Prototype–GVA–IDAC: Prototype Localization and Dual-Branch Scoring for Cross-Dataset Industrial Anomaly Detection

Anonymous authors
Paper under double-blind review

Abstract

Industrial visual anomaly detection models need not only to localize small defects in unseen products and acquisition environments but also to provide reliable image-level decisions. However, appearance shifts across datasets can weaken the correspondence between text semantics and local visual evidence; MultiADS directly averages the global text-based anomaly probability and the maximum response of the anomaly map, making it difficult to achieve both fine-grained localization and stable discrimination. To address this, we propose the Prototype–GVA–IDAC model based on MultiADS with frozen CLIP encoders. Prototype enhances multi-level patch representations through soft prototype assignment, context reconstruction, and residual refinement, allowing local anomaly cues to transfer more reliably to unseen datasets; GVA independently aggregates global image features and multi-level local statistics to provide additional visual discriminative evidence for image-level prediction; IDAC separately calibrates normal and abnormal text anchors under identity-preservation and normal/abnormal separation constraints, reducing the mismatch between fixed text semantics and target-domain images. The model is trained on MVTec AD without using any target-domain data for intervention, hyperparameter selection, or threshold adjustment during the process, and is tested on BTAD, MPDD, KSDD2, VisA, and WFDD. Experimental results show that, under the same protocol, the model offers significant advantages in cross-dataset defect localization and anomaly discrimination.

1 Introduction

When an industrial inspection model is deployed on a new production line, it often encounters different product shapes, surface textures, and imaging conditions, while defects themselves may occupy only small regions. Collecting annotations and retraining the model for each change increases deployment and maintenance costs. Cross-dataset anomaly detection therefore poses a practical requirement: a model trained with source-domain supervision alone must both accurately locate defects and determine whether an entire image is anomalous on datasets not used for training.

This paper focuses on how localization features can support reliable image-level decisions. MultiADS averages the global text-based anomaly probability and the maximum anomaly-map response: one summarizes image–text semantics, while the other represents the strongest local response. This aggregation does not explicitly learn the relationship between multiple local responses and global appearance. We therefore model patch-context refinement and image-level evidence aggregation separately, allowing the localization branch to preserve spatial details and the discrimination branch to learn the feature combinations needed for image-level decisions. Figure 1 presents localization comparisons; Appendices F and L provide the complete candidate analysis and additional examples.

Prototype–GVA–IDAC defines its module interfaces around this distinction. Prototype aggregates within-image prototypes from four levels of patch tokens, supplements local representations with reconstructed context, and preserves patch differences through gated residuals. GVA reads global CLIP features and multi-level local statistics to learn image anomaly scores directly; IDAC adjusts normal and abnormal text anchors to calibrate image-level semantic decisions.

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.

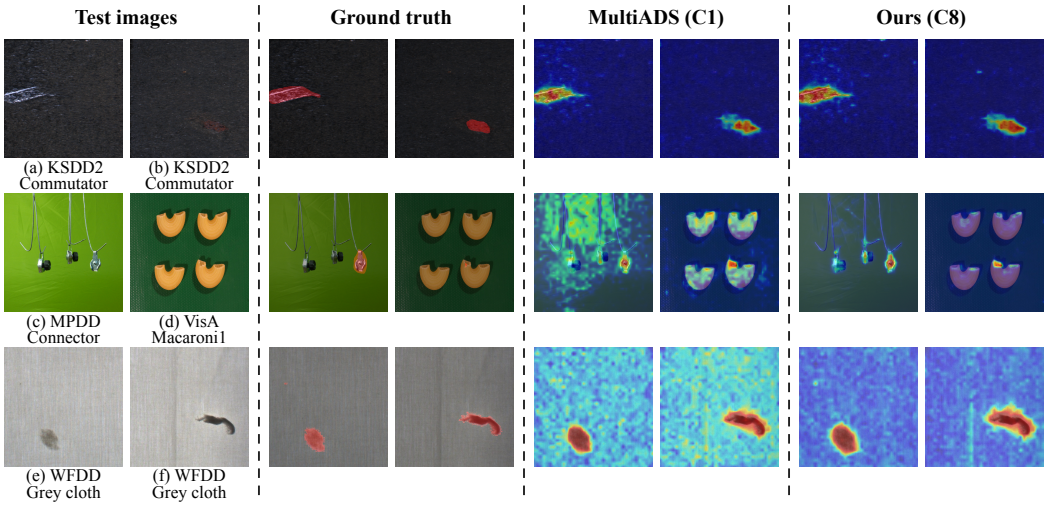


Figure 1: Actual localization comparisons grouped by method. The four groups show inputs, ground truth, MultiADS (C1), and our method (C8). The two columns and three rows within each group are aligned, with KSDD2, MPDD/VisA, and WFDD examples in the top, middle, and bottom rows, respectively. Dashed lines separate the method/data groups. The complete images and archived heatmaps retain their original visualizations; predictions and color mappings have not been modified. These are selected successful examples, not estimates of overall performance.

2 Related Work

Modeling normal distributions. Local features of normal samples provide an important reference for anomaly detection. PatchCore organizes them into a memory bank, whereas PaDiM estimates their statistical distributions (Roth et al., 2022; Defard et al., 2021); reconstruction and distillation approaches locate defects through reconstruction errors or feature discrepancies (Zavrtanik et al., 2021; Deng & Li, 2022; Batzner et al., 2024). These methods generally require normal images from the target category. We use a different source of context: Prototype generates its prototypes from the current image without building a target-domain normal memory bank, while training supervision still comes from the source domain.

Vision–language adaptation. WinCLIP combines prompts with window features (Jeong et al., 2023), while AnomalyCLIP, AdaCLIP, and MGVCLIP adapt prompts or multi-level features (Zhou et al., 2024; Cao et al., 2024; Cheng et al., 2025). MultiADS uses defect-aware source segmentation supervision (Sadikaj et al., 2025). AA-CLIP adapts anomaly-aware text anchors and patch features, and AdaptCLIP combines textual, visual, and prompt-query adapters (Ma et al., 2025; Gao et al., 2026). These are close precedents for separating visual and textual adaptation. IDAC instead applies a small BCE-trained residual after text encoding, only to the image-scoring anchors. Proximity-CLIP studies visually calibrated semantic margins and bounded regularization (Yang et al., 2026); its objectives differ from the evaluated residual correction.

Visual prototypes and separate discrimination. Compact context aggregation predates this work: A^2-Nets gathers and redistributes features, and EMANet reconstructs features from a compact set of bases (Chen et al., 2018; Li et al., 2019). Prototype follows this aggregation–reconstruction family, using image-conditioned assignments and gated residuals at four frozen-CLIP levels. VisualAD uses normal/abnormal visual tokens, while Tipsomaly separates image and pixel prompts (Hou et al., 2026; Salehi et al., 2026). Our contribution is the evaluated integration of local refinement and independent image scoring within MultiADS, not the general idea of compact context reconstruction.

Scoring and evaluation protocols. How image-level scores are formed also affects detection. GVA learns a decision rule from source-domain global features and patch statistics, whereas RareCLIP adapts during testing through online memory updates (He et al., 2025). Our main results use offline test-batch normalization. Appendix A.1 distinguishes source supervision, target adapta-

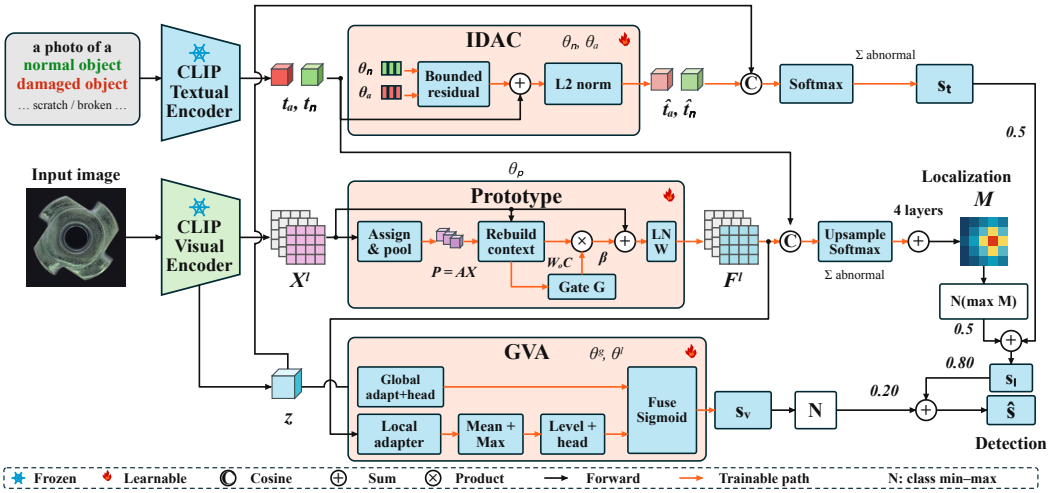


Figure 2: Inference structure of the proposed method. The input illustration is a normal metal-nut image from the MVTec AD training set. Prototype expands prototype aggregation, context reconstruction, and gated residual refinement; LN/W includes normalization and the subsequent visual projection. GVA adapts features before pooling, while IDAC calibrates text anchors after encoding; the original anchors are still used for localization. X^ℓ, F^ℓ, z denote raw tokens, projected tokens, and global features, respectively; N denotes within-category min–max normalization. Colored grids are schematic tensors, not predictions for the illustrated training image.

tion, and access to target statistics, defining the respective comparison boundaries for same-protocol results and externally evaluated native configurations.

3 Method

3.1 Problem Setting and the MultiADS Base Model

The model receives source images with image-level labels and pixel annotations and learns to output an anomaly map $M(x)$ and an image score $\hat{s}(x)$ for a test image x from other datasets. Target images are not used for gradient-based training and do not trigger test-time parameter updates. Prompt construction follows MultiADS and retains product and defect names. Thus, zero-shot in this paper specifically means *zero target-domain training images*; source anomaly annotations and category semantics remain available information.

MultiADS freezes the CLIP image and text encoders (Radford et al., 2021) and extracts global image features z and patch features X^ℓ from layers $\ell \in \{6, 12, 18, 24\}$. On the text side, prompts are grouped by the normal state and each defect type. Averaging the embeddings within each group yields a normal anchor t_n and abnormal anchors $\{t_{a,j}\}_{j=1}^{J}$. A linear adapter projects each layer's visual tokens into the text space, where comparison with these anchors produces multi-channel similarity maps. Multi-class Focal loss supervises defect types; the complement of the normal channel forms a binary anomaly probability, with Dice loss supervising localization (Lin et al., 2017; Milletari et al., 2016). The original paper averages the multi-level anomaly maps and uses an equal-weight average of their maximum response and the global text anomaly score for image-level detection (Sadikaj et al., 2025).

3.2 Prototype: Contextual Prototype Refinement

Different locations on the same product often contain similar local structures, a relationship that is difficult to exploit when patches are adapted independently. Prototype constructs $K = 32$ contextual prototypes at each layer. Given flattened patch features $X \in \mathbb{R}^{N\times d}$, the assignment network g applies depthwise 3×3 convolution, GELU, and 1×1 convolution in sequence to predict spatial assignment weights for the prototypes and aggregate tokens accordingly.

$$A = \mathrm{softmax}_N(g(X)), \qquad P = AX, \tag{1}$$

The spatial assignment matrix is $A \in \mathbb{R}^{K\times N}$, and the aggregated prototypes are $P \in \mathbb{R}^{K\times d}$. Both vary with the tokens of the current image rather than originating from stored normal images. Each token then combines the prototypes according to its correlations with them to obtain its contextual representation.

$$C = \text{softmax}_K(XP^\top/\sqrt{d})P. \tag{2}$$

Writing $T = \text{softmax}_K(XP^\top/\sqrt{d})$ gives $C = TAX$, with $\text{rank}(TA) \le K$ for fixed X. Context supplements rather than replaces X: global and local MLPs process its spatial mean and individual tokens, respectively, and their summed outputs form a sigmoid residual gate.

$$G = \sigma\big(h_g(\text{mean}_N\, C) + h_l(C)\big), \qquad R = \text{LN}(X + \beta G \odot W_o C). \tag{3}$$

Both MLPs use a reduction ratio of $r = 4$. The global term is broadcast over spatial positions, and the learnable coefficient β is initialized to 0.1. The refined features from the four layers are projected into the text space through their respective visual adapters to obtain F^ℓ. Their similarities to the original text anchors undergo upsampling and channel-wise softmax, after which the abnormal channels and the four layers are summed to form M. This layer sum differs from the original paper's layer mean by a fixed factor; subsequent within-category normalization removes that factor's effect on image-level fusion.

3.3 GVA: Global and Local Visual Discrimination

GVA supplies information beyond the maximum anomaly-map response for image-level discrimination. It applies residual adaptation separately to the normalized global feature z and each layer's features F^ℓ, enabling both global appearance and local responses to participate in image-level learning.

$$\mathcal{R}(u) = u + 0.10\, W_2\, \text{GELU}(W_1\, \text{LN}(u)). \tag{4}$$

The adapter bottleneck dimension is 192. The global branch maps the adapted features to a logit q_g through a linear head. The local branch first adapts tokens and then concatenates their spatial mean and maximum, retaining both overall response strength and salient local responses.

$$v_\ell = [\text{mean}_N\, \mathcal{R}_\ell(F^\ell); \max_N\, \mathcal{R}_\ell(F^\ell)], \quad v = \sum_\ell \alpha_\ell v_\ell, \quad \boldsymbol{\alpha} = \text{softmax}(\boldsymbol{a}). \tag{5}$$

The local statistics v pass through layer normalization, a linear layer with a hidden dimension of 256, GELU, and a single-output linear layer to produce q_l. Another set of learnable softmax weights combines the global and local logits, which are then converted into a visual anomaly score.

$$q_v = b_g q_g + b_l q_l, \quad (b_g, b_l) = \text{softmax}(\boldsymbol{b}), \quad s_v = \sigma(q_v). \tag{6}$$

3.4 IDAC: Bounded Dual Text-Anchor Calibration

Normal and abnormal descriptions are calibrated separately. IDAC learns residual vectors θ_n, θ_a for the two anchor types, applies bounded corrections to the normalized original anchors, and normalizes them again.

$$\delta_k = \frac{0.02}{\sqrt{d}} \tanh \theta_k, \quad k \in \{n, a\}; \qquad \hat{t}_n = \text{norm}(t_n + \delta_n), \quad \hat{t}_{a,j} = \text{norm}(t_{a,j} + \delta_a). \tag{7}$$

Abnormal anchors share δ_a, avoiding separate parameters for each defect type. The bound limits perturbation before renormalization; it does not enforce normal/abnormal separation. Similarities between normalized global image features and the calibrated anchors produce the text-based image score.

$$p = \text{softmax}\left(100\bar{z}^\top[\hat{t}_n, \hat{t}_{a,1}, \ldots, \hat{t}_{a,J}]\right), \qquad s_t = \sum_{j=1}^{J} p_{a,j}. \tag{8}$$

The evaluated configuration uses source-label weighted BCE, with identity-preservation, separation, and ranking loss weights all set to 0. Thus, measured IDAC gains concern bounded dual-residual adaptation, not these auxiliary objectives. The pixel branch retains the original anchors. Here calibration denotes anchor adjustment; AUROC/AP measure ranking, not probability calibration.

Table 1: Same-batch overall results on five target datasets and three seeds (%). Each row contains 15 evaluation records.

Method	AUROC-PX	AUPRO	AUROC-SP	AP-SP
MultiADS (Base)	92.019	80.602	85.270	86.014
Prototype+GVA+IDAC	95.148	85.945	89.756	90.763
Gain (pp)	+3.130	+5.342	+4.486	+4.749

3.5 Training and Image Scoring

Training proceeds in the order of localization, text calibration, and visual discrimination. The Prototype stage learns visual adaptation with equally weighted Dice/Focal losses; the visual adapters are then frozen to train IDAC, and all existing modules are finally frozen to train GVA. The pixel output already learned remains unchanged as later stages optimize image scoring.

