SYNAD: SYNERGISTIC DUAL-VIEW ADAPTATION FOR GENERALIZED ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot anomaly detection (ZSAD) aims to identify and localize structural defects without target-domain supervision. While vision–language models such as CLIP have gained traction via prompt engineering, cross-modal alignment fundamentally struggles with subtle, non-semantic physical flaws that lack explicit linguistic descriptors. This work introduces **SynAD** (**Syn**ergistic Dual-View Adaptation for Generalized **A**nomaly **D**etection), a purely visual adaptation framework built upon frozen vision foundation models (VFMs). Rather than deploying separate view-specific experts or brittle cross-scale attention, SynAD conceptualizes macroscopic context and native localized details as complementary observation distributions of a single unified detector. By injecting shared Weight-Decomposed Low-Rank Adaptation (DoRA) modules into selected self-attention projections, both global and local views are forwarded through identical adapted parameters and decoded by a shared lightweight anomaly head via independent forward passes. During inference, overlapping local predictions are seamlessly stitched at native image resolution and harmonized with global logits using an evidence-balancing coefficient calibrated strictly on the corresponding held-out source validation split. Comprehensive evaluations across 15 industrial and medical benchmarks demonstrate that SynAD achieves highly competitive anomaly detection alongside robust, fine-grained localization, delivering pronounced gains on precision-sensitive metrics on challenging benchmarks (e.g. MVTec AD 2). Furthermore, we verify architectural versatility across representative visual backbones and substantiate practical real-world utility by introducing **BoltAD**, our field-acquired industrial fastener benchmark characterized by intricate non-semantic geometric flaws.

1 INTRODUCTION

Visual anomaly detection is central to modern manufacturing and medical image analysis, where rare defects must be identified and localized with high spatial precision. Public benchmarks such as MVTec AD have established a standard evaluation protocol for unsupervised inspection (Bergmann et al., 2019). Conventional methods typically learn target-domain normality from normal reference samples using memory banks, patch-distribution modeling, self-supervised reconstruction objectives, normalizing flows, or knowledge distillation (Roth et al., 2022; Defard et al., 2021; Li et al., 2021; Zavrtanik et al., 2021; Yu et al., 2021; Deng & Li, 2022; You et al., 2022; Liu et al., 2023; Batzner et al., 2024). Such target-data reliance becomes problematic when defects are scarce, production lines change rapidly, privacy restrictions apply, or expert annotation is costly. Zero-shot anomaly detection (ZSAD) addresses this limitation by transferring anomaly representations learned from an auxiliary source to previously unseen domains, without using target-domain data for training, calibration, or model selection.

Contemporary ZSAD methods predominantly adapt large vision–language models (VLMs) such as CLIP (Radford et al., 2021) through prompt tuning or template ensembles (Jeong et al., 2023; Zhou et al., 2024; Cao et al., 2024; Zhu et al., 2025; Ma et al., 2025). This approach can be limited when defects are geometric and lack stable textual descriptors. In industrial inspection, critical defects—such as microscopic surface abrasions, subtle thread stripping, and minute fissures—are continuous geometric irregularities. Consequently, text-conditioned classifiers often rely on overly generic prompts (e.g., "damaged object"), which lack spatial geometric constraints and lead to dif-

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

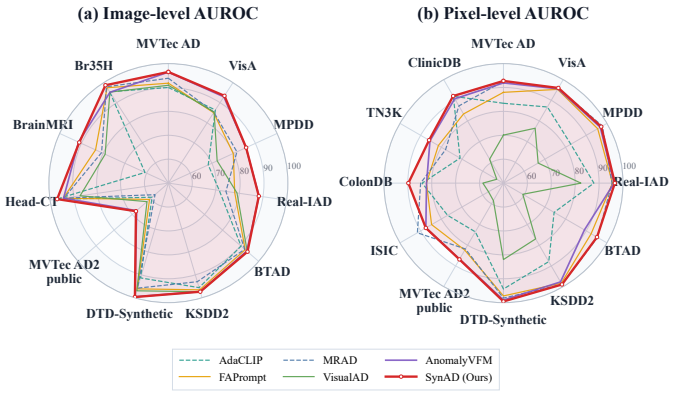


Figure 1: **Zero-shot generalization across industrial and medical benchmarks.** Image- and pixel-level AUROC are shown on representative targets under standard cross-dataset protocols. The radial scale spans 50–100%; SynAD is shown in red.

fuse localization maps. Moreover, benign visual variations such as specular reflections, component occlusions, and severe illumination shifts are easily conflated with semantic anomalies, triggering false positives in defect-free areas. This limitation motivates an adaptation strategy that learns geometry-sensitive anomaly representations without a text-conditioning bottleneck.

While vision foundation models (VFMs) offer strong visual priors for language-free inspection (Ranzinger et al., 2024), processing a single resized image introduces a field-of-view trade-off: downsampling preserves object context but weakens few-pixel defect evidence, whereas native-resolution crops retain local boundaries but omit the context needed to interpret them. To address this trade-off without separate view-specific experts or heavy cross-scale attention, we formulate source-only ZSAD as a shared dual-view adaptation problem and propose **SynAD** (**Syn**ergistic Dual-View **A**daptation for Generalized Anomaly **D**etection). SynAD processes resized global and native-resolution local views through one frozen RADIO v2.5-L VFM with shared Weight-Decomposed Low-Rank Adaptation (DoRA) (Liu et al., 2024) and a shared lightweight decoder. The views are forwarded independently; overlapping local predictions are reconstructed in native coordinates and fused with global logits using a coefficient calibrated on a held-out source split. As summarized in Figure 1, SynAD achieves competitive image-level detection and strong pixel-level localization across industrial and medical benchmarks, with notable pixel-level AUROC gains over the competing single-scale VFM method AnomalyVFM (Fučka et al., 2026) on BTAD and ColonDB.

In summary, our contributions are threefold:

- **A unified dual-view adaptation paradigm.** We formulate source-only ZSAD as a shared representation problem across observational scales. SynAD processes resized global and overlapping native-resolution views with a frozen VFM, shared DoRA adapters, and a shared lightweight decoder. The streams are forwarded independently and coupled through the shared parameterization and source-calibrated logit fusion, without view-specific experts or cross-attention.
- **A field-grounded fastener benchmark.** We introduce **BoltAD**, a benchmark for evaluating zero-shot defect localization and out-of-distribution transfer in industrial fastener imagery. It combines normal inspections with real anomaly references and controlled cases produced by a component-aware construction pipeline, with pixel-level masks for evaluation, and is reserved as an unseen target domain.
- **Cross-domain evaluation.** With one auxiliary source dataset per transfer setting, SynAD achieves leading or competitive localization across 15 public industrial and medical benchmarks and BoltAD. On MVTec AD 2, it surpasses the single-scale VFM baseline by 7.67, 7.47, and 18.42 percentage points in pixel F1-max, AP, and AUPRO@0.30, respectively. These results assess the shared visual adaptation under domain and imaging shifts rather than a target-specific calibration.

2 Related Work

Zero-shot anomaly detection (ZSAD). ZSAD seeks to recognize and localize defects in previously unobserved target domains without target-domain samples for training or calibration. The dominant route adapts CLIP (Radford et al., 2021): WinCLIP combines prompt ensembles with local windows, while AnomalyCLIP, PromptAD, AdaCLIP, FAPrompt, and AA-CLIP refine object-agnostic, fine-grained, or anomaly-aware prompting (Jeong et al., 2023; Zhou et al., 2024; Li et al., 2024; Cao et al., 2024; Zhu et al., 2025; Ma et al., 2025). Later work models prompt uncertainty, diffusion priors, online rarity, or VLM reasoning (Qu et al., 2025; Zhang et al., 2025b; He et al., 2025; Gu et al., 2024); training-free variants remove gradient updates but may still use normal references or score calibration (Zhang et al., 2025a). A visual route removes linguistic mediation: AnomalyVFM adapts a pretrained VFM with synthetic auxiliary anomalies, whereas VisualAD learns normal and abnormal tokens in a frozen ViT (Fučka et al., 2026; Hou et al., 2026). The cited visual studies use different adaptation and view protocols; they do not report the matched-budget test of one shared adapter and decoder across global and native-resolution observations used here. Encoders such as DINOv2 and AM-RADIO provide language-free representations (Oquab et al., 2024; Ranzinger et al., 2024); AnomalyDINO uses few-shot target references, whereas ResAD studies class-generalizable detection (Damm et al., 2025; Yao et al., 2024). Parameter-efficient updates adapt frozen encoders with limited trainable capacity (Hu et al., 2022; Jia et al., 2022; Liu et al., 2024; Chen et al., 2022). SynAD studies whether one shared adaptation can make global and local evidence scale-consistent under source-only transfer.

Global–local anomaly detection and industrial benchmarks. Global observations preserve object structure, whereas native-resolution windows expose weak local defects. PatchCore and PaDiM model local feature distributions (Roth et al., 2022; Defard et al., 2021), and patch or sliding-window methods retain contextual local evidence (Damm et al., 2025; Dong et al., 2023). Multiscale detectors employ parallel branches and learned cross-scale fusion (Zhou et al., 2025; Zhang et al., 2025c); their scale-specific parameters differ from SynAD's single shared adapter and decoder. Non-overlapping tiling can cut a defect at a patch boundary and produce discontinuous maps (Zhang et al., 2025c). Overlapping tiled ensembles and sliding windows improve coverage and reconstruction (Rolih et al., 2024; Dong et al., 2023), but do not by themselves impose a shared source-trained representation across views. SynAD varies the observation while sharing the VFM, adapter, and decoder, and combines the views only at the output level. MANTA provides a multi-view tiny-object benchmark (Fan et al., 2025); MVTec AD and MVTec AD 2 offer standardized industrial evaluations (Bergmann et al., 2019; Heckler-Kram et al., 2026). BoltAD complements them with a field-grounded fastener benchmark that uses component-aware anomaly construction.

3 Methodology

3.1 Overview & Motivation

A Resolution–Context Trade-Off in Frozen VFMs. For a patch-based frozen VFM, the input window determines how much native image evidence contributes to each patch token. A global resize preserves coarse object layout and long-range context, but reduces the image area occupied by a small defect; its contribution to the resulting tokens can therefore become weak. A native-coordinate crop preserves local sampling density and boundary evidence, but may omit the object-level context needed to distinguish a defect from a normal edge or illumination change. The two views thus expose complementary failure modes.

Adding another scale does not by itself resolve this trade-off when a detector is trained on one dataset and applied to unseen domains. Representative multiscale detectors use scale-specific branches or learned cross-scale fusion (Zhou et al., 2025; Zhang et al., 2025c), introducing parameters whose relationship must be learned from source data. Non-overlapping tiling may also truncate defects at crop boundaries and create discontinuous maps (Zhang et al., 2025c). We therefore test the hypothesis that one adapted detector can learn a common anomaly representation from both observations while overlap preserves spatial coverage.

SynAD implements this hypothesis: global and overlapping local views are processed by the same VFM, DoRA adapters, and decoder; local logits are reconstructed in native coordinates and fused with global logits using a coefficient selected on held-out source-validation data. The matched-

budget control in Section 4.3 tests this shared-view choice with the tile count and fusion path held fixed.

Method Overview. Figure 2 summarizes SynAD, a source-trained adaptation framework for generalized transfer without target-domain parameter updates. Each input produces a resized global observation and overlapping native-resolution local observations. Both streams are forwarded independently through one frozen VFM with shared DoRA modules in the Query, Value, and Output projections, while the Key projection remains fixed. A shared decoder predicts dense logits for both views, and an image-prediction head uses the global summary representation. During source training, global and local logits receive mask supervision and the global representation receives image-label supervision. At inference, local logits are reconstructed in native coordinates and fused with upsampled global logits using α^* selected only on held-out source-validation data. The coefficient and all model parameters are fixed before target evaluation; the two views exchange no intermediate features, and no target sample is used for training or calibration.

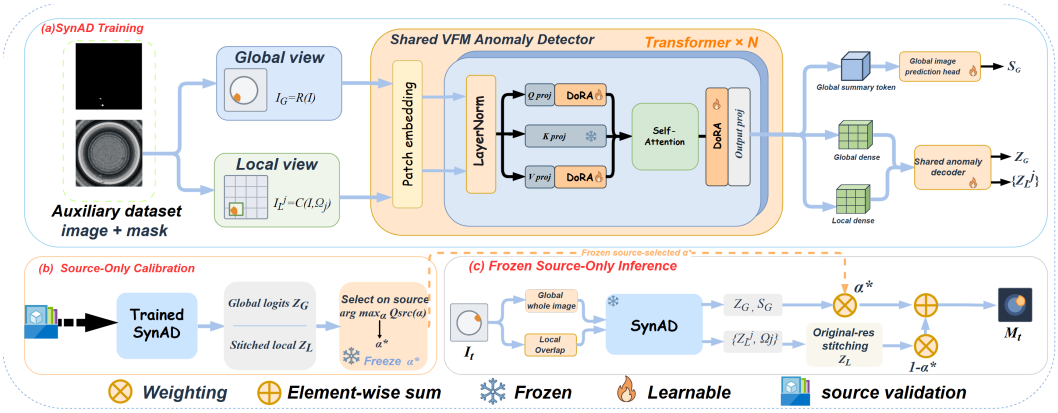


Figure 2: Overview of SynAD. Global and local observations share a frozen VFM with DoRA applied to the query, value, and output projections, together with a shared anomaly decoder and a global image-prediction head. The fusion coefficient α^* is selected on held-out source-validation data and then fixed for inference on unseen targets, where overlapping local logits are mapped to native coordinates and fused with the upsampled global logits.

