ReMAP-AD: Unifying Visual Memory Reconstruction and Prompt Learning for Few-Shot Anomaly Detection

Anshuo Yin[1], Jiong Yu[1*], and Jiangqi Shi[2]

[1]School of Computer Science and Technology, Xinjiang University, Urumqi, China
591534599@qq.com, yujiong@xju.edu.cn
[2]Kashi University, Kashgar, China
3177366587@qq.com

***Abstract*—Few-shot industrial anomaly detection is typically trained with only a small K-shot set of normal images, where reliable localization is easily confounded by repetitive textures and background patterns. Recent CLIP-based prompting introduces semantic priors through text, yet its semantic responses can vary substantially with prompt templates, leading to over-activation and spatial drift. This paper presents ReMAP-AD, an evidence-calibrated vision–language framework that constrains prompt-driven semantics with explicit residual visual evidence derived from the normal support set. ReMAP-AD first stabilizes the abnormal semantic anchor by aggregating complementary abnormal prompts into prototypes. It then constructs a reconstructive visual memory over multi-scale patch features to softly reconstruct query patches from the K-shot normal gallery, and converts reconstruction discrepancies into a residual evidence map. Guided by this evidence, a residual evidence-guided semantic alignment module calibrates the raw semantic anomaly map via an evidence-adaptive gate, and the final prediction is obtained by harmonic-mean fusion to emphasize agreement between semantic and evidence cues. Experiments on MVTec AD and VisA under few-shot protocols demonstrate consistent gains in image-level detection and pixel-level localization, and robustness evaluations under template perturbations show reduced performance dispersion and smaller worst-case degradation.**

***Index Terms*—Few-shot Learning, Visual Anomaly Detection, Prompt Learning, Visual Memory Reconstruction, Prompt Robustness.**

I. Introduction

Visual anomaly detection (VAD) is a key component of industrial visual inspection, aiming to identify deviations from normal patterns in manufacturing imagery. In practice, defect samples are scarce and visually diverse, so anomaly detection is often formulated as one-class learning with only normal data. Recent benchmarks such as MVTec AD and VisA provide reproducible protocols for image-level detection and pixel-level localization, enabling systematic comparison under few-shot settings [1], [2].

With limited normal support samples, conventional few-shot anomaly detection models normality in pretrained feature spaces and scores deviations via distances or reconstruction residuals. Representative methods include PaDiM for local feature distribution modeling [3] and PatchCore for memory-based nearest-neighbor retrieval [4]. Reconstruction and distillation paradigms further amplify anomaly cues, e.g., DRAEM [5] and reverse distillation [6]. While these approaches are often stable in industrial environments, they may produce false activations and localization drift when anomalies closely resemble normal textures or when domain shifts occur.

Vision–language foundation models offer an alternative by injecting semantic concepts through text prompts. CLIP aligns images and texts via large-scale contrastive pretraining [7], and has been adopted for zero-shot and few-shot anomaly detection, e.g., WinCLIP [8] and PromptAD [9]. Stronger visual backbones can further improve this paradigm, such as DiCLIP integrating DINOv2 with CLIP [10], [11]. Nevertheless, prompt-based semantic scoring can be sensitive to prompt templates, and background textures may trigger semantically overconfident responses, which distorts anomaly maps and degrades localization consistency. In this work, we treat prompt robustness as a measurable property and report both the mean performance and its dispersion (e.g., standard deviation) across multiple prompt templates.

To improve both image-level detection and pixel-level localization, we propose ReMAP-AD, an evidence-calibrated vision–language framework for few-shot anomaly detection (Fig. 1). We follow a CLIP-based prompt learning setup under normal-only supervision [9]. ReMAP-AD stabilizes the abnormal semantic anchor by fusing generic and class-specific abnormal prompts into prototypes. It further introduces a reconstructive visual memory (RVM) to softly reconstruct multi-scale patch features from a K-shot support set and converts reconstruction discrepancies into an explicit residual evidence map. Based on this evidence, a residual evidence-guided semantic alignment module (ReSA) gates the semantic map, suppressing activations unsupported by visual cues. Finally, we harmonically fuse the calibrated semantic map and the evidence map to produce the anomaly prediction, and provide

* Corresponding author: Jiong Yu.

defect-type likelihoods via an explanation-only routing head (ReRoute) without modifying the detection pathway. Our main contributions are:

- We propose ReMAP-AD, an evidence-calibrated vision–language framework for few-shot industrial anomaly detection that constrains prompt-driven semantics with residual visual evidence, and evaluates prompt robustness via mean and dispersion across templates.
- We introduce RVM, a reconstructive visual memory that performs support-conditioned soft reconstruction and yields an interpretable residual evidence map from a K-shot normal set.
- We develop ReSA and ReRoute: ReSA calibrates semantic maps with an evidence-derived gate, while ReRoute provides defect-type likelihoods for explanation without altering the detection pathway.

II. Related Work

Few-shot industrial anomaly detection. Few-shot industrial anomaly detection models normality using only normal training data and detects anomalies by measuring deviations. PaDiM models local feature distributions [3], while PatchCore performs nearest-neighbor retrieval with a patch memory bank [4]. Reconstruction and distillation further enhance separability, e.g., DRAEM [5], STFPM [12], and reverse distillation [6]. For efficient deployment, CFLOW-AD and EfficientAD reduce latency via conditional flows and lightweight designs [13], [14].

