Co-Evolutionary Prompt Optimization with Cross-Category Transfer for Zero-Shot Anomaly Detection

Anonymous ACL submission

Abstract

Zero-shot anomaly detection (ZSAD) has gained significant attention for its practical value in industrial inspection. Recently, CLIP-based approaches have been widely adopted in ZSAD due to their strong vision-language generalization capabilities. However, existing methods commonly employ continuous prompt embeddings for prompt optimization and encode semantics in latent vectors, which lack interpretability and scalability. To this end, we propose CoEvoAD, a co-evolutionary framework for discrete prompt selection. CoEvoAD performs prompt search in the discrete natural-language space using an evolutionary algorithm. Candidate prompts are iteratively generated, evaluated, and selected throughout population evolution, thus preserving the interpretability and composability of natural language. Furthermore, we introduce a Cross-Category Transfer Objective (CCTO), which treats held-out source categories as proxies for unseen categories and scores prompt rules based on their estimated cross-category transferability, effectively improving cross-category generalization. Extensive experiments are conducted to validate the effectiveness of CoEvoAD, and the results show that it achieves state-of-the-art performance across multiple anomaly detection datasets.

1 Introduction

Industrial anomaly detection aims to identify defective samples and localize anomalous regions in visual inspection images. In practice, deploying a detector to a new product category often requires normal images from the target category and, in supervised settings, defect labels or pixel-level masks, followed by category-specific retraining or calibration. This requirement is especially costly when production lines change frequently and real defects are scarce. Built on CLIP (Radford et al., 2021), recent zero-shot anomaly detection (ZSAD) methods

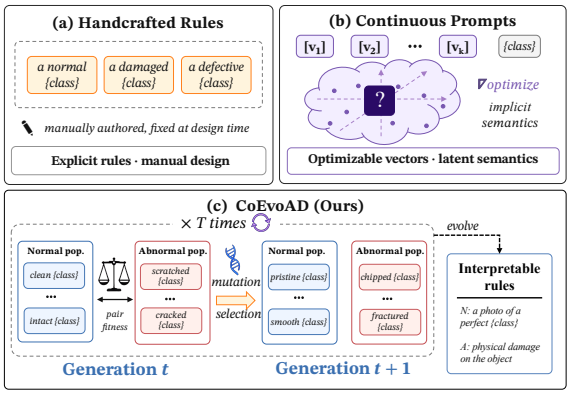

Figure 1: Prompt regimes for ZSAD: handcrafted templates, continuous prompts, and CoEvoAD.

reduce this dependency by comparing visual features with normal and abnormal text prompts, enabling detection in unseen categories (Jeong et al., 2023; Zhou et al., 2024; Qu et al., 2025). In this setting, prompt selection becomes central to specifying transferable normal–abnormal rules. We study how interpretable prompt rules can be automatically generated and selected for cross-category transfer, as illustrated in Figure 1.

Despite recent progress in CLIP-based ZSAD, prompt construction and selection still face two limitations for cross-category transfer. First, existing prompt construction still struggles to jointly achieve prompt interpretability and scalability. Handcrafted rules remain explicit but fixed after design, while prompt-learning methods such as AnomalyCLIP (Zhou et al., 2024), AdaCLIP (Cao et al., 2024), and Bayes-PFL (Qu et al., 2025) improve adaptation by optimizing learnable prompt embeddings or distributions over them. Yet the learned prompt semantics are encoded in uninterpretable continuous representations rather than explicit natural-language rules, limiting prompt interpretability and rule-level reuse. Second, source-category performance alone does not provide reli-

able evidence of cross-category generalization. A prompt may perform well on the source categories used for selection yet fail on unseen categories, because its transferability across categories has not been evaluated. In summary, current prompt-selection methods in ZSAD still fail to achieve both interpretability and scalability. Moreover, candidate prompt selection relies solely on performance within the optimization categories, lacking cross-category validation signals.

To address the lack of prompt interpretability and scalability, we introduce CoEvoAD, a co-evolutionary discrete prompt-rule search framework. This framework performs search in the discrete natural language space using an evolutionary algorithm rather than relying on opaque black-box vector representations. Candidate prompt rules are continuously generated, evaluated, and selected throughout population evolution, while maintaining the interpretability, composability, and auditability of natural language, which provides strong potential for scalability. The fitness function combines the evaluation score on source data with the semantic contrast between normal and abnormal descriptions, thereby enabling a transparent and controllable optimization process without gradients. To avoid the objective conflict caused by mixing normal and abnormal rules in traditional methods, we maintain independent populations for the two roles and achieve co-evolution through a joint fitness design. This allows them to expand along their respective semantic directions in the large-scale natural language space while preserving the necessary semantic contrast. Overall, the framework transforms prompt-rule search from a single black-box vector into a structured, multi-role, multi-candidate co-evolutionary system, effectively improving interpretability and scalability.

To address the cross-category generalization issue, we introduce the Cross-Category Transfer Objective (CCTO), a prompt-rule selection criterion specifically designed for cross-category transfer. The key idea behind CCTO is to construct systematic held-out splits within the source domain, where a subset of source categories is treated as proxy unseen categories. This design enables the evaluation of candidate prompt rules under a simulated zero-shot transfer scenario, allowing us to assess their consistency and robustness when applied across categories. By leveraging only source-domain data, CCTO provides an evaluation signal that aligns more closely with the requirements of zero-shot generalization, without relying on any target-domain supervision. As a result, the proposed objective significantly improves the cross-category generalization ability of the selected rules, offering a more reliable selection criterion under the zero-shot setting.

2 Related Work

Traditional Anomaly Detection. Traditional anomaly detection usually learns category-specific normal patterns from target-category normal images. In feature space, PaDiM (Defard et al., 2021) models normal statistics, while PatchCore (Roth et al., 2022) stores patch-level memories. For discriminative training, DRAEM (Zavrtanik et al., 2021) synthesizes defects, whereas SimpleNet (Liu et al., 2023) perturbs features. Reverse Distillation (Deng and Li, 2022) uses teacher–student discrepancies, with UniAD (You et al., 2022) relying on reconstruction. However, these methods depend on target-category normal images, which limits deployment when new categories appear without prior normal data.

Zero-Shot Anomaly Detection. Zero-shot anomaly detection targets anomaly classification and localization on categories unseen during training. WinCLIP (Jeong et al., 2023) introduces hand-crafted normal/abnormal prompts for training-free anomaly detection. Subsequent methods shift from manual state words to optimized prompt representations. AnomalyCLIP (Zhou et al., 2024) addresses category dependence with object-agnostic prompts, while AdaCLIP (Cao et al., 2024) adopts static-dynamic prompt design. Bayes-PFL (Qu et al., 2025) further models prompt uncertainty through Bayesian prompt flow. Recent work broadens CLIP adaptation beyond text-side prompting. VCP-CLIP (Qu et al., 2024) uses visual context to guide prompt construction, whereas AA-CLIP (Ma et al., 2025) strengthens anomaly-aware textual anchors. AF-CLIP (Fang et al., 2025) and AdaptCLIP (Gao et al., 2026) further adapt visual features or CLIP representations for anomaly detection. MRAD (Xu et al., 2026) introduces memory-driven retrieval. These methods improve CLIP-based anomaly reasoning through prompt learning, visual context, feature adaptation, or memory retrieval.

Prompt Tuning and Discrete Prompt Optimization. Prompt tuning reduces manual prompt en-

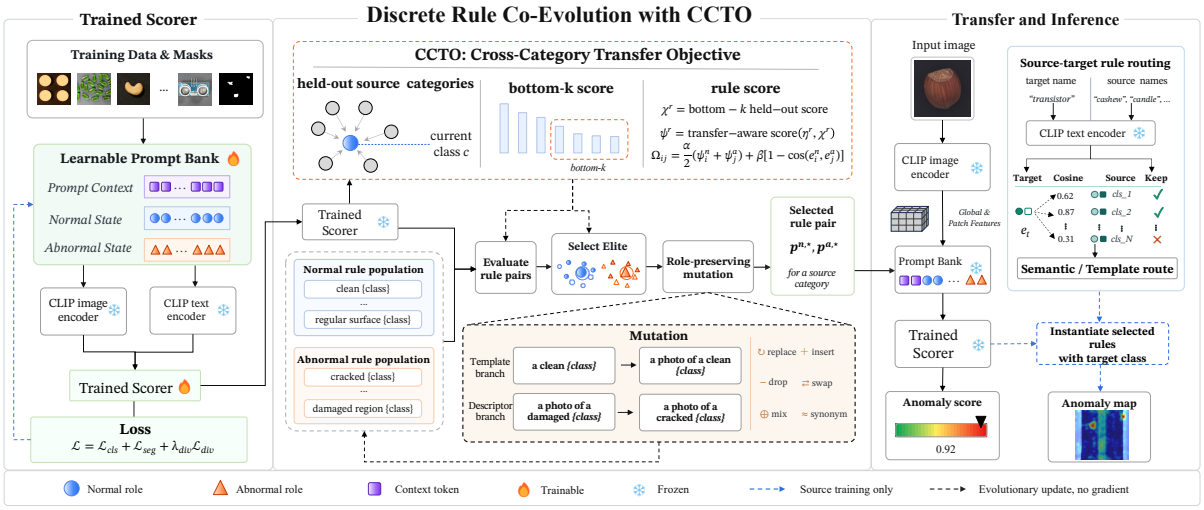

Figure 2: Overview of CoEvoAD. A frozen source-trained scorer evaluates role-separated normal and abnormal rule populations; CCTO uses held-out source categories to select rules for unseen-category inference.

gineering by optimizing prompt representations. CoOp (Zhou et al., 2022b) learns continuous context vectors for vision-language recognition. CoCoOp (Zhou et al., 2022a) makes the learned context condition-dependent. BBT (Sun et al., 2022) studies black-box prompt tuning without gradient access. Discrete prompt optimization keeps the search in token space. AutoPrompt (Shin et al., 2020) uses gradient-guided word substitution, RLPrompt (Deng et al., 2022) formulates prompt search as reinforcement learning, and EvoPrompt (Guo et al., 2024) applies population-based search to LLM prompts. CoEvoAD instead searches discrete normal and abnormal rule populations for ZSAD. Rule pairs are co-evolved with joint fitness and selected using cross-category evidence drawn from source-domain data.

3 Method

3.1 Overview

CoEvoAD treats cross-category ZSAD as discrete prompt-rule selection, as shown in Figure 2. The prompt-conditioned scorer is kept fixed during rule search, so the optimization only changes the natural-language rule strings used to render prompts. CoEvoAD maintains separate normal and abnormal populations, updates them through role-preserving mutation, and selects rule pairs with a joint fitness that combines evaluation scores and normal–abnormal semantic separation. CCTO further evaluates candidate rules on held-out categories from $\mathcal{C}_s$, adding a transfer-oriented signal without target-domain feedback. At inference, selected rules are instantiated for unseen categories through a pre-specified transfer policy and scored by the same prompt-conditioned scorer to produce image-level anomaly scores and pixel-level anomaly maps; each rule remains interpretable through its template, role-specific descriptors, and class-name slot.

