP methods and the language-free VisualAD baseline. Notable improvements include an increase from 75.26% with MadCLIP to 78.40% on Chest and from 95.40% with MVFA to 96.58% on OCT17.

Setting	Metric	Dataset	APRIL-GAN	MediCLIP	MVFA	MadCLIP	VisualAD	ReCAP
			✗	✗	✗	✗	✓	✓
Zero-shot	Image-level (AUROC, F_1-max, AP)	HIS	(59.36, 66.21, 56.72)	(60.60, 64.20, 62.31)	(76.89, 73.73, 70.95)	(77.32, 74.80, 80.16)	(50.50, 64.21, 50.82)	(**79.32**, **75.86**, **82.41**)
		Chest	(57.49, 95.20, 95.84)	(58.20, 94.90, 94.85)	(71.11, 96.42, 97.19)	(75.26, 96.85, 97.94)	(45.30, 96.40, 94.31)	(**78.40**, **96.92**, **98.21**)
		OCT17	(92.61, 89.69, 96.97)	(87.89, 88.72, 93.43)	(95.40, 93.52, 98.12)	(91.48, 90.80, 96.32)	(88.37, 88.20, 96.62)	(**96.58**, **95.73**, **98.84**)
		BrainMRI	(76.63, 86.40, 90.35)	(79.32, 87.92, 91.61)	(79.80, **92.18**, 93.78)	(74.43, 90.50, 93.06)	(72.36, 83.28, 86.98)	(**80.24**, 91.05, **94.20**)
		LiverCT	(54.62, 60.21, 63.45)	(62.13, 67.64, 70.25)	(81.19, 72.27, 76.94)	(73.48, **74.92**, 82.40)	(70.20, 69.32, 80.15)	(**82.60**, 73.64, **83.43**)
		RESC	(75.67, 76.12, 80.35)	(80.50, 74.90, 82.15)	(89.00, 80.48, 84.12)	(84.10, 82.35, 88.40)	(81.70, 78.12, 84.23)	(**90.23**, **82.78**, **88.95**)
	Pixel-level (AUROC, F_1-max, PRO)	BrainMRI	(91.79, 28.50, 69.42)	(90.21, 30.80, 70.20)	(89.68, 14.49, 59.26)	(91.40, 32.60, 74.50)	(80.40, 12.89, 67.21)	(**92.30**, **36.85**, **79.20**)
		LiverCT	(93.05, 32.10, 66.45)	(94.72, 38.65, 71.30)	(93.93, 18.94, 65.46)	(98.30, 43.50, 75.20)	(97.20, 30.12, 69.40)	(**98.72**, **58.32**, **78.46**)
		RESC	(85.23, 52.12, 63.42)	(89.62, 54.82, 65.30)	(90.44, 35.99, 61.42)	(90.10, 58.46, 68.70)	(84.30, 50.24, 62.60)	(**92.60**, **60.42**, **70.35**)

Table 2: Comparisons with state-of-the-art zero-shot anomaly detection methods. Image-level results are reported as (AUROC, F_1-max, AP), and pixel-level results as (AUROC, F_1-max, PRO). ✓denotes language-free methods, while ✗denotes methods using text prompts. All values are reported in percentage.

For pixel-level segmentation, ReCAP achieves the best AUROC and F_1-max on all three segmentation datasets. These results demonstrate that the proposed language-free conditional prototypes generalize effectively to unseen medical domains while retaining accurate anomaly localization. Appendix E further reports results for $K = 8, 16$ and provides mean$\pm$std AUROC over five random seeds for all shot settings ($K = 2, 4, 8, 16$), showing that the improvements are robust to random variation.

Qualitative Comparison. Figure 2 compares anomaly localization results on BrainMRI, LiverCT, and retinal OCT images. APRIL-GAN produces diffuse heatmaps that extend substantially into normal regions. MVFA and MadCLIP yield more localized responses but occasionally show incomplete lesion coverage or off-target activations, while VisualAD incorrectly highlights normal anatomical structures in several BrainMRI examples. In contrast, ReCAP produces more compact anomaly responses that are visually better aligned with the annotated lesions and exhibit fewer background activations across the three modalities. These results qualitatively support the localization performance reported in our quantitative evaluation. Additional visualizations are provided in Appendix F.

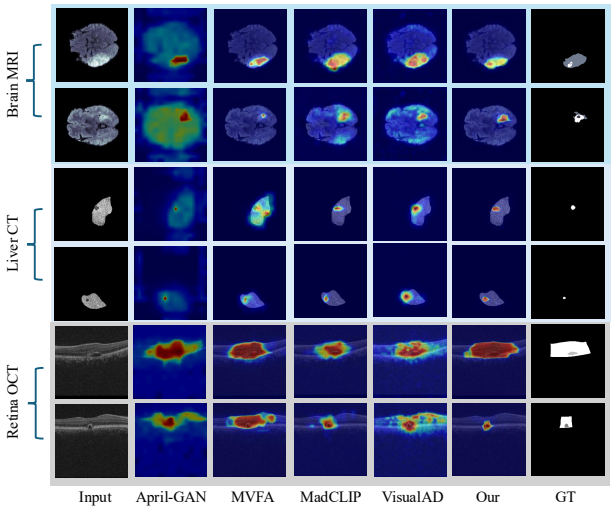

Figure 2: Qualitative comparison of anomaly localization. We compare heatmaps produced by APRIL-GAN, MVFA, MadCLIP, VisualAD, and ReCAP on representative samples from BrainMRI, LiverCT, and retinal OCT datasets.

Anomaly Score Distribution Analysis. Figure 3 shows representative image-level score distributions on OCT17, RESC, BrainMRI, and LiverCT, with the other two datasets reported in Appendix F. ReCAP assigns low scores to most normal samples and high scores to abnormal samples. The separation is particularly clear on OCT17, RESC, and BrainMRI, demonstrating well-separated anomaly responses across retinal and brain imaging domains. Although the distributions partially overlap on the more challenging LiverCT dataset, abnormal samples remain concentrated toward higher scores.

Efficiency Comparison. Table 3 compares model complexity and inference latency under the same input resolution, hardware, and measurement protocol. ReCAP achieves the fastest inference in both zero-shot and four-shot settings, requiring only **17.4** ms and **19.0** ms per image, respectively. The comparison with VisualAD is especially informative because both methods are language-free and use the same ViT-L/14@336px backbone. ReCAP is approximately $3.8\times$ faster in zero-shot inference (17.4 vs. 66.1 ms) and $3.5\times$ faster in the four-shot setting (19.0 vs. 65.8 ms), while using fewer method-specific parameters (22.0M vs. 31.5M). This efficiency gain stems from the different inference designs: VisualAD propagates learnable normal and abnormal tokens through multi-layer token–patch interactions and additional spatial recalibration modules, whereas ReCAP performs lightweight bounded prototype modulation followed by direct patch–prototype cosine scoring. Compared with prompt-based methods, ReCAP further avoids text encoding and image–text matching. This makes ReCAP more suitable for time-sensitive clinical screening workflows. Additional analyses of the visual backbone, fusion weight λ, and top-k ratio are provided in Appendix G.