Vision–language prompting for anomaly detection. Vision–language models extend anomaly detection toward open-category settings by using prompts as semantic interfaces. CLIP provides cross-modal alignment [7], and CoCoOp introduces conditional prompt learning for vision–language models [15]. In anomaly detection, WinCLIP [8], PromptAD [9], and AdaCLIP [16] adapt CLIP-based prompting for zero-shot or few-shot settings, and CLIP-AD and AnomalyCLIP explore staged or object-agnostic prompting [17], [18]. Incorporating stronger visual representations further improves performance, as in DiCLIP [10], [11]. Despite these advances, prompt-based anomaly detection can remain template-sensitive and be confounded by background textures, yielding unstable anomaly maps and localization drift. In contrast, our work introduces an explicit residual visual evidence pathway and uses it to calibrate and fuse semantic responses, improving reliability while yielding enhanced prompt robustness as an additional benefit.

III. Problem Formulation

Few-shot anomaly detection (FSAD) is considered for industrial visual inspection under *normal-only* supervision. For each object category, only a small K-shot support set containing normal images is provided, reflecting the practical difficulty of collecting and exhaustively annotating defects on production lines. Given a query image, the task is two-fold: (i) determine whether anomalies are present at the image level and (ii) localize anomalous regions at the pixel level. In addition to standard detection and localization accuracy, *prompt robustness* is evaluated, i.e., the stability of predictions under variations of prompt templates and phrasings. This criterion is particularly relevant for CLIP-based prompting, where subtle linguistic changes may shift semantic responses and induce unstable anomaly maps in texture-rich industrial scenes.

A. Few-shot Anomaly Detection Setup

Let $\mathcal{S} = \{x_i^{\mathrm{sup}}\}_{i=1}^{K}$ denote the K-shot *normal* support set for one category, and let x^{qry} be a query image from the same category that may be normal or anomalous. The FSAD model outputs (i) an image-level anomaly score $s_{\mathrm{img}} \in [0,1]$ and (ii) a pixel-level anomaly map $\mathbf{M} \in [0,1]^{H\times W}$. During training, only normal images are accessible; anomalous samples are not required, which aligns with the one-class nature of industrial anomaly detection.

B. CLIP-based Representations

A frozen vision–language model is adopted, consisting of a visual encoder $\Phi(\cdot)$ and a text encoder $\Psi(\cdot)$. Given an image x, the visual encoder produces a global embedding and patch-level tokens:

$$\mathbf{z} = \Phi_{\mathrm{cls}}(x) \in \mathbb{R}^{C}, \quad \mathbf{U} = \Phi_{\mathrm{pat}}(x) \in \mathbb{R}^{P\times C}, \tag{1}$$

where $P = H_g W_g$ is the number of patches on the CLIP grid. Following the multi-scale design in Fig. 1, patch features from two visual stages are further extracted as

$$\mathbf{F}^{(1)} = \Phi_1(x) \in \mathbb{R}^{P\times C_1}, \quad \mathbf{F}^{(2)} = \Phi_2(x) \in \mathbb{R}^{P\times C_2}. \tag{2}$$

Cosine similarity is used throughout:

$$\mathrm{sim}(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}^{\top}\mathbf{b}}{\|\mathbf{a}\|_2\,\|\mathbf{b}\|_2 + \epsilon}. \tag{3}$$

C. Prompt Variants and Robustness Evaluation

Prompt-based anomaly detection injects semantic concepts through textual prompts. Let $\mathcal{T} = \{\pi^{(t)}\}_{t=1}^{T}$ be a set of prompt templates with different phrasings. Each prompt is encoded into a unit text feature:

$$\mathbf{t}^{(t)} = \frac{\Psi(\pi^{(t)})}{\|\Psi(\pi^{(t)})\|_2} \in \mathbb{R}^{C}. \tag{4}$$

For the same query image, different prompt templates may yield noticeably different image-level scores $\{s_{\mathrm{img}}^{(t)}\}_{t=1}^{T}$ and pixel-wise maps $\{\mathbf{M}^{(t)}\}_{t=1}^{T}$, resulting in template-sensitive predictions. Accordingly, prompt robustness is quantified by reporting both the central tendency and the dispersion across $\mathcal{T}$. For image-level detection, robustness over prompt variants is summarized using mean and standard deviation:

$$\mu_s = \frac{1}{T}\sum_{t=1}^{T} s_{\mathrm{img}}^{(t)}, \quad \sigma_s = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(s_{\mathrm{img}}^{(t)} - \mu_s\right)^2}. \tag{5}$$

A robust method is expected to maintain competitive μ_s while exhibiting reduced sensitivity to template variations (small σ_s). For pixel-level localization, each anomaly map