3.2 Discrete Prompt-Rule Co-Evolution

CoEvoAD formulates prompt optimization as a discrete search over natural-language prompt rules. It uses the fixed prompt-conditioned scorer as the evaluator and varies only rule strings.

Role-typed rule representation and populations. CoEvoAD searches over structured natural-language rules rather than unconstrained prompt strings. Since anomaly scoring is defined over a normal–abnormal prompt pair, we organize the candidate space into two role-typed rule pools. For role $r \in \{\mathrm{n}, \mathrm{a}\}$ (normal and abnormal), a rule is represented as

$$\begin{aligned} p^r &= (\tau^r, D^r, \langle \text{class} \rangle), \\ D^r &= (d_1^r, \ldots, d_m^r), \end{aligned} \tag{1}$$

where τ^r is a role-compatible template, D^r is a sequence of descriptors drawn from the corresponding role vocabulary, and $\langle \text{class} \rangle$ is a class-name slot. The rendering function $z^r = R(p^r, c)$ fills the slot with category name c and concatenates the template and descriptors into a CLIP text prompt. The normal role uses descriptors of acceptable appearance (*e.g.*, *clean*, *intact*, *regular*), while the abnormal role uses descriptors of defect evidence (*e.g.*,

damaged, *cracked*, *contaminated*). This representation only defines the admissible candidate space; it does not manually determine the final prompt rule. Candidates are generated through role-preserving mutation and selected by the paired fitness function and CCTO, while keeping the role type fixed prevents normal and abnormal descriptions from being mixed during co-evolution.

We maintain two separate populations,

$$\begin{aligned} \mathcal{P}^{\mathrm{n}} &= \{p_1^{\mathrm{n}}, \ldots, p_N^{\mathrm{n}}\}, \\ \mathcal{P}^{\mathrm{a}} &= \{p_1^{\mathrm{a}}, \ldots, p_N^{\mathrm{a}}\}, \end{aligned} \tag{2}$$

where N is the population size and each p_i^r is a role-typed rule. Initial populations are constructed from role-specific templates and descriptor pools; since each candidate is an explicit natural-language rule, the same rule can be instantiated for another category by substituting the class-name placeholder.

Algorithm 1 summarizes the rule search. The scorer is frozen during this stage; only the rendered natural-language rules are changed. Candidate rules are scored by CCTO-adjusted role scores and selected through pair-level fitness on rule pairs.

Candidate evaluation and pair fitness. For a fixed optimization category c, candidates are scored on the validation split with the fixed prompt-conditioned scorer. Each candidate is evaluated with sampled partners from the opposite population.

Given a candidate pair $(p_i^{\mathrm{n}}, p_j^{\mathrm{a}})$, the pair fitness is

$$\begin{aligned} \Omega_{ij} &= \alpha \frac{\eta_i^{\mathrm{n}} + \eta_j^{\mathrm{a}}}{2} + \beta \delta_{ij}, \\ \delta_{ij} &= 1 - \cos(e_i^{\mathrm{n}}, e_j^{\mathrm{a}}), \end{aligned} \tag{3}$$

where η_i^r is the resulting candidate score, e_i^r is the normalized text embedding of the instantiated rule $R(p_i^r, c)$, and $r \in \{\mathrm{n}, \mathrm{a}\}$. The first term measures evaluation scores computed by the scorer, while the second keeps normal and abnormal descriptions separated in text space. Pairs with identical instantiated normal and abnormal strings are marked invalid and excluded from selection. When CCTO is enabled, η_i^r is replaced by the adjusted score $\psi^r(p_i^r, c)$ defined in Section 3.3; the pair-fitness form is unchanged.

Role-preserving mutation and selection. CoEvoAD updates the two populations through mutation only; no recombination across roles is used. Each new candidate is produced by mutating a same-role parent, and the mutation operates only on the mutable fields of the rule grammar in Eq. 1. For descriptors, we apply replacement, insertion, deletion, local reordering, and synonym substitution within the same role-specific vocabulary. For templates, we sample from a small role-compatible template pool. All mutations preserve the role type of the parent rule, so a normal candidate remains a description of acceptable appearance and an abnormal candidate remains a description of visible defect evidence.

Algorithm 1 Prompt-rule co-evolution with CCTO selection.

Input: Frozen scorer S_θ; source categories $\mathcal{C}_s$; role-specific rule pools $\mathcal{V}^{\mathrm{n}}, \mathcal{V}^{\mathrm{a}}$; population size N, generations G, elite size K_e, partner budget K
Output: Rule cache $\Pi = \{(c,\, p_c^{\mathrm{n}\star},\, p_c^{\mathrm{a}\star})\}_{c \in \mathcal{C}_s}$

```
for c ∈ C_s do
    H_c ← C_s \ {c}                          ▷ held-out source categories
    Initialize P_c^n and P_c^a from V^n and V^a
    for g = 1, ..., G do
        for r ∈ {n, a} do
            ψ^r(·, c) ← CCTOScore(P_c^r, c, H_c)   ▷ Eq. 5
        Sample K opposite-role partners per candidate
        Score sampled pairs by Ω(p_i^n, p_j^a)     ▷ Eq. 3
        Rank candidates by mean pair fitness
        Keep the top K_e candidates as role-wise elites
        Refill P_c^n and P_c^a with role-preserving mutations
    Exhaustively score all elite normal–abnormal pairs
    Π[c] ← arg max_{p^n, p^a} Ω(p^n, p^a)
return Π
```

At each generation, candidates are paired with a small diverse subset from the opposite population and ranked by their average pair fitness. Selection is performed separately within the normal and abnormal populations. We retain K_{elite} high-fitness candidates in each role and fill the remaining slots with mutated candidates. After G generations, we evaluate all top-K_{elite} normal–abnormal pairs with Eq. 3 and select the best pair as $\Pi_c = (p_c^{\mathrm{n}\star}, p_c^{\mathrm{a}\star})$.

Rule transfer and inference. After search, CoEvoAD stores one selected normal–abnormal rule pair $\Pi_c = (p_c^{\mathrm{n}\star}, p_c^{\mathrm{a}\star})$ for each source category $c \in \mathcal{C}_s$. For an unseen target category c_t, a pre-specified rule-transfer policy selects one stored pair using only category names and the saved source rules. The selected pair $(\tilde{p}_t^{\mathrm{n}}, \tilde{p}_t^{\mathrm{a}})$ is rendered with c_t, yielding $z_t^r = R(\tilde{p}_t^r, c_t)$ for $r \in \{\mathrm{n}, \mathrm{a}\}$. The fixed prompt-conditioned scorer then produces the image-level anomaly score and pixel-level anomaly map. The policy is fixed before target evaluation.

3.3 Cross-Category Transfer Objective

Prompt rules selected only on the optimization category may capture category-specific visual patterns

that do not transfer to unseen categories. CCTO turns prompt-rule selection into a category-held-out transfer test. For a rule optimized on category c, we render the same rule on each held-out source category $c' \in \mathcal{C}_s \setminus \{c\}$ and evaluate it with the frozen scorer. The resulting held-out scores measure whether the rule preserves the normal–abnormal distinction after category substitution, rather than only fitting the category on which it is optimized. Thus, candidate selection is guided by both within-category evaluation scores and cross-category consistency.

For a candidate rule p^r on category c, CCTO defines the cross-category score as

$$\chi^r(p, c) = \frac{1}{\kappa} \sum_{j=1}^{\kappa} \xi_{(j)}, \tag{4}$$

where $\xi_{(1)} \leq \cdots \leq \xi_{(m)}$ are the held-out scores sorted in ascending order after evaluating p^r on each remaining source category, and $\kappa = \min(k, m)$. We aggregate the held-out scores with a bottom-k operator. A simple mean can be dominated by easy held-out categories and may overlook rules that fail on harder transfers. The bottom-k score instead emphasizes the least transferable held-out categories while remaining less brittle than a strict minimum when the number of source categories is limited.

Normal and abnormal rules use the held-out scores in different ways. We define the role-adjusted selection scores as

$$\begin{aligned} \psi^{\mathrm{n}}(p, c) &= (1 - \lambda_{\mathrm{n}})\eta^{\mathrm{n}} + \lambda_{\mathrm{n}}\chi^{\mathrm{n}}, \\ \psi^{\mathrm{a}}(p, c) &= \eta^{\mathrm{a}} + \lambda_{\mathrm{a}}\left(\eta^{\mathrm{a}} - \chi^{\mathrm{a}}\right), \end{aligned} \tag{5}$$

where $\eta^r = \eta^r(p, c)$ is the in-category score, $\chi^r = \chi^r(p, c)$ is the cross-category score in Eq. 4, and $\lambda_{\mathrm{n}}, \lambda_{\mathrm{a}}$ control the two role-specific adjustments. The normal branch interpolates between in-category performance and cross-category consistency. The abnormal branch uses the margin $\eta^{\mathrm{a}} - \chi^{\mathrm{a}}$ to retain defect-discriminative abnormal rules and suppress overly generic anomaly descriptions. The adjusted score $\psi^r(p, c)$ replaces η_i^r in Eq. 3. All held-out categories are drawn from $\mathcal{C}_s$.

3.4 Overall Loss

The prompt bank and scoring head are trained on the source dataset, while the CLIP image and text encoders remain frozen. For each source sample, the category name is used as the textual anchor; evolved prompt rules are introduced only during prompt-rule search. The scorer is optimized with an image-level binary classification loss, a pixel-level localization loss combining focal and Dice terms, and a same-role prompt-group diversity regularizer. We write the objective compactly as $\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{seg}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}$. In implementation, scorer training further includes mask-guided crop augmentation, inter-role margin regularization, and category-agnostic regularization. This training recipe is identical for the class-name control and all main CoEvoAD comparisons, so matched-control gains can be attributed to prompt-rule selection rather than scorer updates.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate CoEvoAD on six industrial anomaly detection benchmarks: MVTec-AD (Bergmann et al., 2019) (15 categories of manufactured objects), VisA (Zou et al., 2022) (12 categories of complex inspection objects), BTAD (Mishra et al., 2021) (three industrial product categories), KSDD2 (Božič et al., 2021) (industrial steel surface defects), DAGM (Wieler and Hahn, 2007) (manually re-annotated synthetic-texture defects), and RSDD (Niu et al., 2021) (rail surface defects).

Protocol. Following prior CLIP-based ZSAD settings (Zhou et al., 2024; Qu et al., 2025), CoEvoAD is evaluated under strict cross-dataset transfer: the scorer is trained on a source dataset and evaluated on target datasets without target-domain images, labels, or supervision. The two primary transfer directions are VisA→MVTec-AD and MVTec-AD→VisA. The remaining datasets (BTAD, KSDD2, DAGM, RSDD) serve as external industrial targets evaluated under the same source-only protocol.

