SPECIFICATION-CONDITIONED COUNTERFACTUAL REPAIR FOR INDUSTRIAL ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Industrial anomaly detectors can achieve strong benchmark performance while relying on visual mechanisms that do not match the intended inspection logic. For example, a detector may respond strongly to a tolerated appearance change, overlook the specified defect, or depend on interactions between the two. We study this problem as post-training repair under an explicit inspection specification rather than as learning from a new set of anomaly labels. Our approach uses matched counterfactual probes to separate the detector's responses to the target defect, tolerated variation, and their interaction, and evaluates these responses against a user-defined behavioral contract. Based on this formulation, we introduce Specification-Conditioned Counterfactual Gap Repair (SC-CGR), which adjusts a fitted detector through a lightweight head on frozen features while using typed replay and ranking constraints to preserve trusted behavior on real images and clean controls. Across five industrial anomaly-detection datasets, we systematically screen a 71-category inventory to obtain 37 feasible inspection contracts, revealing semantic response violations that are not reflected by the area under the receiver operating characteristic curve or average precision. Evaluation of the 13 contracts taken to the repair stage shows that SC-CGR substantially improves strict contract satisfaction on MVTec AD bottle and metal nut and generalizes to counterfactual probes from held-out Real-IAD bottle-cap anchors. The broader evaluation also identifies cases where pairwise constraints alone are sufficient and cases where enforcing the specification conflicts with retaining real-data behavior. These results show that specification-conditioned counterfactual responses provide a practical interface for auditing and repairing fitted industrial inspection systems beyond conventional anomaly-detection accuracy.

1 INTRODUCTION

Industrial anomaly detection is commonly evaluated by how well anomaly scores rank observed defective images above normal images. Patch-memory and patch-distribution methods such as PatchCore and PaDiM achieve strong AUROC and average precision (AP) (Roth et al., 2022; Defard et al., 2021), especially when paired with foundation-model features such as DINOv2 (Oquab et al., 2024). Ranking performance remains silent about the visual mechanism that produced the score.

Inspection decisions are defined by a specification. A crack or contamination mark may require rejection, while illumination, reflection, or benign texture variation is tolerated. Correlation in a test set can allow a detector to rank anomalies correctly while depending on a tolerated factor. Such a detector performs well on the benchmark and implements the wrong inspection rule.

We expose this failure with matched counterfactual response probes. Starting from one anchor image, two controlled factors produce the tuple $(x_{00}, x_{10}, x_{01}, x_{11})$. Its scores identify the main effect of the target defect, the main effect of the tolerated factor, and their interaction. An inspection specification maps these effects to a response contract, such as a positive target response with neutral nuisance and interaction responses. The resulting audit evaluates how the fitted detector responds, complementing its ranking on observed images.

We then formulate *specification-conditioned counterfactual repair*: given a fitted detector, a response contract, matched probes, and trusted retention data, modify the score response to satisfy the contract

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.

while preserving useful behavior. SC-CGR implements this formulation as response-contract projection. A small repair head operates on frozen detector features; a hinge-relaxed factorial objective moves diagnosed violations toward the feasible response region; typed replay, real-anomaly ranking, and parameter regularization encode preservation constraints. Hard-anchor refinement concentrates updates on the probes that still violate strict margins.

Figure 1 summarizes the system. The method changes the semantic response of an existing detector without refitting its feature extractor or treating counterfactual variants as ordinary independent labels.

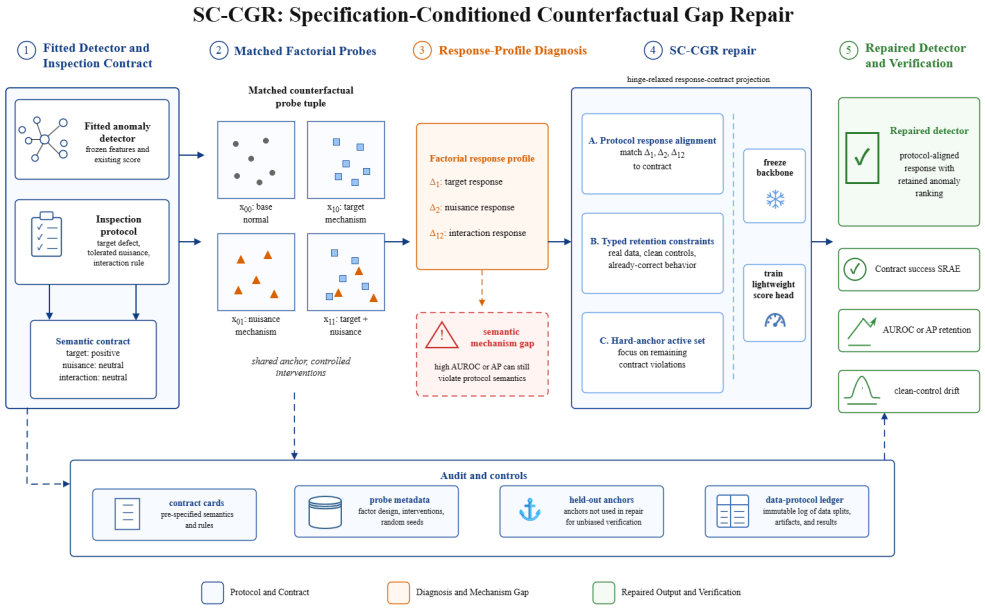

Figure 1: SC-CGR audits and repairs a fitted anomaly detector against an inspection response contract. Matched probes isolate target, tolerated-factor, and interaction effects. A lightweight repair head aligns this response profile, while typed retention constraints preserve trusted real and clean-control behavior.

Our contributions are:

1. We formulate industrial detector repair as projection onto a specification-defined response contract, separating benchmark ranking from mechanism-level correctness.
2. We introduce a factorial counterfactual representation and SC-CGR, which aligns target, tolerated-factor, and interaction responses under typed retention constraints.
3. We evaluate fitted DINOv2 detectors on five industrial datasets, demonstrating hidden response gaps, strict-contract repair, held-out generalization, and empirical boundaries of the full factorial objective.

2 Related Work

Industrial anomaly detection. Normal-only industrial anomaly detection spans teacher–student matching, density estimation, reconstruction, and patch-memory methods. STFPM matches multi-scale teacher and student features (Wang et al., 2021); CFLOW-AD estimates feature likelihoods with conditional normalizing flows (Gudovskiy et al., 2022); and reverse distillation reconstructs teacher features from a compact one-class embedding (Deng & Li, 2022). PaDiM (Defard et al., 2021) models spatial feature distributions, while PatchCore (Roth et al., 2022) retrieves nearby nominal patches from a coreset. Subsequent methods improve feature adaptation, efficiency, or matching, including SimpleNet (Liu et al., 2023), EfficientAD (Batzner et al., 2024), ReConPatch (Hyun et al., 2024), and CostFilter-AD (Zhang et al., 2025). UniAD (You et al., 2022) and Dinomaly (Guo et al., 2025) further study unified multi-class detection. These approaches primarily optimize detection

and localization. Our study instead asks whether an already fitted score obeys a category-specific inspection rule under controlled changes, and how to repair the score when it does not.

Foundation-model anomaly detection. Large pretrained representations have strengthened low-data inspection. DINOv2 provides general-purpose self-supervised visual features (Oquab et al., 2024); AnomalyDINO adapts them to patch-based few-shot detection (Damm et al., 2025). Vision-language methods transfer semantic supervision through prompts or language-conditioned representations, including WinCLIP (Jeong et al., 2023), AnomalyCLIP (Zhou et al., 2024), AdaCLIP (Cao et al., 2024), and AnomalyGPT (Gu et al., 2024). SC-CGR operates after representation learning: the backbone is frozen, and the object of diagnosis is the response implemented by the fitted anomaly score. This separates whether the representation contains useful visual evidence from how the scoring function uses that evidence.

Synthetic anomalies and matched counterfactual probes. CutPaste (Li et al., 2021) creates local cut-and-paste defects for self-supervision, DRAEM (Zavrtanik et al., 2021) learns reconstruction and discrimination from synthetic corruption, AnomalyDiffusion (Hu et al., 2024) generates few-shot defects, and GLASS broadens anomaly synthesis in feature and image space (Chen et al., 2024). Counterfactual Generative Networks use controlled generative mechanisms to expose and reduce shortcut dependence (Sauer & Geiger, 2021). In contrast, our synthetic variants are not independent training anomalies: the four members of each tuple share an anchor and are compared jointly to identify target, tolerated-factor, and interaction responses. The inspection contract then assigns the desired response to these controlled effects.

