ALIGNING DEVIATIONS, NOT STATES, IN CLIP-BASED FEW-SHOT ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot Generalist Anomaly Detection (GAD) detects and localizes anomalies from novel categories using only a few normal references without target-class retraining. Existing CLIP-based methods typically align query features directly with normal and anomalous textual semantics, while using normal references separately for visual matching. However, anomalies are inherently deviations from an expected normal state rather than absolute visual states, making state-based alignment sensitive to normal variations across regions and categories. We therefore exploit normal references in few-shot GAD to shift CLIP alignment from states to deviations. Specifically, we calibrate the query's semantic response with that of a matched normal reference, measuring how the query deviates from its matched normal state and relating this visual deviation to the semantic shift from normal to anomalous text. This naturally yields Residual-Residual Alignment. Based on this insight, we propose Res^{2}CLIP, which combines visual deviation magnitude with its semantic alignment for training-free anomaly scoring and further performs progressive spatial calibration followed by semantic refinement when auxiliary supervision is available. Experiments on multiple benchmarks demonstrate the effectiveness of Res^{2}CLIP in both training-free and fine-tune settings.

1 INTRODUCTION

Anomaly Detection (AD) is widely used in real-world applications such as industrial inspection. However, collecting sufficient training data and retraining a dedicated model for every new product or scenario is often impractical. Few-shot Generalist Anomaly Detection (GAD) addresses this problem by detecting and localizing anomalies from unseen categories using only a few normal reference images, without target-class retraining. Pre-trained vision-language models such as CLIP (Radford et al., 2021) provide transferable visual and semantic representations and have therefore been widely adopted for this setting. Recent CLIP-based methods (Jeong et al., 2023; Zhu & Pang, 2024) exploit both textual anomaly semantics and normal visual references, but typically use them separately: textual semantics provide anomaly-related responses, while normal references mainly serve as visual exemplars for feature matching.

With these two sources used separately, semantic alignment still evaluates the query through its absolute visual state. The resulting response reflects not only anomaly-related changes, but also variations in normal visual content. For localization, foregrounds, backgrounds, and different object parts correspond to distinct normal states and may therefore respond differently to the same textual semantics. Such region-dependent responses may be mistaken for anomaly evidence and lead to false positives in normal regions. For optimization, visual states also differ substantially across object categories. Directly adapting these representations on seen categories therefore couples anomaly-related learning with category-specific appearance variations. In both cases, anomaly-related changes are mixed with the normal state from which they arise.

Few-shot GAD already provides a natural baseline for measuring such changes: normal references describe the expected normal state of the query. We therefore use the semantic response of a matched normal reference to calibrate that of the query. Instead of evaluating whether the query state itself matches anomalous semantics, we measure how its semantic response changes from the corresponding normal state. This calibration naturally yields an alignment between two deviations: the visual deviation from the normal reference and the textual deviation from normal to anomalous semantics,

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

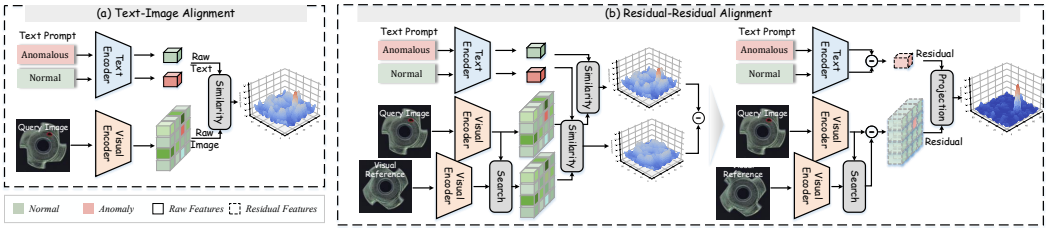


Figure 1: Motivation of normal-reference calibration for CLIP alignment. (a) Conventional Text-Image Alignment is affected by normal variations, which causes false positives. (b) Calibration with matched normal references suppresses normal-state variations and yields Residual-Residual Alignment between visual and textual deviations. All anomaly maps are training-free responses for the illustrated sample.

which we term **Residual-Residual Alignment**. It suppresses much of the region-dependent normal response and produces more focused anomaly responses with fewer background false positives, as shown in Fig. 1(b). This leads to the central motivation of our method: *align deviations, not states*.

This calibration also provides a more suitable representation for optimization. Rather than directly adapting category-dependent visual states, the model learns from deviations measured relative to the corresponding normal state. These reference-relative deviations reduce the influence of inter-category appearance differences and allow supervision from seen categories to focus more directly on anomaly-related changes. Thus, the normal reference serves not only as a baseline for local semantic responses, but also as a reference point for learning transferable anomaly deviations.

Based on Residual-Residual Alignment, we propose **Res^{2}CLIP** for few-shot GAD. Res^{2}CLIP constructs visual residuals from matched normal references and relates them to the normal-to-anomaly text residual. For anomaly scoring, residual magnitude captures visual deviations beyond the selected textual semantics, while semantic projection measures their consistency with the anomaly-related textual direction. This formulation directly supports training-free inference. When auxiliary seen-category supervision is available, we further optimize only the local visual residuals while keeping the CLIP encoders and text residual fixed. The optimization consists of two stages: Stage I uses spatial context to suppress residual responses caused by normal variations while preserving anomaly-related information, and Stage II further refines the semantic relation between visual and text residuals by encouraging anomalous residuals toward the anomaly-related textual direction while suppressing such alignment for normal residuals. Experiments on multiple anomaly detection benchmarks demonstrate the effectiveness of Res^{2}CLIP in both training-free and fine-tune settings.

2 Rethinking CLIP Alignment for Few-Shot Anomaly Detection

2.1 Revisiting Text-Image Alignment

CLIP-based anomaly detection typically identifies anomalies through Text-Image (**`T-I`**) Alignment, where a query visual feature f_q is compared with generic normal and anomalous textual semantics, i.e., $f_q \leftrightarrow (F_t^n, F_t^a)$. Let F_t^n and F_t^a denote the text features encoded from the normal and anomalous prompts, respectively. The corresponding **`T-I`** anomaly response is defined as

$$\mathbf{S}_{\mathrm{T-I}}(f_q) = \hat{f}_q^\top \hat{F}_t^a - \hat{f}_q^\top \hat{F}_t^n, \tag{1}$$

where $\hat{\cdot}$ denotes ℓ_2 normalization.

However, directly aligning the query visual feature with normal and anomalous textual semantics $f_q \leftrightarrow (F_t^n, F_t^a)$ introduces two issues in Few-Shot AD.
(1) **Cross-Granularity Mismatch**: normality is region-dependent, as foreground, background, and different object parts correspond to distinct normal visual states. Thus, even two normal local features $f_q^{(i)}$ and $f_q^{(j)}$ may respond differently to the same textual semantics, $\mathbf{S}_{\mathrm{T-I}}(f_q^{(i)}) \neq \mathbf{S}_{\mathrm{T-I}}(f_q^{(j)})$. Using shared textual semantics for these heterogeneous regions may therefore mix inherent normal variations with anomaly-induced semantic changes, hindering fine-grained localization.
(2) **Category-Dependent Optimization**: when CLIP is further adapted on training categories, optimization is performed on category-dependent visual representations $f_q^{(c)}$. These representations

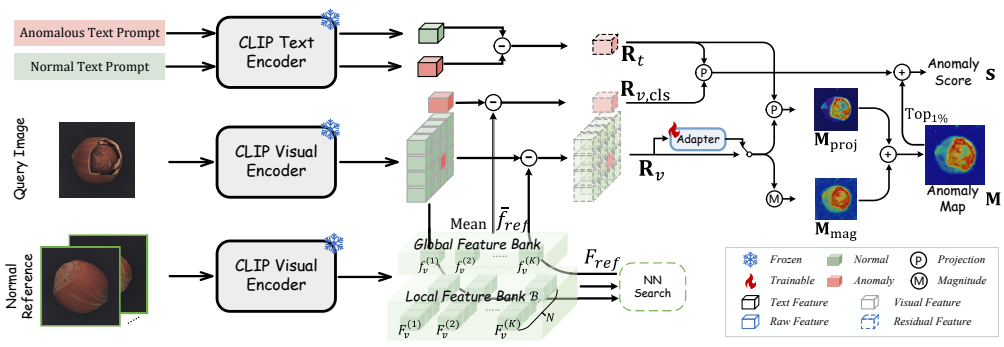


Figure 2: Overall architecture of Res^{2}CLIP.

encode both anomaly-related information and category-specific visual states. Consequently, learning anomaly-related changes is coupled with modeling inter-category representation differences.

2.2 Calibrating CLIP Alignment with Normal References

Few-Shot AD provides normal references that describe the expected normal state of the query. We therefore calibrate the query's **T−I** response using the response of its normal reference as a semantic baseline. Given a query feature f_q and its corresponding normal reference feature f_{ref}, we define

$$\Delta \mathbf{S}_{\text{T-I}} = \mathbf{S}_{\text{T-I}}(f_q) - \mathbf{S}_{\text{T-I}}(f_{\text{ref}}) = (\hat{f}_q - \hat{f}_{\text{ref}})^\top (\hat{F}_t^a - \hat{F}_t^n) = \mathbf{R}_v^\top \mathbf{R}_t = \|\mathbf{R}_t\|_2 \underbrace{\mathbf{R}_v^\top \hat{\mathbf{R}}_t}_{\mathbf{S}_{\text{R-R}}}, \tag{2}$$

where $\mathbf{R}_v = \hat{f}_q - \hat{f}_{\text{ref}}$ and $\mathbf{R}_t = \hat{F}_t^a - \hat{F}_t^n$ denote the visual and text residuals, respectively. Since $\|\mathbf{R}_t\|_2$ is fixed for given text anchors, we omit this constant scale and use $\mathbf{S}_{\text{R-R}} = \mathbf{R}_v^\top \hat{\mathbf{R}}_t$ as the Residual-Residual (**R−R**) Alignment score.

After calibration, the alignment is performed between deviations rather than raw states: $\mathbf{R}_v$ describes the visual change from the normal reference, while $\mathbf{R}_t$ describes the semantic change from normal to anomalous text. For localization, subtracting the normal-reference response evaluates each local feature by its deviation from the corresponding normal baseline, reducing the influence of region-dependent normal responses. For optimization, using $\mathbf{R}_v$ instead of f_q shifts learning from category-dependent visual states to reference-relative deviations with reduced category dependence, allowing the model to focus more directly on deviations from normality.

3 Res^{2}CLIP

Problem Definition. Few-shot GAD aims to detect and localize anomalies from unseen classes using only a few normal reference images, without target-class retraining. Let $\mathcal{C}_{\text{seen}}$ and $\mathcal{C}_{\text{unseen}}$ denote the seen training classes and unseen test classes, respectively, with $\mathcal{C}_{\text{seen}} \cap \mathcal{C}_{\text{unseen}} = \emptyset$. When training is involved, an auxiliary set $\mathcal{I}_{\text{train}}$ from $\mathcal{C}_{\text{seen}}$ is used, with image-level labels $y^{gt} \in \{0,1\}$ and pixel-level masks $\mathbf{M}^{gt} \in \{0,1\}^{H\times W}$. At inference, given a query image $I_q \in \mathbb{R}^{H\times W\times 3}$ from an unseen class and K normal references $\mathcal{P}_{\text{ref}} = \{I_{\text{ref}}^{(1)}, \ldots, I_{\text{ref}}^{(K)}\}$ from the same class, the model predicts an image-level anomaly score s and a pixel-level anomaly map $\mathbf{M} \in \mathbb{R}^{H\times W}$.

3.1 Overview

Building on R-R Alignment, Res^{2}CLIP constructs visual deviations from matched normal references and relates them to the normal-to-anomaly textual deviation, as in Fig. 2. Given a query image I_q and K normal reference images, the frozen CLIP visual encoder extracts the global `[CLS]` feature $f_{v,\text{cls}}^q$ and multi-layer local features $\{F_{v,l}^q\}_{l=1}^L$ for the query, together with $f_{v,\text{cls}}^{(k)}$ and $\{F_{v,l}^{(k)}\}_{l=1}^L$ for the k-th normal reference. The text encoder produces the normal and anomalous semantic anchors F_t^n and F_t^a. Res^{2}CLIP constructs visual and text residuals from these features and directly exploits them for training-free inference. In the fine-tune mode, only the local visual residuals are further optimized in the residual space, while the CLIP features and text residual remain fixed.

Residual Construction. The text residual is defined as $\mathbf{R}_t = \hat{F}_t^a - \hat{F}_t^n$, representing the semantic deviation from normal to anomalous text. For local visual features, we construct a normal feature bank at each selected layer as $B_l = \bigcup_{k=1}^{K} \hat{F}_{v,l}^{(k)}$. For the i-th query token $F_{v,l}^q(i)$ at the l-th layer, we retrieve its κ nearest neighbors $\mathcal{N}_\kappa(i)$ from B_l according to cosine similarity and aggregate them with similarity-based softmax weights:

$$\alpha_{ij} = \mathrm{Softmax}_{j\in\mathcal{N}_\kappa(i)}\left(\frac{\hat{F}_{v,l}^q(i)^\top B_l(j)}{\tau_{nn}}\right), \quad F_{\mathrm{ref},l}(i) = \sum_{j\in\mathcal{N}_\kappa(i)} \alpha_{ij} B_l(j), \tag{3}$$

where $B_l(j)$ denotes the j-th token in the reference bank and τ_{nn} is the softmax temperature. The local visual residual of the i-th query token at the l-th layer is then given by $\mathbf{R}_{v,l}(i) = \hat{F}_{v,l}^q(i) - \hat{F}_{\mathrm{ref},l}(i)$. For global representations, we average the `[CLS]` features of the normal references as $\bar{f}_{\mathrm{ref}} = \frac{1}{K}\sum_{k=1}^{K} f_{v,\mathrm{cls}}^{(k)}$ and construct the global visual residual $\mathbf{R}_{v,\mathrm{cls}} = \hat{f}_{v,\mathrm{cls}}^q - \hat{\bar{f}}_{\mathrm{ref}}$.

Training-Free and Fine-Tune. We use $\widetilde{\mathbf{R}}_{v,l}$ to denote the local visual residual representation used for anomaly scoring. In the training-free mode, no residual optimization is performed and $\widetilde{\mathbf{R}}_{v,l} = \mathbf{R}_{v,l}$. In the fine-tune mode, $\widetilde{\mathbf{R}}_{v,l}$ denotes the output of the residual-space optimization introduced in Sec. 3.3. The text residual $\mathbf{R}_t$ and the global visual residual $\mathbf{R}_{v,\mathrm{cls}}$ remain fixed in both modes.

3.2 Residual-based Anomaly Scoring

R-R Alignment leverages the semantic knowledge inherited from pre-trained CLIP by measuring the alignment between visual residuals and the fixed text residual $\mathbf{R}_t$, which provides an anomaly-related semantic direction. However, real-world anomalies are diverse and open-ended, and previously unseen anomaly patterns may not be sufficiently characterized by the selected textual semantics. Relying solely on semantic projection may therefore overlook visual deviations that do not strongly align with $\mathbf{R}_t$. To retain such anomaly evidence, we jointly consider the magnitude of the visual residual and its projection onto the text residual.

For the i-th token at the l-th selected layer, we define the magnitude and semantic projection of the corresponding local visual residual as

$$m_l(i) = \left\|\widetilde{\mathbf{R}}_{v,l}(i)\right\|_2, \qquad p_l(i) = \widetilde{\mathbf{R}}_{v,l}(i)^\top \hat{\mathbf{R}}_t, \tag{4}$$

where $m_l(i)$ measures the visual deviation of the token from its matched normal reference, while $p_l(i)$ measures the consistency of this deviation with the anomaly-related semantic direction. The local anomaly score is defined as

$$S_l(i) = m_l(i) + \lambda_p [p_l(i)]_+, \qquad [x]_+ = \max(x, 0), \tag{5}$$

where the projection weight λ_p is fixed for each inference mode and shared across all datasets. The magnitude term provides anomaly evidence independent of the textual semantics, while the positive projection enhances deviations that are consistent with the anomaly-related semantic direction.

The token-wise scores are rearranged into the spatial score map $\mathbf{S}_l$ at each selected layer. We aggregate the score maps across layers and upsample them to the input resolution to obtain the pixel-level anomaly map:

$$\mathbf{M} = \mathrm{Upsample}\left(\frac{1}{L}\sum_{l=1}^{L} \mathbf{S}_l\right). \tag{6}$$

For image-level detection, the `[CLS]` token mainly captures global image content and category-level information. Therefore, localized anomalies may induce only limited changes in the magnitude of the global visual residual. However, the residual direction still indicates whether this global change is consistent with anomaly-related semantics. We thus retain only the signed projection $\mathbf{R}_{v,\mathrm{cls}}^\top \hat{\mathbf{R}}_t$ as complementary global semantic evidence. We combine this response with the strongest local anomaly responses in $\mathbf{M}$ to obtain

$$s = \mathbf{R}_{v,\mathrm{cls}}^\top \hat{\mathbf{R}}_t + \mathrm{Mean}\left(\mathrm{Top}_{1\%}(\mathbf{M})\right). \tag{7}$$

where $\mathrm{Top}_{1\%}(\cdot)$ selects the highest 1% responses in the anomaly map.