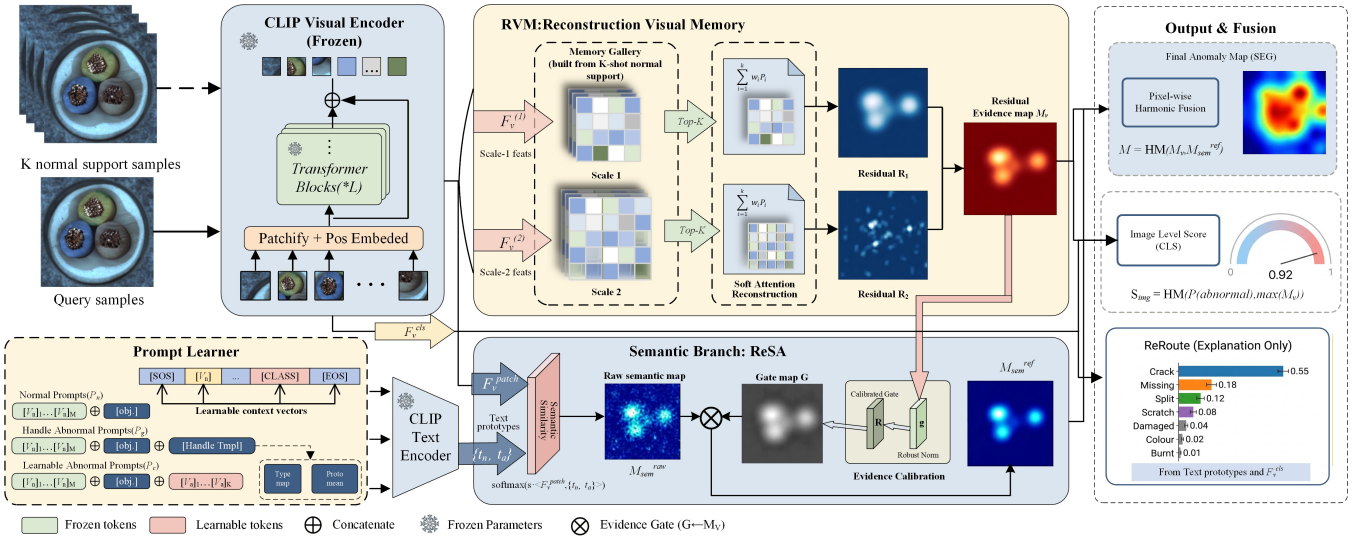

Fig. 1. Overall framework of ReMAP-AD. A frozen CLIP visual encoder extracts multi-scale patch features. RVM reconstructs query patches from a K-shot normal memory gallery and outputs a residual evidence map $\mathbf{M}_v$. The semantic branch computes a raw semantic map $\mathbf{M}_{\text{sem}}^{\text{raw}}$ via prompt-based similarity, which is calibrated by ReSA using an evidence-derived gate to obtain $\mathbf{M}_{\text{sem}}^{\text{ref}}$. The final anomaly map is produced by pixel-wise harmonic fusion, while ReRoute provides defect-type likelihoods for explanation only.

can be optionally summarized by a scalar statistic (e.g., peak response) and analyzed analogously:

$$\hat{s}_{\text{seg}}^{(t)} = \max_p \mathbf{M}^{(t)}(p). \tag{6}$$

Based on the above formulation, Sec. IV presents ReMAP-AD, which improves reliability by stabilizing semantic anchors and calibrating prompt-based semantic responses with residual visual evidence, while prompt robustness is treated as an additional desirable property.

IV. Methodology

A. Overview

Fig. 1 summarizes the ReMAP-AD pipeline. Given a K-shot normal support set $\mathcal{S} = \{x_i^{\text{sup}}\}_{i=1}^{K}$ and a query image x^{qry}, a frozen CLIP visual encoder extracts a global feature and patch-level tokens. The framework comprises: (i) a prompt-learning semantic branch that produces a raw semantic anomaly map, (ii) a reconstructive visual memory (RVM) that yields residual visual evidence $\mathbf{M}_v$ via support-conditioned reconstruction, (iii) a residual evidence-guided semantic alignment module (ReSA) that calibrates semantic responses using $\mathbf{M}_v$, (iv) harmonic fusion for pixel-level localization (and image-level detection), and (v) an explanation-only defect-type routing head (ReRoute). The design principle is to expose an explicit, support-derived visual evidence pathway and to use it as a constraint on prompt-driven semantics, thereby reducing spurious activations and improving localization reliability.

B. Prompt Prototypes and Raw Semantic Map

A CLIP-based prompt learning setup is adopted under normal-only supervision. To reduce sensitivity to prompt templates, multiple prompts are aggregated into stable *prototypes*, which suppress idiosyncratic effects of individual phrasings and provide more consistent semantic anchors. Let $\mathcal{P}$ be a prompt set; the prototype operator is defined as

$$\mathbf{t}(\mathcal{P}) = \text{norm}\left(\frac{1}{|\mathcal{P}|}\sum_{\pi\in\mathcal{P}}\Psi(\pi)\right), \tag{7}$$

where $\Psi(\cdot)$ is the frozen CLIP text encoder and $\text{norm}(\mathbf{v}) = \mathbf{v}/(\|\mathbf{v}\|_2 + \epsilon)$.

