symmetry

MDPI

Article

HiFRA-AD: High-Frequency Residual-Aware Sparse Refinement for Unsupervised Industrial Anomaly Detection

Yangu Chen

School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; 011825110@sues.edu.cn

Abstract

Symmetry and repetitive patterns are common in industrial products, where small defects often disrupt local structural regularity. Fixed-resolution reconstruction can weaken this evidence, while high-resolution cropping changes visual context and anomaly-score distributions. We propose HiFRA-AD, a high-frequency residual-aware sparse refinement framework that couples local feature sensitivity with calibrated evidence aggregation. A fixed-coefficient token-grid residual operator enhances local spatial variation before reconstruction using a frozen DINOv2 encoder. At inference, the global anomaly map selects four peak-suppressed regions for re-evaluation by the shared model. Robust median/quantile alignment and a soft evidence gate convert the local responses into nonnegative residual updates. HiFRA-AD-S achieves P-AP/AUPRO scores of 71.6/95.5, 55.1/96.1, and 46.8/59.8 on MVTec AD, VisA, and MVTec AD 2, respectively. The MVTec AD 2 evaluation uses the official 2045-image TESTpriv partition and AU-PRO@0.30. In backbone-matched MVTec AD comparisons, P-AP improves from 67.5 to 71.6 with ViT-S and from 69.3 to 72.8 with ViT-B. Five paired ViT-S training runs yield a 4.10-point P-AP gain (95% CI [3.95, 4.25]) and a 1.10-point AUPRO gain (95% CI [0.98, 1.22]). With four local regions, ViT-S inference takes 20.8 ms per image on an RTX 4090, compared with 10.0 ms for the global baseline. A defect-subtype analysis records larger paired per-image AP gains for structural disruptions (+4.99 points) than for other anomalies (+2.68 points). These results support selective local refinement for fine-grained industrial anomaly localization.

Keywords: industrial anomaly detection; unsupervised anomaly detection; DINOv2; high-frequency residual; sparse high-resolution refinement; structural symmetry; local asymmetry

Check for updates

Academic Editor: Zhixun Su

Received: 9 October 2026
Revised: 3 November 2026
Accepted: 12 November 2026
Published: 17 November 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Modern manufacturing is increasingly organized around automated, flexible, and data-driven production, where visual inspection is expected to operate continuously with high accuracy and limited human intervention. Industrial defects, however, remain difficult to model because they are rare, heterogeneous, and often impossible to enumerate before deployment. A production line may provide abundant defect-free samples, whereas defective samples are scarce and newly emerging defect types are inherently unknown during training. This setting makes unsupervised anomaly detection (UAD), which learns normality from anomaly-free data and identifies deviations at test time, particularly attractive for industrial quality control. MVTec AD established a widely used benchmark for this paradigm [1], while VisA [2] and Real-IAD [3] extended evaluation toward more complex

structures, larger product diversity, high-resolution imagery, and realistic acquisition conditions. More recently, MVTec AD 2 introduced advanced scenarios involving transparent and overlapping objects, challenging illumination, large normal appearance variation, and extremely small defects [4]. These developments reveal a shift in the central challenge of industrial UAD: strong performance on controlled benchmarks is no longer sufficient; practical inspection increasingly depends on whether subtle defect evidence survives the visual pipeline and remains distinguishable from benign local variation.

Industrial UAD has evolved from class-specific normality modeling toward unified and transferable representations. Feature-distribution and memory-based approaches characterize normal patch embeddings and detect deviations through distance or density estimation [5–10]. Reconstruction and distillation methods instead use discrepancies between observed and reconstructed representations as anomaly evidence [11–15]. The move to multi-class UAD further exposed the identical shortcutproblem, where a sufficiently expressive reconstruction network can also reproduce anomalous inputs [16]. Subsequent work has addressed this issue from complementary directions: SimpleNet introduces feature-space pseudo anomalies [17], EfficientAD combines efficient student–teacher learning with global autoencoding [18], RealNet and GLASS improve anomaly synthesis and discriminative learning [19,20], and MambaAD explores state-space modeling for unified multi-class inspection [21]. Together, these studies mark a clear transition from independent per-category detectors toward one-for-all models built on stronger representations and more general normality-learning objectives.

This transition has been accelerated by self-supervised vision foundation models. DINOv2 provides generic visual features that transfer effectively across image distributions and dense prediction tasks [22], while register tokens improve the spatial behavior of Vision Transformer representations [23]. Building on these advances, Dinomaly shows that a simple Transformer reconstruction framework equipped with a noisy bottleneck, unfocused attention, and loose reconstruction can substantially narrow the gap between unified and class-separated UAD [24]. Dinomaly2 further extends this minimalist principle across multi-class, inference-unified, few-shot, multi-view, and multi-modal settings [25]. Its success suggests that, once a sufficiently strong visual representation is available, a remaining bottleneck is not simply model capacity. More important is how defect-sensitive information is preserved and how reconstruction evidence is interpreted when the spatial scale and visual context of the input change.

Two coupled challenges are central to high-resolution industrial inspection. First, fixed-resolution processing weakens microscopic defect evidence before anomaly scoring.Thin scratches, pinholes, edge chipping, weak cracks, and localized texture corruption often occupy only a few pixels in a multi-megapixel image. Standard UAD pipelines resize or crop the image to a fixed spatial resolution before feature extraction, which attenuates subtle structures. Uniformly increasing global resolution rapidly expands the number of visual tokens and spends most of the additional computation on normal or easy regions. The practical conflict is therefore global structural context versus local defect fidelity under a finite computation budget.

Second, high-resolution local re-evaluation requires explicit alignment with the global prediction. Cropping a candidate region and resizing it to the model input size increases model-coordinate sampling density, while simultaneously changing object context, boundary composition, and local score statistics seen by the reconstruction network. Direct score addition therefore mixes genuine local evidence with crop-induced score shift. A reliable refinement mechanism must answer two questions jointly: where should local re-evaluation be allocated, and how should the resulting local response be calibrated before fusion?

Symmetry and local structural regularity in industrial inspection. Manufactured surfaces exhibit approximate mirror balance, repeated grooves, periodic grids, and regularly spaced printed elements. Interrupted stripes, edge chips, and missing repeated elements disturb these nominal relationships. HiFRA-AD enhances local spatial deviations and evaluates them against learned normal reconstruction across both repetitive textures and irregular objects. Section 3.3 derives the residual operator's relation to geometric transformations.

A related observation concerns spatial frequency. Many small industrial defects manifest as sharp boundaries, narrow discontinuities, or local texture changes. In a token-grid representation, these structures correspond to local spatial variation that is weakened by feature fusion and reconstruction. Increasing decoder capacity strengthens reconstruction of both normal and abnormal patterns, whereas HiFRA-AD preserves the semantic representation of the frozen encoder and directly increases the reconstruction pathway sensitivity to local token variation without changing feature dimensionality or adding a heavy auxiliary network.

Based on these observations, we propose HiFRA-AD, a high-frequency residual-aware anomaly-detection framework with sparse high-resolution refinement. HiFRA-AD is built on the unified DINOv2 reconstruction pipeline of Dinomaly2 and separates feature sensitivity from evidence aggregation. During training, a fixed-coefficient high-frequency residual operator with no additional learnable weights is applied to the fused spatial token field before the bottleneck. The operator preserves spatially constant responses while mildly amplifying token variations that are attenuated by local averaging. During inference, the global anomaly map is first used to select a small number of peak-suppressed candidate regions. Only these regions are mapped back to the original image, cropped, and re-evaluated at higher effective sampling density by the same reconstruction model.

HiFRA-AD first standardizes each local anomaly map with robust within-ROI statistics and aligns it to the location and scale of the corresponding global patch. A continuous evidence gate retains positive residual evidence above the global response, and a center-weighted fusion window combines overlapping updates. The resulting refinement is pointwise non-suppressive by construction: it preserves existing global anomaly evidence and adds only supported local residuals. This formulation gives the global–local fusion a direct mathematical interpretation.

The main contributions of this work are summarized as follows:

1. **We formulate a coupled scale–context problem in high-resolution reconstruction-based UAD.** Fixed-resolution inference can attenuate small defects, whereas local magnification changes reconstruction context and score statistics. This motivates explicit alignment of local and global anomaly scores.
2. We propose HiFRA-AD, which combines high-frequency token enhancement without extra learnable weights with sparse high-resolution re-evaluation. The frequency module strengthens local token variation without modifying the frozen encoder or feature shape, while the global anomaly map allocates local computation only to a small set of suspicious regions instead of performing exhaustive sliding-window inference.
3. We develop a residual-aware local evidence calibration and non-suppressive fusion rule with explicit algebraic properties. Robust location–scale alignment removes spatially constant crop-score offsets, a soft evidence gate filters weak local responses, and the raw fusion never decreases any value in the global anomaly map. This algebraic guarantee complements the empirical evaluation of detection performance.

2. Related Work

2.1. Feature-Distribution and Memory-Based Normality Modeling

A major branch of industrial UAD models normality directly in a pretrained feature space. PaDiM estimates multivariate Gaussian statistics for spatial patch embeddings [6], while PatchCore uses a representative coreset of nominal patch features and nearest-neighbor matching to preserve fine-grained localization [5]. Conditional normalizing-flow methods such as CFLOW-AD map pretrained features to tractable likelihood distributions [7], whereas CFA adapts patch descriptors toward compact target-oriented normal representations [8]. SANFlow further introduces semantic-aware flow modeling to avoid forcing heterogeneous features into a single distribution [9]. SoftPatch addresses a complementary practical issue by suppressing suspicious patches before coreset construction, improving robustness when nominal training data contain noise or contamination [10].

These methods avoid explicit modeling of unknown defect appearance and often provide strong localization. Their fixed feature scale, however, ties anomaly evidence to the spatial resolution established during preprocessing. Global resizing can weaken microscopic structures before memory matching or density estimation, while storage and retrieval costs also grow with the number of categories or normal samples. HiFRA-AD complements this paradigm by preserving local variation and allocating higher-resolution evidence only where it is informative.

2.2. Reconstruction, Distillation, and Synthetic-Anomaly Learning

A second line of research learns anomaly-sensitive mappings by reconstructing normal images or representations. Reverse distillation reconstructs multi-scale teacher features through a one-class bottleneck and detects anomalies from teacher–student discrepancies [11]. RD++ revisits this design with improved representation coupling [13], while memory-guided knowledge distillation incorporates normal memories into the distillation pathway [14]. Template-guided hierarchical feature restoration introduces normal templates to guide multi-level reconstruction [15], and ReContrast jointly optimizes feature extractors through contrastive reconstruction to reduce domain mismatch [12]. These methods show that feature reconstruction can combine pretrained semantics with dense spatial comparison more effectively than direct pixel reconstruction.

The central difficulty of reconstruction-based UAD is over-generalization. A powerful reconstruction pathway may reproduce unseen anomalies nearly as well as normal regions, reducing the residual needed for detection. One response is to convert anomaly detection into restoration or discrimination using synthetic anomalies. CutPaste creates self-supervised local perturbations [26], DRAEM combines synthetic corruption with discriminative reconstruction [27], and Natural Synthetic Anomalies improves realism by drawing on natural image statistics [28]. DeSTSeg trains a denoising student–teacher segmentation model with synthesized defects [29], while OmniAL unifies synthesis, reconstruction, and multi-level localization [30]. Recent methods further increase the controllability of generated defects through realistic feature synthesis [19,20] or diffusion models [31,32].

Synthetic anomalies provide explicit negative signals but also introduce a modeling assumption about what unseen defects should look like. Their generalization depends on how well the synthetic distribution overlaps with real abnormalities. HiFRA-AD instead retains normal-only reconstruction and introduces a shape-preserving residual filter that changes the sensitivity of the bottleneck input to local token variation. This choice is also related to frequency-aware image modeling. Focal Frequency Loss demonstrates that difficult frequency components can be under-optimized by conventional reconstruction objectives [33]. In HiFRA-AD, however, frequency analysis is applied to the spatial token

field, not directly to image pixels, and is used as a lightweight feature transformation rather than a supervised frequency-domain loss.

2.3. Unified Multi-Class and Scalable Anomaly Detection

Industrial deployment increasingly favors a single detector that serves many products rather than one model per category. UniAD formalized the unified multi-class UAD setting and identified reconstruction shortcuts as a major obstacle [16]. Hierarchical Vector Quantized Transformer constrains reconstruction through discrete normal prototypes [34], RLR learns unified reference representations to anchor reconstruction to nominal patterns [35], and MambaAD explores state-space models for long-range dependencies in multi-class inspection [21]. GeneralAD seeks broader cross-domain anomaly modeling by discriminating distorted patch features [36], while IUF considers incremental unified inspection as new product categories are introduced [37]. Continual diffusion methods such as One-for-More further study how shared generative models can acquire new normal patterns without catastrophic forgetting [38].

The unified trend has recently converged with foundation-model reconstruction. Dinomaly shows that strong self-supervised ViT features and deliberately imperfect reconstruction can produce a highly competitive one-for-all detector [24]. Dinomaly2 extends this philosophy across multi-class, inference-unified, few-shot, multi-view, and multi-modal scenarios [25]. This progression substantially reduces the maintenance burden of industrial inspection systems.

Scale fidelity remains a separate challenge from model unification. A unified model can represent many products with one parameter set while still operating on globally resized images, and stronger reconstruction alone leaves the input-scale bottleneck unchanged. HiFRA-AD preserves the one-for-all reconstruction setting and addresses this complementary axis of scalability by allocating additional computation to a fixed number of globally proposed regions.

2.4. Vision and Vision–Language Foundation Models

Foundation models have changed both the representation quality and data regime of anomaly detection. WinCLIP adapts CLIP to zero- and few-shot anomaly classification and segmentation through window-level visual–text matching [39]. AnomalyCLIP learns object-agnostic prompts for generic normality and abnormality [40], PromptAD learns anomaly-sensitive prompts from only a few normal samples [41], and AdaCLIP combines static and dynamic learnable prompts for zero-shot transfer [42]. AnomalyGPT further explores large vision–language models for industrial anomaly reasoning [43], while AA-CLIP adapts both visual and textual spaces toward anomaly-aware representations [44].

Recent work also shows that language is not the only route to strong few-shot generalization. FastRecon reconstructs query features from a small normal support set [45]. AnomalyDINO demonstrates that DINOv2 patch representations alone can provide strong training-free few-shot anomaly detection [46]. UniVAD uses vision foundation models and a few normal references to unify structural, logical, and medical anomaly detection [47], whereas LogSAD combines vision and language foundation models to detect both structural and logical anomalies without target-specific training [48].

