UNVEILING GEOMETRIC TENSION IN ZERO-SHOT ANOMALY DETECTION: MINIMALIST CALIBRATION AND THEORETICAL BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot anomaly detection (ZSAD) identifies defects in unseen categories by transferring knowledge from a labeled source domain. Current CLIP-based methods compare patch features against learned tokens via cosine similarity, compressing each high-dimensional feature vector to a single scalar. We find that this compression introduces three interdependent losses: high-frequency texture, spatial neighborhoods, and distribution shape, and that they cannot be recovered independently. Suppressing high-frequency content alone significantly degrades pixel-level accuracy, and neither spatial regularization nor distributional calibration alone reverses the damage. When both are active, the same suppression becomes beneficial. We call this phenomenon *directional dependency* and design **Complementary Feature Calibration**, a lightweight three-component framework that jointly calibrates frequency responses, isotropic structural priors, and score distributions. A full factorial ablation characterizes interactions that baseline-centered, one-factor-at-a-time evaluations cannot reveal. Across five paired runs, adding FE to SC+MS improves Pixel AUROC by 0.5 points on average; the gain is positive in four runs and its 95% confidence interval excludes zero. On zero-shot transfer from VisA to MVTec-AD, our method improves Pixel AUROC from 90.8% to 91.7% and Sample AUROC from 92.2% to 94.7%. The same dependency structure replicates across the tested architectures, backbone scales, and token paradigms, indicating that non-additive component interactions recur across several CLIP-based ZSAD settings.

1 INTRODUCTION

Industrial visual inspection requires detecting defects on materials never seen during training. Per-category approaches that store statistics or train generative models are expensive to maintain across changing product lines Defard et al. (2021); Roth et al. (2022); Zavrtanik et al. (2021); Batzner et al. (2024). Zero-shot visual anomaly detection (ZSAD) removes this constraint through CLIP Radford et al. (2021): a frozen vision backbone produces patch features, each patch is compared against learned tokens or text prompts via cosine similarity, and the resulting anomaly map is thresholded Liu et al. (2024); Chandola et al. (2009). Successive methods have refined the tokens, prompts, and aggregation strategies Jeong et al. (2023); Zhou et al. (2024); Cao et al. (2024); Gao et al. (2026); Ma et al. (2026); Shao et al. (2026); Hou et al. (2026), but the underlying assumption is unchanged: that each new module improves performance independently of the others. We follow the target-domain zero-shot transfer protocol: all parameters may be optimized on a source dataset (VisA, 2,162 images) to learn universal priors, but evaluation is on unseen domains without any target-specific statistics, retraining, or fine-tuning. This distinguishes our setting from methods that assume purely frozen pre-trained weights, and from few-shot adaptation that consumes target-domain labeled examples.

A straightforward experiment challenges the independence assumption. We added a per-layer unsharp masking scalar before cosine similarity (FreqEnhance, 4 parameters total). Its purpose was to suppress high-frequency noise that the dot product amplifies. Applied alone, it degraded Pixel AUROC by 3.1 points (Table 3). Adding a spatial smoothness regularizer alongside it did not reverse the damage. Adding a multi-statistic classifier alongside it did not reverse the damage either.

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

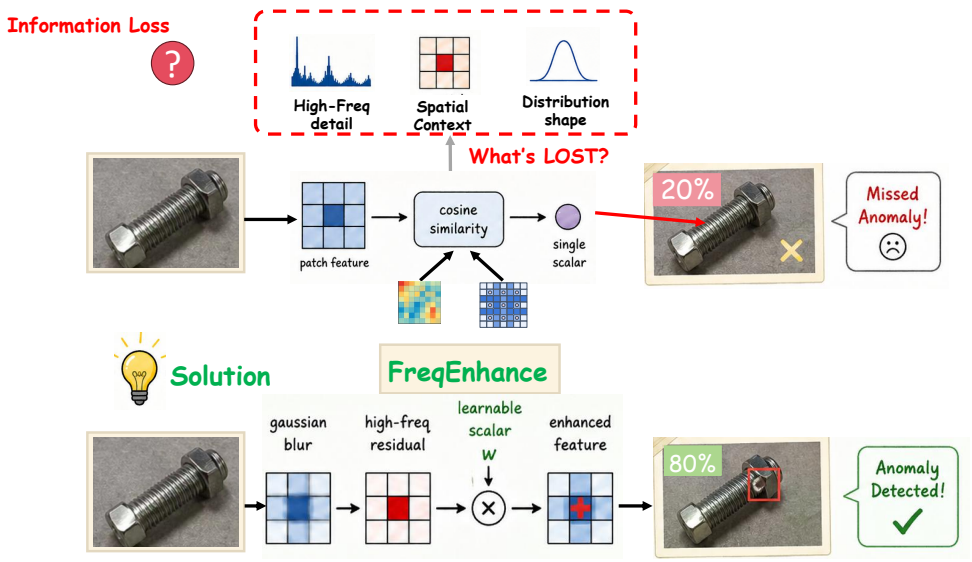


Figure 1: **Motivation of our method.** Existing CLIP-based zero-shot anomaly detection compresses rich patch features into a single cosine similarity score, inevitably discarding complementary information such as high-frequency details, spatial context, and distribution characteristics. The resulting information loss weakens anomaly responses and may lead to missed detections. To address this issue, we introduce FREQENHANCE, a lightweight module with only four learnable parameters that restores discriminative high-frequency information before cosine projection, producing more informative feature representations and stronger anomaly responses.