Baselines. We compare against representative CLIP-based ZSAD baselines: WinCLIP (Jeong et al., 2023), AnomalyCLIP (Zhou et al., 2024), AdaCLIP (Cao et al., 2024), Bayes-PFL (Qu et al., 2025), and MRAD (Xu et al., 2026). Baseline numbers in Tables 1 and 2 are taken from the original papers or their official codebases where available.

Evaluation metrics. Following common ZSAD evaluation, we report image-level AUROC and AP and pixel-level AUROC and Per-Region Overlap (PRO) for benchmark comparison. For matched

Table 1: Image-level benchmark comparison (%) across six anomaly detection datasets. Bold/underline denote best/second-best; "–" marks entries not reported by the original paper.

Method → Dataset ↓	WinCLIP (CVPR'23)		AnomalyCLIP (ICLR'24)		AdaCLIP (ECCV'24)		Bayes-PFL (CVPR'25)		MRAD (ICLR'26)		CoEvoAD (Ours)	
	AUROC↑	AP↑	AUROC↑	AP↑	AUROC↑	AP↑	AUROC↑	AP↑	AUROC↑	AP↑	AUROC↑	AP↑
MVTec-AD	91.8	95.1	91.5	96.2	92.0	96.4	92.3	96.7	**94.0**	**97.4**	93.4	96.8
VisA	78.1	77.5	82.1	85.4	83.0	84.9	87.0	89.2	85.7	88.3	**87.4**	**89.7**
BTAD	83.3	84.1	89.1	91.1	91.6	92.4	93.2	**96.5**	92.4	94.2	**94.4**	96.1
KSDD2	93.5	77.9	92.1	77.8	95.9	95.9	97.3	97.9	95.1	88.9	**97.4**	**97.9**
DAGM	89.6	90.4	95.6	94.6	96.5	95.7	97.7	97.0	**98.4**	**98.6**	98.2	97.6
RSDD	85.3	65.3	73.5	55.0	89.1	70.8	94.1	92.3	–	–	**98.9**	**98.9**
Mean	86.9	81.7	87.3	83.4	91.4	89.4	93.6	94.9	–	–	**95.0**	**96.2**

Table 2: Pixel-level benchmark comparison (%; AUROC and Per-Region Overlap (PRO)). Bold/underline denote best/second-best across the six datasets; "–" marks entries not reported by the original paper.

Method → Dataset ↓	WinCLIP (CVPR'23)		AnomalyCLIP (ICLR'24)		AdaCLIP (ECCV'24)		Bayes-PFL (CVPR'25)		MRAD (ICLR'26)		CoEvoAD (Ours)	
	AUROC↑	PRO↑	AUROC↑	PRO↑	AUROC↑	PRO↑	AUROC↑	PRO↑	AUROC↑	PRO↑	AUROC↑	PRO↑
MVTec-AD	85.1	64.6	91.1	81.4	86.8	33.8	91.8	87.4	**93.0**	86.8	92.2	**87.9**
VisA	79.6	56.8	95.5	87.0	95.1	71.3	95.6	88.9	**95.9**	88.0	95.8	**89.4**
BTAD	71.4	32.8	93.3	69.3	87.7	17.1	93.9	76.6	**95.4**	72.8	94.7	**81.9**
KSDD2	97.9	91.2	99.1	85.6	99.4	92.7	96.1	70.8	98.9	95.6	**99.6**	**98.5**
DAGM	83.2	55.4	99.1	93.6	97.0	40.9	99.3	98.0	97.4	90.3	**99.5**	**98.3**
RSDD	95.1	75.4	99.1	92.0	99.5	50.5	99.6	98.0	–	–	**99.8**	**99.0**
Mean	85.4	62.7	96.2	84.8	94.3	51.1	96.1	86.6	–	–	**96.9**	**92.5**

controls and ablations, we additionally report pixel-level AP and F1, averaged over categories.

Implementation details. CoEvoAD uses frozen CLIP ViT-L/14@336px (Radford et al., 2021), with images resized to 518×518. The prompt-bank scorer is trained for 30 epochs and frozen before prompt-rule search. Unless otherwise stated, we use $M = 3$ prompt groups, context length $L_c = 5$, state length $L_s = 5$, population size $N = 16$, $G = 5$ search generations, $K_e = 4$ elites, and $K = 3$ sampled partners.

4.2 Main Results

We compare CoEvoAD against representative CLIP-based ZSAD baselines (Tables 1 and 2) and against a matched class-name prompt control under the same scorer and protocol (Table 3). The two comparisons cover cross-paper benchmark context and within-paper attribution, respectively.

CoEvoAD separates normal and anomalous samples more clearly than the baselines on representative MVTec-AD categories (Figure 3).

CoEvoAD attains the highest mean image AUROC/AP (95.0/96.2) and pixel AUROC/PRO (96.9/92.5) among the compared CLIP-based ZSAD baselines (Tables 1 and 2), with the largest image gain on RSDD. Table 3 reports the matched-protocol comparison: CoEvoAD improves the class-name control on both primary transfer directions and external industrial targets, yielding consistent pixel-level AP and F1 gains.

Table 3: Matched comparison. Both systems use the same frozen source-trained scorer and protocol; the control uses class-name prompts without rule search. $\Delta = \text{CoEvoAD} - \text{Control}$.

Target	Metric	Control	CoEvoAD	Δ
MVTec-AD	Image AUROC ↑	93.01	93.39	**+0.38**
	Pixel AP ↑	47.48	47.72	**+0.24**
	Pixel F1 ↑	47.69	47.96	**+0.27**
VisA	Image AUROC ↑	87.05	87.43	**+0.38**
	Pixel AP ↑	30.51	31.11	**+0.60**
	Pixel F1 ↑	36.21	36.72	**+0.51**
BTAD	Image AUROC ↑	92.81	92.68	−0.13
	Pixel AP ↑	41.05	41.66	**+0.61**
	Pixel F1 ↑	44.69	46.25	**+1.56**
RSDD	Image AUROC ↑	97.69	97.62	−0.07
	Pixel AP ↑	46.39	49.30	**+2.91**
	Pixel F1 ↑	49.11	50.10	**+0.99**

Robustness across categories. Table 4 tests whether the matched-control pixel gains are driven

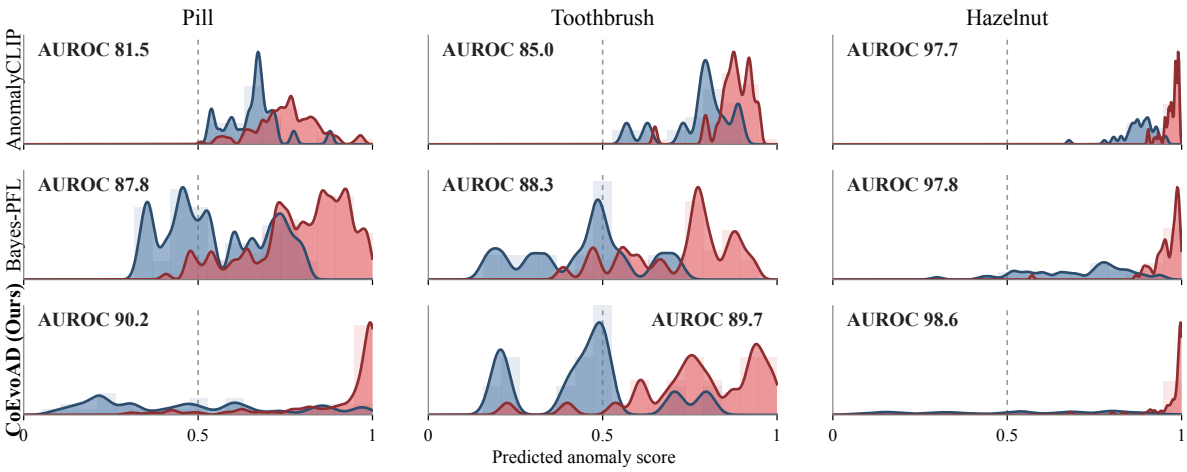

Figure 3: Per-image anomaly score distributions on three representative MVTec-AD categories (Pill, Toothbrush, Hazelnut) under the VisA→MVTec-AD matched protocol. Rows compare AnomalyCLIP, Bayes-PFL, and CoEvoAD; blue/red denote normal/abnormal samples; per-cell AUROC is shown.

Table 4: Per-category Δ statistics (mean, bottom-k, worst) vs the matched control.

Dir.	Metric	Mean Δ	Bot-k Δ	Worst Δ
MVTec	Pixel AP↑	+0.24	**+0.51**	−0.48
	Pixel F1↑	+0.27	**+0.45**	−0.41
VisA	Pixel AP↑	+0.60	**+0.78**	−3.27
	Pixel F1↑	+0.51	**+0.77**	−3.37

Table 5: Same-budget search controls. R0 is the class-name reference; search rows report percentage-point changes over R0.

	MVTec-AD		**VisA**	
Method	Image AUROC ↑	Pixel AP ↑	Image AUROC ↑	Pixel AP ↑
R0 class-name	93.01	47.48	87.05	30.51
Random search	+0.08	−0.01	−0.95	+0.52
Single-pop. EA	+0.14	−0.05	−0.24	+0.07
GA crossover	+0.17	−0.01	−0.23	**+0.74**
Role-sep. CoEvo	+0.05	+0.04	−0.03	+0.47
CoEvoAD	**+0.38**	**+0.24**	**+0.38**	+0.60

by high-baseline categories alone. Bottom-k Δ (k=3) exceeds the mean Δ in all four rows, indicating that the average lift is not concentrated on easy high-baseline categories.

Same-budget search controls. CoEvoAD denotes role-separated co-evolution with CCTO. Random search and single-population variants do not consistently improve both transfer directions, suggesting that the gain is not explained by candidate budget alone. CoEvoAD gives the strongest overall balance across Image AUROC and Pixel AP, although the crossover control obtains the highest Pixel AP on MVTec-AD→VisA.

CoEvoAD produces responses more concentrated around annotated defect regions on representative MVTec-AD and VisA categories (Figure 4).

Table 6: Representative natural-language rules selected by CoEvoAD (the leading learned soft-context token is omitted for compactness).