Let $m(x) = \max M(x)$ denote the maximum anomaly-map response and $\mathcal{N}_c$ denote min–max normalization within category c. Image scoring first combines text evidence with the maximum map response and then fuses the independent visual branch. All weights follow the finalized configuration.

$$s_l(x) = 0.5 s_t(x) + 0.5 \mathcal{N}_c(m(x)), \tag{9}$$

$$\hat{s}(x) = 0.80 s_l(x) + 0.20 \mathcal{N}_c(s_v(x)). \tag{10}$$

The main protocol normalizes using extrema of unlabeled scores from test images of the same category, with a zero-output safeguard when the range is close to zero. It is therefore an offline batch-scoring procedure. To isolate the effect of test statistics, the fixed-scale control replaces $\mathcal{N}_c(m)$ and $\mathcal{N}_c(s_v)$ with $m/4$ and the raw s_v, respectively. Both scoring rules use the same checkpoint and retain the fusion weights; their comparison is reported in Appendix B.

4 Experiments

4.1 Experimental Setup

Data and implementation. MVTec AD supplies source supervision (Bergmann et al., 2021); targets are BTAD (Mishra et al., 2021), MPDD (Jezek et al., 2021), KSDD2 (Božič et al., 2021), VisA (Zou et al., 2022), and WFDD (Chen et al., 2024). Frozen CLIP ViT-L/14@336 uses 518×518 inputs and layers 6, 12, 18, and 24. Each stage runs 10 epochs; the same-batch evaluation reads epoch 8 with effective batch size 8, learning rate 10^{-3}, and seeds 42, 111, and 333. Defaults are $K = 32, r = 4$, lightweight Prototype, IDAC scale 0.02, and GVA weight 0.20. Targets supply no gradient training images, but historical target feedback informed design choices. The fixed rerun settings do not make these targets development-independent. Appendix H.1 specifies source splits, optimization, augmentation, and evaluation code.

Evaluation and statistics. Localization is measured by AUROC-PX and AUPRO, and image-level discrimination by AUROC-SP and AP-SP, with all values reported as percentages. Results are computed per dataset and then averaged equally across the five datasets and three seeds, preventing datasets with more images from dominating the mean. Seed stability is expressed by the sample standard deviation of the respective five-domain means. The four-metric average is used only to summarize paired contributions or sensitivity; individual metrics remain separately reported. Native configurations and seed coverage of external methods are specified separately.

4.2 Main Results and Per-Dataset Comparisons

The same-batch results in Table 1 show that localization refinement and independent image scoring can yield simultaneous gains. Across five target domains and three random seeds, AUROC-PX, AUPRO, AUROC-SP, and AP-SP improve over MultiADS by 3.130, 5.342, 4.486, and 4.749 percentage points, respectively.

These gains extend across all target domains rather than being driven by a single dataset. In Table 2 and Appendix Figure 8, the seed means of all four metrics exceed the corresponding Base results on

Table 2: Three-seed means on the five target datasets (%).

Dataset	Method	AUROC-PX	AUPRO	AUROC-SP	AP-SP
BTAD	Base	90.144	71.196	89.621	89.792
	Ours	94.932	80.662	93.988	97.259
MPDD	Base	94.322	84.722	73.208	76.419
	Ours	96.886	90.394	81.013	85.087
KSDD2	Base	96.198	94.049	96.126	89.411
	Ours	97.190	95.399	97.289	91.042
VisA	Base	94.694	87.245	78.088	83.375
	Ours	95.374	90.697	84.926	87.814
WFDD	Base	84.735	65.800	89.305	91.073
	Ours	91.358	72.571	91.563	92.615

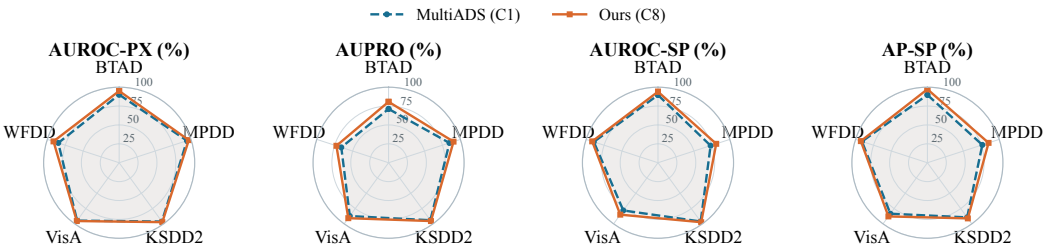


Figure 3: Radar plots of four metrics across five target domains. Each point is the mean of three same-batch seeds. All radial axes span 0–100%, without separate normalization for each model. Dashed lines with circles denote MultiADS (C1), and solid lines with squares denote the final model (C8). Polygon area is not used as an additional evaluation metric.

each of the five domains. AUPRO increases by 9.466 and 5.672 percentage points on BTAD and MPDD, respectively, while image AUROC increases by 6.838 points on VisA. KSDD2 and WFDD, where the baseline is already strong, also show gains.

4.3 Structural Ablation

Table 3: Eight same-batch ablations. The upper half reports absolute metrics (%), with best and tied-best values in bold; the lower half directly reports conditional increments from adding an individual module, in percentage points. P/G/I denote Prototype/GVA/IDAC. The background modules held fixed differ across pairs.

ID	Config.	P	G	I	AUROC-PX	AUPRO	AUROC-SP	AP-SP
C1	Base	–	–	–	92.019	80.602	85.270	86.014
C2	P	✓	–	–	**95.148**	**85.945**	88.436	89.318
C3	G	–	✓	–	92.019	80.602	86.792	87.947
C4	I	–	–	✓	92.019	80.602	85.739	86.499
C5	P+G	✓	✓	–	**95.148**	**85.945**	89.531	90.494
C6	P+I	✓	–	✓	**95.148**	**85.945**	88.835	89.744
C7	G+I	–	✓	✓	92.019	80.602	87.090	88.282
C8	P+G+I	✓	✓	✓	**95.148**	**85.945**	**89.756**	**90.763**
Module	Pair				ΔPX	ΔPRO	ΔSP	ΔAP
+P	C5–C3				+3.130	+5.342	+2.739	+2.547
+G	C7–C4				0.000	0.000	+1.351	+1.783
+I	C8–C5				0.000	0.000	+0.225	+0.269

The structural ablation examines the three modules through their $2^3 = 8$ combinations. Complete metrics and three conditional paired increments appear in the upper and lower halves of Table 3, respectively. C1–C8 use the same execution round of code and evaluation protocol, with three seeds and five target domains per configuration, yielding 120 records. The analysis uses only this con-

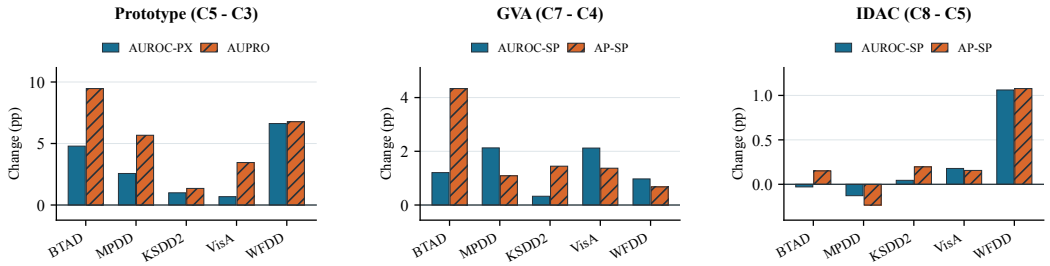


Figure 4: Conditional increments of the three modules, in percentage points: Prototype (C5–C3, holding GVA fixed), GVA (C7–C4, holding IDAC fixed), and IDAC (C8–C5, holding Prototype+GVA fixed). Pixel AUROC/AUPRO are shown for Prototype and image AUROC/AP for the other modules. All per-domain changes are retained, and vertical-axis scales differ across panels.

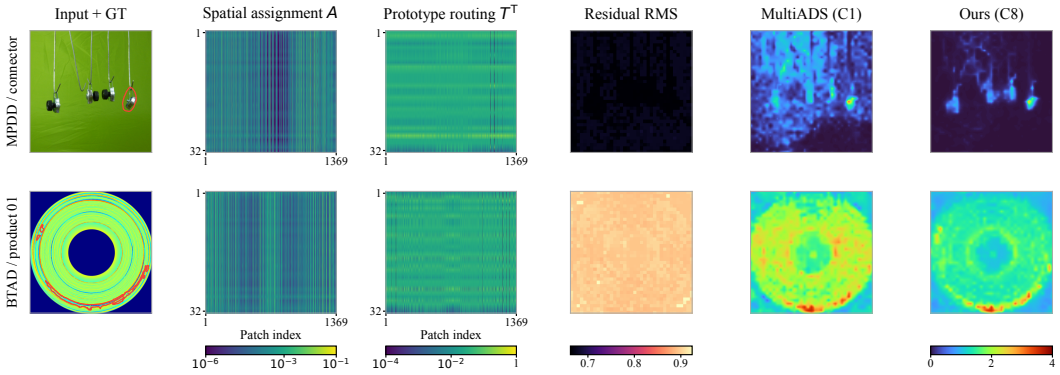


Figure 5: Prototype internals at layer 24 of the fixed C8 checkpoint (seed 42, epoch 8). Inputs with ground-truth contours, spatial assignments A, token-to-prototype weights, and residual RMS precede archived C1/C8 maps. Matrix columns retain raster-order patches; all 32 prototypes are shown. These internal weights are not defect probabilities.

trolled batch; historical mixed-batch summaries are archived separately. Execution identifiers, raw records, and the increment heatmap are available in Appendices A and I and Figure 9.

Effect of prototype context. With GVA fixed, C5 exceeds C3 by 3.130/5.342 points in AUROC-PX/AUPRO, while both image metrics improve. The four-metric mean gain is 3.439, positive for each seed and dataset. This establishes the benefit of the complete block; matched-capacity, alternative-context, and gate-removal controls are needed to isolate its internal mechanisms.

Figure 5 exposes the image-conditioned aggregation and residual path on two archived examples without retraining. Appendix H.3 defines the plotted quantities and records the extraction checks.

Effects of the visual and textual image branches. With pixel outputs held fixed, GVA improves AUROC-SP/AP-SP by 1.351/1.783 points (C7–C4). IDAC adds 0.225/0.269 points over Prototype+GVA (C8–C5), with positive five-domain means for all seeds. It is a modest refinement: gains vary by domain and are concentrated on WFDD. Appendix E gives signed domain results and leave-one-domain-out summaries, without changing the five-domain evaluation set.

4.4 Three-Seed Stability

All three seeds achieve positive aggregate gains. In Table 4, sample standard deviations of the complete model's four metrics are 0.296, 1.049, 0.061, and 0.154, respectively, with relatively small between-run variation in image-level performance. Per-seed results for the other combinations are listed in Appendix Table 20. These statistics describe consistency among the evaluated runs and do not constitute a formal significance test.

Table 4: Per-seed results for Base (C1) and the final model (C8), in percent. Each row is an equally weighted five-dataset mean. Sample standard deviations in the summary rows are computed over three seeds ($n = 3$), not over 15 evaluations.

Config.	Seed	AUROC-PX	AUPRO	AUROC-SP	AP-SP
C1	42	92.542	80.705	84.696	85.400
C1	111	91.296	80.291	86.570	86.842
C1	333	92.218	80.810	84.544	85.800
C1	Mean±s.d.	92.019±0.647	80.602±0.275	85.270±1.129	86.014±0.745
C8	42	95.199	86.064	89.823	90.599
C8	111	95.416	86.928	89.706	90.786
C8	333	94.830	84.841	89.738	90.905
C8	Mean±s.d.	95.148±0.296	85.945±1.049	89.756±0.061	90.763±0.154

4.5 Hyperparameter Sensitivity

The hyperparameter analysis distinguishes historical configuration grids from scoring changes at a fixed checkpoint. The earlier nine-configuration grid contains 135 evaluations, as reported in Appendix Table 25; the additional inference evaluates seven fusion weights under two scoring protocols, as detailed in Appendix B.3. At the same C8 checkpoint, the default $w = 0.20$ increases image AUROC/AP by 0.921/1.019 percentage points relative to scoring without GVA fusion. The post-hoc grid attains higher values at $w = 0.50$, but the main results retain the prespecified default weight.

4.6 Scoring Controls and Inference Cost

Table 5: Image-scoring controls at fixed checkpoints (%). The means and sample standard deviations are computed over the three seed-specific five-target means. Fixed-scale scoring does not use test-batch statistics; C8 retains $w = 0.20$.

Scoring	Model	AUROC-SP	AP-SP
Test-batch	C1	85.270±1.129	86.014±0.745
Test-batch	C8	**89.756±0.061**	**90.763±0.154**
Fixed-scale	C1	84.836±0.988	84.333±0.697
Fixed-scale	C8	**89.754±0.076**	**90.246±0.233**

Fixed-scale controls separate structural gains from the effects of test statistics. In Table 5, removing test-batch statistics yields image AUROC/AP of 89.754/90.246, improving on MultiADS under the same protocol by 4.917/5.912 percentage points; the main-protocol results are 89.756/90.763. Under both protocols, the three-seed means on all five target datasets retain an advantage over MultiADS. Appendix B reports the per-domain changes, while maximum, Top-1%, whole-map mean, and text-scoring controls further examine the role of learning a visual image score.

The additional inference cost accompanying these gains is also measured. On an A40, with 518×518 inputs and batch size 1, mean model and anomaly-map processing time is 58.63 ms, compared with 52.67 ms for MultiADS, an increase of 11.3%; peak allocated memory is 2547/2464 MiB. Timing includes model inference and anomaly-map processing but excludes loading, input transfer, prompt precomputation, and score export.

4.7 External Generalization Comparisons

Table 6 shows competitive pixel AUROC and higher aggregate AUPRO and image metrics among the listed runs. The controlled superiority claim concerns same-batch MultiADS; other rows retain native protocols. Matching source supervision, target-image access, test splits, and metric implementations is essential for stronger comparisons, beyond matching compute alone. AA-CLIP and AdaptCLIP are relevant untested comparators. Appendix G separates these evidence levels and audits the archived AdaCLIP AUPRO convention. The RareCLIP row covers only its offline variant.

Table 6: External generalization results on five target datasets (%). Our method and Base use new same-batch three-seed results; RareCLIP uses an existing offline summary from a different execution batch. Group B retains each method's native configuration and is not an equal-compute ranking.

Method	Seeds	AUROC-PX	AUPRO	AUROC-SP	AP-SP
A: three-seed offline summaries					
MultiADS (current batch)	3	92.019	80.602	85.270	86.014
RareCLIP (offline)	3	94.377	84.143	88.371	89.669
Prototype+GVA+IDAC(current batch)	3	95.148	85.945	89.756	90.763
B: offline evaluation with native configurations					
WinCLIP	1	77.160	48.000	78.280	78.120
APRIL-GAN	1	91.600	77.800	82.140	81.000
AnomalyCLIP	1	95.260	84.700	86.840	86.740
AdaCLIP (MVTec only)	1	96.226	58.638	83.774	86.748
MGVCLIP	3	81.567	52.717	78.338	76.129

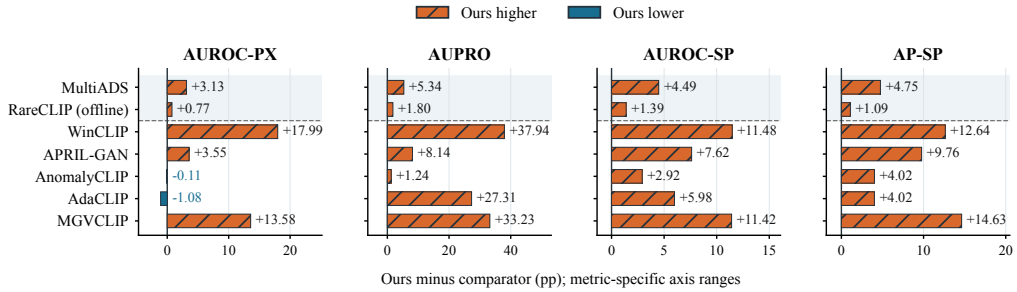


Figure 6: Differences between our method and each comparator in the five-target mean metrics, in percentage points, calculated from Table 6. Orange hatching indicates increases and blue indicates decreases. Horizontal-axis ranges differ by metric. Dashed lines separate offline summaries from the native-configuration group; the comparison is not an equal-compute ranking.