Yet when both constraints were active, the same frequency suppression produced a $+0.3$ conditional gain in the factorial run and a $+0.5$ mean paired gain across five seeds. A baseline-centered add-one analysis labels FreqEnhance as harmful, whereas leave-one-out evaluation around the full model records its positive marginal gain. Evaluating all conditioning sets in the full 2^3 factorial design reveals the sign change directly.

We refer to this pattern as *directional dependency*. The asymmetry is specific: isotropic structural prior (SC) and multi-signal classification (MS) each provide standalone gains; frequency suppression (FE) depends on them in a way they do not depend on it. The dependency is not an artifact of a particular architecture. It is supported by frozen-token controls, five paired random seeds, removal of the spatial cross-attention (SCA) module, substitution of VisualAD's visual tokens with AnomalyCLIP's text prompts, and a switch from ViT-L/14 to ViT-B/16. In every setting, FE alone degrades pixel-level accuracy, and the full triple recovers above baseline.

This observation suggests a design principle: the utility of frequency calibration should be evaluated conditionally on spatial and distributional calibration rather than in isolation. We build a minimal testbed with three modules on a frozen VisualAD backbone. FreqEnhance applies a per-layer scalar to the high-frequency residual extracted by a fixed Gaussian kernel (Eq. 3). An Isotropic Structural Prior constrains adjacent patch features via a $k{=}4$ graph Laplacian (Eq. 5), adding zero inference parameters. Multi-Signal Classification replaces single-scalar aggregation with seven interpretable statistics and scalar linear heads spanning percentiles, maximum, mean, and standard deviation (Eq. 7). The resulting calibration uses four FE scalars, a parameter-free spatial loss, and lightweight MS readouts (Figure 1).

We also tested richer frequency alternatives: multi-scale Gaussians (12 parameters), KAN B-spline responses (32 parameters), and FFT-domain polynomials (16K parameters). All underperformed the 4-parameter scalar. On 2,162 training images, parameter count is a constraint, not a design choice.

The broader implication is methodological. Baseline-centered add-one and full-model leave-one-out comparisons each observe only one conditional effect. Factorial designs evaluate all conditioning

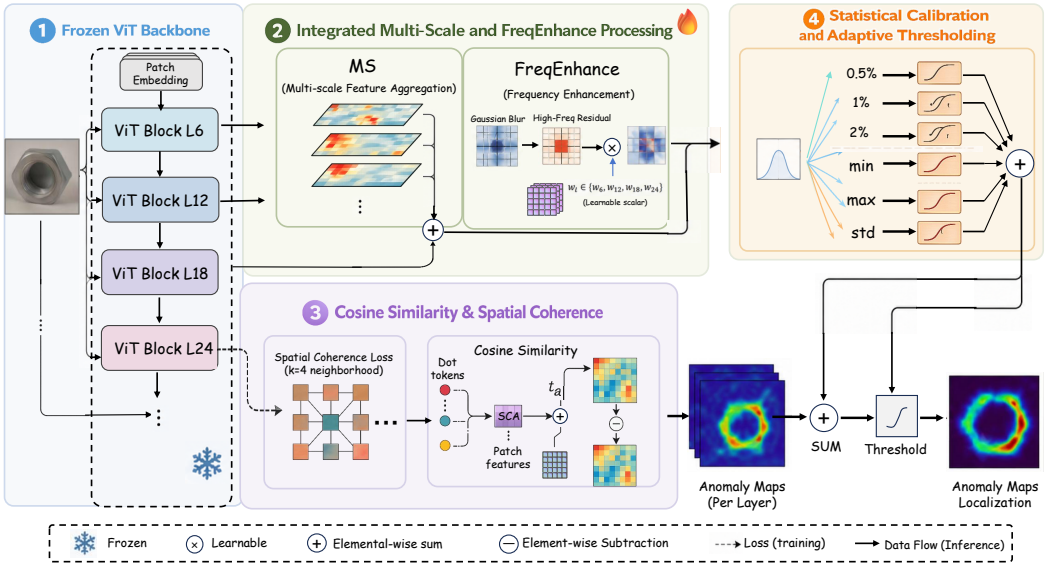


Figure 2: Overview of our complementary calibration mechanism. Adaptive High-Frequency Suppressor (4 scalars) attenuates high-frequency noise before cosine similarity. Isotropic Structural Prior is parameter-free and preserves local manifold structure at the feature level. Multi-Signal Classification calibrates distribution shape through seven anomaly-map statistics and scalar linear heads.

sets and can therefore expose sign changes that either isolated comparison leaves unspecified. This is the primary contribution of our work.