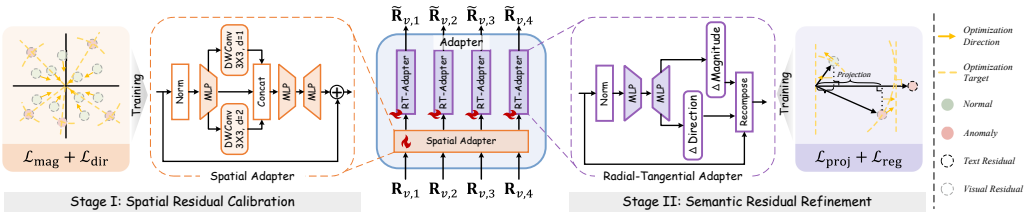


Figure 3: Overview of the two-stage residual optimization strategy.

3.3 Optimization in Residual Space

While residual representations derived from pre-trained CLIP already provide effective anomaly evidence, supervision from seen classes allows their responses to be further adapted to observed anomaly patterns. We therefore optimize local visual residuals while keeping the CLIP encoders and the text residual $\mathbf{R}_t$ fixed. As in Fig. 3, we adopt a two-stage strategy: Stage I calibrates residual responses using spatial context, while Stage II refines their semantic relation to the text residual.

3.3.1 Stage I: Spatial Residual Calibration

Normal-reference matching provides a local baseline for each query token, but limited reference coverage and token-wise retrieval may leave residual responses caused by normal inter-sample variations. Stage I therefore performs spatial residual calibration to suppress these responses while preserving anomaly-related information in the original residuals.

Spatial Adapter. To achieve the goal, we introduce a spatial-aware adapter $\mathcal{A}_{\text{SP}}$, shared across all selected layers, to calibrate the visual residuals before semantic refinement:

$$\mathbf{R}_{v,l}^{(1)} = \mathcal{A}_{\text{SP}}\left(\mathbf{R}_{v,l}\right), \tag{8}$$

where $\mathbf{R}_{v,l}^{(1)}$ denotes the Stage-I residual. $\mathcal{A}_{\text{sp}}$ aggregates multi-scale spatial context with lightweight convolutional operations, with architectural details provided in *Appendix*.

Training Objective. $\mathcal{A}_{\text{sp}}$ is optimized with complementary constraints on residual magnitude and direction. Normal residual magnitudes are minimized, whereas anomalous residual magnitudes are constrained to remain close to their original values:

$$\mathcal{L}_{\text{mag}} = \sum_{l=1}^{L}\left[\frac{1}{N_0}\sum_{\mathbf{M}^{gt}(i)=0}\left\|\mathbf{R}_{v,l}^{(1)}(i)\right\|_2 + \frac{1}{N_1}\sum_{\mathbf{M}^{gt}(i)=1}\left|\left\|\mathbf{R}_{v,l}^{(1)}(i)\right\|_2 - \left\|\mathbf{R}_{v,l}(i)\right\|_2\right|\right], \tag{9}$$

where N_0 and N_1 are the numbers of normal and anomalous tokens, and $N = N_0 + N_1$. Since both the magnitude and direction of anomalous residuals carry useful information from the original CLIP space, we further introduce a directional constraint. As normal residuals have no explicit directional target and are already suppressed toward zero in magnitude, this constraint is applied only to anomalous residuals:

$$\mathcal{L}_{\text{dir}} = \sum_{l=1}^{L}\frac{1}{N_1}\sum_{\mathbf{M}^{gt}(i)=1}\left[1 - \hat{\mathbf{R}}_{v,l}^{(1)}(i)^{\top}\hat{\mathbf{R}}_{v,l}(i)\right]. \tag{10}$$

Combining the two constraints, the overall objective of Stage I is

$$\mathcal{L}_{\text{I}} = \mathcal{L}_{\text{mag}} + \mathcal{L}_{\text{dir}}. \tag{11}$$

3.3.2 Stage II: Semantic Residual Refinement

Although pre-trained CLIP provides transferable image-text representations, its alignment is learned for general semantic recognition rather than fine-grained anomaly discrimination. In particular, the relation between anomaly-induced visual residuals and the normal-to-anomaly text residual is not explicitly optimized during pre-training. With supervision from seen classes, Stage II further adapts this relation by encouraging anomalous visual residuals to align with the fixed text residual $\mathbf{R}_t$, while suppressing such alignment for normal residuals.

Radial-Tangential Adapter. Stage II primarily learns how visual residuals should be directionally refined with respect to anomaly-related semantics. A straightforward approach is to predict a feature update and directly add it to the current residual. However, under direct addition, the angular change induced by the same tangential update depends on the residual magnitude: smaller residuals rotate more, whereas larger residuals require larger updates to achieve the same rotation. The predictor must therefore learn both the desired semantic correction and a magnitude-dependent compensation for realizing that correction. To avoid this implicit coupling, we introduce a Radial-Tangential Adapter (RT-Adapter), which performs directional refinement in the normalized residual coordinate and parameterizes radial adjustment separately. A detailed analysis is provided in *Appendix D.3*.

For a Stage-I residual $\mathbf{R}_{v,l}^{(1)}(i)$, we denote its magnitude by $m_{v,l}^{(1)}(i) = \left\|\mathbf{R}_{v,l}^{(1)}(i)\right\|_2$. The RT-Adapter predicts a refinement vector $\mathbf{h}_{v,l}(i)$, whose detailed network architecture is provided in the *Appendix*. Relative to the current residual direction $\hat{\mathbf{R}}_{v,l}^{(1)}(i)$, this refinement is represented by a radial coefficient and a tangential component:

$$a_{v,l}(i) = \hat{\mathbf{R}}_{v,l}^{(1)}(i)^\top \mathbf{h}_{v,l}(i), \qquad \mathbf{h}_{v,l}^{\perp}(i) = \mathbf{h}_{v,l}(i) - a_{v,l}(i)\hat{\mathbf{R}}_{v,l}^{(1)}(i). \tag{12}$$

Here, $a_{v,l}(i)$ controls the radial adjustment of residual magnitude, whereas $\mathbf{h}_{v,l}^{\perp}(i)$ controls its directional adjustment. The two components are applied separately to update the residual magnitude and normalized direction:

$$\widetilde{m}_{v,l}(i) = \max\left(\epsilon,\, m_{v,l}^{(1)}(i) + a_{v,l}(i)\right), \tag{13}$$

$$\widetilde{\mathbf{R}}_{v,l}(i) = \widetilde{m}_{v,l}(i)\,\mathrm{Norm}\left(\hat{\mathbf{R}}_{v,l}^{(1)}(i) + \mathbf{h}_{v,l}^{\perp}(i)\right), \tag{14}$$

where $\epsilon = 10^{-6}$ ensures numerical stability and a positiveresidual magnitude. By performing the directional update in the normalized residual coordinate, this parameterization prevents the original residual magnitude from implicitly determining the induced angular change, while retaining an explicit radial adjustment.

Training Objective. The fixed text residual $\mathbf{R}_t$ provides the semantic target for optimizing the refined visual residuals. For normal regions, positive projections toward the anomaly-related textual direction are suppressed. For anomalous regions, the R-R projection is encouraged to reach the magnitude of the corresponding Stage-I residual feature:

$$\mathcal{L}_{\mathrm{proj}} = \sum_{l=1}^{L}\left[\frac{1}{N_0}\sum_{\mathbf{M}^{gt}(i)=0}[p_l(i)]_+ + \frac{1}{N_1}\sum_{\mathbf{M}^{gt}(i)=1}\left[m_{v,l}^{(1)}(i) - p_l(i)\right]_+\right]. \tag{15}$$

Since the projection of a residual with magnitude $m_{v,l}^{(1)}(i)$ reaches $m_{v,l}^{(1)}(i)$ when it is fully aligned with $\hat{\mathbf{R}}_t$, the Stage-I magnitude provides a residual-specific alignment target without introducing a fixed global margin.

Since the projection depends on both residual magnitude and direction, the projection objective can be improved either by rotating a residual toward $\hat{\mathbf{R}}_t$ or by increasing its magnitude. The latter provides a trivial radial shortcut that does not necessarily improve semantic alignment. We therefore regularize the radial variation:

$$\mathcal{L}_{\mathrm{reg}} = \sum_{l=1}^{L}\frac{1}{N}\sum_{i}\left|\widetilde{m}_{v,l}(i) - m_{v,l}^{(1)}(i)\right|. \tag{16}$$

Combining semantic alignment with radial regularization gives

$$\mathcal{L}_{\mathrm{II}} = \mathcal{L}_{\mathrm{proj}} + \mathcal{L}_{\mathrm{reg}}. \tag{17}$$

4 Experiments

4.1 Experimental Setup

Datasets. Our evaluation covers five anomaly detection datasets: MVTecAD (Bergmann et al., 2019), VisA (Zou et al., 2022), BTAD (Mishra et al., 2021), MPDD (Jezek et al., 2021), and DTD-Synthetic (Aota et al., 2023) (abbreviated as DTD). Following the few-shot generalist anomaly detection protocol, we use the test set of MVTecAD as the auxiliary training set for all target datasets,

Table 1: Comprehensive quantitative comparison on five anomaly detection datasets under few-shot settings (I-AUC / P-AP / PRO), with the best results in **bold** and the second-best underlined.

Shots	Dataset	*Training-free*		*Fine-tune*				
		WinCLIP+	Res^{2}CLIP* (Ours)	APRIL-GAN	AnomalyCLIP+	ReMP-AD	AdaptCLIP	Res^{2}CLIP† (Ours)
1	MVTecAD	93.1 / 38.5 / 84.1	95.3 / 54.3 / 89.5	92.2 / 52.5 / 90.8	93.4 / 46.2 / 89.5	95.2 / 57.3 / **91.6**	95.0 / 53.9 / 89.7	**95.6 / 58.4** / 91.3
	VisA	82.9 / 15.5 / 80.2	88.3 / 33.3 / 89.0	91.6 / 31.0 / 90.1	80.0 / 28.2 / 89.6	**92.1** / 35.3 / 91.8	91.2 / 40.0 / 91.2	89.7 / **41.0 / 92.0**
	BTAD	85.0 / 42.0 / 66.7	95.9 / 62.9 / 81.3	91.7 / 50.1 / 78.6	92.7 / 58.4 / 79.4	94.9 / 53.9 / 82.3	93.1 / 61.4 / 77.7	**96.0 / 64.8 / 82.7**
	MPDD	69.1 / 30.7 / 88.4	**85.6** / 32.4 / 92.5	82.6 / 35.6 / 91.2	79.7 / 36.8 / 93.5	82.8 / **38.9** / 93.7	84.1 / 34.7 / 93.7	83.6 / 36.6 / **93.8**
	DTD	98.1 / 48.8 / 90.3	94.3 / 66.6 / 86.6	98.4 / 76.5 / 92.0	95.3 / 70.6 / 91.5	**99.3 / 79.2** / 92.5	98.0 / 76.9 / 93.1	97.0 / 77.5 / **93.6**
	Average	85.6 / 35.1 / 81.9	92.4 / 49.7 / 89.6	91.3 / 49.1 / 88.5	88.2 / 48.0 / 88.7	**92.9** / 52.9 / 90.4	92.3 / 53.4 / 89.1	92.4 / **55.6 / 90.7**
2	MVTecAD	94.7 / 40.0 / 84.6	95.5 / 56.9 / 90.9	92.4 / 53.5 / 91.3	93.7 / 47.7 / 90.2	95.2 / 57.2 / 91.9	95.7 / 55.5 / 90.3	**96.6 / 61.2 / 92.4**
	VisA	83.5 / 16.5 / 80.8	89.4 / 35.9 / 89.5	92.3 / 31.6 / 90.1	82.2 / 28.8 / 89.4	92.6 / 35.1 / 91.9	**92.7** / 41.1 / 91.6	90.5 / **42.5 / 92.7**
	BTAD	85.7 / 43.3 / 66.7	95.9 / 63.6 / 81.2	91.9 / 50.6 / 78.3	92.5 / 59.1 / 79.0	95.2 / 54.4 / 82.3	92.9 / 61.7 / 77.9	**96.5 / 66.0 / 83.2**
	MPDD	70.1 / 31.9 / 89.5	85.6 / 32.4 / 92.5	84.6 / 36.9 / 91.3	78.1 / 37.6 / 94.1	**87.0 / 41.4** / 93.8	85.4 / 35.7 / **94.3**	85.0 / 37.8 / 94.0
	DTD	98.2 / 49.3 / 90.4	95.0 / 69.2 / 87.7	98.4 / 76.9 / 92.2	95.0 / 71.2 / 91.7	**99.4 / 80.7** / 93.1	98.2 / 78.0 / 93.2	97.3 / 78.5 / **94.2**
	Average	86.4 / 36.2 / 82.4	93.0 / 51.5 / 90.1	91.9 / 49.9 / 88.6	88.3 / 48.9 / 88.9	**93.9** / 53.8 / 90.6	93.0 / 54.4 / 89.5	93.2 / **57.2 / 91.3**
4	MVTecAD	95.1 / 42.0 / 85.5	97.3 / 59.7 / 92.2	92.7 / 54.8 / 91.9	93.8 / 48.6 / 90.7	95.8 / 58.0 / 92.3	96.8 / 57.8 / 91.1	**97.7 / 63.8 / 93.5**
	VisA	84.4 / 17.9 / 81.3	90.1 / 37.2 / 90.3	92.8 / 31.8 / 90.1	81.7 / 29.3 / 89.7	**93.3** / 34.8 / 91.9	**93.3** / 42.0 / 92.0	91.1 / **42.9 / 92.9**
	BTAD	86.4 / 45.0 / 67.0	96.5 / 64.6 / 81.3	91.8 / 51.1 / 78.5	92.4 / 59.3 / 79.1	94.4 / 55.6 / 82.0	93.1 / 62.8 / 78.3	**96.7 / 66.7 / 82.9**
	MPDD	71.8 / 32.2 / 88.5	88.1 / 38.0 / 94.2	85.9 / 37.3 / 91.1	78.4 / 38.6 / 94.2	**89.6 / 43.1** / 93.6	86.3 / 39.5 / 94.9	87.2 / 41.2 / **95.1**
	DTD	98.3 / 50.0 / 90.7	95.7 / 70.9 / 89.2	98.6 / 77.3 / 92.5	95.3 / 71.8 / 92.2	**99.4 / 81.2** / 93.5	98.2 / 78.4 / 93.6	97.4 / 79.4 / **94.6**
	Average	87.2 / 37.4 / 82.6	94.0 / 53.1 / 90.8	92.4 / 50.5 / 88.8	88.3 / 49.5 / 89.2	**94.5** / 54.5 / 90.7	93.5 / 56.1 / 90.0	94.0 / **58.8 / 91.8**

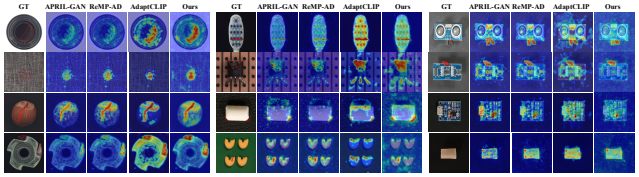


Figure 4: Anomaly map comparison with state-of-the-art methods under the 1-shot setting.

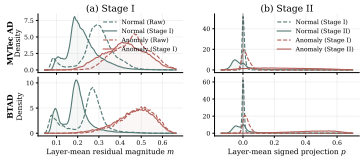


Figure 5: Evolution of residual magnitude and semantic projection across optimization stages.

and switch to the test set of VisA when MVTecAD itself is the target. This ensures that no images from the target domain are seen during training.

Evaluation Metrics. For each setting we evaluate both anomaly classification and anomaly localization. Image-level performance is reported with the area under the ROC curve (I-AUC). Pixel-level performance is reported with the average precision (P-AP), and the per-region overlap (PRO).

Implementation Details. We adopt the CLIP model ViT-L/14@336px pretrained by OpenAI (Radford et al., 2021) as the backbone, with V-V attention (Li et al., 2023) applied to the last 20 transformer layers of the visual encoder. The CLIP backbone is kept frozen throughout training. The input images are all resized to 518×518. We extract patch features from layers $\{6, 12, 18, 24\}$ ($L = 4$). All experiments are run on a single NVIDIA RTX 3090. We set $\lambda_p = 1$ for the training-free setting and $\lambda_p = 0.1$ for the fine-tune setting, and use the same values across all datasets. For normal-reference matching, we set the number of nearest neighbors to $\kappa = 5$ and the softmax temperature to $\tau_{\text{nn}} = 0.05$ in all experiments.

4.2 Main Results

We evaluate two settings of our method, **Res^{2}CLIP*** and **Res^{2}CLIP†**, corresponding to the training-free mode and the fine-tune mode.

Quantitative Comparison. Table 1 compares Res^{2}CLIP with WinCLIP (Jeong et al., 2023), APRIL-GAN (Chen et al., 2023), AnomalyCLIP (Zhou et al., 2023), ReMP-AD (Ma et al., 2025a), and AdaptCLIP (Gao et al., 2026) across five datasets under 1-, 2-, and 4-shot settings. Res^{2}CLIP† achieves the best average P-AP and PRO under all three settings, reaching 55.6%/90.7%, 57.2%/91.3%, and 58.8%/91.8%, respectively, while remaining competitive on I-AUC. The training-free Res^{2}CLIP* also shows strong localization performance, indicating that calibrated **R−R** Alignment is already effective without optimization, while fine-tuning further improves localization.

Qualitative Comparison. Fig. 4 shows representative anomaly maps on MVTec AD and VisA under the 1-shot setting. Res^{2}CLIP consistently produces tighter localization around the ground-truth defect regions, demonstrating superior localization precision.