5 Scope and Evaluation Boundaries

The five-domain results establish consistent gains over the reproduced MultiADS under two scoring rules, including one without test statistics. They are transfer results on development-visible targets, not sealed confirmation: target feedback informed structural choices, and multiple seeds control training variation rather than this selection effect. Independent target validation and mechanism-replacement controls remain open. IDAC is evaluated as BCE-trained anchor adjustment, and external native-protocol comparisons support competitiveness rather than a protocol-matched SOTA claim. Deployment thresholds and probability reliability require separate validation.

6 Conclusion

In this work, we proposed Prototype–GVA–IDAC for cross-dataset industrial anomaly detection. Built on MultiADS with frozen CLIP encoders, the method refines patch features through within-image prototype context, aggregates global and local visual evidence for image decisions, and calibrates text anchors through bounded residuals. This design separates defect localization from image-level scoring while allowing the two tasks to share useful visual information. Experiments on five target datasets show improvements over MultiADS in both localization and detection under the same protocol. The eight-configuration ablation supports the complementary roles of the three modules, and fixed-scale scoring retains the overall image-level advantage without test-batch statistics. These results demonstrate the value of combining contextual patch refinement with independent visual image scoring for transfer beyond the source dataset.

AI Usage Statement

Generative AI tools were used to assist with and polish the writing, support information retrieval and related literature discovery, aid research execution, and contribute to drafting portions of the manuscript.

Ethics Statement

This study uses existing industrial anomaly-detection benchmarks and does not involve human-subject experiments or the collection of personal data. The method is intended to support quality inspection, not to replace safety certification. Before deployment, performance should be validated under the intended operating conditions, with appropriate human oversight to manage missed defects and false alarms.

Reproducibility Statement

Appendix H.1 specifies the source split, training settings, transformations, and metric implementation. The accompanying anonymous archive `PGISupplementEvidence.zip` provides 120 structural records, 30 score caches, all 1,800 scoring-control results, A40 timing records, and mechanism-figure data. It contains experimental evidence, not source code or executable scripts; full datasets and model weights are not included.

References

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 (WACV)*, pp. 128–138, January 2024.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021. doi: 10.1007/s11263-020-01400-4.

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 *Computer Vision – ECCV 2024*, pp. 55–72. Springer, 2024. doi: 10.1007/978-3-031-72761-0_4.

Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In *Computer Vision – ECCV 2024*, pp. 37–54. Springer, 2024. Introduces the WFDD dataset.

Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A^2-Nets: Double attention networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL `https://papers.nips.cc/paper_files/paper/2018/hash/e165421110ba03099a1c0393373c5b43-Abstract.html`.

Yahui Cheng, Guojun Wen, Aoshuang Luo, Shuang Mei, Hongbo Dong, and Xingyue Liu. An efficient and scale-aware zero-shot industrial anomaly detection technique based on optimized CLIP. *Measurement*, 253:117443, 2025. doi: 10.1016/j.measurement.2025.117443.

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*, volume 12664 of *Lecture Notes in Computer Science*, pp. 475–489. Springer, 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 (CVPR)*, pp. 9737–9746, June 2022.

Bin-Bin Gao, Yue Zhou, Jiangtao Yan, Yuezhi Cai, Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, and Chengjie Wang. AdaptCLIP: Adapting CLIP for universal visual anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4095–4103, 2026. doi: 10.1609/aaai.v40i6.42404.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. doi: 10.1007/s11263-023-01891-x.

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

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 98–107, January 2022.

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.

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, 2026. URL `https://openaccess.thecvf.com/content/CVPR2026/html/Hou_VisualAD_Language-Free_Zero-Shot_Anomaly_Detection_via_Vision_Transformer_CVPR_2026_paper.html`.

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

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

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.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122, June 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15190–15200, October 2023b.

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 (CVPR)*, pp. 9664–9674, June 2021.

Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9167–9176, 2019.

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16838–16848, June 2024.

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. 2980–2988, 2017.

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 (CVPR)*, pp. 20402–20411, June 2023.

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 (CVPR)*, pp. 4744–4754, June 2025.

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.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics*, pp. 1–6, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, 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. URL `https://openreview.net/forum?id=a68SUt6zFt`.

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.

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 (CVPR)*, pp. 14318–14328, June 2022.

Ylli Sadikaj, Hongkuan Zhou, Lavdim Halilaj, Stefan Schmid, Steffen Staab, and Claudia Plant. MultiADS: Defect-aware supervision for multi-type anomaly detection and segmentation in zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22978–22988, 2025.

Alireza Salehi, Ehsan Karami, Sepehr Noey, Sahand Noey, Makoto Yamada, Reshad Hosseini, and Mohammad Sabokrou. TIPS over tricks: Simple prompts for effective zero-shot anomaly detection. In *ICASSP 2026 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2911–2915. IEEE, 2026. doi: 10.1109/ICASSP55912.2026.11461758.

Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan T.M. Duong, Chanh D. Tr. Nguyen, and Steven Q. H. Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24511–24520, June 2023.

Manwen Yang, Leqian Ding, Yu Guo, and Fei Wang. Proximity-CLIP: Text-guided semantic proximity learning for zero-shot anomaly detection. In *Computer Vision – ECCV 2026*, volume 17052 of *Lecture Notes in Computer Science*, pp. 671–687. Springer, 2026. doi: 10.1007/978-3-032-37235-2_38.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem - a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8330–8339, October 2021.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *Computer Vision – ECCV 2022*, pp. 493–510, 2022. doi: 10.1007/978-3-031-19833-5_29.

Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3914–3923, June 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, June 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022b. doi: 10.1007/s11263-022-01653-1.

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

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

A Supplementary Material: Experimental Conventions and Result Sources

This appendix uses saved experimental records and traceable images. The eight structural combinations cover seeds 42, 111, and 333 and five target datasets, giving $8 \times 3 \times 5 = 120$ records. Paper identifiers correspond to execution identifiers as follows: C1/C00, C2/C01, C3/C03, C4/C04, C5/C06, C6/C07, C7/C10, and C8/C13. Renumbering does not change the records. The accompanying evidence archive retains all unrounded values in `evidence/structural/full8_raw_120.csv`; the historical server-side JSON files need not be available to recalculate the tables. Metrics are averaged over categories within each dataset, then equally over datasets and seeds.

Gradient training uses MVTec AD, while target datasets supply offline evaluation. The same-batch run reads epoch 8 without target-label gradient updates or threshold fitting; main scoring uses unlabeled test-score extrema. Earlier target feedback informed structural choices, so the five domains are development-visible transfer benchmarks. Appendix J separates this development stage from the fixed rerun and later scoring audit. The historical grid in Table 25 does not replace the current C8 checkpoint, default weights, or means.

A.1 Supervision, Target Information Access, and Comparison Boundaries

After supervised training on the source domain, this study directly evaluates target domains that provide no training images. Available information includes normal/abnormal source images and pixel annotations, target category names used in prompts, and unlabeled target-score statistics used by the main scoring rule. Table 7 organizes these conditions by information source to distinguish them from training-free inference and independent per-image prediction.

Table 7: Information access and comparison conventions in the current experiments. Unverified external conditions are not assumed to match.

Dimension	Current model and comparison boundary
Gradient supervision	MVTec AD image labels and segmentation masks; distinct from methods without source anomaly supervision.
Target training images	Not used to fit the model; target-normal memory methods access additional information.
Category semantics	Product and defect names may enter prompts; category information is not absent.
Test-batch statistics	Used in the primary scoring rule; the fixed-scale control in Appendix B removes this access.
Labels and development	Target metrics informed historical structural choices; these are development-visible transfer benchmarks, not sealed confirmation.
Internal comparisons	C1–C8 share the training execution batch, backbone, 518 inputs and three seeds. New fixed C1/C8 inference controls also measure A40 cost.
External comparisons	Native settings are retained; source supervision, target access, splits, and metric code are not fully matched.

Single-image versus batch scoring. Let $L = 4$ be the number of summed anomaly-map layers. The measured fixed-scale control uses m/L rather than directly averaging a layer sum with an approximate range of $[0, 4]$ and a text probability in $[0, 1]$. This rule retains the raw GVA sigmoid score without changing model parameters; the complete comparison with the original test-batch protocol is provided in Appendix B. Neither rule fits a source-held-out calibration, which would require a source split genuinely excluded from training.

Relation to recent mechanisms. VisualAD uses normal/abnormal tokens for anomaly recognition, whereas within-image prototypes reconstruct local context (Hou et al., 2026). Proximity-CLIP learns visually calibrated semantic margins (Yang et al., 2026); the evaluated IDAC configuration uses BCE-trained residual corrections, with identity and separation terms inactive. TIPS Over Tricks changes the pretrained backbone and is therefore discussed under its native protocol rather than included in same-protocol comparisons (Salehi et al., 2026). CoOp, CoCoOp, MaPLe, and AA-CLIP provide further context for prompt learning and anchor adaptation (Zhou et al., 2022b;a; Khattak et al., 2023a; Ma et al., 2025).

A.2 Reaggregation of Existing Evidence

Table 8: All 12 conditional paired increments, in percentage points. Means and sample standard deviations are derived from the five-domain averages of three seeds.

Module	Pair	AUROC-PX	AUPRO	AUROC-SP	AP-SP
P	C2-C1	+3.130±0.858	+5.342±1.303	+3.166±1.334	+3.304±0.912
P	C5-C3	+3.130±0.858	+5.342±1.303	+2.739±1.178	+2.547±1.122
P	C6-C4	+3.130±0.858	+5.342±1.303	+3.096±1.105	+3.245±0.761
P	C8-C7	+3.130±0.858	+5.342±1.303	+2.666±0.954	+2.481±0.898
G	C3-C1	+0.000±0.000	+0.000±0.000	+1.522±0.178	+1.933±0.499
G	C5-C2	+0.000±0.000	+0.000±0.000	+1.095±0.149	+1.176±0.202
G	C7-C4	+0.000±0.000	+0.000±0.000	+1.351±0.160	+1.783±0.466
G	C8-C6	+0.000±0.000	+0.000±0.000	+0.921±0.103	+1.019±0.216
I	C4-C1	+0.000±0.000	+0.000±0.000	+0.470±0.152	+0.485±0.149
I	C6-C2	+0.000±0.000	+0.000±0.000	+0.400±0.083	+0.427±0.031
I	C7-C3	+0.000±0.000	+0.000±0.000	+0.298±0.208	+0.335±0.193
I	C8-C5	+0.000±0.000	+0.000±0.000	+0.225±0.028	+0.269±0.037

The four paired backgrounds characterize how each module works with other active components. The statistical unit is each seed's five-domain mean, from which the three-seed mean and sample standard deviation are computed; the 15 results from five domains and three seeds are not treated

Table 9: The earlier nine-configuration grid (%). The final column identifies cases in which image-branch parameters are accompanied by drift in pixel metrics. These results do not replace the current C8.

Setting	AUROC-PX	AUPRO	AUROC-SP	AP-SP	Drift
default	95.185	86.052	89.740	90.785	–
K16	95.270	86.259	89.704	90.914	–
K64	95.229	86.317	89.628	90.821	–
r2	95.156	86.041	89.829	90.967	–
r8	95.277	86.219	89.797	90.967	–
scale0.01	95.186	86.071	89.579	90.612	Yes
scale0.03	95.186	86.071	89.789	90.882	Yes
gva_weight0.10	95.186	86.071	89.400	90.458	Yes
gva_weight0.30	95.186	86.071	90.006	91.062	Yes

as independent replicates. The old grid also contains pixel-metric drift accompanying changes in image-branch settings, so it is presented as a historical sensitivity record rather than as a parameter effect at a fixed checkpoint. Both positive and negative differences are retained.

B Fixed-Checkpoint Scoring Controls and A40 Efficiency

B.1 Measurement Setup and Reproduction Checks

Supplementary inference uses existing checkpoints: epoch 8 parameters of C1 and C8 are evaluated with seeds 42, 111, and 333 on five target datasets, yielding $2 \times 3 \times 5 = 30$ tasks. BTAD, MPDD, KSDD2, VisA, and WFDD contain 741, 458, 1,004, 2,162, and 444 test images, respectively, totaling 4,809 images and 28,854 model–image evaluations. All C8 components and the existing text calibration remain fixed throughout, with no additional training, checkpoint selection, IDAC ablations, or default-parameter adjustment.

For each image, the inference cache retains the text score, anomaly-map maximum, Top-1% mean, whole-map mean, and C8's raw GVA sigmoid score. Checkpoint and code hashes are checked against the original archive, and image paths, categories, and labels are matched individually. Recalculation exactly reproduces the main protocol's final per-image scores, with image-metric discrepancies at floating-point rounding scale. All 1,800 scoring controls and their summaries are independently verified again locally. Labels are used only for AUROC/AP evaluation, not for score transformations or random batch partitions.

All scoring rules use the same aggregation procedure: average over categories within each dataset, weight the five target datasets equally, and then compute the three-seed mean and sample standard deviation. The 1,800 controls arise from rescoring complete prediction caches, not from independently trained models. Anomaly maps remain fixed during these operations, so this section evaluates image-level metrics only.

B.2 Scoring Without Test-Batch Statistics

The fixed-scale control replaces $\mathcal{N}_c(m)$ with $m/4$ and $\mathcal{N}_c(s_v)$ with s_v in Equations (9)–(10). Dividing the four-layer summed map by 4 follows the known architectural scale, not a target-fitted estimate. C8 retains $w = 0.20$, while C1 uses the corresponding equal-weight text/map scoring without GVA. Neither performs test-time updates or label-based calibration.

Without test-batch statistics, C8 still achieves image AUROC/AP of 89.754/90.246, exceeding C1 by 4.917/5.912 percentage points. In Table 10, the three-seed means on all five target datasets retain a positive C8 advantage over C1. Relative to the original main protocol, C8's five-domain means change by $-0.002/-0.518$ points. This control shows that the overall detection advantage does not require test-batch normalization. Its scope is dataset-averaged performance; per-domain sensitivity, individual-seed behavior, deployment thresholds, and cross-category streams should still be evaluated separately.

Table 10: Complete per-target scoring controls (%), reported as three-seed means. Each metric pair lists image AUROC/AP. Both protocols use the same checkpoints and test images.

Scoring	Target	C1 AUROC	C1 AP	C8 AUROC	C8 AP
Test-batch	BTAD	89.621	89.792	93.988	97.259
Test-batch	MPDD	73.208	76.419	81.013	85.087
Test-batch	KSDD2	96.126	89.411	97.289	91.042
Test-batch	VisA	78.088	83.375	84.926	87.814
Test-batch	WFDD	89.305	91.073	91.563	92.615
Fixed-scale	BTAD	83.727	79.257	93.875	96.874
Fixed-scale	MPDD	71.452	74.409	77.819	80.065
Fixed-scale	KSDD2	95.968	89.172	97.292	91.135
Fixed-scale	VisA	78.396	83.277	84.526	86.979
Fixed-scale	WFDD	94.639	95.552	95.256	96.174

Statistical batch size. Within each category, unlabeled test images are randomly partitioned into batches of size 1, 8, 32, or 128, and min–max normalization is applied separately to each batch. The final undersized batch is retained. Ten deterministic random permutations are independent of labels and scores, and each model retains its original fusion weights. Table 11 first averages random partitions within each seed and then reports variation across three seeds. Larger statistical batches approach the original offline mean; at size 1, both normalized image branches become zero, reducing the ranking to text scoring. This is a degenerate case of the original min–max rule, not the performance of the fixed-scale per-image scheme. The GPU forward batch size remains 1 throughout.

Table 11: Sensitivity to the statistical batch size used for scoring (%). Each seed first averages over five target datasets and ten within-category random partitions, followed by the three-seed mean and standard deviation. The GPU forward batch size remains 1.