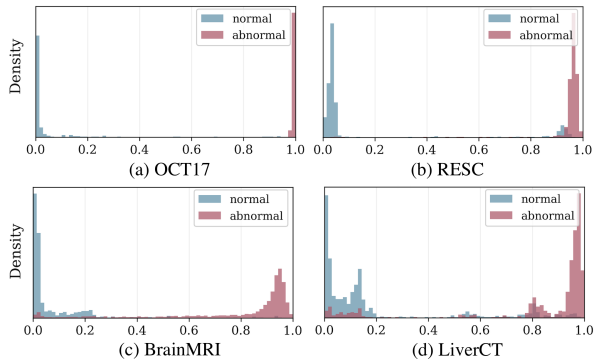

Figure 3: Image-level anomaly score distributions under the 16-shot setting. Normalized scores for normal (blue) and abnormal (red) test samples on four representative datasets.

Shots	Methods	CLIP Models	Input Size	# Params (M)	Inf. Time (ms)
0	APRIL-GAN	ViT-B/16	240×240	149.6 + 1.2	114.5 ± 3.7
0	MVFA	ViT-L/14	240×240	427.7 + 12.6	125.7 ± 5.3
0	MadCLIP	ViT-L/14	240×240	427.7 + 106.9	265.3 ± 11.0
0	VisualAD	ViT-L/14@336px	240×240	427.2 + 31.5	66.1 ± 0.9
0	ReCAP	ViT-L/14@336px	240×240	427.7 + 22.0	$\mathbf{17.4 \pm 2.0}$
4	APRIL-GAN	ViT-B/16	240×240	152.1 + 4.6	120.1 ± 2.4
4	MVFA	ViT-L/14	240×240	427.7 + 32.3	126.2 ± 4.2
4	MadCLIP	ViT-L/14	240×240	427.7 + 92.3	268.5 ± 9.3
4	VisualAD	ViT-L/14@336px	240×240	427.2 + 31.5	65.8 ± 0.9
4	ReCAP	ViT-L/14@336px	240×240	427.7 + 22.0	$\mathbf{19.0 \pm 4.2}$

Table 3: Complexity and inference efficiency under zero- and four-shot settings. Parameter counts are reported as pre-trained backbone parameters + additional method-specific parameters. Inference time is reported as mean $\pm$ standard deviation over 100 runs after 20 warm-up iterations.

4.3 Ablation Studies

Component Analysis. Table 4 evaluates the contributions of the normal memory bank (MB), context-conditioning module (CCM), and learnable layer fusion (LF). The full model achieves the best AC and AS on both datasets, indicating that all three components contribute to the final performance. On LiverCT, removing LF produces the largest reductions, decreasing AC and AS by 4.00 and 2.89 percentage points, respectively, which demonstrates the importance of aggregating anomaly evidence across multiple layers. Removing MB reduces AC and AS by 2.36 and 2.72 points, while removing CCM results in drops of 1.60 and 1.48 points. On RESC, every component removal also degrades both metrics, although the margins are smaller because performance is already close to saturation. Overall, the results support the joint use of target-domain normal references, context-conditioned prototypes, and multi-layer score fusion.

Effect of Controlled Prototype Conditioning. Table 5 compares different language-free prototype designs. Static prototypes outperform the VisualAD-style anchors on all eight metrics, showing the benefit of ReCAP's normal/abnormal prototype representation even without query conditioning. However, these prototypes remain identical for all query images and cannot adapt to anatomy-dependent appearance variations. Adding an unbounded context residual decreases all four zero-shot metrics and produces mixed results under the 4-shot setting, indicating that unconstrained modulation does not reliably improve the anomaly reference. Bounding the residual improves six of eight metrics over its unbounded counterpart, including all AC and RESC results. Introducing the gate further improves every metric, with the bounded gated residual achieving the best AC and AS across both datasets and shot settings. These results support the use of controlled query-conditioned modulation rather than either static or unconstrained anomaly prototypes.

Variant	MB	CCM	LF	LiverCT		RESC	
				AC	AS	AC	AS
w/o MB		✓	✓	85.20	97.05	95.54	98.82
w/o CCM	✓		✓	85.96	98.29	96.84	98.94
w/o LF	✓	✓		83.56	96.88	96.91	98.97
ReCAP	✓	✓	✓	**87.56**	**99.77**	**96.94**	**99.11**

Table 4: Component ablation under the 4-shot setting. AC and AS are reported as AUROC (%).

Variant	Zero-shot				4-shot			
	LiverCT		RESC		LiverCT		RESC	
	AC	AS	AC	AS	AC	AS	AC	AS
VisualAD-style anchors	68.92	96.84	80.35	82.58	58.64	97.42	90.82	96.85
Static prototypes	70.06	98.19	84.77	90.67	80.71	98.91	95.46	97.42
Unbounded residual	67.43	97.99	81.83	88.41	77.70	98.98	95.77	96.93
Bounded residual w/o gate	70.51	97.61	86.84	91.35	81.42	98.92	96.12	98.21
Bounded gated residual	**82.60**	**98.72**	**90.23**	**92.60**	**87.56**	**99.77**	**96.94**	**99.11**

Table 5: Ablation of prototype conditioning under zero- and 4-shot settings. The first two variants omit query conditioning: VisualAD-style uses shared learnable anchors, whereas Static prototypes use ReCAP's normal/abnormal prototypes.

5 Conclusion

This paper presented ReCAP, a language-free framework for zero- and few-shot medical anomaly detection. ReCAP replaces static visual anchors with input-conditioned normal/abnormal prototypes whose updates are bounded and gated to provide adaptive anomaly references. For few-shot transfer, it further combines a normal-reference memory with learnable multi-layer fusion to capture target-domain normal variation and aggregate complementary anomaly cues. Across six medical datasets, ReCAP achieves the best zero-shot image-level AUROC on all six datasets and pixel-level AUROC on all three segmentation datasets, while maintaining strong performance across few-shot settings. Ablation studies confirm the contributions of controlled prototype conditioning, normal-reference memory, and layer fusion, and the efficiency analysis demonstrates fast inference without text prompts. These results establish controlled visual prototype adaptation as an effective alternative to text-based anomaly modeling. Current evaluation focuses on 2D image-based benchmarks; extending ReCAP to volumetric and multimodal clinical data remains an important direction.

References

Baid, U.; Ghodasara, S.; Mohan, S.; Bilello, M.; Calabrese, E.; Colak, E.; Farahani, K.; Kalpathy-Cramer, J.; Kitamura, F. C.; and Pati, S. 2021. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *ArXiv:2107.02314*.

Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J. S.; Freymann, J. B.; Farahani, K.; and Davatzikos, C. 2017. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1): 170117.

Bilic, P.; Christ, P.; Li, H. B.; Vorontsov, E.; Ben-Cohen, A.; Kaissis, G.; Szeskin, A.; Jacobs, C.; Mamani, G. E. H.; Chartrand, G.; et al. 2023. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84: 102680.

Cai, Y.; Zhang, W.; Chen, H.; and Cheng, K.-T. 2025. MedIAnomaly: A comparative study of anomaly detection in medical images. *Medical Image Analysis*, 102: 103500.

Cao, Y.; Zhang, J.; Frittoli, L.; Cheng, Y.; Shen, W.; and Boracchi, G. 2024. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, 55–72. Springer.