These methods demonstrate the broad transferability of foundation features, but industrial defect localization remains spatially fine-grained. Patch size and fixed-resolution preprocessing can limit the visibility of sub-patch anomalies. HiFRA-AD uses a frozen DINOv2 encoder as a stable semantic backbone and focuses on a complementary question: how can local spatial variation be emphasized and high-resolution evidence acquired selectively while retaining the pretrained representation?

2.5. High-Resolution and Realistic Industrial Inspection

Recent datasets and methods increasingly expose practical conditions that are weakly represented by early anomaly benchmarks. MVTec LOCO AD extends inspection from local surface defects to logical constraint violations such as missing or misplaced components [49]. MVTec 3D-AD introduces high-resolution 3D sensing for geometric defects [50], and cross-modal feature mapping learns normal correspondences between RGB and 3D representations [51]. Distribution-shift studies further show that a normal test distribution may differ from the training distribution even when no true anomaly is present [52,53]. These works collectively emphasize that practical anomaly detection must remain reliable when sensing conditions, scene composition, or normal appearance change.

High-resolution inspection has more recently become a first-class research problem. HiAD argues that global downsampling can erase subtle anomalies and adopts a multi-resolution framework to combine global and local evidence [54]. This direction confirms the importance of preserving small structures, but it also exposes an efficiency–coverage trade-off: dense tiling or multi-resolution processing can become expensive, whereas sparse processing risks missing defects outside selected regions.

Complementary advances in detection-oriented perception show how resource allocation and spatial detail preservation can support reliable visual decisions in challenging environments. Chen et al. proposed a geometry-guided sparse-to-dense fusion approach for railway three-dimensional object detection, using sparse structural anchors to restrict expensive image-feature retrieval [55]. Sun et al. investigated a super-resolution-enhanced, lightweight YOLOv13n pipeline for maturity-aware fruit detection, showing the relevance of recovering fine image details for small-object recognition under occlusion and variable illumination [56]. These supervised object-detection studies illustrate selective computation and resolution-aware perception in related inspection tasks.

Recent industrial anomaly-detection studies examine related issues more directly. Samele et al. explored scalability through segmentation-based anomaly detection for large and high-resolution product images [57], while Qiao and Liu introduced frequency-guided local semantic prompting for zero-shot industrial anomaly localization [58]. Unlike the latter's semantic prompting with vision–language representations, HiFRA-AD operates on DINOv2 reconstruction tokens and combines fixed-coefficient residual enhancement with peak-guided local re-evaluation and score calibration. Taken together, these studies sharpen the distinction between improving local evidence sensitivity, deciding where to allocate computation, and reconciling anomaly scores across global and local views.

HiFRA-AD studies this trade-off from a reconstruction perspective. The global anomaly map proposes a small number of candidate regions, which are revisited at higher model-coordinate sampling density to strengthen spatially weak anomaly evidence. The crop-level anomaly map is then treated as a context-shifted local measurement, and robust location–scale alignment separates crop-induced score shift from additional defect evidence before fusion.

2.6. Research Gap

Existing literature has made substantial progress along several largely parallel axes: accurate normality modeling, unified reconstruction, foundation-model transfer, synthetic anomaly generation, and high-resolution inspection. The remaining gap addressed in this work lies in their interaction. In particular, three questions are rarely treated as one coupled problem:

1. **Feature sensitivity:** after global resizing and token reconstruction, does a small local structural change still produce sufficiently strong spatial variation?

2. **Computation allocation:** can higher-resolution re-evaluation be restricted to a small set of globally suspicious regions instead of using dense sliding windows?
3. **Evidence compatibility:** once a crop is enlarged and re-evaluated, how should its anomaly map be aligned with the corresponding global response before fusion?

HiFRA-AD couples these questions through three operations. A fixed-coefficient high-frequency residual operator increases local token variation before reconstruction; a peak-suppressed proposal mechanism selects sparse high-resolution regions; and robust residual-aware calibration converts local responses into nonnegative evidence relative to the global map. The method therefore forms a direct chain from scale-induced information loss, to selective local re-sampling, to mathematically controlled global–local fusion.

3. Method

3.1. Overview of HiFRA-AD

HiFRA-AD is a reconstruction-based framework for unsupervised industrial anomaly detection. It is built on the 2D Dinomaly2 pipeline and retains its frozen DINOv2 encoder, lightweight bottleneck, Transformer decoder, context recentering, and loose reconstruction strategy [25]. HiFRA-AD targets the scale–context mismatch: fixed-resolution reconstruction can weaken small spatial variations, whereas re-evaluating a suspicious region at higher resolution changes its context and makes the resulting anomaly scores difficult to compare directly with the global prediction. The framework therefore separates feature sensitivity from evidence aggregation. A high-frequency residual operator without additional learnable weights is introduced before the reconstruction bottleneck, while a sparse global-to-local refinement mechanism is used only at inference.

Figure 1 summarizes the complete HiFRA-AD pipeline. The training branch modifies feature sensitivity through high-frequency residual enhancement, whereas the inference branch uses the global anomaly response to allocate sparse local re-evaluation and subsequently calibrates the local evidence before non-suppressive fusion.

Given an input image $x \in \mathbb{R}^{3\times H_0\times W_0}$, the frozen DINOv2 encoder $\mathcal{E}$ extracts intermediate token representations

$$\mathcal{F} = \mathcal{E}(x) = \{f_\ell \mid \ell \in \mathcal{M}\}, \tag{1}$$

where $\mathcal{M}$ contains eight intermediate Transformer layers. In the default ViT-S/14 configuration, the 392 × 392 model crop produces a 28 × 28 spatial token grid. The selected features are averaged before entering the bottleneck,

$$X = \frac{1}{|\mathcal{M}|}\sum_{\ell\in\mathcal{M}} f_\ell. \tag{2}$$

The class and register tokens are retained during feature processing, while only spatial tokens participate in dense anomaly scoring. Following Dinomaly2, encoder-side spatial features are recentered by the class token and normalized before comparison. Let g_k and $\hat{g}_k$ denote the k-th grouped encoder feature and its decoder reconstruction. The reconstruction objective is the grouped cosine distance

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\left[1 - \frac{\operatorname{vec}(g_k)^\top \operatorname{vec}(\hat{g}_k)}{\|\operatorname{vec}(g_k)\|_2\|\operatorname{vec}(\hat{g}_k)\|_2}\right]. \tag{3}$$

The loose optimization policy of Dinomaly2 is retained: well-reconstructed positions receive attenuated gradients, while difficult normal positions remain emphasized during training. The DINOv2 encoder is frozen and only the bottleneck and decoder are optimized.

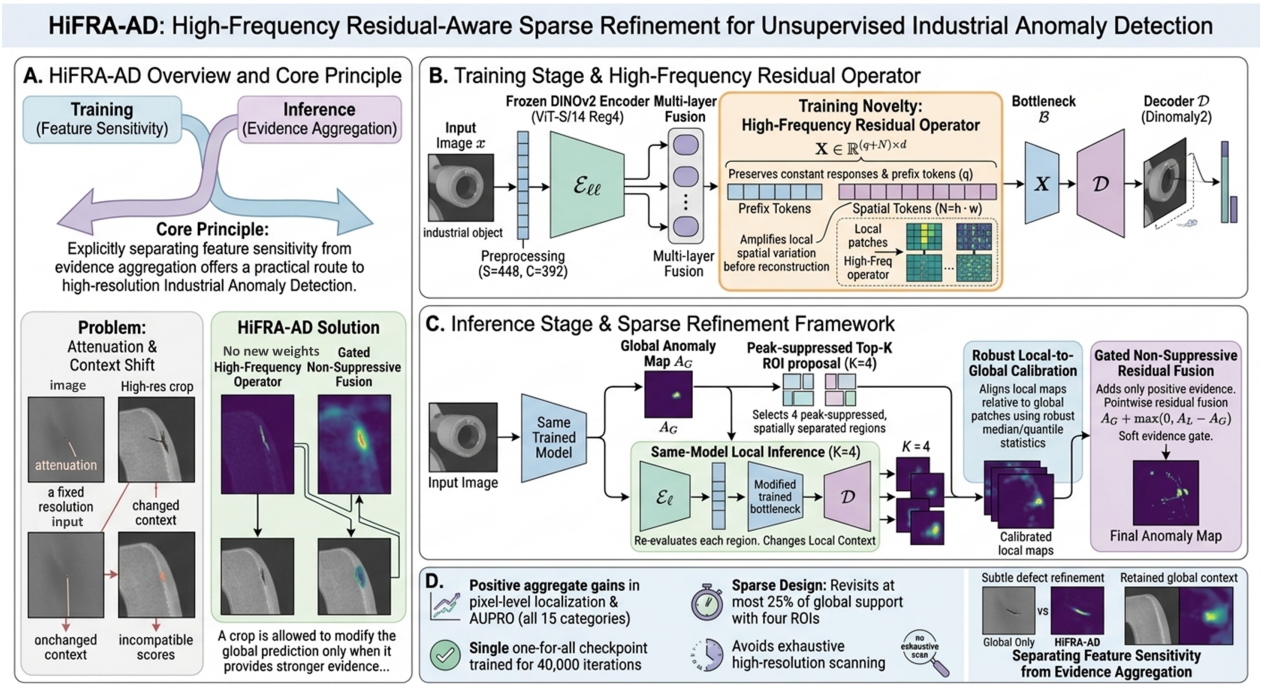


Figure 1. Overview of HiFRA-AD. (**A**) The framework separates training-time feature sensitivity from inference-time evidence aggregation. (**B**) A high-frequency residual operator with no additional learnable weights enhances spatial token variation before the Dinomaly2 reconstruction bottleneck. (**C**) During inference, the global anomaly map proposes four peak-suppressed regions, which are re-evaluated by the shared model and calibrated before gated non-suppressive fusion. (**D**) The design uses a single one-for-all checkpoint and sparse local refinement instead of exhaustive high-resolution scanning. Panel C depicts positive residual fusion schematically; Equations (26)–(31) specify the gate, spatial weights, fusion coefficient, and overlap normalization. Blue and purple distinguish the feature-processing and reconstruction/fusion branches; orange highlights the residual operator, and green highlights sparse local re-evaluation. In the anomaly maps, brighter colors indicate stronger responses.

At inference, token-wise reconstruction distances are upsampled and averaged to form the global anomaly map

$$A_G(p) = \frac{1}{|\mathcal{G}|} \sum_{k \in \mathcal{G}} \mathrm{Up}\left(1 - \frac{g_k(p)^\top \hat{g}_k(p)}{\|g_k(p)\|_2 \|\hat{g}_k(p)\|_2}\right). \tag{4}$$

HiFRA-AD then uses A_G to select a small number of suspicious regions from the original image. These regions are re-evaluated by the same reconstruction model at a higher effective sampling density, and their local responses are calibrated against the corresponding global patches before fusion. The overall pipeline therefore consists of three tightly connected operations: high-frequency residual enhancement, sparse high-resolution refinement, and residual-aware global-local fusion.

3.2. High-Frequency Residual Enhancement

The reconstruction bottleneck receives a fused token sequence containing q prefix tokens and $N = h \times w$ spatial tokens. We leave the prefix tokens unchanged and reshape the spatial part into

$$F \in \mathbb{R}^{d \times h \times w}. \tag{5}$$

A stride-one 3 × 3 average-pooling operator $\mathcal{A}(\cdot)$ is used to estimate the locally smooth component,

$$F_{\mathrm{low}} = \mathcal{A}(F), \qquad F_{\mathrm{high}} = F - F_{\mathrm{low}}. \tag{6}$$

Instead of introducing an additional attention block or trainable enhancement network, we directly strengthen the residual spatial variation,

$$F^{+} = F + \alpha F_{\text{high}} = (1 + \alpha)F - \alpha \mathcal{A}(F), \tag{7}$$

where $\alpha = 0.2$ by default. The averaging kernel and scalar coefficient are fixed configuration choices. Setting $\alpha = 0$ recovers the original spatial features. The transformation preserves the model's learnable parameter count and changes the input to the bottleneck and decoder. The enhanced spatial tokens are restored to sequence layout and concatenated with the unchanged prefix tokens.

The design exploits the observation that small scratches, cracks, edge damage, and weak local texture changes often produce sharper spatial variation than their surrounding normal regions. Feature fusion and reconstruction can attenuate these variations. The residual operator compensates for this attenuation without modifying the frozen foundation encoder or changing the feature dimensionality.

The transformation also has a direct frequency interpretation. Under a locally stationary approximation, the response of the operator to a token-grid spatial frequency ω is

$$H_{\alpha}(\omega) = 1 + \alpha(1 - \hat{a}(\omega)), \tag{8}$$

where $\hat{a}(\omega)$ is the frequency response of the averaging kernel. For the default 3×3 box filter,

$$\hat{a}(\omega_x, \omega_y) = \frac{(1 + 2\cos\omega_x)(1 + 2\cos\omega_y)}{9}. \tag{9}$$

The DC component satisfies $H_{\alpha}(0) = 1$, while spatial modes attenuated by local averaging receive a larger gain. Therefore, the module preserves a spatially constant token field and selectively increases local variation. The operation is applied to the token feature field rather than directly to image pixels: it increases reconstruction sensitivity to local structure without treating all high-frequency content as anomalous.

3.3. Symmetry Interpretation of Local Residual Enhancement

Define $\mathcal{R}_{\alpha} = (1 + \alpha)I - \alpha\mathcal{A}$. For g in the dihedral group D_4 of quarter-turn rotations and mirror reflections on the square token grid, let $(\mathcal{T}_g F)(p) = F(g^{-1}p)$. The uniform 3×3 stencil and a transformation-symmetric boundary rule give

$$\mathcal{A}\mathcal{T}_g = \mathcal{T}_g\mathcal{A}, \qquad \mathcal{R}_{\alpha}\mathcal{T}_g = \mathcal{T}_g\mathcal{R}_{\alpha}. \tag{10}$$

This operator-level equivariance preserves an ideal nominal symmetry: if $\mathcal{T}_g F_N = F_N$, then $\mathcal{T}_g\mathcal{R}_{\alpha}(F_N) = \mathcal{R}_{\alpha}(F_N)$. For a locally perturbed field $F = F_N + \delta$,

$$(I - \mathcal{T}_g)\mathcal{R}_{\alpha}(F) = \mathcal{R}_{\alpha}(I - \mathcal{T}_g)\delta. \tag{11}$$

The symmetry-deviation component therefore passes through the frequency response $H_{\alpha}(\omega)$. High-frequency departures from a regular pattern receive increased gain, while constant modes are preserved. The derivation concerns the fixed token-grid filter; the encoder and reconstruction pathway determine the eventual anomaly scores.