Batch size	C1 AUROC	C1 AP	C8 AUROC	C8 AP
1	77.482±0.000	77.345±0.000	77.933±0.046	77.414±0.053
8	82.766±1.410	80.367±1.002	87.928±0.031	86.129±0.262
32	84.527±1.255	84.117±0.960	89.303±0.135	89.638±0.044
128	85.142±1.192	85.619±0.929	89.693±0.078	90.608±0.115

B.3 Fusion Sensitivity at a Fixed Checkpoint

Fusion sensitivity is measured at the same C8 checkpoint, testing $w \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 1\}$ under both scoring protocols. Table 12 uses directly recorded branch scores, preserving their original ties and the measured GVA-only endpoint instead of recovering branches algebraically from final mixed scores. Default scoring results, the backbone, and IDAC remain unchanged. Under the main protocol, default fusion at the same C8 checkpoint increases AUROC/AP by 0.921/1.019 percentage points relative to $w = 0$. Within the grid, $w = 0.50$ reaches 90.444/91.396, exceeding the default by 0.688/0.632 points; the GVA-only endpoint is 89.394/90.118 and shows greater seed variation. These results support combining the two types of evidence rather than continuously increasing the weight of one branch. Because the grid is a post-hoc analysis on target test sets, the main model retains the prespecified $w = 0.20$.

B.4 Simple Image-Scoring Controls

Table 13 uses simple scoring rules to test whether independent visual discrimination provides additional information. Controls include text-only scoring, three pure anomaly-map summaries, and equal-weight combinations of each summary with text. Top-1% denotes the average of the highest-scoring one percent of pixels. The simple rules do not incorporate GVA; final scoring follows each model's original fusion settings. Underlying model components remain fixed, so differences arise from scoring rather than retraining.

Table 12: C8 fusion sensitivity using directly measured branch scores (%). Means and standard deviations are computed from the five-target means of three seeds. Bold marks the unchanged default $w = 0.20$, not the post-hoc optimum; $w = 1$ uses GVA alone.

Scoring	Weight	AUROC-SP	AP-SP
Test-batch	0.00	88.835±0.164	89.744±0.147
Test-batch	0.10	89.447±0.121	90.458±0.176
Test-batch	**0.20**	**89.756±0.061**	**90.763±0.154**
Test-batch	0.30	90.041±0.037	91.003±0.158
Test-batch	0.40	90.253±0.097	91.230±0.190
Test-batch	0.50	90.444±0.118	91.396±0.267
Test-batch	1.00	89.394±1.244	90.118±1.359
Fixed-scale	0.00	87.787±1.478	87.106±1.771
Fixed-scale	0.10	89.440±0.180	89.914±0.371
Fixed-scale	**0.20**	**89.754±0.076**	**90.246±0.233**
Fixed-scale	0.30	89.909±0.065	90.373±0.205
Fixed-scale	0.40	90.053±0.088	90.494±0.199
Fixed-scale	0.50	90.146±0.132	90.604±0.189
Fixed-scale	1.00	89.394±1.244	90.118±1.359

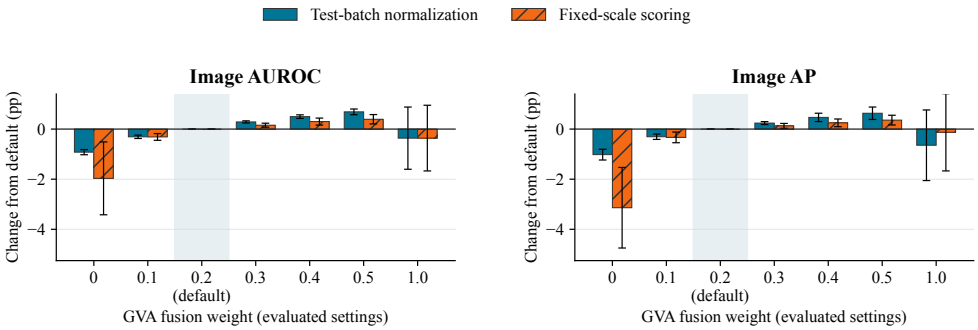


Figure 7: C8 fusion sensitivity relative to $w = 0.20$ under each scoring protocol, in percentage points. Bar heights are the means of paired changes in each seed's five-target mean relative to its own default, with sample standard deviations as error bars. Both protocols use the same checkpoints; shading marks the unchanged default. All seven settings and both positive and negative changes are retained. Absolute values are provided in Table 12.

Among the scoring rules actually tested, C8's final score achieves the highest five-domain mean image AUROC and AP under both protocols. Main-protocol results are 89.756/90.763, compared with 88.835/89.744 for the combination of text and the anomaly-map maximum, indicating that the learned visual score provides information beyond a single spatial summary. This comparison is limited to the implemented rules in the table and is not extended to external structured scoring methods that were not run.

B.5 Measured Inference Cost

The hardware is an NVIDIA A40, the software is PyTorch 2.0.1+cu118, the input is 518×518, and the forward batch size is 1. CUDA-synchronized timing covers image encoding, adapters, text scoring, and anomaly-map computation, including transfer of the anomaly map to the CPU. It excludes data loading, input transfer to the GPU, prompt precomputation, CPU score-export statistics, and disk writes. The first 50 images of each task are warm-up, leaving 13,677 timed evaluations per model. Latency statistics pool the actual post-warm-up samples rather than weighting differently sized datasets equally.

Relative to the same-backbone baseline, C8 obtains the localization and detection gains reported in the main text while adding 5.961 ms per image (11.32%), 82.81 MiB to the maximum allocated

Table 13: Image-scoring rule controls (%), averaged over five target datasets and three seeds. Simple scoring does not fuse GVA. The established C1/C8 scoring weights are 0/0.20, respectively, while all other model components remain unchanged.

Scoring	Rule	C1 AUROC	C1 AP	C8 AUROC	C8 AP
Test-batch	Text only	77.482	77.345	77.933	77.414
Test-batch	Map max	82.927	84.662	86.897	88.545
Test-batch	Map top-1%	81.582	84.489	86.375	88.480
Test-batch	Map mean	60.438	64.672	65.627	69.096
Test-batch	Text + max	85.270	86.014	88.835	89.744
Test-batch	Text + top-1%	84.174	86.146	87.802	89.412
Test-batch	Text + mean	69.052	71.889	75.501	77.416
Test-batch	Default score	85.270	86.014	89.756	90.763
Fixed-scale	Text only	77.482	77.345	77.933	77.414
Fixed-scale	Map max	82.927	84.662	86.897	88.545
Fixed-scale	Map top-1%	81.582	84.489	86.375	88.480
Fixed-scale	Map mean	60.438	64.672	65.627	69.096
Fixed-scale	Text + max	84.836	84.333	87.787	87.106
Fixed-scale	Text + top-1%	84.542	84.758	87.160	87.370
Fixed-scale	Text + mean	77.872	77.387	79.803	78.621
Fixed-scale	Default score	84.836	84.333	89.754	90.246

Table 14: Measured cost on an NVIDIA A40 with 518×518 inputs and GPU batch size 1. Timing pools all post-warm-up images; memory is the maximum allocated peak across each model's 15 tasks; parameters include frozen CLIP.

Measurement	C1	C8
Mean latency (ms)	52.669	58.630
Median latency (ms)	52.652	58.578
P95 latency (ms)	53.053	59.232
Peak allocated (MiB)	2464.36	2547.17
Loaded parameters (M)	431.905	442.379
Timed images	13,677	13,677

peak (3.36%), and 10.474 M loaded parameters. Parameter counts include frozen CLIP and must not be interpreted as trainable-parameter counts. Timing describes complete-model latency within the specified processing scope, not standalone module costs, batch throughput, FLOPs, or deployment end-to-end cost. This permits the additional resources required for the performance gains to be examined under a common measurement convention.

C Complete Results of the Eight Structural Ablations

Tables 15–19 report the three-seed means of the eight combinations on each target dataset; Table 20 lists the per-seed five-dataset means of all eight combinations. These tables retain each original metric rather than replacing localization and image-level results with a four-metric average. Figure 9 presents changes in the four metrics of all eight ablations relative to C1. GVA and IDAC do not participate in pixel-map computation; combinations with the same pixel path therefore share AUROC-PX and AUPRO. Pixel differences across different backgrounds in the table should not be attributed to the image-level branches.

D Category-Level Performance and Cross-Domain Differences

Category-level results examine the distribution of dataset-averaged gains. Following the organization of fine-grained results in AnomalyCLIP (Zhou et al., 2024), we extract C1/C8 metrics for all 26 categories from same-batch records, first averaging three seeds within each category and then computing method differences. This analysis is presented separately and does not alter the weighting of the five-domain mean in the main text.

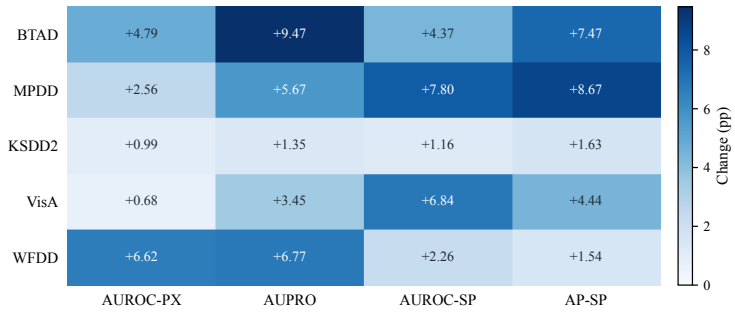


Figure 8: Four-metric increments of C8 over same-batch C1 on the five target domains, in percentage points. Color intensity and cell values jointly encode increments, computed from the per-domain three-seed means.

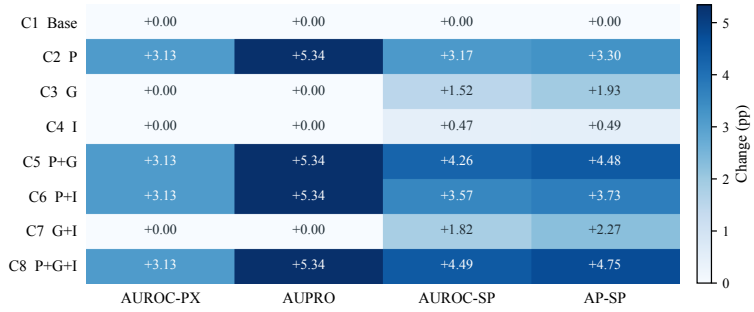


Figure 9: Metric changes of the eight same-batch ablations relative to C1, in percentage points. Color intensity and cell values jointly encode increments for verification in black-and-white printing.

Table 15: Three-seed means of the eight combinations on BTAD (%).

ID	Config.	AUROC-PX	AUPRO	AUROC-SP	AP-SP
C1	Base	90.144	71.196	89.621	89.792
C2	P	94.932	80.662	93.470	95.323
C3	G	90.144	71.196	90.974	94.398
C4	I	90.144	71.196	89.782	90.330
C5	P+G	94.932	80.662	94.017	97.108
C6	P+I	94.932	80.662	93.501	95.689
C7	G+I	90.144	71.196	90.989	94.665
C8	P+G+I	94.932	80.662	93.988	97.259

Table 16: Three-seed means of the eight combinations on MPDD (%).

ID	Config.	AUROC-PX	AUPRO	AUROC-SP	AP-SP
C1	Base	94.322	84.722	73.208	76.419
C2	P	96.886	90.394	79.997	84.424
C3	G	94.322	84.722	75.398	77.488
C4	I	94.322	84.722	73.370	76.482
C5	P+G	96.886	90.394	81.141	85.320
C6	P+I	96.886	90.394	80.109	84.285
C7	G+I	94.322	84.722	75.497	77.570
C8	P+G+I	96.886	90.394	81.013	85.087

Table 17: Three-seed means of the eight combinations on KSDD2 (%).

ID	Config.	AUROC-PX	AUPRO	AUROC-SP	AP-SP
C1	Base	96.198	94.049	96.126	89.411
C2	P	97.190	95.399	96.359	89.590
C3	G	96.198	94.049	96.494	91.028
C4	I	96.198	94.049	96.226	89.742
C5	P+G	97.190	95.399	97.245	90.846
C6	P+I	97.190	95.399	96.449	89.915
C7	G+I	96.198	94.049	96.552	91.188
C8	P+G+I	97.190	95.399	97.289	91.042

Table 18: Three-seed means of the eight combinations on VisA (%).

ID	Config.	AUROC-PX	AUPRO	AUROC-SP	AP-SP
C1	Base	94.694	87.245	78.088	83.375
C2	P	95.374	90.697	82.612	86.316
C3	G	94.694	87.245	80.444	84.921
C4	I	94.694	87.245	78.795	83.933
C5	P+G	95.374	90.697	84.748	87.658
C6	P+I	95.374	90.697	82.952	86.582
C7	G+I	94.694	87.245	80.916	85.304
C8	P+G+I	95.374	90.697	84.926	87.814

Table 19: Three-seed means of the eight combinations on WFDD (%).

ID	Config.	AUROC-PX	AUPRO	AUROC-SP	AP-SP
C1	Base	84.735	65.800	89.305	91.073
C2	P	91.358	72.571	89.741	90.936
C3	G	84.735	65.800	90.647	91.897
C4	I	84.735	65.800	90.523	92.009
C5	P+G	91.358	72.571	90.502	91.538
C6	P+I	91.358	72.571	91.165	92.250
C7	G+I	84.735	65.800	91.495	92.683
C8	P+G+I	91.358	72.571	91.563	92.615

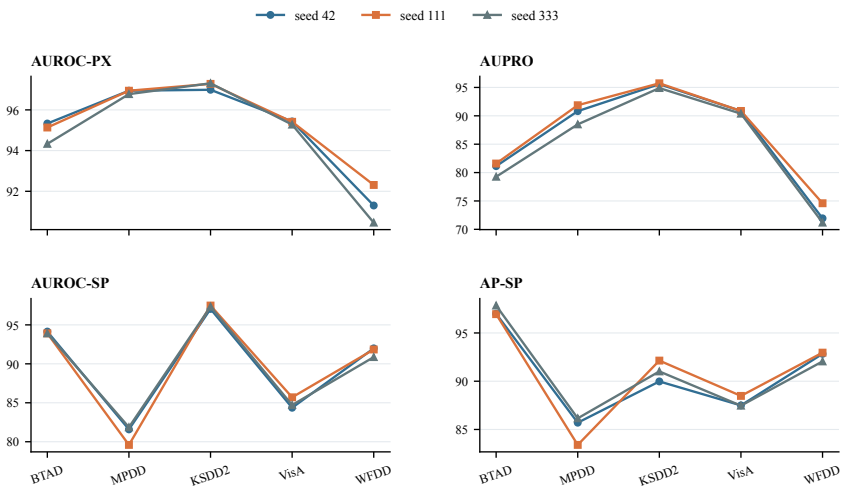


Figure 10: Per-metric three-seed results of C8 on the five target datasets. Each line connects values for the same seed across datasets solely to display target-domain and seed variation; connecting lines do not represent a continuous data domain.

Table 20: Per-seed five-dataset means of the eight combinations (%). Each value equally averages the five datasets.

ID	Config.	Seed	AUROC-PX	AUPRO	AUROC-SP	AP-SP
C1	Base	42	92.542	80.705	84.696	85.400
C1		111	91.296	80.291	86.570	86.842
C1		333	92.218	80.810	84.544	85.800
C2	P	42	95.199	86.064	88.659	89.402
C2		111	95.416	86.928	88.196	89.114
C2		333	94.830	84.841	88.452	89.437
C3	G	42	92.542	80.705	86.410	87.489
C3		111	91.296	80.291	88.056	89.176
C3		333	92.218	80.810	85.908	87.174
C4	I	42	92.542	80.705	85.213	85.914
C4		111	91.296	80.291	86.869	87.166
C4		333	92.218	80.810	85.136	86.418
C5	P+G	42	95.199	86.064	89.626	90.362
C5		111	95.416	86.928	89.454	90.475
C5		333	94.830	84.841	89.512	90.645
C6	P+I	42	95.199	86.064	89.014	89.814
C6		111	95.416	86.928	88.691	89.576
C6		333	94.830	84.841	88.801	89.843
C7	G+I	42	92.542	80.705	86.748	87.877
C7		111	91.296	80.291	88.130	89.298
C7		333	92.218	80.810	86.392	87.671
C8	P+G+I	42	95.199	86.064	89.823	90.599
C8		111	95.416	86.928	89.706	90.786
C8		333	94.830	84.841	89.738	90.905

In Figure 11, AUROC-PX, AUPRO, AUROC-SP, and AP-SP improve in 23, 24, 24, and 23 categories, respectively, indicating gains across most categories. Tables 21 and 22 report baseline values, final values, and signed differences together, retaining the different changes in the remaining categories. Each metric here aggregates all test samples in a category and therefore has a different statistical scope from the localization gallery selected by per-image quality.