Category	Normal rule	Abnormal rule
screw	clean surface screw	broken shape with missing parts
bottle	normal bottle	irregular visible surface bottle
grid	regular standard grid	faulty flawed abnormal grid
toothbrush	a photo of a healthy toothbrush	faulty toothbrush
pcb2	flawless fine pcb2	structural defect like a hole or cut
cashew	a photo of healthy cashew	faulty flawed cashew
candle	clean intact candle	visible anomaly on object

4.3 Rule Inspection

Normal rules retain explicit category anchors, while abnormal rules combine category-specific defect cues with general defect descriptors (Table 6), consistent with CCTO's role-aware selection. For instance, abnormal rules name concrete defect modes for structured objects (*e.g.*, "structural defect like a hole or cut" for `pcb2`; "broken shape with missing parts" for `screw`), while less structured categories (*e.g.*, `candle`) retain generic descriptors like "visible anomaly on object". The selected rules are interpretable text and

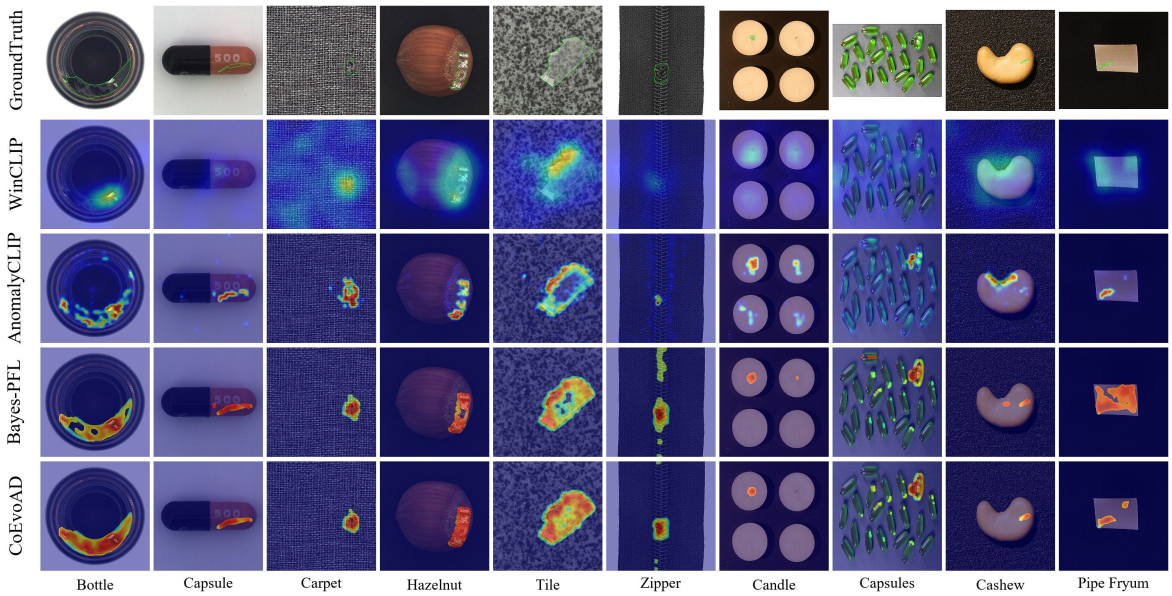

Figure 4: Qualitative localization examples on MVTec-AD and VisA. Columns show representative target categories; rows compare ground-truth contours and anomaly maps from CLIP-based baselines and CoEvoAD. Green contours denote annotated anomaly regions; warmer colors indicate higher anomaly scores.

Table 7: Component ablation comparing matched class-name prompt control, role-separated co-evolution (CoEvo), and CoEvoAD (CoEvo + CCTO). Bold marks the best score per row.

Target	Metric	Baseline	CoEvo	CoEvo +CCTO
MVTec-AD	Image AUROC↑	93.01	93.06	**93.39**
	Pixel AP↑	47.48	47.52	**47.72**
	Pixel F1↑	47.69	47.67	**47.96**
VisA	Image AUROC↑	87.05	87.02	**87.43**
	Pixel AP↑	30.51	30.98	**31.11**
	Pixel F1↑	36.21	36.33	**36.72**
BTAD	Image AUROC↑	92.81	**92.87**	92.68
	Pixel AP↑	41.05	40.85	**41.66**
	Pixel F1↑	44.69	45.22	**46.25**
RSDD	Image AUROC↑	97.69	**97.92**	97.62
	Pixel AP↑	46.39	43.37	**49.30**
	Pixel F1↑	49.11	46.69	**50.10**

can be edited, audited, and instantiated with another category name through the class-name slot, the same mechanism used by the rule-transfer policy (Section 3.2).

4.4 Ablation Studies

Component analysis. Role-separated co-evolution alone yields mixed effects on image-level AUROC, suggesting role separation is necessary but not sufficient. Adding CCTO produces consistent positive gains on pixel-level localization (Pixel AP and Pixel F1 improve over the class-name control across all four targets), with weaker image-level effects. This supports our claim that held-out source categories serve as a more transfer-oriented selection signal than performance on the optimization categories. In the same-budget search controls (Table 5), the non-role-separated baselines (Random / EA / GA) and role-separated co-evolution alone each fall short of the CoEvoAD pairing, which reaches +0.38 image AUROC under the matched protocol, indicating that the gain comes from the substrate–selection pairing rather than either component in isolation. This pattern reflects CCTO's role as a selection criterion rather than an additive scoring term.

5 Conclusion

We presented CoEvoAD, a framework that searches discrete, interpretable normal and abnormal prompt rules in natural language and uses a Cross-Category Transfer Objective to favor rules generalizing across unseen target categories. Across six industrial anomaly detection benchmarks, CoEvoAD attains the highest mean image-level and pixel-level scores among compared CLIP-based ZSAD baselines, yielding positive matched-control gains attributable to prompt-rule search. The framework shows that interpretable prompt search is feasible for ZSAD via role-separated co-evolution guided by held-out cross-category signals. The saved rules remain auditable as natural-language strings, complementing continuous prompt embeddings.

Limitations

The benefits of CoEvoAD concentrate on pixel-level localization rather than image-level AUROC: matched-control gains in image AUROC are modest and not uniformly positive, and CCTO gains vary across target datasets. The MVTec-AD→VisA direction shows a -3.27 pp worst-category pixel-AP Δ on `pcb4`, so the lift is not uniform across all categories. CoEvoAD also inherits the limitations of the underlying CLIP backbone for defects that are not well described by natural language.

Ethics Statement

This work targets zero-shot anomaly detection for industrial visual inspection. We use public benchmark datasets and do not collect human-subject data or personally identifiable information. The method is intended to assist defect screening rather than replace human inspection in safety-critical production. Since anomaly detectors may produce false positives or false negatives under distribution shift, practical deployment should require human oversight, site-specific validation, and continuous monitoring.

Although CoEvoAD searches interpretable natural-language prompt rules, the selected rules may still reflect biases of the source datasets and the pretrained vision-language model. We therefore recommend auditing selected rules before deployment and avoiding deployment in high-stakes settings without further validation.

References

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. MVTec AD—a comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9592–9600.

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

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. 2024. AdaCLIP: Adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. 2021. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition Workshops (ICPR Workshops)*, pages 475–489.

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

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P. Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391.

Qingqing Fang, Wenxi Lv, and Qinliang Su. 2025. AF-CLIP: Zero-shot anomaly detection via anomaly-focused CLIP adaptation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4846–4855.

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

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *International Conference on Learning Representations*.

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

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. 2023. SimpleNet: A simple network for image anomaly detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20402–20411.

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

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

Menghui Niu, Kechen Song, Liming Huang, Qi Wang, Yunhui Yan, and Qinggang Meng. 2021. Unsupervised saliency detection of rail surface defects using stereoscopic images. *IEEE Transactions on Industrial Informatics*, 17(3):2271–2281.

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

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

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. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763.

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

Taylor Shin, Yasaman Razeghi, Robert L. Logan, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4222–4235.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International Conference on Machine Learning*, pages 20841–20855.

Matthias Wieler and Tobias Hahn. 2007. Weakly supervised learning for industrial optical inspection. DAGM 2007 Competition Dataset.

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

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. 2022. A unified model for multi-class anomaly detection. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. 2021. DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8330–8339.

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

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

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

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

A Experimental Protocol and Dataset Details

Datasets. We evaluate CoEvoAD on six industrial anomaly detection benchmarks. MVTec-AD (Bergmann et al., 2019) contains 15 object/texture categories with pixel-level defect masks; VisA (Zou et al., 2022) provides 12 categories of small electronics and food items; BTAD (Mishra et al., 2021) covers three industrial product categories; KSDD2 (Božič et al., 2021) and RSDD (Niu et al., 2021) focus on surface-defect inspection on metal and rail-surface, respectively; DAGM (Wieler and Hahn, 2007) consists of synthetic texture patches. Two additional texture-oriented external targets, DTD-Synthetic and a 7-class DAGM subset, are reported only as diagnostic transfers (Section E). All datasets are publicly available and we use their official train/test splits without modification.

Cross-dataset zero-shot setting. We evaluate two primary transfer directions: VisA→MVTec-AD and MVTec-AD→VisA. We refer to these directions by their target dataset name, namely *MVTec* (for VisA→MVTec-AD) and *VisA* (for MVTec-AD→VisA), when context is unambiguous. For external industrial targets (BTAD, KSDD2, RSDD, DAGM, DTD-Synthetic), we use two source conventions tied to the reporting purpose. The VisA-source convention applies a VisA-trained checkpoint to non-VisA targets and is used in the headline comparison, matching prior CLIP-based ZSAD baselines (Zhou et al., 2024; Qu et al., 2025). The MVTec-AD-source convention applies an MVTec-AD-trained checkpoint to non-MVTec-AD targets and is used in the matched-ablation tables, so that the control and CoEvoAD share a single frozen scorer with the MVTec-AD→VisA primary direction. For the VisA target, an MVTec-AD-trained checkpoint is used in both settings. Source category labels are used during source-side prompt-bank training and prompt-rule search; target category names are used only as a frozen source-only transfer-policy key. No target-domain image, label, mask, statistic, or metric ever enters training, validation, calibration, prompt search, or model selection. Target data are observed only at final inference and final evaluation.

Table 8: Full implementation and hyperparameters.

Component	Setting
Backbone	CLIP ViT-L/14@336px
Input resolution	518×518
Stage 1 epochs / optimiser	30 / AdamW, cosine annealing
Stage 1 learning rates	prompt 1×10^{-3}, other 1×10^{-4}
Prompt groups M / context L_c / state L_s	3 / 5 / 5
Search population N	16
Search generations G	5
Search elite count K_e	4
Search pairing count K	3
CCTO bottom-k	3
CCTO shared-role coefficient α_{ccto}	0.6
CCTO role-aware coefficients λ_n / λ_a	0.35 / 0.20
Fitness weights α / β	0.85 / 0.15
Routing (MVTec / VisA)	semantic fallback / template transfer
Semantic-fallback template, min-sim	`a photo of {}`, 0.45
Seeds	111, 222, 333
Hardware	single NVIDIA RTX 4090

Table 9: Ablation of Stage 1 training components on the class-name control. Each row removes one component from the fixed scorer training.