Structural regularity. For a normal token field F_N, a non-identity transformation g, and overlapping support Ω_g, define the normalized discrepancy

$$D_g(F_N) = \frac{1}{|\Omega_g|} \sum_{p \in \Omega_g} \frac{\|F_N(p) - F_N(g^{-1}p)\|_2^2}{\sigma^2(F_N) + \epsilon}. \tag{12}$$

A normal structure is approximately regular under g when $D_g(F_N)$ is small. Local perturbations can disrupt this correspondence. The discrepancy provides a descriptive measure of structural regularity, and HiFRA-AD uses the reconstruction score for detection. Translation-related repetition and reflection/rotation symmetry define distinct forms of regularity. Section 4.7 evaluates the observed gains on predefined structural-disruption subtypes.

3.4. Sparse High-Resolution Refinement

A straightforward way to preserve small defects is to process the full image at a higher resolution or scan it using dense overlapping windows. Both strategies are expensive because most high-resolution computation is spent on normal regions. HiFRA-AD instead uses the global reconstruction response as a proposal signal and revisits only a small number of candidate regions.

Let the global anomaly map have size $H_A \times W_A$. All ROI operations use the pixel grid of this map, after the token-resolution anomaly scores have been bilinearly upsampled to the global $C \times C$ model-crop resolution; thus, $H_A = W_A = C = 392$ in the reported implementation. The map is computed before final Gaussian smoothing and image-level score aggregation. For an ROI side ratio ρ, the proposal size is

$$b_h = \mathrm{round}(\rho H_A), \qquad b_w = \mathrm{round}(\rho W_A). \tag{13}$$

Starting from $P_0 = A_G$, the j-th proposal center is selected as

$$p_j = \arg\max_p P_{j-1}(p). \tag{14}$$

Peak suppression uses greedy rectangular masking. At iteration j, let $p_j = (x_j, y_j)$ denote the highest remaining map pixel, with row-major order resolving exact ties. The zero-indexed, half-open proposal rectangle is

$$\begin{aligned} x_{1j} &= \max\big(0, x_j - \lfloor b_w/2 \rfloor\big), \quad x_{2j} = \min\big(W_A, x_j + \lceil b_w/2 \rceil\big), \\ y_{1j} &= \max\big(0, y_j - \lfloor b_h/2 \rfloor\big), \quad y_{2j} = \min\big(H_A, y_j + \lceil b_h/2 \rceil\big). \end{aligned} \tag{15}$$

$$r_j = [x_{1j}, x_{2j}) \times [y_{1j}, y_{2j}), \qquad P_j(p) = \begin{cases} -\infty, & p \in r_j, \\ P_{j-1}(p), & p \notin r_j. \end{cases} \tag{16}$$

Only the temporary proposal map P_j is masked: the original A_G is left unchanged for subsequent calibration and fusion. Boxes touching image boundaries are clipped, and distinct boxes may still overlap although their selected peak coordinates cannot lie within a previously suppressed rectangle. Repeating this process yields K peak-suppressed candidate regions. We use $K = 4$ and $\rho = 0.25$ by default. Up to boundary clipping and overlap, the selected regions occupy at most $K\rho^2$ of the global map, corresponding to approximately 25% for the default setting.

Each selected box is mapped back to the original image instead of being cropped from the already resized model input. Global preprocessing first resizes the source image of width W_0 and height H_0 to a fixed $S \times S$ canvas and then extracts its centered $C \times C$ crop; no aspect-ratio-preserving padding or letterboxing is used. The inverse therefore undoes the centered crop offset and then the horizontal and vertical resizing scales, separately. Let S be the resize canvas, C the center-crop size, and

$$o = \frac{S - C}{2} \tag{17}$$

the crop offset. For a global ROI $r = [x_1, x_2) \times [y_1, y_2)$, the corresponding source-image coordinates are

$$\begin{aligned} \bar{x}_1 &= \left\lfloor (x_1 + o)\frac{W_0}{S} \right\rfloor, \quad \bar{y}_1 = \left\lfloor (y_1 + o)\frac{H_0}{S} \right\rfloor, \\ \bar{x}_2 &= \left\lceil (x_2 + o)\frac{W_0}{S} \right\rceil, \quad \bar{y}_2 = \left\lceil (y_2 + o)\frac{H_0}{S} \right\rceil. \end{aligned} \tag{18}$$

All four bounds are first clamped to $[0, W_0]$ or $[0, H_0]$ after taking the floor/ceiling, and the source-image crop uses the corresponding half-open integer rectangle. This coordinate

rule is exact for the stated fixed-canvas resize and $C \times C$ global map; if proposals are instead computed on an $H_A \times W_A$ grid, their corner coordinates must first be multiplied by C/W_A and C/H_A in the horizontal and vertical directions, respectively, before applying the equations.

The source crop is resized directly to the $C \times C$ model input (without a second global center crop) and passed through the same HiFRA-AD reconstruction network. If the ROI side length is approximately ρC, resizing it to C increases the model-coordinate sampling density by approximately $1/\rho$ along each spatial axis.

The local reconstruction map is also upsampled to $C \times C$ before spatial re-projection. To preserve correspondence after integer rounding of the source crop, local values are warped onto the original ROI's pixel support instead of being added to the full global map at crop resolution. For a global pixel center $(x + \frac{1}{2}, y + \frac{1}{2}) \in r_j$, its continuous source-image coordinates are

$$\begin{aligned} u_0(x) &= \frac{W_0}{S}\left(o + \frac{C}{W_A}\left(x + \frac{1}{2}\right)\right), \\ v_0(y) &= \frac{H_0}{S}\left(o + \frac{C}{H_A}\left(y + \frac{1}{2}\right)\right). \end{aligned} \tag{19}$$

The corresponding sampling locations in the $C \times C$ local anomaly map $A_{\text{loc}}^{(j)}$ are

$$\begin{aligned} u_j(x) &= \frac{C\,[u_0(x) - \bar{x}_1]}{\bar{x}_2 - \bar{x}_1} - \frac{1}{2}, \\ v_j(y) &= \frac{C\,[v_0(y) - \bar{y}_1]}{\bar{y}_2 - \bar{y}_1} - \frac{1}{2}, \\ L_j(x,y) &= \text{BilinearSample}\left(A_{\text{loc}}^{(j)}, u_j(x), v_j(y)\right), \quad (x,y) \in r_j. \end{aligned} \tag{20}$$

This uses the half-pixel convention (equivalent to bilinear grid sampling with `align_corners = False`) and border-value padding for coordinates outside the local grid. The resulting L_j has exactly the height and width of r_j; pixels outside r_j receive no local update. No anomaly-score normalization or thresholding is applied before this geometric re-projection. With $\rho = 0.25$, the selected region receives roughly four times denser sampling per axis than in the global pass.

The global reconstruction map remains the proposal and contextual backbone, while the local branch refines selected regions with spatially weak evidence. The dominant model-evaluation work is approximately

$$C_{\text{HiFRA}} \approx (1 + K)C_{\text{model}} + C_{\text{crop}} + C_{\text{fuse}}, \tag{21}$$

which scales with a fixed number of local evaluations rather than the number of windows required to cover the complete high-resolution image. This expression counts forward-pass work; the wall-clock ratios reported in Section 4.4 additionally reflect GPU utilization and the lightweight crop and fusion operations.

3.5. Residual-Aware Calibration and Anomaly Scoring

Local re-evaluation increases spatial sampling density and simultaneously changes object context, boundary composition, and normal score statistics. HiFRA-AD therefore aligns each local map with its corresponding global patch before fusion, separating crop-induced score shift from stronger anomaly evidence.

For ROI r_j, let $G_j = A_G|_{r_j}$ denote the cropped global anomaly map and L_j the local anomaly map re-projected by the pixel-center bilinear sampling rule in Section 3.4. Thus, both maps have identical $|r_j|$ spatial support, including for boundary-clipped ROIs, before

any quantile statistics are computed. We estimate robust location and scale using the median and the upper-tail contrast,

$$m_G = Q_{0.50}(G_j), \qquad s_G = \max\{Q_{0.90}(G_j) - m_G, \varepsilon\}, \tag{22}$$

$$m_L = Q_{0.50}(L_j), \qquad s_L = \max\{Q_{0.90}(L_j) - m_L, \varepsilon\}, \tag{23}$$

where $Q_q(\cdot)$ denotes the empirical q-quantile and $\varepsilon = 10^{-6}$. The standardized local evidence is

$$z_j(p) = \mathrm{clip}_{[0,z_{\max}]}\left(\frac{L_j(p) - m_L}{s_L}\right), \qquad z_{\max} = 4. \tag{24}$$

It is then mapped to the location and scale of the global patch,

$$\hat{L}_j(p) = m_G + s_G z_j(p). \tag{25}$$

This location-scale alignment removes a spatially constant additive bias from the local score map. Indeed, replacing L_j by $L_j + c$ shifts both the local median and upper quantile by the same constant and therefore leaves z_j unchanged.

Interpretation of calibration. Median/quantile alignment compensates for crop-induced score shifts. Under $L'_j = aL_j + b$ with $a > 0$, the median and upper-tail contrast become $m'_L = am_L + b$ and $Q_{0.90}(L'_j) - m'_L = a[Q_{0.90}(L_j) - m_L]$. Thus, $z'_j = z_j$ when the scale floor is inactive; additive-offset invariance also holds with the floor. The evidence gate then attenuates weak standardized local peaks. Calibration is an optional operation on existing ROI scores. The positive-part residual supplies the raw fusion's nonnegativity.

To avoid injecting weak or unsupported local responses, we introduce a continuous evidence gate,

$$g_j(p) = \sigma\left(\frac{z_j(p) - \tau}{T}\right)\frac{z_j(p)}{z_{\max}}, \tag{26}$$

with $\tau = 1.5$ and $T = 0.5$. The additional evidence contributed by ROI j is

$$\Delta_j(p) = g_j(p)\left[\max\{\hat{L}_j(p), G_j(p)\} - G_j(p)\right]. \tag{27}$$

Thus, a local estimate contributes only when it is stronger than the already available global response. A spatially constant local map gives $z_j = 0$, hence $\Delta_j = 0$.

Finally, a center-weighted window reduces crop-boundary artifacts,

$$W_j(\xi, \eta) = \mu + (1 - \mu)\cos^2\left(\frac{\pi\xi}{2}\right)\cos^2\left(\frac{\pi\eta}{2}\right), \tag{28}$$

where $\xi, \eta \in [-1, 1]$ and $\mu = 0.25$. The accumulated residual and weight maps are

$$U(p) = \sum_{j=1}^{K} W_j(p)\Delta_j(p)\mathbf{1}_{r_j}(p), \tag{29}$$

$$V(p) = \sum_{j=1}^{K} W_j(p)\mathbf{1}_{r_j}(p), \tag{30}$$

and the refined anomaly map is

$$A_R^{\mathrm{raw}}(p) = A_G(p) + \beta\frac{U(p)}{\max\{1, V(p)\}}, \qquad \beta = 0.5. \tag{31}$$

Because $g_j \geq 0$, $W_j \geq 0$, and $\Delta_j \geq 0$, the raw fusion satisfies

$$A_R^{\mathrm{raw}}(p) \geq A_G(p), \qquad \forall p. \tag{32}$$

The local branch therefore preserves anomaly evidence already produced by the global branch and adds only nonnegative residuals. This pointwise non-suppression property provides a direct interpretation of the raw fusion rule.

Calibration at low global contrast. For a constant global patch $G_j(p) = c$, its median and upper quantile equal c, giving $s_G = \varepsilon$ and

$$\hat{L}_j(p) = c + \varepsilon z_j(p), \qquad \Delta_j(p) = g_j(p)\varepsilon z_j(p). \tag{33}$$

With $0 \le z_j \le 4$, $0 \le g_j, W_j \le 1$, and $\beta = 0.5$, a pixel covered by one ROI satisfies

$$0 \le A_R^{\mathrm{raw}}(p) - A_G(p) = \beta W_j(p)\Delta_j(p) \le \beta z_{\max}\varepsilon = 2\varepsilon = 2 \times 10^{-6}. \tag{34}$$

The correction is consequently negligible in the constant-global-patch limit, even for a strong standardized local response. More generally, let $d_j = Q_{0.90}(G_j) - m_G$. The one-ROI correction satisfies

$$0 \le A_R^{\mathrm{raw}}(p) - A_G(p) \le \beta[z_{\max} \max\{d_j, \varepsilon\} + [m_G - G_j(p)]_+]. \tag{35}$$

For $G_j(p) \ge m_G$, this reduces to $2\max\{d_j, \varepsilon\}$; points below the median retain the lower-tail term. For overlapping ROIs, the correction is bounded by the largest corresponding single-ROI bound. Calibration thus scales the local contribution to the global patch's contrast. The ε floor provides numerical stability, while proposal coverage determines which defects receive local re-evaluation.

During training, only the high-frequency token enhancement is active; no local crops are required and no additional local detector is learned. During inference, one global pass is followed by K local evaluations, robust calibration, and residual fusion. The final map is resized and smoothed using the same post-processing as the baseline. The image-level anomaly score follows the standard top-response aggregation,

$$s_{\mathrm{img}} = \mathrm{mean}\big(\mathrm{Top}_{1\%}(A_R)\big). \tag{36}$$

HiFRA-AD retains the Dinomaly2 reconstruction pathway and adds a fixed-coefficient local-variation transformation with no additional learnable parameters during training and sparse residual refinement during inference.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate HiFRA-AD on three representative industrial anomaly-detection benchmarks. MVTec AD [1] contains 15 industrial categories, including five texture classes and ten object classes, and remains the most widely used benchmark for image-level anomaly detection and pixel-level localization. VisA [2] contains 12 object categories with more complex structures and diverse defect patterns, providing a complementary test of cross-category generalization. MVTec AD 2 [4] extends the evaluation to more challenging acquisition conditions and small defects. For all datasets, only normal images from the official training split are used for optimization; abnormal images and pixel annotations are used exclusively for evaluation. Following the unified protocol of Dinomaly2, all categories within a dataset are jointly trained with a single model rather than fitting one detector per category. This setting reflects deployment with one shared detector and avoids category-specific model maintenance.