Invariance, logical constraints, and behavioral correctness. MVTec LOCO shows that industrial anomalies can violate logical constraints even when their local appearance is individually plausible (Bergmann et al., 2022). More generally, "right for the right reasons" regularization constrains sensitivity to specified input evidence (Ross et al., 2017), invariant risk minimization seeks stable predictive relationships across environments (Arjovsky et al., 2019), and counterfactual invariance formalizes insensitivity to factors that should not affect a prediction (Veitch et al., 2021). Pairwise ranking directly constrains relative scores (Burges et al., 2005). These ideas motivate useful components of our comparison family, but they do not by themselves specify the complete 2×2 response geometry. SC-CGR measures target, second-factor, and interaction effects on matched tuples and aligns only the effects required by the inspection contract.

Post-training repair and retention. Model editing changes selected behaviors of trained predictors (Santurkar et al., 2021), while NNrepair formulates classifier repair under explicit behavioral constraints (Usman et al., 2021). Calibration transforms score confidence without changing the underlying representation or ranking geometry (Platt, 2000; Guo et al., 2017), and continual anomaly detection uses replay to retain earlier knowledge during updates (Hu et al., 2025). SC-CGR specializes post-training repair to anomaly-score response contracts. Its preservation conditions are source-specific: real good images, benign controls, already-correct tuples, and, when used in split-controlled studies, labeled real anomalies are retained differently from diagnosed gap tuples.

3 Specification-Conditioned Repair

3.1 Inspection contracts and matched probes

Let a fitted detector produce score $s_0(x)$ from frozen features $f(x)$. An inspection contract is

$$\mathcal{C} = (\mathcal{M}, r, \tau^\star, \Omega), \tag{1}$$

where $\mathcal{M}$ contains controlled visual factors, r assigns each factor an inspection role (e.g., target defect or tolerated factor), $\tau^\star$ specifies the desired signed or margin-constrained response, and Ω records category, region, support mask, and provenance. Contracts are instantiated from defect definitions, intervention metadata, and category-specific acceptance rules before model diagnosis; they are not inferred from the detector scores used in the audit.

Two controlled factors applied to anchor u form $(x_{00}, x_{10}, x_{01}, x_{11})$, where each binary index records the absence (0) or presence (1) of its factor. The first factor is the target mechanism; the second is a

protocol-defined factor, typically a tolerated nuisance. For any score function s, the tuple induces

$$\Delta_1 s = \tfrac{1}{2}\big[(s(x_{10}) - s(x_{00})) + (s(x_{11}) - s(x_{01}))\big], \tag{2}$$

$$\Delta_2 s = \tfrac{1}{2}\big[(s(x_{01}) - s(x_{00})) + (s(x_{11}) - s(x_{10}))\big], \tag{3}$$

$$\Delta_{12} s = s(x_{11}) - s(x_{10}) - s(x_{01}) + s(x_{00}). \tag{4}$$

These terms measure the target main effect, second-factor main effect, and interaction. We write the response for anchor u as $\rho(s, u) = (\Delta_1 s_u, \Delta_2 s_u, \Delta_{12} s_u)$.

The semantic response alignment error (SRAE) is

$$\mathrm{SRAE}(s, \mathcal{C}) = \mathbb{E}_u[d_{\mathcal{C}}(\rho(s, u), \tau^\star)], \tag{5}$$

where $\mathbb{E}_u$ averages over anchors and $d_{\mathcal{C}}$ sums contract-inequality violations. Contract success is the fraction of anchors satisfying every inequality. A semantic mechanism gap is a low-success or high-SRAE response despite useful real-image ranking.

3.2 Constrained repair view

SC-CGR learns $s_\phi(x) = g_\phi(f(x))$, where g_ϕ is the repair head with parameters ϕ and f remains fixed. Let $\mathcal{G}$ contain diagnosed gap probes, $\mathcal{T}$ contain typed retention sources, and $\mathcal{Q}$ contain real-anomaly/real-good ranking pairs. The ideal repair is

$$\begin{aligned} \min_{\phi} \quad & \mathbb{E}_{u \in \mathcal{G}} d_{\mathcal{C}}(\rho(s_\phi, u), \tau^\star) \\ \text{s.t.} \quad & s_\phi \in \mathcal{P}_t(s_0), && t \in \mathcal{T}, \\ & s_\phi(x^+) \geq s_\phi(x^-) + m, && (x^+, x^-) \in \mathcal{Q}, \\ & \|\phi\|_2^2 \leq B. \end{aligned} \tag{6}$$

The objective aligns the diagnosed response. Here $\mathcal{P}_t(s_0)$ is the preservation set for source type t, m is the anomaly-ranking margin, and B bounds the repair-head norm.

3.3 Protocol response alignment

For the common target-positive, nuisance-neutral contract, the feasible response region is

$$\Delta_1 s_\phi \geq \delta, \qquad |\Delta_2 s_\phi| \leq \epsilon, \qquad |\Delta_{12} s_\phi| \leq \epsilon. \tag{7}$$

Its hinge relaxation is

$$\mathcal{L}_{\text{align}} = [\delta - \Delta_1 s_\phi]_+ + \lambda_2[|\Delta_2 s_\phi| - \epsilon]_+ + \lambda_{12}[|\Delta_{12} s_\phi| - \epsilon]_+. \tag{8}$$

where $[z]_+ = \max(z, 0)$, δ is the required target margin, ϵ is the neutral-response tolerance, and λ_2, λ_{12} weight the second-factor and interaction violations. Other contracts change the signs, margins, or interaction target while retaining the same response representation. Per-image binary supervision discards tuple structure; pairwise supervision constrains selected orderings but leaves parts of the factorial profile unspecified.

3.4 Typed retention and active-set refinement

Replay examples are typed as real good, real anomaly, clean-control counterfactual, already-correct counterfactual, or diagnosed-gap counterfactual. Each trusted source receives a preservation rule:

$$\mathcal{L}_{\text{replay}} = \sum_{t \in \mathcal{T}} \lambda_t \mathbb{E}_{x \sim t} \ell_t(s_\phi(x), s_0(x), y_t). \tag{9}$$

Here λ_t is a per-source weight, ℓ_t is its preservation loss, and y_t is an optional source label. Real-good and clean-control sources preserve the corresponding trusted score or label behavior, while already-correct tuples preserve contract-compliant responses. Diagnosed gap tuples are excluded from old-score matching and are optimized toward Equation 8. When labeled real anomalies are available in a split-controlled diagnostic, ranking is retained by

$$\mathcal{L}_{\text{rank}} = \mathbb{E}_{(x^+, x^-)}[m + s_\phi(x^-) - s_\phi(x^+)]_+. \tag{10}$$

Strict success can remain low after mean SRAE becomes small. We rank training anchors by current violation, form an active set $\mathcal{H}$ from the hardest fraction, and add

$$\mathcal{L}_{\text{hard}} = \mathbb{E}_{u \in \mathcal{H}} d_{\mathcal{C}}(\rho(s_\phi, u), \tau^\star). \tag{11}$$

The final hinge-relaxed objective is

$$\mathcal{L} = \lambda_a \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{replay}} + \lambda_q \mathcal{L}_{\text{rank}} + \lambda_h \mathcal{L}_{\text{hard}} + \lambda_w \|\phi\|_2^2. \tag{12}$$

The coefficients $\lambda_a, \lambda_q, \lambda_h$, and λ_w weight protocol alignment, real-data ranking, hard-anchor refinement, and parameter regularization, respectively.

3.5 Repair procedure

SC-CGR fixes the contract and anchor split before diagnosis, assigns source-specific retention constraints, optimizes Equation 12 with a refreshed hard-anchor set, and evaluates the selected head on held-out probes and retention sets. Appendix A.3 gives the complete repair procedure.