Variant	AUROC	AP_{px}
Full training setup (control)	93.01	47.48
− mask-guided crop	92.94 (−0.07)	46.35 (−1.13)
− inter-role margin	92.83 (−0.18)	47.04 (−0.44)
− agnostic reg.	93.01 (±0.00)	47.09 (−0.39)

B Implementation Details

Table 8 lists the full configuration used for every CoEvoAD row in the main paper, shared across both transfer directions unless stated otherwise.

Stage 1 training design. The prompt-conditioned scorer applies three training components on top of the base classification + segmentation losses. *Mask-guided crop augmentation* samples crops biased toward annotated defect regions during training, ensuring the pixel-level head sees fine-grained defect patterns. *Inter-role margin regularization* adds a hinge-style term that pushes apart the per-image average embeddings of the normal and abnormal prompt groups, preventing the two roles from collapsing into a shared direction. *Category-agnostic regularization* encourages the prompt-group representations to be transferable across categories by penalizing per-category specialization on the source set. Their isolated effects on VisA→MVTec-AD are quantified in Table 9.

C Prompt-Rule Search Details

Search algorithm. The rule search follows Algorithm 1, applied independently for each source category $c \in \mathcal{C}_{\text{src}}$ with hyperparameters listed in

Table 10: Rule-search cost. $|\mathcal{C}_{\text{src}}|$: number of source categories. Evals $= N \times G \times |\mathcal{C}_{\text{src}}|$.

Dir.	$\lvert\mathcal{C}_{\text{src}}\rvert$	N	G	K	Evals	Wall-clock
MVTec (src VisA)	12	16	5	3	960	∼3 h 47 m
VisA (src MVTec-AD)	15	16	5	3	1200	∼5 h 34 m

Table 8.

Search cost. Table 10 reports the rule-search budget and observed wall-clock. The number of prompt-pair evaluations is $N \times G \times |\mathcal{C}_{\text{src}}|$ plus elite carry-over; the search is per source category and scales with the source category count. Wall-clock is the observed search span from the log of one representative seed on a single RTX 4090.

Same-budget controls with CCTO-augmented variants. Table 11 extends the main paper same-budget control table (Table 5) with three CCTO-augmented variants that pair CCTO selection with non-role-separated search backbones (random prompt-pair search, single-population evolutionary algorithm, and genetic algorithm with crossover). Adding CCTO on top of these non-role-separated backbones does not consistently improve over the corresponding backbone-only rows, indicating that CCTO's selection benefit is coupled to the role-separated co-evolution substrate rather than functioning as a generic add-on. Values are absolute scores; the budget and frozen-scorer setting are identical to the main paper version.

D Cross-Category Transfer Objective Details

Bottom-k aggregation. Let $\mathcal{C}_{\text{src}}$ be the source category set used during prompt-bank training, and let $c \in \mathcal{C}_{\text{src}}$ be the category currently being optimized. The held-out source set is $\mathcal{C}_{\text{src}} \setminus \{c\}$. For a candidate pair (p^n, p^a), we compute its in-category fitness $\eta^r(p, c)$ on c and its cross-category fitness $\chi^r(p, c')$ on each held-out $c' \in \mathcal{C}_{\text{src}} \setminus \{c\}$. We aggregate per-$c'$ scores with the bottom-k operator described in Sec. 3.3, using k=3 given the small source-category counts ($|\mathcal{C}_{\text{src}}| = 12$ for VisA→MVTec-AD, 15 for MVTec-AD→VisA).

Role-aware CCTO scoring. The aggregated CCTO score is combined with the in-category score through role-specific weights λ^n, λ^a. We use two scoring configurations, both selected once on a source-only validation split. A shared-role configuration uses a single coefficient α_{ccto}=0.6 for both roles. A role-aware configuration uses λ_n=0.35 for the normal role and λ_a=0.20 for the abnormal role, motivated by the empirical observation that the abnormal vocabulary tends to transfer more uniformly across categories than the normal vocabulary.

E Complete Quantitative Results

Matched-control details. Per-category 95% confidence intervals on the matched-control gains in Table 3 exclude zero on VisA→MVTec-AD image AUROC ($[+0.10, +0.66]$) and pixel F1 ($[+0.06, +0.47]$), but include zero on MVTec-AD→VisA (dominated by `pipe_fryum` and `pcb4`, see Table 18).

Per-category results. Tables 12 and 13 report every target category for the two primary transfer directions, comparing the matched control (class-name prompts, no evolved rules) against CoEvoAD. The breakdown shows the gain is concentrated in a subset of categories and is not uniform: in the VisA→MVTec-AD direction, image AUROC improves on 8/15 categories and decreases on 2; in the MVTec-AD→VisA direction it improves on 7/12 and decreases on 5.

Multi-seed stability. Table 14 reports the per-seed results and the mean and standard deviation over three random seeds for the two main transfer directions, following the search and transfer settings in Sec. 4.1.

External industrial transfer. Table 15 reports the percentage-point deltas of CoEvoAD over the matched baseline on five external targets under both MVTec-AD and VisA source conventions. The MVTec-source route gives consistent pixel AP and F1 gains while staying image-neutral on average; the VisA-source route is image-neutral and slightly negative on pixel AP and F1, so it must not be cited as a pixel-improvement result.

Texture-oriented external targets. We also examine two texture-oriented external targets, DAGM and DTD-Synthetic. Table 16 reports image-level AUROC. On DTD-Synthetic (standard 12-class split), CoEvoAD remains close to AnomalyCLIP but trails Bayes-PFL by about two percentage points. On DAGM (evaluated on the 7-class subset used in our protocol), CoEvoAD is the strongest among methods reproduced on the same subset; AdaCLIP and MRAD numbers are taken from their original publications under the 10-class stan-

Table 11: Extended same-budget prompt-search control table with CCTO-augmented variants. All search rows use the same frozen scorer and candidate budget; R0 is the non-search class-name reference; seed 111.

Method	Role-sep.	CCTO	MVTec-AD		VisA	
			$AUROC_{img}$	AP_{px}	$AUROC_{img}$	AP_{px}
Class-name control (R0)	✗	✗	93.01	47.48	87.05	30.51
Random prompt-pair search	✗	✗	93.09	47.47	86.10	31.03
Random prompt-pair search + CCTO	✗	✓	93.23	47.45	86.90	30.80
Single-population evolutionary algorithm	✗	✗	93.15	47.43	86.81	30.58
Single-population evolutionary algorithm + CCTO	✗	✓	93.24	47.54	86.90	30.62
Genetic algorithm with crossover	✗	✗	93.18	47.47	86.82	**31.25**
Genetic algorithm with crossover + CCTO	✗	✓	93.19	47.45	86.95	30.35
Role-separated co-evolution	✓	✗	93.06	47.52	87.02	30.98
CoEvoAD (CoEvo+CCTO)	✓	✓	**93.39**	**47.72**	**87.43**	31.11

Table 12: Per-category breakdown for the MVTec direction (VisA→MVTec-AD). C: strict class-name control. E: CoEvoAD. Δ =E−C (pp). Image AUROC / pixel AP / pixel F1.

Category	$AUROC_{img}$			AP_{px}			$F1_{px}$		
	C	E	Δ	C	E	Δ	C	E	Δ
bottle	94.76	94.84	+0.08	67.19	66.97	−0.22	63.09	62.92	−0.17
cable	82.27	83.36	+1.09	12.49	12.01	−0.48	20.09	19.98	−0.11
capsule	94.50	94.30	−0.20	36.72	36.67	−0.05	39.85	39.44	−0.41
carpet	100.00	100.00	+0.00	82.69	82.98	+0.29	74.93	75.22	+0.29
grid	100.00	100.00	+0.00	41.38	40.97	−0.41	42.70	42.75	+0.05
hazelnut	97.96	97.96	+0.00	58.13	58.78	+0.65	55.08	55.49	+0.41
leather	100.00	100.00	+0.00	61.53	61.40	−0.13	57.56	57.62	+0.06
metal_nut	74.93	75.61	+0.68	28.65	29.23	+0.58	36.20	36.89	+0.69
pill	89.12	89.06	−0.06	30.39	30.62	+0.23	34.19	34.46	+0.27
screw	90.55	91.29	+0.74	46.18	47.29	+1.11	48.61	49.38	+0.77
tile	99.39	99.39	+0.00	79.69	79.81	+0.12	72.72	72.89	+0.17
toothbrush	93.06	94.72	+1.66	20.98	22.44	+1.46	24.10	25.26	+1.16
transistor	81.96	83.17	+1.21	11.19	11.75	+0.56	17.17	17.47	+0.30
wood	97.28	97.63	+0.35	68.56	68.32	−0.24	63.81	63.71	−0.10
zipper	99.37	99.50	+0.13	66.46	66.63	+0.17	65.30	65.89	+0.59
Mean	93.01	93.39	+0.38	47.48	47.72	+0.24	47.69	47.96	+0.27

Table 13: Per-category breakdown for the MVTec-AD→VisA direction. C: strict class-name control. E: CoEvoAD (MVTec-AD→VisA main config). Δ =E−C (pp). The MVTec-AD→VisA gain is dominated by a few categories (`pipe_fryum` +4.05 AP_{px}, `pcb4` −3.27 AP_{px}).

Category	$AUROC_{img}$			AP_{px}			$F1_{px}$		
	C	E	Δ	C	E	Δ	C	E	Δ
candle	90.02	90.67	+0.65	33.69	33.53	−0.16	44.08	43.55	−0.53
capsules	92.95	91.95	−1.00	49.56	52.27	+2.71	54.92	55.78	+0.86
cashew	93.84	93.36	−0.48	36.35	37.40	+1.05	42.16	42.68	+0.52
chewinggum	97.38	97.00	−0.38	83.18	83.12	−0.06	75.45	75.53	+0.08
fryum	90.68	90.92	+0.24	27.95	27.60	−0.35	33.19	32.92	−0.27
macaroni1	91.61	91.27	−0.34	26.18	26.42	+0.24	34.80	35.99	+1.19
macaroni2	67.43	68.37	+0.94	3.17	2.75	−0.42	9.64	8.56	−1.08
pcb1	79.36	80.08	+0.72	10.40	11.91	+1.51	17.75	19.64	+1.89
pcb2	76.81	77.81	+1.00	13.03	14.28	+1.25	21.64	23.13	+1.49
pcb3	77.09	77.56	+0.47	19.97	20.60	+0.63	26.41	27.72	+1.31
pcb4	89.87	92.83	+2.96	28.09	24.82	−3.27	33.09	29.72	−3.37
pipe_fryum	97.60	97.28	−0.32	34.59	38.64	+4.05	41.41	45.44	+4.03
Mean	87.05	87.43	+0.38	30.51	31.11	+0.60	36.21	36.72	+0.51

dard split. These results are consistent with a boundary of category-keyed prompt transfer: ab-

Table 14: Multi-seed stability over seeds 111, 222, and 333. We report per-seed results together with mean and standard deviation under the matched protocol.