Localization and image decisions are different problems. Pixel localization and image discrimination involve different ranking problems: the former compares spatial positions, while the latter compares scalar scores between normal and anomalous images. Background responses, defect area, and aggregation all affect image decisions, so improvements in local ranking cannot substitute for image-level evaluation. We report both metric types separately, neither ranking categories by the four-metric average nor judging text calibration from a single heatmap.

The average increments of the three conditional pairs are 3.439, 0.784, and 0.124. Because the background modules held fixed differ, these values should be interpreted within their respective pairs, not added as independent effects against a shared baseline. With Prototype fixed, both GVA and IDAC preserve pixel metrics, consistent with their image-scoring-only implementation. Complete eight-configuration results are provided in Figure 9 and Appendix C.

E Conditional Pairs and Hyperparameter Analysis

Table 23 computes three conditional pairs for each domain: C5–C3 for Prototype, C7–C4 for GVA, and C8–C5 for IDAC. Each pair has a positive overall mean; its background modules differ, so these are conditional increments rather than independent main effects. IDAC adds 0.225/0.269 points in image AUROC/AP over Prototype+GVA. Table 24 reports every leave-one-domain-out mean to describe the domain distribution of this increment without selecting a new primary evaluation set. All five-domain results remain the main estimates.

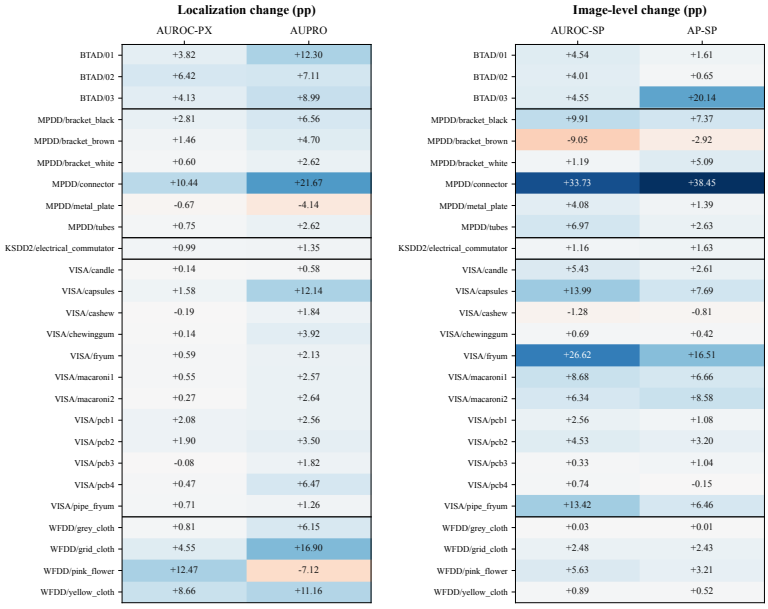


Figure 11: Three-seed mean C8–C1 increments for 26 categories, in percentage points. Left and right panels show pixel and image metrics, respectively. Horizontal lines separate datasets; zero-centered colors and cell values preserve both positive and negative changes.

Table 21: Per-category pixel-level results. Each value is a three-seed mean (%), and the difference is C8 minus C1 in percentage points. All categories and negative changes are retained.

Dataset / Class	AUROC-PX			AUPRO		
	C1	C8	Δ	C1	C8	Δ
BTAD/01	90.552	94.370	+3.818	68.731	81.027	+12.296
BTAD/02	88.091	94.508	+6.418	62.283	69.392	+7.109
BTAD/03	91.789	95.917	+4.128	82.575	91.566	+8.991
MPDD/bracket_black	95.667	98.472	+2.805	89.812	96.372	+6.561
MPDD/bracket_brown	92.674	94.135	+1.461	82.540	87.235	+4.696
MPDD/bracket_white	98.829	99.433	+0.604	94.968	97.591	+2.623
MPDD/connector	84.488	94.926	+10.438	59.807	81.478	+21.671
MPDD/metal_plate	95.817	95.145	-0.672	86.771	82.634	-4.137
MPDD/tubes	98.458	99.205	+0.747	94.434	97.051	+2.617
KSDD2/electrical_commutator	96.198	97.190	+0.993	94.049	95.399	+1.350
VISA/candle	98.603	98.739	+0.136	95.047	95.629	+0.581
VISA/capsules	96.110	97.694	+1.584	78.110	90.249	+12.138
VISA/cashew	92.660	92.475	-0.185	94.723	96.567	+1.843
VISA/chewinggum	99.327	99.467	+0.140	88.400	92.324	+3.924
VISA/fryum	93.628	94.221	+0.593	90.157	92.286	+2.129
VISA/macaroni1	98.898	99.446	+0.549	94.106	96.672	+2.566
VISA/macaroni2	98.653	98.921	+0.267	88.829	91.467	+2.638
VISA/pcb1	88.598	90.680	+2.081	86.988	89.551	+2.562
VISA/pcb2	89.314	91.211	+1.897	74.436	77.937	+3.501
VISA/pcb3	90.503	90.423	-0.080	79.514	81.334	+1.820
VISA/pcb4	93.960	94.426	+0.466	80.557	87.027	+6.470
VISA/pipe_fryum	96.073	96.786	+0.713	96.068	97.324	+1.256
WFDD/grey_cloth	97.686	98.496	+0.810	80.158	86.304	+6.146
WFDD/grid_cloth	87.907	92.456	+4.549	58.291	75.191	+16.900
WFDD/pink_flower	70.015	82.486	+12.472	49.915	42.797	-7.118
WFDD/yellow_cloth	83.332	91.993	+8.662	74.836	85.994	+11.157

Table 22: Per-category image-level results. Each value is a three-seed mean (%), and the difference is C8 minus C1 in percentage points. All categories and negative changes are retained.

Dataset / Class	AUROC-SP			AP-SP		
	C1	C8	Δ	C1	C8	Δ
BTAD/01	93.716	98.251	+4.535	97.687	99.297	+1.610
BTAD/02	80.578	84.589	+4.011	96.782	97.435	+0.653
BTAD/03	94.569	99.124	+4.555	74.907	95.044	+20.137
MPDD/bracket_black	58.511	68.418	+9.907	69.262	76.630	+7.368
MPDD/bracket_brown	63.650	54.600	-9.050	76.795	73.878	-2.917
MPDD/bracket_white	70.815	72.000	+1.185	65.838	70.925	+5.086
MPDD/connector	61.190	94.921	+33.730	52.363	90.809	+38.447
MPDD/metal_plate	95.486	99.567	+4.081	98.454	99.848	+1.394
MPDD/tubes	89.598	96.573	+6.975	95.799	98.430	+2.631
KSDD2/electrical_commutator	96.126	97.289	+1.164	89.411	91.042	+1.631
VISA/candle	84.597	90.030	+5.433	88.686	91.291	+2.605
VISA/capsules	78.683	92.672	+13.989	87.693	95.379	+7.685
VISA/cashew	89.620	88.340	-1.280	95.365	94.550	-0.815
VISA/chewinggum	96.267	96.953	+0.687	98.363	98.781	+0.418
VISA/fryum	56.793	83.413	+26.620	75.374	91.886	+16.512
VISA/macaroni1	81.317	89.993	+8.677	85.365	92.026	+6.661
VISA/macaroni2	63.050	69.393	+6.343	63.686	72.264	+8.578
VISA/pcb1	73.027	75.590	+2.563	78.070	79.146	+1.076
VISA/pcb2	68.657	73.190	+4.533	70.915	74.113	+3.198
VISA/pcb3	72.762	73.089	+0.327	74.617	75.656	+1.040
VISA/pcb4	88.941	89.680	+0.739	90.477	90.323	-0.153
VISA/pipe_fryum	83.347	96.767	+13.420	91.886	98.351	+6.464
WFDD/grey_cloth	99.785	99.812	+0.027	99.923	99.933	+0.010
WFDD/grid_cloth	87.350	89.829	+2.479	88.987	91.415	+2.428
WFDD/pink_flower	72.949	78.581	+5.632	77.690	80.902	+3.213
WFDD/yellow_cloth	97.138	98.031	+0.894	97.693	98.211	+0.518

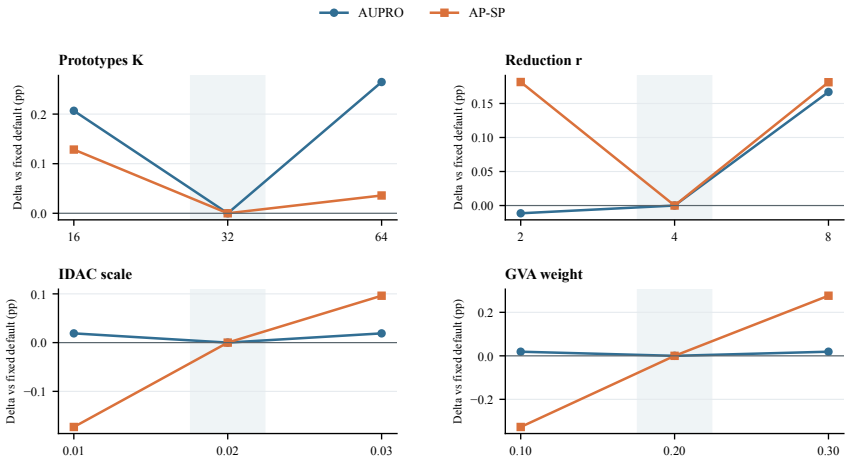


Figure 12: AUPRO and AP-SP differences from the prespecified defaults in four local hyperparameter grids. Shaded regions mark default values. This grid was executed separately from the eight same-batch ablations, and higher post-hoc means are not used to change the model.

Table 23: Mean increments of the three conditional pairs on each target dataset, in percentage points. Negative values are not omitted.

Module	Dataset	ΔPX	ΔPRO	ΔSP	ΔAP	Mean
Prototype	BTAD	+4.788	+9.466	+3.043	+2.710	+5.002
Prototype	MPDD	+2.564	+5.672	+5.743	+7.832	+5.453
Prototype	KSDD2	+0.993	+1.350	+0.750	-0.182	+0.728
Prototype	VisA	+0.680	+3.452	+4.303	+2.736	+2.793
Prototype	WFDD	+6.623	+6.771	-0.145	-0.359	+3.223
GVA	BTAD	+0.000	+0.000	+1.207	+4.335	+1.385
GVA	MPDD	+0.000	+0.000	+2.127	+1.088	+0.804
GVA	KSDD2	+0.000	+0.000	+0.326	+1.446	+0.443
GVA	VisA	+0.000	+0.000	+2.121	+1.371	+0.873
GVA	WFDD	+0.000	+0.000	+0.972	+0.674	+0.412
IDAC	BTAD	+0.000	+0.000	-0.029	+0.150	+0.030
IDAC	MPDD	+0.000	+0.000	-0.128	-0.233	-0.090
IDAC	KSDD2	+0.000	+0.000	+0.045	+0.196	+0.060
IDAC	VisA	+0.000	+0.000	+0.178	+0.156	+0.084
IDAC	WFDD	+0.000	+0.000	+1.061	+1.077	+0.535

Table 24: Domain sensitivity of the existing C8–C5 image-metric increment (pp). Each row averages the retained domains within a seed and then reports the three-seed mean ± sample SD. These are reaggregations, not new training or IDAC experiments.

Domains summarized	ΔAUROC-SP	ΔAP-SP
All five (primary)	$+0.2254 \pm 0.0277$	$+0.2693 \pm 0.0373$
All except BTAD	$+0.2891 \pm 0.0316$	$+0.2991 \pm 0.0307$
All except MPDD	$+0.3138 \pm 0.0092$	$+0.3949 \pm 0.0021$
All except KSDD2	$+0.2705 \pm 0.0403$	$+0.2876 \pm 0.0577$
All except VisA	$+0.2371 \pm 0.0574$	$+0.2976 \pm 0.0663$
All except WFDD	$+0.0164 \pm 0.0115$	$+0.0674 \pm 0.0385$

Table 25: Historical hyperparameter grid (%). Each row averages three seeds and five datasets from an earlier execution. Default refers to that grid's reference run, not current C8 in Table 1. These are separate trained runs, not fixed-checkpoint effects; the main configuration is not replaced by the best post-hoc row.

Setting	AUROC-PX	AUPRO	AUROC-SP	AP-SP	Mean
Default	95.185	86.052	89.740	90.785	90.441
$K = 16$	95.270	86.259	89.704	90.914	90.537
$K = 64$	95.229	86.317	89.628	90.821	90.499
$r = 2$	95.156	86.041	89.829	90.967	90.498
$r = 8$	95.277	86.219	89.797	90.967	90.565
$s = 0.01$	95.186	86.071	89.579	90.612	90.362
$s = 0.03$	95.186	86.071	89.789	90.882	90.482
$w = 0.10$	95.186	86.071	89.400	90.458	90.279
$w = 0.30$	95.186	86.071	90.006	91.062	90.581

F Actual Qualitative Results and Failure Cases

Qualitative analysis addresses three directly observable questions: whether hotspots correspond to ground truth, whether normal regions produce false responses, and whether defect contours are distinguishable. The four-column localization comparisons at the end of the appendix are selected by these criteria, without requiring positive per-image metric increments. Table 26 retains differences for both improving and degrading samples, while Figure 13 gives the complete distribution of the fixed candidate set. All 260 paired originals are archived, heatmaps come from original outputs, and the selected cases are not used to estimate overall error.

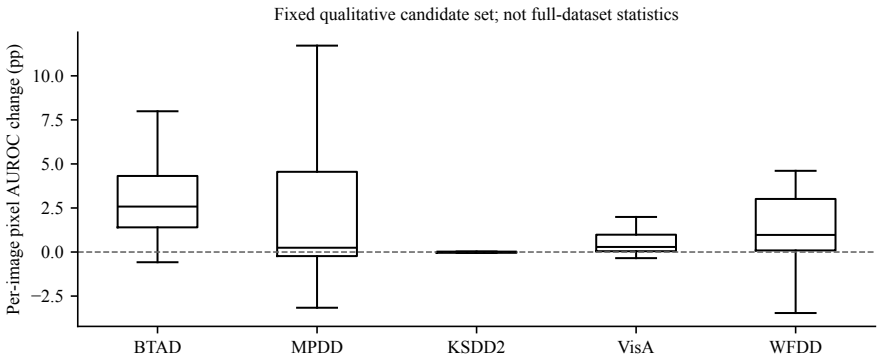


Figure 13: Distribution of per-image pixel AUROC changes within the fixed candidate set. This figure does not replace the full-dataset metrics in Table 2.

Table 26: Per-image pixel AUROC differences for visualization examples, in percentage points. They are used only for case presentation and do not participate in model selection.

Dataset	Class	Case	ΔAUROC-PX
BTAD	01	Improved	+7.990
BTAD	02	Degraded	-0.579
MPDD	connector	Improved	+16.360
MPDD	metal_plate	Degraded	-3.162
KSDD2	electrical_commutator	Improved	+0.029
KSDD2	electrical_commutator	Degraded	-0.486
VisA	capsules	Improved	+10.306
VisA	candle	Degraded	-0.347
WFDD	grid_cloth	Improved	+21.856
WFDD	pink_flower	Degraded	-3.462

G External Results and Structural Details

Table 27 lists the available per-domain records for four external methods together with current C8. Its C8 rows now use exactly the same three-seed means as Table 2, replacing the earlier-batch values previously used in this table. No external measurement is changed. Table 6 additionally contains aggregate MGVCLIP and offline RareCLIP results; their per-domain rows are not reconstructed from these aggregates. The online RareCLIP variant is outside the comparison. WinCLIP, APRIL-GAN, AnomalyCLIP, and AdaCLIP all use seed 42. APRIL-GAN and AnomalyCLIP use author-provided MVTec weights, while AdaCLIP is trained only on MVTec for five epochs. WinCLIP uses ViT-B/16+240; MGVCLIP retains its native 240 resolution and two-stage training; the other three methods use ViT-L/14@336 with 518 inputs. The additional A40 cost evaluation compares only C1 and C8, not these external methods. APRIL-GAN uses the authors' public code (`https://github.com/ByChelsea/VAND-APRIL-GAN`).