4 Experimental Protocol

Datasets, detectors, and contracts. We evaluate on MVTec AD (Bergmann et al., 2019), Real-IAD (Wang et al., 2024), VisA (Zou et al., 2022), MPDD (Jezek et al., 2021), and MVTec AD 2 (Heckler-Kram et al., 2026), covering 71 industrial categories. Fitted detectors use frozen DINOv2 features (Oquab et al., 2024) with PatchCore-style or PaDiM-like scoring (Roth et al., 2022; Defard et al., 2021). Each anchor produces four matched variants that preserve product content and intervention support while isolating the target, tolerated-factor, and interaction responses. Contracts and inclusion criteria are fixed before repair. Of the 71 categories, 37 category–contract pairs admit the required interventions and enter the systematic audit; 13 contracts are evaluated in at least one repair stage. Appendix B gives the complete coverage, selection rules, and probe construction details.

Repair setting and comparisons. SC-CGR and all repair baselines operate on the same frozen representation, feature banks, and matched anchors, isolating post-training behavioral repair from representation learning. Images are resized to 224×224 and encoded by frozen DINOv2-small into a 16×16 grid of 384-dimensional patch features. The repair head has one 64-unit hidden layer and top-5% patch pooling. Anchors are split before feature export, and all four members of a tuple remain in the same split. Retention supervision is source-specific: real-good images preserve nominal behavior, clean controls preserve tolerated responses, and already-correct tuples preserve contract-compliant behavior; diagnosed violations receive contract alignment without old-score matching. Any labeled anomalies used for ranking retention are disjoint from real-test evaluation. Patch-BCE treats variants as independent labeled samples (Li et al., 2021; Zavrtanik et al., 2021; Chen et al., 2024); RankNet-style pairwise repair orders target-present above target-absent variants (Burges et al., 2005); adding invariance also suppresses tolerated-factor responses (Arjovsky et al., 2019; Veitch et al., 2021). We further compare affine calibration (Platt, 2000; Guo et al., 2017), linear binary heads, and scalar response regression, and ablate interaction alignment and clean-control replay. Main cases train for 800 epochs and multi-category studies for 500; Appendix F reports the remaining settings.

Metrics and model selection. Semantic response is evaluated using SRAE, strict contract success, worst-anchor SRAE, and the target, tolerated-factor, and interaction responses $(\Delta_1, \Delta_2, \Delta_{12})$. Retention is measured by real-data AUROC/AP and clean-control SRAE. Hyperparameters are selected on validation anchors by strict contract success and then SRAE, with real AP used only as a retention tie-breaker. Held-out anchors are not used for optimization or selection. When labeled real anomalies are used for retention, the corresponding ranking results are reported separately and do not define the contract or held-out probes. Bold marks the best value among directly compared methods for each metric and category, with ties sharing emphasis. Full split, sensitivity, and implementation details are provided in Appendices A, E, and F.

Table 1: Representative inspection contracts fixed before repair. The three response roles correspond to target, tolerated factor, and interaction.

Dataset and category	Target mechanism	Tolerated factor	Response roles
MVTec bottle	Crack	Illumination/reflection	Positive / neutral / neutral
MVTec metal nut	Surface defect	Boundary reflection	Positive / neutral / neutral
Real-IAD bottle cap	Defect-like cap change	Reflection	Positive / neutral / neutral
VisA capsules	Contamination	Benign appearance change	Positive / neutral / neutral
MPDD metal plate	Scratch	Benign surface variation	Positive / neutral / neutral

5 Results

5.1 Real-image ranking does not determine protocol response

Figure 2 reports expanded audits of representative contracts. Both fitted detector families achieve perfect real AUROC/AP on MVTec bottle and zero contract success. The metal-nut detectors also combine strong ranking with complete response failure. VisA extends the observation beyond the initial MVTec cases. All six detectors have zero contract success. The broader screening audit and complete response decomposition in Appendix C separate this phenomenon from low-performing detectors across datasets and categories.

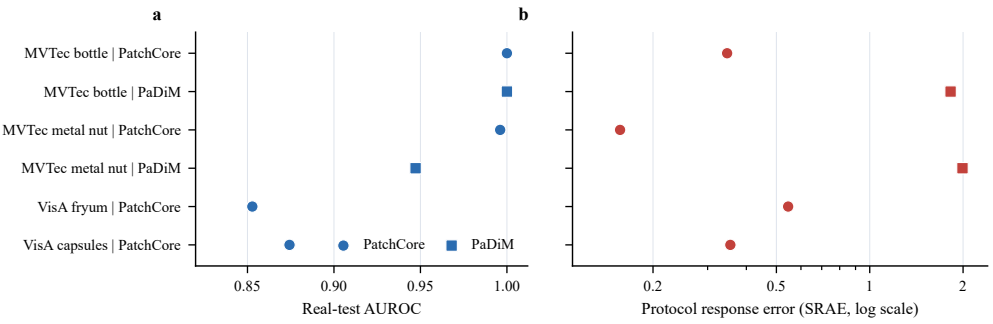

Figure 2: Real-test ranking and protocol response error for representative fitted detectors. Each row identifies the dataset, category, and detector; all six have zero strict contract success. High AUROC can coexist with large SRAE. Complete AUROC/AP and response components are reported in Appendix Tables 13–14.

5.2 Factorial alignment repairs strict contracts

Table 2 compares repair objectives on MVTec bottle. Binary supervision preserves perfect ranking but leaves every anchor outside the contract. Pairwise+invariance reduces average violation and reaches 34.38% success. Removing interaction alignment reaches 56.25%. Full SC-CGR satisfies every anchor with unchanged AUROC/AP. Appendix Figure 4 visualizes the corresponding target, tolerated-factor, and interaction responses.

Table 2: Factorial-response ablation on the MVTec bottle contract.

Method	SRAE ↓	Success ↑	AUROC/AP ↑
Fitted PatchCore-style	0.3468	0.0000	**1.0000 / 1.0000**
Patch-BCE with typed replay	0.0942	0.0000	**1.0000 / 1.0000**
Pairwise + invariance	0.0065	0.3438	**1.0000 / 1.0000**
SC-CGR without interaction	0.0048	0.5625	**1.0000 / 1.0000**
SC-CGR	**0.0000**	**1.0000**	**1.0000 / 1.0000**

5.3 Retention constraints support hard repairs

Metal nut tests a harder surface-defect/reflection contract. Table 3 shows that SC-CGR reaches SRAE 0.0008 and 95.83% success while retaining AUROC/AP of $0.9901/0.9976$. Pairwise+invariance preserves ranking more closely but leaves most anchors in violation. Hard-anchor refinement targets the remaining strict violations after average SRAE has fallen, while typed replay controls drift on clean responses. Figure 3a summarizes every retained outcome from the validation-ranked sweep; multiple configurations exceed 89% success with real AP above 0.997, so the main result is not tied to one isolated configuration. Appendix Table 17 gives every response and retention metric.

Table 3: Retention-constrained repair on the MVTec metal-nut contract.

Method	SRAE ↓	Success ↑	AUROC/AP ↑
Fitted PatchCore-style	0.1566	0.0000	0.9961 / 0.9991
Fitted PaDiM-like	1.9914	0.0000	0.9472 / 0.9869
Patch-BCE with typed replay	0.1022	0.0000	0.9901 / 0.9977
Pairwise + invariance	0.0148	0.3125	**1.0000 / 1.0000**
SC-CGR	**0.0008**	**0.9583**	0.9901 / 0.9976

5.4 Repair generalizes to held-out Real-IAD anchors

Real-IAD bottle cap uses disjoint optimization and evaluation anchors. Table 4 reports the 512-anchor protocol, with 384 anchors for repair and 128 for evaluation. SC-CGR reduces held-out SRAE to 0.0007 and reaches 84.38% strict success. Moving from the contract-balanced to the response-focused configuration reduces SRAE from 0.000694 to 0.000644 and raises strict success by 0.78 percentage points, at the cost of 1.19 AP points on real images. Pairwise+invariance retains the highest AP among repair methods but satisfies fewer than half of the held-out anchors. Figure 3b shows that increasing anchor coverage from 96 to 512 improves success from 35.42% to 84.38% while reducing SRAE by more than an order of magnitude. Appendix Table 20 gives the split sizes.

Table 4: Held-out Real-IAD bottle-cap repair with 384 optimization and 128 evaluation anchors.