For each category, a normal prompt set $\mathcal{P}_n$, a generic abnormal set $\mathcal{P}_g$, and a class-specific abnormal set $\mathcal{P}_c$ (e.g., defect phrases) are constructed, with prototypes $\mathbf{t}_n$, $\mathbf{t}_g$, and $\mathbf{t}_c$, respectively. The abnormal anchor is stabilized by fusing generic and class-specific prototypes with a fixed weight:

$$\mathbf{t}_a = \text{norm}\big(\lambda\,\mathbf{t}_g + (1-\lambda)\,\mathbf{t}_c\big), \quad \lambda \in [0,1]. \tag{8}$$

This fixed fusion reduces degrees of freedom and mitigates overfitting to a particular prompt distribution, while retaining category-relevant defect semantics.

Raw semantic anomaly map. Let $\mathbf{U} = \Phi_{\text{pat}}(x^{\text{qry}}) \in \mathbb{R}^{P\times C}$ be the patch tokens. For patch p, two-class logits are computed using the CLIP logit scale κ:

$$\ell_p^n = \kappa\,\text{sim}(\mathbf{u}_p, \mathbf{t}_n), \quad \ell_p^a = \kappa\,\text{sim}(\mathbf{u}_p, \mathbf{t}_a). \tag{9}$$

The raw semantic anomaly probability is given by a two-class softmax:

$$M_{\text{sem}}^{\text{raw}}(p) = \frac{\exp(\ell_p^a)}{\exp(\ell_p^n) + \exp(\ell_p^a)}, \quad \mathbf{M}_{\text{sem}}^{\text{raw}} \in [0,1]^{H_g\times W_g}. \tag{10}$$

Although $\mathbf{M}_{\text{sem}}^{\text{raw}}$ provides semantic cues, it may become brittle in the presence of repetitive textures or background patterns, and it can drift when prompt wording changes. The following modules therefore introduce an explicit residual evidence map and enforce evidence-consistent semantic responses.

C. RVM: Reconstructive Visual Memory and Residual Evidence

RVM constructs a non-parametric normal memory from the K-shot support set and measures how well each query patch can be explained by normal patches. Two scales of patch features are extracted as $\mathbf{F}^{(\ell)} = \Phi_\ell(x) \in \mathbb{R}^{P\times C_\ell}$ for $\ell \in \{1,2\}$. All support patches are flattened into two galleries $\mathcal{G}^{(\ell)} = \{\mathbf{g}_m^{(\ell)}\}_{m=1}^{M}$ with $M = K \cdot P$.

Top-k retrieval and soft reconstruction. For each query patch feature $\mathbf{f}_p^{(\ell)}$, its top-k neighbors $\mathcal{N}_k(p)$ are retrieved by cosine similarity:

$$s_{p,m}^{(\ell)} = \mathrm{sim}\Big(\mathrm{norm}(\mathbf{f}_p^{(\ell)}), \mathrm{norm}(\mathbf{g}_m^{(\ell)})\Big), \qquad m = 1,\dots,M. \tag{11}$$

A normal reference patch is then reconstructed by temperature-soft weighting:

$$\hat{\mathbf{f}}_p^{(\ell)} = \sum_{m\in\mathcal{N}_k(p)} \underbrace{\frac{\exp(s_{p,m}^{(\ell)}/\tau)}{\sum_{j\in\mathcal{N}_k(p)}\exp(s_{p,j}^{(\ell)}/\tau)}}_{w_{p,m}^{(\ell)}} \mathrm{norm}\big(\mathbf{g}_m^{(\ell)}\big), \tag{12}$$

where $\tau > 0$ controls the sharpness of reconstruction weights. This reconstruction is instantiated on-the-fly from the support set at inference time, maintaining few-shot feasibility without introducing additional learnable parameters.

Residual evidence. Patch-wise residual evidence is defined as the reconstruction discrepancy averaged across scales:

$$r_p = \frac{1}{2}\sum_{\ell\in\{1,2\}} \Big\|\mathrm{norm}(\mathbf{f}_p^{(\ell)}) - \mathrm{norm}(\hat{\mathbf{f}}_p^{(\ell)})\Big\|_2^2, \qquad p = 1,\dots,P. \tag{13}$$

Large r_p indicates that the query patch is poorly explained by the normal memory, providing strong visual evidence for anomalies.

Robust normalization. To obtain a bounded evidence map, $\{r_p\}_{p=1}^{P}$ is normalized using per-image quantiles. Let $q^- = Q_\eta(\{r_p\})$ and $q^+ = Q_{1-\eta}(\{r_p\})$ with a small η (e.g., $\eta = 0.05$). Then

$$M_v(p) = \mathrm{clip}\left(\frac{r_p - q^-}{q^+ - q^- + \epsilon}, 0, 1\right), \quad \mathbf{M}_v \in [0,1]^{H_g\times W_g}. \tag{14}$$