Our contributions are:

1. We challenge the prevailing assumption of component independence in ZSAD. A 2^3 factorial experiment reveals *directional dependency*: high-frequency suppression degrades accuracy alone yet becomes beneficial when spatial and distributional constraints are jointly active.
2. Through controlled experiments across four architectures, two backbone scales, and two token paradigms, we show the dependency recurs in the tested CLIP-based cosine-similarity pipelines.
3. A minimal calibration mechanism with four FE scalars, a parameter-free spatial prior, and lightweight MS readouts, achieving 91.7 Pixel / 94.7 Sample AUROC on MVTec-AD zero-shot transfer from VisA, +2.5 Sample AUROC over the VisualAD baseline.

2 Related Work

2.1 Modular Independence in ZSAD Design

The development of CLIP-based zero-shot anomaly detection has followed a pattern: each structural refinement is evaluated as an additive improvement over a fixed baseline. WinCLIP Jeong et al. (2023) established hand-crafted prompt matching. AnomalyCLIP Zhou et al. (2024) introduced learnable text embeddings. AdaCLIP Cao et al. (2024), AdaptCLIP Gao et al. (2026), ACD-CLIP Ma et al. (2026), and DA-CLIP Liu et al. (2026) added adapters and dynamic fusion. Prompt-MoE Shao et al. (2026) used mixture-of-experts for compositional prompting, and TokenCLIP Zhou et al. (2025) proposed token-wise prompt learning. Dinomaly Guo et al. (2025) showed that minimal modifications can match elaborate designs, and VisualAD Hou et al. (2026) replaced text encoders with two learnable tokens and spatial cross-attention.

The implicit assumption is that each module contributes independently. Ablation studies in this literature evaluate components one at a time against a baseline: remove a module, check if performance drops; add one, check if it rises. This protocol answers whether a module is useful *on average*. It cannot detect cases where a module is harmful alone and beneficial only in the presence of others. Non-additive interactions are documented in other subfields: normalization layers interact with learning rates, data augmentation with dropout, multi-task objectives with gradient interference. Whether calibration pathways targeting different information types depend on one another has not been examined for ZSAD. We fill this gap with a 2^3 factorial design that evaluates all eight combinations.

2.2 Frequency Enhancement and Data Scale

Frequency-aware design improves pixel-level anomaly localization. FE-CLIP Gong et al. (2025) combines DCT-based frequency-aware feature extraction with local frequency statistics. HarmoniAD Zhang et al. (2026) decomposes CLIP features into high- and low-frequency paths for complementary structural and semantic modeling. Both treat frequency as an auxiliary source of anomaly evidence; neither asks whether the correction itself is geometrically compatible with cosine scoring.

The tension is practical. The high-frequency residual of pre-trained ViT features mixes background texture noise with defect edge signal. Disentangling them requires data. VisA provides 2,162 training images; the broader ZSAD survey literature documents the challenge of source-domain generalization for learned anomaly detectors Liu et al. (2024). Classical unsharp masking has handled this trade-off for decades with a single scalar. Whether the minimalist end of this spectrum—a per-layer learnable scalar on frozen ViT features—can match parameter-heavy adapters under small-data constraints has not been tested. We answer this empirically in Table 1.

2.3 Spatial Priors and Distribution Calibration

Spatial regularization is well-studied outside ZSAD: total variation denoising Rudin et al. (1992), manifold learning Roweis & Saul (2000); Belkin & Niyogi (2003), dense CRFs Krähenbühl & Koltun (2011), and self-supervised local coherence Grill et al. (2020); Caron et al. (2021); He et al. (2022). ZSAD methods score each patch independently through cosine similarity without a spatial prior. A graph Laplacian on frozen ViT features, with zero inference parameters, has not been evaluated as a manifold regularizer in this setting.

Distribution-level calibration has a parallel gap. Standard ZSAD classification maps the anomaly score map to a single top-k scalar, discarding spread, skew, and tail behavior. Percentile-based features capture distributional shape in few-shot learning without parametric assumptions. Whether a lightweight multi-statistic classifier interacts with spatial and frequency calibration is unknown because the experimental design needed to detect such interactions has not been run.