Method	SRAE ↓	Success ↑	AUROC/AP ↑	Clean SRAE ↓
Patch-BCE	0.005108	0.8125	0.9600 / 0.9675	0.009426
Pairwise + invariance	0.005401	0.4688	**0.9794 / 0.9826**	0.011970
SC-CGR, retention-balanced	0.000799	0.7734	0.9762 / 0.9817	**0.003770**
SC-CGR, contract-balanced	0.000694	0.8438	0.9581 / 0.9662	0.005740
SC-CGR, response-focused	**0.000644**	**0.8516**	0.9437 / 0.9543	0.005513

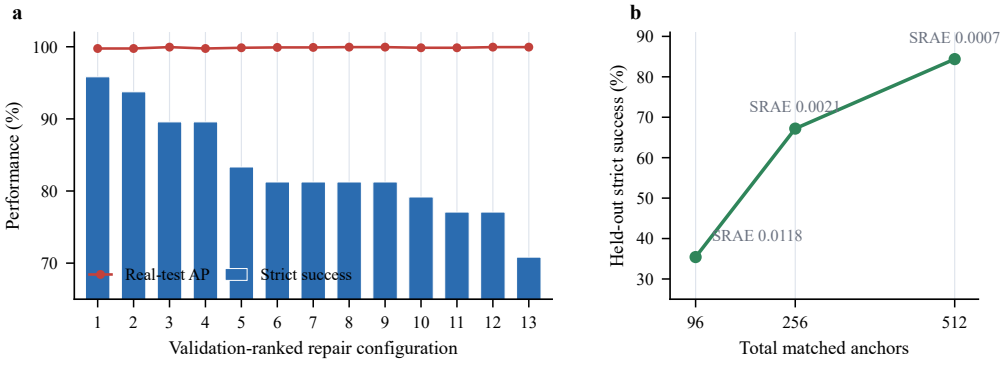

Figure 3: Stability and anchor coverage. (a) Strict success and real-test AP for all 13 retained metal-nut configurations, ordered by the prespecified validation rule. (b) Held-out success on Real-IAD bottle cap as matched-anchor coverage grows; labels report held-out SRAE. Complete values are in Appendix Tables 17 and 20.

5.5 THE FULL OBJECTIVE IS NEEDED ONLY FOR SPECIFIC RESPONSE GEOMETRY

The multi-category study first compares objectives on small matched-anchor sets (Table 5) and then repeats the comparison with disjoint held-out anchors and real-data retention (Table 6). The expansion separates easy optimization from reliable repair. MPDD metal plate and VisA cashew remain pairwise-sufficient. VisA fryum and capsules reduce semantic-response error but retain low strict success and lose real AP, suggesting a trade-off between enforcing the nuisance-neutral contract and preserving ranking on the observed real-image distribution. Method choice therefore depends on both the audited response geometry and the retention requirement.

Table 5: Small-anchor multi-category comparison before held-out expansion.

Dataset and category	Method	SRAE ↓	Success ↑
VisA cashew	Patch-BCE / Pairwise+invariance / SC-CGR	0.2119 / **0.0000** / **0.0000**	0.0833 / **1.0000** / **1.0000**
VisA capsules	Patch-BCE / Pairwise+invariance / SC-CGR	0.8477 / 0.0349 / **0.0220**	0.0000 / 0.1667 / **0.3333**
MPDD metal plate	Patch-BCE / Pairwise+invariance / SC-CGR	0.0210 / **0.0000** / **0.0000**	0.2500 / **1.0000** / **1.0000**

Table 6: Complete held-out multi-category comparison with real-data and clean-control retention.

Dataset and category	Method	SRAE ↓	Success ↑	AUROC ↑	AP ↑	Clean SRAE ↓
MPDD metal plate	Pairwise + invariance	**0.0000**	**1.0000**	0.8722	0.9537	0.0208
	SC-CGR	0.0002	0.9231	**0.9480**	**0.9797**	0.0121
	SC-CGR without interaction	0.0001	0.9231	0.9382	0.9759	0.0127
	Patch-BCE	0.0120	0.3077	0.8348	0.9384	**0.0100**
VisA capsules	SC-CGR	0.0370	**0.1406**	0.5166	0.1666	0.0726
	Pairwise + invariance	0.0478	0.0781	0.4937	0.1446	0.0528
	SC-CGR without interaction	**0.0335**	0.0469	0.5111	0.1598	0.0677
	Patch-BCE	0.6074	0.0000	**0.5820**	**0.2412**	**0.0251**
VisA cashew	Pairwise + invariance	**0.0009**	**0.9219**	0.8040	0.5438	0.0264
	SC-CGR	0.0015	0.7500	0.9014	0.7270	0.0247
	SC-CGR without interaction	0.0017	0.7188	**0.9043**	**0.7460**	0.0183
	Patch-BCE	0.1432	0.0938	0.8401	0.5926	**0.0104**
VisA fryum	SC-CGR	**0.0098**	**0.3125**	0.7018	0.4297	0.0413
	Pairwise + invariance	0.0167	0.2031	0.7701	0.5314	**0.0242**
	SC-CGR without interaction	0.0111	0.2344	0.6865	0.3989	0.0422
	Patch-BCE	0.4406	0.0000	**0.8256**	**0.6295**	0.0417

5.6 CALIBRATION AND ORDINARY HEAD FITTING ARE INSUFFICIENT

Affine calibration cannot alter tuple geometry and obtains zero bottle success. A linear binary head uses the same frozen representation but also reaches zero success. Scalar regression heads remain at or below 3.12% success across the expanded controls in Appendix E.2. Table 7 places these controls beside SC-CGR on the bottle contract. The improvement tracks tuple-level response supervision rather than head capacity or global score scaling.

Table 7: Post-training controls on the expanded MVTec bottle protocol.

Method	SRAE ↓	Success ↑	AUROC ↑	AP ↑	Clean SRAE ↓
Affine score calibration	0.0486	0.0000	**1.0000**	**1.0000**	–
Linear binary head	0.2405	0.0000	0.9903	0.9970	0.0123
SC-CGR	**0.0000**	**1.0000**	**1.0000**	**1.0000**	**0.0095**

The same pattern holds beyond the bottle example. On MVTec metal nut, the strongest ordinary score-regression control reaches 10.42% success, compared with 95.83% for SC-CGR. On Real-IAD bottle cap, all three ordinary heads remain below 1% success despite using the same frozen DINOv2 representation. Table 8 reports these controls together because their shared failure is response-structural: independent sample targets do not specify the tolerated-factor and interaction geometry.

Table 8: Ordinary post-training heads using the same frozen representation. These runs omit labeled-anomaly ranking retention.

Dataset	Category	Method	SRAE ↓	Success ↑	AUROC ↑	AP ↑	Clean SRAE ↓
MVTec AD	Bottle	Linear Patch-BCE	**0.0384**	0.0000	0.9032	0.9620	0.0053
		Nonlinear score regression	0.0681	**0.0312**	**0.9839**	**0.9950**	0.0053
		Linear score regression	0.0434	**0.0312**	0.9032	0.9639	**0.0020**
MVTec AD	Metal nut	Linear Patch-BCE	0.0276	0.0208	0.7885	0.9477	0.0104
		Nonlinear score regression	0.0802	0.0208	0.8043	0.9510	0.0095
		Linear score regression	**0.0210**	**0.1042**	**0.9032**	**0.9771**	**0.0068**
Real-IAD	Bottle cap	Linear Patch-BCE	0.1510	**0.0078**	0.8731	0.8851	0.0143
		Nonlinear score regression	**0.0553**	**0.0078**	0.8369	0.8807	**0.0040**
		Linear score regression	0.1464	**0.0078**	**0.8762**	**0.8909**	0.0121

5.7 Response geometry localizes the remaining errors

The factorial decomposition distinguishes causes behind failed contracts. The PaDiM-like bottle detector has a positive target effect ($\Delta_1 = 2.1645$), alongside a large nuisance response ($\Delta_2 = 0.6664$) and negative interaction ($\Delta_{12} = -1.9657$). On VisA cashew, the tolerated-factor and interaction responses are both negative (-0.5767 and -1.1481). Target-versus-normal ranking alone does not distinguish these geometries.