Table 2: Ablation study of alignment methods and optimization space.

Mode		TF		FT		
Alignment		T–I	R–R	T–I	R–R	R–R
Optimization		–	–	$F_{v,l}$	$F_{v,l}$	$\mathbf{R}_{v,l}$
MVTec AD	I-AUC	89.6	90.4	87.4	94.7	93.6
	P-AP	25.5	30.9	36.8	46.8	48.8
	PRO	73.7	77.9	79.4	87.5	87.9
VisA	I-AUC	79.0	79.7	77.8	87.1	88.1
	P-AP	9.8	13.2	17.1	22.9	30.5
	PRO	73.0	68.6	83.3	84.9	89.3
BTAD	I-AUC	88.7	90.7	93.8	94.2	95.9
	P-AP	20.3	45.2	40.9	61.2	61.4
	PRO	58.0	74.7	71.8	80.9	83.7
Average	I-AUC	85.8	86.9	86.3	92.0	92.5
	P-AP	18.5	29.8	31.6	43.6	46.9
	PRO	68.2	73.7	78.2	84.4	86.9

Table 3: Ablation study of residual scoring under both modes.

Mode		TF			FT		
Scoring		M	P	M + P	M	P	M + P
MVTec AD	I-AUC	90.8	90.4	95.2	94.1	95.6	95.6
	P-AP	53.5	30.9	53.6	57.8	53.0	57.6
	PRO	87.9	77.9	89.1	90.9	89.1	90.9
VisA	I-AUC	86.9	79.7	88.7	89.8	89.5	90.3
	P-AP	33.2	13.2	32.9	41.5	39.1	41.5
	PRO	88.1	68.6	88.9	92.3	90.9	92.4
BTAD	I-AUC	94.4	90.7	96.0	95.6	95.9	96.2
	P-AP	63.4	45.2	65.0	66.6	65.0	66.6
	PRO	81.4	74.7	81.9	83.4	84.0	83.5
Average	I-AUC	90.7	86.9	93.3	93.2	93.7	94.0
	P-AP	50.0	29.8	50.5	55.3	52.4	55.2
	PRO	85.8	73.7	86.6	88.8	88.0	88.9



Figure 6: t-SNE visualization of the residuals before and after optimization.

Table 4: Ablation study of the proposed residual optimization.

Stage I	Stage II	Mode	$\mathcal{L}_{\mathrm{I}}$		$\mathcal{L}_{\mathrm{II}}$		$\mathcal{A}_{\mathrm{I}}$	$\mathcal{A}_{\mathrm{II}}$	MVTec AD			VisA			BTAD			Average		
			$\mathcal{L}_{\mathrm{mag}}$	$\mathcal{L}_{\mathrm{dir}}$	$\mathcal{L}_{\mathrm{proj}}$	$\mathcal{L}_{\mathrm{reg}}$			I-AUC	P-AP	PRO	I-AUC	P-AP	PRO	I-AUC	P-AP	PRO	I-AUC	P-AP	PRO
–	–	TF	–	–	–	–	–	–	95.2	53.6	89.1	88.7	32.9	88.9	96.0	65.0	81.9	93.3	50.5	86.6
✓	–	FT	✓	–	–	–	$\mathcal{A}_{\mathrm{SP}}$	–	95.9	56.0	91.0	90.0	38.0	91.5	96.2	64.4	82.8	94.0	52.8	88.4
✓	–	FT	✓	✓	–	–	$\mathcal{A}_{\mathrm{SP}}$	–	95.6	55.4	90.8	89.9	38.1	92.1	96.3	64.8	82.6	94.0	52.8	88.5
✓	–	FT	✓	✓	–	–	$\mathcal{A}_{\mathrm{MLP}}$	–	95.6	54.6	90.2	89.8	35.4	90.9	96.2	66.1	82.5	93.9	52.0	87.9
–	✓	FT	–	–	✓	✓	–	$\mathcal{A}_{\mathrm{RT}}$	95.5	56.9	90.6	90.5	39.8	91.2	96.1	67.3	83.6	94.0	54.7	88.5
✓	✓	FT	✓	✓	✓	–	$\mathcal{A}_{\mathrm{SP}}$	$\mathcal{A}_{\mathrm{RT}}$	95.2	57.0	90.7	90.0	41.2	92.2	96.0	66.4	83.6	93.7	54.8	88.8
✓	✓	FT	✓	✓	✓	✓	$\mathcal{A}_{\mathrm{SP}}$	$\mathcal{A}_{\mathrm{SP}}$	93.5	55.5	89.5	89.6	40.4	91.9	94.4	65.2	82.8	92.5	53.7	88.1
✓	✓	FT	✓	✓	✓	✓	$\mathcal{A}_{\mathrm{SP}}$	$\mathcal{A}_{\mathrm{MLP}}$	95.5	56.0	90.4	90.3	39.2	92.5	95.7	65.5	84.3	93.9	53.6	89.0
✓	✓	FT	✓	✓	✓	✓	$\mathcal{A}_{\mathrm{SP}}$	$\mathcal{A}_{\mathrm{RT}}$	95.6	57.6	90.9	90.3	41.5	92.4	96.2	66.6	83.5	94.0	55.2	88.9

4.3 Ablation Studies

All ablation experiments are conducted under the 1-shot setting, and for each dataset we keep the same normal reference image across all comparisons to ensure fairness. TF and FT denote the training-free and fine-tune modes, respectively.

Effectiveness of Normal-Reference Calibration and Residual-Space Optimization. Table 2 shows that normal-reference calibration consistently improves performance in both training-free and fine-tune settings, with particularly clear gains in P-AP and PRO. Moreover, directly optimizing the constructed residuals further improves performance over optimizing the original visual features before residual construction. These results validate both normal-reference calibration and explicit residual-space optimization.

Residual Scoring. In Table 3, $\mathbf{M}$ and $\mathbf{P}$ denote the residual magnitude and projection terms, respectively. Both terms contribute to anomaly scoring, while their combination provides the strongest overall performance across the training-free and fine-tune settings. This verifies the effectiveness of jointly modeling residual magnitude and semantic projection.

Benefits of Two-Stage Optimization. We first examine the individual and joint effects of the two optimization stages in Table 4. Either stage improves the training-free baseline, while combining them provides the strongest overall performance. These results support the progressive optimization of spatial residual calibration followed by semantic residual refinement.

Role of Optimization Objectives. The objective ablations further examine whether the proposed constraints behave as intended. In Stage I, $\mathcal{L}_{\mathrm{dir}}$ preserves the original anomaly directions while optimizing residual magnitudes. In Stage II, removing $\mathcal{L}_{\mathrm{reg}}$ consistently degrades performance, indicating that semantic refinement benefits from constraining unnecessary radial changes.

Advantages of Stage-Specific Adapters. In Table 4, $\mathcal{A}_{\mathrm{I}}$ and $\mathcal{A}_{\mathrm{II}}$ denote the adapters used in Stage I and Stage II, respectively. $\mathcal{A}_{\mathrm{SP}}$ is the proposed Spatial Adapter, $\mathcal{A}_{\mathrm{MLP}}$ removes its convolutional spatial modeling while retaining the feature transformation, and $\mathcal{A}_{\mathrm{RT}}$ denotes the proposed Radial-Tangential Adapter. For Stage I, $\mathcal{A}_{\mathrm{SP}}$ consistently outperforms $\mathcal{A}_{\mathrm{MLP}}$, confirming that spatial context is beneficial for calibrating local residual variations. For Stage II, $\mathcal{A}_{\mathrm{RT}}$ performs better than both $\mathcal{A}_{\mathrm{SP}}$ and $\mathcal{A}_{\mathrm{MLP}}$. This supports explicitly separating radial and tangential updates, which allows residual direction to be refined without its angular change being implicitly coupled to the residual magnitude.

Quantitative Evidence of Two-Stage Optimization. We further analyze the residual magnitude and projection in Fig. 5. Stage I mainly suppresses the magnitude of normal residuals while preserving anomaly responses, whereas Stage II further strengthens the projection of anomalous resid-

uals onto the text residual, matching the intended roles of the two stages. The t-SNE visualization in Fig. 6 provides complementary evidence of this process. Even without optimization, anomalous visual residuals already exhibit semantic alignment with the text residual; after optimization, this alignment becomes substantially stronger and more concentrated.

5 Related Work

Few-Shot Anomaly Detection. Unsupervised anomaly detection methods rely on abundant normal samples to learn reconstruction patterns (You et al., 2022; Zhang et al., 2023), feature distributions (Gudovskiy et al., 2022; Zhou et al., 2024), or feature modeling (Deng & Li, 2022; Guo et al., 2023; Gu et al., 2023; Guo et al., 2024; Liu et al., 2025), but their performance degrades sharply in few-shot settings. Few-shot anomaly detection methods address this by discriminating and localizing anomalies from only a handful of normal samples per class. Training-free approaches such as PaDiM (Defard et al., 2021) and PatchCore (Roth et al., 2022) construct memory banks or estimate distributions from the available samples and detect anomalies via nearest-neighbor or Mahalanobis distance, but do not learn transferable detection capabilities. Optimization-based approaches including RegAD (Huang et al., 2022), FastRecon (Fang et al., 2023), PromptAD (Li et al., 2024), IIPAD (Lv et al., 2025), and FoundAD (Zhai et al., 2025) fine-tune registration networks, prompts, or manifold projections, but still require per-class re-modeling. To remove this per-class dependency, the generalist paradigm trains a unified model on auxiliary categories and applies it directly to novel ones without adaptation. WinCLIP (Jeong et al., 2023) and APRIL-GAN (Chen et al., 2023) inject CLIP's language priors into few-shot detection, and InCTRL (Zhu & Pang, 2024) formally defines the generalist task by combining text prompts with query-support residuals. Subsequent methods explore visual and textu perspectives in different ways: ResAD (Yao et al., 2024) leverages residual features within the visual feature space to eliminate domain shifts; MetaUAS (Gao, 2024) unifies anomaly segmentation into change segmentation; DictAS (Qu et al., 2025b) reformulates detection as dictionary lookup; AdaptCLIP (Gao et al., 2026) inserts adaptive adapters for zero-/few-shot generalization; and ReMP-AD (Ma et al., 2025a) fuses retrieval-augmented multimodal prompts. Among these, ResAD is closest to our work in adopting a residual representation, but operates entirely within the visual modality. In contrast, we extend the residual paradigm to a symmetric multimodal alignment between visual and text residuals.

CLIP-based Anomaly Detection Recently, researchers have introduced pre-trained vision-language models such as CLIP (Radford et al., 2021) into anomaly detection, leveraging their massive pre-trained knowledge and strong vision-text alignment capability. WinCLIP (Jeong et al., 2023) pioneered hand-crafted normal/abnormal prompt templates for language-guided zero-/few-shot detection. AnomalyCLIP (Zhou et al., 2023) learns object-agnostic prompts to capture generic normality and abnormality patterns. AdaCLIP (Cao et al., 2024) introduces static-dynamic hybrid prompts for test-time adaptation. AA-CLIP (Ma et al., 2025b) constructs anomaly-aware textual anchors and aligns patch-level visual features with these anchors through a two-stage strategy. Bayes-PFL (Qu et al., 2025a) models the prompt space as a learnable distribution from a Bayesian perspective. VCP-CLIP (Qu et al., 2024) injects global visual context into prompts to eliminate product-specific dependencies, and MRAD (Xu et al., 2026) constructs a memory bank for direct anomaly score retrieval. These methods all perform alignment within CLIP's original feature space, where unified text embeddings struggle to capture fine-grained foreground-background differences and domain-specific optimization risks overfitting to seen classes. Our work instead shifts multimodal alignment into the residual domain.

6 Conclusions

We present Res^{2}CLIP for few-shot Generalist Anomaly Detection by reconsidering how CLIP alignment should incorporate normal references. Rather than aligning raw query states with textual semantics, we calibrate the query response using a matched normal reference, naturally yielding Residual-Residual Alignment between visual and textual deviations. This formulation supports effective training-free anomaly scoring and provides a suitable space for further residual optimization when auxiliary supervision is available. Experiments demonstrate strong performance across multiple benchmarks and validate the proposed calibration, scoring, and optimization designs.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, assist in the writing of proofs, support qualitative and thematic data analysis, and interpret results. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, or clean and reformat dataset, and support for qualitative and thematic data analysis is not applicable to this work.

Additionally, we used generative AI tools to create or edit software code, draft parts of a research paper, and edit a research paper to improve readability. In particular, code assistance was mainly used for routine tasks, such as generating visualization code based on existing implementations and experimental outputs.

We have reviewed all AI-assisted work. In particular, AI-assisted proofs were checked by the authors, and AI-generated visualization code was reviewed and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the formulation of residual construction, anomaly scoring, and the two-stage optimization procedure in the main manuscript. Detailed adapter architectures, training and inference configurations, hyperparameter settings, and additional ablation studies are provided in the *Appendix*. All datasets and evaluation metrics used in our experiments are publicly available, and the corresponding experimental protocols are described in the experimental section and *Appendix*.

REFERENCES

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

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

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European conference on computer vision*, pp. 55–72. Springer, 2024.

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

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pp. 475–489. Springer, 2021.

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

Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17481–17490, 2023.

Bin-Bin Gao. Metauas: Universal anomaly segmentation with one-prompt meta-learning. *Advances in Neural Information Processing Systems*, 37:39812–39836, 2024.

Bin-Bin Gao, Yue Zhou, Jiangtao Yan, Yuezhi Cai, Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, and Chengjie Wang. Adaptclip: Adapting clip for universal visual anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4095–4103, 2026.

Zhihao Gu, Liang Liu, Xu Chen, Ran Yi, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Annan Shu, Guannan Jiang, and Lizhuang Ma. Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16401–16409, 2023.

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

Hewei Guo, Liping Ren, Jingjing Fu, Yuwang Wang, Zhizheng Zhang, Cuiling Lan, Haoqian Wang, and Xinwen Hou. Template-guided hierarchical feature restoration for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6447–6458, 2023.

Jia Guo, Lize Jia, Weihang Zhang, Huiqi Li, et al. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024.

Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *European conference on computer vision*, pp. 303–319. Springer, 2022.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19606–19616, 2023.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pp. 66–71. IEEE, 2021.

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang 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*, pp. 16838–16848, 2024.

Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv e-prints*, pp. arXiv–2304, 2023.

Xinyue Liu, Jianyuan Wang, Biao Leng, and Shuo Zhang. Unlocking the potential of reverse distillation for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5640–5648, 2025.

Wenxi Lv, Qinliang Su, and Wenchao Xu. One-for-all few-shot anomaly detection via instance-induced prompt learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Hongchi Ma, Guanglei Yang, Debin Zhao, Yanli Ji, and Wangmeng Zuo. Remp-ad: Retrieval-enhanced multi-modal prompt fusion for few-shot industrial visual anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20425–20434, 2025a.

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

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pp. 1614–1623. PMLR, 2016.

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

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

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

Zhen Qu, Xian Tao, Xinyi Gong, ShiChen Qu, Xiaopei Zhang, Xingang Wang, Fei Shen, Zhengtao Zhang, Mukesh Prasad, and Guiguang Ding. Dictas: A framework for class-generalizable few-shot anomaly segmentation via dictionary lookup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20519–20528, 2025b.

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

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

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

Xincheng Yao, Zixin Chen, Chao Gao, Guangtao Zhai, and Chongyang Zhang. Resad: A simple framework for class generalizable anomaly detection. *Advances in Neural Information Processing Systems*, 37:125287–125311, 2024.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022.

Guangyao Zhai, Yue Zhou, Xinyan Deng, Lars Heckler, Nassir Navab, and Benjamin Busam. Foundation visual encoders are secretly few-shot anomaly detectors. *arXiv preprint arXiv:2510.01934*, 2025.

Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6782–6791, 2023.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023.

Yixuan Zhou, Xing Xu, Jingkuan Song, Fumin Shen, and Heng Tao Shen. Msflow: Multiscale flow-based framework for unsupervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17826–17836, 2024.

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

APPENDIX

A EQUIVALENCE OF SOFTMAX-BASED SCORING AND T-I ANOMALY RESPONSE

Existing CLIP-based anomaly detection methods commonly convert the similarities to normal and anomalous text anchors into an anomaly probability using a binary Softmax. In the main manuscript, instead of directly using this probability, we define the **`T-I`** anomaly response as

$$\mathbf{S}_{\mathrm{T-I}}(f) = \hat{f}^\top \hat{F}_t^a - \hat{f}^\top \hat{F}_t^n. \tag{18}$$

The conventional binary Softmax-based **`T-I`** anomaly score can be written as

$$\begin{aligned} \mathbf{S}_{\mathrm{T-I}}^{\mathrm{softmax}}(f) &= \frac{\exp(\hat{f}^\top \hat{F}_t^a/\tau)}{\exp(\hat{f}^\top \hat{F}_t^n/\tau) + \exp(\hat{f}^\top \hat{F}_t^a/\tau)} \\ &= \sigma\left(\frac{\hat{f}^\top \hat{F}_t^a - \hat{f}^\top \hat{F}_t^n}{\tau}\right) = \sigma\left(\frac{\mathbf{S}_{\mathrm{T-I}}(f)}{\tau}\right), \end{aligned} \tag{19}$$

where $\tau > 0$ is the temperature and $\sigma(\cdot)$ denotes the sigmoid function. Since the sigmoid function is strictly monotonic, $\mathbf{S}_{\mathrm{T-I}}^{\mathrm{softmax}}$ and $\mathbf{S}_{\mathrm{T-I}}$ preserve the same anomaly ranking. We therefore directly use the simpler **`T-I`** anomaly response $\mathbf{S}_{\mathrm{T-I}}$ in the main manuscript.