	VisA→MVTec-AD		MVTec-AD→VisA	
	$AUROC_{img}$	AP_{px}	$AUROC_{img}$	AP_{px}
Seed 111	93.39	47.72	87.43	31.11
Seed 222	93.13	47.37	87.05	30.79
Seed 333	93.26	47.45	87.46	31.19
Mean±std	93.26 ±0.13	47.51 ±0.18	87.31 ±0.23	31.03 ±0.21

stract texture categories have weaker semantic overlap with the object-defect vocabulary used by the source-side prompt rules. Table 17 drills down to per-class DTD-Synthetic results, showing the aggregate gap is concentrated on two categories whose names have low semantic similarity with any source object-defect class.

F Additional Ablations

Paired category-level significance. Table 18 reports, per direction and metric, the mean per-category Δ, a normal-approximation 95% confidence interval over categories, and the improved/degraded/tied counts. The VisA→MVTec-AD image-AUROC interval excludes zero; the MVTec-AD→VisA image-AUROC interval includes zero, so the MVTec-AD→VisA image-level gain is not distinguishable from category-level noise. This is consistent with the Limitations statement that the strongest evidence is the direction-sensitive ablation and pixel-side robustness, not a uniform lift.

Test-time transfer policy. Table 19 reports the test-time transfer ablation, separating nearest-source rule transfer (Sem) from template transfer (Tmpl).

G Prompt-Rule Interpretability Analysis

Selected prompt rules per source category. Tables 20 and 21 list the evolved prompt-rule pair selected per source category by the locked rule search. "X" denotes the learned soft-context slot of the prompt bank (the prompt to CLIP is hybrid: a learned latent context plus the interpretable rule string), so these are interpretable *rule strings*, not a fully interpretable system. The default column is the class-name baseline the search starts from; the evolved column is the saved, reloadable selection. The search is mutation-only (no recombination).

Target transfer-policy map. Tables 22 and 23 give the transfer-policy decision for every target category. Each target category is mapped to its nearest source category by text-embedding cosine similarity, and the target category name is the only target-side input.

Table 22: VisA→MVTec-AD transfer policy: MVTec-AD target → VisA source. Sim = cosine similarity; Tag = policy tag.

Target	Nearest source	Sim	Route tag
bottle	candle	0.810	semantic (abn. fallback)
cable	macaroni1	0.758	semantic transfer
capsule	capsules	0.903	semantic transfer
carpet	fryum	0.799	semantic transfer
grid	fryum	0.804	semantic transfer
hazelnut	cashew	0.848	semantic transfer
leather	fryum	0.749	semantic transfer
metal_nut	cashew	0.794	semantic transfer
pill	macaroni1	0.826	semantic transfer
screw	macaroni1	0.759	semantic transfer
tile	fryum	0.791	semantic transfer
toothbrush	candle	0.786	semantic (abn. fallback)
transistor	fryum	0.736	semantic transfer
wood	fryum	0.786	semantic transfer
zipper	macaroni1	0.816	semantic transfer

Table 23: MVTec-AD→VisA transfer policy: VisA target → MVTec-AD source.

Target	Nearest source	Sim	Route tag
candle	bottle	0.810	semantic transfer
capsules	capsule	0.903	semantic transfer
cashew	hazelnut	0.848	semantic transfer
chewinggum	toothbrush	0.716	semantic transfer
fryum	pill	0.819	semantic transfer
macaroni1	pill	0.826	semantic transfer
macaroni2	pill	0.806	semantic transfer
pcb1	pill	0.770	semantic transfer
pcb2	pill	0.775	semantic transfer
pcb3	pill	0.756	semantic transfer
pcb4	pill	0.764	semantic transfer
pipe_fryum	pill	0.770	semantic transfer

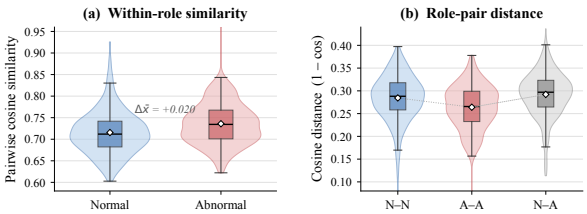

Figure A1: Analysis of selected rule embeddings. We report within-role cosine similarity and cross-role distance for the text embeddings of selected normal and abnormal rules.

Prompt embedding analysis.

Table 15: External industrial benchmarks, all five targets, both source conventions. $\text{AUROC}_{\text{img}}$: control→CoEvoAD (%). Δ columns: CoEvoAD − matched strict control (pp).

Source	Target	$\text{AUROC}_{\text{img}}$ C→E	$\Delta\text{AUROC}_{\text{px}}$	ΔAUPRO	$\Delta\text{AP}_{\text{px}}$	$\Delta\text{F1}_{\text{px}}$
MVTec	BTAD	92.81→92.68	+0.45	+0.47	+0.61	+1.56
MVTec	DAGM	93.82→93.56	−0.12	−0.46	+0.81	+0.35
MVTec	DTD	93.77→93.99	+0.12	+0.25	+0.04	+0.32
MVTec	KSDD2	96.04→95.94	+0.01	+0.09	+0.14	+0.21
MVTec	RSDD	97.69→97.62	+0.02	+0.10	+2.91	+0.99
MVTec	Mean	94.83→94.76	+0.10	+0.09	+0.90	+0.69
VisA	BTAD	94.25→94.40	−0.02	—	+0.04	+0.08
VisA	DAGM	98.17→98.19	+0.00	—	−0.39	−0.32
VisA	DTD	94.25→94.25	+0.02	—	−0.66	−0.55
VisA	KSDD2	97.42→97.44	−0.01	—	−0.55	−0.35
VisA	RSDD	98.89→98.86	−0.01	—	+0.27	−0.19
VisA	Mean	96.60→96.63	+0.00	—	−0.26	−0.27

Table 16: Image-level AUROC (%) on texture-oriented external targets. † denotes evaluation on our 7-class DAGM subset; unmarked DAGM numbers follow the standard 10-class reporting convention.

Method	DAGM	DTD-Synthetic
Same 7-class DAGM subset		
AnomalyCLIP†	96.3	94.3
Bayes-PFL†	97.35	**96.28**
CoEvoAD	**98.19**	94.25
Original publication reports		
WinCLIP	87.6	83.9
AdaCLIP	99.1	95.5
MRAD	98.4	96.0

Table 17: Per-class image-level AUROC (%) on DTD-Synthetic. C: Bayes-PFL (matched VisA-source reproduction). E: CoEvoAD. Δ =E−C.

Category	C (Bayes-PFL)	E (CoEvoAD)	Δ
blotchy	94.06	75.75	−18.31
matted1	88.10	80.95	−7.15
marbled2	99.00	96.50	−2.50
fibrous	99.94	98.69	−1.25
mesh	91.15	92.64	+1.49
perforated	93.00	96.44	+3.44
stratified	99.81	99.69	−0.12
woven1	92.48	93.23	+0.75
woven2	100.00	100.00	±0.00
woven3	99.25	98.31	−0.94
woven4	98.51	98.88	+0.37
woven5	100.00	99.90	−0.10
Mean	96.28	94.25	−2.03

Table 18: Paired category-level sign test. Δ = CoEvoAD − matched control, averaged over categories (n=15 for the VisA→MVTec-AD direction, n=12 for the MVTec-AD→VisA direction). CI is the normal-approximation 95% interval of the per-category Δ. Differences of ±0.01 from Table 3 reflect different rounding paths (aggregate vs per-category).

Dir.	Metric	Mean Δ	95% CI	#Up	#Down	#Tie
MVTec	$\text{AUROC}_{\text{img}}$	+0.38	[+0.10, +0.66]	8	2	5
	AP_{px}	+0.24	[−0.03, +0.51]	9	6	0
	F1_{px}	+0.27	[+0.06, +0.47]	11	4	0
VisA	$\text{AUROC}_{\text{img}}$	+0.37	[−0.19, +0.93]	7	5	0
	AP_{px}	+0.60	[−0.38, +1.58]	7	5	0
	F1_{px}	+0.51	[−0.47, +1.49]	8	4	0

Table 19: Test-time transfer ablation. R0 is the class-name control; Sem/Tmpl denote nearest-source/template transfer.

			MVTec-AD		**VisA**	
Route	Sem	Tmpl	$\text{AUROC}_{\text{img}}$	AP_{px}	$\text{AUROC}_{\text{img}}$	AP_{px}
R0 (control)	✗	✗	93.01	47.48	87.05	30.51
R1 (sem-only)	✓	✗	**93.39**	**47.72**	**87.21**	30.85
R2 (tmpl-only)	✗	✓	93.13	47.49	87.11	**31.51**
R3 (both)	✓	✓	**93.39**	**47.72**	**87.21**	30.85

H Additional Qualitative Results and Failure Cases

Qualitative localization. The main paper (Figure 4) compares anomaly maps for a subset of MVTec-AD categories. For completeness, figs. A2 to A21 report the full per-category anomaly localization produced by CoEvoAD on the two primary cross-dataset directions (MVTec, VisA) and the four external industrial targets (BTAD, DAGM, KSDD2, RSDD). In every figure, the top row shows input images with ground-truth defect contours overlaid in green, and the bottom row shows the CoEvoAD anomaly map (warmer colors indicate higher anomaly scores). All maps are produced by the locked rule set.

Failure modes. Three failure modes are visible across the per-category breakdowns and inform the boundaries of the method.

(i) Texture-only targets with weak object-defect overlap. On DTD-Synthetic, two categories

Table 20: Evolved prompt rules per VisA source category, for the VisA→MVTec-AD direction. Default = `X normal/abnormal {cat}`.

Src cat.	Evolved normal rule	Evolved abnormal rule
candle	X clean intact candle	X visible anomaly on object
capsules	X a photo of a capsules	X plain damaged capsules
cashew	X a photo of healthy cashew	X faulty flawed cashew
chewinggum	X typical intact with chewinggum	X anomalous abnormal textural chewinggum
fryum	X a view of pristine good fryum	X item: defect in fryum
macaroni1	X a picture of macaroni1	X defective visible textural macaroni1
macaroni2	X perfect macaroni2	X object: in macaroni2
pcb1	X typical flawless perfect pcb1	X item: irregular pcb1
pcb2	X flawless fine pcb2	X structural defect like a hole or cut
pcb3	X clean regular pcb3	X a view of compromised imperfect pcb3
pcb4	X a photo of a pcb4	X faulty flawed deteriorated pcb4
pipe_fryum	X object: flawless healthy pipe_fryum	X object: faulty pipe_fryum

Table 21: Evolved prompt rules per MVTec-AD source category, for the MVTec-AD→VisA direction. Default = `X normal/abnormal {cat}`.