Mean SRAE and strict success provide complementary views. Every retained metal-nut configuration has SRAE below 0.002, while success ranges from 70.83% to 95.83%, localizing residual error to boundary anchors. On Real-IAD, increasing optimization anchors from 48 to 384 reduces held-out SRAE from 0.0118 to 0.0007 and raises success from 35.42% to 84.38%.

6 Discussion

SC-CGR turns an inspection specification into a post-training repair target. Matched tuples measure target, tolerated-factor, and interaction effects while cancelling anchor-specific score offsets. The resulting profile supports diagnosis, optimization, and direct evaluation of the intended inspection rule.

Choosing and constraining a repair. The audited geometry indicates the required objective: target ordering addresses a missing target effect, invariance addresses an unwanted tolerated-factor effect, and factorial alignment addresses a residual interaction. Hard-anchor refinement focuses on remaining boundary violations. Typed retention protects fitted behavior on real good images, benign controls, and already-correct tuples, and also on labeled real anomalies when anomaly-ranking retention is enabled in split-controlled studies. The evaluation reports semantic correction together with retention, and reveals both pairwise-sufficient contracts and conflicts with the observed anomaly distribution.

Scope and deployment. Our experiments repair a lightweight score head on frozen DINOv2 features, isolating response correction and keeping the update inexpensive. Deployment requires contracts and interventions that reflect expert-defined acceptance rules and validation with physical data. The present results cover image-level scoring across five public datasets; end-to-end repair, localization, and broader physical interventions remain future directions. Threshold re-analysis and held-out-anchor bootstrap quantify sensitivity and uncertainty.

7 Conclusion

Industrial anomaly ranking and inspection-rule compliance are distinct properties. Matched factorial probes reveal this distinction and provide a structured target for repair. SC-CGR projects a fitted detector's response onto a specification-defined contract while retaining trusted behavior through typed constraints. Results across five datasets show strict-contract repair, held-out generalization, and clear regimes for factorial, pairwise, and retention-conflicted behavior. Specification-conditioned response auditing can complement AUROC and AP when anomaly detectors are expected to implement concrete industrial acceptance rules.

AI USE STATEMENT

In this work, we used generative AI tools for proofreading, improving the clarity and readability of the manuscript, and assisting with the writing, debugging, and refinement of code. We did not use generative AI tools for generating research ideas, designing algorithms, conducting experiments, analyzing results, or producing scientific claims. All AI-assisted text modifications and code were carefully reviewed, verified, and tested by the authors. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 127–137, 2024. doi: 10.1109/WACV57701.2024.00020.

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

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130:947–969, 2022. doi: 10.1007/s11263-022-01578-9.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 89–96, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102363. URL https://doi.org/10.1145/1102351.1102363.

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 Nature Switzerland, 2024.

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 Nature Switzerland, 2024. doi: 10.1007/978-3-031-72855-6_3.

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1319–1329, 2025. doi: 10.1109/WACV61041.2025.00136.

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

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9727–9736, 2022. doi: 10.1109/CVPR52688.2022.00951.

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

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1819–1828, 2022. doi: 10.1109/WACV51458.2022.00188.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20405–20415, 2025. doi: 10.1109/CVPR52734.2025.01900.

Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection. *International Journal of Computer Vision*, 134(4):175, 2026. ISSN 1573-1405. doi: 10.1007/s11263-026-02743-0. URL https://doi.org/10.1007/s11263-026-02743-0.

Lei Hu, Zhiyong Gan, Ling Deng, Jinglin Liang, Lingyu Liang, Shuangping Huang, and Tianshui Chen. Replaycad: Generative diffusion replay for continual anomaly detection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 2946–2954. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/328.

Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8526–8534, 2024. doi: 10.1609/aaai.v38i8.28696.

Jeeho Hyun, Sangyun Kim, Giyoung Jeon, Seung Hwan Kim, Kyunghoon Bae, and Byung Jun Kang. Reconpatch : Contrastive patch representation learning for industrial anomaly detection. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2041–2050, 2024. doi: 10.1109/WACV57701.2024.00205.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19606–19616, 2023. URL https://api.semanticscholar.org/CorpusID:257766897.

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

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9659–9669, 2021. doi: 10.1109/CVPR46437.2021.00954.

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.

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, Herve 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. Featured Certification.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10, 06 2000.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL https://doi.org/10.24963/ijcai.2017/371.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14298–14308, 2022. doi: 10.1109/CVPR52688.2022.01392.

Shibani Santurkar, Dimitris Tsipras, Mahalaxmi Elango, David Bau, Antonio Torralba, and Aleksander Madry. Editing a classifier by rewriting its prediction rules. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23359–23373. Curran Associates, Inc., 2021.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. *ArXiv*, abs/2101.06046, 2021. URL https://api.semanticscholar.org/CorpusID:231627872.

Muhammad Usman, Divya Gopinath, Youcheng Sun, Yannic Noller, and Corina S. Pasareanu. Nnrepair: Constraint-based repair of neural network classifiers. In *Computer Aided Verification*, pp. 3–25, Cham, 2021. Springer International Publishing.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:248497852.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22883–22892, 2024. doi: 10.1109/CVPR52733.2024.02159.

Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for unsupervised anomaly detection. *ArXiv*, abs/2103.04257, 2021. URL https://api.semanticscholar.org/CorpusID:232147174.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 4571–4584. Curran Associates, Inc., 2022. doi: 10.52202/068431-0330.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8310–8319, 2021. doi: 10.1109/ICCV48922.2021.00822.

Zhe Zhang, Mingxiu Cai, Hanxiao Wang, Gaochang Wu, Tianyou Chai, and Xiatian Zhu. CostFilter-AD: Enhancing anomaly detection through matching cost filtering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74540–74564. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zhang25e.html.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, pp. 49705–49737, 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. doi: 10.1007/978-3-031-20056-4_23.

A Protocol and Dataset Details

A.1 Dataset inventory

The study covers 71 categories across five datasets. Table 9 reports dataset-level image inventories from the dataset releases, rather than feature records, generated counterfactuals, or the number of images sampled by an individual experiment. Detector fitting and real-image evaluation follow the split metadata supplied with each dataset. For MVTec AD and MPDD, we use the released anomaly-free training and labeled evaluation partitions. For VisA and Real-IAD, we read the provided one-class and JSON train/test metadata, respectively. MVTec AD 2 uses its anomaly-free training set for detector fitting; only label-bearing validation and public-test data contribute to local screening metrics, while the private test partitions are excluded. The broad screening stage audits fitted detectors and does not optimize a repair head. Expanded repair protocols use larger matched-anchor sets and separate validation or held-out evaluation where indicated.

Table 9: Dataset inventory and partition scope. Dataset-scale counts follow the corresponding releases (Bergmann et al., 2019; Heckler-Kram et al., 2026; Wang et al., 2024; Zou et al., 2022; Jezek et al., 2021). "Approx." denotes the scale reported by the dataset authors.

Dataset	Categories	Official image inventory	Released partition structure and study scope
MPDD	6	1,346 images: 888 train-normal; 176 validation-normal; 282 validation-anomalous	Released train/validation split; pixel masks accompany anomalous validation images.
MVTec AD	15	5,354 images: 3,629 train-normal; 467 test-normal; 1,258 test-anomalous	Released train/test split; pixel masks accompany anomalous test images.
MVTec AD 2	8	More than 8,000 images	Defect-free train and validation sets; public and private test partitions. We use train plus label-bearing validation/public-test data and do not report private-test metrics.
Real-IAD	30	Approx. 150,000 multi-view images	Supplied JSON train/test split; views retain their sample grouping for evaluation. Directory-level OK/NG file totals are not treated as test-set counts.
VisA	12	10,821 images: 9,621 normal; 1,200 anomalous	Supplied one-class split; normal images are divided between train and test, and anomalous images include pixel masks.

A.2 Contract construction and data separation

Each contract card records the category, target mechanism, tolerated factor, desired response signs and margins, support region, intervention source, and review status. The contract is fixed before repair and is never inferred from model scores. For the target-positive, nuisance-neutral contracts used in the main experiments, "tolerated" means that adding the factor should change the anomaly score by at most the preset tolerance ϵ; it does not require the visual change itself to be imperceptible. Matched tuple members share one anchor and differ only by recorded interventions. Held-out evaluations split anchors before optimization.