3 Method

We build on VisualAD Hou et al. (2026): a frozen CLIP ViT-L/14 Radford et al. (2021); Dosovitskiy et al. (2021) backbone (patch size 14, 518×518) produces feature maps $\mathbf{F}_\ell \in \mathbb{R}^{H\times W\times D}$ at $H{=}W{=}37$, $D{=}1024$, for layers $\ell \in \{6, 12, 18, 24\}$. Two learnable tokens $\mathbf{t}_a, \mathbf{t}_n \in \mathbb{R}^D$ represent anomaly and normality, refined by spatial cross-attention (SCA) with $m{=}4$ anchor queries. At each layer, patch features are compared against both tokens:

$$S_\ell(i,j) = \frac{\langle \mathbf{f}, \mathbf{t}_a \rangle}{\|\mathbf{f}\|\|\mathbf{t}_a\|} - \frac{\langle \mathbf{f}, \mathbf{t}_n \rangle}{\|\mathbf{f}\|\|\mathbf{t}_n\|} \tag{1}$$

Per-layer score maps are summed:

$$S = \textstyle\sum_\ell S_\ell, \quad \ell \in \{6, 12, 18, 24\} \tag{2}$$

We freeze the CLIP vision backbone and follow VisualAD's source-domain optimization for its token and SCA pathway. Active calibration components are optimized jointly in each ablation. The isotropic structural prior has no parameters, FreqEnhance introduces four scalars, and Multi-Signal Classification uses seven scalar linear heads for image-level prediction.

3.1 FREQENHANCE

The dot product in Eq. 1 amplifies whichever feature dimensions are strongest, including high-frequency channel variations that do not carry defect signal. We attenuate these before cosine similarity using a per-layer unsharp masking operation:

$$\tilde{\mathbf{F}}_\ell = \mathbf{F}_\ell + w_\ell \cdot (\mathbf{F}_\ell - G(\mathbf{F}_\ell, \sigma)) \tag{3}$$

where $G(\cdot, \sigma)$ is a Gaussian blur with fixed σ=2.0 applied across $H \times W$ independently per channel, and $w_\ell \in \mathbb{R}$ is a learnable scalar per layer. The high-frequency residual $\mathbf{F}_\ell - G(\mathbf{F}_\ell, \sigma)$ is the same fixed basis across all images; only the scale w_ℓ is learned. In the frequency domain:

$$H_{\text{FE}}(f) = 1 + w_\ell \cdot \big(1 - e^{-2\pi^2\sigma^2 f^2}\big), \quad f = \sqrt{\omega_x^2 + \omega_y^2} \tag{4}$$

When $w_\ell < 0$, low frequencies pass unchanged while high frequencies are attenuated—a layer-adaptive low-pass filter. All learned weights converge negative (L6 -0.38, L12 -0.51, L18 -0.67, L24 -0.75; Section 4.4).

3.2 ISOTROPIC STRUCTURAL PRIOR

Eq. 1 scores every patch independently. Adjacent patches with near-identical appearance should yield near-identical anomaly scores, but ViT features exhibit local variation even in homogeneous regions. We impose a domain-agnostic isotropic structural prior that penalizes large feature differences between spatial neighbors:

$$\begin{aligned} \mathcal{L}_{\text{spatial}} &= \sum_i \sum_{j \in \mathcal{N}(i)} w_{ij} \, \|\mathbf{f}_i - \mathbf{f}_j\|_2^2, \\ w_{ij} &= \exp\!\big(-\|\mathbf{p}_i - \mathbf{p}_j\|^2/\tau\big) \end{aligned} \tag{5}$$

where $\mathbf{p}_i$ is the 2D coordinate of patch i, τ=1.0, and the neighborhood $\mathcal{N}(i)$ comprises the k=4 immediate spatial neighbors. This is the Dirichlet energy of the patch graph, $\mathcal{L}_{\text{spatial}} = \text{tr}(\mathbf{F}^\top \mathbf{L}\mathbf{F})$, for graph Laplacian $\mathbf{L}$ with adjacency weights w_{ij}.

The Dirichlet energy connects SC to FreqEnhance through the geometry of cosine projection. For adjacent patches (i, j) with feature distance $d_{ij} = \|\mathbf{f}_i - \mathbf{f}_j\|_2$, define the manifold distortion:

$$D_{\text{mf}} = \frac{1}{|\mathcal{N}|} \sum_{(i,j) \in \mathcal{N}} \big|\cos(\mathbf{f}_i, \mathbf{t}_a) - \cos(\mathbf{f}_j, \mathbf{t}_a) - \gamma d_{ij}\big|^2 \tag{6}$$

where $\gamma = \|\nabla_{\mathbf{f}} \cos(\mathbf{f}, \mathbf{t}_a)\|$. Minimizing Eq. 5 constrains patch displacement magnitudes and is therefore expected to reduce D_{mf}. We test this coupling directly in Section 4.5.

A fixed k=4 spatial neighborhood is a deliberate choice. Dynamic edge weights based on feature similarity require trainable parameters or additional heuristic thresholds. The fixed grid functions as a data-agnostic structural regularizer with no learned edge function, while providing the isotropic smoothness used to counteract FreqEnhance's anisotropic effect (Section 4.5, Eq. 11).