D. ReSA: Evidence-guided Semantic Alignment

ReSA promotes *evidence-consistent* semantics by attenuating prompt-induced responses in regions with weak residual evidence. Rather than directly trusting the raw semantic map, residual evidence is first transformed into a smooth gate through a saturating mapping:

$$G(p) = \left(\frac{M_v(p)}{M_v(p) + c}\right)^{\alpha}, \tag{15}$$

where $c > 0$ controls the saturation level and $\alpha > 0$ adjusts the gate sharpness. The calibrated semantic map is then obtained via evidence-adaptive rescaling:

$$M_{\mathrm{sem}}^{\mathrm{ref}}(p) = M_{\mathrm{sem}}^{\mathrm{raw}}(p)\Big(\beta + (1-\beta)G(p)\Big), \tag{16}$$

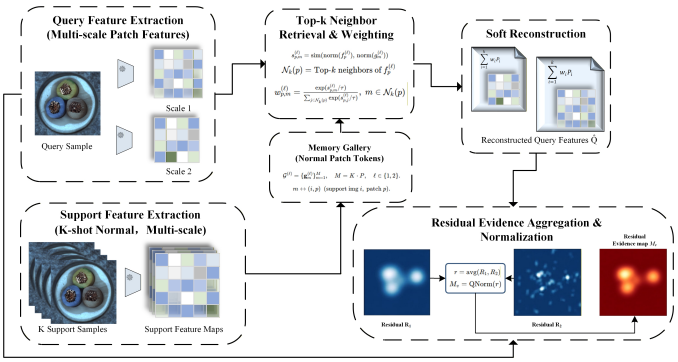

Fig. 2. RVM module. Multi-scale query patch features are reconstructed from a normal memory gallery built from the K-shot support set via top-k retrieval and soft weighting. The reconstruction discrepancy yields a residual evidence map $\mathbf{M}_v$ on the CLIP patch grid, which serves as a visual anchor for semantic calibration and fusion.

with $\beta \in [0,1]$ providing a conservative lower bound to prevent excessive suppression. This calibration suppresses semantically confident activations that are not supported by residual visual evidence, improving robustness to both background textures and prompt variations.

E. Fusion, Image-level Score, and Normal-only Training

Pixel-level fusion. The calibrated semantic map and residual evidence map are combined using harmonic mean:

$$M(p) = \frac{2\,M_v(p)\,M_{\mathrm{sem}}^{\mathrm{ref}}(p)}{M_v(p) + M_{\mathrm{sem}}^{\mathrm{ref}}(p) + \epsilon}. \tag{17}$$

Compared with arithmetic averaging, harmonic fusion emphasizes *agreement* between two sources and down-weights predictions that are confident in only one branch. The fused map is upsampled (bilinear interpolation) to the input resolution to obtain the final pixel-level anomaly map.

Image-level detection and ReRoute. For image-level detection, an abnormal probability is computed from the CLS feature $\mathbf{z}$ using the same normal/abnormal prototypes, and it can be optionally fused with the peak evidence $\max_p M_v(p)$ in the same scoring space. ReRoute provides defect-type likelihoods by grouping abnormal prompts into defect types and applying a softmax over the corresponding type prototypes; this head is used for attribution only and does not modify the anomaly map.

Normal-only training. The CLIP encoders are frozen and only the prompt learner is optimized. Training encourages normal images to align with $\mathbf{t}_n$ while remaining conservative in low-evidence regions:

$$\mathcal{L} = -\log\frac{\exp\big(\kappa\,\mathrm{sim}(\mathbf{z},\mathbf{t}_n)\big)}{\exp\big(\kappa\,\mathrm{sim}(\mathbf{z},\mathbf{t}_n)\big) + \exp\big(\kappa\,\mathrm{sim}(\mathbf{z},\mathbf{t}_a)\big)} + \lambda_s\cdot\frac{1}{P}\sum_{p=1}^{P}\big(1 - M_v(p)\big)\Big[-\log\big(1 - M_{\mathrm{sem}}^{\mathrm{raw}}(p) + \epsilon\big)\Big], \tag{18}$$

where $\lambda_s \geq 0$ controls the regularization strength. At inference, RVM is instantiated from the K-shot support set and no anomalous samples are required.

TABLE I
IMAGE-LEVEL AUROC (MAX-F1) (%) ON MVTEC AD AND VISA; BEST/SECOND-BEST ARE **BOLD**/UNDERLINED (AUROC, MAX-F1).

Type	Method	MVTec AD			VisA		
		1-shot	2-shot	4-shot	1-shot	2-shot	4-shot
Non-CLIP	SPADE [19]	81.00 (90.2)	82.90 (90.4)	84.80 (91.1)	79.50 (80.3)	80.70 (81.1)	81.70 (82.3)
	PaDiM [3]	76.60 (87.1)	78.90 (87.2)	80.40 (88.3)	62.80 (77.2)	67.40 (78.1)	72.80 (79.3)
	PatchCore [4]	83.40 (91.0)	86.30 (91.5)	88.80 (92.6)	79.90 (79.6)	81.60 (80.1)	85.30 (84.3)
CLIP-based	WinCLIP+ [8]	93.10 (93.7)	94.40 (93.9)	95.20 (94.7)	83.80 (83.1)	84.60 (83.5)	87.30 (84.3)
	PromptAD [9]	93.15 (95.9)	93.73 (96.0)	95.71 (96.2)	85.98 (86.8)	88.68 (87.1)	88.80 (**88.3**)
CLIP-based	ReMAP-AD (ours)	**95.27** (**96.1**)	**95.62** (**96.6**)	**96.83** (**96.7**)	**87.26** (**87.2**)	**89.15** (**87.4**)	**90.01** (87.5)

TABLE II
PIXEL-LEVEL AUROC (MAX-F1) (%) ON MVTEC AD AND VISA; BEST/SECOND-BEST ARE **BOLD**/UNDERLINED (AUROC, MAX-F1).