Src cat.	Evolved normal rule	Evolved abnormal rule
bottle	X normal bottle	X irregular visible surface bottle
cable	X typical cable	X an anomalous cable sample
capsule	X a standard capsule product	X a deformed structure capsule sample
carpet	X pristine surface carpet	X faulty blemished carpet
grid	X regular standard grid	X faulty flawed abnormal grid
hazelnut	X hazelnut	X abnormal faulty flawed hazelnut
leather	X perfect leather	X faulty damaged region with structural leather
metal_nut	X perfect metal_nut	X broken faulty compromised metal_nut
pill	X pill	X item: irregular deteriorated pill
screw	X clean surface screw	X broken shape with missing parts
tile	X the normal tile	X plain flawed tile
toothbrush	X a photo of a healthy toothbrush	X faulty toothbrush
transistor	X clean transistor	X defect in defective irregular transistor
wood	X standard good wood	X defective irregular wood
zipper	X regular zipper	X defective bad deformed zipper

(`blotchy`, `matted1`) drive the entire aggregate gap (Table 17). Their category names are abstract texture descriptors with no clear semantic counterpart in the source object-defect vocabulary used during rule search, so neither nearest-source rule transfer nor the template fallback finds a strong source anchor.

(ii) PCB-style targets with high intra-category variability. On MVTec-AD→VisA the pixel-AP delta is dominated by `pipe_fryum` ($+4.05$ pp) and `pcb4` (-3.27 pp), and the per-category sign-test interval includes zero (Table 18). The two-direction asymmetry (VisA→MVTec-AD image-AUROC interval excludes zero; MVTec-AD→VisA image-AUROC interval does not) is consistent with this category-level instability and is reported honestly in the Limitations.

(iii) Image-saturated categories. On targets where the matched control already exceeds 99 image AUROC (*e.g.*, several MVTec textures), CoEvoAD has no remaining headroom and reports ± 0.00 pp. These rows are not failures of the search but reflect the ceiling under the current backbone.

These boundaries define the scope of the claim (see Limitations).

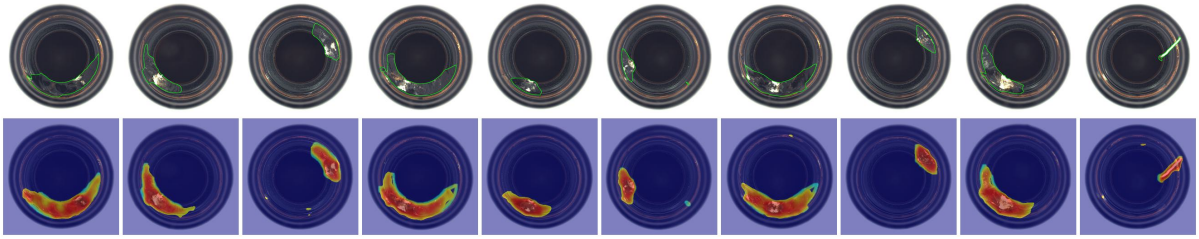

Figure A2: Qualitative localization on MVTec-AD *bottle*. Top: input with ground-truth contours; bottom: CoEvoAD anomaly map.

Figure A3: Qualitative localization on MVTec-AD *capsule*.

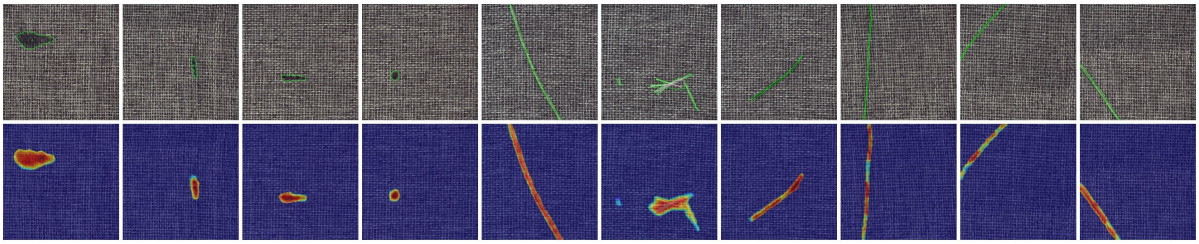

Figure A4: Qualitative localization on MVTec-AD *carpet*.

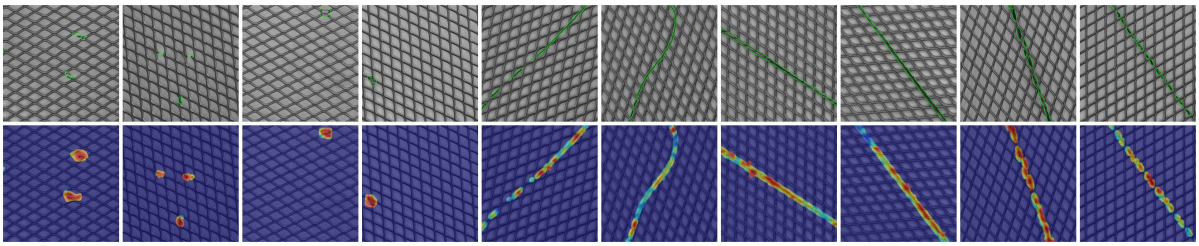

Figure A5: Qualitative localization on MVTec-AD *grid*.

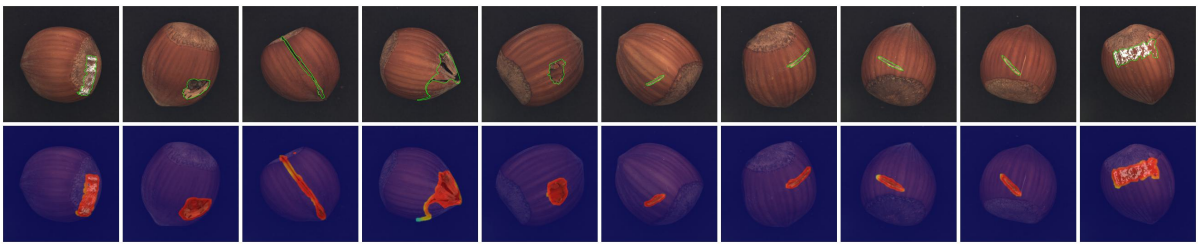

Figure A6: Qualitative localization on MVTec-AD *hazelnut*.

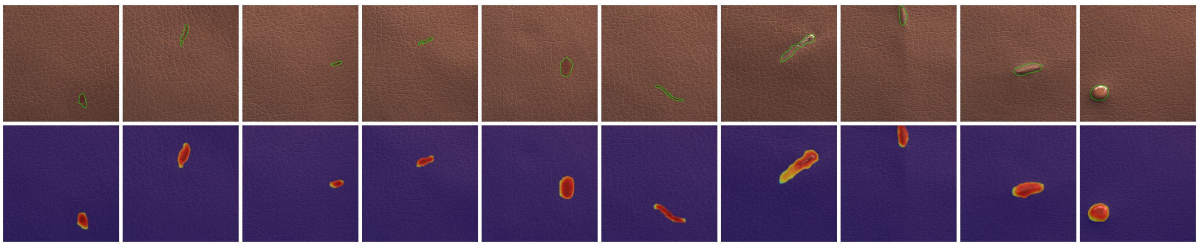

Figure A7: Qualitative localization on MVTec-AD *leather*.

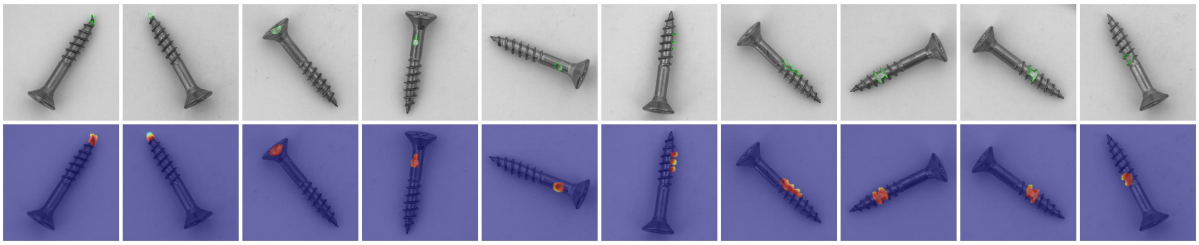

Figure A8: Qualitative localization on MVTec-AD *screw*.

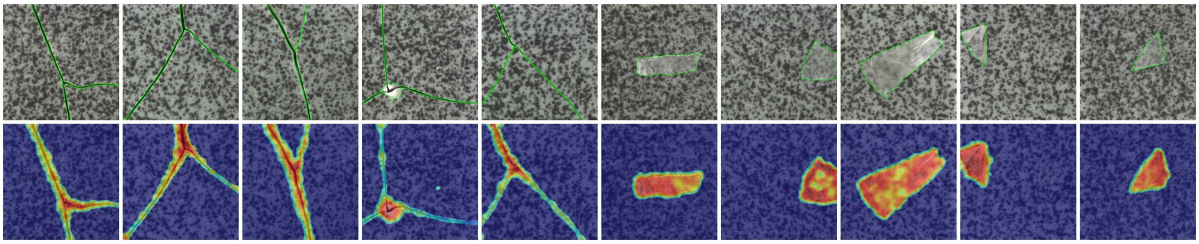

Figure A9: Qualitative localization on MVTec-AD *tile*.

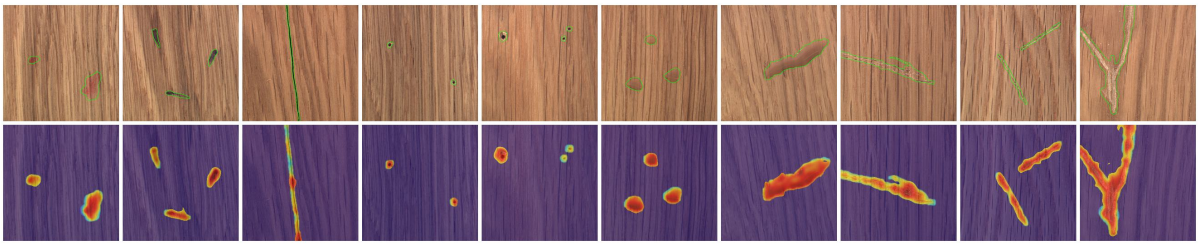

Figure A10: Qualitative localization on MVTec-AD *wood*.