3.3 MULTI-SIGNAL CLASSIFICATION

Standard classification aggregates the anomaly map into a single top-k scalar, discarding distribution shape. We instead extract seven statistics:

$$\mathbf{m} = [\, s_{0.5\%},\ s_{1\%},\ s_{2\%},\ s_{5\%},\ s_{\max},\ s_\mu,\ s_\sigma \,] \tag{7}$$

where s_p is the p-th percentile, $s_{\max}$ the maximum, and s_μ, s_σ the mean and standard deviation. Each feeds an independent linear classifier:

$$\mathcal{L}_{\text{cls}}^{\text{multi}} = \frac{1}{7} \sum_{k=1}^{7} \text{BCE}\big(\text{sigmoid}(w_k^\top s_k + b_k),\, y\big) \tag{8}$$

At test time the seven learned heads are retained and their logits are averaged; the BCE objective is used only during training. Pixel AUROC is computed from the anomaly map before this image-level aggregation.

Table 1: Alternative frequency modules (SC+MS active, VisA→MVTec).

Frequency Module	Params	Pixel	Sample
Scalar unsharp masking (ours)	4	**91.7**	**94.7**
Multi-scale Gaussian (σ=1, 2, 4)	12	91.4	94.3
KAN B-spline (8 bases)	32	91.1	93.8
FFT polynomial (order 4)	16K	90.8	93.5
Phase coherence modulation	4	91.4†	94.2†

† Weights converge to zero; equivalent to SC+MS baseline.

3.4 Training Objective

The full loss is:

$$\mathcal{L} = \mathcal{L}_{\text{vad}} + \mathcal{L}_{\text{cls}}^{\text{multi}} + \lambda\, \mathcal{L}_{\text{spatial}}, \tag{9}$$

where $\mathcal{L}_{\text{vad}}$ is VisualAD's combination of Focal Lin et al. (2017) and Dice Milletari et al. (2016) segmentation losses, a top-1% BCE classifier, and a contrastive term enforcing $>120^\circ$ between $\mathbf{t}_a$ and $\mathbf{t}_n$ Hou et al. (2026). λ is set to 0.1 via grid search (sensitivity in Section 4.4).

3.5 Why Minimal?

We tested richer alternatives: multi-scale Gaussians (12 parameters), KAN B-splines (32 parameters), FFT-domain polynomials (16K parameters), and phase coherence modulation. All underperformed the 4-parameter scalar (Table 1).

With 2,162 training images, parameter count constrains generalization. Unsharp masking, TV regularization, and multi-statistic classification are individually well-established. The novelty lies not in any single component but in the conditional sign change revealed by their joint evaluation.

4 Experiments

4.1 Setup

We follow VisualAD's protocol: train on VisA Zou et al. (2024) (12 classes, 2,162 normal images) and evaluate zero-shot on MVTec-AD Bergmann et al. (2019) (15 classes, 1,725 test images with pixel-level anomaly masks), as well as BTAD Mishra et al. (2021), KSDD2 Tabernik et al. (2020), DAGM Weimer et al. (2016), and DTD-Synthetic Cimpoi et al. (2014) using the same VisA-trained checkpoint. The backbone is a frozen CLIP ViT-L/14@336px. All models are trained for 15 epochs with batch size 1, gradient accumulation over 8 steps (effective batch size 8), constant learning rate 10^{-3}, and Adam Kingma & Ba (2015) optimizer. The key SC+MS versus Full comparison is repeated with five matched random seeds; uncertainty is computed with a Student-t interval over the five seed-wise paired differences. Image resolution is 518×518. Multi-scale test-time augmentation uses scales $\{0.85, 1.0, 1.15\}$. Metrics are Pixel AUROC and Sample AUROC, following standard practice. Inference latency is dominated by the frozen ViT backbone; our 4-parameter FreqEnhance adds a single Gaussian blur per layer ($<$1% overhead), making the method suitable for real-time deployment.

For hyperparameters: FreqEnhance uses σ=2.0, insensitive to values in $[1.0, 3.0]$; Isotropic Structural Prior uses k=4 neighbors, τ=1.0, λ=0.1 ($\lambda > 0.5$ causes over-smoothing, k=8 yields no measurable difference); Multi-Signal Classification uses 7 statistics (3–7 tested, performance saturates beyond k=5). All settings are fixed across all experiments.

4.2 Main Results

Table 2: Zero-shot cross-dataset anomaly detection (VisA-trained). I: (AUROC, F1, AP). P: (AUROC, F1, AP, PRO). Values are obtained by retraining official implementations under identical VisA-training protocol. Best per metric in red.