Type	Method	MVTec AD			VisA		
		1-shot	2-shot	4-shot	1-shot	2-shot	4-shot
Non-CLIP	SPADE [19]	91.20 (44.5)	92.00 (45.3)	92.70 (46.2)	95.60 (36.3)	96.20 (37.2)	96.60 (41.0)
	PaDiM [3]	89.30 (39.1)	91.30 (41.1)	92.60 (42.3)	89.90 (32.1)	92.00 (33.9)	93.20 (36.7)
	PatchCore [4]	92.00 (53.1)	93.30 (54.7)	94.30 (56.2)	95.40 (38.2)	96.10 (41.3)	96.80 (46.4)
CLIP-based	WinCLIP+ [8]	95.20 (55.9)	96.03 (57.2)	96.20 (60.1)	96.40 (42.1)	96.80 (45.2)	97.20 (47.5)
	PromptAD [9]	95.73 (59.8)	96.26 (**60.2**)	96.50 (**61.4**)	95.97 (47.1)	96.48 (49.0)	96.85 (49.3)
CLIP-based	ReMAP-AD (ours)	**96.65** (**60.1**)	**96.44** (**60.2**)	**97.03** (61.3)	**96.71** (**49.2**)	**97.35** (**49.4**)	**97.70** (**49.9**)

V. EXPERIMENTS

A. Experimental Setup

Datasets and few-shot protocol. Experiments are conducted on MVTec AD [1] and VisA [2], which provide category-wise splits with image-level anomaly labels and pixel-level masks. For each category, a K-shot support set $\mathcal{S}$ is formed by uniformly sampling $K \in \{1, 2, 4\}$ normal training images, and each test image is evaluated conditioned on $\mathcal{S}$. Results are averaged over five independent support samplings per category unless otherwise specified.

Metrics and evaluation practice. Performance is measured by AUROC at both image and pixel levels, i.e., image-level AUROC for anomaly detection and pixel-level AUROC for anomaly localization. All methods are evaluated under the same few-shot protocol and identical preprocessing/evaluation scripts for controlled comparison. Representative baselines include SPADE [19], PaDiM [3], PatchCore [4], WinCLIP [8], and PromptAD [9].

Implementation details and ReMAP-AD configuration. We use OpenCLIP ViT-B/16 pretrained on LAION-400M and follow the standard CLIP preprocessing. Images are resized and center-cropped to 240×240. For prompt learning, we optimize learnable context tokens for the normal/abnormal branches and learnable anomaly suffix tokens (E_n=4, E_a=1, L=4), and include a small handcrafted abnormal suffix pool (M=4 on MVTec AD and M=2 on VisA). ReMAP-AD introduces a residual-evidence pathway where RVM reconstructs multi-scale query patches from the K-shot support memory to obtain $\mathbf{M}_v$, and ReSA calibrates the semantic map via an evidence-derived gate before fusion. All hyperparameters are fixed across datasets and shot numbers. Prompt robustness is evaluated by perturbing only the handcrafted abnormal templates at test time while keeping learned parameters unchanged, and we report both i_{text} and i_{fuse}.

B. Comparison with State-of-the-art Methods

We compare ReMAP-AD with representative few-shot industrial anomaly detection methods from both non-vision–language and vision–language paradigms, including distribution/memory-based approaches (SPADE [19], PaDiM [3], PatchCore [4]) and CLIP-based prompting methods (WinCLIP [8], PromptAD [9]). All methods in Tables I–II are evaluated under the unified few-shot protocol and preprocessing in Sec. V-A, with identical support sampling and evaluation scripts.

Table I reports image-level detection performance. ReMAP-AD achieves the best AUROC across all shot numbers on both datasets. The gain is most evident in the low-shot regime: on MVTec AD (1-shot), AUROC improves from 93.15 to 95.27 compared with PromptAD; on VisA (1-shot), AUROC increases from 85.98 to 87.26. As K increases, ReMAP-AD maintains its advantage, suggesting that evidence-calibrated fusion remains effective without category-specific tuning.

At the pixel level (Table II), ReMAP-AD consistently improves localization over both non-CLIP and CLIP-based baselines. Compared with PromptAD, ReMAP-AD yields higher pixel-level AUROC under all shots (e.g., $95.73 \rightarrow 96.65$ on MVTec AD and $95.97 \rightarrow 96.71$ on VisA for 1-shot). Reconstruction/memory baselines such as PatchCore are strong in localization but lack semantic interfaces, whereas CLIP-

TABLE III

PROMPT-TEMPLATE ROBUSTNESS AT IMAGE LEVEL (AUROC %). WE REPORT DEFAULT, MEAN±STD, WORST, AND DROP (DEFAULT−WORST) OVER MULTIPLE TEMPLATE VARIANTS. HIGHER IS BETTER FOR AUROC AND LOWER IS BETTER FOR DROP.