Chen, X.; Han, Y.; and Zhang, J. 2023. A zero-/few-shot anomaly classification and segmentation method for CVPR 2023 (VAND) workshop challenge tracks 1 &2. *1st Place on Zero-shot AD and 4th Place on Few-shot AD*, 2305(17382): 2.

Deng, H.; and Li, X. 2022. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9737–9746.

Ding, C.; Pang, G.; and Shen, C. 2022. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7388–7398.

Ehteshami Bejnordi, B.; Veta, M.; Johannes van Diest, P.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J. A.; consortium, C.; Hermsen, M.; Manson, Q. F.; et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22): 2199–2210.

Gao, B.-B.; and Wang, C. 2026. One Language-Free Foundation Model Is Enough for Universal Vision Anomaly Detection. *ArXiv:2601.05552*.

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and Hengel, A. v. d. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1705–1714.

Gudovskiy, D.; Ishizaka, S.; and Kozuka, K. 2022. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 98–107.

He, H.; Zhang, J.; Chen, H.; Chen, X.; Li, Z.; Chen, X.; Wang, Y.; Wang, C.; and Xie, L. 2024. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8472–8480.

Hou, Y.; Li, P.; Liu, Z.; Wang, Y.; Ruan, Y.; Qiu, J.; and Xu, K. 2026. VisualAD: Language-Free Zero-Shot Anomaly Detection via Vision Transformer. *ArXiv:2603.07952*.

Hu, J.; Chen, Y.; and Yi, Z. 2019. Automated segmentation of macular edema in OCT using deep neural networks. *Medical Image Analysis*, 55: 216–227.

Huang, C.; Jiang, A.; Feng, J.; Zhang, Y.; Wang, X.; and Wang, Y. 2024. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11375–11385.

Huang, H.; Sun, W.; Wu, Z.; Niu, J.; Lu, D.; Wu, X.; and Zheng, Y. 2025. IQE-CLIP: Instance-aware Query Embedding for Zero-/Few-shot Anomaly Detection in Medical Domain. *ArXiv:2506.10730*.

Kermany, D. S.; Goldbaum, M.; Cai, W.; Valentim, C. C.; Liang, H.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5): 1122–1131.

Koshil, M.; Wegener, T.; Mentrup, D.; Frintrop, S.; and Wilms, C. 2024. Anomalouspatchcore: exploring the use of anomalous samples in industrial anomaly detection. In *European Conference on Computer Vision*, 294–309. Springer.

Landman, B.; Xu, Z.; Igelsias, J.; Styner, M.; Langerak, T.; and Klein, A. 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proceedings of MICCAI Multi-atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, 12. Munich, Germany.

Li, X.; Tan, X.; Chen, Z.; Zhang, Z.; Zhang, R.; Guo, R.; Jiang, G.; Chen, Y.; Qu, Y.; Ma, L.; et al. 2025. One-for-more: Continual diffusion model for anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4766–4775.

Liu, Z.; Zhou, Y.; Xu, Y.; and Wang, Z. 2023. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20402–20411.

Ma, W.; Zhang, X.; Yao, Q.; Tang, F.; Wu, C.; Li, Y.; Yan, R.; Jiang, Z.; and Zhou, S. K. 2025. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4744–4754.

Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. 2014. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10): 1993–2024.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from

natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PmLR.

Rolih, B.; Fučka, M.; and Skočaj, D. 2024. Supersimplenet: Unifying unsupervised and supervised learning for fast and reliable surface defect detection. In *International Conference on Pattern Recognition*, 47–65. Springer.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.

Sabokrou, M.; Khalooei, M.; Fathy, M.; and Adeli, E. 2018. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3379–3388.

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Schmidt-Erfurth, U.; and Langs, G. 2017. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. ArXiv:1703.05921.

Shih, G.; Wu, C. C.; Halabi, S. S.; Kohli, M. D.; Prevedello, L. M.; Cook, T. S.; Sharma, A.; Amorosa, J. K.; Arteaga, V.; Galperin-Aizenberg, M.; et al. 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1): e180041.

Shiri, M.; Beyan, C.; and Murino, V. 2025. MadCLIP: few-shot medical anomaly detection with CLIP. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 416–426. Springer.

Wang, F.; Zhang, T.; Wang, Y.; Qiu, Y.; Liu, X.; Guo, X.; and Cui, Z. 2025. Distribution prototype diffusion learning for open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20416–20426.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2097–2106.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3876–3887.

Wolleb, J.; Bieder, F.; Sandkühler, R.; and Cattin, P. C. 2022. Diffusion models for medical anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 35–45. Springer.

Zhang, X.; Xu, M.; Qiu, D.; Yan, R.; Lang, N.; and Zhou, X. 2024. Mediclip: Adapting clip for few-shot medical image anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 458–468. Springer.

Zhou, Q.; Pang, G.; Tian, Y.; He, S.; and Chen, J. 2023. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *ArXiv:2310.18961*.

Zhou, Y.; Bi, Y.; Tong, W.; Wang, W.; Navab, N.; and Jiang, Z. 2025. Ultraad: Fine-grained ultrasound anomaly classification via few-shot clip adaptation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 625–635. Springer.

Zong, B.; Song, Q.; Min, M. R.; Cheng, W.; Lumezanu, C.; Cho, D.; and Chen, H. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *Proceedings of the International Conference on Learning Representations*.

A Detailed Algorithm of ReCAP

Algorithm 1: Training and inference procedures of ReCAP

Input: Source domains $\mathcal{D}_{\mathrm{src}}$, optional support set $\mathcal{S}$, query image x, frozen encoder E_v, layers Ω, and conditioning strength η.
Output: Anomaly map $A(x)$ and image-level score $s(x)$.
Freeze E_v and initialize adapters, prototypes, conditioning modules, and fusion weights.

Training:
if $\mathcal{S} \neq \varnothing$ **then**
  Initialize support prototypes and normal memory banks from $\mathcal{S}$.
  Optimize ReCAP on $\mathcal{S}$ with Eqs. (10)–(13).
else
  Optimize ReCAP on $\mathcal{D}_{\mathrm{src}}$ with Eqs. (10)–(13).

Inference:
Extract multi-level patch tokens from x and generate conditioned prototypes by Eqs. (1)–(2).
Compute prototype responses and layer-fused scores by Eqs. (3), (4), and (7).
if $\mathcal{S} \neq \varnothing$ **then**
  Compute memory responses by Eqs. (5)–(6) and fuse both branches by Eqs. (8)–(9).
else
  Set $\lambda = 0$ and use the prototype branch only.
return $A(x)$ *and* $s(x)$.

Algorithm 1 summarizes ReCAP in both zero-shot and few-shot settings. In zero-shot inference, ReCAP uses source-trained visual prototypes and input-conditioned modulation to produce anomaly predictions without target-domain support. In few-shot inference, support images are additionally used to initialize target-domain prototypes and build a normal-reference memory. In both cases, query-specific adaptation is performed by bounded gated prototype conditioning in a single forward pass, without text prompts or test-time gradient updates.