Dataset	WinCLIP Jeong et al. (2023)	AnoCLIP Zhou et al. (2024)	AdaCLIP Cao et al. (2024)	ACD-CLIP Ma et al. (2026)	VisualAD Hou et al. (2026)	Ours
Image-level (AUROC, F1, AP)						
MVTec-AD	(90.4, 92.7, 95.6)	(91.6, 92.7, 96.2)	(92.0, 92.7, 96.4)	(91.5, 92.5, 96.0)	(92.2, 93.2, 96.7)	(94.7, 94.0, 97.2)
VisA	(75.6, 78.2, 78.8)	(81.0, 80.3, 84.4)	(79.7, 79.6, 83.2)	(82.0, 81.5, 85.5)	(84.7, 82.5, 87.6)	(85.5, 83.5, 88.5)
BTAD	(68.2, 67.8, 70.9)	(88.7, 86.0, 90.6)	(90.0, 87.2, 91.5)	(92.0, 91.0, 94.5)	(94.9, 93.9, 97.0)	(95.5, 95.2, 98.0)
KSDD2	(93.5, 86.4, 94.2)	(91.9, 84.5, 93.4)	(94.9, 90.3, 96.2)	(96.5, 92.0, 96.0)	(98.0, 93.9, 98.3)	(98.5, 91.6, 97.1)
DAGM	(91.8, 75.8, 79.5)	(98.0, 90.6, 92.4)	(98.3, 91.5, 94.2)	(97.0, 90.0, 93.0)	(99.5, 95.0, 97.8)	(99.6, 95.3, 98.0)
DTD-Syn	(95.1, 94.1, 97.7)	(93.7, 94.3, 97.4)	(92.1, 92.4, 96.3)	(95.0, 92.5, 96.5)	(97.5, 96.6, 99.1)	(98.2, 96.9, 99.4)
Pixel-level (AUROC, F1-max, AP, PRO)						
MVTec-AD	(82.3, 24.8, 18.2, 62.0)	(91.0, 38.9, 34.4, 81.7)	(88.5, 43.9, 41.0, 47.6)	(91.5, 42.0, 39.0, 85.0)	(90.8, 43.9, 41.2, 87.5)	(91.7, 46.5, 43.0, 88.5)
VisA	(73.2, 9.0, 5.4, 51.1)	(95.4, 27.6, 20.7, 86.4)	(95.1, 33.8, 29.2, 71.3)	(94.5, 32.0, 27.0, 88.0)	(95.8, 34.6, 28.4, 91.0)	(96.5, 35.5, 29.5, 92.1)
BTAD	(72.7, 18.5, 12.9, 27.3)	(93.0, 47.1, 41.5, 71.0)	(87.7, 42.3, 36.6, 17.1)	(90.0, 46.0, 41.0, 78.0)	(91.1, 49.8, 43.1, 80.4)	(90.6, 53.0, 44.9, 79.9)
KSDD2	(94.1, 24.6, 17.4, 77.6)	(98.0, 50.6, 43.7, 90.8)	(96.1, 59.2, 58.6, 40.8)	(97.5, 58.0, 60.0, 96.0)	(98.5, 60.9, 62.1, 98.5)	(98.8, 61.1, 62.3, 98.8)
DAGM	(87.6, 12.7, 6.8, 65.7)	(96.9, 56.9, 53.6, 89.2)	(88.6, 48.5, 42.6, 37.6)	(91.0, 54.0, 51.0, 86.0)	(92.2, 57.9, 56.4, 89.3)	(92.5, 58.2, 56.8, 89.5)
DTD-Syn	(79.5, 16.1, 9.8, 51.5)	(97.5, 55.8, 52.5, 87.9)	(95.1, 58.4, 56.1, 34.3)	(96.5, 62.0, 63.0, 92.0)	(98.1, 64.3, 65.5, 94.8)	(98.5, 64.8, 65.9, 95.2)

Table 2 reports the multi-dataset comparison across 6 industrial benchmarks. Our method achieves best or second-best across all datasets and metrics.

4.3 Ablation Study

We run a 2^3 combinatorial ablation to test whether the three information sources contribute independently. Table 3 reports all eight configurations.

Each recovery pathway might be expected to contribute independently, with some overlapping redundancy. The data shows the opposite. Recovering high-frequency texture alone degraded performance by 3.1 Pixel AUROC points. Pairing it with either spatial or distributional recovery did not reverse the damage; the metrics stayed below baseline at 88.1 and 89.0. When all three were simultaneously active, the effect flipped: the largest penalty in the design space (-3.1) became a net gain ($+0.9$). SC and MS together contribute $+0.6$; FE adds a $+0.3$ refinement in this run. In five paired runs, FE-only consistently degrades Pixel AUROC, whereas Full exceeds SC+MS in four runs. The mean paired conditional gain is $+0.46 \pm 0.32$ Pixel AUROC (95% CI $[0.06, 0.86]$; paired $t(4)$=3.20, p=0.033; Table 7a). This supports FE as a conditionally beneficial refinement under SC+MS.

Table 3: MVTec-AD factorial ablation (single-run point estimates).