B A LIKELIHOOD PERSPECTIVE ON RESIDUAL-BASED ANOMALY SCORING

We provide a likelihood-based interpretation of the two complementary cues used in residual-based anomaly scoring. Since normal visual residuals are expected to concentrate around the normal-reference state, while anomalous residuals exhibit both larger deviations and diverse anomaly-related variations, we model the normal and anomalous residual distributions as

$$p_n(\mathbf{R}_v) = \mathcal{N}\left(\mathbf{0}, \sigma_n^2 \mathbf{I}\right), \qquad p_a(\mathbf{R}_v) = \mathcal{N}\left(\mu \hat{\mathbf{R}}_t, \sigma_a^2 \mathbf{I}\right), \tag{20}$$

where $\mu > 0$ denotes the anomaly-related semantic shift along the text residual direction, and $\sigma_a > \sigma_n$ reflects the broader variation of anomalous residuals.

Given a visual residual $\mathbf{R}_v$, its log-likelihood ratio between the anomalous and normal hypotheses is

$$\Lambda(\mathbf{R}_v) = \log \frac{p_a(\mathbf{R}_v)}{p_n(\mathbf{R}_v)}. \tag{21}$$

According to the multivariate Gaussian density,

$$\log p_n(\mathbf{R}_v) = -\frac{D}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\|\mathbf{R}_v\|_2^2, \tag{22}$$

and

$$\log p_a(\mathbf{R}_v) = -\frac{D}{2}\log(2\pi\sigma_a^2) - \frac{1}{2\sigma_a^2}\left\|\mathbf{R}_v - \mu\hat{\mathbf{R}}_t\right\|_2^2. \tag{23}$$

The quadratic term in Eq. (23) can be expanded as

$$\left\|\mathbf{R}_v - \mu\hat{\mathbf{R}}_t\right\|_2^2 = \|\mathbf{R}_v\|_2^2 - 2\mu\mathbf{R}_v^\top\hat{\mathbf{R}}_t + \mu^2, \tag{24}$$

where $\|\hat{\mathbf{R}}_t\|_2 = 1$. Substituting Eqs. (22)–(24) into Eq. (21) yields

$$\Lambda(\mathbf{R}_v) = \left(\frac{1}{2\sigma_n^2} - \frac{1}{2\sigma_a^2}\right)\|\mathbf{R}_v\|_2^2 + \frac{\mu}{\sigma_a^2}\mathbf{R}_v^\top\hat{\mathbf{R}}_t + C, \tag{25}$$

where

$$C = \frac{D}{2}\log\frac{\sigma_n^2}{\sigma_a^2} - \frac{\mu^2}{2\sigma_a^2} \tag{26}$$

is independent of $\mathbf{R}_v$. Denoting

$$\alpha = \frac{1}{2\sigma_n^2} - \frac{1}{2\sigma_a^2}, \qquad \beta = \frac{\mu}{\sigma_a^2}, \tag{27}$$

we have $\alpha > 0$ and $\beta > 0$ under $\sigma_a > \sigma_n$ and $\mu > 0$, and Eq. (25) becomes

$$\Lambda(\mathbf{R}_v) = \alpha \|\mathbf{R}_v\|_2^2 + \beta \mathbf{R}_v^\top \hat{\mathbf{R}}_t + C. \tag{28}$$

Remark. Eq. (28) shows that the likelihood ratio jointly depends on two complementary statistics: the **residual magnitude**, which measures the overall deviation from the normal reference, and the **semantic projection**, which measures the deviation explained by the anomaly-related textual direction. The former captures visual deviations beyond the selected textual semantics, while the latter exploits the semantic prior encoded by $\mathbf{R}_t$. This motivates retaining both residual magnitude and semantic projection in anomaly scoring.

C Text Prompt Design

This section provides full details of the text prompts used to construct the normal and anomalous text anchor features F_t^n and F_t^a.

C.1 Prompt Construction

Each prompt is constructed by combining three components:

- A state word `[state]` describing the normal or anomalous condition, such as `flawless` or `damaged`.
- A generic noun `[noun]` referring to the specific category or a generic object term.
- A template $\mathcal{T}(\cdot)$ that wraps the state-noun phrase into a natural sentence.

A complete prompt is formed as $\mathcal{T}($`[state] [noun]`$)$. For example, $\mathcal{T}(\cdot) =$ `a photo of a` $\cdot$ with `[state]` $=$ `flawless` and `[noun]` $=$ `object` yields the prompt `a photo of a flawless object`.

C.2 State Words

We use small fixed sets of state words for the normal and anomalous conditions:

- Normal: {empty, `perfect`, `flawless`, `clean`, `undamaged`}. The empty state produces a noun-only prompt that captures the neutral semantic baseline.
- Anomalous: {`damaged`}.

C.3 Generic Noun

Following prior work (Zhou et al., 2023), we use a single generic noun `[noun]` $=$ `object` for all classes.

C.4 Templates

We use the following 15 templates, following common practice in CLIP-based zero-shot classification:

- `a photo of a {}.`
- `a photo of the {}.`
- `a cropped photo of a {}.`
- `a cropped photo of the {}.`
- `a close-up photo of a {}.`
- `a close-up photo of the {}.`

- `the {} in the image.`
- `the {} in the scene.`
- `an image of a {}.`
- `an image of the {}.`
- `a rendering of a {}.`
- `a photo of one {}.`
- `showing the {}.`
- `showing a {}.`
- `itap of a {}.`

C.5 Anchor Computation

Let $\mathcal{T}_{\text{tmpl}}$ denote the set of templates listed above, and let $\mathcal{W}^n$ and $\mathcal{W}^a$ denote the sets of normal and anomalous state words, respectively. The full prompt sets are:

$$\mathcal{P}_t^n = \left\{\mathcal{T}(\texttt{[state] [noun]}) : \texttt{[state]} \in \mathcal{W}^n,\ \mathcal{T} \in \mathcal{T}_{\text{tmpl}}\right\}, \tag{29}$$

$$\mathcal{P}_t^a = \left\{\mathcal{T}(\texttt{[state] [noun]}) : \texttt{[state]} \in \mathcal{W}^a,\ \mathcal{T} \in \mathcal{T}_{\text{tmpl}}\right\}, \tag{30}$$

where `[noun]` = `object`. Each prompt is encoded by the CLIP text encoder $\mathcal{E}_t$, and all encodings within a set are averaged to yield the final anchor features:

$$F_t^n = \frac{1}{|\mathcal{P}_t^n|} \sum_{T \in \mathcal{P}_t^n} \mathcal{E}_t(T), \quad F_t^a = \frac{1}{|\mathcal{P}_t^a|} \sum_{T \in \mathcal{P}_t^a} \mathcal{E}_t(T). \tag{31}$$

Remark. The same prompt sets and templates are used across all datasets. No per-dataset prompt tuning is performed.

D Adapter Architecture

This section provides the architectural details of the two adapters used in the fine-tuning setting. Stage I employs a shared Spatial Adapter to calibrate local visual residuals by exploiting residual patterns and spatial context. Stage II uses layer-specific Radial-Tangential Adapters (RT-Adapters) for semantic residual refinement. Both adapters adopt a bottleneck ratio of 4; with CLIP ViT-L/14, $D = 768$ and the bottleneck dimension is $D/4 = 192$.

D.1 Spatial Adapter

Stage I uses a single multi-scale convolutional adapter $\mathcal{A}_{\text{SP}}$, whose parameters are shared across all selected CLIP layers. Given a patch-level visual residual map $\mathbf{R}_{v,l} \in \mathbb{R}^{N \times D}$, the adapter first applies LayerNorm and projects the feature dimension into a $D/4$-dimensional bottleneck:

$$\mathbf{X}_l = \text{ReLU}\left(W_{\text{down}}\,\text{LN}(\mathbf{R}_{v,l})\right), \qquad \mathbf{X}_l \in \mathbb{R}^{N \times D/4}. \tag{32}$$

The resulting sequence is reshaped into a two-dimensional feature map to exploit the spatial organization of patch residuals.

Two depthwise convolutional branches are then applied in parallel:

$$\mathbf{X}_l^f = \text{DWConv}_{3\times3}(\mathbf{X}_l), \qquad \mathbf{X}_l^c = \text{DWConv}_{3\times3,d=2}(\mathbf{X}_l), \tag{33}$$

where the first branch captures fine local context with dilation 1, while the second enlarges the receptive field using dilation 2. The original bottleneck feature is retained as an identity branch and concatenated with the two convolutional responses:

$$\mathbf{X}_l^{\text{multi}} = \text{Concat}\left[\mathbf{X}_l, \mathbf{X}_l^f, \mathbf{X}_l^c\right]. \tag{34}$$

After flattening the feature map back to a sequence, the three branches are normalized and fused:

$$\mathbf{X}_l^{\text{fuse}} = \text{ReLU}\left(W_{\text{blend}}\,\text{LN}(\mathbf{X}_l^{\text{multi}})\right), \qquad \mathbf{X}_l^{\text{fuse}} \in \mathbb{R}^{N \times D/4}. \tag{35}$$

Finally, the fused feature is projected back to D dimensions and added to the input residual:

$$\mathbf{R}_{v,l}^{(1)} = \mathbf{R}_{v,l} + W_{\mathrm{up}}\mathbf{X}_l^{\mathrm{fuse}}. \tag{36}$$

All linear projections and depthwise convolutions are bias-free. W_{up} is initialized to zero, so the adapter starts from an identity mapping. The same $\mathcal{A}_{\mathrm{SP}}$ is applied to all selected layers, keeping Stage-I spatial calibration lightweight.

D.2 Radial-Tangential Adapter

The RT-Adapter in Stage II uses a lightweight bottleneck MLP to predict the refinement vector $\mathbf{h}_{v,l}(i)$ defined in the main paper. A separate predictor is instantiated for each selected CLIP layer. Given a Stage-I residual $\mathbf{R}_{v,l}^{(1)}(i)$, the predictor operates on its normalized direction:

$$\mathbf{h}_{v,l}(i) = W_{2,l}\,\mathrm{GELU}\left(W_{1,l}\,\mathrm{LN}\left(\hat{\mathbf{R}}_{v,l}^{(1)}(i)\right)\right), \tag{37}$$

where $W_{1,l} \in \mathbb{R}^{D/4\times D}$ and $W_{2,l} \in \mathbb{R}^{D\times D/4}$ are bias-free linear projections. The output projection $W_{2,l}$ is initialized to zero, so $\mathbf{h}_{v,l}(i) = \mathbf{0}$ at the beginning of Stage II.

The predictor contains no separate radial and tangential branches. It outputs a single D-dimensional refinement vector $\mathbf{h}_{v,l}(i)$, from which the radial coefficient $a_{v,l}(i)$ and tangential component $\mathbf{h}_{v,l}^{\perp}(i)$ are analytically obtained as defined in the main paper. Thus, RT-Adapter uses the same bottleneck predictor as a standard MLP adapter; its distinction lies in how the predicted refinement is applied to the residual rather than in additional prediction heads.

D.3 Analysis of Radial-Tangential Refinement

The primary purpose of Stage II is to refine the semantic direction of visual residuals toward the anomaly-related text residual. A straightforward implementation is to predict a feature-space update $\mathbf{h}$ and directly add it to the current residual. However, the angular effect of such an additive update depends strongly on the magnitude of the residual itself. This makes the predicted update an indirect parameterization of the semantic rotation that Stage II aims to learn.

Magnitude-dependent angular effect of additive refinement. Consider a visual residual

$$\mathbf{R} = m\hat{\mathbf{R}}, \qquad m = \|\mathbf{R}\|_2, \qquad \|\hat{\mathbf{R}}\|_2 = 1, \tag{38}$$

and decompose a predicted refinement into its radial and tangential components:

$$\mathbf{h} = a\hat{\mathbf{R}} + \mathbf{h}^{\perp}, \qquad \hat{\mathbf{R}}^{\top}\mathbf{h}^{\perp} = 0. \tag{39}$$

A conventional additive update gives

$$\mathbf{R}_{\mathrm{add}} = \mathbf{R} + \mathbf{h} = (m + a)\hat{\mathbf{R}} + \mathbf{h}^{\perp}. \tag{40}$$

Let δ_{add} denote the angle between the original residual direction $\hat{\mathbf{R}}$ and the updated residual $\mathbf{R}_{\mathrm{add}}$. Assuming $m + a > 0$, we have

$$\tan\delta_{\mathrm{add}} = \frac{\|\mathbf{h}^{\perp}\|_2}{m + a}. \tag{41}$$

For a fixed predicted update $(a, \mathbf{h}^{\perp})$, increasing m increases the denominator and therefore decreases δ_{add}. Hence, the same tangential update produces a large rotation for a weak residual but only a small rotation for a strong residual. Equivalently, producing a prescribed angular change δ requires

$$\|\mathbf{h}^{\perp}\|_2 = (m + a)\tan\delta, \tag{42}$$

so the predictor must generate increasingly large tangential updates as the residual magnitude increases.

This behavior is undesirable for semantic residual refinement. The quantity that Stage II primarily needs to learn is a directional correction toward anomaly-related semantics, yet direct addition makes the resulting angular change depend jointly on the predicted tangential component and the current residual magnitude. For the same predicted update, a residual with a smaller magnitude undergoes a larger angular change, whereas a residual with a larger magnitude undergoes a smaller angular change. Consequently, the predictor must additionally adapt the scale of $\mathbf{h}^{\perp}$ according to the current residual magnitude in order to produce a comparable directional correction across residuals.

Directional refinement in normalized coordinates. RT-Adapter instead applies the tangential refinement to the normalized residual direction:

$$\widetilde{\mathbf{d}} = \mathrm{Norm}\left(\hat{\mathbf{R}} + \mathbf{h}^{\perp}\right). \tag{43}$$

Let δ_{norm} denote the angle between $\hat{\mathbf{R}}$ and $\widetilde{\mathbf{d}}$. Since $\mathbf{h}^{\perp}$ is orthogonal to $\hat{\mathbf{R}}$,

$$\tan \delta_{\mathrm{norm}} = \|\mathbf{h}^{\perp}\|_2. \tag{44}$$

The angular effect is now determined directly by the predicted tangential component rather than being inversely scaled by the residual magnitude. This provides a more direct parameterization for learning semantic directional refinement.

D.4 Discussion

The adapters of two stages are designed for different objectives in residual-space optimization. Stage I focuses on residual calibration. The Spatial Adapter therefore operates on the complete residual features and incorporates multi-scale spatial context, allowing the residual response to be adjusted according to both its feature pattern and neighboring regions. Its objective suppresses residual responses from normal regions while retaining anomaly-related information in the original CLIP residuals.

Stage II instead focuses on semantic refinement. Its main objective is to adjust visual residuals with respect to the anomaly-related textual direction, for which directional correction is particularly important. Under direct feature-space addition, the angular change induced by a predicted update implicitly depends on the current residual magnitude. The RT-Adapter avoids this coupling by performing directional refinement in the normalized residual coordinate and parameterizing radial adjustment separately. Consequently, the tangential component directly controls the directional correction, while the radial component independently adjusts residual magnitude.

Accordingly, the two stages are complementary rather than redundant: Stage I exploits feature and spatial information to calibrate visual residual responses, whereas Stage II explicitly models how these residuals should be adjusted toward anomaly-related semantics. This separation allows spatial residual calibration and semantic alignment to be optimized with adapter structures suited to their respective objectives.

E Details of Experimental Setup

E.1 Details of Datasets

Table 5 summarizes the statistics of all datasets used in our evaluation. For each dataset, we report the number of object categories, the number of normal and anomalous test images.

Table 5: Statistics of evaluation datasets. #Normal and #Anomalous refer to the number of test images in each split.

Dataset	#Categories	#Normal	#Anomalous	Aux. Training
MVTecAD (Bergmann et al., 2019)	15	467	1258	VisA
VisA (Zou et al., 2022)	12	962	1200	MVTecAD
BTAD (Mishra et al., 2021)	3	451	290	MVTecAD
MPDD (Jezek et al., 2021)	6	176	282	MVTecAD
DTD-Synthetic (Aota et al., 2023)	12	357	947	MVTecAD

E.2 More Implementation Details

Feature Extraction. We use OpenAI CLIP ViT-L/14@336px as the backbone, with V-V attention applied to the last 20 transformer layers of the visual encoder. All input images are resized to 518×518, and local visual features are extracted from layers $\{6, 12, 18, 24\}$ ($L = 4$) in both the training-free and fine-tune settings. The CLIP backbone and the text residual $\mathbf{R}_t$ remain fixed throughout training and inference.

Training Configuration. Residual optimization is performed sequentially in two stages. Stage I trains the shared Spatial Adapter for 20 epochs. Stage II then freezes the Spatial Adapter from Stage I and trains four layer-specific Radial-Tangential Adapters for 10 epochs. Both stages use Adam with an initial learning rate of 5×10^{-4}, $\beta_1 = 0.5$, and $\beta_2 = 0.999$. The learning rate is linearly warmed up for the first epoch and then decayed with a cosine schedule. We use a batch size of 16 and gradient clipping with a maximum norm of 1.0. During training, one normal reference from the same category is randomly sampled for each query, corresponding to the 1-shot setting. All experiments are conducted on a single NVIDIA RTX 3090 (24 GB).