3.2 Synergistic Dual-View Parameter-Efficient Adaptation

Dual-View Observation Construction. For a backbone b with patch size P_b, we use a backbone-specific square window $S_b = 48P_b$. Thus, $S_b = 768$ for RADIO v2.5-L and DINOv3-L/16, whereas $S_b = 672$ for DINOv2-L/14; all three encoders consequently produce a 48×48 patch lattice. For an image I_i of original size $H_i \times W_i$, define the padded dimensions

$$H'_{i,b} = \max(H_i, S_b), \qquad W'_{i,b} = \max(W_i, S_b). \tag{1}$$

The image is extended by border-safe reflection to $\bar{I}_{i,b} \in \mathbb{R}^{3\times H'_{i,b}\times W'_{i,b}}$, repeating the reflection operation when the required padding exceeds the source extent. Its binary mask is zero-padded to $\bar{M}_{i,b} \in \{0,1\}^{H'_{i,b}\times W'_{i,b}}$. Padding is appended on the bottom and right, so the original image occupies the top-left support $[0, H_i) \times [0, W_i)$ of the padded canvas. We also define the validity mask

$$V_{i,b}(u,v) = \mathbf{1}[0 \le u < H_i,\ 0 \le v < W_i], \tag{2}$$

which is one on the original support and zero on the padded region. The validity mask is propagated through resizing and reconstruction; padded pixels are removed before scoring. We then construct two complementary observation streams:

$$I_G = \mathcal{R}_b(\bar{I}_{i,b}) \in \mathbb{R}^{3\times S_b\times S_b}, \quad I_L^j = \mathcal{C}(\bar{I}_{i,b}, \Omega_{i,b}^j) \in \mathbb{R}^{3\times S_b\times S_b}, \quad j \in \{1,\dots,N_{L,i,b}\}, \tag{3}$$

where $\mathcal{R}_b$ denotes bilinear interpolation to $S_b \times S_b$, and $\mathcal{C}(\bar{I}_{i,b}, \Omega_{i,b}^j)$ extracts a rectangular region defined by:

$$\Omega_{i,b}^j = [y_{i,b}^j, y_{i,b}^j + S_b) \times [x_{i,b}^j, x_{i,b}^j + S_b). \tag{4}$$

To prevent defects from being truncated across boundaries, crop coordinates use strides $S_h = \mathrm{round}((1-\rho)S_b)$ and $S_w = \mathrm{round}((1-\rho)S_b)$, where ρ is the overlap ratio. At the default $\rho = 0.25$, this gives a stride of 576 for RADIO/DINOv3 and 504 for DINOv2. This overlap-aware construction is motivated by evidence that simple tiling can yield discontinuous anomaly maps, whereas overlapping tiles improve high-resolution anomaly detection (Zhang et al., 2025c; Rolih et al., 2024). The final tile along each axis is end-aligned on the padded canvas:

$$y_{i,b}^{n_y} = \min(n_y S_h, H'_{i,b} - S_b), \qquad x_{i,b}^{n_x} = \min(n_x S_w, W'_{i,b} - S_b), \tag{5}$$

where $n_y \in \{0, \ldots, N_{y,i,b}\}$, $n_x \in \{0, \ldots, N_{x,i,b}\}$,

$$N_{y,i,b} = \left\lceil \frac{H'_{i,b} - S_b}{S_h} \right\rceil, \quad N_{x,i,b} = \left\lceil \frac{W'_{i,b} - S_b}{S_w} \right\rceil, \quad N_{L,i,b} = (N_{y,i,b}+1)(N_{x,i,b}+1), \tag{6}$$

and $j = 1 + n_y(N_{x,i,b}+1) + n_x$, with $y_{i,b}^j = y_{i,b}^{n_y}$ and $x_{i,b}^j = x_{i,b}^{n_x}$. Hence every local crop has exactly $S_b \times S_b$ pixels, and the padded canvas is fully covered while the validity mask identifies its original support.

Shared Weight-Decomposed Low-Rank Adaptation. We freeze the pretrained parameter set Θ_0 of the Vision Transformer backbone Φ. To adapt representations toward anomaly boundaries while retaining general visual priors, we integrate Weight-Decomposed Low-Rank Adaptation (DoRA) (Liu et al., 2024) into self-attention projections across all Transformer layers $\mathcal{L}_{\text{adapt}} = \{1, \ldots, L\}$.

For a pretrained projection matrix $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, DoRA decomposes the updated weight matrix W into a trainable magnitude vector $m \in \mathbb{R}^{d_{\text{out}} \times 1}$ and a row-normalized direction matrix. Let $\mathrm{rowNorm}(X)_{i,:} = X_{i,:}/(\|X_{i,:}\|_2 + \epsilon)$:

$$V = W_0 + \frac{\alpha_{\text{DoRA}}}{r} BA, \qquad W = m \odot \mathrm{rowNorm}(V), \tag{7}$$

where $\odot$ represents row-wise broadcasting multiplication, $r \ll \min(d_{\text{in}}, d_{\text{out}})$ is the low-rank dimension, and α_{DoRA} is a constant scaling hyperparameter. The trainable parameters per projection are $\theta = \{m, A \in \mathbb{R}^{r \times d_{\text{in}}}, B \in \mathbb{R}^{d_{\text{out}} \times r}\}$, where m_i is initialized as $\|W_{0,i,:}\|_2$, A is drawn from a zero-mean Gaussian distribution, and B is initialized to zero.

Selective Projection Adaptation. In each self-attention block comprising Query, Key, Value, and Output projections $\{W^Q, W^K, W^V, W^O\}$, DoRA updates are applied to $\{W^Q, W^V, W^O\}$, while the Key projection remains fixed:

$$W^Q = \text{DoRA}(W_0^Q; \theta_Q), \quad W^V = \text{DoRA}(W_0^V; \theta_V), \quad W^O = \text{DoRA}(W_0^O; \theta_O), \quad W^K = W_0^K. \tag{8}$$

Keeping W_0^K fixed preserves the token association space established during pretraining, while adapting $\{W^Q, W^V, W^O\}$ permits re-weighting and extraction of defect-sensitive features.

Shared Lightweight Anomaly Decoder. Both global and local inputs are processed by the adapted backbone $\Phi_{\Theta_0,\theta}$ and decoded by a shared anomaly head $\mathcal{H}_\psi$. For any observation $X \in \{I_G, I_L^1, \ldots, I_L^{N_{L,i,b}}\}$:

$$F = \Phi_{\Theta_0,\theta}(X) \in \mathbb{R}^{C \times 48 \times 48}, \quad Z = \mathcal{H}_\psi(F) \in \mathbb{R}^{R \times R}, \qquad R = 96, \tag{9}$$

The shared 48×48 lattice makes the decoder interface invariant to the backbone-specific input window. The decoder $\mathcal{H}_\psi$ consists of 1×1 convolutional layers followed by progressive bilinear upsampling and produces a 96×96 logit map. For the global view I_G, a summary feature $T_{\text{summary}} \in \mathbb{R}^C$—obtained via a dedicated class token or global average pooling over spatial tokens—is processed by a linear layer g_ϕ to yield an image-level logit $S_G = g_\phi(T_{\text{summary}}) \in \mathbb{R}$.

Decoupled Forwarding. Global and local observations are evaluated in separate forward passes without cross-attention, token concatenation, or intermediate alignment losses. This constraint encourages the shared parameters θ and ψ to project both macro-structural context and micro-textural evidence into a unified feature space.

3.3 Native Reconstruction & Source-Calibrated Inference

Native-Coordinate Logit Stitching and Fusion. Each local tile produces a 96×96 logit map $Z_L^j = \mathcal{H}_\psi(\Phi_{\Theta_0,\theta}(I_L^j))$. We first restore it to the tile's pixel grid, $\widetilde{Z}_L^j = \mathcal{U}_{S_b,S_b}(Z_L^j)$. Stitching is performed on the padded canvas before the original support is cropped. For a padded-canvas coordinate $q = (u, v) \in \{0, \dots, H'_{i,b} - 1\} \times \{0, \dots, W'_{i,b} - 1\}$, define

$$\mathcal{J}_{i,b}(q) = \{j : q \in \Omega_{i,b}^j\}, \qquad \ell_j(q) = (u - y_{i,b}^j, v - x_{i,b}^j). \tag{10}$$

We use uniform overlap averaging with a validity gate:

$$W_{i,b}^j(q) = \mathbf{1}[q \in \Omega_{i,b}^j] V_{i,b}(q), \qquad \bar{Z}_L(q) = \frac{\sum_{j \in \mathcal{J}_{i,b}(q)} W_{i,b}^j(q)\, \widetilde{Z}_L^j(\ell_j(q))}{\sum_{j \in \mathcal{J}_{i,b}(q)} W_{i,b}^j(q) + \epsilon}, \tag{11}$$

where $\epsilon = 10^{-6}$. Thus, $W_{i,b}^j$ is a uniform window, not a Hann or Gaussian window. The final local map is $Z_L = \mathrm{Crop}_{H_i,W_i}(\bar{Z}_L)$; the denominator counts only valid tiles covering each pixel, and padded pixels contribute neither logits nor weights.

Here $\mathcal{U}_{A,B}$ denotes bilinear interpolation to an $A \times B$ grid, and Crop_{H_i,W_i} retains rows $[0, H_i)$ and columns $[0, W_i)$ of the padded canvas. In parallel, the global view produces $Z_G \in \mathbb{R}^{96\times 96}$. It is first mapped to the padded canvas and then cropped:

$$\bar{Z}_G = \mathcal{U}_{H'_{i,b},W'_{i,b}}(Z_G), \qquad Z_G^{\mathrm{orig}} = \mathrm{Crop}_{H_i,W_i}(\bar{Z}_G). \tag{12}$$

The fused spatial logit map and the resulting pixel anomaly map $\widehat{M}$ are given by:

$$Z_{\text{fused}} = \alpha^* Z_G^{\mathrm{orig}} + (1 - \alpha^*) Z_L, \quad \widehat{M} = \sigma(Z_{\text{fused}}) \in [0,1]^{H_i \times W_i}, \tag{13}$$

where $\alpha^* \in [0, 1]$ balances macroscopic layout cues with fine-grained local evidence. This is a convex interpolation in logit space; no additional probabilistic interpretation is assumed.

The overall image-level anomaly score S_t integrates the global classification confidence with peak local spatial evidence:

$$S_t = \beta\, \sigma(S_G) + (1 - \beta)\, \text{TopKMean}(\widehat{M}, k), \tag{14}$$

where $\beta = 0.5$ balances the two signals, and $\text{TopKMean}(\widehat{M}, k)$ averages the top-k values of $\widehat{M}$ with $k = \lceil \kappa H_i W_i \rceil$ ($\kappa = 0.005$) to mitigate background noise.

Held-Out Source Calibration. To prevent target-domain leakage, α^* is selected before target testing by grid search over $\mathcal{A} = \{0, 0.05, \dots, 1\}$ on the auxiliary source validation split $\mathcal{D}_{\text{src}}^{\text{val}}$. Specifically,

$$\begin{aligned} \mathcal{Q}_{\text{src}}(\alpha; \mathcal{D}_{\text{src}}^{\text{val}}) &= \tfrac{1}{2}\big[\text{F1-max}_{\text{pixel}}(\alpha; \mathcal{D}_{\text{src}}^{\text{val}}) \\ &\qquad + \text{AUPRO}_{0.30}(\alpha; \mathcal{D}_{\text{src}}^{\text{val}})\big], \\ \alpha^* &= \arg\max_{\alpha \in \mathcal{A}} \mathcal{Q}_{\text{src}}(\alpha; \mathcal{D}_{\text{src}}^{\text{val}}). \end{aligned} \tag{15}$$

where the two terms are computed on the same held-out source split. Calibration is performed separately for each source configuration and seed; the selected α^* is stored with that checkpoint and remains fixed for every target evaluated with it.

3.4 Training Objective & Source Supervision Protocol

Source Supervision Protocol. SynAD is trained on labeled source data $\mathcal{D}_{\text{src}} = \{(I_i, M_i, y_i)\}_{i=1}^{N_{\text{src}}}$, where M_i is the binary defect mask and y_i the image-level anomaly label. Target data are excluded from optimization and calibration.

For each source instance, the padded targets are constructed from $\bar{M}_{i,b}$ and $V_{i,b}$. Nearest-neighbor resizing to the decoder resolution $R \times R$ gives

$$\begin{aligned} M_{i,b}^G &= \mathcal{R}_{\text{nn}}(\bar{M}_{i,b}), & V_{i,b}^G &= \mathcal{R}_{\text{nn}}(V_{i,b}), \\ M_{i,b}^{L,j} &= \mathcal{R}_{\text{nn}}\Big(\mathcal{C}(\bar{M}_{i,b}, \Omega_{i,b}^j)\Big), & V_{i,b}^{L,j} &= \mathcal{R}_{\text{nn}}\Big(\mathcal{C}(V_{i,b}, \Omega_{i,b}^j)\Big). \end{aligned} \tag{16}$$

where all four tensors lie in $\{0,1\}^{R\times R}$. At optimization step t, a fixed deterministic sampler selects one local tile $j_i(t) \in \{1,\dots,N_{L,i,b}\}$ for each image. The joint training objective is therefore

$$\mathcal{L}_{\text{train}} = \lambda_{\text{seg}}\mathcal{L}^V_{\text{seg}}(Z_G, M^G_{i,b}; V^G_{i,b}) + \mathcal{L}_{\text{img}}(S_G, y_i) + \lambda_{\text{seg}}\mathcal{L}^V_{\text{seg}}(Z^{j_i(t)}_L, M^{L,j_i(t)}_{i,b}; V^{L,j_i(t)}_{i,b}), \tag{17}$$

where $\mathcal{L}_{\text{img}}$ denotes binary cross-entropy (BCE) loss. The masked dense objective combines Focal loss (Lin et al., 2017) with Dice loss (Milletari et al., 2016):

$$\mathcal{L}^V_{\text{seg}}(Z, M; V) = \frac{\sum_{r,c} V_{r,c}\,\ell_{\text{Focal}}(\sigma(Z_{r,c}), M_{r,c})}{\sum_{r,c} V_{r,c} + \epsilon} + 1 - \frac{2\sum_{r,c} V_{r,c}\sigma(Z_{r,c})M_{r,c} + \epsilon}{\sum_{r,c} V_{r,c}\sigma(Z_{r,c}) + \sum_{r,c} V_{r,c}M_{r,c} + \epsilon}. \tag{18}$$

Thus, padded pixels are excluded from both segmentation terms. During inference, all $N_{L,i,b}$ valid local tiles are enumerated and stitched using the same validity rule. The hyperparameter λ_{seg} balances localization and classification losses; we use $\lambda_{\text{seg}} = 0.5$ in all reported runs. Only the shared decoder ψ, image head ϕ, and DoRA adapters $\{\theta_Q, \theta_V, \theta_O\}$ are trainable; the base VFM Θ_0 remains frozen.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Protocols. We evaluate zero-shot anomaly detection on 15 public industrial and medical benchmarks (Bergmann et al., 2019; Zou et al., 2022; Jezek et al., 2021; Wang et al., 2024; Mishra et al., 2021; Božič et al., 2021; Aota et al., 2023; Heckler-Kram et al., 2026; Salehi et al., 2021; Kanade & Gumaste, 2015; Hamada, 2020; Codella et al., 2018; Tajbakhsh et al., 2016; Gong et al., 2021; Bernal et al., 2015), together with BoltAD as an unseen target. BoltAD is reserved for testing and contains 400 normal and 96 anomalous images; normal masks are all zero. Each run uses one auxiliary source dataset (VisA by default, or MVTec AD when VisA is the target). The fusion coefficient α^* is selected on the corresponding held-out source split and fixed for target evaluation; target images are not used for fitting, calibration, or model selection. Release information and the complete BoltAD report are provided in Appendices A.1–A.9 and C.3.