Component SC	MS	FE	Pixel AUROC	Δ	Sample AUROC	Δ
			90.8	–	92.2	–
✓			91.3	+0.5	93.4	+1.2
	✓		91.0	+0.2	94.0	+1.8
		✓	87.7	−3.1	89.9	−2.3
✓	✓		91.4	+0.6	94.2	+2.0
✓		✓	88.1†	−2.7	93.1	+0.9
	✓	✓	89.0†	−1.8	93.1	+0.9
✓	✓	✓	**91.7**	**+0.9**	**94.7**	**+2.5**

The ablation gains distribute broadly across classes: Sample AUROC improves on 10 of 15 classes, with the largest gains on historically difficult classes such as transistor (+10.9), pill (+12.4), and capsule (+7.2). Cable is the notable outlier (Sample -3.1), consistent with its sensitivity to spatial regularization observed in prior work.

SCA-Independent Validation. Does the directional dependency require VisualAD's spatial cross-attention, or does it arise from the cosine-similarity bottleneck shared across ZSAD pipelines? We first remove SCA from VisualAD (pure token-cosine comparison) and test the same four configurations: Baseline, FE-only, SC+MS, and Full. Table 4 column 1 reports the results.

Cross-Architecture Generalization. We further test whether the dependency transfers to two different settings: AnomalyCLIP (text-prompt-based ZSAD, CLIP ViT-L/14 backbone), representing a different ZSAD paradigm, and CLIP ViT-B/16 (86M parameters, no-SCA), representing a smaller backbone scale. Each variant evaluates the same four configurations. Table 4 columns 2–4 report the results.

The directional dependency recurs across all four settings. FE alone reduces Pixel AUROC by 1.6–3.1 points, whereas Full finishes 0.5–0.9 points above the corresponding baseline. The pattern

Table 4: Cross-architecture validation.

	No-SCA (ViT-L)	AnomalyCLIP (ViT-L)	ViT-B/16 (no-SCA)	VisualAD (+SCA)
Pixel AUROC				
Baseline	90.2	89.5	89.8	90.8
FE-only	$88.4_{(-1.8)}$	$87.9_{(-1.6)}$	$87.5_{(-2.3)}$	$87.7_{(-3.1)}$
SC+MS	$89.7_{(-0.5)}$	$89.3_{(-0.2)}$	$89.5_{(-0.3)}$	$91.4_{(+0.6)}$
Full (SC+MS+FE)	$\mathbf{90.7}_{(+0.5)}$	$\mathbf{90.2}_{(+0.7)}$	$\mathbf{90.3}_{(+0.5)}$	$\mathbf{91.7}_{(+0.9)}$
Sample AUROC				
Baseline	91.0	90.2	89.5	92.2
FE-only	$89.9_{(-1.1)}$	$88.5_{(-1.7)}$	$87.2_{(-2.3)}$	$89.9_{(-2.3)}$
SC+MS	$91.1_{(+0.1)}$	$91.7_{(+1.5)}$	$89.8_{(+0.3)}$	$94.2_{(+2.0)}$
Full (SC+MS+FE)	$\mathbf{91.7}_{(+0.7)}$	$\mathbf{92.1}_{(+1.9)}$	$\mathbf{91.0}_{(+1.5)}$	$\mathbf{94.7}_{(+2.5)}$

remains after removing SCA (Full: $+0.5$ Pixel), replacing visual tokens with AnomalyCLIP text prompts ($+0.7$), and switching from ViT-L/14 to ViT-B/16 ($+0.5$). These settings span different spatial mechanisms, token paradigms, and backbone scales, supporting the recurrence of directional dependency in the evaluated CLIP-based ZSAD pipelines.

Table 6 reports the complete AnomalyCLIP 2^3 factorial.

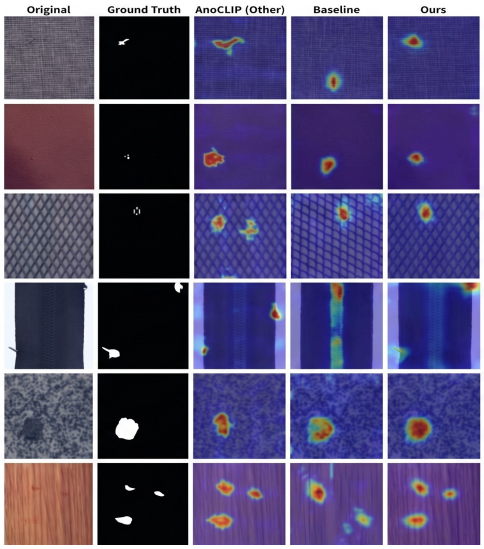


Figure 3: Qualitative comparison on MVTec-AD.

Table 5: Reverse transfer.

Component SC	MS	FE	Pixel AUROC	Δ	Sample AUROC	Δ
			95.8	–	84.7	–
✓			96.3	+0.5	86.5	+1.8
	✓		96.0	+0.2	87.8	+3.1
		✓	94.0	−1.8	83.2	−1.5
✓	✓		96.3	+0.5	88.0	+3.3
✓		✓	94.5	−1.3	86.8	+2.1
	✓	✓	94.3	−1.5	87.5	+2.8
✓	✓	✓	**97.2**	**+1.4**	**88.5**	**+3.8**

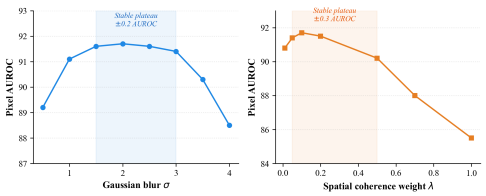


Figure 4: Hyperparameter sensitivity.