Inference Configuration. For each query token, we retrieve the top-κ nearest normal features using cosine similarity and construct its matched normal reference by softmax-weighted aggregation. We set $\kappa = 5$ and $\tau_{\text{nn}} = 0.05$. The projection weight is set to $\lambda_p = 1$ in the training-free setting and $\lambda_p = 0.1$ in the fine-tune setting, with the same values used across all datasets and shot settings. The few-shot reference set is fixed for all test images of the same class. Anomaly maps are smoothed with a Gaussian kernel ($\sigma = 4$) before evaluation.

Evaluation protocol. All few-shot experiments are repeated with $K \in \{1, 2, 4\}$ shots. For each K, the same set of normal reference images is used across all compared methods to ensure a fair comparison. Each metric is averaged over three different runs. To comprehensively evaluate the model across different aspects of anomaly perception, we report diverse metrics. Specifically, image-level metrics (I-AUC, I-AP, I-F1) assess the global classification performance, where I-AP and I-F1 are strictly included to provide robust evaluation under image-level class imbalance. Pixel-level metrics (P-AUC, P-AP, P-F1) measure the precision of fine-grained spatial localization. Given the extreme numerical imbalance between anomalous and normal pixels, P-AP and P-F1 serve as highly stringent indicators that heavily penalize spatial mismatch and background noise. Furthermore, the Per-Region Overlap (PRO) evaluates localization at the connected-component level, ensuring the assessment is not disproportionately dominated by large anomalous regions and accurately reflects sensitivity to subtle structural defects.

F Additional Experimental Results

F.1 Detailed Quantitative Metrics

Table 6, Table 7, Table 8, Table 9, and Table 10 provides the complete metric set on datasets MVTecAD (Bergmann et al., 2019), VisA (Zou et al., 2022), BTAD (Mishra et al., 2021), MPDD (Jezek et al., 2021), and DTD-Synthetic (Aota et al., 2023) including I-AP, I-F1, P-F1 for all datasets and all shot counts.

F.2 Efficiency and Performance Trade-off

To further demonstrate the practical deployment value of our proposed method, we visualize the trade-off between inference speed, localization performance (PRO), and parameter efficiency under the 4-shot setting. As illustrated in Fig. 7, the x-axis represents the inference speed (FPS), the y-axis indicates the average pixel-level localization performance (Average PRO), and the bubble size corresponds to the number of extra learnable parameters.

Res^{2}CLIP* requires no additional learnable parameters and achieves 90.8% PRO at 8.41 FPS. After residual optimization, Res^{2}CLIP† introduces only 1.60M trainable parameters while improving PRO to 91.8%, with inference speed remaining nearly unchanged at 8.28 FPS. For reference, the frozen CLIP backbone alone runs at 8.78 FPS, indicating that both residual construction and the proposed adapters introduce only limited inference overhead. This comparison demonstrates that our design achieves a highly competitive accuracy-efficiency trade-off.

Table 6: Full quantitative results on MVTecAD (Bergmann et al., 2019) across 1, 2, and 4 shots. All metrics are reported as mean ± standard deviation over three random runs. The best results are highlighted in **bold** and the second-best results are underlined.

Shots	Method	I-AUC	I-AP	I-F1	P-AUC	P-AP	P-F1	PRO
1	WinCLIP+ (Jeong et al., 2023)	93.1$_{\pm 0.4}$	96.6$_{\pm 0.1}$	93.7$_{\pm 0.2}$	93.6$_{\pm 0.1}$	38.5$_{\pm 0.4}$	42.7$_{\pm 0.4}$	84.1$_{\pm 0.3}$
	APRIL-GAN (Chen et al., 2023)	92.2$_{\pm 0.1}$	96.0$_{\pm 0.1}$	92.6$_{\pm 0.2}$	95.3$_{\pm 0.1}$	52.5$_{\pm 0.4}$	55.0$_{\pm 0.3}$	90.8$_{\pm 0.2}$
	AnomalyCLIP+ (Zhou et al., 2023)	93.4$_{\pm 0.4}$	97.1$_{\pm 0.3}$	93.7$_{\pm 0.2}$	94.8$_{\pm 0.2}$	46.2$_{\pm 0.5}$	49.1$_{\pm 0.2}$	89.5$_{\pm 0.3}$
	ReMP-AD (Ma et al., 2025a)	95.2$_{\pm 0.4}$	97.5$_{\pm 0.4}$	94.4$_{\pm 0.2}$	95.5$_{\pm 0.1}$	57.3$_{\pm 0.9}$	57.3$_{\pm 0.8}$	**91.6**$_{\pm 0.4}$
	AdaptCLIP (Gao et al., 2026)	95.0$_{\pm 0.6}$	97.6$_{\pm 0.4}$	**95.2**$_{\pm 0.3}$	94.2$_{\pm 0.1}$	53.9$_{\pm 1.3}$	53.8$_{\pm 1.0}$	89.7$_{\pm 0.3}$
	Res^{2}CLIP* (Ours)	95.3$_{\pm 0.2}$	**98.1**$_{\pm 0.1}$	95.1$_{\pm 0.4}$	95.9$_{\pm 0.2}$	54.3$_{\pm 0.7}$	55.2$_{\pm 0.7}$	89.5$_{\pm 0.3}$
	Res^{2}CLIP† (Ours)	**95.6**$_{\pm 0.4}$	98.0$_{\pm 0.1}$	95.0$_{\pm 0.2}$	**96.4**$_{\pm 0.1}$	**58.4**$_{\pm 0.6}$	**58.0**$_{\pm 0.5}$	91.3$_{\pm 0.3}$
2	WinCLIP+ (Jeong et al., 2023)	94.7$_{\pm 0.2}$	97.3$_{\pm 0.0}$	94.6$_{\pm 0.2}$	93.9$_{\pm 0.1}$	40.0$_{\pm 0.7}$	43.9$_{\pm 0.4}$	84.6$_{\pm 0.3}$
	APRIL-GAN (Chen et al., 2023)	92.4$_{\pm 0.3}$	96.1$_{\pm 0.1}$	92.5$_{\pm 0.2}$	95.6$_{\pm 0.1}$	53.5$_{\pm 0.2}$	55.8$_{\pm 0.2}$	91.3$_{\pm 0.1}$
	AnomalyCLIP+ (Zhou et al., 2023)	93.7$_{\pm 0.2}$	97.2$_{\pm 0.2}$	93.6$_{\pm 0.1}$	95.0$_{\pm 0.1}$	47.7$_{\pm 0.4}$	50.1$_{\pm 0.5}$	90.2$_{\pm 0.4}$
	ReMP-AD (Ma et al., 2025a)	95.2$_{\pm 0.2}$	97.4$_{\pm 0.1}$	94.4$_{\pm 0.2}$	95.5$_{\pm 0.1}$	57.2$_{\pm 0.1}$	57.5$_{\pm 0.2}$	91.9$_{\pm 0.2}$
	AdaptCLIP (Gao et al., 2026)	95.7$_{\pm 0.1}$	97.8$_{\pm 0.1}$	95.4$_{\pm 0.1}$	94.5$_{\pm 0.1}$	55.5$_{\pm 0.2}$	55.3$_{\pm 0.1}$	90.3$_{\pm 0.2}$
	Res^{2}CLIP* (Ours)	95.5$_{\pm 0.2}$	98.1$_{\pm 0.2}$	95.5$_{\pm 0.3}$	96.4$_{\pm 0.1}$	56.9$_{\pm 0.2}$	57.5$_{\pm 0.1}$	90.9$_{\pm 0.4}$
	Res^{2}CLIP† (Ours)	**96.6**$_{\pm 0.3}$	**98.4**$_{\pm 0.1}$	**95.8**$_{\pm 0.1}$	**96.9**$_{\pm 0.0}$	**61.2**$_{\pm 0.2}$	**60.5**$_{\pm 0.2}$	**92.4**$_{\pm 0.1}$
4	WinCLIP+ (Jeong et al., 2023)	95.1$_{\pm 0.2}$	97.5$_{\pm 0.1}$	95.1$_{\pm 0.2}$	94.3$_{\pm 0.1}$	42.0$_{\pm 0.2}$	45.7$_{\pm 0.2}$	85.5$_{\pm 0.5}$
	APRIL-GAN (Chen et al., 2023)	92.7$_{\pm 0.1}$	96.1$_{\pm 0.1}$	92.8$_{\pm 0.1}$	96.0$_{\pm 0.1}$	54.8$_{\pm 0.1}$	57.1$_{\pm 0.2}$	91.9$_{\pm 0.1}$
	AnomalyCLIP+ (Zhou et al., 2023)	93.8$_{\pm 0.2}$	97.3$_{\pm 0.2}$	93.8$_{\pm 0.1}$	95.2$_{\pm 0.1}$	48.6$_{\pm 0.2}$	50.9$_{\pm 0.3}$	90.7$_{\pm 0.1}$
	ReMP-AD (Ma et al., 2025a)	95.8$_{\pm 0.2}$	97.7$_{\pm 0.2}$	94.7$_{\pm 0.1}$	95.7$_{\pm 0.0}$	58.0$_{\pm 0.2}$	58.3$_{\pm 0.1}$	92.3$_{\pm 0.1}$
	AdaptCLIP (Gao et al., 2026)	96.8$_{\pm 0.2}$	98.5$_{\pm 0.0}$	96.2$_{\pm 0.2}$	94.8$_{\pm 0.1}$	57.8$_{\pm 0.1}$	57.5$_{\pm 0.1}$	91.1$_{\pm 0.0}$
	Res^{2}CLIP* (Ours)	97.3$_{\pm 0.3}$	98.8$_{\pm 0.2}$	96.8$_{\pm 0.1}$	96.9$_{\pm 0.0}$	59.7$_{\pm 0.1}$	60.6$_{\pm 0.1}$	92.2$_{\pm 0.1}$
	Res^{2}CLIP† (Ours)	**97.7**$_{\pm 0.1}$	**98.9**$_{\pm 0.1}$	**97.0**$_{\pm 0.2}$	**97.3**$_{\pm 0.1}$	**63.8**$_{\pm 0.3}$	**62.9**$_{\pm 0.3}$	**93.5**$_{\pm 0.0}$

Table 7: Full quantitative results on VisA (Zou et al., 2022) across 1, 2, and 4 shots. All metrics are reported as mean ± standard deviation over three random runs. The best results are highlighted in **bold** and the second-best results are underlined.

Shots	Method	I-AUC	I-AP	I-F1	P-AUC	P-AP	P-F1	PRO
1	WinCLIP+ (Jeong et al., 2023)	82.9$_{\pm 0.1}$	84.5$_{\pm 0.5}$	82.1$_{\pm 0.2}$	94.8$_{\pm 0.1}$	15.5$_{\pm 0.2}$	22.6$_{\pm 0.1}$	80.2$_{\pm 0.7}$
	APRIL-GAN (Chen et al., 2023)	91.6$_{\pm 0.4}$	93.6$_{\pm 0.3}$	87.1$_{\pm 0.4}$	96.1$_{\pm 0.0}$	31.0$_{\pm 0.3}$	38.5$_{\pm 0.2}$	90.1$_{\pm 0.1}$
	AnomalyCLIP+ (Zhou et al., 2023)	80.0$_{\pm 0.8}$	82.2$_{\pm 0.4}$	81.0$_{\pm 0.2}$	96.5$_{\pm 0.1}$	28.2$_{\pm 0.3}$	37.1$_{\pm 0.5}$	89.6$_{\pm 0.2}$
	ReMP-AD (Ma et al., 2025a)	**92.1**$_{\pm 0.6}$	**94.0**$_{\pm 0.3}$	**87.8**$_{\pm 0.5}$	96.9$_{\pm 0.1}$	35.3$_{\pm 0.4}$	42.1$_{\pm 0.5}$	91.8$_{\pm 0.1}$
	AdaptCLIP (Gao et al., 2026)	91.2$_{\pm 0.9}$	93.0$_{\pm 0.7}$	87.0$_{\pm 1.0}$	97.0$_{\pm 0.1}$	40.0$_{\pm 0.4}$	45.2$_{\pm 0.4}$	91.2$_{\pm 0.1}$
	Res^{2}CLIP* (Ours)	88.3$_{\pm 0.3}$	89.8$_{\pm 0.1}$	85.4$_{\pm 0.3}$	96.8$_{\pm 0.0}$	33.3$_{\pm 0.4}$	39.9$_{\pm 0.3}$	89.0$_{\pm 0.2}$
	Res^{2}CLIP† (Ours)	89.7$_{\pm 0.5}$	91.2$_{\pm 0.1}$	86.2$_{\pm 0.4}$	**97.7**$_{\pm 0.1}$	**41.0**$_{\pm 0.4}$	**46.2**$_{\pm 0.3}$	**92.0**$_{\pm 0.3}$
2	WinCLIP+ (Jeong et al., 2023)	83.5$_{\pm 0.4}$	84.9$_{\pm 0.6}$	82.5$_{\pm 0.3}$	95.0$_{\pm 0.1}$	16.5$_{\pm 0.6}$	23.9$_{\pm 0.7}$	80.8$_{\pm 0.3}$
	APRIL-GAN (Chen et al., 2023)	92.3$_{\pm 0.2}$	94.2$_{\pm 0.2}$	87.7$_{\pm 0.3}$	96.1$_{\pm 0.1}$	31.6$_{\pm 0.1}$	39.1$_{\pm 0.2}$	90.1$_{\pm 0.0}$
	AnomalyCLIP+ (Zhou et al., 2023)	82.2$_{\pm 0.7}$	84.6$_{\pm 0.2}$	81.1$_{\pm 0.3}$	96.4$_{\pm 0.1}$	28.8$_{\pm 0.3}$	37.6$_{\pm 0.2}$	89.4$_{\pm 0.3}$
	ReMP-AD (Ma et al., 2025a)	92.6$_{\pm 0.4}$	**94.3**$_{\pm 0.5}$	**88.7**$_{\pm 0.5}$	96.8$_{\pm 0.1}$	35.1$_{\pm 0.2}$	42.1$_{\pm 0.2}$	91.9$_{\pm 0.1}$
	AdaptCLIP (Gao et al., 2026)	**92.7**$_{\pm 0.0}$	94.1$_{\pm 0.2}$	88.2$_{\pm 0.1}$	97.2$_{\pm 0.1}$	41.1$_{\pm 0.4}$	46.4$_{\pm 0.5}$	91.6$_{\pm 0.4}$
	Res^{2}CLIP* (Ours)	89.4$_{\pm 0.1}$	90.9$_{\pm 0.2}$	85.7$_{\pm 0.0}$	97.0$_{\pm 0.0}$	35.9$_{\pm 1.0}$	42.1$_{\pm 0.9}$	89.5$_{\pm 0.3}$
	Res^{2}CLIP† (Ours)	90.5$_{\pm 0.4}$	91.9$_{\pm 0.1}$	86.5$_{\pm 0.5}$	**97.8**$_{\pm 0.0}$	**42.5**$_{\pm 0.5}$	**47.6**$_{\pm 0.5}$	**92.7**$_{\pm 0.3}$
4	WinCLIP+ (Jeong et al., 2023)	84.4$_{\pm 1.0}$	85.6$_{\pm 1.6}$	82.9$_{\pm 0.9}$	95.2$_{\pm 0.1}$	17.9$_{\pm 1.2}$	25.4$_{\pm 1.1}$	81.3$_{\pm 0.2}$
	APRIL-GAN (Chen et al., 2023)	92.8$_{\pm 0.2}$	94.6$_{\pm 0.2}$	88.3$_{\pm 0.2}$	96.2$_{\pm 0.0}$	31.8$_{\pm 0.0}$	39.5$_{\pm 0.3}$	90.1$_{\pm 0.1}$
	AnomalyCLIP+ (Zhou et al., 2023)	81.7$_{\pm 0.7}$	84.0$_{\pm 0.7}$	81.3$_{\pm 0.6}$	96.5$_{\pm 0.1}$	29.3$_{\pm 0.1}$	38.0$_{\pm 0.1}$	89.7$_{\pm 0.2}$
	ReMP-AD (Ma et al., 2025a)	**93.3**$_{\pm 0.3}$	**94.8**$_{\pm 0.3}$	**89.0**$_{\pm 0.2}$	96.9$_{\pm 0.1}$	34.8$_{\pm 0.2}$	42.5$_{\pm 0.2}$	91.9$_{\pm 0.1}$
	AdaptCLIP (Gao et al., 2026)	**93.3**$_{\pm 0.3}$	94.5$_{\pm 0.3}$	88.8$_{\pm 0.3}$	97.3$_{\pm 0.1}$	42.0$_{\pm 0.4}$	47.2$_{\pm 0.3}$	92.0$_{\pm 0.2}$
	Res^{2}CLIP* (Ours)	90.1$_{\pm 0.6}$	91.1$_{\pm 0.8}$	86.5$_{\pm 0.3}$	97.3$_{\pm 0.0}$	37.2$_{\pm 0.5}$	43.6$_{\pm 0.5}$	90.3$_{\pm 0.1}$
	Res^{2}CLIP† (Ours)	91.1$_{\pm 0.3}$	92.2$_{\pm 0.2}$	87.3$_{\pm 0.3}$	**98.0**$_{\pm 0.0}$	**42.9**$_{\pm 0.3}$	**48.2**$_{\pm 0.2}$	**92.9**$_{\pm 0.1}$

Table 8: Full quantitative results on BTAD (Mishra et al., 2021) across 1, 2, and 4 shots. All metrics are reported as mean ± standard deviation over three random runs. The best results are highlighted in **bold** and the second-best results are underlined.