Evaluation metrics. Table 1 reports image-level AUROC/AP and pixel-level AUROC/AUPRO@0.30; AUPRO@0.30 is the area under the per-region-overlap (PRO) curve up to a false-positive-rate limit of 0.30. The earlier RADIO full-metric comparison is given in Tables C.1–C.3. Table 2 reports image AUROC and pixel AUROC/F1-max/AUPRO@0.30, with the remaining BoltAD metrics in Appendix C.3. Head-CT, BrainMRI, and Br35H are image-level benchmarks, whereas ISIC, ColonDB, TN3K, and ClinicDB are pixel-level benchmarks. Backbone interfaces, view construction, optimization, and runtime settings are specified in Appendix B.

4.2 Comparison with State-of-the-Art Methods

Quantitative comparison. Table 1 summarizes image-level AUROC/AP and pixel-level AUROC/AUPRO@0.30 across 15 public benchmarks. On MVTec AD and VisA, SynAD obtains AUPRO@0.30 scores of 89.14% and 90.88%, with pixel F1-max scores of 50.20% and 35.81% in the full-metric comparison (Tables C.1–C.3). On MVTec AD 2, it reaches 86.83% pixel AUROC and 68.67% AUPRO@0.30, compared with 86.72% and 50.25% for AnomalyVFM; its pixel F1-max/AP are higher by 7.67/7.47 points. On medical targets, SynAD leads ColonDB in pixel AUROC/AUPRO@0.30 (89.77%/83.17%) and exceeds 90% image AUROC on Head-CT, BrainMRI, and Br35H.

Evaluation on BoltAD. Table 2 summarizes the field-grounded fastener benchmark.

Table 1: **Zero-shot anomaly detection results across industrial and medical benchmarks.** For each dataset, we report image-level AUROC and AP and pixel-level AUROC and AUPRO@0.30. Best and second-best results are highlighted in red and blue, respectively. Full metrics for the earlier RADIO comparison are reported in Tables C.1–C.3.

Domain	Metric	Dataset	WinCLIP CVPR'23	AnomalyCLIP ICLR'24	AdaCLIP ECCV'24	FAPrompt ICCV'25	MRAD ICLR'26	VisualAD CVPR'26	AnomalyVFM CVPR'26	SynAD Ours
Industrial	Image-level (AUROC, AP)	MVTec AD	(70.06, 84.57)	(93.89, 97.05)	(89.99, 95.67)	(91.80, 95.62)	(93.87, 97.30)	(90.92, 96.30)	(96.42, 97.88)	(**96.50**, **98.00**)
		VisA	(64.29, 70.20)	(88.76, 90.14)	(86.37, 88.51)	(85.01, 86.71)	(85.21, 87.97)	(85.37, 87.33)	(92.91, 94.63)	(**93.40**, **95.26**)
		MPDD	(43.86, 59.40)	(73.66, 76.52)	(68.33, 73.87)	(79.83, 82.95)	(81.52, 83.01)	(72.43, 77.16)	(85.70, 89.83)	(**85.75**, **90.34**)
		Real-IAD	(60.12, 43.63)	(70.01, 55.40)	(71.52, 58.07)	(78.04, 66.74)	(74.32, 62.49)	(79.00, 68.80)	(**88.31**, 86.59)	(88.22, **86.86**)
		BTAD	(57.56, 63.62)	(89.12, 91.05)	(90.24, 93.24)	(92.52, 93.70)	(92.24, 94.43)	(93.63, 92.28)	(93.41, **95.80**)	(**93.79**, 95.65)
		KSDD2	(70.85, 36.16)	(90.81, 76.46)	(95.56, 89.31)	(96.51, 89.18)	(92.97, 86.35)	(**97.34**, 87.36)	(97.30, 88.93)	(**97.34**, **89.56**)
		DTD-Synthetic	(64.00, 82.97)	(94.46, 97.70)	(91.32, 95.41)	(96.10, 98.34)	(95.79, 98.32)	(96.94, 99.02)	(99.55, 99.79)	(**99.66**, **99.85**)
		MVTec AD2 public	(54.11, 71.30)	(58.07, 73.83)	(59.02, 72.89)	(60.42, 75.69)	(57.38, 72.37)	(61.82, 75.59)	(67.58, 80.52)	(**67.94**, **81.13**)
	Pixel-level (AUROC, AUPRO@0.30)	MVTec AD	(80.61, 61.66)	(90.92, 84.95)	(83.39, 78.15)	(87.97, 83.82)	(92.60, 87.73)	(70.09, 66.21)	(91.91, 88.59)	(**92.74**, **89.14**)
		VisA	(70.04, 48.80)	(95.07, 86.58)	(86.77, 81.79)	(95.27, 86.88)	(95.79, 87.98)	(76.51, 63.69)	(95.55, 89.60)	(**95.95**, **90.88**)
		MPDD	(66.47, 39.14)	(92.88, 81.89)	(82.53, 69.59)	(95.51, 87.14)	(**97.78**, 91.33)	(66.76, 40.64)	(96.75, 90.57)	(97.22, **91.83**)
		Real-IAD	(84.95, 57.31)	(86.27, 73.02)	(87.83, 71.58)	(95.60, 87.00)	(95.41, 83.38)	(82.40, 67.41)	(96.26, 86.06)	(**96.59**, **87.70**)
		ʼTAD	(64.34, 23.72)	(80.97, 57.30)	(74.47, 48.76)	(92.34, 68.77)	(**95.42**, 74.67)	(59.34, 31.64)	(89.12, 60.50)	(95.18, **78.82**)
		KSDD2	(78.52, 61.91)	(98.04, 95.87)	(87.90, 87.23)	(98.33, **96.57**)	(98.93, 96.10)	(76.93, 79.55)	(97.72, 95.18)	(**98.95**, 94.77)
		DTD-Synthetic	(79.23, 53.91)	(95.93, 89.56)	(94.29, 89.36)	(97.18, 91.83)	(98.10, 93.24)	(81.96, 75.24)	(98.91, 94.88)	(**99.48**, **97.03**)
		MVTec AD2 public	(66.97, 30.01)	(72.06, 41.66)	(73.53, 47.48)	(82.36, 49.56)	(81.80, 48.47)	(58.42, 31.80)	(86.72, 50.25)	(**86.83**, **68.67**)
Medical	Image-level (AUROC, AP)	Head-CT	(65.32, 60.48)	(95.24, 95.73)	(**97.36**, 97.55)	(94.39, 91.18)	(94.15, 93.10)	(87.31, 83.53)	(94.29, 93.14)	(97.00, **97.58**)
		BrainMRI	(47.98, 29.66)	(75.12, 51.23)	(60.75, 45.67)	(83.48, 69.60)	(80.71, 66.04)	(79.06, 63.60)	(**91.30**, **85.73**)	(90.97, 85.59)
		Br35H	(68.29, 65.55)	(97.37, 97.18)	(95.43, 93.47)	(97.41, 96.75)	(97.89, 97.69)	(95.25, 91.57)	(95.16, 93.12)	(**98.66**, **98.66**)
	Pixel-level (AUROC, AUPRO@0.30)	ISIC	(82.78, 56.42)	(87.11, 77.18)	(76.95, 68.29)	(84.70, 72.86)	(**91.56**, **83.66**)	(57.22, 41.84)	(87.48, 77.19)	(87.49, 79.56)
		ColonDB	(57.46, 21.23)	(81.87, 72.41)	(84.01, 77.00)	(82.21, 72.33)	(84.64, 73.51)	(58.61, 43.49)	(82.17, 77.04)	(**89.77**, **83.17**)
		ʼN3K	(71.28, 43.28)	(77.50, 49.26)	(70.95, 42.10)	(81.44, 50.07)	(77.96, 51.53)	(53.23, 20.69)	(85.48, 46.30)	(**85.92**, **52.61**)
		ClinicDB	(64.44, 21.97)	(84.24, 70.49)	(91.95, **88.15**)	(83.40, 68.73)	(87.28, 73.87)	(61.57, 45.74)	(90.75, 82.21)	(**92.06**, 82.03)

Table 2: **Zero-shot anomaly detection on BoltAD (%).** The main-text summary reports image AUROC and pixel AUROC, F1-max, and AUPRO@0.30. Best and second-best results are shown in red and blue, respectively. SynAD reports mean ± standard deviation over three independent runs; complete metrics are given in Appendix C.3.

Method	Image AUROC	Pixel AUROC	Pixel F1-max	AUPRO@0.30
WinCLIP (Jeong et al., 2023)	64.52	73.92	9.55	45.85
AdaCLIP (Cao et al., 2024)	60.26	81.52	25.67	69.38
AnomalyCLIP (Zhou et al., 2024)	**66.89**	90.10	23.66	68.88
FAPrompt (Zhu et al., 2025)	65.14	90.32	23.98	64.40
VisualAD (Hou et al., 2026)	61.27	90.65	25.58	75.25
MRAD (Xu et al., 2026)	64.12	**91.88**	24.38	69.92
AnomalyVFM (Fučka et al., 2026)	60.15	**91.53**	**27.83**	**79.75**
SynAD (Ours)	**65.24 ± 1.06**	91.06 ± 0.21	**29.54 ± 0.29**	**80.69 ± 0.83**

SynAD reaches 65.24±1.06% image AUROC, 91.06±0.21% pixel AUROC, 29.54±0.29% pixel F1-max, and 80.69±0.83% AUPRO@0.30. MRAD and AnomalyVFM obtain higher pixel AUROC (91.88% and 91.53% versus 91.06%), while AnomalyCLIP obtains the highest image AUROC. Because pixel AUROC aggregates rankings over all pixels, we also consider metrics that emphasize precision and region overlap: SynAD exceeds AnomalyVFM by 1.71 percentage points in pixel F1-max and 0.94 percentage points in AUPRO@0.30. Complete metrics, release structure, and evaluation protocol are provided in Appendices A.1–A.9 and C.3.

Qualitative comparison. Figure 3 shows representative industrial and medical cases. Prompt-based methods and VisualAD sometimes activate on normal texture or reflected contours, whereas SynAD responses are more concentrated around annotated defects. Additional backbone examples and inference profiles are provided in Appendices C.4 and C.5.

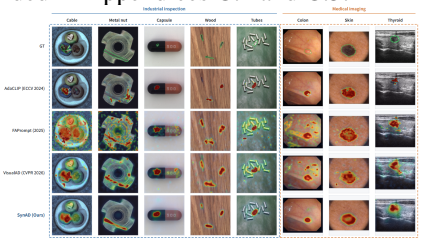


Figure 3: **Qualitative examples.** Ground-truth overlays and anomaly maps on representative industrial and medical images.

4.3 Ablation Study

The ablations test the shared dual-view hypothesis. Table 3 is the primary matched-budget control; Table 4 separately diagnoses view formulation and projection policy. We compute each dataset macro within a run and report its average over three independent runs.

Matched-budget view supervision. To separate local-view supervision from extra test-time tiles, we run the Global-only checkpoint through the complete O25 (25% overlap) tile list, native-coordinate stitching, and the same fusion code as SynAD. This independent run keeps the backbone, decoder, source split, optimizer, and test budget fixed; only Global-only omits local-view segmentation. Each row is calibrated independently on its source validation split ($\alpha^* = 0.35$ for Global-only and 0.30 for SynAD). Table 3 shows higher values for the jointly trained model on every metric, including gains of 0.0235 in pixel AUROC, 0.0155 in pixel F1-max, and 0.0198 in AUPRO@0.30. This supports a contribution from local-view supervision beyond the extra tiles.

Table 3: **Matched-budget view-supervision control.** Means over three independent runs on 11 image- and 12 pixel-level benchmarks. Both rows use the same O25 tile path; source-only coefficients are $\alpha^* = 0.35$ for Global-only and 0.30 for SynAD.

Configuration	I-AUROC	I-AP	P-AUROC	P-F1	P-AP	AUPRO@0.30	AUPRO@0.05
Global-only checkpoint + O25	0.9018	0.8944	0.9025	0.4621	0.4474	0.7977	0.5599
SynAD (matched-budget)	**0.9114**	**0.9156**	**0.9260**	**0.4775**	**0.4583**	**0.8175**	**0.5806**

View-formulation comparison. The upper block of Table 4 is an independent diagnostic: Global-only uses the resized view, Local-only samples one tile per update and enumerates all O25 tiles at test time, and Fused SynAD trains with both streams and fuses their outputs. Because both training and test views change, this block measures view complementarity rather than the matched-budget attribution. Fusion improves all five pixel metrics; its 0.9037 pixel-AUROC value is a separate view-formulation result, not a duplicate of Table 3 or the primary QVO row.

Table 4: **View and projection ablations.** Unweighted macro averages over 12 pixel-localization benchmarks. The upper block changes the view protocol; the lower block changes only the attention projection policy. The primary DoRA-QVO row matches the pixel macro in Table 1 and is distinct from the matched-budget row in Table 3.