B Additional Method Details

B.1 Multi-level Patch Token Extraction

ReCAP extracts multi-level patch representations from a frozen CLIP visual encoder with lightweight task-specific adapters. Let $\Omega = \{l_1, \ldots, l_M\}$ denote the selected transformer layers. For each branch $b \in \{\mathrm{seg}, \mathrm{det}\}$ and layer $l \in \Omega$, the adapter produces a sequence of patch tokens:

$$P_l^b(x) = \left[p_{l,1}^b(x), \ldots, p_{l,R_l}^b(x)\right]^\top \in \mathbb{R}^{R_l \times C}, \qquad \text{(A1)}$$

where R_l is the number of patch tokens and C is the feature dimension. We use ℓ_2-normalized tokens for prototype matching and memory retrieval:

$$\hat{p}_{l,r}^b(x) = \frac{p_{l,r}^b(x)}{\|p_{l,r}^b(x)\|_2}. \qquad \text{(A2)}$$

The segmentation branch keeps spatially faithful patch responses for pixel-level anomaly maps, while the detection branch produces patch responses that are aggregated into an image-level anomaly score.

B.2 Support-based Prototype Initialization

In the few-shot setting, normal and abnormal visual prototypes are initialized from support-set patch statistics. For each branch b, layer l, and class $z \in \{n, a\}$, we collect the corresponding support patch tokens into $\mathcal{T}_l^{b,z}$. Each prototype is initialized as the normalized mean:

$$q_l^{b,z} = \mathrm{Norm}\left(\frac{1}{|\mathcal{T}_l^{b,z}|}\sum_{u \in \mathcal{T}_l^{b,z}} u\right), \qquad z \in \{n, a\}, \quad \text{(A3)}$$

where $\mathrm{Norm}(v) = v/\|v\|_2$. The sets $\mathcal{T}_l^{b,n}$ and $\mathcal{T}_l^{b,a}$ contain all patch tokens from normal and abnormal support images, respectively. Lesion masks, when available, are used only for pixel-level supervision and not for prototype initialization. In the zero-shot setting, no target-domain support set is available; source-trained prototypes are directly used as base anchors for the unseen target domain.

C Details of Medical Anomaly Detection Datasets

We evaluate ReCAP on six medical anomaly detection datasets covering diverse modalities, anatomical regions, and evaluation settings. As summarized in Table 6, BrainMRI, LiverCT, and RESC provide both image-level anomaly classification (AC) and pixel-level anomaly segmentation (AS) annotations, whereas OCT17, Chest, and HIS are evaluated for image-level anomaly classification only. The labeled pool is used only for source-domain training or few-shot support construction, depending on the evaluation protocol, while the test split is kept fixed for evaluation. Normal-only training splits provided by some datasets are not used under our few-shot or zero-shot protocols. To avoid patient-level leakage, samples from the same patient are assigned to the same split whenever patient/case identifiers are available. For datasets without exposed patient identifiers, we follow the official benchmark splits and do not mix samples across the labeled pool and test set.

BrainMRI. The BrainMRI dataset is derived from BraTS2021 (Baid et al. 2021; Bakas et al. 2017; Menze et al. 2014) and uses FLAIR volumes for brain anomaly detection. We extract 2D slices from selected depth ranges of the original 3D volumes and resize them to 240×240. The benchmark contains 83 labeled pool samples and 3,715 test images for evaluating both anomaly classification and localization.

LiverCT. The LiverCT dataset is constructed from abdominal CT scans, where normal samples are taken from BTCV (Landman et al. 2015) and abnormal testing samples are obtained from LiTS (Bilic et al. 2023). The CT volumes are converted into grayscale images using an abdominal window and then cropped into 2D axial slices. The benchmark includes 166 labeled pool samples and 1,493 test slices. Since

Dataset	Modality	Task	Pool	Test	Size
BrainMRI	Brain MRI	AC+AS	83	3,715	240×240
LiverCT	Liver CT	AC+AS	166	1,493	512×512
RESC	Retinal OCT	AC+AS	115	1,805	512×1024
OCT17	Retinal OCT	AC	32	968	512×496
Chest	Chest X-ray	AC	1,490	17,194	1024×1024
HIS	Histopathology	AC	236	2,000	256×256

Table 6: **Summary of the six medical AD datasets.** The labeled pool is used for few-shot support construction, and the test split is used for evaluation.

pixel-level lesion annotations are available, LiverCT is used for both AC and AS evaluation.

RESC. RESC is a retinal OCT dataset designed for retinal edema analysis. It provides pixel-level annotations of abnormal regions, making it suitable for both image-level and pixel-level anomaly detection. The benchmark includes 115 labeled pool samples and 1,805 test images, with an original image size of 512×1024.

OCT17. OCT17 is a retinal OCT dataset mainly used for image-level anomaly classification. It contains normal OCT images and multiple abnormal retinal disease categories. The benchmark includes 32 labeled pool samples and 968 test images, with an image size of 512×496.

Chest. The Chest dataset is based on RSNA chest radiographs (Wang et al. 2017; Shih et al. 2019), where abnormal cases are associated with lung opacity or other non-normal findings. The benchmark includes 1,490 labeled pool samples and 17,194 test images. This dataset is evaluated under the image-level anomaly classification setting.

HIS. The HIS dataset is constructed from Camelyon16 (Ehteshami Bejnordi et al. 2017), which consists of hematoxylin and eosin stained histopathology whole-slide images of lymph node tissue. Following the patch-based setting, normal and abnormal image patches are extracted from the original WSIs. The benchmark includes 236 labeled pool samples and 2,000 test patches, and is used for image-level anomaly classification.

D More Implementation Details

Here we provide descriptions of more implementation details:

Zero-shot Protocol. In the zero-shot setting, no target-domain support images are used for prototype initialization, memory construction, or model adaptation. For each run, one dataset is held out as the unseen target domain, and the remaining datasets are used as source domains. The trained model is then evaluated directly on the held-out test set using the learned visual prototypes, adapters, conditioning modules, and layer-fusion weights.

Few-shot Protocol. In the few-shot setting, we construct a small target-domain support set from the labeled pool of the target dataset. For a K-shot experiment, the support set contains K normal images and K abnormal images. For datasets with pixel-level annotations, abnormal support images are paired with lesion masks. The support images are used to initialize normal/abnormal visual prototypes, while only normal support images are used to construct the normal-reference memory bank. Unless otherwise specified, we evaluate $K \in \{2, 4, 8, 16\}$ following the same support-shot protocol as prior work.

Preprocessing and Optimization Settings. All input images are converted to RGB and resized to 240×240. ReCAP is trained for 50 epochs using Adam with $\beta_1 = 0.5$, $\beta_2 = 0.999$, batch size 1. The learning rate is set to 5×10^{-4} for zero-shot training and 1×10^{-3} for few-shot training. For datasets without pixel-level annotations, the segmentation loss is omitted. The default logit scale is $\alpha = 100$, and the top-k ratio for image-level aggregation is 10%. The memory-fusion weight is set to $\lambda = 0.5$ in the few-shot setting. The bounded conditioning strength η is set to 0.05 for both zero-shot and few-shot training, unless otherwise specified.

Support Sampling and Augmentation. For each target dataset and shot setting, we use a fixed predefined support set selected from the validation split. All compared methods use identical support indices, and no repeated random sampling is performed, ensuring fair and reproducible comparisons. To reduce overfitting, we apply lightweight geometric augmentation, including small rotations, translations, and horizontal and vertical flips. When pixel-level masks are available, the same transformations are applied to the corresponding masks. Otherwise, zero masks are used only as placeholders, without segmentation supervision.