Interpretation and directly relevant comparators. The same-batch C1/C8 comparison matches source supervision, target data access, test metadata, and metric code. Native external runs provide complementary benchmarks, not this level of control. AA-CLIP's two-stage text-anchor and visual adaptation and AdaptCLIP's complementary adapters are especially relevant (Ma et al., 2025; Gao et al., 2026); neither was rerun here, so no numerical superiority over them is claimed. A direct comparison should match source anomaly supervision, target-normal reference access, test splits, and metric implementation, and separately report backbone and runtime. Matching FLOPs alone would not resolve these protocol differences.

AdaCLIP AUPRO code audit. The archived AdaCLIP `tools/metrics.py` and our evaluator both use a nominal 200-step score-threshold grid, connected-component overlap, FPR below 0.30, and normalization by the retained FPR span. They differ in grid construction (`linspace` versus

`arange`) and degenerate-case handling. Consequently, a different nominal FPR cutoff is not supported as an explanation of the low archived AdaCLIP values. This source inspection is not a prediction-level rerun: inputs, preprocessing, and the complete training/inference execution would need to be aligned before drawing a matched-comparison conclusion. The original measurements are retained without declaring them erroneous.

Table 27: Per-dataset offline comparisons (%). C8 uses the current same-batch three-seed means, identical to Table 2; external rows retain their archived native configurations. These are descriptive, not protocol-matched comparisons.

Method	Dataset	AUROC-PX	AUPRO	AUROC-SP	AP-SP
APRIL-GAN	BTAD	89.300	68.700	73.900	70.100
APRIL-GAN	MPDD	94.300	83.800	76.700	82.500
APRIL-GAN	KSDD2	95.500	93.600	89.000	77.000
APRIL-GAN	VisA	94.200	86.600	77.500	80.900
APRIL-GAN	WFDD	84.700	56.300	93.600	94.500
AdaCLIP (MVTec-only)	BTAD	94.840	28.910	86.170	95.160
AdaCLIP (MVTec-only)	MPDD	97.000	67.700	77.610	82.070
AdaCLIP (MVTec-only)	KSDD2	98.800	78.710	97.700	92.880
AdaCLIP (MVTec-only)	VisA	96.050	80.240	75.340	79.340
AdaCLIP (MVTec-only)	WFDD	94.440	37.630	82.050	84.290
AnomalyCLIP	BTAD	94.200	75.400	88.200	88.200
AnomalyCLIP	MPDD	96.500	88.700	77.500	82.500
AnomalyCLIP	KSDD2	98.500	96.400	94.100	84.300
AnomalyCLIP	VisA	95.500	86.700	82.000	85.300
AnomalyCLIP	WFDD	91.600	76.300	92.400	93.400
C8 (current, 3 seeds)	BTAD	94.932	80.662	93.988	97.259
C8 (current, 3 seeds)	MPDD	96.886	90.394	81.013	85.087
C8 (current, 3 seeds)	KSDD2	97.190	95.399	97.289	91.042
C8 (current, 3 seeds)	VisA	95.374	90.697	84.926	87.814
C8 (current, 3 seeds)	WFDD	91.358	72.571	91.563	92.615
WinCLIP	BTAD	72.700	27.300	68.200	70.900
WinCLIP	MPDD	71.200	40.500	61.500	69.200
WinCLIP	KSDD2	94.000	77.600	93.900	78.100
WinCLIP	VisA	73.200	51.000	75.500	78.700
WinCLIP	WFDD	74.700	43.600	92.300	93.700

Module interfaces and tensor dependencies. Prototype operates separately on each layer's $N \times d$ patch tensor. The spatial assignment matrix has shape $32 \times N$, the prototypes have shape $32 \times d$, and the redistributed context returns to $N \times d$. Residual refinement therefore preserves the number of patches and the channel interface. GVA reads the global vector and the four levels of refined tokens; local means and maxima are concatenated into $2d$-dimensional statistics and converted into a scalar through level-wise weighting and an image head. IDAC changes only the text anchors used for image scoring. All abnormal texts share one residual, and the original anchors continue to serve the pixel branch.

Training boundaries and checkpoints. When Prototype, IDAC, and GVA are trained in sequence, later stages freeze previously learned modules. Final evaluation reads the prespecified epoch 8 parameters from each stage rather than selecting the best epoch on target domains. This separation of interfaces also supports an implementation check: with Prototype fixed, adding GVA or IDAC should preserve pixel outputs and metrics. Repeated pixel values among the relevant configurations in Table 28 reflect this dependency, not multiple independent pieces of localization evidence. Operators and losses are detailed in Equations (1)–(10) of the main text.

GVA training objective. Training constrains both the fused output and the two subbranches: fused BCE learns the final decision, global/local auxiliary BCE supervises the respective outputs, and

pairwise ranking constrains the relative scores of anomalous and normal samples.

$$\mathcal{L}_{\text{GVA}} = \mathcal{L}_{\text{BCE}}(q_v, y) + \frac{0.25}{2} \sum_{u \in \{g,l\}} \mathcal{L}_{\text{BCE}}(q_u, y) + 0.25\mathcal{L}_{\text{rank}}. \tag{11}$$

BCE uses a positive-class weight of 1.5. Pairwise ranking uses a margin of 0.15 to compare anomalous and normal samples within the same batch. GVA contributes only to image scoring and does not feed back into the pixel anomaly map.

H Reproducibility and Evidence Boundaries

H.1 Source Data, Training, and Evaluation Implementation

Source split. The archived `MVTecDataset` reads the `test` entry of the MVTec AD metadata for supervised source training: 1,725 images across all 15 categories, comprising 467 normal and 1,258 anomalous images with defect masks. These are the source dataset's original test-directory images, repurposed for source supervision; no within-MVTec test performance is claimed. The normal-only original training split is not the input to this loader. The run does not enable `exclude_combined`, so all 55 combined-defect images are retained. Source metadata matches the run-manifest SHA256 `db66cf552367cdb7a36f1c5673f61aa73521d05d449b047a63c12e140fd22790`. This is not a normal-only unsupervised training regime.

Optimization and transformations. All stages use Adam with learning rate 10^{-3}, $(\beta_1, \beta_2) = (0.5, 0.999)$, default $\epsilon = 10^{-8}$ and zero weight decay; no learning-rate scheduler is used. Training runs ten epochs per stage and reads epoch 8. Four-image microbatches with two-step accumulation give effective batch size eight (the last accumulation window uses its actual size). The loader shuffles source images; frozen CLIP encoding uses CUDA autocast. With probability 0.20, the dataset constructs a 2×2 mosaic from four randomly drawn source images of the same category and their masks; otherwise it uses the indexed source image. The returned validation preprocessing transform resizes RGB images to 518×518 using bicubic interpolation, center-crops, converts to tensors, and applies CLIP channel normalization. Masks are resized, center-cropped, and converted to tensors by the separate mask transform. The active path does not add random flips or random-resized crops.

Stages and parameter settings. Prototype training uses equally weighted Dice and Focal terms. The text stage loads and freezes the Prototype checkpoint; GVA then loads and freezes the resulting model. The run manifest specifies $K = 32$, reduction ratio 4, lightweight Prototype, text-residual scale 0.02, GVA bottleneck 192, hidden dimension 256, residual scale 0.10, and fusion weight 0.20. The text-stage identity, separation, and ranking weights are zero; GVA uses the BCE, auxiliary, and ranking terms specified above. The three seeds are 42, 111, and 333. These settings identify the recorded configuration, not a new tuning procedure.

Metric implementation. The entry point is `test_prior_knowledge.py`. AUROC uses `roc_auc_score`, and image AP uses `average_precision_score`; category metrics are averaged within each dataset before the equal-domain and seed averages. AUPRO is computed by `cal_pro_score`: it samples 200 equally spaced thresholds from the category map minimum toward its maximum, labels connected ground-truth regions, and averages their overlap fractions. False-positive rate is measured over all background pixels. Only sampled points with FPR strictly below 0.30 are retained; their observed FPR range is rescaled to $[0, 1]$ before trapezoidal integration. The code does not interpolate an endpoint at FPR $= 0.30$. Thus, an implementation integrating exactly over $[0, 0.30]$ and dividing by 0.30 is not assumed numerically identical. Main and fixed-scale image scoring use the same checkpoints; only score transformations change.

H.2 Anonymous Evidence Supplement

The accompanying `PGISupplementEvidence.zip` contains experimental records and figure data. Paths below are relative to its `PGISupplement/` root. The manuscript, including this appendix, is submitted separately. The archive does not distribute source code, executable scripts, complete datasets, or pretrained/trained weights, and does not provide a runnable training or inference environment.

Structural results. `evidence/structural/full8_raw_120.csv` contains all eight combinations, three seeds, and five datasets; `evidence/structural/full8_summary.csv` contains their aggregates. The C1–C8 mapping appears in Appendix A. Each reported dataset value averages its three seeds, and each overall value weights the five datasets equally. The two CSV files in `evidence/external/` use these same current C8 records for our method, matching Tables 2 and 27; external-method measurements are unchanged. Historical C8 rows are not mixed into this comparison.

Scoring and timing evidence. `evidence/inference/` contains 30 C1/C8 score caches and their individual hashes, all 1,800 scoring-control rows and their summary, pooled A40 timings, and the batch-one timing record. These are saved measurements and numerical inputs for checking the reported aggregates, not newly executed experiments. The scoring definitions and timing scope are given in Appendix B.

Cache rows identify the test image, category, and anomaly label alongside the saved scalar scores. They permit inspection of image-level score transformations, but do not include full anomaly maps for every image; pixel-level AUROC and AUPRO are provided as recorded metrics, not recomputed from these scalar caches.

Figure data and integrity. `evidence/qualitative/mechanism/` contains the mechanism figure, the plotted arrays for its two samples, and provenance information. It does not include the complete feature tensors required to repeat the internal forward check. `documents/evidence_index.json` identifies the separately submitted PDF by filename, page count, and SHA256. `MANIFEST.sha256` covers every packaged file except itself, enabling integrity checks without distributing the model implementation.

H.3 Prototype Mechanism and Diagnostic Scope

Connection to compact attention. Equations (1)–(2) factor the context operator as $C = TAX$, where A is $K \times N$ and T is $N \times K$. For a fixed input, $\text{rank}(TA) \leq K$; both factors nevertheless depend on that input. This bound applies to the context operator, not to the full residual output, which retains X and applies gating and layer normalization. A^2-Nets previously introduced two-step feature aggregation and redistribution, and EMANet uses iteratively estimated compact bases for reconstruction (Chen et al., 2018; Li et al., 2019). Our implementation uses spatial convolutional assignments, one correlation-based redistribution, and global/local residual gates at four CLIP layers. The present ablation establishes the usefulness of this complete implementation within MultiADS, not the necessity of each operation or novelty of the aggregation principle.

A parameter-matched residual adapter, an existing aggregation/reconstruction block at the same four interfaces, and Prototype without its gate would distinguish capacity, generic context, and gate-specific effects. These replacement controls have not been run; the internal visualizations below are diagnostic illustrations, not substitutes for them. Likewise, the IDAC results evaluate separate bounded normal/abnormal residuals trained with BCE. Shared-residual controls and probability-calibration measurements would address different questions from the available AUROC/AP comparisons.

Prototype tensor extraction. Figure 5 uses two previously archived examples: MPDD connector, parts_mismatch/005.png, and BTAD product 01, ko/0002.bmp. The original C8 seed-42 epoch-8 adapter and frozen CLIP weights are loaded without retraining or changes to IDAC. The final selected layer (24) gives $N = 37 \times 37 = 1369$ spatial tokens and $K = 32$ prototypes. We export $A = \text{softmax}_N(g(X))$ and $T = \text{softmax}_K(XP^\top/\sqrt{d})$, showing A and $T^\top$ with original prototype order and raster-order patches. The residual image is the channel root-mean-square of $\beta G \odot W_o C$ before layer normalization. All quantities are computed from actual forward tensors, not inferred from anomaly scores. Independently recomputing the hooked layer reproduces its output exactly. Input and mask hashes match the archived visualization manifest. Extraction uses CPU float32; the two rightmost panels retain the archived C1/C8 anomaly maps rather than claiming a new GPU metric evaluation. Internal weights and residual magnitude are not defect probabilities.

These examples were selected for visibly improved localization, not as estimates of typical performance. On the BTAD example, archived per-image pixel AUROC is 86.697% for C1 and 94.687%

for C8; these values are distinct from dataset-level or seed-averaged metrics. The replacement changes only the illustrative sample, not the checkpoint, reported aggregate results, or scoring protocol.

Each column uses the same color scale for both examples. Logarithmic scales display A over $[10^{-6}, 10^{-1}]$ and T over $[10^{-4}, 1]$; weights outside the displayed ranges saturate at the endpoint colors. Residual RMS uses the joint observed minimum and maximum of the two examples, whereas both methods' anomaly maps use the fixed range $[0, 4]$ for the sum of four layer scores. No per-image contrast normalization is applied. Sample identifiers, checkpoint hashes, and raw arrays are retained with the figure source.

Internal figure and table construction checks verified the 120 structural records, aggregate values, and sources of paired images, retaining the mapping between C1–C8 and the original execution identifiers. The supplementary archive provides the resulting records and plotted arrays, not those internal scripts. Additional C1/C8 inference caches save scalar text, anomaly-map, and GVA scores together with timing records, as described in Appendix B. They do not contain complete per-anchor similarity tensors, attention, or feature embeddings, so these visualizations are not inferred from aggregate AUROC. Ground truth for target visualizations is used only for offline analysis and example selection, not for gradient training or checkpoint selection.

I Complete Same-Batch Individual Metrics

To allow the structural ablation to be independently recalculated, Table 28 lists all 120 rows of four metrics for 8 combinations, 3 seeds, and 5 target datasets rather than only eight rows of means. The mapping from C1–C8 to original execution identifiers is provided in Appendix A. Each row corresponds to an archived server-side `metrics_summary.json`. Percentages are printed to three decimal places, so the final digits recalculated from the printed table may differ from the unrounded CSV only within rounding error.

In the complete records, the four configurations without Prototype share pixel metrics, as do the four configurations with Prototype, consistent with GVA and IDAC not entering the anomaly-map path. Individual records can also be used to examine whether gains concentrate in a particular target domain or seed and how negative changes are distributed within conditional pairs.

Table 28: All 120 raw dataset-level metric records for the current eight same-batch combinations. Each row corresponds to one combination, random seed, and target dataset. Percentages are retained to three decimal places.

Config.	Seed	Dataset	AUROC-PX	AUPRO	AUROC-SP	AP-SP
C1	42	btad	90.709	71.295	89.874	88.875
C1	42	ksdd2	96.750	94.508	95.480	89.182
C1	42	mpdd	94.084	84.004	76.054	79.206
C1	42	visa	94.759	86.910	76.834	82.249
C1	42	wfdd	86.408	66.808	85.235	87.488
C1	111	btad	89.112	69.846	88.856	89.130
C1	111	ksdd2	95.752	93.568	96.834	89.510
C1	111	mpdd	94.805	85.789	71.728	74.607
C1	111	visa	94.336	87.255	79.673	84.607
C1	111	wfdd	82.473	64.998	95.758	96.357
C1	333	btad	90.611	72.447	90.132	91.371
C1	333	ksdd2	96.090	94.070	96.063	89.541
C1	333	mpdd	94.077	84.372	71.843	75.443
C1	333	visa	94.986	87.569	77.757	83.269
C1	333	wfdd	85.324	65.594	86.923	89.374
C2	42	btad	95.330	81.130	93.669	95.379
C2	42	ksdd2	96.988	95.576	96.312	89.074
C2	42	mpdd	96.943	90.823	80.350	84.958
C2	42	visa	95.433	90.854	82.052	86.048
C2	42	wfdd	91.302	71.938	90.913	91.550
C2	111	btad	95.134	81.591	93.527	95.160
C2	111	ksdd2	97.277	95.731	97.046	90.729

Table 28: All 120 raw dataset-level metric records for the current eight same-batch combinations. Each row corresponds to one combination, random seed, and target dataset. Percentages are retained to three decimal places. (continued).