Ablation Setting	Pixel AUROC	Pixel F1-max	Pixel AP	AUPRO@0.30	AUPRO@0.05
Dual-View Formulation					
Global-only (global-view protocol)	0.8650	0.3450	0.2962	0.6783	0.3783
Local-only (local-view protocol)	0.8587	0.3299	0.2898	0.6680	0.3718
Fused SynAD (view-formulation run)	**0.9037**	**0.3906**	**0.3567**	**0.7573**	**0.4736**
Selective Attention Projections (12-Pixel-Benchmark Macro)					
DoRA-QV (Key/Out frozen)	0.9204	0.4579	0.4356	0.8023	0.5509
DoRA-QKVO (all adapted)	0.9214	0.4621	0.4407	0.8083	0.5607
LoRA-QVO (control)	0.9221	0.4648	0.4435	0.8092	0.5624
Primary SynAD (DoRA-QVO, Key frozen)	**0.9318**	**0.4817**	**0.4642**	**0.8302**	**0.5988**

The two blocks answer complementary questions. Table 3 holds the complete O25 test path fixed and therefore isolates the effect of local-view supervision. The upper block of Table 4 instead changes the view protocol during both training and testing, so its Fused SynAD row measures the benefit of combining the two observations under that independent formulation. The lower block keeps the primary view path fixed and changes only the attention projections: adapting Q, V, and O while leaving K frozen gives the best result, whereas adapting K as well or replacing DoRA with the LoRA control is less effective. Thus the matched-budget and projection studies support two distinct parts of the design: local supervision adds information beyond the extra tiles, and selective parameter-efficient adaptation is a better fit for the shared dual-view representation.

Cross-backbone evaluation. The same dual-view interface transfers across RADIO, DINOv2, and DINOv3: each encoder is mapped to the common 48×48 token lattice and uses the same decoder, fusion rule, and source-only protocol. Across the three encoders, the macro spread is at most 0.77 points for image AUROC, 0.63 for pixel AUROC, 0.43 for pixel AP, and 0.71 for AUPRO@0.30. RADIO obtains the highest image AUROC, whereas DINOv2 obtains the highest pixel AP and AUPRO@0.30. The narrow spread supports portability of the shared dual-view adaptation beyond the primary encoder; complete dataset-wise results and qualitative comparisons are provided in Appendix C.4.

5 CONCLUSION

SynAD combines global and native-resolution views through shared DoRA adaptation. Across the evaluated benchmarks, it provides competitive zero-shot detection and localization, with ablations supporting dual-view supervision and selective QVO adaptation.

AI USE STATEMENT

Generative AI was used only for localized candidate editing in BoltAD and for language assistance during drafting. The authors manually reviewed all outputs; AI was not used for model training, checkpoint selection, target calibration, or reported metrics, and the authors take responsibility for the final content.

REFERENCES

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

Kilian Batzner, Lars Heckler, and Rebecca König. EfficientAD: Accurate visual anomaly detection at millisecond-level latencies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 128–138, 2024. doi: 10.1109/WACV57701.2024.00020.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9584–9592, 2019. doi: 10.1109/CVPR.2019.00982.

Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation versus saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015. doi: 10.1016/j.compmedimag.2015.02.007.

Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459, 2021. doi: 10.1016/j.compind.2021.103459.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. AdaCLIP: Adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pp. 55–72, 2024. doi: 10.1007/978-3-031-72761-0_4.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16664–16678, 2022. doi: 10.52202/068431-1212.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging, hosted by the international skin imaging collaboration. In *2018 IEEE 15th International Symposium on Biomedical Imaging*, pp. 168–172, 2018. doi: 10.1109/ISBI.2018.8363547.

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. AnomalyDINO: Boosting patch-based few-shot anomaly detection with DINOv2. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1319–1329, 2025. doi: 10.1109/WACV61041.2025.00136.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition: ICPR International Workshops and Challenges*, pp. 475–489, 2021. doi: 10.1007/978-3-030-68799-1_35.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9737–9746, 2022. doi: 10.1109/CVPR52688.2022.00951.

Haoyu Dong, Yifan Zhang, Hanxue Gu, Nicholas Konz, Yixin Zhang, and Maciej A. Mazurowski. SWSSL: Sliding window-based self-supervised learning for anomaly detection in high-resolution images. *IEEE Transactions on Medical Imaging*, 42(12):3860–3870, 2023. doi: 10.1109/TMI.2023.3314318.

Lei Fan, Dongdong Fan, Zhiguang Hu, Yiwen Ding, Donglin Di, Kai Yi, Maurice Pagnucco, and Yang Song. MANTA: A large-scale multi-view and visual-text anomaly detection dataset for tiny objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25518–25527, 2025. doi: 10.1109/CVPR52734.2025.02376.

Matic Fučka, Vitjan Zavrtanik, and Danijel Skočaj. AnomalyVFM—transforming vision foundation models into zero-shot anomaly detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35555–35566, June 2026.

Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th International Symposium on Biomedical Imaging*, pp. 257–261, 2021. doi: 10.1109/ISBI48211.2021.9434087.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. AnomalyGPT: Detecting industrial anomalies using large vision-language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):1932–1940, 2024. doi: 10.1609/aaai.v38i3.27963.

Ahmed Hamada. Br35H: Brain tumor detection. Kaggle dataset, 2020. URL `https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection`.

Jianfang He, Min Cao, Silong Peng, and Qiong Xie. RareCLIP: Rarity-aware online zero-shot industrial anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24478–24487, 2025. doi: 10.1109/ICCV51701.2025.02269.

Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. The MVTec AD 2 dataset: Advanced scenarios for unsupervised anomaly detection. *International Journal of Computer Vision*, 134(4):175, 2026. doi: 10.1007/s11263-026-02743-0.

Yanning Hou, Peiyuan Li, Zirui Liu, Yitong Wang, Yanran Ruan, Jianfeng Qiu, and Ke Xu. VisualAD: Language-free zero-shot anomaly detection via vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21346–21356, June 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19606–19616, 2023. doi: 10.1109/CVPR52729.2023.01878.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: Evaluating current methods in complex conditions. In *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*, pp. 66–71, 2021. doi: 10.1109/ICUMT54235.2021.9631567.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision—ECCV 2022*, pp. 709–727, 2022. doi: 10.1007/978-3-031-19827-4_41.

Pranita Balaji Kanade and P. P. Gumaste. Brain tumor detection using MRI images. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, 3(2):146–150, 2015.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. CutPaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9664–9674, 2021. doi: 10.1109/CVPR46437.2021.00954.

Yiting Li, Adam Goodge, Fayao Liu, and Chuan-Sheng Foo. PromptAD: Zero-shot anomaly detection using text prompts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1093–1102, 2024. doi: 10.1109/WACV57701.2024.00113.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2999–3007, 2017. doi: 10.1109/ICCV.2017.324.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121, 2024.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20402–20411, 2023. doi: 10.1109/CVPR52729.2023.01954.

Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S. Kevin Zhou. AA-CLIP: Enhancing zero-shot anomaly detection via anomaly-aware CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4744–4754, 2025. doi: 10.1109/CVPR52734.2025.00447.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision*, pp. 565–571, 2016. doi: 10.1109/3DV.2016.79.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: A vision transformer network for image anomaly detection and localization. In *30th IEEE/IES International Symposium on Industrial Electronics*, 2021. doi: 10.1109/ISIE45552.2021.9576231.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30398–30408, 2025. doi: 10.1109/CVPR52734.2025.02830.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12490–12500, 2024. doi: 10.1109/CVPR52733.2024.01187.

Blaž Rolih, Dick Ameln, Ashwin Vaidya, and Samet Akcay. Divide and conquer: High-resolution industrial anomaly detection via memory efficient tiled ensemble. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3866–3875, 2024. doi: 10.1109/CVPRW63382.2024.00391.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022. doi: 10.1109/CVPR52688.2022.01392.

Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14902–14912, 2021. URL https://arxiv.org/abs/2011.11108.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, 35(2):630–644, 2016. doi: 10.1109/TMI.2015.2487997.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22883–22892, 2024.

Chaoran Xu, Chengkan Lv, Qiyu Chen, Feng Zhang, and Zhengtao Zhang. MRAD: Zero-shot anomaly detection with memory-driven retrieval. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xincheng Yao, Zixin Chen, Chao Gao, Guangtao Zhai, and Chongyang Zhang. ResAD: A simple framework for class generalizable anomaly detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125287–125311, 2024. doi: 10.52202/079017-3979.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4571–4584, 2022. doi: 10.52202/068431-0330.

Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. FastFlow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021. doi: 10.48550/arXiv.2111.07677.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRÆM: A discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2021. doi: 10.1109/ICCV48922.2021.00822.

Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15204–15213, 2025a. doi: 10.1109/CVPR52734.2025.01416.

Tianrui Zhang, Liang Gao, Xinyu Li, and Yiping Gao. DZAD: Diffusion-based zero-shot anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10):10131–10138, 2025b. doi: 10.1609/aaai.v39i10.33099.

Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Towards high-resolution industrial image anomaly detection. *arXiv preprint arXiv:2508.12931*, 2025c. doi: 10.48550/arXiv.2508.12931.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2024.

Yixuan Zhou, Xing Xu, Jingkuan Song, Fumin Shen, and Heng Tao Shen. MSFlow: Multiscale flow-based framework for unsupervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):2437–2450, 2025. doi: 10.1109/TNNLS.2023.3344118.

Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22241–22251, 2025. doi: 10.1109/ICCV51701.2025.02065.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. SPot-the-Difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pp. 392–408, 2022. doi: 10.1007/978-3-031-20056-4_23.

APPENDIX

A THE BOLTAD INDUSTRIAL BENCHMARK: CONSTRUCTION, RELEASE, AND EVALUATION

A.1 BOLTAD OVERVIEW

BoltAD is a field-grounded benchmark for zero-shot defect localization and out-of-distribution evaluation on industrial fastening components. The expanded canonical release contains 496 test images: 400 normal inspections under `test/good` and 96 anomaly cases under `test/anomaly`. The `train` directory is empty. Each anomaly image is paired with a same-stem LabelMe JSON record and an annotated pixel-level mask under `ground_truth/anomaly`; each normal image has a corresponding all-zero mask under `ground_truth/good`. The 96 anomaly cases comprise 18 real anomaly references and 78 controlled cases spanning four categories: *Bolt Loosening*, *Bolt Missing*, *Bolt Damage*, and *Bolt Corrosion*.

The benchmark is used only as an unseen target domain. No BoltAD image is used for training, adapter fitting, checkpoint selection, threshold selection, or score calibration. The normal images make image-level transfer measurable, while the annotated anomaly cases preserve the fine-grained localization task. BoltAD is designed around real inspection imagery, real anomaly references, component-aware localized construction, and pixel-level ground truth rather than as an unconstrained synthetic collection.

A.2 RELEASE STRUCTURE AND LABELS

The released directory keeps `train/` empty. The 400 normal images are under `test/good/`; the 96 anomaly images and their LabelMe records are under `test/anomaly/`; and the corresponding masks are under `ground_truth/good/` and `ground_truth/anomaly/`, respectively. The accompanying `metadata.csv` contains one row for each of the 496 canonical identifiers and records the split, image-level label, image path, mask path, and original image dimensions. Identifiers `BOLTAD_000001`–`BOLTAD_000096` denote anomalous cases, and `BOLTAD_000097`–`BOLTAD_000496` denote normal cases.

Table A.1: **BoltAD release structure.** Counts describe the expanded test-only release and its role in evaluation.

Path or field	Count	Role
`train/`	0	No target-domain training data
`test/good/`	400	Image-level normal cases
`test/anomaly/`	96	Anomaly cases with annotations
`ground_truth/good/`	400	All-zero masks for normal cases
`ground_truth/anomaly/`	96	Pixel masks for anomaly cases
`metadata.csv`	496 rows	Canonical IDs, labels, and paths
`protocol.md`	1	Unseen-target evaluation rules

A.3 COLLECTION AND DE-IDENTIFICATION

The images are local views of fastening components acquired in real industrial inspection settings. Before release, raw filenames, equipment-specific information, positional codes, business identifiers, and other site-related metadata were removed, and each retained record was assigned an anonymous canonical identifier. We describe the source only at the level needed to establish the inspection setting; the originating industry, organization, equipment, and location are not disclosed.

A.4 REAL ANOMALY REFERENCES

The benchmark contains 18 real anomaly references. These cases primarily cover *Bolt Loosening (Alignment-Line Misalignment)* and *Bolt Missing*. Bolt loosening is identified by a loss of continuity between alignment marks on a fastening component and its neighboring structure after a small rotation or displacement. Bolt missing denotes the absence of a bolt, nut, or other principal fastener from a location where it is structurally expected.

Based on practical feedback from frontline maintenance personnel, these two conditions account for approximately 99% of the fastener anomalies encountered in routine inspection. We treat this figure

as an operational observation rather than a population-level estimate. The real anomaly cases serve both as evaluation examples and as visual and structural references for the controlled construction process.

A.5 Anomaly Taxonomy

BoltAD uses four anomaly categories: *Bolt Loosening*, *Bolt Missing*, *Bolt Damage*, and *Bolt Corrosion*. The first two categories are grounded in the real anomaly references. Damage and corrosion broaden the visual and spatial forms represented by the benchmark, including local material loss, thread or edge damage, wear, oxidation, pitting, and roughened metallic surfaces. They are principally introduced through controlled construction and should not be interpreted as having the same number of real cases as loosening and missing.

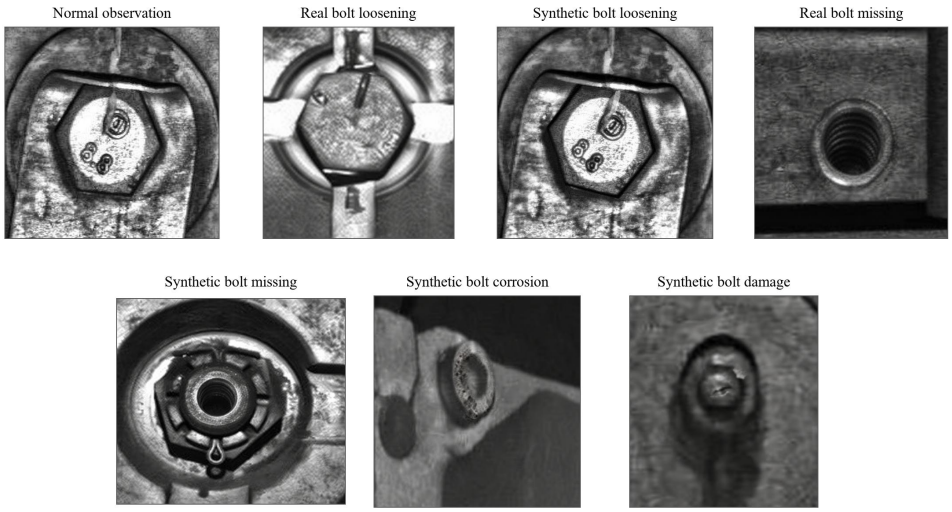


Figure A.1: **Illustrative BoltAD cases.** The gallery contains a normal inspection image, real and controlled examples of bolt loosening and bolt missing, and controlled examples of bolt corrosion and bolt damage. These examples illustrate the visual range of the benchmark and do not encode category proportions.