Shots	Method	I-AUC	I-AP	I-F1	P-AUC	P-AP	P-F1	PRO
1	WinCLIP+ (Jeong et al., 2023)	85.0$_{\pm 3.7}$	80.6$_{\pm 4.4}$	78.2$_{\pm 3.9}$	95.6$_{\pm 0.4}$	42.0$_{\pm 1.5}$	47.4$_{\pm 1.6}$	66.7$_{\pm 1.4}$
	APRIL-GAN (Chen et al., 2023)	91.7$_{\pm 1.2}$	94.7$_{\pm 1.5}$	91.3$_{\pm 0.9}$	94.1$_{\pm 0.2}$	50.1$_{\pm 0.5}$	53.2$_{\pm 0.5}$	78.6$_{\pm 1.1}$
	AnomalyCLIP+ (Zhou et al., 2023)	92.7$_{\pm 0.9}$	93.8$_{\pm 1.2}$	89.6$_{\pm 1.5}$	96.6$_{\pm 0.3}$	58.4$_{\pm 0.9}$	57.6$_{\pm 0.6}$	79.4$_{\pm 0.6}$
	ReMP-AD (Ma et al., 2025a)	94.9$_{\pm 0.8}$	96.1$_{\pm 1.1}$	92.1$_{\pm 1.5}$	95.5$_{\pm 0.3}$	53.9$_{\pm 1.0}$	56.7$_{\pm 0.7}$	82.3$_{\pm 1.3}$
	AdaptCLIP (Gao et al., 2026)	93.1$_{\pm 0.6}$	96.2$_{\pm 1.4}$	91.7$_{\pm 2.0}$	96.7$_{\pm 0.3}$	61.4$_{\pm 0.9}$	59.3$_{\pm 0.5}$	77.7$_{\pm 0.8}$
	Res^{2}CLIP* (Ours)	95.9$_{\pm 0.3}$	95.0$_{\pm 1.8}$	89.9$_{\pm 2.4}$	97.5$_{\pm 0.3}$	62.9$_{\pm 1.6}$	60.3$_{\pm 1.2}$	81.3$_{\pm 0.6}$
	Res^{2}CLIP† (Ours)	**96.0**$_{\pm 0.2}$	**97.1**$_{\pm 0.8}$	**93.4**$_{\pm 1.7}$	**97.9**$_{\pm 0.2}$	**64.8**$_{\pm 1.4}$	**62.0**$_{\pm 1.1}$	**82.7**$_{\pm 0.9}$
2	WinCLIP+ (Jeong et al., 2023)	85.7$_{\pm 0.5}$	85.3$_{\pm 1.4}$	80.4$_{\pm 1.9}$	95.8$_{\pm 0.1}$	43.3$_{\pm 1.5}$	49.1$_{\pm 1.1}$	66.7$_{\pm 0.6}$
	APRIL-GAN (Chen et al., 2023)	91.9$_{\pm 0.7}$	95.2$_{\pm 0.7}$	91.6$_{\pm 0.3}$	94.2$_{\pm 0.1}$	50.6$_{\pm 0.4}$	53.5$_{\pm 0.3}$	78.3$_{\pm 0.7}$
	AnomalyCLIP+ (Zhou et al., 2023)	92.5$_{\pm 0.3}$	93.4$_{\pm 0.6}$	89.0$_{\pm 1.3}$	96.6$_{\pm 0.1}$	59.1$_{\pm 0.5}$	58.1$_{\pm 0.5}$	79.0$_{\pm 0.4}$
	ReMP-AD (Ma et al., 2025a)	95.2$_{\pm 0.5}$	96.8$_{\pm 0.4}$	92.3$_{\pm 0.3}$	95.4$_{\pm 0.1}$	54.4$_{\pm 0.3}$	57.2$_{\pm 0.3}$	82.3$_{\pm 0.6}$
	AdaptCLIP (Gao et al., 2026)	92.9$_{\pm 0.7}$	96.9$_{\pm 0.7}$	93.3$_{\pm 1.3}$	96.7$_{\pm 0.1}$	61.7$_{\pm 0.6}$	59.3$_{\pm 0.4}$	77.9$_{\pm 0.4}$
	Res^{2}CLIP* (Ours)	95.9$_{\pm 0.3}$	96.1$_{\pm 0.9}$	91.1$_{\pm 1.7}$	97.6$_{\pm 0.1}$	63.6$_{\pm 0.5}$	60.8$_{\pm 0.3}$	81.2$_{\pm 0.6}$
	Res^{2}CLIP† (Ours)	**96.5**$_{\pm 0.1}$	**98.0**$_{\pm 0.4}$	**94.5**$_{\pm 0.9}$	**98.1**$_{\pm 0.1}$	**66.0**$_{\pm 0.7}$	**62.7**$_{\pm 0.4}$	**83.2**$_{\pm 0.8}$
4	WinCLIP+ (Jeong et al., 2023)	86.4$_{\pm 1.5}$	85.4$_{\pm 3.9}$	83.6$_{\pm 1.0}$	95.9$_{\pm 0.0}$	45.0$_{\pm 0.4}$	50.5$_{\pm 0.2}$	67.0$_{\pm 0.4}$
	APRIL-GAN (Chen et al., 2023)	91.8$_{\pm 0.6}$	94.7$_{\pm 0.4}$	91.1$_{\pm 0.3}$	94.4$_{\pm 0.2}$	51.1$_{\pm 0.1}$	53.9$_{\pm 0.2}$	78.5$_{\pm 0.8}$
	AnomalyCLIP+ (Zhou et al., 2023)	92.4$_{\pm 0.2}$	93.4$_{\pm 0.5}$	88.8$_{\pm 0.7}$	96.7$_{\pm 0.1}$	59.3$_{\pm 0.1}$	58.2$_{\pm 0.2}$	79.1$_{\pm 0.2}$
	ReMP-AD (Ma et al., 2025a)	94.4$_{\pm 0.5}$	96.4$_{\pm 0.7}$	91.9$_{\pm 1.0}$	95.5$_{\pm 0.2}$	55.6$_{\pm 0.1}$	57.8$_{\pm 0.1}$	82.0$_{\pm 0.8}$
	AdaptCLIP (Gao et al., 2026)	93.1$_{\pm 0.1}$	96.7$_{\pm 0.3}$	92.6$_{\pm 0.6}$	96.8$_{\pm 0.1}$	62.8$_{\pm 0.2}$	60.0$_{\pm 0.1}$	78.3$_{\pm 0.1}$
	Res^{2}CLIP* (Ours)	96.5$_{\pm 0.4}$	96.5$_{\pm 0.7}$	92.4$_{\pm 0.9}$	97.6$_{\pm 0.1}$	64.6$_{\pm 0.2}$	61.5$_{\pm 0.0}$	81.3$_{\pm 0.3}$
	Res^{2}CLIP† (Ours)	**96.7**$_{\pm 0.4}$	**98.2**$_{\pm 0.4}$	**95.1**$_{\pm 0.8}$	**98.1**$_{\pm 0.1}$	**66.7**$_{\pm 0.4}$	**63.2**$_{\pm 0.4}$	**82.9**$_{\pm 0.5}$

Table 9: Full quantitative results on MPDD (Jezek et al., 2021) across 1, 2, and 4 shots. All metrics are reported as mean ± standard deviation over three random runs. The best results are highlighted in **bold** and the second-best results are underlined.

Shots	Method	I-AUC	I-AP	I-F1	P-AUC	P-AP	P-F1	PRO
1	WinCLIP+ (Jeong et al., 2023)	69.1$_{\pm 0.8}$	73.7$_{\pm 1.0}$	81.1$_{\pm 0.4}$	96.2$_{\pm 0.3}$	30.7$_{\pm 0.6}$	32.6$_{\pm 0.9}$	88.4$_{\pm 1.2}$
	APRIL-GAN (Chen et al., 2023)	82.6$_{\pm 1.5}$	84.9$_{\pm 1.5}$	84.4$_{\pm 3.2}$	96.7$_{\pm 0.0}$	35.6$_{\pm 0.5}$	40.5$_{\pm 0.7}$	91.2$_{\pm 0.1}$
	AnomalyCLIP+ (Zhou et al., 2023)	79.7$_{\pm 2.1}$	80.5$_{\pm 2.3}$	82.1$_{\pm 0.9}$	97.4$_{\pm 0.1}$	36.8$_{\pm 0.7}$	41.5$_{\pm 0.6}$	93.5$_{\pm 0.3}$
	ReMP-AD (Ma et al., 2025a)	82.8$_{\pm 1.5}$	85.4$_{\pm 1.4}$	83.0$_{\pm 1.0}$	97.6$_{\pm 0.1}$	**38.9**$_{\pm 0.9}$	**42.9**$_{\pm 1.2}$	93.7$_{\pm 0.2}$
	AdaptCLIP (Gao et al., 2026)	84.1$_{\pm 0.3}$	85.7$_{\pm 0.6}$	85.1$_{\pm 1.5}$	97.7$_{\pm 0.1}$	34.7$_{\pm 2.0}$	37.9$_{\pm 2.0}$	93.7$_{\pm 0.3}$
	Res^{2}CLIP* (Ours)	**85.6**$_{\pm 0.9}$	**87.8**$_{\pm 0.8}$	**85.2**$_{\pm 0.8}$	97.3$_{\pm 0.0}$	32.4$_{\pm 1.2}$	36.3$_{\pm 1.3}$	92.5$_{\pm 0.4}$
	Res^{2}CLIP† (Ours)	83.6$_{\pm 0.9}$	85.2$_{\pm 0.6}$	82.7$_{\pm 1.4}$	**97.9**$_{\pm 0.1}$	36.6$_{\pm 1.3}$	39.9$_{\pm 1.2}$	**93.8**$_{\pm 0.3}$
2	WinCLIP+ (Jeong et al., 2023)	70.1$_{\pm 0.7}$	74.0$_{\pm 0.9}$	81.9$_{\pm 0.6}$	95.1$_{\pm 1.0}$	31.9$_{\pm 0.6}$	33.7$_{\pm 0.6}$	89.5$_{\pm 0.3}$
	APRIL-GAN (Chen et al., 2023)	84.6$_{\pm 0.9}$	89.2$_{\pm 1.1}$	85.3$_{\pm 1.2}$	96.7$_{\pm 0.1}$	36.9$_{\pm 0.3}$	41.5$_{\pm 0.7}$	91.3$_{\pm 0.2}$
	AnomalyCLIP+ (Zhou et al., 2023)	78.1$_{\pm 0.9}$	82.4$_{\pm 6.8}$	82.6$_{\pm 0.6}$	97.5$_{\pm 0.1}$	37.6$_{\pm 0.5}$	42.3$_{\pm 0.6}$	94.1$_{\pm 0.1}$
	ReMP-AD (Ma et al., 2025a)	**87.0**$_{\pm 1.2}$	**89.5**$_{\pm 0.8}$	86.9$_{\pm 0.6}$	97.7$_{\pm 0.1}$	**41.4**$_{\pm 0.7}$	**45.3**$_{\pm 1.0}$	93.8$_{\pm 0.2}$
	AdaptCLIP (Gao et al., 2026)	85.4$_{\pm 1.1}$	87.1$_{\pm 1.6}$	86.2$_{\pm 0.3}$	97.8$_{\pm 0.1}$	35.7$_{\pm 0.1}$	37.8$_{\pm 0.3}$	**94.3**$_{\pm 0.1}$
	Res^{2}CLIP* (Ours)	**87.0**$_{\pm 0.2}$	**89.5**$_{\pm 1.0}$	**87.3**$_{\pm 0.5}$	97.5$_{\pm 0.2}$	34.3$_{\pm 0.2}$	38.1$_{\pm 0.4}$	93.3$_{\pm 0.2}$
	Res^{2}CLIP† (Ours)	85.0$_{\pm 0.9}$	86.8$_{\pm 1.1}$	85.3$_{\pm 1.1}$	**98.0**$_{\pm 0.0}$	37.8$_{\pm 0.2}$	41.1$_{\pm 0.2}$	94.0$_{\pm 0.1}$
4	WinCLIP+ (Jeong et al., 2023)	71.8$_{\pm 1.3}$	75.9$_{\pm 0.9}$	82.5$_{\pm 0.2}$	94.3$_{\pm 0.2}$	32.2$_{\pm 0.3}$	33.8$_{\pm 0.3}$	88.5$_{\pm 0.6}$
	APRIL-GAN (Chen et al., 2023)	85.9$_{\pm 0.5}$	90.1$_{\pm 1.1}$	86.4$_{\pm 0.2}$	96.7$_{\pm 0.0}$	37.3$_{\pm 0.3}$	41.6$_{\pm 0.4}$	91.1$_{\pm 0.2}$
	AnomalyCLIP+ (Zhou et al., 2023)	78.4$_{\pm 0.1}$	79.4$_{\pm 0.5}$	82.7$_{\pm 0.5}$	97.5$_{\pm 0.1}$	38.6$_{\pm 0.3}$	43.2$_{\pm 0.1}$	94.2$_{\pm 0.2}$
	ReMP-AD (Ma et al., 2025a)	**89.6**$_{\pm 0.3}$	**91.8**$_{\pm 0.3}$	87.5$_{\pm 0.3}$	97.7$_{\pm 0.0}$	**43.1**$_{\pm 0.2}$	**47.0**$_{\pm 0.6}$	93.6$_{\pm 0.1}$
	AdaptCLIP (Gao et al., 2026)	86.3$_{\pm 0.4}$	87.8$_{\pm 0.8}$	87.6$_{\pm 0.5}$	98.1$_{\pm 0.1}$	39.5$_{\pm 1.1}$	41.8$_{\pm 0.8}$	94.9$_{\pm 0.1}$
	Res^{2}CLIP* (Ours)	88.1$_{\pm 0.9}$	90.9$_{\pm 0.9}$	**87.9**$_{\pm 0.7}$	97.7$_{\pm 0.0}$	38.0$_{\pm 1.5}$	42.5$_{\pm 1.5}$	94.2$_{\pm 0.0}$
	Res^{2}CLIP† (Ours)	87.2$_{\pm 1.7}$	89.5$_{\pm 1.8}$	87.6$_{\pm 1.6}$	**98.2**$_{\pm 0.1}$	41.2$_{\pm 1.7}$	44.7$_{\pm 1.9}$	**95.1**$_{\pm 0.2}$

Table 10: Full quantitative results on DTD-Synthetic (Aota et al., 2023) across 1, 2, and 4 shots. All metrics are reported as mean ± standard deviation over three random runs. The best results are highlighted in **bold** and the second-best results are underlined.