Config.	Seed	Dataset	AUROC-PX	AUPRO	AUROC-SP	AP-SP
C2	111	mpdd	96.941	91.864	78.410	82.536
C2	111	visa	95.414	90.858	82.952	86.528
C2	111	wfdd	92.311	74.597	89.042	90.616
C2	333	btad	94.331	79.264	93.212	95.429
C2	333	ksdd2	97.306	94.890	95.718	88.965
C2	333	mpdd	96.774	88.493	81.230	85.779
C2	333	visa	95.276	90.380	82.832	86.371
C2	333	wfdd	90.460	71.179	89.268	90.642
C3	42	btad	90.709	71.295	90.441	93.207
C3	42	ksdd2	96.750	94.508	96.067	90.597
C3	42	mpdd	94.084	84.004	77.867	80.490
C3	42	visa	94.759	86.910	79.256	83.727
C3	42	wfdd	86.408	66.808	88.422	89.426
C3	111	btad	89.112	69.846	90.892	95.158
C3	111	ksdd2	95.752	93.568	97.474	92.223
C3	111	mpdd	94.805	85.789	73.349	75.263
C3	111	visa	94.336	87.255	83.447	87.284
C3	111	wfdd	82.473	64.998	95.120	95.955
C3	333	btad	90.611	72.447	91.589	94.831
C3	333	ksdd2	96.090	94.070	95.942	90.264
C3	333	mpdd	94.077	84.372	74.980	76.710
C3	333	visa	94.986	87.569	78.630	83.754
C3	333	wfdd	85.324	65.594	88.398	90.312
C4	42	btad	90.709	71.295	90.119	89.512
C4	42	ksdd2	96.750	94.508	95.587	89.404
C4	42	mpdd	94.084	84.004	75.945	79.103
C4	42	visa	94.759	86.910	77.655	82.831
C4	42	wfdd	86.408	66.808	86.758	88.719
C4	111	btad	89.112	69.846	89.007	89.529
C4	111	ksdd2	95.752	93.568	96.938	89.948
C4	111	mpdd	94.805	85.789	72.203	74.725
C4	111	visa	94.336	87.255	80.276	85.117
C4	111	wfdd	82.473	64.998	95.922	96.511
C4	333	btad	90.611	72.447	90.219	91.949
C4	333	ksdd2	96.090	94.070	96.152	89.875
C4	333	mpdd	94.077	84.372	71.962	75.618
C4	333	visa	94.986	87.569	78.455	83.849
C4	333	wfdd	85.324	65.594	88.890	90.797
C5	42	btad	95.330	81.130	94.166	96.859
C5	42	ksdd2	96.988	95.576	96.992	89.784
C5	42	mpdd	96.943	90.823	81.812	86.104
C5	42	visa	95.433	90.854	84.092	87.263
C5	42	wfdd	91.302	71.938	91.071	91.800
C5	111	btad	95.134	81.591	93.889	96.685
C5	111	ksdd2	97.277	95.731	97.471	92.000
C5	111	mpdd	96.941	91.864	79.605	83.432
C5	111	visa	95.414	90.858	85.615	88.405
C5	111	wfdd	92.311	74.597	90.688	91.854
C5	333	btad	94.331	79.264	93.997	97.781
C5	333	ksdd2	97.306	94.890	97.271	90.755
C5	333	mpdd	96.774	88.493	82.007	86.423
C5	333	visa	95.276	90.380	84.536	87.305
C5	333	wfdd	90.460	71.179	89.747	90.961
C6	42	btad	95.330	81.130	93.703	95.699
C6	42	ksdd2	96.988	95.576	96.406	89.384
C6	42	mpdd	96.943	90.823	80.319	84.849
C6	42	visa	95.433	90.854	82.461	86.371
C6	42	wfdd	91.302	71.938	92.179	92.764
C6	111	btad	95.134	81.591	93.595	95.677

Table 28: All 120 raw dataset-level metric records for the current eight same-batch combinations. Each row corresponds to one combination, random seed, and target dataset. Percentages are retained to three decimal places. (continued).

Config.	Seed	Dataset	AUROC-PX	AUPRO	AUROC-SP	AP-SP
C6	111	ksdd2	97.277	95.731	97.066	90.957
C6	111	mpdd	96.941	91.864	78.714	82.305
C6	111	visa	95.414	90.858	83.282	86.788
C6	111	wfdd	92.311	74.597	90.799	92.151
C6	333	btad	94.331	79.264	93.206	95.691
C6	333	ksdd2	97.306	94.890	95.875	89.403
C6	333	mpdd	96.774	88.493	81.293	85.699
C6	333	visa	95.276	90.380	83.112	86.587
C6	333	wfdd	90.460	71.179	90.518	91.836
C7	42	btad	90.709	71.295	90.492	93.604
C7	42	ksdd2	96.750	94.508	96.136	90.741
C7	42	mpdd	94.084	84.004	77.714	80.376
C7	42	visa	94.759	86.910	79.837	84.151
C7	42	wfdd	86.408	66.808	89.560	90.514
C7	111	btad	89.112	69.846	90.877	95.309
C7	111	ksdd2	95.752	93.568	97.516	92.313
C7	111	mpdd	94.805	85.789	73.450	75.415
C7	111	visa	94.336	87.255	83.632	87.450
C7	111	wfdd	82.473	64.998	95.175	96.002
C7	333	btad	90.611	72.447	91.597	95.082
C7	333	ksdd2	96.090	94.070	96.005	90.511
C7	333	mpdd	94.077	84.372	75.328	76.920
C7	333	visa	94.986	87.569	79.279	84.310
C7	333	wfdd	85.324	65.594	89.751	91.532
C8	42	btad	95.330	81.130	94.140	96.987
C8	42	ksdd2	96.988	95.576	97.045	89.977
C8	42	mpdd	96.943	90.823	81.584	85.703
C8	42	visa	95.433	90.854	84.363	87.495
C8	42	wfdd	91.302	71.938	91.986	92.834
C8	111	btad	95.134	81.591	93.912	96.949
C8	111	ksdd2	97.277	95.731	97.480	92.136
C8	111	mpdd	96.941	91.864	79.594	83.401
C8	111	visa	95.414	90.858	85.703	88.479
C8	111	wfdd	92.311	74.597	91.841	92.961
C8	333	btad	94.331	79.264	93.912	97.840
C8	333	ksdd2	97.306	94.890	97.343	91.014
C8	333	mpdd	96.774	88.493	81.861	86.155
C8	333	visa	95.276	90.380	84.712	87.467
C8	333	wfdd	90.460	71.179	90.863	92.051

J Development History and Interpretation of Transfer Results

The current comparison is the dedicated eight-configuration rerun identified by `iclr2027_full8_samebatch_20260914`. Its manifest fixes the backbone, input size, seeds, ten training epochs, epoch-8 evaluation, $K = 32$, and $w = 0.20$. These settings are fixed within that execution; the archive does not establish a prospective lock before the target domains were first inspected.

Earlier target-domain evaluations influenced whether local-refinement and side-branch variants were retained, revised, or discarded, and motivated a separate visual image-scoring branch. In particular, HGA/FB development used per-domain and cross-seed performance thresholds, while a historical GVA branch provided the signal for further image-scoring development. These trials used different code snapshots and baselines. Their detailed discarded-module catalog is omitted here; this does not change their role in development. The earlier 75-record batch is not pooled with the current 120 records.

The later fixed-checkpoint audit reuses C1/C8 from the current batch and changes only scoring rules. Its seven-weight fusion scan is post-hoc; $w = 0.20$ remains unchanged. The resulting five-domain comparisons therefore measure transfer to domains excluded from gradient training but visible during development. Repeated seeds assess training variability, not independence from this design feedback. No untouched, prospectively held-out target-domain confirmation is reported. Such a test would require locking the architecture, checkpoints, and scoring rule before accessing the new domain's results.

K Supplementary Literature Comparisons and Protocol Boundaries

Memory banks and current-image prototypes. Both PatchCore's representative normal patches (Roth et al., 2022) and PaDiM's normal-feature statistics (Defard et al., 2021) identify anomalies through local references. We retain the role of local context but change its source: the K prototypes are aggregated solely from the current image, have no normality labels, and do not access a target-normal reference set. The prototype count therefore does not correspond to the memory capacity of these methods.

Reconstruction, distillation, and discriminative adaptation. Teacher–student discrepancies form the anomaly signal in Uninformed Students (Bergmann et al., 2020); Reverse Distillation and subsequent work further investigate one-class embeddings and reverse reconstruction (Deng & Li, 2022; Tien et al., 2023). DeSTSeg combines a denoising student with segmentation supervision, while EfficientAD considers both low-latency local discrepancies and global information (Zhang et al., 2023; Batzner et al., 2024). These studies motivate separate discussion of local responses, global decisions, and inference cost. Because matched-throughput tests were not conducted, these approaches are not included in an efficiency ranking.

Anomaly synthesis and density modeling. Anomaly evidence can also be obtained through synthesis and density modeling. CutPaste constructs self-supervised tasks by locally cutting and pasting image regions; DRAEM jointly learns reconstruction and anomaly discrimination; and SimpleNet generates anomalous features in an adapted feature space (Li et al., 2021; Zavrtanik et al., 2021; Liu et al., 2023). CFLOW-AD describes normal-feature densities with conditional invertible transformations (Gudovskiy et al., 2022). We directly supervise multi-level pixel responses using source defect annotations without introducing an additional generative model. Results from the original papers under their normal-sample training settings are not merged into the current five-domain comparison.

Prompt optimization and text anchors. Prompt learning adapts pretrained models by changing their encoded inputs. CoOp learns prompt context, CoCoOp introduces image conditioning, and MaPLe jointly optimizes visual and language prompts (Zhou et al., 2022b;a; Khattak et al., 2023a); PromptSRC regularizes adaptation to reduce forgetting (Khattak et al., 2023b). IDAC instead adjusts the image-scoring anchors after encoding, using a different parameter interface. This paper evaluates that residual-calibration interface rather than reproducing a complete prompt-learning pipeline.

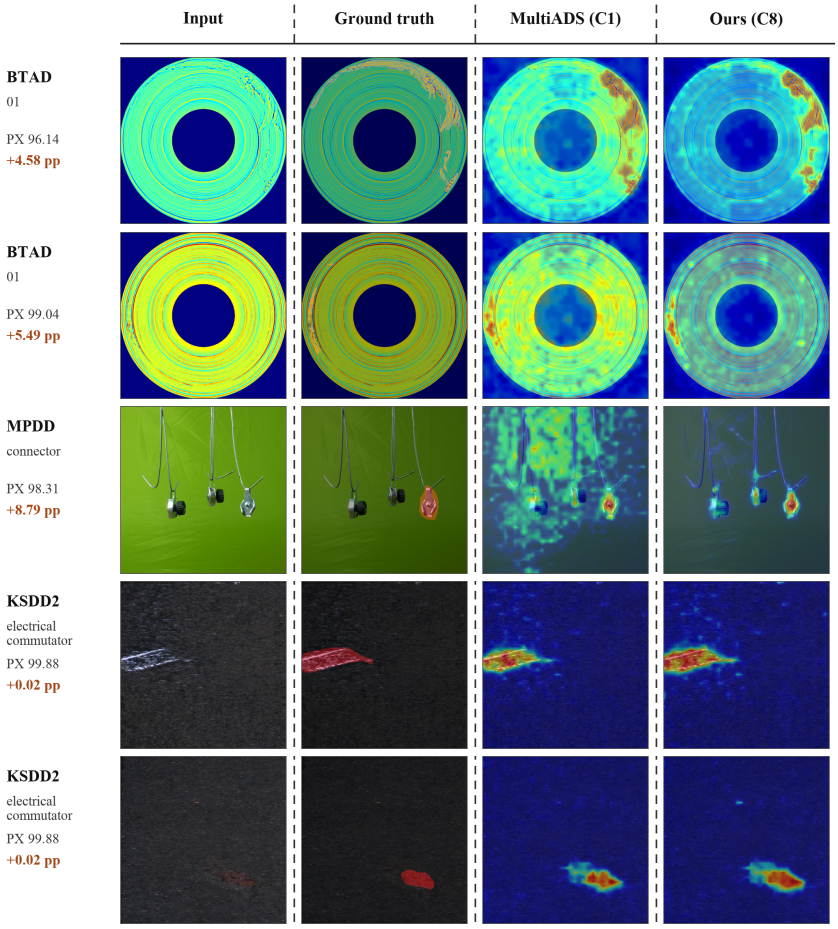
Zero-shot, few-shot, and explanatory detection. Feature adaptation also requires distinguishing residual modules from reference-sample retrieval. CLIP-Adapter adds bottleneck residuals to pretrained representations, whereas Tip-Adapter corrects predictions with a few-shot key–value cache (Gao et al., 2024; Zhang et al., 2022); VPT inserts prompts into visual encoder inputs (Jia et al., 2022). We neither use a target-domain cache nor insert visual prompts into the frozen encoder. DINOv2 (Oquab et al., 2024), used in the historical FDSA side branch, is an independent self-supervised visual model. That single-seed trial did not enter the final method and cannot establish the general effectiveness of the pretrained model.

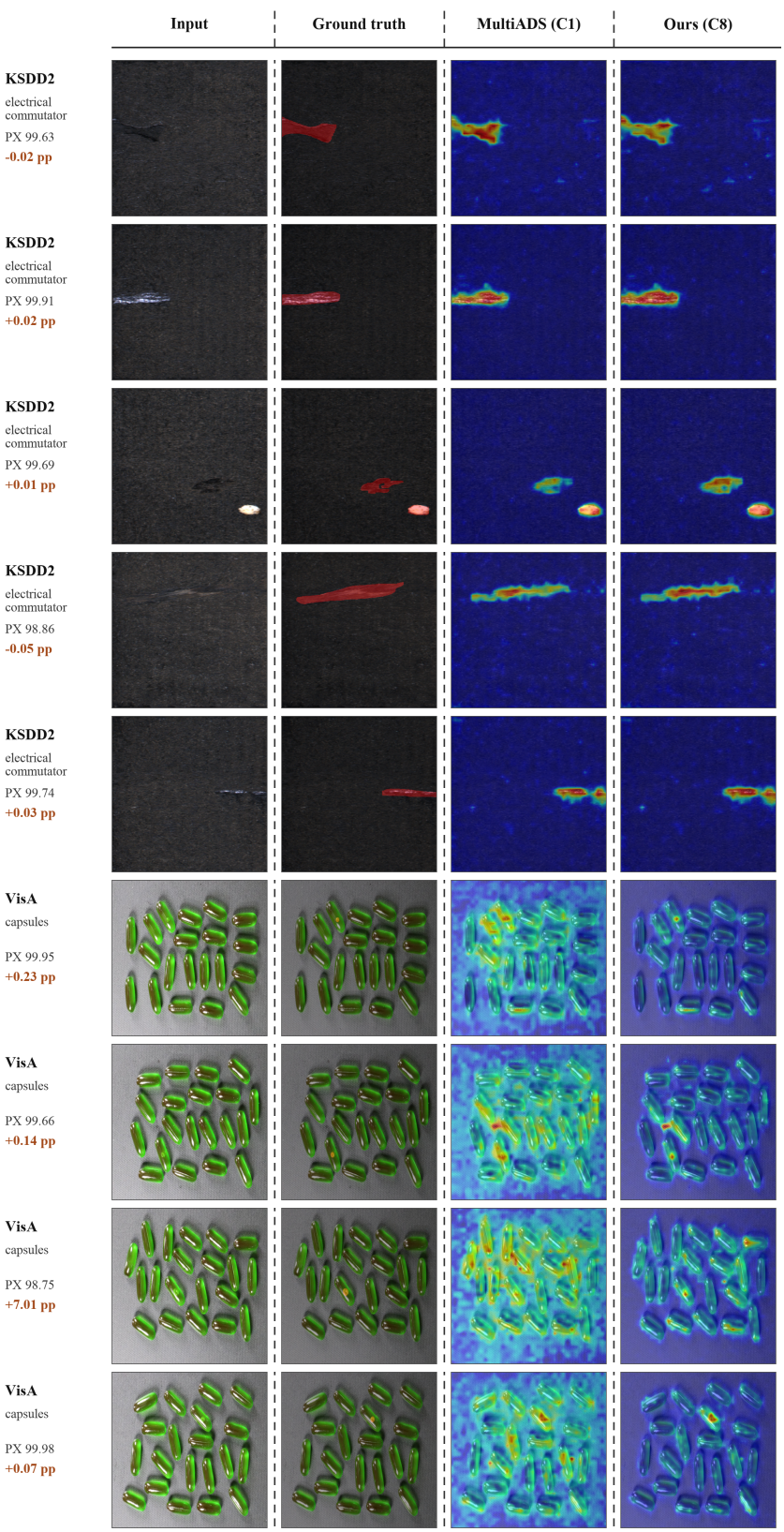
AnomalyCLIP's object-agnostic prompts and AA-CLIP's anomaly-aware anchors address cross-category anomaly semantics (Zhou et al., 2024; Ma et al., 2025); PromptAD studies few-shot prompt learning using only normal samples (Li et al., 2024). AnomalyGPT extends industrial anomaly detection to large vision–language models and conversational interaction (Gu et al., 2024). We use neither target-normal support samples nor generated textual anomaly explanations, so these references define the method's position rather than supplying results for baselines that were not run.