Dataset composition and evaluation splits. The precise category-wise benchmark sizes and the split composition are summarized in Table 1. MVTec AD comprises 5354 images across 15 categories, including 3629 defect-free training samples and 1725 test samples (467 normal and 1258 anomalous). For VisA, the standard one-class split uses 8659 of its

9621 normal images for training, with 962 normal and 1200 anomalous images held out for testing; this gives 10,821 images across 12 categories. MVTec AD 2 contains 8004 images from eight scenarios, partitioned into 2528 normal training images, 302 normal validation images, and 5174 test images. Its test partition consists of 1084 publicly annotated images (379 normal, 705 anomalous) and 4090 privately evaluated images across two separate 2045-image test subsets. Private-test ground truth is unavailable for local scoring and can only be evaluated via the official benchmark server. The MVTec AD 2 validation partition is not used for gradient updates. Ground-truth masks are never used for model training.

Table 1. Image counts, evaluation splits, and server-recorded training epochs for the three industrial anomaly-detection benchmarks. The public test column gives normal/anomalous counts; private-test labels for MVTec AD 2 are held by the official benchmark server. MVTec AD 2 results in the comparison tables use the 2045-image TESTpriv partition; the 4090 private-test total also includes the distinct 2045-image TESTpriv, mix partition. The VisA counts follow the standard one-class split.

Dataset	Cats.	Train Normal	Val. Normal	Public Test N/A	Private Test	Total	Epochs Logged
MVTec AD	15	3629	–	467/1258	–	5354	176
VisA (one-class)	12	8659	–	962/1200	–	10,821	73.974
MVTec AD 2	8	2528	302	379/705	4090	8004	253

Metrics. We report image-level AUROC (I-AUROC), pixel-level AUROC (P-AUROC), pixel-level average precision (P-AP), and area under the per-region overlap curve (AUPRO). P-AP and AUPRO characterize localization when defects occupy a small image fraction. Dataset results are macro-averages of category-wise metrics, following Dinomaly2. Statistical comparisons use independently trained, seed-matched runs as their replication unit.

MVTec AD 2 partition and metric conventions. The benchmark distinguishes the publicly annotated $\mathrm{TEST}_{\mathrm{pub}}$, which permits local scoring, from $\mathrm{TEST}_{\mathrm{priv}}$ (training-matched illumination) and $\mathrm{TEST}_{\mathrm{priv,mix}}$ (mixed seen/unseen illumination). Pixel-level ground truth for both private partitions is withheld and their scores require submission to the official evaluation server [4]. AUPRO must additionally be indexed by its false-positive-rate integration limit. For $a \in \{0.05, 0.30\}$, the normalized score is

$$\mathrm{AUPRO}_a = \frac{1}{a}\int_0^a \mathrm{PRO}(u)\,\mathrm{d}u, \tag{37}$$

where PRO averages the detected fraction over connected ground-truth defect regions at false-positive rate u, and the curve is integrated over the specified interval. The MVTec AD 2 benchmark establishes $\mathrm{AUPRO}_{0.05}$ as its stricter setting while also showing $\mathrm{AUPRO}_{0.30}$ results in the dataset paper. The integration limit is therefore part of the metric definition [4].

MVTec AD 2 evaluation protocol. All eight methods in the MVTec AD 2 comparison use the 2045-image $\mathrm{TEST}_{\mathrm{priv}}$ partition with training-matched illumination. Anomaly maps are evaluated through the official server (https://benchmark.mvtec.com/ (accessed on 11 November 2026)), which holds the private ground-truth masks. AUPRO denotes normalized $\mathrm{AUPRO}_{0.30}$ from Equation (37). Image AUROC, pixel AUROC, pixel AP, and category-macro aggregation follow the common server-side scoring implementation. The partition and metric settings follow the submission and scoring records described for these experiments.

Implementation details. All experiments are implemented in PyTorch 2.10 and run on a single NVIDIA GeForce RTX 4090 with 24 GB memory. We use DINOv2 ViT-S/14 with register tokens as the frozen encoder. Eight intermediate Transformer layers are extracted for reconstruction. The input image is first resized to 448 × 448 and center-cropped to 392 × 392, producing a 28 × 28 token grid. The bottleneck and eight-layer decoder are optimized with StableAdamW for 40,000 iterations. The initial learning rate is 2×10^{-3},

with 100 iterations of linear warm-up followed by cosine decay to 2×10^{-4}; weight decay is 10^{-4}, $\beta = (0.9, 0.999)$, and $\epsilon = 10^{-10}$. The batch size is 16 whenever memory permits, and mixed-precision training is used on the RTX 4090. These settings follow the optimization regime of Dinomaly2, whose reported implementation also uses StableAdamW, a 392-pixel input, top-response image scoring, and RTX 3090/4090 GPUs.

Public implementation resources. The backbone and reconstruction baselines are available from the official DINOv2 (https://github.com/facebookresearch/dinov2 (accessed on 11 November 2026)), Dinomaly (https://github.com/guojiajeremy/Dinomaly (accessed on 11 November 2026)), and Dinomaly2 (https://github.com/guojiajeremy/Dinomaly2 (accessed on 11 November 2026)) repositories. These links identify the third-party implementations used as methodological references.

Training epochs and iteration budget. Each dataset-specific one-for-all model is trained for 40,000 optimizer updates. The reported run counters are 176 epochs for MVTec AD, 73.974 for VisA, and 253 for MVTec AD 2 (Table 1); the fixed update budget is the stopping criterion.

For HiFRA-AD, the high-frequency residual coefficient is fixed to $\alpha = 0.2$. During sparse refinement, we select $K = 4$ candidate regions with an ROI side ratio $\rho = 0.25$. The residual fusion coefficient is $\beta = 0.5$, while the calibration parameters are $\tau = 1.5$, $T = 0.5$, $z_{\max} = 4$, and $\mu = 0.25$. The same settings are shared by all categories; no category-specific hyperparameter tuning is used. For each image, candidate centers are obtained by iterative rectangle masking with suppressed values set to $-\infty$. ROIs are cropped from the original image by floor/ceiling inverse coordinates, and local maps are returned to the global anomaly-map support using half-pixel bilinear re-projection before calibration. These deterministic steps are identical across datasets. Unless otherwise stated, the image-level score is the mean of the top 1% anomaly-map responses, consistent with the underlying reconstruction baseline.

Hyperparameter selection. We select α from $\{0, 0.1, 0.2, 0.3, 0.5\}$ using category-averaged P-AP in the annotated MVTec AD development sweep (see the sensitivity analysis in Section 4.5). The selected $\alpha = 0.2$ yields 71.6% P-AP and 95.50% AUPRO. With $\rho = 0.25$, we select $K = 4$ from $\{0, 1, 2, 4, 6\}$ at the accuracy–cost knee: four ROIs reach 71.6% P-AP at 2.08× inference time, while six reach 71.5% at 2.71× (see Section 4.5). These choices use evaluation annotations and may introduce selection optimism in the MVTec AD scores. The selected settings are shared across categories and transferred unchanged to VisA and MVTec AD 2; model optimization uses normal training images.

Repeated-run evaluation. The MVTec AD component study uses five independent training seeds (42, 123, 3407, 2026, and 4444), matched across configurations. The pretrained encoder, splits, preprocessing, optimization schedule, and refinement settings are held fixed. ROI and calibration variants share the corresponding Base checkpoint, while HF variants use HF-trained models. The repeated-run configurations are summarized in Section 4.5. The HF + ROI control and ViT-B comparisons are configuration-level evaluations.

Statistical analysis. Category-wise scores are macro-averaged separately for each seed. We report the mean and sample standard deviation over five runs and use two-sided paired t-tests on seed-aligned differences, with 95% Student-t confidence intervals ($df = 4$). Holm adjustment covers eight localization tests: four component contrasts for each of P-AP and AUPRO. The principal paired observations and inferential results are reported in Section 4.5. I-AUROC and P-AUROC are supplementary outcomes.

Backbone-controlled training protocol. To distinguish improvements attributable to HiFRA-AD from changes in encoder capacity, the same experimental protocol is applied to matched Dinomaly2-S/HiFRA-AD-S and Dinomaly2-B/HiFRA-AD-B pairs. Each pair uses the corresponding identical pretrained DINOv2 checkpoint (ViT-S/14 or ViT-B/14

with register tokens), feature extraction layers, bottleneck and decoder design, training splits, optimizer, augmentation, input size, 40,000-step schedule, effective batch size, and evaluation procedure. The residual operation is applied only in HiFRA-AD training, and the local re-evaluation and calibration are used only for HiFRA-AD inference. When needed to accommodate ViT-B on a single 24-GB GPU, gradient accumulation retains the matched effective batch size. Neither baseline is given additional local-region inference, and neither HiFRA-AD variant uses a separate ROI network.

Efficiency evaluation protocol. The efficiency comparison is performed on the same RTX 4090 platform. We report total model parameters, arithmetic computation in MACs and multiply–add-equivalent FLOPs, end-to-end time for 40,000 training updates, batch-one end-to-end inference latency, corresponding single-stream throughput, and peak training and inference GPU memory. The latency includes the global forward pass, ROI proposal selection, all four local re-evaluations, score calibration, and map fusion. Single-stream throughput is computed as $1000/t_{\mathrm{ms}}$ images/s from batch-one latency and is not presented as batch-16 throughput. Computational complexity includes the five reconstruction-network evaluations for $K = 4$; the additional proposal and fusion operations contribute smaller terms, separately acknowledged below. All four variants use the same input resolution and numerical precision settings in each matched comparison.

Dense-scan comparison. Analytical dense scans reuse the ViT-S reconstruction model and local crop size of HiFRA-AD on the normalized global canvas. Non-overlapping windows use a stride equal to the ROI side; half-overlapping windows use half that stride. Section 4.4 reports the analytical network-pass counts and MACs, together with hardware measurements for the matched Dinomaly2 and HiFRA-AD models.

4.2. Comparison with State-of-the-Art Methods

We compare HiFRA-AD with representative methods covering the major technical routes in industrial UAD: RD4AD [11] for reverse distillation, SimpleNet [17] and DeSTSeg [29] for pseudo-anomaly learning, UniAD [16] for unified reconstruction, ReContrast [12] for contrastive feature reconstruction, DiAD [31] for diffusion-based multi-class anomaly detection, MambaAD [21] for state-space modeling, and the recent foundation-model reconstruction methods Dinomaly [24] and Dinomaly2 [25]. Together, these nine baselines cover distillation, synthetic-anomaly learning, unified reconstruction, diffusion, state-space modeling, and foundation-model reconstruction. For consistency across the three benchmarks, the overlapping five-method baseline subset is also compared in one consolidated table; other methods are included as additional dataset-specific comparisons.

Results on MVTec AD and VisA. Table 2 summarizes the unified multi-class comparison. The published Dinomaly2 benchmark shows that foundation-model reconstruction has already pushed image-level performance close to saturation on MVTec AD, making pixel-level localization a more informative discriminator. HiFRA-AD preserves this strong image-level discrimination while improving P-AP and AUPRO. On MVTec AD, the gain is concentrated in dense localization metrics rather than I-AUROC, consistent with the design goal of recovering weak local evidence instead of increasing category-level semantic capacity. VisA exhibits the same pattern: localization improves more than image-level discrimination.

Table 2. Multi-class UAD comparison on MVTec AD and VisA (%). Values are category-wise averages; higher is better.

MVTec AD				
Method	**I-AUROC**	**P-AUROC**	**P-AP**	**AUPRO**
RD4AD [11]	94.6	96.1	48.6	91.1
SimpleNet [17]	95.3	96.9	45.9	86.5
DeSTSeg [29]	89.2	93.1	54.3	64.8
UniAD [16]	96.5	96.8	43.4	90.7
ReContrast [12]	98.3	97.1	60.2	93.2
DiAD [31]	97.2	96.8	52.6	90.7
MambaAD [21]	98.6	97.7	56.3	93.1
Dinomaly [24]	99.6	98.4	69.3	94.8
Dinomaly2-B [25]	99.8	98.4	69.3	94.8
HiFRA-AD (ViT-S)	**99.8**	**98.7**	**71.6**	**95.5**
VisA				
Method	**I-AUROC**	**P-AUROC**	**P-AP**	**AUPRO**
RD4AD [11]	92.4	98.1	38.0	91.8
SimpleNet [17]	87.2	96.8	34.7	81.4
DeSTSeg [29]	88.9	96.1	39.6	67.4
UniAD [16]	88.8	98.3	33.7	85.5
ReContrast [12]	95.5	98.5	47.9	91.9
DiAD [31]	86.8	96.0	26.1	75.2
MambaAD [21]	94.3	98.5	39.4	91.0
Dinomaly [24]	98.7	98.7	53.2	94.5
Dinomaly2-B [25]	99.2	99.1	52.9	95.5
HiFRA-AD (ViT-S)	**99.3**	**99.2**	**55.1**	**96.1**

Bold numerical entries indicate the best results within each dataset; ties are included. The bold method label identifies HiFRA-AD.

Figure 2 complements the aggregate metrics with representative localization maps from MVTec AD. It compares Dinomaly2 and HiFRA-AD under the same visualization protocol and shows how the refinement changes spatial responses across textures and objects. The grid and tile examples reveal interrupted repetitive structures, while the capsule and screw cases illustrate fine local surface or boundary changes; these examples also support the symmetry interpretation in Section 3.3.

The MVTec AD comparison between HiFRA-AD-S and Dinomaly2-B yields P-AP and AUPRO differences of +2.3 and +0.7 percentage points, with equal 99.8% I-AUROC. This cross-backbone benchmark comparison complements the matched-backbone evaluation in Section 4.4.

Results on MVTec AD 2. MVTec AD 2 is particularly relevant to HiFRA-AD because it places greater emphasis on difficult imaging conditions and subtle defects. Table 3 collects the reported aggregate values. Unlike the near-saturated image-level results on MVTec AD, this benchmark leaves substantially more room for improvement. All methods in Table 3 use the official 2045-image $\text{TEST}_{\text{priv}}$ evaluation partition and $\text{AUPRO}_{0.30}$, with identical server-side metric implementation and category aggregation.

Common baseline comparison. Table 4 reports RD4AD, ReContrast, MambaAD, Dinomaly, and Dinomaly2-B across all three datasets, using the P-AP/AUPRO entries from the full comparison tables. Additional methods remain in their dataset-specific tables. HiFRA-AD-S uses ViT-S throughout. The MVTec AD 2 entries share the official 2045-image TESTpriv evaluation and AU-PRO@0.30.