Shots	Method	I-AUC	I-AP	I-F1	P-AUC	P-AP	P-F1	PRO
1	WinCLIP+ (Jeong et al., 2023)	$98.1_{\pm 0.2}$	$99.1_{\pm 0.1}$	$96.9_{\pm 0.1}$	$96.8_{\pm 0.1}$	$48.8_{\pm 0.4}$	$52.0_{\pm 0.4}$	$90.3_{\pm 0.1}$
	APRIL-GAN (Chen et al., 2023)	$98.4_{\pm 0.1}$	$99.4_{\pm 0.0}$	$97.6_{\pm 0.1}$	$96.5_{\pm 0.1}$	$76.5_{\pm 0.1}$	$73.9_{\pm 0.1}$	$92.0_{\pm 0.2}$
	AnomalyCLIP+ (Zhou et al., 2023)	$95.3_{\pm 0.6}$	$98.1_{\pm 0.3}$	$94.5_{\pm 0.4}$	$96.3_{\pm 0.2}$	$70.6_{\pm 0.6}$	$68.0_{\pm 0.4}$	$91.5_{\pm 0.3}$
	ReMP-AD (Ma et al., 2025a)	$\mathbf{99.3}_{\pm 0.1}$	$\mathbf{99.7}_{\pm 0.1}$	$\mathbf{98.7}_{\pm 0.2}$	$96.6_{\pm 0.2}$	$\mathbf{79.2}_{\pm 0.4}$	$\mathbf{76.3}_{\pm 0.3}$	$92.5_{\pm 0.4}$
	AdaptCLIP (Gao et al., 2026)	$98.0_{\pm 0.2}$	$99.3_{\pm 0.1}$	$97.4_{\pm 0.3}$	$\mathbf{97.3}_{\pm 0.2}$	$76.9_{\pm 0.3}$	$72.5_{\pm 0.2}$	$93.1_{\pm 0.2}$
	Res^{2}CLIP* (Ours)	$94.3_{\pm 0.3}$	$97.5_{\pm 0.2}$	$93.3_{\pm 0.1}$	$93.5_{\pm 0.1}$	$66.6_{\pm 0.8}$	$66.7_{\pm 0.7}$	$86.6_{\pm 0.5}$
	Res^{2}CLIP† (Ours)	$97.0_{\pm 0.1}$	$98.9_{\pm 0.1}$	$95.8_{\pm 0.1}$	$97.2_{\pm 0.2}$	$77.5_{\pm 0.4}$	$74.2_{\pm 0.2}$	$\mathbf{93.6}_{\pm 0.3}$
2	WinCLIP+ (Jeong et al., 2023)	$98.2_{\pm 0.1}$	$99.2_{\pm 0.1}$	$97.2_{\pm 0.2}$	$96.8_{\pm 0.1}$	$49.3_{\pm 0.3}$	$52.3_{\pm 0.3}$	$90.4_{\pm 0.3}$
	APRIL-GAN (Chen et al., 2023)	$98.4_{\pm 0.2}$	$99.4_{\pm 0.1}$	$97.8_{\pm 0.2}$	$96.7_{\pm 0.0}$	$76.9_{\pm 0.1}$	$74.2_{\pm 0.0}$	$92.2_{\pm 0.1}$
	AnomalyCLIP+ (Zhou et al., 2023)	$95.0_{\pm 0.6}$	$98.0_{\pm 0.2}$	$94.6_{\pm 0.1}$	$96.7_{\pm 0.0}$	$71.2_{\pm 0.2}$	$68.4_{\pm 0.1}$	$91.7_{\pm 0.1}$
	ReMP-AD (Ma et al., 2025a)	$\mathbf{99.4}_{\pm 0.1}$	$\mathbf{99.7}_{\pm 0.0}$	$\mathbf{98.8}_{\pm 0.1}$	$97.0_{\pm 0.0}$	$\mathbf{80.7}_{\pm 0.2}$	$\mathbf{77.3}_{\pm 0.1}$	$93.1_{\pm 0.1}$
	AdaptCLIP (Gao et al., 2026)	$98.2_{\pm 0.2}$	$99.3_{\pm 0.1}$	$97.5_{\pm 0.2}$	$\mathbf{97.7}_{\pm 0.2}$	$78.0_{\pm 0.3}$	$73.3_{\pm 0.1}$	$93.2_{\pm 0.2}$
	Res^{2}CLIP* (Ours)	$95.0_{\pm 0.3}$	$98.0_{\pm 0.1}$	$94.2_{\pm 0.3}$	$94.5_{\pm 0.0}$	$69.2_{\pm 0.1}$	$68.8_{\pm 0.2}$	$87.7_{\pm 0.3}$
	Res^{2}CLIP† (Ours)	$97.3_{\pm 0.1}$	$99.0_{\pm 0.0}$	$96.4_{\pm 0.1}$	$97.6_{\pm 0.0}$	$78.5_{\pm 0.1}$	$74.9_{\pm 0.2}$	$\mathbf{94.2}_{\pm 0.2}$
4	WinCLIP+ (Jeong et al., 2023)	$98.3_{\pm 0.1}$	$99.2_{\pm 0.1}$	$97.1_{\pm 0.1}$	$97.0_{\pm 0.1}$	$50.0_{\pm 0.2}$	$52.7_{\pm 0.2}$	$90.7_{\pm 0.1}$
	APRIL-GAN (Chen et al., 2023)	$98.6_{\pm 0.1}$	$99.5_{\pm 0.1}$	$97.8_{\pm 0.2}$	$96.8_{\pm 0.1}$	$77.3_{\pm 0.1}$	$74.5_{\pm 0.1}$	$92.5_{\pm 0.1}$
	AnomalyCLIP+ (Zhou et al., 2023)	$95.3_{\pm 0.2}$	$98.1_{\pm 0.1}$	$94.6_{\pm 0.2}$	$96.9_{\pm 0.1}$	$71.8_{\pm 0.1}$	$68.7_{\pm 0.1}$	$92.2_{\pm 0.1}$
	ReMP-AD (Ma et al., 2025a)	$\mathbf{99.4}_{\pm 0.1}$	$\mathbf{99.7}_{\pm 0.1}$	$\mathbf{98.9}_{\pm 0.1}$	$97.1_{\pm 0.1}$	$\mathbf{81.2}_{\pm 0.1}$	$\mathbf{77.7}_{\pm 0.1}$	$93.5_{\pm 0.1}$
	AdaptCLIP (Gao et al., 2026)	$98.2_{\pm 0.1}$	$99.3_{\pm 0.1}$	$97.7_{\pm 0.3}$	$97.8_{\pm 0.1}$	$78.4_{\pm 0.2}$	$73.8_{\pm 0.2}$	$93.6_{\pm 0.1}$
	Res^{2}CLIP* (Ours)	$95.7_{\pm 0.3}$	$98.3_{\pm 0.2}$	$94.8_{\pm 0.2}$	$95.2_{\pm 0.0}$	$70.9_{\pm 0.2}$	$70.2_{\pm 0.2}$	$89.2_{\pm 0.2}$
	Res^{2}CLIP† (Ours)	$97.4_{\pm 0.1}$	$99.0_{\pm 0.0}$	$96.5_{\pm 0.1}$	$\mathbf{97.9}_{\pm 0.0}$	$79.4_{\pm 0.2}$	$75.4_{\pm 0.2}$	$\mathbf{94.6}_{\pm 0.0}$

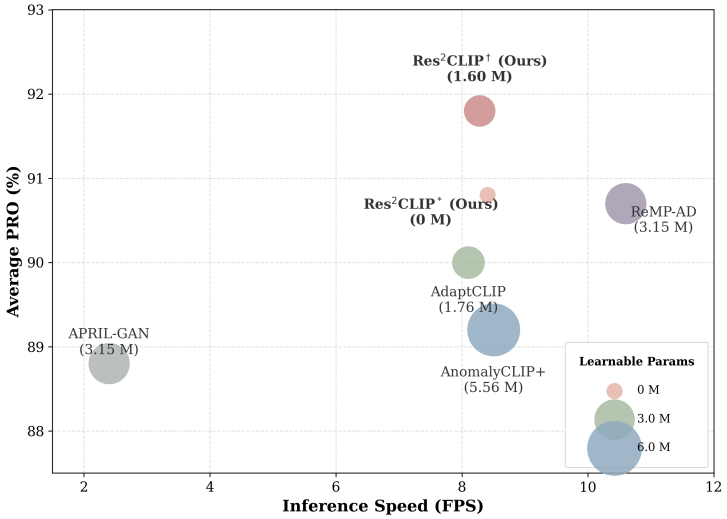


Figure 7: Trade-off between inference speed, localization performance, and parameter efficiency. The bubble size denotes the number of extra learnable parameters (e.g., our Res^{2}CLIP* requires 0M). Our method achieves the optimal balance, delivering state-of-the-art performance with minimal additional parameters and competitive inference speed.

G Additional Ablation Studies

G.1 Normal-Reference Matching

The construction of the local visual residual depends on how the retrieved normal features are used to form the corresponding reference. As defined in the main text, for the i-th query token $\hat{F}^q_{v,l}(i)$, we retrieve its κ nearest neighbors $\mathcal{N}_\kappa(i)$ from the normalized reference bank B_l according to cosine similarity. Here, we investigate two aspects of normal-reference matching: how the retrieved neighbors are weighted and whether their information is aggregated before or after residual scor-

ing. To isolate these design choices from the number of retrieved neighbors, we fix $\kappa = 5$ for all multi-neighbor variants in this experiment.

Feature-Level Aggregation. Our default formulation first aggregates the retrieved normal features into a single matched reference and then constructs the visual residual. For convenience, we denote the cosine similarity between the query token and the j-th retrieved feature as

$$s_{ij} = \hat{F}^q_{v,l}(i)^\top B_l(j), \qquad j \in \mathcal{N}_\kappa(i). \tag{45}$$

Given aggregation weights α_{ij}, the matched normal feature and the resulting visual residual follow the same formulation as in the main manuscript:

$$F_{\mathrm{ref},l}(i) = \sum_{j \in \mathcal{N}_\kappa(i)} \alpha_{ij} B_l(j), \qquad \mathbf{R}_{v,l}(i) = \hat{F}^q_{v,l}(i) - \hat{F}_{\mathrm{ref},l}(i), \tag{46}$$

where $\hat{F}_{\mathrm{ref},l}(i)$ denotes the ℓ_2-normalized matched reference feature.

We compare four variants. *1-NN* directly uses the nearest normal feature, corresponding to $\kappa = 1$. For multi-neighbor matching, *Mean* uniformly averages the retrieved features with $\alpha_{ij} = 1/\kappa$. *Soft* instead assigns similarity-dependent weights:

$$\alpha_{ij} = \frac{\exp\left(s_{ij}/\tau_{\mathrm{nn}}\right)}{\sum_{r \in \mathcal{N}_\kappa(i)} \exp\left(s_{ir}/\tau_{\mathrm{nn}}\right)}, \tag{47}$$

where $\tau_{\mathrm{nn}} = 0.05$ is fixed for this comparison. We additionally evaluate *Sparsemax*, which replaces the softmax normalization with sparsemax (Martins & Astudillo, 2016) over the same retrieved neighbors.

Score-Level Aggregation. We further examine whether aggregation should instead be performed after residual scoring. Rather than constructing a single matched normal feature, each retrieved neighbor independently forms a visual residual:

$$\mathbf{R}^{(j)}_{v,l}(i) = \hat{F}^q_{v,l}(i) - B_l(j), \qquad j \in \mathcal{N}_\kappa(i). \tag{48}$$

Following the residual scoring formulation in the main text, each residual produces

$$S^{(j)}_l(i) = \left\| \mathbf{R}^{(j)}_{v,l}(i) \right\|_2 + \lambda_p \left[\mathbf{R}^{(j)}_{v,l}(i)^\top \hat{\mathbf{R}}_t \right]_+ . \tag{49}$$

The resulting anomaly scores are then aggregated as

$$S_l(i) = \sum_{j \in \mathcal{N}_\kappa(i)} \alpha_{ij} S^{(j)}_l(i), \tag{50}$$

where *Mean* uses $\alpha_{ij} = 1/\kappa$, while *Soft* uses the similarity-based weights in equation (47). Therefore, feature-level aggregation combines normal evidence before residual construction, whereas score-level aggregation first obtains an anomaly response from each retrieved neighbor and then combines these responses.

Table 11 shows that the overall performance is stable across different normal-reference aggregation strategies. Compared with directly using a single nearest neighbor, aggregating multiple normal features improves the average P-AP from 49.7% to 50.5% for mean and soft feature-level aggregation. The differences among mean, softmax, and sparsemax weighting are relatively small, indicating that residual construction is not strongly dependent on a particular weighting function. When using the same weighting scheme, feature-level aggregation yields higher average P-AP than score-level aggregation, suggesting that combining normal evidence before residual construction is more suitable for fine-grained localization. Based on these results, we use soft feature-level aggregation as the default aggregation formulation.

G.2 Hyperparameter Sensitivity

We further investigate the sensitivity of Res^{2}CLIP to three hyperparameters: the number of retrieved neighbors κ, the normal-reference matching temperature τ_{nn}, and the projection weight λ_p. For κ, we evaluate $\{1, 3, 5, 10, 20\}$; for τ_{nn}, we evaluate $\{0.01, 0.03, 0.05, 0.10, 0.20\}$; and for λ_p, we evaluate $\{0.05, 0.1, 0.2, 0.5, 1.0\}$. The sensitivity to κ and τ_{nn} is evaluated under the training-free setting, while λ_p is analyzed for the fine-tune model.

Table 11: Ablation study of normal-reference matching under the training-free setting. Feature-level aggregation combines retrieved normal features before residual construction, whereas score-level aggregation combines anomaly scores obtained from individual retrieved neighbors. For all multi-neighbor variants, $\kappa = 5$ is fixed for a controlled comparison. Results are reported on MVTec AD, VisA, and BTAD.

Aggregation Level		Feature-Level				Score-Level	
Weighting		1-NN	Mean	Soft	Sparsemax	Mean	Soft
MVTec AD	I-AUC	95.4	95.3	95.2	95.4	95.6	95.5
	P-AP	52.5	53.5	53.6	53.4	52.2	52.4
	PRO	88.9	89.1	89.1	89.1	89.3	89.3
VisA	I-AUC	88.4	88.6	88.7	88.4	88.2	88.4
	P-AP	31.2	32.8	32.9	31.7	31.0	31.4
	PRO	87.8	89.0	88.9	88.8	88.5	88.5
BTAD	I-AUC	95.9	96.0	96.0	95.6	96.1	96.1
	P-AP	65.3	64.9	65.0	64.2	65.3	65.3
	PRO	81.6	82.0	81.9	81.8	82.6	82.3
Average	I-AUC	93.2	93.3	93.3	93.1	93.3	93.3
	P-AP	49.7	50.4	50.5	49.8	49.5	49.7
	PRO	86.1	86.7	86.6	86.6	86.8	86.7

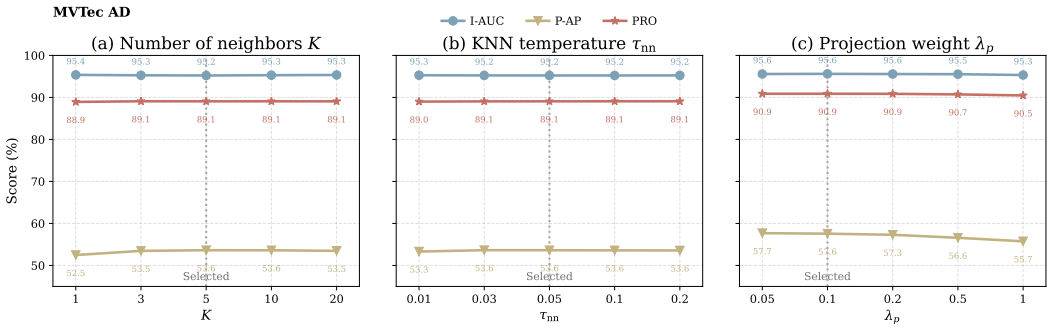


Figure 8: Hyperparameter sensitivity on MVTec AD with respect to κ, τ_{nn}, and λ_p.

As shown in Figs. 8 to 11, the three datasets exhibit generally consistent trends across the tested hyperparameter ranges. The performance remains stable over a broad range of κ and τ_{nn}, indicating that normal-reference matching does not rely on a finely tuned neighborhood size or softmax temperature. The fine-tuned model is also relatively stable under small values of λ_p, while larger projection weights tend to reduce P-AP. Overall, these results show that Res^{2}CLIP is not sensitive to precise hyperparameter tuning. We use $\kappa = 5$ and $\tau_{\text{nn}} = 0.05$ throughout the experiments, and set $\lambda_p = 0.1$ for the fine-tune mode.

G.3 Adapter Sharing Strategy

We further examine whether the adapters in the two optimization stages should be shared across feature levels or instantiated independently for each layer. All other training and inference settings are kept unchanged.

As shown in Table 12, using layer-specific Spatial Adapters in Stage I does not provide a consistent advantage over the shared design. We therefore adopt a shared Spatial Adapter, which maintains comparable performance while using fewer trainable parameters.

For Stage II, layer-specific RT-Adapters provide consistently competitive performance across the three datasets and show a modest overall advantage over the shared variant. We therefore use layer-specific RT-Adapters to allow semantic residual refinement to adapt to different feature levels. Ac-

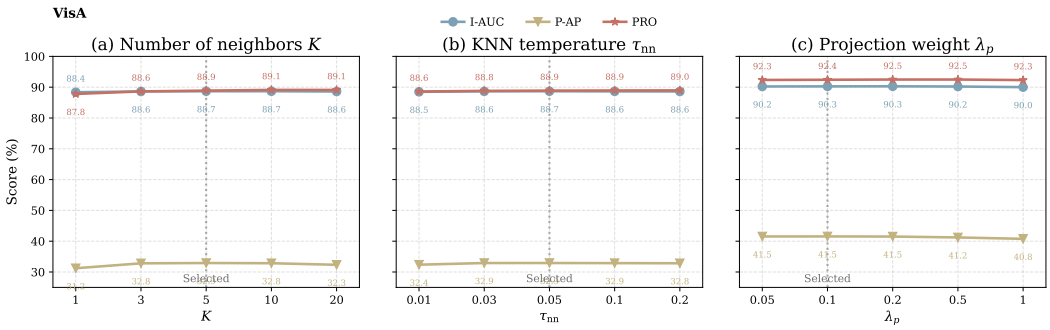


Figure 9: Hyperparameter sensitivity on VisA with respect to κ, τ_{nn}, and λ_p.

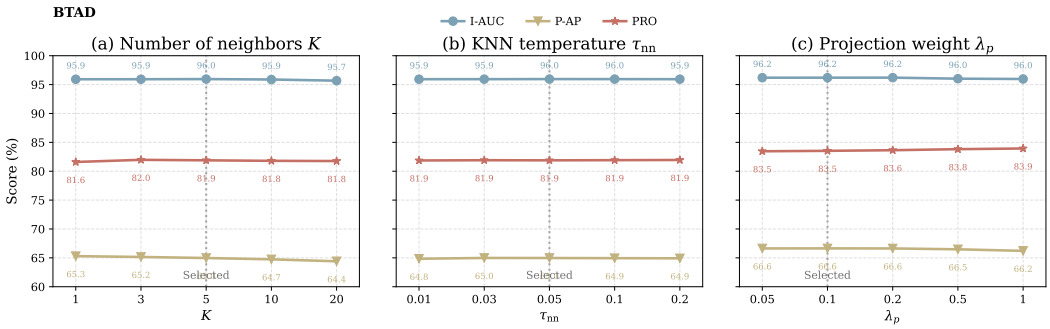


Figure 10: Hyperparameter sensitivity on BTAD with respect to κ, τ_{nn}, and λ_p.

cordingly, the final model adopts a shared Spatial Adapter in Stage I and layer-specific RT-Adapters in Stage II.

Table 12: Ablation study of adapter sharing strategies in the two optimization stages. Stage I compares shared and layer-specific Spatial Adapters, while Stage II compares shared and layer-specific Radial-Tangential Adapters. Results are reported on MVTec AD, VisA, and BTAD.