A.6 Component-Aware Anomaly Construction

Controlled cases begin with a real normal industrial image I_n, rather than an image generated from text. This preserves the original background, illumination, viewing geometry, imaging noise, and component context. For each candidate, a complete component region $M_{\text{component}}$ is first localized with an interactive segmentation model. The real anomaly references are converted into a structured anomaly plan specifying the component, anomaly type, location, visible cue, local structural change, expected scale, regions to preserve, and prohibited artifacts. The plan is then translated into a localized editing operation.

The three masks used by the process have distinct meanings. $M_{\text{component}}$ denotes the complete target fastener, M_{edit} denotes the region in which an edit is permitted, and M_{GT} denotes the final visible anomalous pixels. They are not interchangeable. For example, for a local corrosion case, $M_{\text{GT}} \subset M_{\text{edit}} \subset M_{\text{component}}$; for missing and loosening, the final anomalous area is determined by the visible structural change rather than by labeling the entire component.

Given a normal image, an edit region, and a structured plan P, the candidate image is written as

$$I_a = G(I_n, M_{\text{edit}}, P). \tag{19}$$

The edit is constrained to retain the original scene outside the target region. For loosening, the plan requires a visible alignment-line misalignment; for missing, it requires a structurally plausible

absence with a locally restored background; for damage and corrosion, it limits the modification to physically relevant component surfaces.

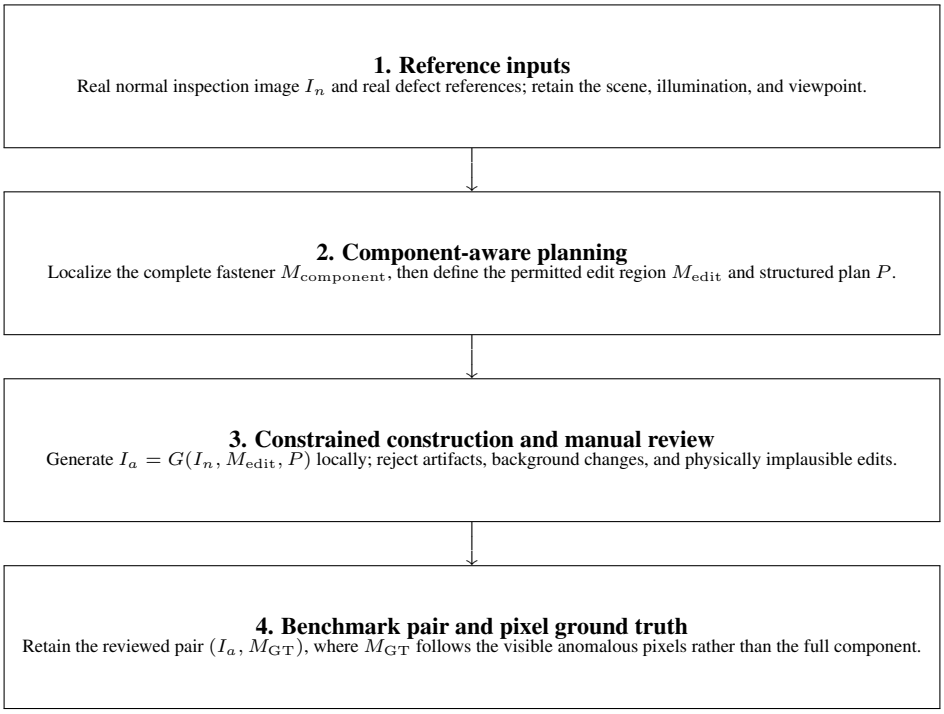


Figure A.2: **Component-aware anomaly construction.** The process combines real inspection context and defect references with localized editing, manual review, and a pixel-level mask that follows the visible structural change.

A.7 Annotation Records and Rasterization

For every anomalous identifier, the image, JSON record, and mask share the same six-digit stem. The JSON files use the LabelMe format, and the released masks have the same spatial dimensions as their corresponding images. The JSON files preserve annotation geometry, whereas the raster masks are the targets used for pixel-level scoring.

A.8 Quality Control and Pixel Annotation

Candidate constructions are not included automatically. Each candidate is screened for unintended background changes, black-hole or dark-spot artifacts, excessive structural deformation, implausible physical appearance, mismatch with the target component, unrealistic anomaly size, near-invisible edits, and additional unspecified anomalies. All controlled cases are manually reviewed before inclusion in BoltAD. This review retains candidates whose changes are localized, visually consistent with the underlying inspection image, and aligned with the planned anomaly type.

For real anomalies, annotators delineate the visible anomalous region directly. For controlled cases, the editing region provides an initial reference, after which the mask is manually checked and revised to follow the visible change. Each anomalous case retains the same canonical identifier across its image, LabelMe record, and mask. Mask values are encoded as $M_i(x, y) \in \{0, 255\}$, where 0 denotes normal pixels and 255 denotes anomalous pixels. Loosening masks cover the misaligned alignment line and associated local structural change; missing masks cover the region expected to contain the absent fastener; damage masks cover material loss, cracks, damaged threads, or broken edges; and corrosion masks cover only the visible oxidized, pitted, or degraded surface. An entire component is not labeled anomalous merely because a local defect is present.

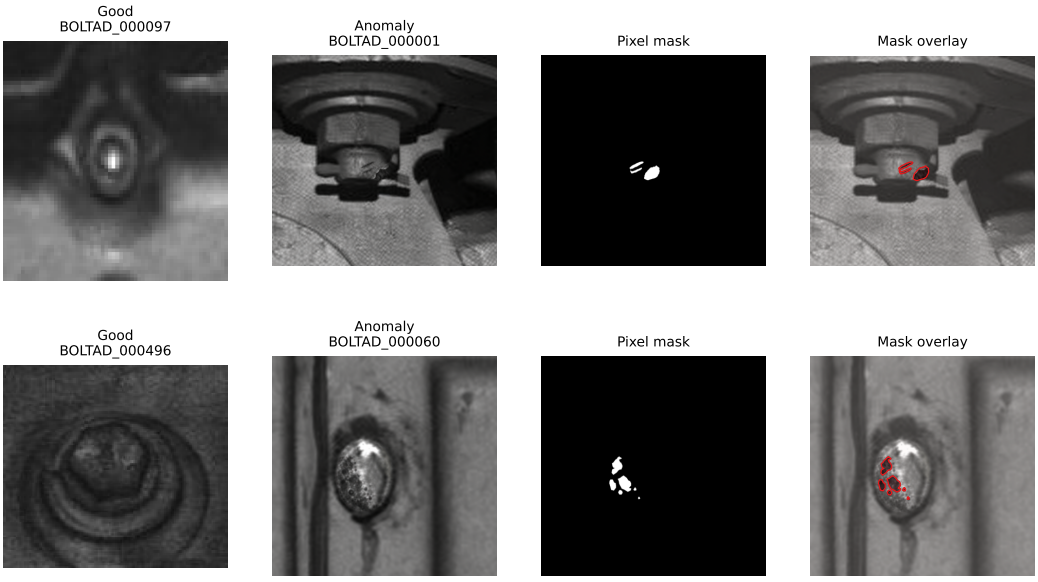


Figure A.3: **Examples from the expanded BoltAD release.** Each row pairs a normal inspection with an anomalous inspection, its pixel mask, and an overlay of the annotated region. The released protocol assigns all-zero masks to normal cases and includes all 496 masks in pixel-level evaluation; only representative anomalous masks are visualized here.

A.9 Source-only Protocol and Metrics

BoltAD is evaluated as an unseen target domain with $\mathcal{D}_{\text{train}} = \emptyset$. Image-level scores are computed on all 496 test images using the image labels in `metadata.csv`. Pixel-level predictions are evaluated on all 496 test images using the corresponding released masks; the 400 normal masks are all zero and therefore contribute normal-pixel false positives. No target image is used for model fitting, adapter learning, checkpoint selection, threshold tuning, or score calibration.

A.10 Evaluation Metrics

The complete BoltAD report contains image-level AUROC, F1-max, and AP, together with pixel-level AUROC, F1-max, AP, and AUPRO@0.30. Here, PRO denotes the per-region-overlap curve, and AUPRO@0.30 denotes the area under this curve up to a false-positive-rate limit of 0.30. Image-level metrics use one scalar score per test image; pixel-level metrics use anomaly maps at the original image support and all 496 released masks, including the all-zero masks for normal cases. The complete values are reported in Appendix C.3.

A.11 Dataset Composition

Table A.2 summarizes the expanded release. The 400 normal images support image-level evaluation and contribute all-zero masks to pixel-level evaluation, while the 96 anomaly cases retain the real-reference and controlled-construction design with annotated defect masks. The construction counts are descriptive dataset metadata; the source-only protocol uses only the prescribed image labels and masks during evaluation.

Table A.2: **Expanded BoltAD composition.** The release contains 400 normal images with all-zero masks and 96 pixel-annotated anomaly cases.

Item	Value
Total test images	496
Normal inspection images	400
Anomalous inspection images	96
Real anomaly references	18
Controlled anomaly cases	78
Anomaly categories	4
Bolt Loosening	58 (60.4%)
Bolt Damage	19 (19.8%)
Bolt Missing	10 (10.4%)
Bolt Corrosion	9 (9.4%)
Images with pixel-level masks	496 / 496 test cases
Target-domain training images	0

B Extended Implementation Details & Experimental Protocols

B.1 Vision Foundation Model Interfaces and Checkpoint Verification

SynAD uses a common encoder interface that returns an image-level summary vector and a sequence of patch tokens. For an input image x, we write $(s, F) = f_{\mathrm{VFM}}(x)$, where $s \in \mathbb{R}^C$ is the summary representation and $F \in \mathbb{R}^{N \times C}$ contains the patch representations. Each backbone uses a window equal to 48 patch strides: RADIO and DINOv3 process 768×768 global views and local tiles, while DINOv2 processes 672×672 views and tiles. Thus, every interface supplies a 48×48 token lattice to the shared decoder, which outputs a 96×96 logit map. Images use a border-safe reflection extension to the relevant window or padded canvas, repeating reflection when needed, while masks are zero-padded and accompanied by a binary validity mask. Padding is appended on the bottom and right, leaving the original image at the top-left of the padded canvas. Boundary tiles are end-aligned and retain the full backbone-specific window. During native-coordinate reconstruction, all valid tiles are uniformly averaged on the padded canvas, the validity mask excludes padded locations, and the result is cropped exactly to the original image support.

Our primary encoder is RADIO v2.5-L (Ranzinger et al., 2024). The implementation uses checkpoint `radio-v2.5-l_half.pth.tar`; RADIO uses patch size $P = 16$ and channel dimension $C = 1024$, and its local interface provides `output.summary` as s and `output.features` as F.

For the cross-backbone experiments, we use DINOv2-L/14-reg (Oquab et al., 2024) through the official Torch Hub entry, and DINOv3-L/16 (Siméoni et al., 2025) through its released `dinov3_vitl16` interface and weights. Both expose `x_norm_clstoken` as s and `x_norm_patchtokens` as F, with $C = 1024$. The input windows differ, but $672/14 = 768/16 = 48$, so all three encoders expose the same spatial lattice to the decoder.

Table B.1: **Encoder interfaces used by SynAD.** The summary token feeds the image-level predictor, and patch tokens are reshaped for the shared decoder.

Encoder	Window	Patch P	Grid	Summary / patch interface	
RADIO v2.5-L	768^2	16	48^2	`output.summary` `output.features`	/
DINOv2-L/14-reg	672^2	14	48^2	`x_norm_clstoken` `x_norm_patchtokens`	/
DINOv3-L/16	768^2	16	48^2	`x_norm_clstoken` `x_norm_patchtokens`	/

B.2 Parameter-Efficient Adaptation Configurations and Formulas

For an attention projection $p \in \{Q, V, O\}$ with frozen base weight W_0^p, the DoRA update is

$$\bar{W}^p = W_0^p + \gamma B^p A^p, \qquad \widehat{W}^p_{i,:} = m_i^p \frac{\bar{W}^p_{i,:}}{\left\|\bar{W}^p_{i,:}\right\|_2 + \epsilon}, \tag{20}$$

where $A^p \in \mathbb{R}^{r \times d_{\text{in}}}$, $B^p \in \mathbb{R}^{d_{\text{out}} \times r}$, and m^p is the learnable row-wise magnitude vector. We use $r = 64$ and $\alpha_{\text{DoRA}} = 1.0$, hence $\gamma = \alpha_{\text{DoRA}}/r = 1/64 = 0.015625$. The Key projection remains frozen, $\widehat{W}^K = W_0^K$.

All base RADIO parameters are frozen before adapter injection. The shared decoder, image-level predictor, and DoRA parameters are optimized jointly; the RADIO configuration has $35{,}476{,}483$ trainable parameters. In the fused QKV layer, the QVO policy reconstructs Query and Value with DoRA while retaining the original Key weights:

```
if name == "qkv":
    wrapper = DoRAQKV(child, adapt_key=False, rank=64)
elif name == "proj":
    wrapper = DoRALinear(child, rank=64)

final_k = base_k
weight = torch.cat((final_q, final_k, final_v), dim=0)
return F.linear(value, weight, bias)
```

Figure B.1 visualizes the controlled projection study that compares QVO with DoRA-QV, DoRA-QKVO, and a LoRA-QVO control.