Computing Resources. All experiments are conducted on NVIDIA A100 GPUs. Inference-time measurements are evaluated on the same GPU platform using the same input resolution of 240×240.

E Additional Quantitative Results

Table 7 reports additional few-shot results under larger support settings ($K = 8, 16$). The results show that ReCAP maintains strong performance as the number of support samples increases. Under $K = 8$, ReCAP achieves the best image-level AUROC on all six datasets, with particularly strong results on HIS (90.01%), LiverCT (90.77%), and RESC (98.28%). For pixel-level segmentation, ReCAP obtains the best AUROC on BrainMRI and RESC, and remains nearly tied with the best method on LiverCT. Under $K = 16$, ReCAP again achieves the best image-level AUROC across all six datasets, reaching 100.00% on OCT17 and 99.32% on RESC. On pixel-level segmentation, ReCAP delivers the best AUROC on LiverCT and competitive results on BrainMRI and RESC, while achieving strong F_1-max and PRO scores across all three segmentation datasets. These additional results further confirm that ReCAP benefits consistently from increased target-domain support and remains competitive against both prompt-based and language-free baselines.

Robustness Across Random Seeds. Table 8 reports the mean and standard deviation over five random seeds under

Few-shot	Metric	Dataset	DRA	APRIL-GAN	MediCLIP	MVFA	MadCLIP	VisualAD	ReCAP
			N/A	✗	✗	✗	✗	✓	✓
K=8	Image-level (AUROC, F_1-max, AP)	HIS	(74.33, 72.10, 73.62)	(81.70, 76.84, 80.52)	(69.80, 71.22, 72.42)	(80.29, 76.34, 78.34)	(87.45, 80.24, 86.32)	(64.20, 66.50, 65.10)	(**90.01**, **81.21**, **90.85**)
		Chest	(82.70, 97.02, 98.34)	(73.69, 97.22, 98.36)	(72.08, 96.52, 97.22)	(75.82, 97.68, 98.23)	(83.90, 97.74, 99.10)	(62.50, 97.93, 97.00)	(**89.82**, **98.74**, **99.02**)
		OCT17	(99.13, 98.02, 99.45)	(99.75, 98.42, 99.70)	(95.69, 93.52, 96.82)	(99.54, 98.41, 99.86)	(99.14, 98.30, 99.55)	(99.80, 98.60, 99.90)	(**99.99**, **99.72**, **100.00**)
		BrainMRI	(85.94, 90.14, 94.92)	(88.41, 92.94, 97.34)	(92.29, 93.02, 97.32)	(87.09, 90.94, 97.01)	(95.17, **95.90**, 99.00)	(93.90, 93.40, 95.70)	(**95.85**, 95.84, **98.96**)
		LiverCT	(72.53, 67.24, 73.62)	(62.38, 65.84, 70.10)	(86.32, 78.42, 86.24)	(78.54, 70.67, 80.33)	(89.31, 81.24, 90.22)	(84.53, 74.00, 84.90)	(**90.77**, **86.41**, **91.90**)
		RESC	(93.06, 84.20, 91.80)	(91.36, 83.64, 90.52)	(88.82, 81.34, 88.24)	(97.57, **92.50**, 96.43)	(97.16, 89.20, 96.04)	(92.20, 82.00, 91.50)	(**98.28**, 92.42, **97.99**)
	Pixel-level (AUROC, F_1-max, PRO)	BrainMRI	(75.32, 22.42, 58.40)	(95.50, 40.20, 81.26)	(98.02, 43.50, 86.92)	(95.11, 40.69, 77.89)	(98.02, 46.10, 89.20)	(86.80, 45.30, 83.86)	(**98.12**, **53.60**, **90.15**)
		LiverCT	(81.78, 30.42, 60.82)	(97.56, 49.20, 80.64)	(98.32, 56.80, 85.52)	(94.59, 57.30, 73.27)	(**99.81**, 60.80, 88.62)	(99.10, 34.90, 82.68)	(99.80, **69.28**, **92.27**)
		RESC	(83.07, 38.24, 58.40)	(97.36, 68.20, 81.40)	(95.98, 64.80, 77.20)	(99.33, **85.23**, **96.12**)	(98.85, 72.10, 86.30)	(98.10, 72.20, 91.06)	(**99.58**, 83.38, 94.85)
K=16	Image-level (AUROC, F_1-max, AP)	HIS	(79.16, 74.52, 78.20)	(81.16, 78.90, 81.40)	(70.22, 71.82, 74.00)	(83.69, 77.55, 82.39)	(90.14, 82.00, 89.50)	(63.10, 67.20, 60.30)	(**90.28**, **82.17**, **90.92**)
		Chest	(85.01, 97.10, 98.55)	(78.62, 97.35, 98.50)	(69.74, 96.22, 96.82)	(84.40, 97.69, 98.96)	(88.15, 97.80, 99.20)	(63.80, **98.12**, 97.00)	(**90.19**, 98.07, **99.25**)
		OCT17	(99.87, 98.50, 99.85)	(99.93, 98.70, 99.90)	(96.37, 94.10, 97.20)	(98.85, 98.96, 99.48)	(99.71, 98.60, 99.70)	(99.40, 97.60, 99.80)	(**100.00**, **99.93**, **100.00**)
		BrainMRI	(82.99, 86.30, 91.50)	(94.03, 94.20, 98.10)	(91.56, 92.50, 97.00)	(94.33, 94.11, 98.74)	(95.90, 95.68, 99.20)	(97.10, 92.90, 96.40)	(**97.59**, **96.21**, **99.97**)
		LiverCT	(80.89, 72.80, 82.10)	(82.94, 76.40, 85.30)	(79.31, 76.20, 84.60)	(85.35, 75.08, 85.97)	(91.46, 83.20, 91.50)	(90.70, 82.80, 86.90)	(**92.15**, **88.82**, **92.13**)
		RESC	(94.88, 86.00, 92.60)	(95.96, 86.20, 94.00)	(86.51, 80.00, 86.80)	(98.81, 94.84, 98.65)	(99.11, 95.50, 99.00)	(98.80, 85.50, 93.90)	(**99.32**, **95.97**, **99.13**)
	Pixel-level (AUROC, F_1-max, PRO)	BrainMRI	(80.45, 25.80, 62.20)	(96.17, 42.84, 84.20)	(98.08, 44.10, 87.40)	(97.45, 40.77, 76.93)	(97.97, 47.20, 89.80)	(**99.50**, 51.70, 85.92)	(98.96, **58.40**, **90.36**)
		LiverCT	(93.00, 36.20, 68.40)	(99.64, 54.80, 86.42)	(98.95, 58.40, 87.10)	(99.35, 70.22, 74.51)	(99.74, 64.20, 90.20)	(99.40, 45.30, 84.56)	(**99.75**, **72.42**, **95.01**)
		RESC	(84.01, 40.20, 60.80)	(98.47, 73.40, 85.20)	(94.07, 66.80, 78.00)	(99.49, 88.56, **96.50**)	(99.10, 78.00, 91.20)	(95.00, 76.70, 92.67)	(**99.45**, **89.11**, 95.56)

Table 7: **Comparisons with state-of-the-art few-shot anomaly detection methods under different few-shot settings** ($K = 8, 16$)**.** Image-level results are reported as (AUROC, F_1-max, AP), and pixel-level results as (AUROC, F_1-max, PRO). ✓denotes language-free methods, while ✗ denotes methods using text prompts, and N/A denotes non-CLIP methods. All values are reported in percentage.