Dataset	Method	Fused image score (i_{fuse})				Text-only image score (i_{text})			
		Default	Mean±Std	Worst	Drop↓	Default	Mean±Std	Worst	Drop↓
MVTec (1-shot)	PromptAD [9]	93.15	92.86±1.14	91.72	1.43	87.21	86.42±1.21	84.69	2.52
	ReMAP-AD (ours)	**95.27**	**95.03±0.31**	**94.69**	**0.44**	**87.43**	**86.98±0.34**	**85.46**	**1.97**
VisA (1-shot)	PromptAD [9]	85.98	86.13±2.21	83.92	2.06	80.31	78.36±1.50	76.00	4.31
	ReMAP-AD (ours)	**87.26**	**87.01±0.59**	**86.59**	**0.67**	**81.07**	**79.91±0.96**	**78.93**	**2.14**

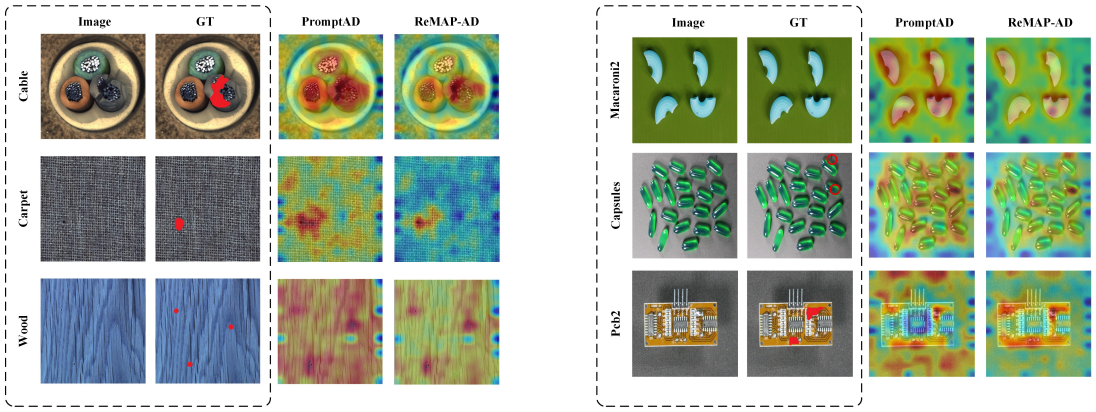

Fig. 3. Qualitative comparison of anomaly localization on MVTec AD and VisA. Each row shows the input image, ground-truth mask, and predicted anomaly maps from PromptAD [9] and ReMAP-AD. ReMAP-AD calibrates the semantic response using residual visual evidence, which reduces background activations and improves spatial alignment to annotated defects.

based methods provide semantics but may over-activate on textures. ReMAP-AD bridges these strengths by exposing support-conditioned residual evidence (RVM) and using it to suppress unsupported semantic responses (ReSA), leading to more reliable anomaly maps under limited normal support.

Qualitative results in Fig. 3 further corroborate the quantitative trends, showing reduced background activations and improved spatial alignment to defect regions.

C. Prompt templates and template perturbations

a) Class name canonicalization: Industrial categories may contain non-natural tokens or dataset-specific naming artifacts. To reduce prompt variance, each class label c is mapped to a canonical name $\tilde{c} = m(c)$ (e.g., `pcb1--pcb4`↦"printed circuit board" and `chewinggum`↦"chewing gum"). All prompts use $\tilde{c}$.

b) Handcrafted abnormal prompt pool: A handcrafted abnormal prompt pool is the union of a generic template set and an object-customized template set. Let $\mathcal{G}$ denote generic anomaly templates and $\mathcal{S}_{\tilde{c}}$ denote class-specific templates for class $\tilde{c}$. The pool is defined as

$$\mathcal{A}_{\tilde{c}} = \{\, g(\tilde{c}) \mid g \in \mathcal{G} \,\} \cup \{\, s(\tilde{c}) \mid s \in \mathcal{S}_{\tilde{c}} \,\}. \tag{19}$$

$\mathcal{G}$ provides broad anomaly descriptors, while $\mathcal{S}_{\tilde{c}}$ adds category-specific defect phrases.

Templates are assigned to defect types using keyword-matching rules; otherwise they are mapped to GENERIC. This taxonomy is used for attribution only and does not affect anomaly scoring.

c) Template perturbations and metrics: Robustness is evaluated by replacing only the handcrafted abnormal template pool at test time while keeping learned parameters fixed. We consider DEFAULT, GENERIC_ONLY, CUSTOM_ONLY, REWRITE_SYN, and SUBSAMPLE_K (multiple seeds). For each score head $s \in \{\texttt{i\_text}, \texttt{i\_fuse}\}$, we report Default, Mean±Std, Worst, and Drop $\Delta s = s_{\text{def}} - \min_j s_j$ over perturbations.

Table III summarizes robustness under template perturbations. ReMAP-AD consistently reduces dispersion and worst-case degradation, with more pronounced gains on the fused score i_{fuse} than on the text-only score i_{text}. These results support that residual evidence provides an effective constraint that stabilizes prompt-driven predictions.