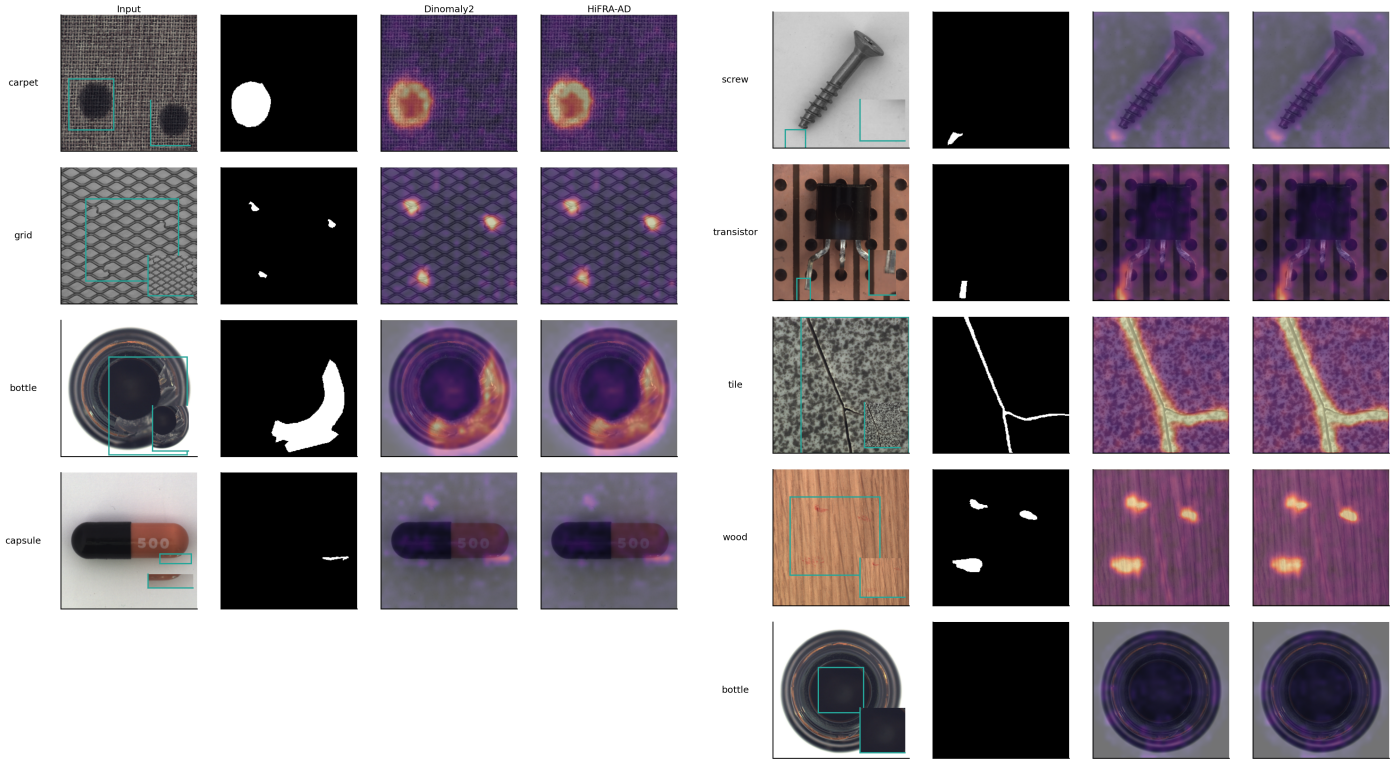


Figure 2. Qualitative anomaly localization on representative MVTec AD categories. Columns show the input image, ground-truth anomaly mask, Dinomaly2 response, and HiFRA-AD response under the same visualization protocol. Repeated-texture and fine-boundary defects are presented together in this consolidated figure. The teal rectangles identify the image regions shown in the magnified insets. In the binary ground-truth masks, white denotes anomalous pixels and black denotes the remaining pixels. Response overlays progress from dark purple for weaker anomaly responses to orange/yellow for stronger responses.

Table 3. MVTec AD 2 comparison (%) on the official 2045-image TESTpriv partition. All methods were evaluated through the benchmark server with identical metric implementation and category averaging. AUPRO denotes AU-PRO@0.30. Higher is better.

Method	I-AUROC	P-AUROC	P-AP	AUPRO
PatchCore [5]	86.8	92.1	31.4	46.5
RD4AD [11]	85.4	91.3	29.8	44.8
EfficientAD [18]	89.9	93.7	36.2	50.7
ReContrast [12]	91.3	94.4	39.5	53.8
MambaAD [21]	90.8	94.0	38.7	52.1
Dinomaly [24]	92.6	94.9	42.1	55.3
Dinomaly2-B [25]	93.4	95.4	44.0	57.1
HiFRA-AD-S (ViT-S)	**94.2**	**96.0**	**46.8**	**59.8**

Bold numerical entries indicate the best result in each metric column. The bold method label identifies HiFRA-AD-S.

Figure 3 traces the corresponding refinement process from global proposals through calibrated local evidence to the fused anomaly map.

The larger gain in region-sensitive metrics than in I-AUROC supports the central hypothesis: frequency-sensitive reconstruction and sparse re-sampling primarily strengthen fine-grained spatial evidence.

Table 4. Common-baseline localization comparison across three datasets (%). Identical baseline methods are shown for every dataset; P-AP and AUPRO are copied from the full-metric tables. MVTec AD 2 values are from official-server TESTpriv evaluation (2045 images), with AU-PRO@0.30. Higher is better.

	MVTec AD		VisA		MVTec AD 2	
Method	**P-AP**	**AUPRO**	**P-AP**	**AUPRO**	**P-AP**	**AUPRO**
RD4AD [11]	48.6	91.1	38.0	91.8	29.8	44.8
ReContrast [12]	60.2	93.2	47.9	91.9	39.5	53.8
MambaAD [21]	56.3	93.1	39.4	91.0	38.7	52.1
Dinomaly [24]	69.3	94.8	53.2	94.5	42.1	55.3
Dinomaly2-B [25]	69.3	94.8	52.9	95.5	44.0	57.1
HiFRA-AD-S	**71.6**	**95.5**	**55.1**	**96.1**	**46.8**	**59.8**

Bold numerical entries indicate the best result in each dataset/metric column. The bold method label identifies HiFRA-AD-S. Short horizontal rules group the two metric columns for each dataset.

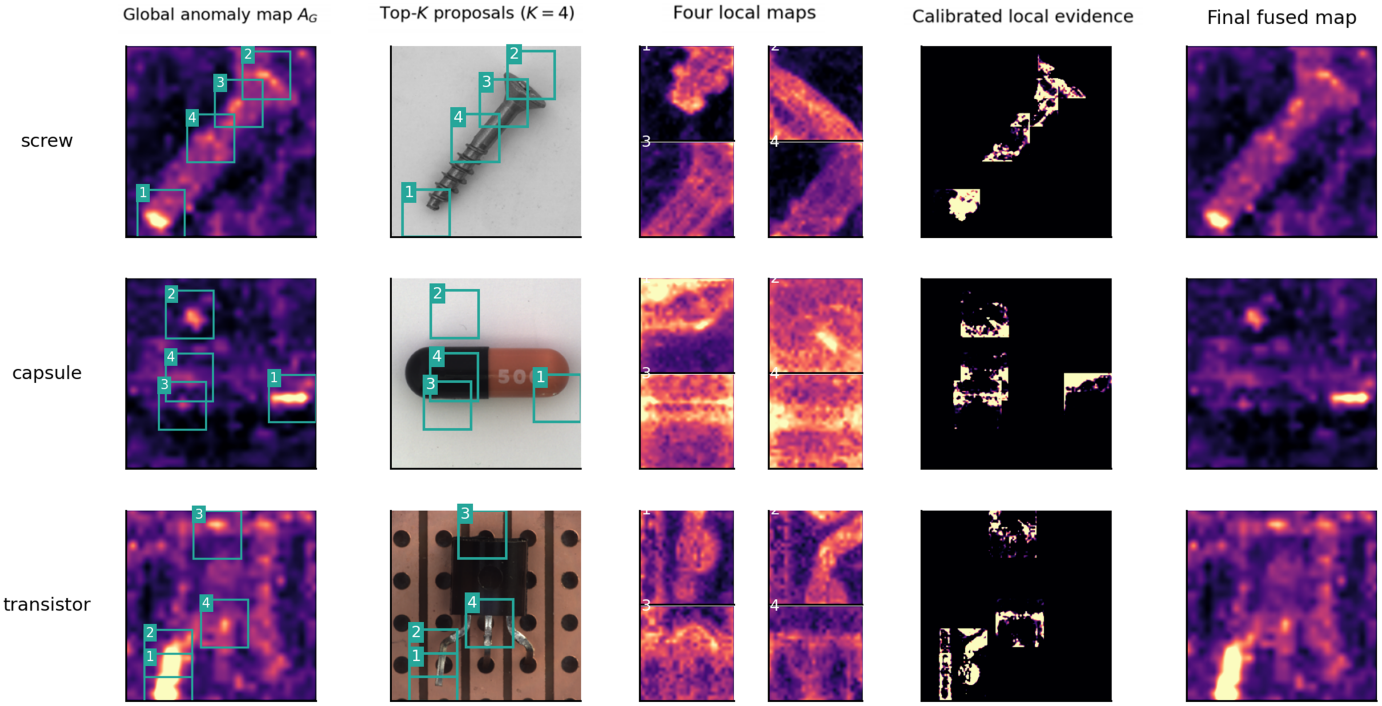


Figure 3. Sparse high-resolution refinement on representative MVTec AD samples. The global anomaly map proposes four peak-suppressed ROIs, which are re-evaluated locally, calibrated to the corresponding global regions, and fused into the final anomaly map. Teal rectangles numbered 1–4 identify the proposed ROIs and correspond to the numbered local maps. Brighter orange/yellow colors indicate stronger responses, while dark purple indicates weaker responses; black regions in the calibrated-evidence column receive no positive local update.

4.3. Statistical Inference for Close-Performing Competitors

The backbone-matched MVTec AD comparison yields a mean P-AP gain of 4.10 percentage points (paired 95% CI [3.95, 4.25], Holm-adjusted $p = 1.53 \times 10^{-6}$) and an AUPRO gain of 1.10 points (95% CI [0.98, 1.22], adjusted $p = 8.88 \times 10^{-5}$). These tests use the five seed-aligned observations and the eight-test adjustment family described in Section 4.5. Table 5 also summarizes the descriptive differences against external aggregate benchmark results.

The five-seed analysis measures training variability for the matched ViT-S comparison on MVTec AD. External aggregate comparisons, including the 0.70-point MVTec AD and 0.60-point VisA AUPRO margins over Dinomaly2-B, have descriptive evidence. Paired

external inference requires corresponding category scores or aligned image predictions. Seed-, category-, and image-based analyses quantify different sources of variation.

For run s, the dataset score is $M^{(s)} = C^{-1} \sum_{c=1}^{C} M_c^{(s)}$, with M denoting P-AP or AUPRO. The available repeated-run summaries establish aggregate MVTec AD variability; a complete category-wise comparison requires the underlying category exports. The defect-subtype analysis in Section 4.7 adds category-balanced per-image AP evidence for four categories, using a different aggregation from the benchmark P-AP.

Table 5. Localization comparison with the nearest competitors (%). Score pairs are given as P-AP/AUPRO, and the deltas are in percentage points (HiFRA-AD-S minus comparator). Only the ViT-S matched-backbone row has five-seed paired significance evidence; other entries are descriptive aggregate benchmarks. The MVTec AD 2 entry uses the official-server TESTpriv protocol with AUPRO@0.30, but has no paired significance analysis.

Dataset	Comparator	Baseline	HiFRA-AD	Δ (P-AP/AUPRO)	Evidence
MVTec AD	Dinomaly2-S (ViT-S)	67.5/94.40	71.6/95.50	+4.10/+1.10	5-seed paired
MVTec AD	Dinomaly2-B (ViT-B)	69.3/94.80	71.6/95.50	+2.30/+0.70	Aggregate
VisA	Dinomaly	53.2/94.50	55.1/96.10	+1.90/+1.60	Aggregate
VisA	Dinomaly2-B (ViT-B)	52.9/95.50	55.1/96.10	+2.20/+0.60	Aggregate
MVTec AD 2	Dinomaly2-B (ViT-B)	44.0/57.10	46.8/59.80	+2.80/+2.70	Aggregate

4.4. Backbone-Controlled Comparison and Computational Efficiency

Matched-backbone performance. Table 6 holds backbone capacity and training settings fixed within each pair. HiFRA-AD-S raises MVTec AD P-AP from 67.5% to 71.6% and AUPRO from 94.40% to 95.50%. HiFRA-AD-B raises P-AP from 69.3% to 72.8% and AUPRO from 94.80% to 95.82%. These gains occur at unchanged parameter counts within each backbone pair.

Table 6. Backbone-controlled comparison on MVTec AD (%). The training and evaluation protocols are identical within each matched backbone pair. The ViT-S mean-score repeatability is presented in Section 4.5; the ViT-B comparison follows the original matched-run evaluation. Higher is better.

Method	Backbone	I-AUROC	P-AUROC	P-AP	AUPRO
Dinomaly2-S [25]	ViT-S/14	99.55	98.31	67.5	94.40
HiFRA-AD-S	ViT-S/14	**99.80**	**98.70**	**71.6**	**95.50**
Dinomaly2-B [25]	ViT-B/14	99.80	98.40	69.3	94.80
HiFRA-AD-B	ViT-B/14	**99.83**	**98.78**	**72.8**	**95.82**

Bold numerical entries indicate the better result within each matched backbone pair. Bold method labels identify the proposed HiFRA-AD variants.

Computational characteristics. High-frequency residual enhancement and ROI refinement share the pretrained encoder and reconstruction weights without introducing additional trainable model parameters. For four selected regions, inference carries out one full-image forward pass and four resized ROI forward passes. The network cost is $C_{\text{network}} = (1 + K)C_{\text{single}} = 5C_{\text{single}}$, whereas the complete procedure additionally includes ROI proposals and calibration/fusion: $C_{\text{total}} = C_{\text{network}} + C_{\text{proposal}} + C_{\text{calibration/fusion}}$.