(a) Stage-I Spatial Adapter

Sharing Strategy		Shared	Layer-Specific
MVTec AD	I-AUC	95.6	96.0
	P-AP	55.4	55.6
	PRO	90.8	90.8
VisA	I-AUC	89.9	90.4
	P-AP	38.1	38.8
	PRO	92.1	92.2
BTAD	I-AUC	96.3	96.5
	P-AP	64.8	63.0
	PRO	82.6	82.6
Average	I-AUC	94.0	94.3
	P-AP	52.8	52.4
	PRO	88.5	88.5

(b) Stage-II RT-Adapter

Sharing Strategy		Shared	Layer-Specific
MVTec AD	I-AUC	95.4	95.6
	P-AP	57.5	57.6
	PRO	90.8	90.9
VisA	I-AUC	89.9	90.3
	P-AP	41.2	41.5
	PRO	92.3	92.4
BTAD	I-AUC	96.0	96.2
	P-AP	65.4	66.6
	PRO	83.3	83.5
Average	I-AUC	93.8	94.0
	P-AP	54.7	55.2
	PRO	88.8	88.9

G.4 Sequential vs. Joint Optimization

We further compare the proposed sequential optimization with joint training, where the Spatial Adapter and RT-Adapters are optimized simultaneously for 20 epochs. As shown in Table 13, joint

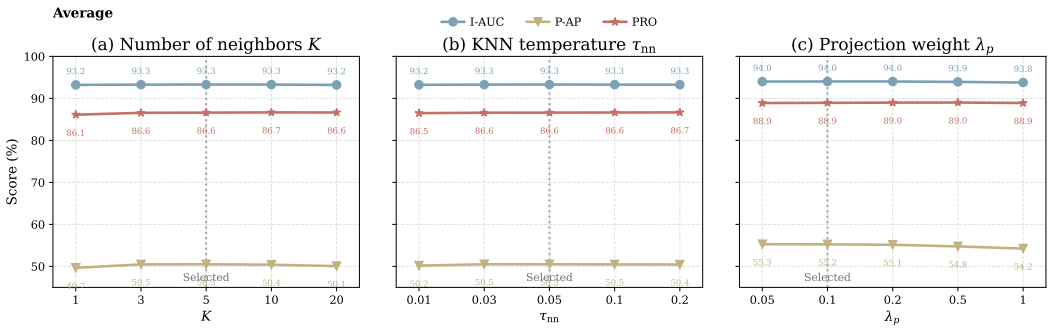


Figure 11: Average hyperparameter sensitivity across MVTec AD, VisA, and BTAD with respect to κ, τ_{nn}, and λ_p.

optimization achieves comparable I-AUC and P-AP, but consistently yields lower PRO across all three datasets. Notably, the RT-Adapters are optimized for 20 epochs under joint training, compared with 10 epochs in the sequential setting. These results suggest that performing semantic refinement on a fixed, spatially calibrated residual representation provides more stable regional localization, motivating the sequential optimization adopted in our final model.

Table 13: Comparison between joint and sequential optimization of the two stages.

Optimization Strategy		Joint	Sequential
MVTec AD	I-AUC	95.0	95.6
	P-AP	58.1	57.6
	PRO	90.2	90.9
VisA	I-AUC	90.4	90.3
	P-AP	41.1	41.5
	PRO	91.4	92.4
BTAD	I-AUC	95.9	96.2
	P-AP	67.0	66.6
	PRO	82.9	83.5
Average	I-AUC	93.8	94.0
	P-AP	55.4	55.2
	PRO	88.2	88.9

G.5 Ablation of Radial-Tangential Refinement

Residual Refinement Formulation. We further examine how residual magnitude and direction should be refined in Stage II. All variants use the same Stage-I features, predictor architecture, training objective, and optimization settings, differing only in how the predicted refinement $\mathbf{h}$ is applied. Direct addition $\widetilde{\mathbf{R}} = \mathbf{R} + \mathbf{h}$ jointly changes magnitude and direction, while the direction-only variant $\widetilde{\mathbf{R}} = m\ \mathrm{Norm}(\hat{\mathbf{R}} + \mathbf{h}^{\perp})$ keeps the residual magnitude fixed. The proposed radial–tangential formulation instead updates the magnitude through $\widetilde{m} = \max(\epsilon, m + a)$ and the direction through $\mathrm{Norm}(\hat{\mathbf{R}} + \mathbf{h}^{\perp})$, thereby explicitly separating the two types of refinement.

Direction-only refinement does not consistently improve over direct addition, indicating that fixing the residual magnitude throughout Stage II is overly restrictive. In contrast, radial–tangential refinement increases the average P-AP from 53.6% to 55.2% over direct addition, while maintaining comparable I-AUC and PRO. Moreover, its P-AP improvement is consistent across all three datasets. These results support separately parameterizing radial and directional adjustments for semantic residual refinement.

Table 14: Ablation study of magnitude and directional refinement in Stage II. Results are reported on MVTec AD, VisA, and BTAD.

Residual Update		$\mathbf{R}+\mathbf{h}$	$m\ \mathrm{Norm}(\hat{\mathbf{R}}+\mathbf{h}^{\perp})$	$\widetilde{m}\ \mathrm{Norm}(\hat{\mathbf{R}}+\mathbf{h}^{\perp})$
Magnitude		Coupled	Fixed	Explicit
Direction		Coupled	Explicit	Explicit
MVTec AD	I-AUC	95.5	95.1	95.6
	P-AP	56.0	55.5	57.6
	PRO	90.4	90.7	90.9
VisA	I-AUC	90.3	90.1	90.3
	P-AP	39.2	40.1	41.5
	PRO	92.5	92.2	92.4
BTAD	I-AUC	95.7	96.0	96.2
	P-AP	65.5	63.7	66.6
	PRO	84.3	82.5	83.5
Average	I-AUC	93.9	93.8	94.0
	P-AP	53.6	53.1	55.2
	PRO	89.0	88.4	88.9

H Additional Qualitative Visualizations

H.1 Residual Feature Distributions

To delve deeper into the mechanism of residual representations in multimodal alignment, we conduct comprehensive t-SNE visualizations of the feature space before and after residual extraction.

Reduction of Category-Dependent Structure. To visualize normal and anomalous textual states separately, we construct visualization-only text residuals by subtracting a neutral text feature, obtained without state-specific descriptors, from the corresponding normal or anomalous text feature. This construction is used only for the analysis in this section and does not alter the text residual $\mathbf{R}_t$ used in the proposed method.

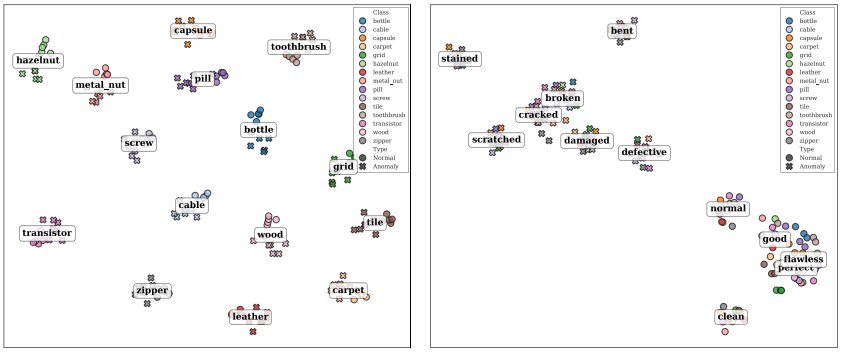
As in Figs. 12 and 13, the original CLIP feature space exhibits pronounced category-dependent structure, where features from the same object category tend to form distinct clusters. After converting the representations into residual form, such category-wise clustering becomes substantially weaker, and features from different categories are more strongly mixed. These results suggest that residualization reduces the influence of category-specific appearance and semantics, allowing the representation to focus more on deviations relative to the corresponding normal or neutral state.

Cross-Modal Alignment in the Residual Space. Fig. 14 further illustrates the relationship between text and visual residuals across multiple datasets. Notably, even in the training-free setting (Fig. 14, left), anomalous visual residuals already exhibit a clear tendency toward the text residual anchor without any parameter optimization. This observation suggests that the residual formulation naturally exposes a degree of cross-modal correspondence between visual anomaly deviations and the normal-to-anomaly textual deviation.

Effectiveness of Residual Optimization. While the quantitative improvements brought by residual optimization are reported in Section 4.3, Fig. 14 provides further qualitative evidence. Comparing the left and right panels of Fig. 14, it is evident that after optimization in the residual space (w/ Adapter), anomalous visual residuals become more concentrated around the text residual anchor, while the separation between normal and anomalous residuals becomes clearer.

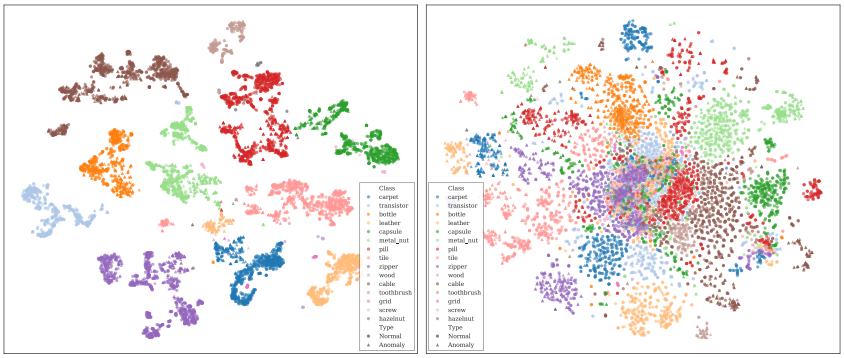
H.2 Visualization of Residual Scoring

Fig. 15 visualizes the anomaly maps produced by residual magnitude, semantic projection, and their combination. Magnitude and projection exhibit complementary response patterns: the former can highlight anomaly deviations missed by semantic projection, whereas the latter can suppress some



(a) Raw Features (b) Residual Features

Figure 12: t-SNE visualization of text features on MVTecAD.



(a) Raw Features (b) Residual Features

Figure 13: t-SNE visualization of visual features on MVTecAD.

irrelevant responses. Their combination integrates these complementary cues, consistent with the overall results in Table 3.

H.3 Visualization of Two-Stage Optimization

Fig. 16 visualizes the final anomaly maps under the 1-shot setting for the training-free model, after Stage I, and after the full two-stage optimization. A clear progressive refinement can be observed. Stage I mainly suppresses false responses in normal regions, while Stage II further concentrates the responses on anomalous regions. These examples illustrate the progressive effect of the two-stage optimization on the final anomaly maps.

I Limitations and Future Work

Limitations. Res^{2}CLIP still depends on the coverage of the few-shot normal reference set for constructing visual residuals. Although the spatial residual calibration in Stage I can suppress residual responses caused by some normal variations, it cannot fully compensate for normal patterns that are insufficiently represented by the available references. Such variations may therefore remain in the residuals and affect anomaly localization, particularly in the training-free setting. In addition, the current formulation uses a single fixed text residual $\mathbf{R}_t$ to represent the semantic deviation from normal to anomalous, which may not fully capture the diversity of anomaly patterns.

Future Work. In future work, we will explore more robust normal-reference modeling and adaptive reference aggregation to improve residual construction under limited normal coverage. Another promising direction is to extend the fixed text residual to multiple or adaptive semantic residuals,

allowing richer anomaly semantics while preserving the cross-category generalization of the current formulation.

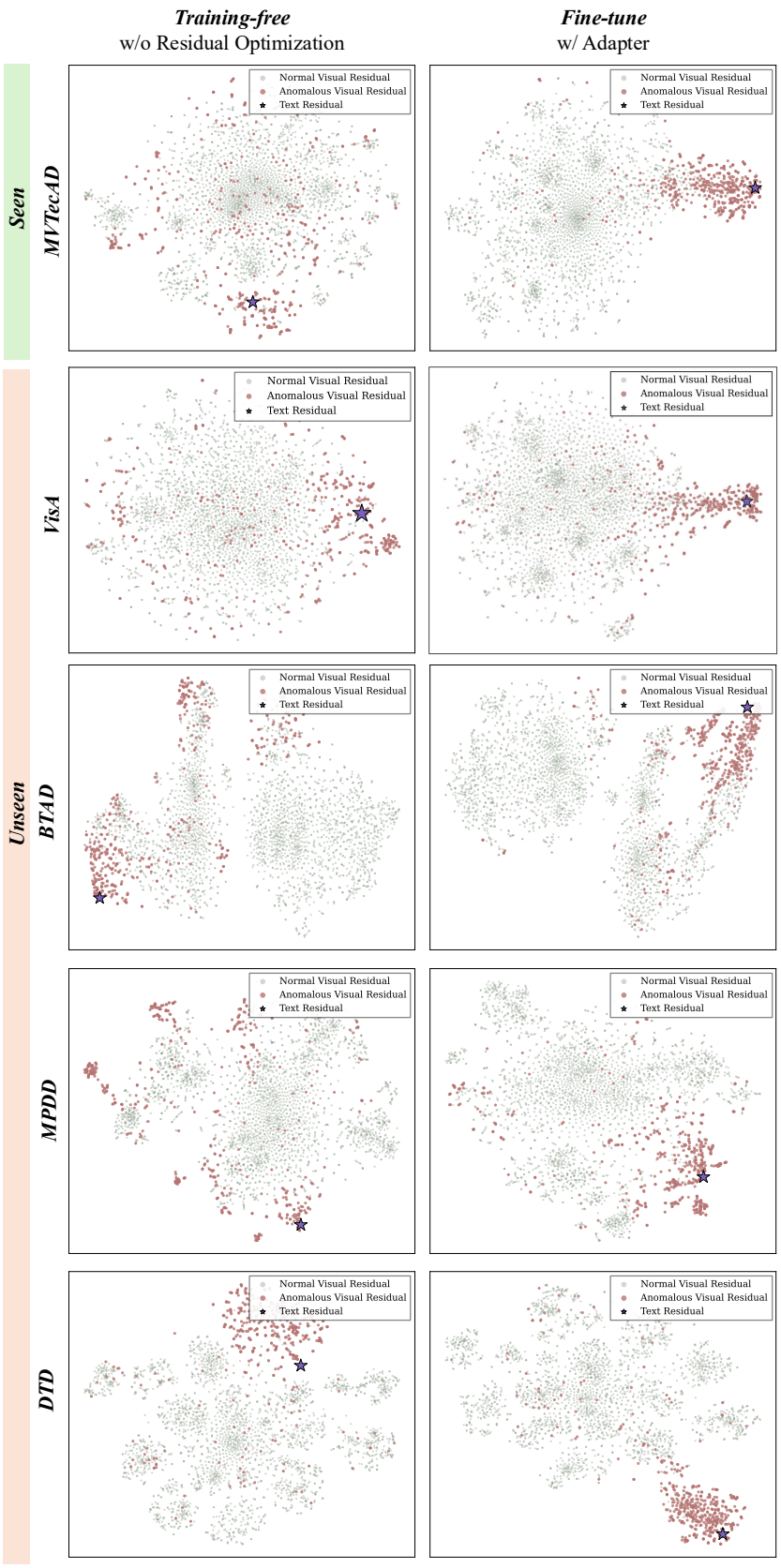


Figure 14: More t-SNE visualizations of residual distributions before and after optimization across multiple datasets.

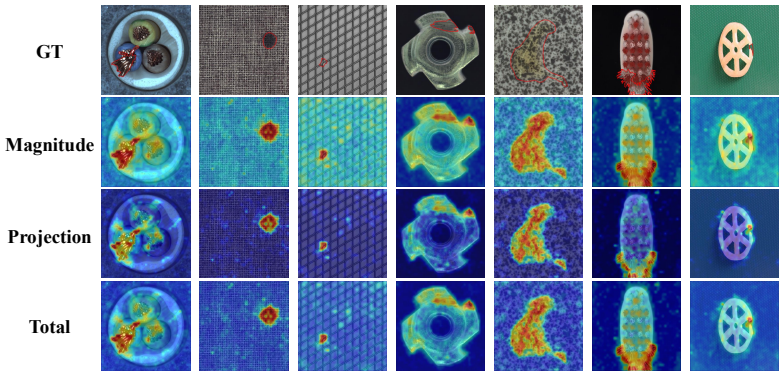


Figure 15: Visualization of anomaly maps produced by residual magnitude, semantic projection, and their combination under the 1-shot setting after fine-tuning. The two scoring terms exhibit complementary response patterns, while their combination integrates both cues.

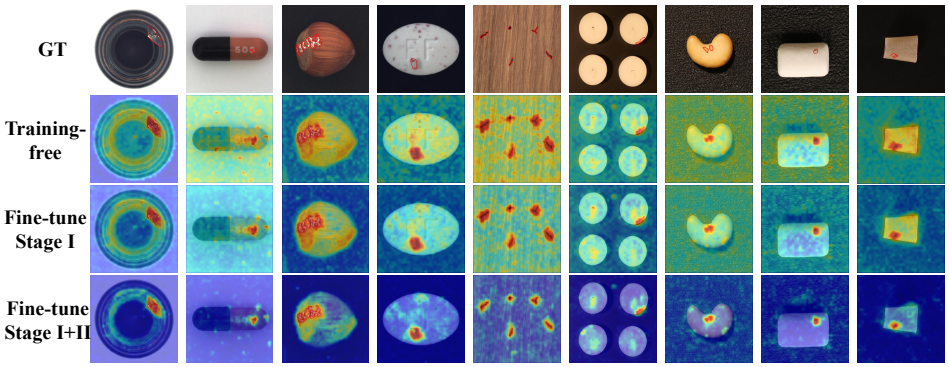


Figure 16: Visualization of the final anomaly maps under the 1-shot setting at different optimization stages. Stage I mainly suppresses false responses in normal regions, while Stage II further concentrates the responses on anomalous regions.