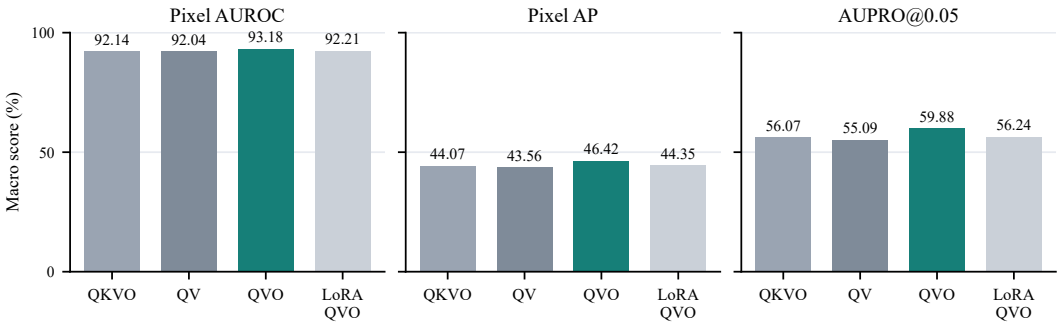


Figure B.1: **Selective projection adaptation.** All variants use the same backbone, source data, seeds, optimization schedule, and target evaluation protocol. QVO denotes DoRA on Query, Value, and Output with Key frozen.

B.3 Optimization Schedule, Batch Dynamics, and Infrastructure Profile

We optimize all trainable parameters with AdamW using learning rate 10^{-4} and weight decay 10^{-2}, without a learning-rate scheduler. Each optimizer update accumulates eight micro-batches of four images, giving an effective batch size of

$$B_{\text{eff}} = B_{\text{micro}} \times N_{\text{acc}} = 4 \times 8 = 32. \tag{21}$$

Training stops after exactly 200 optimizer updates rather than a fixed number of epochs. The source pair stream is shuffled and restarted when necessary, so the update count is independent of source-set size and incomplete final epochs. At each update, a fixed deterministic sampler selects one local tile per image; inference instead enumerates all valid local tiles before stitching.

Each model instance is trained on one NVIDIA GeForce RTX 3090 GPU with 24 GB memory; the measured peak allocation is approximately 13.11 GiB per run. The two source configurations and three random seeds produce six independent single-GPU training jobs. These jobs use task-level parallelism, not distributed data-parallel training, and exchange no gradients. Eight RTX 3090 GPUs are subsequently used only to shard target-domain inference. Target data are never used for training, source calibration, checkpoint selection, or threshold adaptation.

B.4 Source Calibration Search Dynamics and Stability of α^*

Let $\widetilde{Z}_G$ and Z_L denote the upsampled global and reconstructed local logit maps. The fusion coefficient controls their logit-space combination,

$$Z_\alpha = \alpha \widetilde{Z}_G + (1-\alpha) Z_L. \tag{22}$$

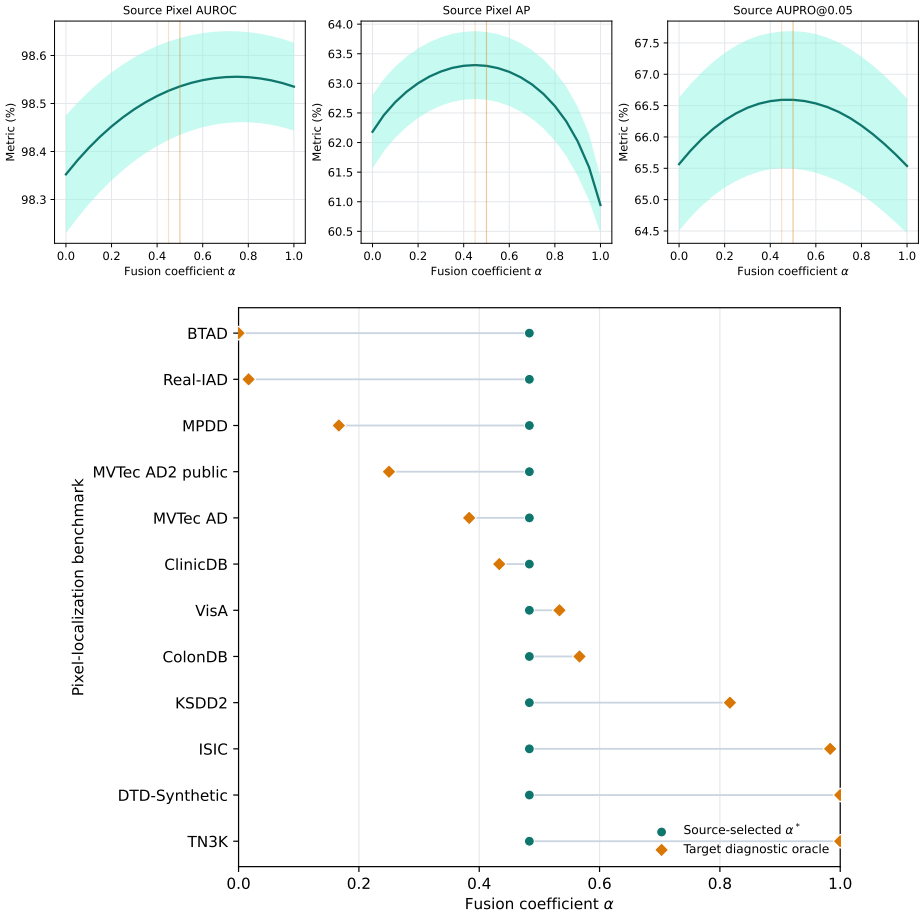


Figure B.2: **Source calibration sensitivity and target diagnostics.** Top: QVO source-validation curves over the fusion grid; vertical lines mark the source-selected coefficients for the primary experimental runs. Bottom: each row compares a source-selected coefficient with its target diagnostic oracle for the pixel-localization benchmarks. The target diagnostic is computed after prediction freezing and is not used for calibration.

For every source configuration and seed, calibration evaluates $\mathcal{A} = \{0, 0.05, \ldots, 1\}$ on the registered held-out source split. Consistent with Section 3.3, the selected coefficient maximizes $\mathcal{Q}_{\text{src}}$, which averages pixel F1-max and AUPRO@0.30, and is then frozen for downstream target evaluation.

For the seed-12 deployment runs used in the efficiency analysis, the source-selected values are $\alpha^* = 0.45$ for MVTec AD-trained checkpoints (used when VisA is the target) and $\alpha^* = 0.50$ for VisA-trained checkpoints (used for the remaining targets). Thus, those runs assign 55% or 50% of the fusion weight to local evidence, respectively. Figure B.2 shows the corresponding source searches and their improvement in AUPRO@0.05 over the single-view endpoints.

For analysis only, we also sweep α on deterministic category-stratified target subsets after predictions are frozen. Across 36 pixel-dataset–seed pairs, the mean AUPRO@0.05 hindsight regret relative to the diagnostic target oracle is 0.0154. This diagnostic neither changes the source-selected coefficient nor uses target data in the operating protocol.

B.5 Source-Domain Split Manifest and Leakage Prevention

SynAD uses only original labeled images from the public auxiliary source datasets. Under the source-only cross-dataset transfer protocol in Section 4.1, the eligible source pool comprises 2,162 VisA images or 1,725 MVTec AD images, depending on the target benchmark. We construct a fixed image-level 80/20 source-train/source-validation split with seed 3407, stratified jointly by category and image-level label. The resulting partitions are summarized in Table B.2.

The fixed source split manifests accompany the anonymous supplementary material. Each row records `image_id`, `dataset`, `category`, `label`, and `split`. Consequently, source-train and source-validation image IDs are disjoint. Training pairs are formed only after split assignment: both normal and anomalous images are drawn exclusively from source-train IDs. Source-validation IDs are used only to select α^*; they do not participate in gradient updates, training-pair construction, checkpoint selection, or threshold tuning.

SynAD does not generate synthetic source anomalies from VisA or MVTec AD. Hence, no normal base image has a synthesized derivative that can cross the source train/validation boundary. All BoltAD images, annotations, and masks are kept outside source training, source validation, calibration, checkpoint selection, and target adaptation.

Table B.2: **Fixed source-domain split manifests.** Candidate source pools contain original labeled public images used for source-only cross-dataset transfer. All counts are image-level.

Source dataset	Eligible pool	Train IDs	Validation IDs	Split rule
VisA	2,162	1,729	433	80/20 image-level; category × label stratified; seed 3407
MVTec AD	1,725	1,380	345	80/20 image-level; category × label stratified; seed 3407

C Per-Category Breakdown & Extended Quantitative Analyses

The full-metric results from the earlier RADIO comparison are split into portrait-page panels for legibility. Image-level tuples are ordered as (AUROC, F1-max, AP), and pixel-level tuples as (AUROC, F1-max, AP, AUPRO@0.30). WinCLIP is not included in this earlier comparison.

Table C.1: **Full-metric comparison, panel 1 of 3.** Tuples are ordered as (AUROC, F1-max, AP) for image-level and (AUROC, F1-max, AP, AUPRO@0.30) for pixel-level results; red/blue denote the best/second-best values.

Dataset	AnomalyCLIP	AdaCLIP	SynAD (Ours)
Industrial image-level (AUROC, F1-max, AP)			
MVTec AD	(93.89, 93.72, 97.05)	(89.99, 92.27, 95.67)	(**96.50**, **95.38**, **98.00**)
VisA	(88.76, 85.17, 90.14)	(86.37, 83.99, 88.51)	(**93.40**, **89.39**, **95.26**)
MPDD	(73.66, 78.95, 76.52)	(68.33, 77.88, 73.87)	(**85.75**, **87.59**, **90.34**)
Real-IAD	(70.01, 57.39, 55.40)	(71.52, 58.32, 58.07)	(**88.22**, **81.84**, **86.86**)
BTAD	(89.12, 85.99, 91.05)	(90.24, 88.55, 93.24)	(**93.79**, **91.14**, **95.65**)
KSDD2	(90.81, 71.79, 76.46)	(95.56, **85.15**, **89.31**)	(**97.34**, 82.94, **89.56**)
DTD-Synthetic	(94.46, 94.53, 97.70)	(91.32, 92.62, 95.41)	(**99.66**, **99.35**, **99.85**)
MVTec AD 2 public	(58.07, 80.79, 73.83)	(59.02, 79.98, 72.89)	(**67.94**, **81.91**, **81.13**)
Industrial pixel-level (AUROC, F1-max, AP, AUPRO@0.30)			
MVTec AD	(90.92, 47.24, 43.64, 84.95)	(83.39, 43.50, 37.44, 78.15)	(**92.74**, **50.20**, **51.49**, **89.14**)
VisA	(95.07, 27.25, 19.86, 86.58)	(86.77, **36.82**, **29.71**, 81.79)	(**95.95**, **35.81**, **30.63**, **90.88**)
MPDD	(92.88, 30.15, 24.69, 81.89)	(82.53, 29.97, 25.31, 69.59)	(**97.22**, **38.37**, **35.99**, **91.83**)
Real-IAD	(86.27, 32.54, 24.06, 73.02)	(87.83, 37.27, 32.42, 71.58)	(**96.59**, **42.60**, **38.95**, **87.70**)
BTAD	(80.97, 35.11, 29.38, 57.30)	(74.47, 43.82, 31.17, 48.76)	(**95.18**, **51.35**, **48.11**, **78.82**)
KSDD2	(98.04, 54.72, 49.61, 95.87)	(87.90, **66.72**, **64.76**, 87.23)	(**98.95**, 56.29, 59.49, 94.77)
DTD-Synthetic	(95.93, 56.54, 52.98, 89.56)	(94.29, **68.66**, **69.73**, 89.36)	(**99.48**, **71.55**, **78.52**, **97.03**)
MVTec AD 2 public	(72.06, 15.67, 9.22, 41.66)	(73.53, **20.03**, 12.93, 47.48)	(**86.83**, **25.37**, **21.11**, **68.67**)
Medical image-level (AUROC, F1-max, AP)			
Head-CT	(95.24, 90.73, 95.73)	(**97.36**, **92.68**, **97.55**)	(**97.00**, **93.33**, **97.58**)
BrainMRI	(75.12, 62.61, 51.23)	(60.75, 50.20, 45.67)	(**90.97**, **78.01**, **85.59**)
Br35H	(97.37, 92.38, 97.18)	(95.43, 89.70, 93.47)	(**98.66**, **94.87**, **98.66**)
Medical pixel-level (AUROC, F1-max, AP, AUPRO@0.30)			
ISIC	(87.11, **69.60**, **72.96**, 77.18)	(76.95, 67.35, 62.35, 68.29)	(**87.49**, 50.90, 41.55, **79.56**)
ColonDB	(81.87, 39.49, 33.38, 72.41)	(84.01, 30.62, 21.39, 77.00)	(**89.77**, **56.36**, **59.69**, **83.17**)
TN3K	(77.50, **45.15**, **42.18**, 49.26)	(70.95, 44.71, 31.92, 42.10)	(**85.92**, 36.73, 27.32, **52.61**)
ClinicDB	(84.24, 45.36, 40.42, 70.49)	(**91.95**, **75.61**, **79.37**, **88.15**)	(**92.06**, **62.56**, **64.23**, 82.03)

Table C.2: **Full-metric comparison, panel 2 of 3.** Tuples follow the order defined above; red and blue mark the best and second-best values.