Table 7 reports localization accuracy with absolute processing costs under the same single-RTX-4090, input-size, and precision settings. MACs count reconstruction-network evaluations; FLOPs use two operations per multiply–accumulate. End-to-end latency includes proposal selection, the global and local forward passes, calibration, and fusion. Throughput is $1000/t_{\text{ms}}$ at batch size one. Batching the four local crops allows the 5× network-operation budget to incur a 2.08× ViT-S latency increase. Section 4.5 reports the ROI-budget sensitivity.

Matched-backbone accuracy–cost trade-off. HiFRA-AD-S gains 4.1 P-AP points with an additional 10.8 ms per image (10.0 to 20.8 ms); HiFRA-AD-B gains 3.5 points with an additional 25.3 ms (19.5 to 44.8 ms). Throughput changes from 100.0 to 48.1 images/s for ViT-S and from 51.3 to 22.3 images/s for ViT-B. The five-pass inference retains model parameter counts and increases memory as shown in Table 8.

Table 7. Matched-backbone accuracy and computational efficiency on MVTec AD (single RTX 4090). Latency is measured end-to-end at batch size 1; throughput is its reciprocal. MACs/FLOPs count network passes. S/B indicates ViT-S/ViT-B.

Method	P-AP (%)	Params (M)	MACs (G)	FLOPs (G)	Train (h)	Latency (ms)	Throughput (img/s)
Dinomaly2-S	67.5	37.3	26.2	52.4	7.3	10.0	100.0
HiFRA-AD-S	71.6	37.3	131.0	262.0	7.6	20.8	48.1
Dinomaly2-B	69.3	146.6	103.5	207.0	18.6	19.5	51.3
HiFRA-AD-B	72.8	146.6	517.5	1035.0	19.1	44.8	22.3

Table 8. Peak GPU memory for the matched-backbone configurations on one RTX 4090.

Method	Training Memory (GB)	Inference Memory (GB)
Dinomaly2-S	12.5	2.1
HiFRA-AD-S	12.7	5.5
Dinomaly2-B	22.3	4.6
HiFRA-AD-B	22.4	11.9

Sparse refinement versus exhaustive local scanning. For a fixed local side ratio of $\rho = 0.25$, exhaustive non-overlapping coverage of a square global canvas requires $N_{\text{dense}} = (1/\rho)^2 = 16$ local crops; a half-ROI stride requires $N_{\text{overlap}} = \left(\lfloor (1-\rho)/(\rho/2) \rfloor + 1\right)^2 = 49$ local crops. Including one global evaluation, these settings correspond to 17 and 50 model forwards, respectively, compared with five for HiFRA-AD at $K = 4$. Table 9 presents exact network-pass counts and corresponding MAC budgets for the matched ViT-S reconstruction model. These entries quantify network operation budgets.

Table 9. Analytical inference-work comparison for identical ViT-S/14 reconstruction passes and ROI side ratio $\rho = 0.25$. Network MACs exclude ROI selection and map fusion; dense-scan entries are analytical operation budgets.

Method	Global	Local Crops	Total Passes	Network MACs (G)
Dinomaly2-S (global)	1	0	1	26.2
HiFRA-AD-S ($K = 4$)	1	4	5	131.0
Dense non-overlap (stride ρ)	1	16	17	445.4
Dense 50% overlap (stride $\rho/2$)	1	49	50	1310.0

Relative to non-overlapping and half-overlapping exhaustive scans, sparse selection reduces network evaluations by 70.6% and 90.0%, respectively. These analytical savings depend on the proposal budget, while wall-clock time also depends on batching and preprocessing. Dense scans provide broader spatial coverage; the sparse policy exchanges coverage for fewer model evaluations.

Relation to HiAD. HiAD [54] combines global–local multi-resolution inference with detector-pool assignment on high-resolution benchmarks. Its published HiAD/DeSTSeg setting reports 1.79 s per 4K image on four RTX 4090 GPUs. The differing resolution, detector, and hardware define a separate operating point from our 392-pixel, single-GPU

measurements. The identical-detector scan budgets above isolate the computational effect of sparse region selection.

Accuracy and efficiency. The fixed ROI budget provides a configurable operating point for local-detail recovery. With four ROIs, inference memory rises from 2.1 to 5.5 GB for ViT-S and from 4.6 to 11.9 GB for ViT-B. The accuracy, latency, and memory measurements together characterize this trade-off.

4.5. Ablation Studies

We conduct ablation experiments on MVTec AD under the same one-for-all training protocol to isolate the contribution of each component. All variants use the same DINOv2 ViT-S/14 encoder, reconstruction decoder, input size, training schedule, and post-processing. Table 10 therefore attributes the measured differences to the frequency residual, sparse ROI refinement, and residual calibration.

Table 10. Component ablation on MVTec AD (%). The HF + ROI configuration isolates the calibration contribution conditional on high-frequency enhancement and sparse refinement. Five-run stability for the configurations in the repeated-run study is reported separately below.

Variant	HF	ROI	Cal	I-AUROC	P-AUROC	P-AP	AUPRO
Base				99.55	98.31	67.5	94.40
Base + HF	✓			99.58	98.43	68.8	94.72
Base + ROI		✓		99.57	98.45	69.2	94.81
Base + ROI + Cal		✓	✓	99.59	98.52	70.0	95.05
Base + HF + ROI	✓	✓		99.61	98.58	70.7	95.20
HiFRA-AD	✓	✓	✓	**99.80**	**98.70**	**71.6**	**95.50**

HF: high-frequency residual enhancement; ROI: sparse local re-evaluation; Cal: residual-aware calibration. Check marks indicate enabled components. Bold numerical entries indicate the best result in each metric column; the bold method label identifies the complete method.

Component analysis.

The Base variant corresponds to the backbone-matched Dinomaly2-S reconstruction baseline in Table 6, with the same ViT-S/14 encoder and training protocol.

Base + HF adds the high-frequency residual enhancement. Base + ROI introduces sparse high-resolution re-evaluation with direct local–global fusion. Base + ROI + Cal adds residual-aware calibration without high-frequency enhancement. Base + HF + ROI evaluates the uncalibrated combination, allowing a direct matched-configuration isolation of calibration in the complete pipeline. HiFRA-AD combines all three components.

Calibration control. Base + HF + ROI uses the HF-trained model and the same proposals ($K = 4$, $\rho = 0.25$), with direct local–global fusion. It achieves 99.61% I-AUROC, 98.58% P-AUROC, 70.7% P-AP, and 95.20% AUPRO. Enabling calibration raises P-AP by 0.90 points and AUPRO by 0.30 points, isolating its contribution with HF and ROI active. Calibration operates on paired local/global responses and is therefore evaluated within the ROI branch.

Table 11 summarizes the isolated metric changes from enabling calibration with the other components held fixed.

Table 11. Isolated change caused by residual calibration on MVTec AD (percentage points). Positive values indicate improvement when Cal is enabled while the other components are held fixed.

Controlled Contrast	Δ I-AUROC	Δ P-AUROC	Δ P-AP	Δ AUPRO
ROI → ROI + Cal	+0.02	+0.07	+0.80	+0.24
HF + ROI → HF + ROI + Cal	+0.19	+0.12	+0.90	+0.30

Stability across five independent runs. Table 12 summarizes the five-run variation for the MVTec AD component analysis, with means and sample standard deviations reported for each configuration. The values are dataset-level category-averaged results: one independently trained model, rather than an individual image or pixel, is the unit of replication.

Across five seeds, HiFRA-AD-S improves over Base by 0.25 percentage points in I-AUROC, 4.10 points in P-AP, and 1.10 points in AUPRO. The 95% marginal confidence intervals for the mean P-AP are [66.95, 68.05] (Base) and [71.06, 72.14] (HiFRA-AD-S), whereas those for mean AUPRO are [94.24, 94.56] and [95.41, 95.59], respectively. These marginal intervals describe each method separately; the more relevant uncertainty of the seed-matched improvement is estimated from the paired differences reported below.

Table 12. Five-run MVTec AD component stability (%, mean ± sample standard deviation; $n = 5$ independent training seeds) for the five configurations in the repeated-run analysis. Higher is better.

Method	I-AUROC ↑	P-AP ↑	AUPRO ↑
Base	99.55 ± 0.04	67.50 ± 0.44	94.40 ± 0.13
Base + HF	99.58 ± 0.04	68.80 ± 0.42	94.72 ± 0.11
Base + ROI	99.57 ± 0.06	69.20 ± 0.37	94.81 ± 0.11
Base + ROI + Cal	99.59 ± 0.05	70.00 ± 0.31	95.05 ± 0.10
HiFRA-AD-S	**99.80 ± 0.05**	**71.60 ± 0.44**	**95.50 ± 0.07**

Bold numerical entries identify the highest mean scores and include their accompanying sample standard deviations. The bold method label identifies HiFRA-AD-S. Upward arrows indicate that higher scores are better.

Seed-aligned significance analysis. For reproducibility, Table 13 provides the five matched MVTec AD scores for the principal Base versus HiFRA-AD-S comparison. Table 14 reports the resulting paired mean differences, 95% confidence intervals, two-sided p-values and Holm-adjusted p-values. For each comparison, the unit of analysis is one independently trained seed; pairwise differences are computed prior to significance testing.

Table 13. Five seed-aligned MVTec AD results (%) for Base and HiFRA-AD-S. Identical seed IDs denote paired independent training runs.

Seed	Base I-AUROC	HiFRA I-AUROC	Base P-AP	HiFRA P-AP	Base AUPRO	HiFRA AUPRO
42	99.48	99.73	66.90	70.90	94.22	95.43
123	99.57	99.84	67.90	71.90	94.53	95.59
3407	99.54	99.81	67.60	71.70	94.34	95.54
2026	99.60	99.85	67.20	71.50	94.51	95.52
4444	99.56	99.77	67.90	72.00	94.40	95.42

For the near-ceiling I-AUROC endpoint, the HiFRA-AD-S versus Base paired difference is $\Delta = 0.25$ percentage points (paired 95% CI [0.22, 0.28]; exploratory two-sided, unadjusted $p = 2.18 \times 10^{-5}$). This secondary test is not part of the eight-test Holm family; its small absolute effect should be interpreted in light of image-level performance saturation. All eight prespecified localization contrasts remain significant after Holm adjustment at 0.05 under the stated paired-t assumptions.

Comparing Base with Base + HF evaluates whether the fixed-coefficient residual transformation without extra learned weights improves reconstruction sensitivity without changing spatial resolution. The gain is modest at image level but clearer in P-AP, consistent with the module's focus on localized variation. Comparing Base + ROI with Base + ROI + Cal isolates the effect of local-to-global calibration. Direct ROI re-evaluation can amplify both defects and benign local changes; aligning the local map to the correspond-

ing global patch reduces this context-induced score shift before fusion. HiFRA-AD combines frequency enhancement, sparse re-sampling, and calibration and yields the strongest localization result, indicating that the components act at different stages of the same information-loss problem.

Table 14. Paired statistical tests for MVTec AD component comparisons. Differences are in percentage points; two-sided paired t-tests ($n = 5$, $df = 4$), 95% confidence intervals, and Holm-adjusted p-values over eight prespecified localization tests.

Contrast	Metric	Mean Δ (pp)	95% Paired CI (pp)	Raw p	Holm p
HF vs. Base	P-AP	1.30	[1.18, 1.42]	8.34×10^{-6}	5.84×10^{-5}
HF vs. Base	AUPRO	0.32	[0.23, 0.41]	6.57×10^{-4}	2.11×10^{-3}
ROI vs. Base	P-AP	1.70	[1.24, 2.16]	5.29×10^{-4}	2.11×10^{-3}
ROI vs. Base	AUPRO	0.41	[0.34, 0.48]	1.01×10^{-4}	5.06×10^{-4}
ROI+Cal vs. ROI	P-AP	0.80	[0.48, 1.12]	2.17×10^{-3}	4.07×10^{-3}
ROI+Cal vs. ROI	AUPRO	0.24	[0.15, 0.33]	2.04×10^{-3}	4.07×10^{-3}
HiFRA-S vs. Base	P-AP	4.10	[3.95, 4.25]	1.91×10^{-7}	1.53×10^{-6}
HiFRA-S vs. Base	AUPRO	1.10	[0.98, 1.22]	1.48×10^{-5}	8.88×10^{-5}

Calibration effect. Table 11 gives two conditional contrasts. Adding calibration to ROI raises P-AP by 0.80 points and AUPRO by 0.24 points; with HF active, the increases are 0.90 and 0.30 points. The latter P-AP increment accounts for approximately 22% of the full 4.10-point gain over Base. The ROI-only five-seed contrast yields Holm-adjusted $p = 0.00407$ for both localization endpoints. The HF + ROI contrast is a configuration-level measurement. Calibration adds ROI score statistics, gating, and fusion while reusing the same reconstruction passes.

Calibration and global contrast. The bounds in Section 3.5 relate correction magnitude to the global patch's contrast. A constant patch permits at most a 2ε increment per singly covering ROI. This explains the conservative behavior of calibrated fusion in low-contrast regions.

Figure 4 visualizes the fused token field, its low- and high-frequency components, the enhanced representation, and the final anomaly map for representative samples.

Sensitivity to the frequency coefficient. Table 15 varies α while holding the remaining settings fixed. Category-mean P-AP and AUPRO peak at 71.6% and 95.50% for $\alpha = 0.2$. Smaller coefficients weaken enhancement; larger coefficients can amplify benign texture variation. The $\alpha = 0$ row matches Base + ROI + Cal in Table 10.

Sensitivity to the number of ROIs. Table 16 fixes $\rho = 0.25$ and varies K. The $K = 0$ row matches Base + HF in Table 10. Localization improves from one to four proposals and then saturates, while inference time continues to increase. Section 4.6 measures the corresponding defect-coverage behavior.

Selection provenance. The sweeps in Tables 15 and 16 use annotated MVTec AD evaluation images for model development. The selected $\alpha = 0.2$ and $K = 4$ are shared across categories and transferred unchanged to the other datasets, as specified in Section 4.1.

Table 15. Sensitivity to the frequency coefficient α on MVTec AD (%).

α	I-AUROC	P-AUROC	P-AP	AUPRO
0.0	99.59	98.52	70.0	95.05
0.1	99.61	98.60	70.8	95.24
0.2	**99.80**	**98.70**	**71.6**	**95.50**
0.3	99.61	98.61	71.0	95.31
0.5	99.57	98.48	69.9	95.02

The bold coefficient identifies the selected setting; bold metric values are the best results in their respective columns.