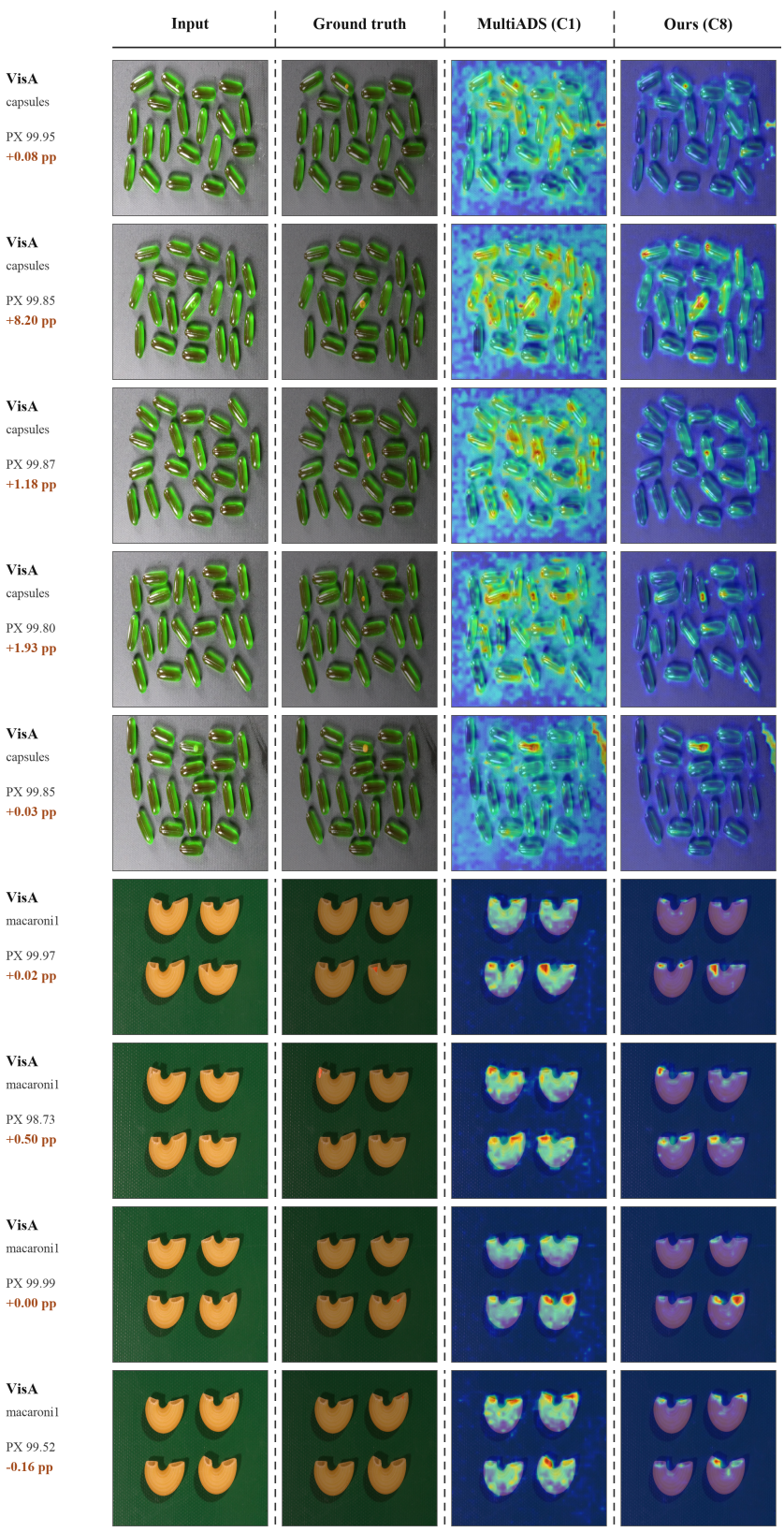
Questions that visualizations can answer. Localization images and category statistics address spatial responses and performance distributions, respectively. We follow AnomalyCLIP's organizational approach, but all visual outputs and experimental numbers come from this study. Paired images reveal hotspot coverage and background responses, while category tables check the distribution of average gains. Text-similarity distributions, attention, and image-score calibration require the corresponding tensors; existing aggregate metrics cannot substitute for these measurements.

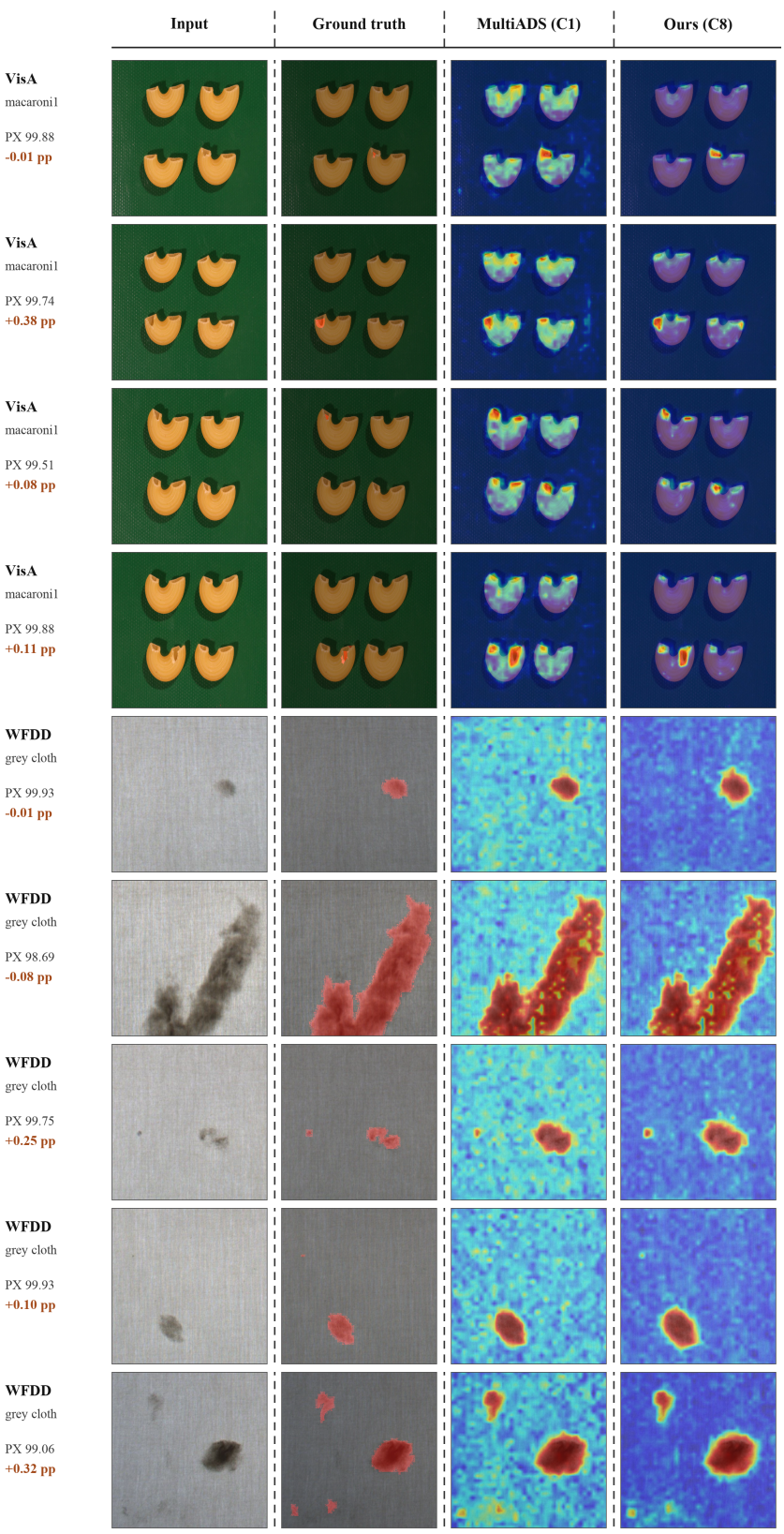
L Selected High-Quality Localization Gallery

We manually selected 40 of the 260 actual paired images: 2 from BTAD, 1 from MPDD, 7 from KSDD2, 17 from VisA, and 13 from WFDD. Selection focuses on correspondence between hotspots and defects, normal-region responses, and the visibility of localization contours rather than ranking solely by pixel AUROC. Samples from the same dataset, category, and defect type are arranged consecutively. The examples favor clear localization and do not represent the complete test set or category-average performance; negative changes and the full candidate distribution remain available in Figure 13. KSDD2 examples demonstrate high absolute quality rather than pronounced improvement. The gallery also retains six clearly localized examples with slightly lower pixel AUROC, without equating visual improvement with improvement in every metric. Each row shows the input, ground truth, C1, and C8 in that order, annotated with C8 pixel AUROC and its signed increment. All originals are verified against the manifest SHA256 hashes. Only proportional resizing and removal of repeated headers are applied; heatmap values, colors, and normalization are unchanged.




Input
Ground truth
MultiADS (C1)
Ours (C8)
KSDD2
electrical commutator
PX 99.63
-0.02 pp
KSDD2
electrical commutator
PX 99.91
+0.02 pp
KSDD2
electrical commutator
PX 99.69
+0.01 pp
KSDD2
electrical commutator
PX 98.86
-0.05 pp
KSDD2
electrical commutator
PX 99.74
+0.03 pp
VisA
capsules
PX 99.95
+0.23 pp
VisA
capsules
PX 99.66
+0.14 pp
VisA
capsules
PX 98.75
+7.01 pp
VisA
capsules
PX 99.98
+0.07 pp


Input
Ground truth
MultiADS (C1)
Ours (C8)
VisA
capsules
PX 99.95
+0.08 pp
VisA
capsules
PX 99.85
+8.20 pp
VisA
capsules
PX 99.87
+1.18 pp
VisA
capsules
PX 99.80
+1.93 pp
VisA
capsules
PX 99.85
+0.03 pp
VisA
macaroni1
PX 99.97
+0.02 pp
VisA
macaroni1
PX 98.73
+0.50 pp
VisA
macaroni1
PX 99.99
+0.00 pp
VisA
macaroni1
PX 99.52
-0.16 pp


Input
Ground truth
MultiADS (C1)
Ours (C8)
VisA
macaroni1
PX 99.88
-0.01 pp
VisA
macaroni1
PX 99.74
+0.38 pp
VisA
macaroni1
PX 99.51
+0.08 pp
VisA
macaroni1
PX 99.88
+0.11 pp
WFDD
grey cloth
PX 99.93
-0.01 pp
WFDD
grey cloth
PX 98.69
-0.08 pp
WFDD
grey cloth
PX 99.75
+0.25 pp
WFDD
grey cloth
PX 99.93
+0.10 pp
WFDD
grey cloth
PX 99.06
+0.32 pp

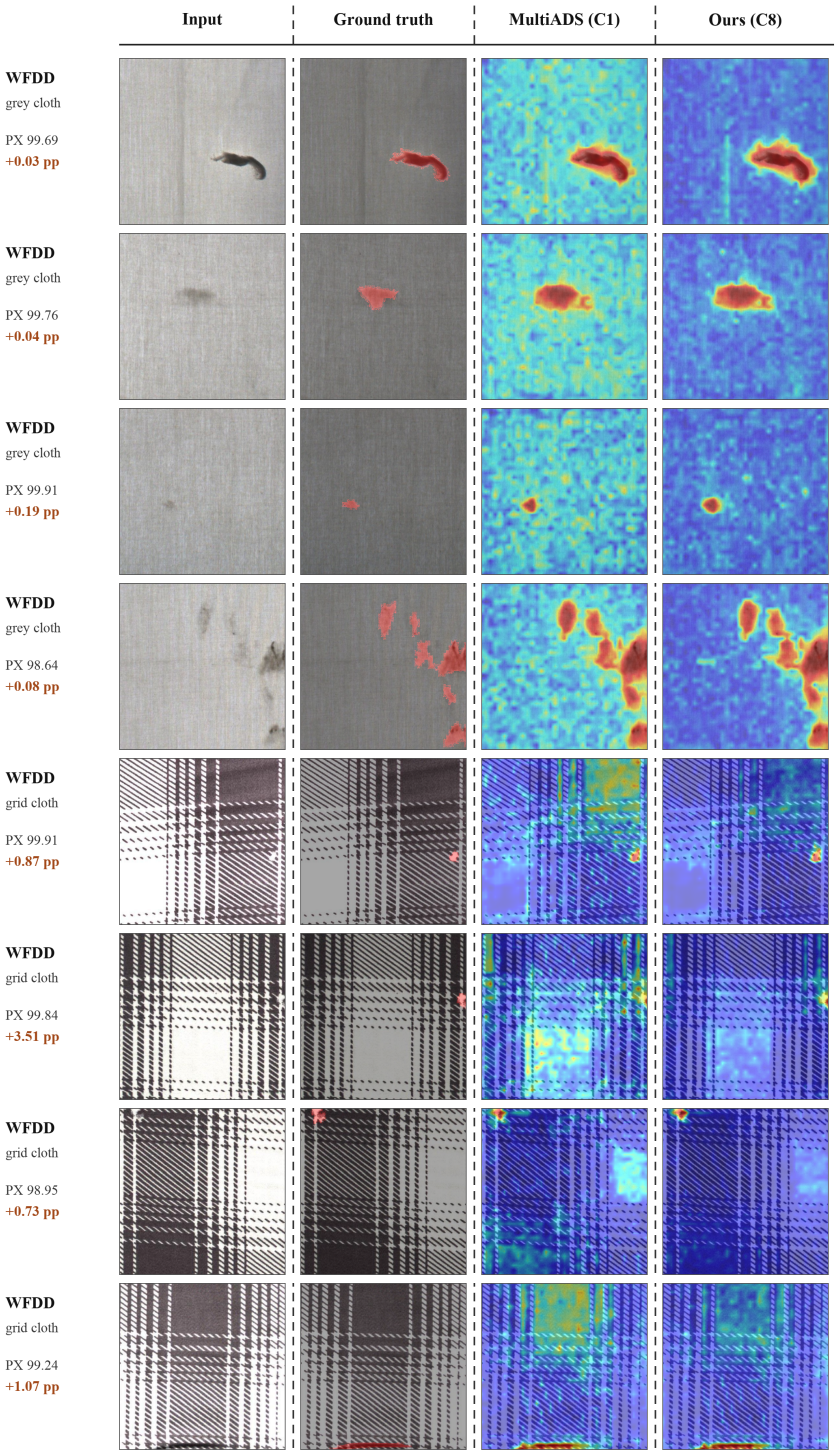


Figure 14: Selected localization comparisons: input, ground truth, C1, and C8. Selection criteria are described in this section. These examples do not represent the overall distribution; negative changes are reported in Appendix F.