$K \in \{2, 4, 8, 16\}$. ReCAP achieves the best mean AC in 23 of the 24 dataset–shot combinations. For AS, ReCAP obtains the best mean AUROC in 7 of the 12 comparisons and remains competitive in the others. Its standard deviation is at most 0.45 percentage points across all settings, decreasing to at most 0.30 and 0.25 points under $K = 8$ and $K = 16$, respectively. These results demonstrate that ReCAP maintains consistent performance across random seeds rather than relying on a favorable run.

F Additional Visualizations

Additional Qualitative Localization. Figure 4 shows additional localization results on LiverCT, RESC, and BrainMRI. ReCAP produces compact anomaly responses that largely overlap with the ground-truth regions while suppressing activations on normal tissues. These examples further show that ReCAP can localize both small focal lesions and larger structural abnormalities across different medical imaging modalities.

Additional Image-level Anomaly Score Distributions. Figure 5 presents score distributions for HIS and Chest, complementing the four datasets shown in the main paper. On both datasets, normal samples generally receive lower scores, whereas abnormal samples shift toward higher values. HIS exhibits broader distributions and greater overlap, while Chest shows clearer separation with limited overlap at high scores. These results remain consistent with the image-level detection performance. More challenging datasets, such as Histopathology, Chest, and LiverCT, show partial overlap between normal and abnormal distributions. This overlap reflects the difficulty of medical anomaly detection under subtle pathological patterns, imaging noise, and complex anatomical variation. Nevertheless, abnormal samples consistently shift toward higher anomaly scores on all datasets, indicating that ReCAP learns a robust image-level scoring function across heterogeneous medical domains.

G Additional Parameter Analysis

Effects of Pre-trained Visual Backbones. Table 9 examines ReCAP with different pre-trained visual backbones on LiverCT. Among CLIP backbones, ViT-L/14@336px substantially outperforms ViT-B/16 under 2-shot supervision, improving AC from 70.56% to 90.45% and AS from 97.59% to 99.64%. More importantly, even when using the lighter ViT-B/16 backbone, ReCAP still outperforms APRIL-GAN, which uses the same backbone scale, on both AC and AS under $K = 2$ and $K = 4$. This indicates that the gains are not merely due to a larger visual encoder, but also come from the proposed re-centered anomaly prototypes and normal-reference modeling. Under 4-shot supervision, the gap between ViT-B/16 and ViT-L/14@336px becomes smaller, suggesting reduced sensitivity to backbone capacity as target-domain support increases. ReCAP also transfers effectively to DINOv2 backbones: ViT-B/14 achieves 88.99% AC and 99.42% AS with four shots, approaching the performance of ViT-L/14.

Sensitivity to the Bounded Conditioning Strength η. The parameter η controls the strength of the bounded conditioning constraint in our method. We therefore study its influence under the 4-shot setting by varying η while keeping all other hyperparameters fixed. As shown in Figure 6, moderate values of η yield better or comparable mean performance on both Liver and RESC across Image-AUROC, Pixel-AUROC, and PRO. In particular, $\eta = 0.025$ and $\eta = 0.05$ consistently achieve competitive results, suggesting that a softly bounded conditioning signal can provide useful anatomical guidance without dominating the learned representation. In contrast, larger values such as $\eta = 0.1$ and $\eta = 0.2$ tend to degrade

Setting	Method	Source	HIS	Chest	OCT17	BrainMRI		LiverCT		RESC	
			AC	AC	AC	AC	AS	AC	AS	AC	AS
K=2	DRA	CVPR22	72.84±0.86	72.36±0.94	98.04±0.21	71.64±1.12	72.31±1.38	57.42±1.46	63.38±1.72	85.57±0.68	65.84±1.55
	APRIL-GAN	CVPRw23	69.76±1.08	69.62±1.02	99.18±0.14	78.28±0.96	94.15±0.64	57.64±1.35	96.02±0.52	89.61±0.74	96.24±0.48
	MediCLIP	MICCAI24	64.72±1.18	61.44±1.30	93.52±0.56	85.02±0.82	97.28±0.36	68.62±1.08	97.18±0.44	83.79±0.86	96.16±0.62
	MVFA	CVPR24	79.68±0.72	78.52±0.84	96.61±0.42	91.82±0.56	97.34±0.32	87.12±0.64	96.82±0.48	94.58±0.42	97.02±0.50
	MadCLIP	MICCAI25	83.48±0.46	84.43±0.39	99.03±0.16	93.84±0.38	**97.86±0.28**	84.36±0.58	99.34±0.15	95.02±0.34	97.24±0.42
	VisualAD	CVPR26	53.66±1.42	53.92±1.35	92.54±0.68	83.76±0.76	95.12±0.58	84.28±0.82	99.04±0.24	91.36±0.64	97.02±0.46
	ReCAP	–	**85.76±0.32**	**84.83±0.35**	**99.36±0.08**	**94.28±0.28**	97.48±0.26	**90.36±0.41**	**99.58±0.10**	**95.38±0.30**	**98.29±0.22**
K=4	DRA	CVPR22	68.91±0.92	75.62±0.80	99.02±0.18	80.46±0.82	74.95±1.10	59.48±1.20	72.03±1.32	90.74±0.54	77.46±1.05
	APRIL-GAN	CVPRw23	76.28±0.78	77.31±0.72	99.39±0.12	89.06±0.55	94.81±0.52	52.86±1.18	96.38±0.45	94.58±0.38	98.06±0.34
	MediCLIP	MICCAI24	70.62±0.84	57.04±1.26	89.21±0.72	83.96±0.68	96.74±0.42	81.68±0.74	98.52±0.26	87.69±0.76	96.78±0.52
	MVFA	CVPR24	80.12±0.48	82.38±0.56	99.34±0.13	88.76±0.54	95.18±0.44	81.94±0.68	99.61±0.10	94.82±0.42	98.66±0.30
	MadCLIP	MICCAI25	79.96±0.44	88.06±0.31	99.32±0.14	**95.18±0.26**	**97.84±0.24**	82.86±0.60	99.23±0.16	96.54±0.28	98.84±0.26
	VisualAD	CVPR26	67.46±0.96	65.04±1.10	98.42±0.28	77.66±0.92	84.74±0.80	59.58±1.24	98.31±0.32	91.34±0.58	97.38±0.36
	ReCAP	–	**82.18±0.36**	**88.76±0.28**	**99.82±0.05**	95.06±0.24	97.50±0.22	**87.42±0.45**	**99.72±0.07**	**96.86±0.24**	**99.05±0.18**
K=8	DRA	CVPR22	74.48±0.70	82.54±0.62	99.10±0.14	85.82±0.58	75.46±0.92	72.38±0.82	81.94±0.86	92.94±0.42	83.22±0.72
	APRIL-GAN	CVPRw23	81.56±0.54	73.84±0.78	99.71±0.08	88.28±0.46	95.62±0.38	62.52±0.95	97.48±0.30	91.52±0.52	97.48±0.32
	MediCLIP	MICCAI24	69.96±0.80	71.92±0.86	95.82±0.42	92.16±0.38	97.96±0.20	86.18±0.50	98.44±0.24	88.96±0.62	96.12±0.44
	MVFA	CVPR24	80.42±0.56	75.68±0.66	99.49±0.10	87.22±0.52	95.02±0.36	78.68±0.72	94.72±0.48	97.50±0.22	99.26±0.16
	MadCLIP	MICCAI25	87.32±0.30	83.78±0.36	99.20±0.15	95.08±0.22	98.04±0.19	89.22±0.38	**99.78±0.06**	97.04±0.26	98.94±0.24
	VisualAD	CVPR26	64.38±0.92	62.34±0.98	99.76±0.07	94.02±0.34	86.62±0.70	84.66±0.56	99.02±0.18	92.38±0.48	98.24±0.28
	ReCAP	–	**89.86±0.24**	**89.64±0.22**	**99.96±0.03**	**95.76±0.18**	**98.10±0.16**	**90.62±0.30**	99.73±0.05	**98.16±0.18**	**99.52±0.12**
K=16	DRA	CVPR22	79.02±0.56	84.86±0.48	99.84±0.06	83.12±0.62	80.62±0.80	80.74±0.60	93.16±0.52	94.72±0.34	84.18±0.64
	APRIL-GAN	CVPRw23	81.28±0.48	78.48±0.62	99.91±0.04	93.92±0.28	96.30±0.30	83.08±0.52	99.58±0.08	96.08±0.28	98.56±0.22
	MediCLIP	MICCAI24	70.36±0.72	69.58±0.78	96.52±0.34	91.42±0.36	98.02±0.18	79.48±0.58	99.02±0.15	86.70±0.56	94.22±0.50
	MVFA	CVPR24	83.56±0.42	84.56±0.44	98.92±0.18	94.20±0.26	97.58±0.22	85.48±0.44	99.28±0.11	98.70±0.18	**99.42±0.09**
	MadCLIP	MICCAI25	90.02±0.20	88.04±0.26	99.68±0.08	95.82±0.20	97.90±0.20	91.34±0.32	99.70±0.05	99.04±0.14	99.02±0.16
	VisualAD	CVPR26	63.28±0.86	63.64±0.92	99.34±0.10	96.98±0.18	**99.46±0.07**	90.56±0.38	99.34±0.09	98.66±0.20	95.18±0.42
	ReCAP	–	**90.18±0.18**	**90.04±0.20**	**99.98±0.03**	**97.48±0.14**	98.88±0.12	**92.02±0.25**	**99.72±0.04**	**99.24±0.10**	99.38±0.08