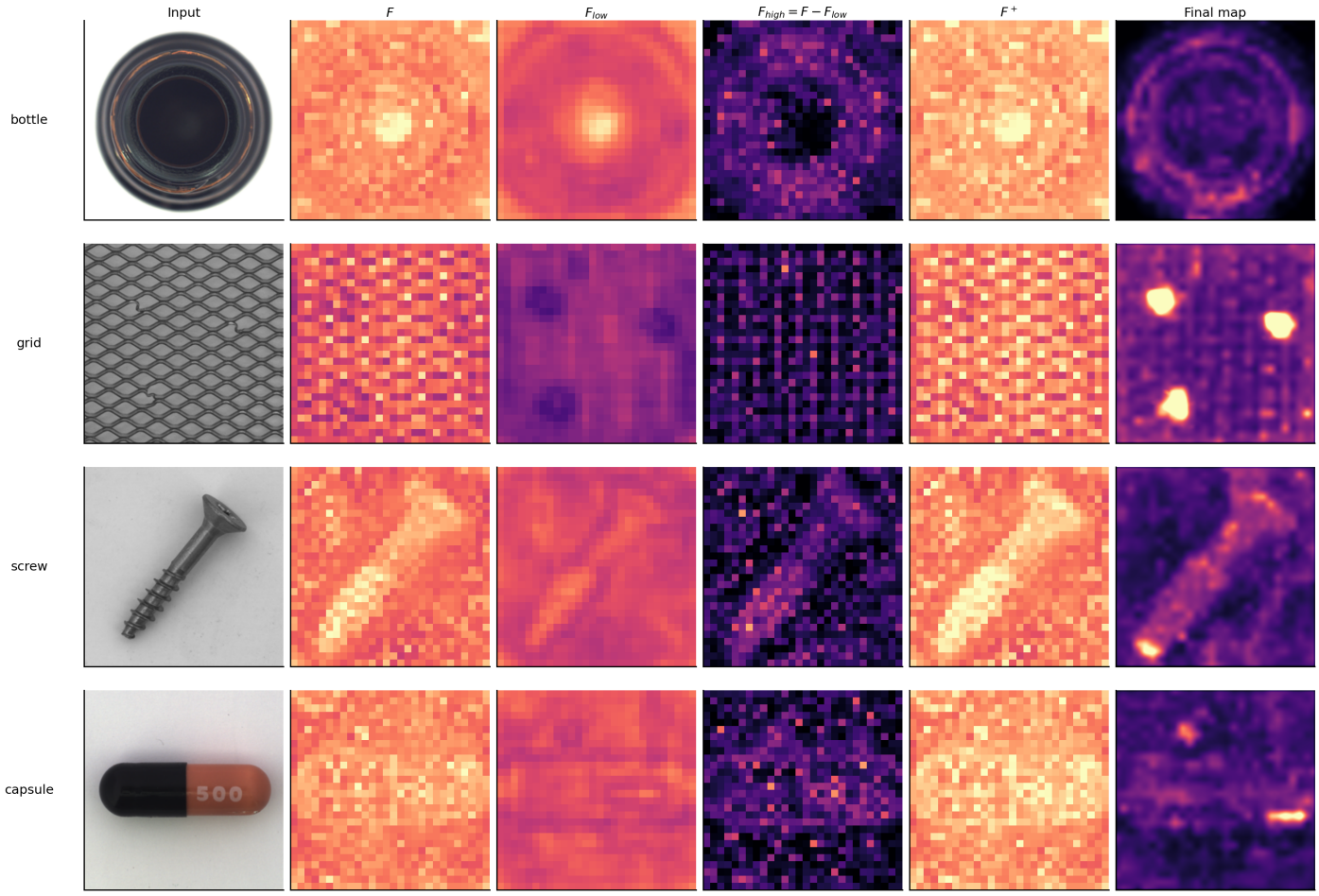


Figure 4. High-frequency residual analysis on representative MVTec AD samples. From **left** to **right**: input, fused token field F, locally smoothed component F_{low}, residual component F_{high}, enhanced field F^+, and final anomaly map. False colors in the feature panels visualize spatial variation in the respective representations rather than ground-truth anomaly labels. In the final anomaly maps, dark purple indicates weaker responses and bright orange/yellow indicates stronger responses.

Table 16. Effect of the number of local proposals K.

K	P-AUROC	P-AP	AUPRO	Relative Inference Time
0	98.43	68.8	94.72	1.00×
1	98.51	69.7	94.96	1.35×
2	98.59	70.6	95.19	1.63×
4	**98.70**	**71.6**	**95.50**	2.08×
6	98.69	71.5	95.44	2.71×

The bold proposal count identifies the selected budget; bold metric values are the best localization scores. Relative inference time is measured against the global-only setting ($K = 0$); lower inference time is better.

4.6. Ground-Truth Coverage of Sparse ROI Selection

Proposal-coverage evaluation. We evaluate which ground-truth defects enter the ROI union before calibration or fusion. The evaluation uses the inference pipeline's unsmoothed global map, greedy rectangular suppression, inverse coordinate transform, and $\rho = 0.25$. Ground-truth masks measure coverage offline; proposal selection uses the global anomaly response.

Defect-level overlap and adequate-coverage recall. Let $D_{i\ell} \subseteq \{1, \dots, W_{0,i}\} \times \{1, \dots, H_{0,i}\}$ be the ℓ-th connected ground-truth defect component in test image i, and let $U_i^{(K)}$ be the

union of the K selected rectangles after mapping them back to the original image. The defect-level pixel coverage and adequate-coverage indicator are

$$c_{i\ell}^{(K)} = \frac{|D_{i\ell} \cap U_i^{(K)}|}{|D_{i\ell}|}, \qquad h_{i\ell}^{(K)} = 1\left[c_{i\ell}^{(K)} \geq 0.5\right]. \tag{38}$$

Within each subgroup, mean coverage is the mean of $c_{i\ell}^{(K)}$, and adequate-coverage recall is the mean of $h_{i\ell}^{(K)}$. Every ground-truth component remains eligible for size-stratified analysis, including components outside the globally visible center crop, for which $c_{i\ell}^{(K)} = 0$; only components having projected pixels inside the crop can receive an initial-response rank. This distinction prevents inflation of area-stratified coverage by silently excluding nonvisible defects. Overlapping ROIs are counted once through their union $U_i^{(K)}$.

Defect size and initial response. The relative defect area is $a_{i\ell} = |D_{i\ell}|/(H_{0,i}W_{0,i})$. The area strata are small ($a < 0.25\%$), medium ($0.25\% \leq a < 1\%$), and large ($a \geq 1\%$). For each defect visible after global resizing and center cropping, let $t_{i\ell}$ be the 90th percentile of its unsmoothed global response. Its within-image mid-percentile rank is

$$r_{i\ell} = \frac{1}{|\Omega_{A,i}|} \sum_{p \in \Omega_{A,i}} \left[1\big(A_{G,i}(p) < t_{i\ell}\big) + \tfrac{1}{2}1\big(A_{G,i}(p) = t_{i\ell}\big)\right]. \tag{39}$$

The response groups are weak ($r < 0.75$), intermediate ($0.75 \leq r < 0.90$), and strong ($r \geq 0.90$). A constant map has midrank 0.5 and falls in the weak group. Defects outside the global center crop retain zero coverage in the area analysis and have undefined response rank. We also report the intersection of small and weak defects. These prespecified strata describe proposal behavior.

Geometric proposal budget. For $K = 4$ and $\rho = 0.25$, the nominal union area of the global selected boxes is at most $K\rho^2 = 25\%$ of the 392-pixel global crop. Accounting for the preceding 448-pixel resize canvas, this corresponds to approximately $K\rho^2(C/S)^2 = 19.1\%$ of the full original image area before integer rounding, boundary clipping, and box overlap. This spatial-computation budget covers a selected portion of the visible crop; defect coverage depends on the locations and rankings of the proposed regions.

Coverage results and interpretation. Table 17 quantifies the proposal stage independently of final detection scores. For the default four ROIs, mean ground-truth defect coverage rises from 45.8% for small defects to 78.5% for large ones; the corresponding adequate-coverage recalls are 52.3% and 87.4%. The dependence on initial global-map evidence is more pronounced: mean coverage is 42.1% for weak responses versus 90.2% for strong responses, with adequate-coverage recall of 52.0% versus 92.5%. The joint small-and-weak subgroup is most challenging, at 28.5% mean coverage and 33.3% adequate-coverage recall. Area and response strata overlap and describe different conditional subsets.

Sparse refinement is most effective when defect evidence survives global resizing. Defects outside the selected ROIs retain the global prediction, while covered defects receive an opportunity for additional local discrimination. The small-and-weak subgroup's 28.5% coverage identifies a principal source of missed refinement in this proposal policy.

Structural regularity. Figure 2 shows local interruptions of repeated grid and tile patterns alongside surface and boundary defects. Section 4.7 complements these examples with a within-category comparison of structural-disruption and other anomaly subtypes.

Table 17. Ground-truth coverage of sparse ROI selection on MVTec AD for the default $K = 4$ proposals ($\rho = 0.25$). Coverage is the mean proportion of each connected defect component inside the ROI union; adequate-coverage recall is the percentage of components with at least 50% of pixels covered. Groups are formed from defect area and the initial, pre-refinement anomaly-map rank.

Grouping	Defect Subgroup	Coverage (%) ↑	Recall@50% (%) ↑
Area	Small (<0.25%)	45.8	52.3
Area	Medium (0.25–1%)	63.1	69.7
Area	Large (≥1%)	78.5	87.4
Initial response	Weak ($r < 0.75$)	42.1	52.0
Initial response	Intermediate ($0.75 \leq r < 0.90$)	65.0	74.0
Initial response	Strong ($r \geq 0.90$)	90.2	92.5
Joint subgroup	**Small and weak**	**28.5**	**33.3**

Upward arrows indicate that higher coverage and recall are better. Bold entries draw attention to the small-and-weak joint subgroup and do not denote the highest values.

4.7. Symmetry-Related Defect Subgroup Evaluation

Operational defect grouping. We conduct a focused MVTec AD evaluation using the official defect-subtype folders in four product categories. The structural-disruption group contains grid/bent and grid/broken; zipper/broken_teeth, zipper/split_teeth, and zipper/squeezed_teeth; screw/thread_side and screw/thread_top; and transistor/bent_lead and transistor/cut_lead. These subtypes interrupt repeated lines, teeth, threads, or component arrangements. The comparison group includes the remaining anomalous subtypes from the same four product categories. The groups are defined from defect semantics rather than model outputs, and the matched comparison uses the same test images and ground-truth masks for both models.

Paired metric and category-balanced aggregation. For anomalous test image i, pixel-level average precision (AP) is evaluated over its ground-truth mask from the spatially aligned output of Dinomaly2-S and HiFRA-AD-S. The paired gain in percentage points is $d_i = \mathrm{AP}_{\mathrm{HiFRA},i} - \mathrm{AP}_{\mathrm{Dinomaly2\text{-}S},i}$. We first average gains within each operational group and product category and then weight the four product categories equally. The contrast of interest is

$$\Gamma = \bar{d}_{\text{structural disruption}} - \bar{d}_{\text{other anomaly}}. \tag{40}$$

Findings. Table 18 records paired per-image AP gains of 4.99 points for structural disruptions and 2.68 points for other defects in the same four product categories, giving a 2.31-point contrast. The larger gain is consistent with enhanced local-detail sensitivity on interrupted regular structures. Defect scale, contrast, and local geometry may also contribute to this descriptive subgroup difference.

Table 18. Symmetry-related defect-subtype analysis on MVTec AD. Values are category-balanced, paired per-image pixel-AP improvements of HiFRA-AD-S relative to Dinomaly2-S, in percentage points (pp). The subtype-level AP aggregation is distinct from the pooled category P-AP in the main benchmark.

Defect Group	AP Gain (pp) ↑
Structural-disruption subtypes	**+4.99**
Other anomalous subtypes (same product categories)	+2.68
Between-group gain difference Γ	**+2.31**

The upward arrow indicates that a larger positive AP gain is better. Bold values emphasize the structural-disruption gain and the between-group contrast, rather than column maxima. A positive contrast indicates a larger gain for structural disruptions.

Overall, the ablation study verifies three claims. First, local token residual enhancement contributes independently of high-resolution inference. Second, sparse ROI re-evaluation recovers additional spatial evidence without scanning the complete image. Third, both the ROI-only comparison and the HF + ROI control support the contribution of residual calibration, with an additional 0.9-point P-AP and 0.30-point AUPRO improvement in the HF + ROI setting. The full model combines these effects while retaining the original one-for-all reconstruction training paradigm.

5. Conclusions

HiFRA-AD combines fixed-coefficient high-frequency token enhancement with sparse high-resolution re-evaluation for one-for-all industrial anomaly detection. The residual operator strengthens local spatial variation, and calibrated nonnegative fusion integrates evidence from four selected regions. Its token-grid equivariance connects the operator to the preservation of nominal structural regularity.

HiFRA-AD-S achieves P-AP/AUPRO of 71.6/95.5 on MVTec AD, 55.1/96.1 on VisA, and 46.8/59.8 on the official MVTec AD 2 TESTpriv partition with AU-PRO@0.30. Backbone-matched MVTec AD comparisons yield P-AP gains of 4.1 points for ViT-S and 3.5 points for ViT-B. Five paired ViT-S runs give a P-AP gain of 4.10 points (95% CI [3.95, 4.25], Holm-adjusted $p = 1.53 \times 10^{-6}$) and an AUPRO gain of 1.10 points (95% CI [0.98, 1.22], adjusted $p = 8.88 \times 10^{-5}$).

The additional local evidence incurs measurable cost: ViT-S latency rises from 10.0 to 20.8 ms per image, and ViT-B latency from 19.5 to 44.8 ms, at unchanged model parameter counts. Component controls isolate calibration gains of 0.80/0.24 P-AP/AUPRO points without HF and 0.90/0.30 points with HF. Structural-disruption subtypes show a larger paired per-image AP gain than other anomalies in matched categories (4.99 versus 2.68 points).

The proposal study measures 45.8% mean coverage for small defects and 28.5% for the small-and-weak subgroup, identifying the coverage dependence of global-peak selection. The calibration bound further explains the small corrections in globally flat regions. The annotated MVTec AD sweeps set the shared coefficient and ROI budget; their transfer to VisA and MVTec AD 2 uses the same settings. These results characterize selective refinement as an accuracy–cost trade-off whose behavior depends on proposal coverage and global score contrast.