Dataset	FAPrompt	MRAD	SynAD (Ours)
Industrial image-level (AUROC, F1-max, AP)			
MVTec AD	(91.80, 92.32, 95.62)	(93.87, 93.71, 97.30)	(**96.50**, **95.38**, **98.00**)
VisA	(85.01, 82.60, 86.71)	(85.21, 82.48, 87.97)	(**93.40**, **89.39**, **95.26**)
MPDD	(79.83, 82.25, 82.95)	(81.52, 82.92, 83.01)	(**85.75**, **87.59**, **90.34**)
Real-IAD	(78.04, 63.93, 66.74)	(74.32, 60.60, 62.49)	(**88.22**, **81.84**, **86.86**)
BTAD	(92.52, 89.93, 93.70)	(92.24, 88.27, 94.43)	(**93.79**, **91.14**, **95.65**)
KSDD2	(96.51, 84.54, 89.18)	(92.97, **86.29**, 86.35)	(**97.34**, 82.94, **89.56**)
DTD-Synthetic	(96.10, 95.08, 98.34)	(95.79, 94.77, 98.32)	(**99.66**, **99.35**, **99.85**)
MVTec AD 2 public	(60.42, 81.00, 75.69)	(57.38, 80.40, 72.37)	(**67.94**, **81.91**, **81.13**)
Industrial pixel-level (AUROC, F1-max, AP, AUPRO@0.30)			
MVTec AD	(87.97, 39.15, 33.60, 83.82)	(**92.60**, 45.27, 43.20, 87.73)	(**92.74**, **50.20**, **51.49**, **89.14**)
VisA	(95.27, 27.14, 20.28, 86.88)	(**95.79**, 32.08, 25.50, 87.98)	(**95.95**, **35.81**, **30.63**, **90.88**)
MPDD	(95.51, 32.67, 27.18, 87.14)	(**97.78**, **39.11**, 34.59, **91.33**)	(**97.22**, **38.37**, **35.99**, **91.83**)
Real-IAD	(95.60, 35.18, 27.45, **87.00**)	(95.41, 32.85, 25.67, 83.38)	(**96.59**, **42.60**, **38.95**, **87.70**)
BTAD	(92.34, 45.77, 40.40, 68.77)	(**95.42**, **47.73**, **43.61**, **74.67**)	(**95.18**, **51.35**, **48.11**, **78.82**)
KSDD2	(98.33, 60.60, 56.15, **96.57**)	(**98.93**, 60.84, 60.79, **96.10**)	(**98.95**, 56.29, 59.49, 94.77)
DTD-Synthetic	(97.18, 61.56, 60.22, 91.83)	(98.10, 60.75, 61.02, 93.24)	(**99.48**, **71.55**, **78.52**, **97.03**)
MVTec AD 2 public	(82.36, 19.75, 12.92, 49.56)	(81.80, 18.27, 13.01, 48.47)	(**86.83**, **25.37**, **21.11**, **68.67**)
Medical image-level (AUROC, F1-max, AP)			
Head-CT	(94.39, 90.53, 91.18)	(94.15, 89.76, 93.10)	(**97.00**, **93.33**, **97.58**)
BrainMRI	(83.48, 70.11, 69.60)	(80.71, 65.42, 66.04)	(**90.97**, **78.01**, **85.59**)
Br35H	(97.41, 93.52, 96.75)	(**97.89**, **93.91**, **97.69**)	(**98.66**, **94.87**, **98.66**)
Medical pixel-level (AUROC, F1-max, AP, AUPRO@0.30)			
ISIC	(84.70, 66.10, 69.70, 72.86)	(**91.56**, **74.53**, **81.76**, **83.66**)	(**87.49**, 50.90, 41.55, **79.56**)
ColonDB	(82.21, 38.25, 31.87, 72.33)	(**84.64**, **41.53**, **34.94**, 73.51)	(**89.77**, **56.36**, **59.69**, **83.17**)
TN3K	(81.44, **47.69**, **45.43**, 50.07)	(77.96, 43.68, 42.15, **51.53**)	(**85.92**, 36.73, 27.32, **52.61**)
ClinicDB	(83.40, 43.04, 37.22, 68.73)	(87.28, 50.23, 49.57, 73.87)	(**92.06**, **62.56**, **64.23**, 82.03)

Table C.3: **Full-metric comparison, panel 3 of 3.** Tuples follow the order defined above; red and blue mark the best and second-best values.

Dataset	VisualAD	AnomalyVFM	SynAD (Ours)
Industrial image-level (AUROC, F1-max, AP)			
MVTec AD	(90.92, 92.21, 96.30)	(**96.42**, **95.09**, **97.88**)	(**96.50**, **95.38**, **98.00**)
VisA	(85.37, 82.65, 87.33)	(**92.91**, **88.70**, **94.63**)	(**93.40**, **89.39**, **95.26**)
MPDD	(72.43, 80.13, 77.16)	(**85.70**, **87.00**, **89.83**)	(**85.75**, **87.59**, **90.34**)
Real-IAD	(79.00, 64.91, 68.80)	(**88.31**, **81.73**, **86.59**)	(**88.22**, **81.84**, **86.86**)
BTAD	(**93.63**, 87.54, 92.28)	(93.41, **91.38**, **95.80**)	(**93.79**, **91.14**, **95.65**)
KSDD2	(**97.34**, 79.15, 87.36)	(**97.30**, 83.00, 88.93)	(**97.34**, 82.94, **89.56**)
DTD-Synthetic	(96.94, 96.98, 99.02)	(**99.55**, **99.26**, **99.79**)	(**99.66**, **99.35**, **99.85**)
MVTec AD 2 public	(61.82, 80.55, 75.59)	(**67.58**, **81.79**, **80.52**)	(**67.94**, **81.91**, **81.13**)
Industrial pixel-level (AUROC, F1-max, AP, AUPRO@0.30)			
MVTec AD	(70.09, 36.23, 25.66, 66.21)	(91.91, **49.14**, **49.78**, **88.59**)	(**92.74**, **50.20**, **51.49**, **89.14**)
VisA	(76.51, 29.68, 20.23, 63.69)	(95.55, 32.53, 27.12, **89.60**)	(**95.95**, **35.81**, **30.63**, **90.88**)
MPDD	(66.76, 23.92, 13.52, 40.64)	(96.75, 37.46, **35.46**, 90.57)	(**97.22**, **38.37**, **35.99**, **91.83**)
Real-IAD	(82.40, 32.49, 24.46, 67.41)	(**96.26**, **40.86**, **36.30**, 86.06)	(**96.59**, **42.60**, **38.95**, **87.70**)
BTAD	(59.34, 26.44, 13.00, 31.64)	(89.12, 39.09, 33.37, 60.50)	(**95.18**, **51.35**, **48.11**, **78.82**)
KSDD2	(76.93, 42.79, 29.46, 79.55)	(97.72, **62.22**, **62.17**, 95.18)	(**98.95**, 56.29, 59.49, 94.77)
DTD-Synthetic	(81.96, 57.08, 44.95, 75.24)	(**98.91**, 64.64, 68.14, **94.88**)	(**99.48**, **71.55**, **78.52**, **97.03**)
MVTec AD 2 public	(58.42, 14.51, 7.55, 31.80)	(**86.72**, 17.70, **13.64**, **50.25**)	(**86.83**, **25.37**, **21.11**, **68.67**)
Medical image-level (AUROC, F1-max, AP)			
Head-CT	(87.31, 82.67, 83.53)	(94.29, 88.99, 93.14)	(**97.00**, **93.33**, **97.58**)
BrainMRI	(79.06, 65.55, 63.60)	(**91.30**, **78.23**, **85.73**)	(**90.97**, **78.01**, **85.59**)
Br35H	(95.25, 90.11, 91.57)	(95.16, 89.31, 93.12)	(**98.66**, **94.87**, **98.66**)
Medical pixel-level (AUROC, F1-max, AP, AUPRO@0.30)			
ISIC	(57.22, 43.97, 37.34, 41.84)	(87.48, 49.98, 41.57, 77.19)	(**87.49**, 50.90, 41.55, **79.56**)
ColonDB	(58.61, 26.30, 14.61, 43.49)	(82.17, 30.09, 20.13, **77.04**)	(**89.77**, **56.36**, **59.69**, **83.17**)
TN3K	(53.23, 22.76, 16.91, 20.69)	(**85.48**, 35.72, 27.13, 46.30)	(**85.92**, 36.73, 27.32, **52.61**)
ClinicDB	(61.57, 33.96, 20.46, 45.74)	(90.75, 59.39, 58.25, **82.21**)	(**92.06**, **62.56**, **64.23**, 82.03)

C.3 FULL BOLTAD BENCHMARK RESULTS

The following tables report the server results for the expanded 496-image test release. Image-level metrics use all 400 normal and 96 anomalous image labels. Pixel-level metrics are computed on all 496 test cases using the released masks, including all-zero masks for the 400 normal cases. SynAD is averaged over three seeds, whereas the other methods are single reported runs. The construction and review procedure are described in Appendix A.6–A.8; these tables report pooled results for the benchmark.

Table C.4: **BoltAD image-level results on the 496-image test protocol (%).** The three image metrics are AUROC, F1-max, and AP. Red and blue indicate the best and second-best values, respectively.

Method	AUROC	F1-max	AP
AdaCLIP (Cao et al., 2024)	60.26	36.05	28.56
AnomalyCLIP (Zhou et al., 2024)	**66.89**	**43.15**	34.18
FAPrompt (Zhu et al., 2025)	65.14	38.30	32.80
VisualAD (Hou et al., 2026)	61.27	36.32	**36.67**
WinCLIP (Jeong et al., 2023)	64.52	39.51	25.69
AnomalyVFM (Fučka et al., 2026)	60.15	35.33	32.15
MRAD (Xu et al., 2026)	64.12	**40.41**	34.49
SynAD (Ours)	**65.24 ± 1.06**	39.33 ± 1.20	**42.57 ± 0.79**

Table C.5: **BoltAD pixel-level results on the complete test release (%).** Pixel metrics are AUROC, F1-max, AP, and AUPRO@0.30, evaluated on all 496 test cases using the released masks, including all-zero masks for normal cases.

Method	AUROC	F1-max	AP	AUPRO@0.30
AdaCLIP (Cao et al., 2024)	81.52	25.67	15.72	69.38
AnomalyCLIP (Zhou et al., 2024)	90.10	23.66	14.75	68.88
FAPrompt (Zhu et al., 2025)	90.32	23.98	16.73	64.40
VisualAD (Hou et al., 2026)	90.65	25.58	**18.67**	75.25
WinCLIP (Jeong et al., 2023)	73.92	9.55	4.30	45.85
AnomalyVFM (Fučka et al., 2026)	**91.53**	**27.83**	18.33	**79.75**
MRAD (Xu et al., 2026)	**91.88**	24.38	17.91	69.92
SynAD (Ours)	91.06 ± 0.21	**29.54 ± 0.29**	**20.39 ± 0.52**	**80.69 ± 0.83**

C.4 Extended Multi-Backbone Comparison

Tables C.6 and C.7 report the dataset-wise results for each encoder under the same source-only transfer protocol. Table C.6 contains the 11 image-level benchmarks, and Table C.7 contains the 12 pixel-level benchmarks: MVTec AD, VisA, MPDD, Real-IAD, BTAD, KSDD2, DTD-Synthetic, MVTec AD 2 public, ISIC, ColonDB, TN3K, and ClinicDB. Figure C.1 visualizes the corresponding macro averages. All macros use the unweighted arithmetic mean over the respective dataset sets, rather than a sample-weighted average.

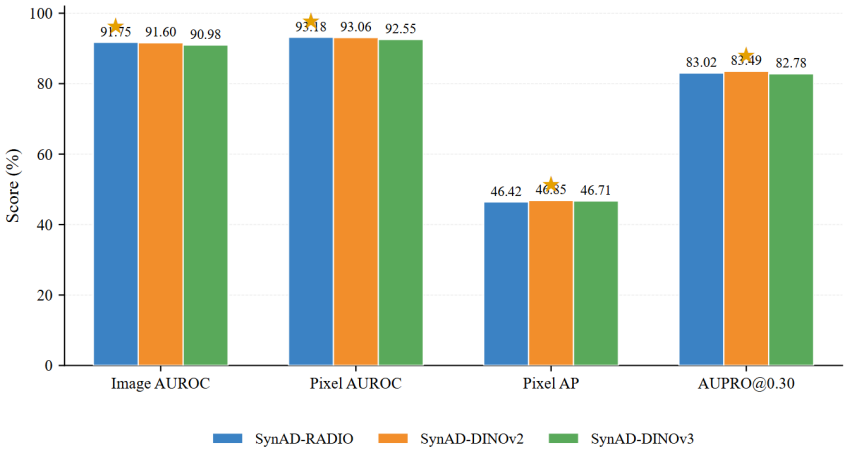


Figure C.1: **Cross-backbone macro comparison.** Bars show unweighted dataset means for SynAD-RADIO, SynAD-DINOv2, and SynAD-DINOv3 on image AUROC and pixel AUROC, AP, and AUPRO@0.30.

Table C.6: **Image-level results across visual backbones.** Values are percentages; metrics within each tuple follow the order (AUROC, F1-max, AP).

Dataset	SynAD-RADIO	SynAD-DINOv2	SynAD-DINOv3
Industrial benchmarks			
MVTec AD	(96.50, 95.38, 98.00)	(96.42, 95.11, 98.23)	(94.15, 95.07, 95.89)
VisA	(93.40, 89.39, 95.26)	(93.12, 89.88, 95.01)	(92.35, 88.12, 93.94)
MPDD	(85.75, 87.59, 90.34)	(85.59, 86.72, 90.21)	(84.54, 88.81, 88.56)
Real-IAD	(88.22, 81.84, 86.86)	(88.01, 82.11, 85.75)	(87.45, 79.55, 85.12)
BTAD	(93.79, 91.14, 95.65)	(94.35, 89.65, 95.38)	(94.02, 90.89, 95.73)
KSDD2	(97.34, 82.94, 89.56)	(96.31, 83.02, 89.22)	(94.35, 80.56, 88.18)
DTD-Synthetic	(99.66, 99.35, 99.85)	(99.85, 99.54, 99.87)	(99.72, 99.72, 99.91)
MVTec AD 2 public	(67.94, 81.91, 81.13)	(69.22, 80.75, 81.84)	(67.78, 82.51, 82.68)
Medical benchmarks			
Head-CT	(97.00, 93.33, 97.58)	(96.38, 93.65, 98.36)	(97.12, 93.15, 96.27)
BrainMRI	(90.97, 78.01, 85.59)	(90.57, 78.82, 85.79)	(91.21, 77.18, 85.13)
Br35H	(98.66, 94.87, 98.66)	(97.75, 93.01, 98.44)	(98.12, 94.26, 98.78)

Table C.7: **Pixel-level results across visual backbones.** Values are percentages; metrics within each tuple follow the order (AUROC, F1-max, AP, AUPRO@0.30).