Table 8: **Comparisons with state-of-the-art few-shot anomaly detection methods under different few-shot settings (K=2, 4, 8, 16)**. The AUROC (%, mean±std over five seeds) for anomaly classification (AC) and anomaly segmentation (AS) are reported. The best mean result is in bold, and the second-best mean result is underlined.

Model	Backbone	Shots	AC	AS
CLIP	ViT-B/16	2	70.56	97.59
	ViT-L/14@336px	2	**90.45**	**99.64**
	ViT-B/16	4	85.32	97.98
	ViT-L/14@336px	4	87.56	**99.77**
DINOv2	ViT-B/14	2	87.46	99.18
	ViT-L/14@336px	2	89.85	99.40
	ViT-B/14	4	88.99	99.42
	ViT-L/14@336px	4	**89.20**	99.54

Table 9: Effects of pre-trained visual backbones on LiverCT. All models use 240×240 inputs. AC and AS denote image- and pixel-level AUROC (%), respectively.

the performance, especially on localization-related metrics. These results indicate that the bounded conditioning strength should be moderate rather than overly strong.

Parameter analysis on the top-k ratio. We analyze the sensitivity of ReCAP to the top-k ratio used for image-level score aggregation under the 4-shot setting. As shown in Figure 7, performance remains stable when the ratio varies from 5% to 20% across all datasets. The best average performance is obtained at 10%, which is used as the default value in our experiments. A smaller ratio may rely on too few highly activated patches and become sensitive to local noise, whereas a larger ratio may include more normal regions and dilute anomaly evidence. These results show that ReCAP is not sensitive to this hyperparameter within a reasonable range.

Sensitivity to Fusion Weight λ. Figure 8 analyzes the effect of the fusion weight λ between the conditional prototype branch and the normal-reference memory branch under the few-shot setting. For classification-only datasets, we report image-level AUROC, while for datasets with pixel-level annotations, we report both image-level and pixel-level AUROC. Across most datasets, the best or near-best performance is achieved around $\lambda = 0.5$, indicating that relying solely on either the prototype branch ($\lambda = 0$) or the memory branch ($\lambda = 1$) is suboptimal. In particular, the segmentation datasets show consistently strong pixel-level AUROC within $\lambda \in [0.25, 0.75]$, suggesting that memory-based normal references complement the conditional prototype scores for localization. Overall, the performance curves remain relatively stable around the default value, demonstrating that ReCAP is robust to the exact fusion weight. We therefore set $\lambda = 0.5$ in all few-shot experiments.