Real-anomaly ranking is enabled only in split-controlled retention studies. Labeled anomalies used for ranking are disjoint from real-test evaluation data and never determine the contract, anchor selection, or counterfactual target. Unsupervised-compatible runs use real-good and counterfactual replay without labeled anomaly ranking.

A.3 COMPLETE REPAIR PROCEDURE

SC-CGR repair procedure

Input: fitted score s_0, frozen representation f, inspection contract $\mathcal{C}$, matched anchors, and typed retention sources.

1. Split anchors into optimization, validation, and held-out sets; instantiate the four matched variants for each anchor.
2. Audit s_0 using the response profile, SRAE, strict success, real AUROC/AP, and clean-control response. Mark violating tuples as gap probes.
3. Export frozen patch features. Attach each replay source to its preservation rule, and exclude gap probes from old-score preservation.
4. Optimize Equation 12. Recompute anchor violations periodically and refresh the hard-anchor active set.
5. Select a configuration on validation anchors by strict success and then SRAE, using real AP only to break retention ties.
6. Evaluate the selected head once on held-out probes, real images, and clean controls; report semantic repair and retention together.

Output: repaired score s_ϕ and its contract/retention audit.

Table 10: Representative inspection contracts.

Dataset and category	Target mechanism	Tolerated factor	Response contract
MVTec bottle	Crack	Illumination/reflection	Positive / neutral / neutral
MVTec metal nut	Surface defect	Boundary reflection	Positive / neutral / neutral
Real-IAD bottle cap	Defect-like cap change	Reflection	Positive / neutral / neutral
VisA capsules	Contamination	Benign appearance change	Positive / neutral / neutral
MPDD metal plate	Scratch	Benign surface variation	Positive / neutral / neutral

B COVERAGE AND STAGED SELECTION

The 71-category dataset inventory defines the available data pool, not a claim that every category admits a meaningful factorial contract. A category enters screening only when the target mechanism and tolerated factor can be stated before scoring, both interventions can be generated on shared support, the four tuple members pass the same construction checks, and enough source images are available. Thirty-seven category–contract pairs meet these conditions. Table 12 reports all of them, including low-gap and ambiguous controls.

For each scan-selected expansion, candidate selection is frozen from the fitted detector's pre-repair audit. Selection uses response-gap strata, dataset and mechanism coverage, counterfactual quality, and anchor availability; no repaired score is used. The previously established MVTec bottle, MVTec metal-nut, and Real-IAD bottle-cap cases remain a separately identified reference set. Small-anchor probes test objective sufficiency before expensive expansion. Strict expansion then uses disjoint optimization and held-out anchors. This staged design yields four scan-selected strict protocols: MPDD metal plate and VisA cashew, capsules, and fryum.

Table 11 lists every dataset/category contract that entered the reported DINOv2 repair study. Each contract appears once at its strongest completed evaluation regime; smaller precursor experiments remain in Tables 15 and 16. The two numeric columns are independently the lowest SRAE and highest success over the tested objectives, so they need not come from the same method. The ledger retains pairwise-sufficient, non-discriminative, unresolved, and retention-conflicted outcomes alongside the headline cases.

Table 11: Complete contract-level coverage of the reported DINOv2 repair study.

Dataset and category	Contract	Evaluation regime	Best SRAE	Best success	Observed outcome
MVTec AD bottle	Reflection	32-anchor expanded	0.0000	1.0000	Full factorial alignment required for strict success.
MVTec AD metal nut	Reflection	48-anchor expanded	0.0008	0.9583	Full SC-CGR gives the strongest strict repair.
Real-IAD bottle cap	Reflection	128 held out	0.0006	0.8516	SC-CGR leads semantic repair; retention varies by configuration.
MVTec AD capsule	Contamination	12-anchor probe	0.0000	1.0000	Pairwise invariance is sufficient.
MVTec AD hazelnut	Contamination	12-anchor probe	0.0000	1.0000	Pairwise invariance is sufficient.
MVTec AD leather	Scratch	12-anchor probe	0.0000	1.0000	Pairwise invariance is sufficient.
MVTec AD 2 vial	Reflection	12-anchor external probe	0.0000	1.0000	Pairwise objectives are sufficient.
VisA candle	Reflection	12-anchor external probe	0.0000	1.0000	All tested objectives succeed; the case is non-discriminative.
VisA pipe fryum	Reflection	12-anchor probe	0.0000	1.0000	Pairwise invariance is sufficient.
VisA cashew	Contamination	64 held out	0.0009	0.9219	Pairwise invariance gives the strongest strict result.
VisA capsules	Contamination	64 held out	0.0335	0.1406	Strict repair remains unresolved and competes with retention.
VisA fryum	Contamination	64 held out	0.0098	0.3125	SC-CGR improves response alignment with a real-data retention cost.
MPDD metal plate	Scratch	13 held out	0.0000	1.0000	Pairwise invariance is sufficient.

C Screening Audit

Table 12 reports the complete 37-contract audit. Tables 13 and 14 retain compact views of the largest gaps and their response decomposition. Screening uses smaller probe sets than the expanded main experiments and supports candidate diagnosis rather than final repair comparison.

Table 12: Complete pre-repair PatchCore-style broad screening ledger. AUROC/AP are real-image ranking metrics; SRAE and success are computed on the small matched-probe audit.

Dataset	Category	Contract	AUROC / AP	SRAE	Success	Audit stratum
MVTec AD	Bottle	Reflection	1.0000 / 1.0000	0.4868	0.0000	High-rank gap
MVTec AD	Metal nut	Reflection	0.9956 / 0.9990	0.2154	0.0000	High-rank gap
MVTec AD	Capsule	Contamination	0.8947 / 0.9772	0.5190	0.0000	Hard gap
MVTec AD	Pill	Contamination	– / –	0.3394	0.0000	Hard gap
MVTec AD	Wood	Scratch	0.9623 / 0.9880	0.0180	0.2500	High-rank gap
MVTec AD	Leather	Scratch	1.0000 / 1.0000	0.3146	0.0000	High-rank gap
MVTec AD	Cable	Scratch	0.9597 / 0.9782	0.0037	0.7500	No-gap control
MVTec AD	Hazelnut	Contamination	0.9946 / 0.9975	0.3399	0.0000	High-rank gap
MVTec AD	Grid	Scratch	0.9766 / 0.9923	0.0092	0.4167	Ambiguous
MVTec AD	Tile	Scratch	1.0000 / 1.0000	0.0226	0.3333	Medium gap
MVTec AD	Transistor	Reflection	0.9254 / 0.9095	0.0563	0.0000	Ambiguous
MVTec AD	Zipper	Scratch	0.9793 / 0.9944	0.0071	0.7500	No-gap control
VisA	Candle	Reflection	0.9027 / 0.8840	0.2608	0.0000	Hard gap
VisA	Cashew	Contamination	0.9028 / 0.9490	0.2340	0.0000	Hard gap
VisA	Capsules	Contamination	0.8743 / 0.9249	0.3552	0.0000	Hard gap
VisA	Chewing gum	Contamination	0.9728 / 0.9893	0.2110	0.0000	High-rank gap
VisA	Fryum	Contamination	0.8528 / 0.9261	0.5456	0.0000	Hard gap
VisA	PCB1	Scratch	0.7884 / 0.7725	0.0000	1.0000	No-gap control
VisA	PCB2	Scratch	0.7766 / 0.7370	0.0000	1.0000	No-gap control
VisA	PCB3	Scratch	0.8090 / 0.8298	0.0000	1.0000	No-gap control
VisA	Pipe fryum	Reflection	0.9298 / 0.9671	0.3240	0.0000	Hard gap
MVTec AD 2	Can	Reflection	0.6298 / 0.6263	0.0488	0.0000	Ambiguous
MVTec AD 2	Fabric	Scratch	0.6453 / 0.7070	0.0354	0.0000	Ambiguous
MVTec AD 2	Rice	Contamination	0.8192 / 0.9026	0.0647	0.0000	Ambiguous
MVTec AD 2	Sheet metal	Scratch	0.7980 / 0.9414	0.0506	0.0000	Ambiguous
MVTec AD 2	Vial	Reflection	0.8236 / 0.8849	0.0556	0.0000	Ambiguous
MPDD	Black bracket	Scratch	0.7832 / 0.7800	0.0797	0.0000	Ambiguous
MPDD	Brown bracket	Scratch	0.8167 / 0.8752	0.3274	0.0000	Hard gap
MPDD	White bracket	Scratch	0.5656 / 0.5065	0.1716	0.0000	Hard gap
MPDD	Connector	Reflection	0.8833 / 0.7498	0.1212	0.0000	Hard gap
MPDD	Metal plate	Scratch	1.0000 / 1.0000	0.0150	0.1667	High-rank gap
MPDD	Tubes	Reflection	0.8238 / 0.9294	0.0317	0.0833	Ambiguous
Real-IAD	Bottle cap	Reflection	0.7573 / 0.7604	0.2440	0.0000	Hard gap
Real-IAD	PCB	Scratch	0.7901 / 0.8017	0.0180	0.2500	Medium gap
Real-IAD	Plastic nut	Reflection	0.6921 / 0.7121	0.2146	0.0000	Hard gap
Real-IAD	USB	Scratch	0.8118 / 0.8335	0.0039	0.8333	No-gap control
Real-IAD	Toy brick	Scratch	0.6503 / 0.6175	0.0480	0.4167	Medium gap