Dataset	SynAD-RADIO	SynAD-DINOv2	SynAD-DINOv3
Industrial benchmarks			
MVTec AD	(92.74, 50.20, 51.49, 89.14)	(91.35, 51.74, 51.22, 90.53)	(91.58, 51.99, 51.58, 88.49)
VisA	(95.95, 35.81, 30.63, 90.88)	(96.47, 36.99, 32.24, 90.02)	(96.12, 35.85, 30.83, 90.67)
MPDD	(97.22, 38.37, 35.99, 91.83)	(98.01, 38.86, 35.77, 92.01)	(97.47, 39.21, 36.62, 90.96)
Real-IAD	(96.59, 42.60, 38.95, 87.70)	(95.86, 41.98, 39.97, 87.29)	(96.78, 42.23, 39.71, 87.94)
BTAD	(95.18, 51.35, 48.11, 78.82)	(95.88, 53.81, 47.93, 79.85)	(93.69, 49.67, 47.85, 77.17)
KSDD2	(98.95, 56.29, 59.49, 94.77)	(96.97, 58.25, 58.52, 93.89)	(97.14, 57.01, 59.87, 95.55)
DTD-Synthetic	(99.48, 71.55, 78.52, 97.03)	(99.33, 71.88, 78.93, 97.22)	(99.26, 72.22, 78.77, 97.88)
MVTec AD 2 public	(86.83, 25.37, 21.11, 68.67)	(86.85, 28.66, 22.33, 70.88)	(87.11, 28.51, 21.68, 70.44)
Medical benchmarks			
ISIC	(87.49, 50.90, 41.55, 79.56)	(88.43, 51.17, 44.53, 79.87)	(87.77, 48.68, 43.57, 77.37)
ColonDB	(89.77, 56.36, 59.69, 83.17)	(88.45, 56.86, 61.54, 83.33)	(86.49, 56.19, 57.87, 82.92)
TN3K	(85.92, 36.73, 27.32, 52.61)	(86.26, 37.99, 28.18, 54.83)	(84.21, 34.44, 27.75, 52.35)
ClinicDB	(92.06, 62.56, 64.23, 82.03)	(92.88, 62.85, 61.01, 82.12)	(92.92, 62.27, 64.46, 81.58)

C.4.1 Qualitative Comparison Across Backbones

Figure C.2 complements the dataset-wise tables with the same industrial and medical cases evaluated by the three SynAD instantiations. Across these examples, all backbones activate on the annotated anomalies, while their spatial concentration differs. DINOv3 produces broader responses in several structured or reflective cases, including Hazelnut, Zipper, Toothbrush, Colon, and Thyroid. RADIO and DINOv2 are often more spatially restrained in these examples, although this qualitative behavior is not uniform across categories. The figure is illustrative; Tables C.6 and C.7 remain the basis for quantitative conclusions.

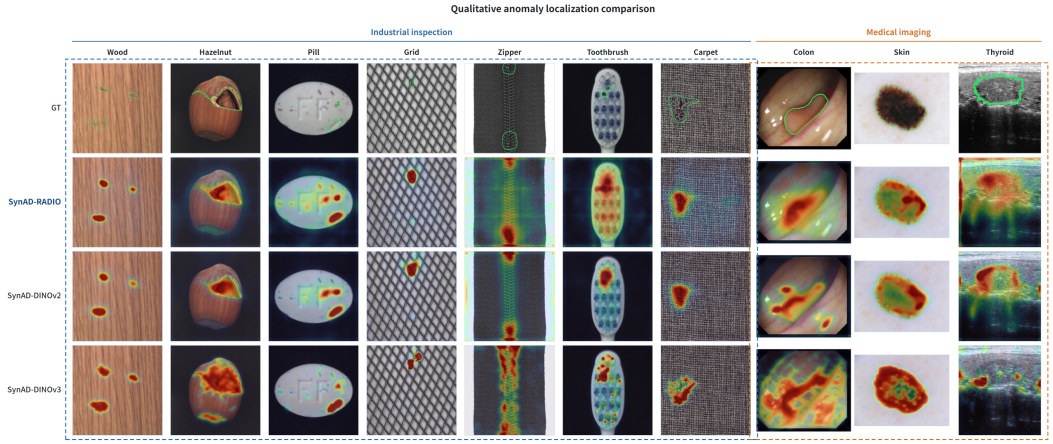


Figure C.2: **Qualitative anomaly localization across visual backbones.** Columns show seven industrial and three medical cases. The first row gives ground-truth overlays; the remaining rows show anomaly maps from SynAD instantiated with RADIO, DINOv2, and DINOv3. Blue and orange dashed frames denote industrial and medical examples, respectively.

C.5 Inference Efficiency and Computational Profile

We measure end-to-end inference under the source-only transfer protocol on one NVIDIA GeForce RTX 3090 GPU (24 GB, `cuda:0`); the remaining GPUs are idle during timing. Models use `eval()` and `torch.inference_mode()` in FP32 with autocast disabled. We use the seed-12 source checkpoints, with VisA-source weights for 11 targets and MVTec-source weights for VisA. Their fixed fusion coefficients are 0.50 and 0.45, respectively.

The timed path includes global resizing, local tiling, local microbatch forwarding (size 8), native-coordinate restoration, logit fusion, and smoothing with coefficient 0.04; RGB file read and decode are excluded. External image batch size is one. For each source checkpoint and inference path, we warm up on 30 images and measure 100 deterministically category-balanced images per dataset, with CUDA synchronization immediately before and after every timed image. Thus, each path pools 1,200 timed images from the 12 pixel-level benchmarks; because every dataset contributes 100 images, the pooled mean equals the unweighted dataset macro mean.

The contract-locked inference manifest contains 30,265 pixel-evaluable RGB images. The evaluation script verifies its SHA-256 digest, uniqueness, and file existence, rejects ground-truth fields, and records `target_gt_read=false` and `target_data_used_for_training_or_calibration=false`. Table C.8 reports the main results. Peak memory is `torch.cuda.max_memory_allocated()`, not reserved memory.

Table C.8: **End-to-end inference efficiency of SynAD.** Timings pool 1,200 images per path at external batch size one. MAC counts cover neural-network forward operators; stitching, fusion, and smoothing are included in latency but excluded from MAC estimates.

Path	Total params	Train. params	Tiles / image	TMAC / image	Peak GiB	Mean ms	Median ms	P95 ms	FPS (B=1)
Global-only	355.358M	35.476M	0.000	1.077	2.691	187.07	183.03	215.77	5.346
Local-only	355.358M	35.476M	4.038	4.349	3.276	643.36	346.51	2366.41	1.554
SynAD	355.358M	35.476M	4.038	5.426	3.297	836.65	532.70	2616.67	1.195

Table C.9 summarizes full-manifest tile counts for the RADIO timing protocol, whose backbone-specific window is 768×768. Small images are padded to one such tile, whereas images with many tiles are processed in local microbatches of at most eight. The timing subset has a mean of 3.782 local tiles because it uses category-balanced sampling; the full-manifest pooled mean is 4.038. The DINOv2 backbone comparison instead uses its 672×672 window, as specified in Appendix B.1.

Table C.9: **Full-manifest local tile-count distribution.** Counts use all 30,265 images in the 12 pixel-level benchmark manifests. P95 follows NumPy's linear percentile convention.

Dataset	Images	Mean	Median	P95	Min	Max
MVTec AD	1,725	3.800	4	4	1	4
VisA	2,162	5.399	6	6	4	6
MPDD	458	4.000	4	4	4	4
Real-IAD	19,802	4.000	4	4	4	4
BTAD	741	2.351	2	9	1	9
KSDD2	1,004	1.000	1	1	1	1
DTD-Synthetic	1,304	1.000	1	1	1	1
ISIC	379	6.897	2	40	1	40
ColonDB	380	1.000	1	1	1	1
TN3K	614	1.252	1	2	1	4
ClinicDB	612	1.000	1	1	1	1
MVTec AD 2 public	1,084	13.395	16	16	8	16
Pooled distribution	**30,265**	**4.038**	**4**	**6**	**1**	**40**
Unweighted dataset mean	**12 datasets**	**3.758**	–	–	–	–

Table C.10 reports the MAC-equivalent estimate. We use 1 MAC $=$ 2 FLOPs and include RADIO/DoRA, decoder, and the global image predictor where applicable. A custom handler cov-

ers scaled-dot-product attention; unsupported elementwise, normalization, and GELU operators are recorded in the supplementary summary and make these approximate MAC-equivalent counts rather than exact hardware instruction counts.

Table C.10: **Neural-network compute estimates.** Local and fused totals use the full-manifest mean of 4.038 local tiles per image.

Quantity	TMAC-equivalent	TFLOPs ($2\times$ MAC)
One 768×768 global branch	1.077	2.154
One 768×768 local tile	1.077	2.154
Local-only image at mean tile count	4.349	8.698
Fused global + local image at mean tile count	5.426	10.852

The anonymous supplementary material contains `resolution_stats_by_dataset.csv`, `latency_summary_by_dataset_mode.csv`, `latency_by_dataset.csv`, and `latency_samples.csv`. We report only unified SynAD measurements: external baselines are not assigned direct latency or throughput comparisons because their public implementations use different preprocessing and inference pipelines.

C.6 Extended Hyperparameter Sensitivity

C.6.1 Overlap-Ratio Sensitivity

For the RADIO configuration, we vary the local-tile overlap ratio $\rho \in \{0.10, 0.25, 0.40\}$. Using $S_h = S_w = \mathrm{round}((1-\rho)S_b)$ gives strides of 691, 576, and 461 pixels, respectively. Each setting repeats the same source-only protocol: its fusion coefficient α^* is selected on the held-out source validation partition and frozen before target evaluation. No target-domain data is used for calibration or model selection. Table C.11 reports localization accuracy together with the resulting tiling cost. The pixel metrics use the 12 pixel-evaluable benchmarks, whereas the Tiles/image column uses the separate 15-benchmark target manifest described below.

The tile-count means in this table and the inference-efficiency section use different populations and aggregation levels. In the efficiency section, 4.038 is the pooled image-level mean over the 12 pixel-evaluable benchmarks (30,265 images). Table C.11 instead reports a dataset-macro mean over the 15-benchmark target manifest: category means are weighted by category sample counts within each dataset, and the resulting dataset means are then averaged equally. This yields 3.215 tiles per image at $\rho = 0.25$. Directly pooling all 34,249 images in the overlap-sensitivity manifest yields 3.695, which is a separate image-weighted statistic.

Table C.11: **Source-only re-calibrated overlap-ratio sensitivity.** Each ρ independently selects α^* on held-out source validation data. Pixel metrics are unweighted macro averages over the 12 pixel-evaluable benchmarks. Tiles/image is a separate dataset-macro mean over the 15-benchmark target manifest, with category means sample-weighted within each dataset. FPS and peak memory are measured in the overlap-sensitivity timing run and are intended for within-sweep comparison; Table C.8 uses a separate end-to-end timing protocol. Peak memory is the maximum allocated GPU memory.

ρ	Stride	Pixel AUROC	Pixel F1-max	Pixel AP	AUPRO@0.30	Tiles / image	FPS	Peak memory
0.10	691	0.92273	0.46596	0.44689	0.81032	2.942	**1.947**	3.586 GB
0.25	576	**0.92311**	**0.46746**	0.44723	**0.81107**	3.215	1.937	3.586 GB
0.40	461	0.92253	0.46636	**0.44735**	0.81082	3.580	1.921	3.586 GB

The deployment setting $\rho = 0.25$ obtains the best Pixel AUROC, F1-max, and AUPRO@0.30. Increasing overlap to $\rho = 0.40$ raises Pixel AP by only 0.00012, while requiring 11.4% more local tiles per image and reducing throughput from 1.937 to 1.921 FPS. Peak allocated memory is unchanged. A fixed-coefficient control that reuses the primary $\rho = 0.25$ source-selected coefficients preserves the same ordering on AUROC, F1-max, and AUPRO@0.30.

C.7 Additional View-Adaptation Ablations

Table C.12 reports the complete five-benchmark comparison between shared DoRA (E_0) and separate global/local DoRA banks (E_1). Both variants share the frozen backbone, decoder, image-level head, source data, and evaluation protocol. Only the adapter parameters are duplicated in E_1. This experiment examines whether additional view-specific capacity improves performance. The separate configuration is lower on every benchmark despite using twice as many adapter parameters.

Table C.12: **Shared versus separate DoRA adaptation.** Values are AUPRO@0.05. E_0 uses one adapter bank for both views (9.51M adapter parameters); E_1 uses separate global and local banks (19.02M adapter parameters). The macro difference is $E_1 - E_0 = -0.0166$ (95% bootstrap CI: $[-0.0184, -0.0149]$); no reported bootstrap draw favors E_1.

Benchmark	**Shared** E_0	**Separate** E_1	$E_1 - E_0$
MVTec AD	0.699633	0.689843	-0.009790
VisA	0.693609	0.673070	-0.020539
MPDD	0.612605	0.594507	-0.018098
Real-IAD	0.613431	0.601983	-0.011448
MVTec AD 2 public	0.356942	0.333832	-0.023110
Macro mean	**0.595244**	**0.578647**	**-0.016597**

C.7.1 Independent Single-View Protocols

Table C.13 specifies the independently trained variants used in the upper block of Table 4. All variants use frozen RADIO v2.5-L, the same source split, DoRA on the selected attention projections, decoder architecture, optimizer, and update count. Local-only has no image-level head and therefore reports pixel-level metrics only.

Table C.13: **Training and evaluation protocols for the view-formulation ablation.** $\mathcal{L}_{\text{seg}}^{G}$ and $\mathcal{L}_{\text{seg}}^{L,j_i(t)}$ denote the global and deterministically sampled local segmentation terms in Equation 17; $\mathcal{L}_{\text{img}}^{G}$ is the global image-level BCE term.

Variant	**Training view**	**Testing view**	**Training loss**
Global-only	Resized global image	Resized global image	$\lambda_{\text{seg}}\mathcal{L}_{\text{seg}}^{G} + \mathcal{L}_{\text{img}}^{G}$
Local-only	One deterministic local tile per image and update	Full $\rho = 0.25$ tiling with native-coordinate reconstruction	$\lambda_{\text{seg}}\mathcal{L}_{\text{seg}}^{L,j_i(t)}$
SynAD	Global image and local tiles	Global–local logit fusion	Joint objective in Equation 17