Future evaluation can extend this local-regularity approach to patterned silicon wafers and complex textiles, where defects interrupt repeated layouts and woven motifs.

Funding: This research received no external funding.

Data Availability Statement: The benchmark datasets analyzed in this study are available from their respective providers [1,2,4]. The study-generated results and implementation code are available from the corresponding author upon request. Pixel-level ground-truth annotations for the private MVTec AD 2 test partitions are held by the benchmark organizers.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Bergmann, P.; Batzner, K.; Fauser, M.; Sattlegger, D.; Steger, C. The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. *Int. J. Comput. Vis.* **2021**, *129*, 1038–1059. [CrossRef]
2. Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; Dabeer, O. SPot-the-Difference Self-Supervised Pre-training for Anomaly Detection and Segmentation. In *European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2022; pp. 392–408. [CrossRef]
3. Wang, C.; Zhu, W.; Gao, B.-B.; Gan, Z.; Zhang, J.; Gu, Z.; Qian, S.; Chen, M.; Ma, L. Real-IAD: A Real-World Multi-View Dataset for Benchmarking Versatile Industrial Anomaly Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 22883–22892. [CrossRef]

4. Heckler-Kram, L.; Neudeck, J.-H.; Scheler, U.; König, R.; Steger, C. The MVTec AD 2 Dataset: Advanced Scenarios for Unsupervised Anomaly Detection. *Int. J. Comput. Vis.* **2026**, *134*, 175. [CrossRef]
5. Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; Gehler, P. Towards Total Recall in Industrial Anomaly Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14318–14328. [CrossRef]
6. Defard, T.; Setkov, A.; Loesch, A.; Audigier, R. PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization. In *Pattern Recognition. ICPR International Workshops and Challenges*; Springer: Cham, Switzerland, 2021; pp. 475–489. [CrossRef]
7. Gudovskiy, D.; Ishizaka, S.; Kozuka, K. CFLOW-AD: Real-Time Unsupervised Anomaly Detection with Localization via Conditional Normalizing Flows. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 98–107. [CrossRef]
8. Lee, S.; Lee, S.; Song, B.C. CFA: Coupled-Hypersphere-Based Feature Adaptation for Target-Oriented Anomaly Localization. *IEEE Access* **2022**, *10*, 78446–78454. [CrossRef]
9. Kim, D.; Baik, S.; Kim, T.H. SANFlow: Semantic-Aware Normalizing Flow for Anomaly Detection. In Proceedings of the Advances in Neural Information Processing Systems 36, New Orleans, LA, USA, 10–16 December 2023; Volume 36. [CrossRef]
10. Jiang, X.; Liu, J.; Wang, J.; Nie, Q.; Wu, K.; Liu, Y.; Wang, C.; Zheng, F. SoftPatch: Unsupervised Anomaly Detection with Noisy Data. In Proceedings of the Advances in Neural Information Processing Systems 35, New Orleans, LA, USA, 28 November–9 December 2022; Volume 35. [CrossRef]
11. Deng, H.; Li, X. Anomaly Detection via Reverse Distillation from One-Class Embedding. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 9737–9746. [CrossRef]
12. Guo, J.; Lu, S.; Jia, L.; Zhang, W.; Li, H. ReContrast: Domain-Specific Anomaly Detection via Contrastive Reconstruction. In Proceedings of the Advances in Neural Information Processing Systems 36, New Orleans, LA, USA, 10–16 December 2023; Volume 36. [CrossRef]
13. Tien, T.D.; Nguyen, A.T.; Tran, N.H.; Huy, T.D.; Duong, S.T.M.; Nguyen, C.D.T.; Truong, S.Q.H. Revisiting Reverse Distillation for Anomaly Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 24511–24520. [CrossRef]
14. Gu, Z.; Liu, L.; Chen, X.; Yi, R.; Zhang, J.; Wang, Y.; Wang, C.; Shu, A.; Jiang, G.; Ma, L. Remembering Normality: Memory-Guided Knowledge Distillation for Unsupervised Anomaly Detection. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 16401–16409. [CrossRef]
15. Guo, H.; Ren, L.; Fu, J.; Wang, Y.; Zhang, Z.; Lan, C.; Wang, H.; Hou, X. Template-Guided Hierarchical Feature Restoration for Anomaly Detection. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 6447–6458. [CrossRef]
16. You, Z.; Cui, L.; Shen, Y.; Yang, K.; Lu, X.; Zheng, Y.; Le, X. A Unified Model for Multi-Class Anomaly Detection. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December 2022; Volume 35. [CrossRef]
17. Liu, Z.; Zhou, Y.; Xu, Y.; Wang, Z. SimpleNet: A Simple Network for Image Anomaly Detection and Localization. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 20402–20411. [CrossRef]
18. Batzner, K.; Heckler, L.; König, R. EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2024; pp. 128–138. [CrossRef]
19. Zhang, X.; Xu, M.; Zhou, X. RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16699–16708. [CrossRef]
20. Chen, Q.; Luo, H.; Lv, C.; Zhang, Z. A Unified Anomaly Synthesis Strategy with Gradient Ascent for Industrial Anomaly Detection and Localization. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 19–36. [CrossRef]
21. He, H.; Bai, Y.; Zhang, J.; He, Q.; Chen, H.; Gan, Z.; Wang, C.; Li, X.; Tian, G.; Xie, L. MambaAD: Exploring State Space Models for Multi-Class Unsupervised Anomaly Detection. In Proceedings of the Advances in Neural Information Processing Systems 37, Vancouver, BC, Canada, 10–15 December 2024; Volume 37. [CrossRef]
22. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv* **2024**, arXiv:2304.07193. [CrossRef]
23. Darcet, T.; Oquab, M.; Mairal, J.; Bojanowski, P. Vision Transformers Need Registers. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024. [CrossRef]

24. Guo, J.; Lu, S.; Zhang, W.; Chen, F.; Li, H.; Liao, H. Dinomaly: The Less Is More Philosophy in Multi-Class Unsupervised Anomaly Detection. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 20405–20415. [CrossRef]
25. Guo, J.; Lu, S.; Fan, L.; Li, Z.; Di, D.; Song, Y.; Zhang, W.; Zhu, W.; Yan, H.; Chen, F.; et al. One Dinomaly2 Detect Them All: A Unified Framework for Full-Spectrum Unsupervised Anomaly Detection. *arXiv* **2025**, arXiv:2510.17611. [CrossRef]
26. Li, C.-L.; Sohn, K.; Yoon, J.; Pfister, T. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9664–9674. [CrossRef]
27. Zavrtanik, V.; Kristan, M.; Skočaj, D. DRAEM: A Discriminatively Trained Reconstruction Embedding for Surface Anomaly Detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 8330–8339. [CrossRef]
28. Schlüter, H.M.; Tan, J.; Hou, B.; Kainz, B. Natural Synthetic Anomalies for Self-Supervised Anomaly Detection and Localization. In *Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; pp. 474–489. [CrossRef]
29. Zhang, X.; Li, S.; Li, X.; Huang, P.; Shan, J.; Chen, T. DeSTSeg: Segmentation Guided Denoising Student-Teacher for Anomaly Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 3914–3923. [CrossRef]
30. Zhao, Y. OmniAL: A Unified CNN Framework for Unsupervised Anomaly Localization. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 3924–3933. [CrossRef]
31. He, H.; Zhang, J.; Chen, H.; Chen, X.; Li, Z.; Chen, X.; Wang, Y.; Wang, C.; Xie, L. A Diffusion-Based Framework for Multi-Class Anomaly Detection. *AAAI Conf. Artif. Intell.* **2024**, *38*, 8472–8480. [CrossRef]
32. Hu, T.; Zhang, J.; Yi, R.; Du, Y.; Chen, X.; Liu, L.; Wang, Y.; Wang, C. AnomalyDiffusion: Few-Shot Anomaly Image Generation with Diffusion Model. *AAAI Conf. Artif. Intell.* **2024**, *38*, 8526–8534. [CrossRef]
33. Jiang, L.; Dai, B.; Wu, W.; Loy, C.C. Focal Frequency Loss for Image Reconstruction and Synthesis. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 13919–13929. [CrossRef]
34. Lu, R.; Wu, Y.; Tian, L.; Wang, D.; Chen, B.; Liu, X.; Hu, R. Hierarchical Vector Quantized Transformer for Multi-Class Unsupervised Anomaly Detection. In Proceedings of the Advances in Neural Information Processing Systems 36, New Orleans, LA, USA, 10–16 December 2023; Volume 36. [CrossRef]
35. He, L.; Jiang, Z.; Peng, J.; Zhu, W.; Liu, L.; Du, Q.; Hu, X.; Chi, M.; Wang, Y.; Wang, C. Learning Unified Reference Representation for Unsupervised Multi-Class Anomaly Detection. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 216–232. [CrossRef]
36. Sträter, L.P.J.; Salehi, M.; Gavves, E.; Snoek, C.G.M.; Asano, Y.M. GeneralAD: Anomaly Detection Across Domains by Attending to Distorted Features. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 448–465. [CrossRef]
37. Tang, J.; Lu, H.; Xu, X.; Wu, R.; Hu, S.; Zhang, T.; Cheng, T.W.; Ge, M.; Chen, Y.-C.; Tsung, F. An Incremental Unified Framework for Small Defect Inspection. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 307–324. [CrossRef]
38. Li, X.; Tan, X.; Chen, Z.; Zhang, Z.; Zhang, R.; Guo, R.; Jiang, G.; Chen, Y.; Qu, Y.; Ma, L.; et al. One-for-More: Continual Diffusion Model for Anomaly Detection. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 4766–4775. [CrossRef]
39. Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; Dabeer, O. WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 19606–19616. [CrossRef]
40. Zhou, Q.; Pang, G.; Tian, Y.; He, S.; Chen, J. AnomalyCLIP: Object-Agnostic Prompt Learning for Zero-Shot Anomaly Detection. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024. [CrossRef]
41. Li, X.; Zhang, Z.; Tan, X.; Chen, C.; Qu, Y.; Xie, Y.; Ma, L. PromptAD: Learning Prompts with Only Normal Samples for Few-Shot Anomaly Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16838–16848. [CrossRef]
42. Cao, Y.; Zhang, J.; Frittoli, L.; Cheng, Y.; Shen, W.; Boracchi, G. AdaCLIP: Adapting CLIP with Hybrid Learnable Prompts for Zero-Shot Anomaly Detection. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 55–72. [CrossRef]
43. Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Tang, M.; Wang, J. AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models. *AAAI Conf. Artif. Intell.* **2024**, *38*, 1932–1940. [CrossRef]
44. Ma, W.; Zhang, X.; Yao, Q.; Tang, F.; Wu, C.; Li, Y.; Yan, R.; Jiang, Z.; Zhou, S.K. AA-CLIP: Enhancing Zero-Shot Anomaly Detection via Anomaly-Aware CLIP. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 4744–4754. [CrossRef]

45. Fang, Z.; Wang, X.; Li, H.; Liu, J.; Hu, Q.; Xiao, J. FastRecon: Few-Shot Industrial Anomaly Detection via Fast Feature Reconstruction. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 17481–17490. [CrossRef]
46. Damm, S.; Laszkiewicz, M.; Lederer, J.; Fischer, A. AnomalyDINO: Boosting Patch-Based Few-Shot Anomaly Detection with DINOv2. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 26 February–6 March 2025; pp. 1319–1329. [CrossRef]
47. Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Tang, M.; Wang, J. UniVAD: A Training-Free Unified Model for Few-Shot Visual Anomaly Detection. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 15194–15203. [CrossRef]
48. Zhang, J.; Wang, G.; Jin, Y.; Huang, D. Towards Training-Free Anomaly Detection with Vision and Language Foundation Models. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 15204–15213. [CrossRef]
49. Bergmann, P.; Batzner, K.; Fauser, M.; Sattlegger, D.; Steger, C. Beyond Dents and Scratches: Logical Constraints in Unsupervised Anomaly Detection and Localization. *Int. J. Comput. Vis.* **2022**, *130*, 947–969. [CrossRef]
50. Bergmann, P.; Jin, X.; Sattlegger, D.; Steger, C. The MVTec 3D-AD Dataset for Unsupervised 3D Anomaly Detection and Localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Volume 5 (VISAPP)*; SciTePress: Setúbal, Portugal, 2022; pp. 202–213. [CrossRef]
51. Costanzino, A.; Ramirez, P.Z.; Lisanti, G.; Di Stefano, L. Multimodal Industrial Anomaly Detection by Crossmodal Feature Mapping. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 17234–17243. [CrossRef]
52. Cao, T.; Zhu, J.; Pang, G. Anomaly Detection Under Distribution Shift. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 6511–6523. [CrossRef]
53. Carvalho, J.B.S.; Zhang, M.; Geyer, R.; Cotrini, C.; Buhmann, J.M. Invariant Anomaly Detection under Distribution Shifts: A Causal Perspective. In Proceedings of the Advances in Neural Information Processing Systems 36, New Orleans, LA, USA, 10–16 December 2023; Volume 36. [CrossRef]
54. Zhang, X.; Xu, M.; Zhou, X. Towards High-Resolution Industrial Image Anomaly Detection. *arXiv* **2025**, arXiv:2508.12931. [CrossRef]
55. Chen, Z.; You, K.; Yang, J.; Chen, L.; Li, F.; Feng, Z.; Jia, L. A Sparse-to-Dense Guided Fusion Framework for Three-Dimensional Object Detection in Railway Environments. *Eng. Appl. Artif. Intell.* **2026**, *178*, 115095. [CrossRef]
56. Sun, S.; Chen, Y.; Wu, H.; Qiao, Q.; Yang, M.; Qin, Z.; Wang, D.; Chu, X.; Xu, N.; Zhao, F. Maturity-Aware Fruit Detection in Real Orchard Scenes Using Super-Resolution Enhanced YOLOv13n. *Adv. Eng. Inform.* **2026**, *76*, 105145. [CrossRef]
57. Samele, S.; Attorre, F.; Matteucci, M. Scaling Anomaly Detection with Segmentation Models. *Digit. Signal Process.* **2026**, *168*, 105613. [CrossRef]
58. Qiao, S.; Liu, C. FreqPrompt-AD: Frequency-Guided Local Semantic Prompting for Zero-Shot Industrial Anomaly Detection and Segmentation. *J. King Saud Univ.–Comput. Inf. Sci.* **2026**, *38*, 696. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.