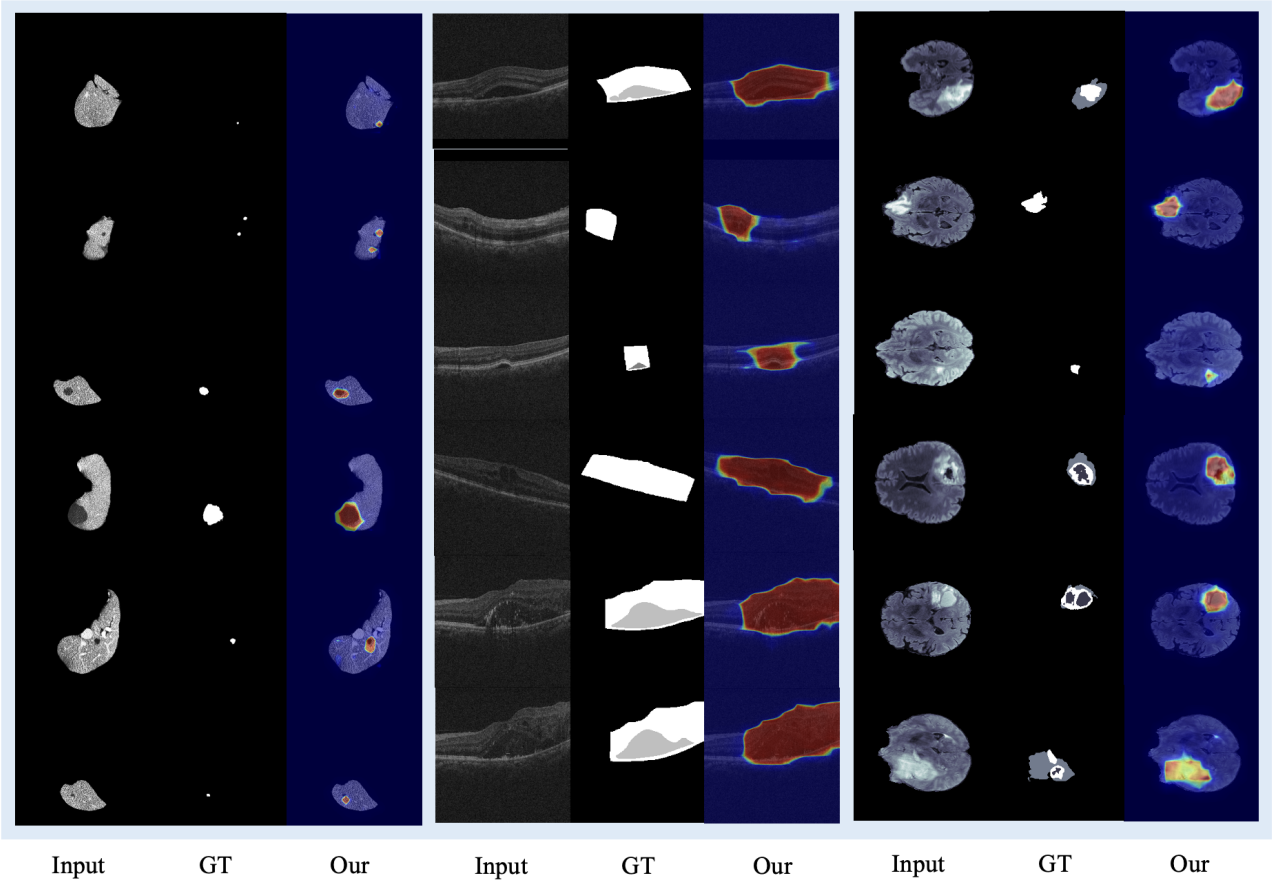

Figure 4: **Additional qualitative visualizations on different medical anomaly detection datasets.** For each case, we show the input image, ground-truth annotation (GT), and the anomaly map predicted by our method.

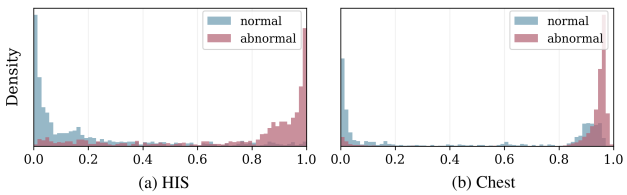

Figure 5: Additional image-level anomaly score distributions under the 16-shot setting. Results are shown for HIS and Chest, the two datasets omitted from the main-paper visualization.

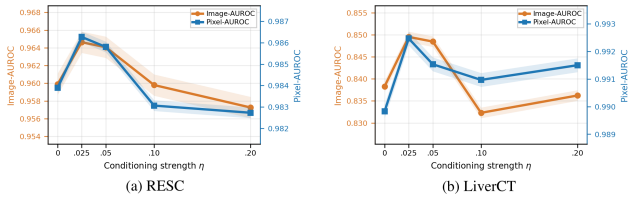

Figure 6: Sensitivity analysis of the bounded conditioning strength η on Liver and RESC under the 4-shot setting. Points and error bars denote the mean and standard deviation over three random seeds. Moderate values of η provide a better trade-off, while overly large η may over-constrain the representation and degrade performance.

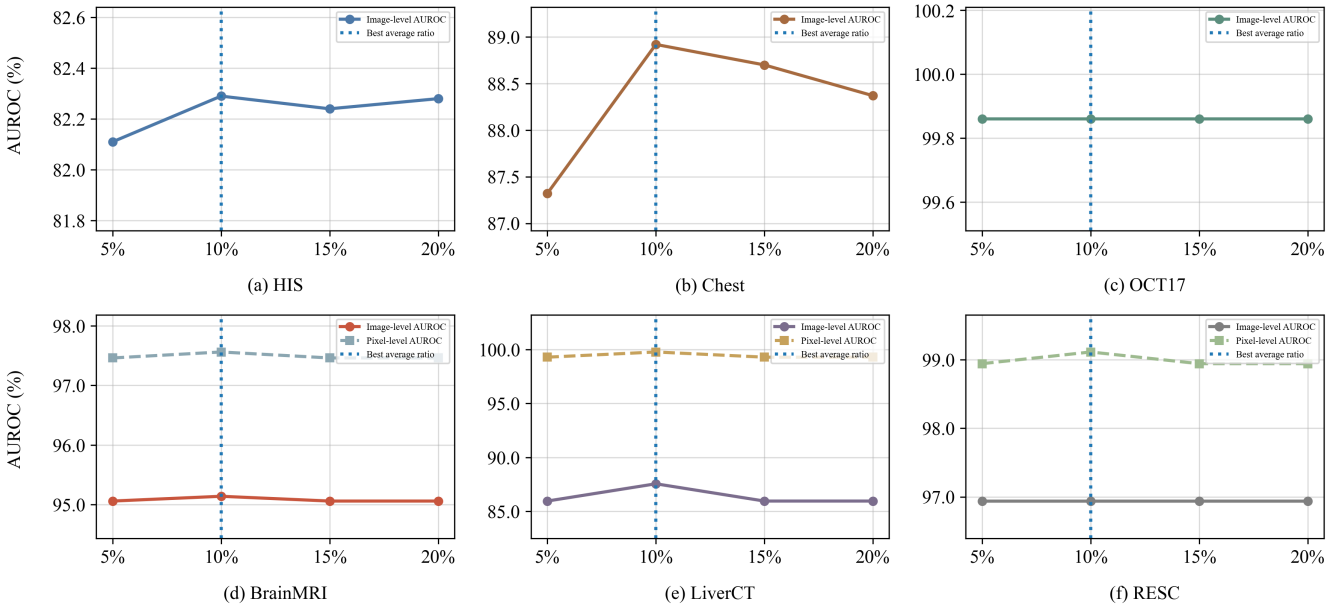

Figure 7: Sensitivity analysis of the top-k ratio under the 4-shot setting. Blue curves report image-level AUROC, and orange dashed curves report pixel-level AUROC when pixel annotations are available. The vertical dotted line marks the default ratio of 10%, which achieves the best average performance while maintaining stable results across datasets.

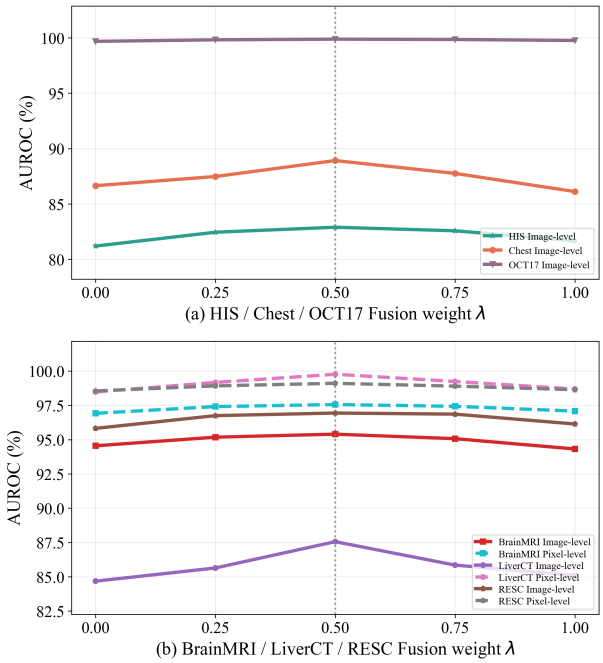

Figure 8: Sensitivity analysis of the fusion weight λ in the few-shot setting. For classification-only datasets, only image-level AUROC is reported. For datasets with pixel-level annotations, both image-level AUROC and pixel-level AUROC are shown.