Table 13: Screening overview across datasets, categories, contracts, and fitted detectors.

Dataset	Category	Contract	Detector	AUROC	AP	SRAE	Success	Audit regime
MVTec AD	Metal nut	Reflection	PaDiM-like	0.9257	0.9816	3.1417	0.0000	Large gap
MVTec AD	Bottle	Reflection	PaDiM-like	1.0000	1.0000	2.6121	0.0000	High ranking, large gap
VisA	Cashew	Contamination	PaDiM-like	0.7836	0.8784	1.7048	0.0000	Large gap
VisA	Capsules	Contamination	PaDiM-like	0.6825	0.7765	0.6683	0.0000	Large gap
MVTec AD	Bottle	Reflection	PatchCore-style	1.0000	1.0000	0.5120	0.0000	High ranking, large gap
VisA	Capsules	Contamination	PatchCore-style	0.8743	0.9249	0.3552	0.0000	Large gap
VisA	Cashew	Contamination	PatchCore-style	0.9028	0.9490	0.2340	0.0000	Large gap
MVTec AD	Metal nut	Reflection	PatchCore-style	0.9956	0.9990	0.2114	0.0000	High ranking, large gap
MPDD	Metal plate	Scratch	PaDiM-like	0.4323	0.7343	0.1137	0.0000	Large gap
MPDD	Black bracket	Scratch	PatchCore-style	0.7832	0.7800	0.0797	0.0000	Small or ambiguous gap
MVTec AD 2	Sheet metal	Scratch	PaDiM-like	0.8129	0.9397	0.0782	0.0000	Small or ambiguous gap
MVTec AD 2	Vial	Reflection	PaDiM-like	0.7129	0.8144	0.0681	0.0000	Small or ambiguous gap
MPDD	Black bracket	Scratch	PaDiM-like	0.6609	0.6875	0.0591	0.0000	Small or ambiguous gap
MVTec AD 2	Vial	Reflection	PatchCore-style	0.8236	0.8849	0.0556	0.0000	Small or ambiguous gap
MVTec AD 2	Sheet metal	Scratch	PatchCore-style	0.7980	0.9414	0.0506	0.0000	Small or ambiguous gap

Table 14: Screening audit of fitted DINOv2 detectors.

Dataset	Category	Detector	AUROC	AP	SRAE	Δ_1	Δ_2	Δ_{12}
MVTec AD	Metal nut	PaDiM-like	0.9257	0.9816	3.1417	2.4567	0.9722	-2.1895
MVTec AD	Bottle	PaDiM-like	1.0000	1.0000	2.6121	2.1645	0.6664	-1.9657
VisA	Cashew	PaDiM-like	0.7836	0.8784	1.7048	1.1049	-0.5767	-1.1481
VisA	Capsules	PaDiM-like	0.6825	0.7765	0.6683	0.2605	-0.2294	-0.4589
VisA	Fryum	PatchCore-style	0.8528	0.9261	0.5456	0.2060	-0.1880	-0.3776
MVTec AD	Capsule	PatchCore-style	0.8947	0.9772	0.5190	0.3738	-0.1157	-0.4234
MVTec AD	Bottle	PatchCore-style	1.0000	1.0000	0.5120	0.4160	0.1142	-0.4178
MVTec AD	Hazelnut	PatchCore-style	0.9946	0.9975	0.3399	0.3190	-0.1015	-0.2584
MPDD	Brown bracket	PatchCore-style	0.8167	0.8752	0.3274	0.1862	0.1610	-0.1833
VisA	Pipe fryum	PatchCore-style	0.9298	0.9671	0.3240	0.4661	0.0996	-0.2444
MVTec AD	Leather	PatchCore-style	1.0000	1.0000	0.3146	0.5321	0.0956	-0.2389
VisA	Candle	PatchCore-style	0.9027	0.8840	0.2608	0.4390	0.0595	-0.2212
Real-IAD	Bottle cap	PatchCore-style	0.7573	0.7604	0.2440	0.3998	0.1011	-0.1629
VisA	Cashew	PatchCore-style	0.9028	0.9490	0.2340	0.3844	-0.0821	-0.1719
MVTec AD	Metal nut	PatchCore-style	0.9956	0.9990	0.2154	0.3675	0.0503	-0.1852
Real-IAD	Plastic nut	PatchCore-style	0.6921	0.7121	0.2146	0.2930	0.1106	-0.1240

D Held-Out Multi-Category Results

Table 15: Small-anchor multi-category repair study used before held-out expansion.

Dataset and category	Contract	Method	SRAE	Success
VisA cashew	Contamination	Patch-BCE	0.2119	0.0833
		Pairwise + invariance	0.0000	1.0000
		SC-CGR without interaction	0.0000	1.0000
		SC-CGR	0.0000	1.0000
VisA capsules	Contamination	Patch-BCE	0.8477	0.0000
		Pairwise + invariance	0.0349	0.1667
		SC-CGR without interaction	0.0062	0.3333
		SC-CGR	0.0220	0.3333
MPDD metal plate	Scratch	Patch-BCE	0.0210	0.2500
		Pairwise + invariance	0.0000	1.0000
		SC-CGR without interaction	0.0000	0.9167
		SC-CGR	0.0000	1.0000

Table 16: Held-out multi-category repair. Real AUROC/AP and clean-control SRAE report retention.

Dataset and category	Method	SRAE	Success	AUROC	AP	Clean SRAE
MPDD metal plate	Pairwise + invariance	0.0000	1.0000	0.8722	0.9537	0.0208
	SC-CGR	0.0002	0.9231	0.9480	0.9797	0.0121
	SC-CGR without interaction	0.0001	0.9231	0.9382	0.9759	0.0127
	Patch-BCE	0.0120	0.3077	0.8348	0.9384	0.0100
VisA capsules	SC-CGR	0.0370	0.1406	0.5166	0.1666	0.0726
	Pairwise + invariance	0.0478	0.0781	0.4937	0.1446	0.0528
	SC-CGR without interaction	0.0335	0.0469	0.5111	0.1598	0.0677
	Patch-BCE	0.6074	0.0000	0.5820	0.2412	0.0251
VisA cashew	Pairwise + invariance	0.0009	0.9219	0.8040	0.5438	0.0264
	SC-CGR	0.0015	0.7500	0.9014	0.7270	0.0247
	SC-CGR without interaction	0.0017	0.7188	0.9043	0.7460	0.0183
	Patch-BCE	0.1432	0.0938	0.8401	0.5926	0.0104
VisA fryum	SC-CGR	0.0098	0.3125	0.7018	0.4297	0.0413
	Pairwise + invariance	0.0167	0.2031	0.7701	0.5314	0.0242
	SC-CGR without interaction	0.0111	0.2344	0.6865	0.3989	0.0422
	Patch-BCE	0.4406	0.0000	0.8256	0.6295	0.0417

E Sensitivity and Additional Controls

Contract success is thresholded by the target margin δ and neutral tolerance ϵ. Saved anchor-level responses are re-evaluated over these thresholds and read together with continuous SRAE, worst-anchor SRAE, AUROC/AP, and clean-control SRAE. Configuration selection uses validation success followed by validation SRAE, with real AP only as a retention tie-breaker.