D. Ablation Study

We perform component-wise ablations under the 1-shot protocol to quantify the effect of each design. All variants follow Sec. V-A and are averaged over the same repeated support samplings. Starting from the semantic-branch baseline, we incrementally add prototype aggregation (Proto), reconstructive visual memory (RVM), harmonic-mean fusion (HM), and evidence-guided semantic alignment (ReSA).

Table IV reports image-/pixel-level AUROC (Img/Pix) on MVTec AD and VisA. Proto consistently improves

TABLE IV
COMPONENT ABLATIONS UNDER THE 1-SHOT PROTOCOL (AUROC, IMG/PIX).

Variant	Proto	RVM	ReSA	Fusion	MVTec	VisA
Semantic branch only	–	–	–	–	93.15/95.73	85.98/95.97
+ Proto	✓	–	–	–	93.45/95.85	86.30/96.05
+ Proto + RVM	✓	✓	–	–	94.20/96.30	86.70/96.40
+ Proto + RVM + HM	✓	✓	–	HM	94.90/96.55	87.05/96.55
Full (ours)	✓	✓	✓	HM	95.27/96.65	87.26/96.71

the baseline, indicatingz a more stable abnormal semantic anchor. Adding RVM yields further gains, showing that support-conditioned residual evidence provides complementary spatial cues. HM improves performance by favoring pixel-wise agreement between semantic and evidence signals (MVTec: $94.20/96.30 \rightarrow 94.90/96.55$; VisA: $86.70/96.40 \rightarrow 87.05/96.55$). Finally, ReSA achieves the best results (MVTec: $94.90/96.55 \rightarrow 95.27/96.65$; VisA: $87.05/96.55 \rightarrow 87.26/96.71$), consistent with suppressing semantic activations unsupported by residual evidence. Overall, these ablations support that residual evidence and evidence-calibrated fusion improve reliability under limited normal support.

VI. CONCLUSION

This paper proposes ReMAP-AD, an evidence-calibrated vision–language framework for few-shot anomaly detection under normal-only supervision. It stabilizes abnormal semantics via prototype aggregation, derives residual evidence from multi-scale reconstruction over a K-shot normal memory, and calibrates semantic responses with evidence-guided gating. The final prediction is obtained by harmonic fusion of semantic and evidence maps, while an auxiliary routing head provides defect-type likelihoods. Experiments on MVTec AD and VisA show consistent improvements in detection, localization, and prompt robustness under template perturbations.

ACKNOWLEDGMENT

This work was supported by the Academy Member (Academician) Major Science and Technology Innovation Project under the "Two Zones" Science and Technology Development Program Grant 2024LQ02001; and the National Natural Science Foundation of China Project under Grant 62262064, 62466057; and the Key R&D projects of Xinjiang Uygur Autonomous Region, grant number 2022295358; Xinjiang University doctor postgraduate innovation project (Grant No.XJDX2025YJS084, XJDX2025YJS086).

REFERENCES

[1] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9592–9600.

[2] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 392–408.

[3] T. Defard, A. Setio, M. Loog, and J. C. van Gemert, "Padim: A patch distribution modeling framework for anomaly detection and localization," in *2020 25th International Conference on Pattern Recognition Workshops (ICPRW)*, 2020, pp. 475–489.

[4] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 318–14 328.

[5] V. Zavrtanik, M. Kristan, and D. Skočaj, "Draem: A discriminatively trained reconstruction embedding for surface anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 8330–8339.

[6] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9737–9746.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 8748–8763.

[8] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "Winclip: Zero-/few-shot anomaly classification and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 606–19 616.

[9] X. Li, Z. Zhang, X. Tan, C. Chen, Y. Qu, Y. Xie, and L. Ma, "Promptad: Learning prompts with only normal samples for few-shot anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[10] J. Cai, G. Wang, X. Wang, Y. Chen, X. Wen, T. Zhao, L. Ye, and C. Yang, "Diclip: Integrating dinov2 and clip for zero-shot and few-shot anomaly detection with versatile combination prompts," in *International Joint Conference on Neural Networks (IJCNN)*, 2025.

[11] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[12] G. Wang, S. Han, E. Ding, and D. Huang, "Student-teacher feature pyramid matching for anomaly detection," in *British Machine Vision Conference (BMVC)*, 2021.

[13] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[14] K. Batzner, L. Heckler, and R. König, "Efficientad: Accurate visual anomaly detection at millisecond-level latencies," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[15] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 816–16 825.

[16] Y. Cao, J. Zhang, L. Frittoli, Y. Cheng, W. Shen, and G. Boracchi, "Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 55–72.

[17] X. Chen, J. Zhang, G. Tian, H. He, W. Zhang, Y. Wang, C. Wang, Y. Wu, Y. Liu *et al.*, "Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection," *arXiv preprint arXiv:2311.00453*, 2023.

[18] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection," in *International Conference on Learning Representations (ICLR)*, 2024.

[19] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," *arXiv preprint arXiv:2005.02357*, 2020.

[20] Z. Fang, X. Wang, H. Li, J. Liu, Q. Hu, and J. Xiao, "Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17 435–17 444.

[21] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.