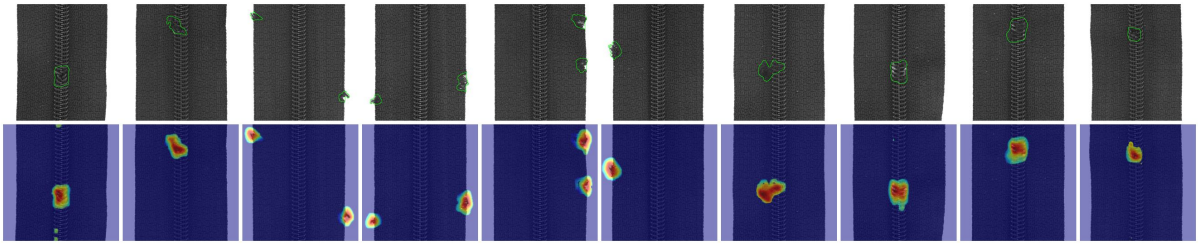

Figure A11: Qualitative localization on MVTec-AD *zipper*.

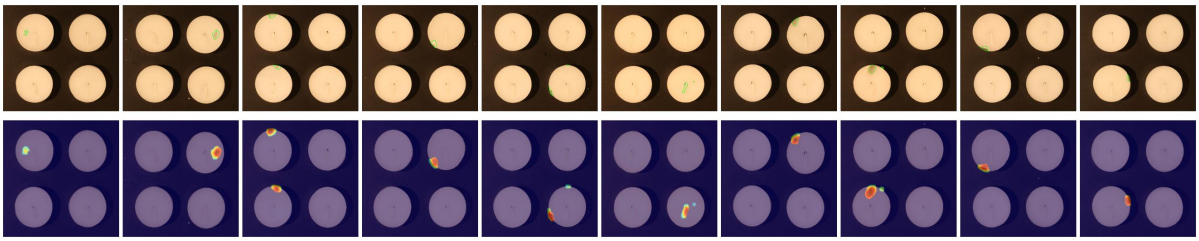

Figure A12: Qualitative localization on VisA *candle*.

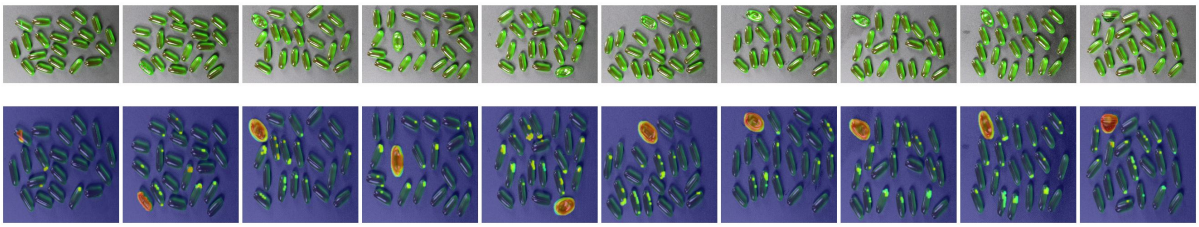

Figure A13: Qualitative localization on VisA *capsules*.

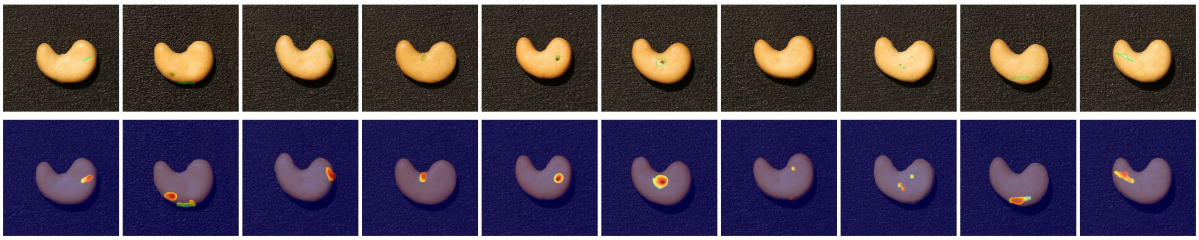

Figure A14: Qualitative localization on VisA *cashew*.

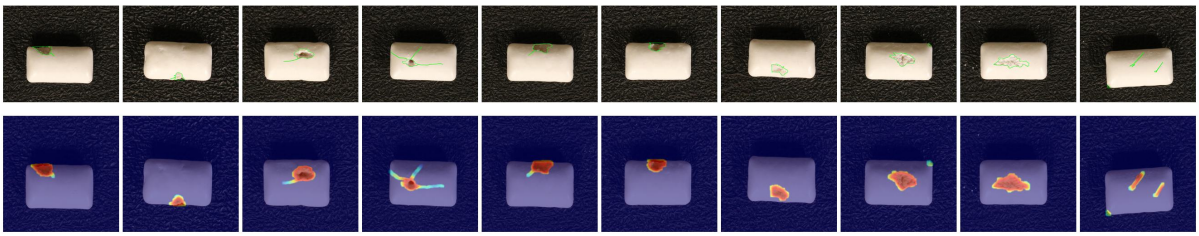

Figure A15: Qualitative localization on VisA *chewinggum*.

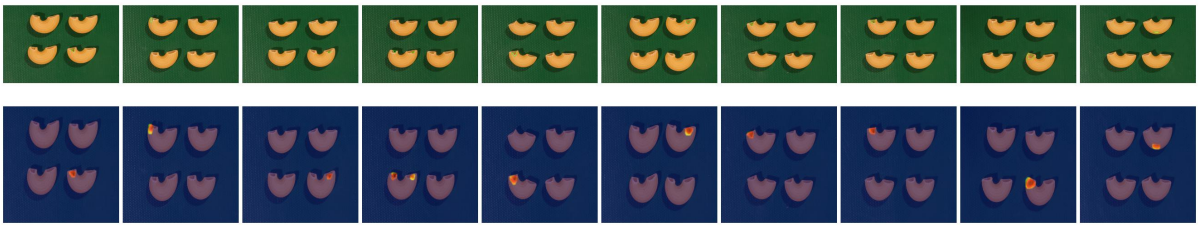

Figure A16: Qualitative localization on VisA *macaroni1*.

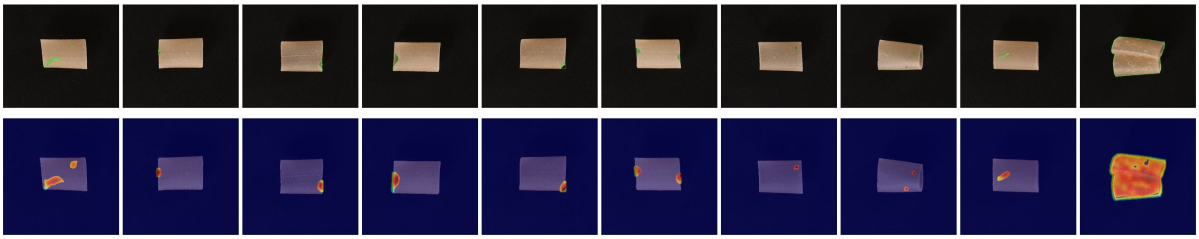

Figure A17: Qualitative localization on VisA *pipe_fryum*.

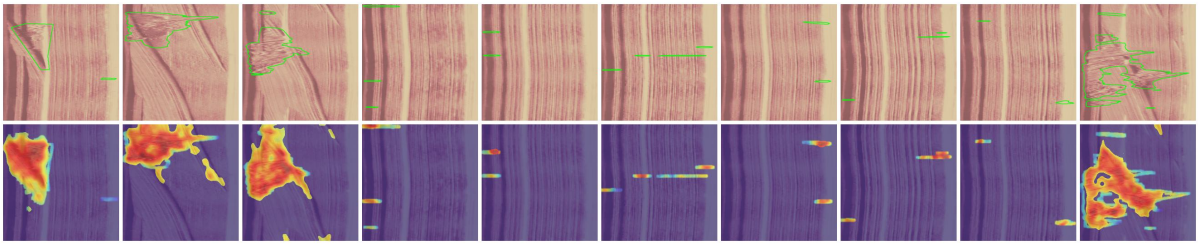

Figure A18: Qualitative localization on BTAD *wood*.

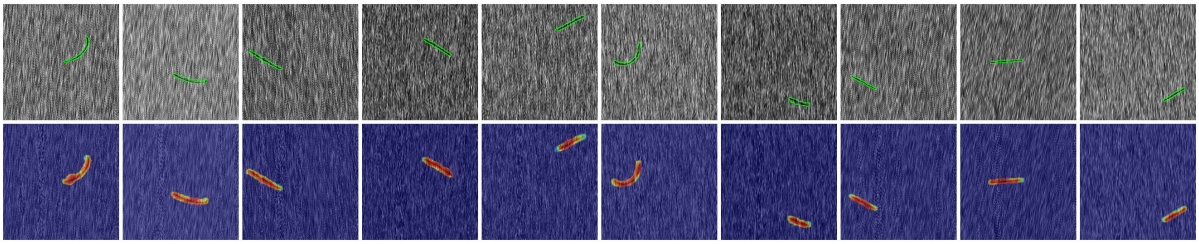

Figure A19: Qualitative localization on DAGM *fabric2*.

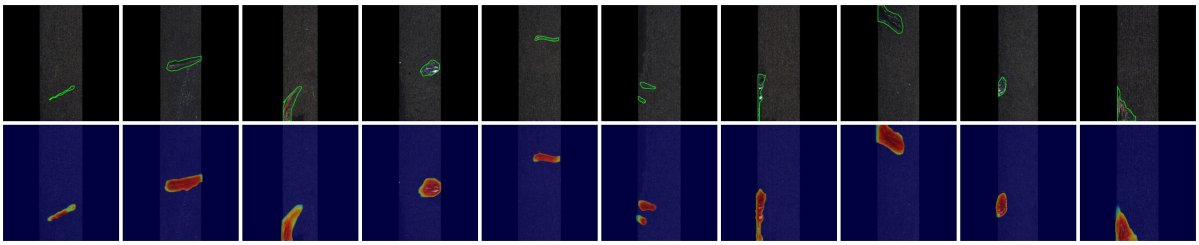

Figure A20: Qualitative localization on KSDD2 *metal*.

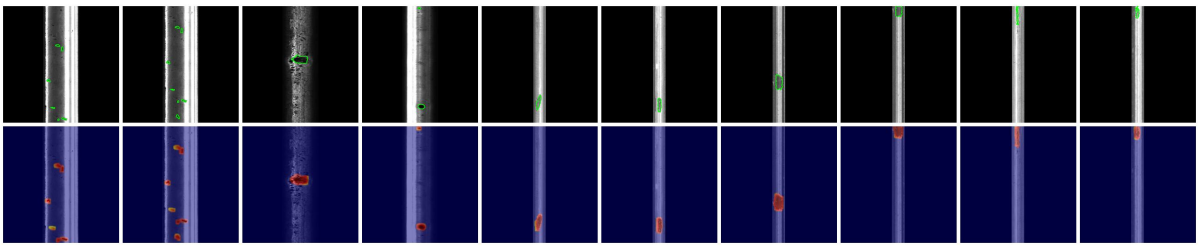

Figure A21: Qualitative localization on RSDD *metal15*.