Table 17: Metal-nut outcomes across the retained hyperparameter sweep, ordered by validation success and then validation SRAE.

Rank	SRAE	Success	AUROC	AP	Clean SRAE	Δ_1	Δ_2	Δ_{12}
1	0.0008	0.9583	0.9901	0.9976	0.0068	0.1032	0.0009	-0.0010
2	0.0009	0.9375	0.9901	0.9976	0.0072	0.1051	0.0016	-0.0009
3	0.0010	0.8958	0.9980	0.9995	0.0069	0.1018	0.0023	-0.0011
4	0.0009	0.8958	0.9901	0.9976	0.0067	0.1029	0.0014	-0.0010
5	0.0009	0.8333	0.9941	0.9986	0.0070	0.1024	0.0021	-0.0007
6	0.0009	0.8125	0.9960	0.9991	0.0071	0.1058	0.0018	-0.0011
7	0.0009	0.8125	0.9960	0.9991	0.0071	0.1058	0.0018	-0.0011
8	0.0018	0.8125	0.9980	0.9995	0.0073	0.0945	0.0024	-0.0004
9	0.0014	0.8125	0.9980	0.9995	0.0079	0.1050	0.0020	-0.0007
10	0.0009	0.7917	0.9941	0.9986	0.0072	0.1034	0.0019	-0.0012
11	0.0010	0.7708	0.9941	0.9986	0.0065	0.1026	0.0010	-0.0007
12	0.0013	0.7708	0.9980	0.9995	0.0072	0.0991	0.0024	-0.0006
13	0.0018	0.7083	0.9980	0.9995	0.0076	0.0967	0.0014	-0.0006

E.1 Additional response visualizations

The following visualizations retain the full response-oriented views used to inspect the representative repair cases. They complement the compact main-text summaries; the corresponding numeric results remain in Tables 2, 3, and 17.

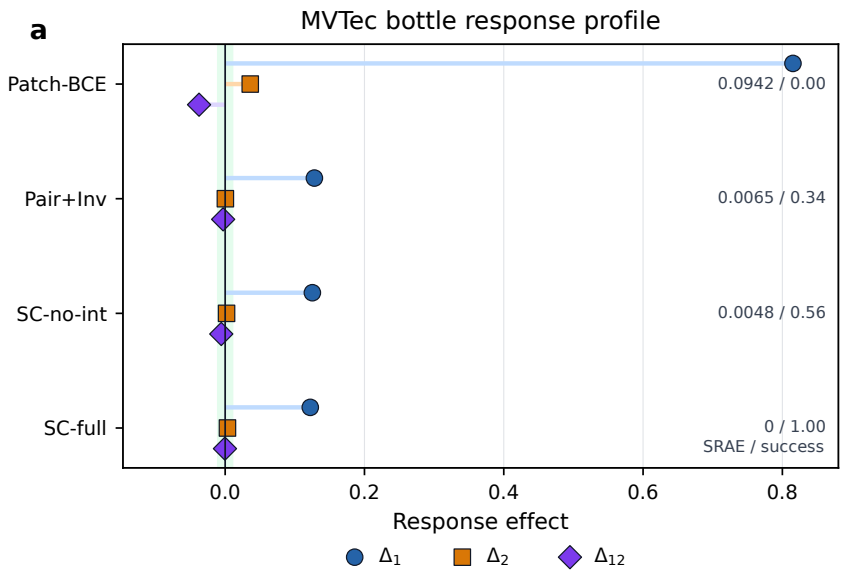

Figure 4: Bottle response ablation. Full SC-CGR jointly aligns target, tolerated-factor, and interaction effects.

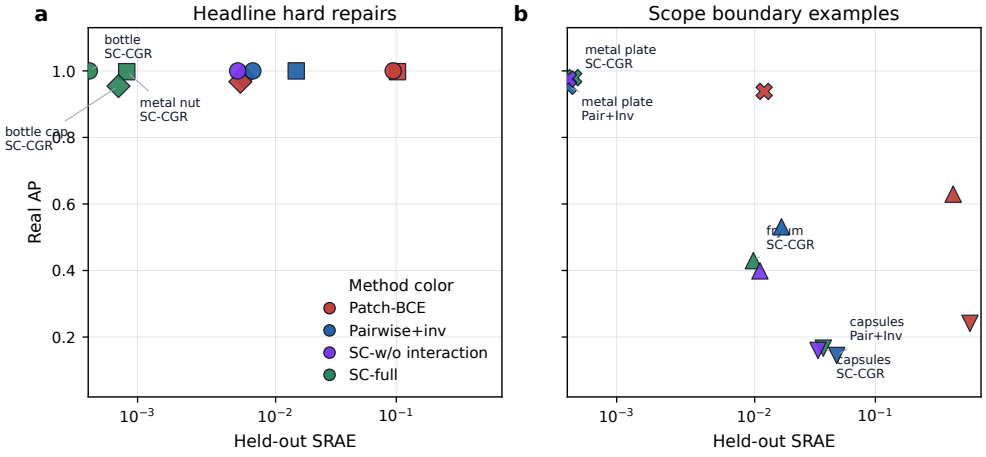

Figure 5: Semantic repair and real-data retention across representative contracts.

E.2 Calibration and ordinary head fitting

Table 18: Post-training controls on the expanded MVTec bottle protocol.

Method	SRAE	Success	AUROC	AP	Clean SRAE
Affine score calibration	0.0486	0.0000	1.0000	1.0000	–
Linear binary head	0.2405	0.0000	0.9903	0.9970	0.0123
SC-CGR	0.0000	1.0000	1.0000	1.0000	0.0095

Table 19: Expanded ordinary-supervision controls using frozen DINOv2 features and no labeled-anomaly ranking retention.

Dataset	Category	Method	SRAE	Success	AUROC	AP	Clean SRAE
MVTec AD	Bottle	Linear Patch-BCE	0.0384	0.0000	0.9032	0.9620	0.0053
		Nonlinear score regression	0.0681	0.0312	0.9839	0.9950	0.0053
		Linear score regression	0.0434	0.0312	0.9032	0.9639	0.0020
MVTec AD	Metal nut	Linear Patch-BCE	0.0276	0.0208	0.7885	0.9477	0.0104
		Nonlinear score regression	0.0802	0.0208	0.8043	0.9510	0.0095
		Linear score regression	0.0210	0.1042	0.9032	0.9771	0.0068
Real-IAD	Bottle cap	Linear Patch-BCE	0.1510	0.0078	0.8731	0.8851	0.0143
		Nonlinear score regression	0.0553	0.0078	0.8369	0.8807	0.0040
		Linear score regression	0.1464	0.0078	0.8762	0.8909	0.0121

F Anchor Scaling and Implementation

Table 20: Anchor scaling on Real-IAD bottle cap.

Total anchors	Optimization anchors	Evaluation anchors	Held-out SRAE / Success
96	48	48	0.0118 / 0.3542
256	192	64	0.0021 / 0.6719
512	384	128	0.0007 / 0.8438

Table 21: Repair architecture and training settings.

Component	Setting
Feature extractor	Pretrained frozen DINOv2 small, patch size 14
Input and patch grid	224×224 pixels; 16×16 tokens
Feature dimension	384
Repair head	Linear$(384, 64)$, ReLU, Linear$(64, 1)$; top-5% patch pooling
Optimization	AdamW, learning rate 10^{-3}
Training length	800 epochs for main cases; 500 for multi-category studies
Regularization	$\lambda_w = 10^{-5}$
Hard-anchor update	SRAE-ranked active set; fraction 0.25 unless varied
Replay types	Real good, clean control, already-correct, optional real anomaly
Inference cost	One small patch head after shared frozen feature extraction

G Additional Limitations

Counterfactual probes approximate controlled interventions and can inherit artifacts from the generation process. Factor isolation, shared-anchor matching, support masks, clean controls, and held-out anchors reduce this risk but do not replace validation with physical inspection data. A contract also represents one inspection context; transferring it across products or factories requires renewed expert review. Finally, the current repair modifies image-level anomaly scoring through a lightweight patch head. Pixel-level localization and end-to-end representation changes require separate evaluation.