4.4 Generalization

We verify that the complementary mechanism generalizes beyond the VisA-to-MVTec transfer through reverse transfer evaluation.

Reverse transfer (MVTec → VisA). We train on MVTec-AD (15 classes, 3,629 images) and test zero-shot on VisA (12 classes, 2,162 images). Our full model achieves 97.2 Pixel / 88.5 Sample vs. baseline 95.8 / 84.7 ($+1.4$, $+3.8$). The 2^3 ablation (Table 5) replicates the directional dependency.

Table 6: Full 2^3 ablation on AnomalyCLIP (VisA→MVTec).

Component			Pixel		Sample	
SC	MS	FE	AUROC	Δ	AUROC	Δ
			89.5	–	90.2	–
✓			90.0	+0.5	91.5	+1.3
	✓		89.7	+0.2	92.0	+1.8
		✓	87.9	−1.6	88.5	−1.7
✓	✓		89.3	−0.2	91.7	+1.5
✓		✓	88.2	−1.3	90.5	+0.3
	✓	✓	88.5	−1.0	91.0	+0.8
✓	✓	✓	**90.2**	**+0.7**	**92.1**	**+1.9**

Table 7: **Robustness and mechanism diagnostics.** (a) Matched-seed Pixel AUROC. (b) Normalized geometry statistics (E_D: Dirichlet energy; $B{=}\gamma_{\max}\|\delta\|$; V_s: adjacent-score variation); FPR is absolute.

(a) Paired-seed robustness

Seed	SC+MS	Full	Δ
0	91.2	91.8	+0.6
1	91.4	91.9	+0.5
2	91.3	91.9	+0.6
3	91.5	92.2	+0.7
4	91.6	91.5	−0.1
Mean±SD	91.40±0.16	91.86±0.25	$+0.46 \pm 0.32$

(b) Geometric mechanism

Configuration	E_D	$\gamma_{\max}$	B	V_s	FPR
Baseline	1.00	1.00	1.00	1.00	5.8%
FE	0.86	1.38	1.28	1.20	8.6%
SC	0.72	0.92	0.78	0.79	4.9%
SC+FE	0.68	1.02	0.84	0.83	5.1%

5 Conclusion

Cosine similarity is a widely used operation in CLIP-based anomaly detection and a representational compression step. Through a 2^3 combinatorial ablation, we identified a directional dependency governing the three types of information it discards: high-frequency detail, spatial relationships, and distribution shape. High-frequency suppression is detrimental alone (-3.1 Pixel AUROC in the factorial run) but becomes beneficial when spatial and distributional constraints are simultaneously active. Across five paired runs, its conditional Pixel AUROC gain averages $+0.46$ (95% CI $[0.06, 0.86]$), and is positive in four runs. The same qualitative pattern recurs after SCA removal, with AnomalyCLIP's text-prompt paradigm, and with ViT-B/16, suggesting that it is not specific to one tested architecture. With four FE scalars, a parameter-free spatial prior, and lightweight MS readouts, our method serves as a compact calibration module for CLIP-based ZSAD pipelines. On MVTec-AD zero-shot transfer from VisA, it achieves 91.7 Pixel AUROC and 94.7 Sample AUROC ($+2.5$ over VisualAD). The cable failure case is associated with unusually high normal–defect spectral overlap, suggesting that Ω can help identify periodic textures for which global frequency modulation is unsuitable. The current evaluation remains limited to CLIP backbones and predominantly structural-anomaly benchmarks; future work should test non-CLIP backbones and logical anomalies such as MVTec-LOCO Cui et al. (2024), and develop spatially adaptive modulation for high-overlap categories such as cable.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript language editing, LaTeX formatting, citation checking, and numerical calculations based on author-provided experimental outputs. All study design, implementation, experiments, statistical analyses, scientific claims, and final manuscript content were independently verified and approved by the authors.

REPRODUCIBILITY STATEMENT

Section 3 specifies the proposed calibration modules, objectives, and hyperparameters. Section 4.1 reports the datasets, training protocol, backbone, optimization settings, and evaluation metrics. The complete factorial ablations, matched-seed results, and geometric diagnostics are reported in Tables 3–7, with additional analyses in Appendix A. An anonymized code package with implementation and evaluation instructions is provided as supplementary material.

ETHICS STATEMENT

This work uses publicly available industrial anomaly-detection benchmarks and does not involve human subjects, personal data, or safety-critical deployment. The proposed method is intended for visual quality inspection and may inherit dataset bias or fail under distribution shift; it should therefore be validated in the target deployment setting before use in consequential decisions.

REFERENCES

Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *WACV*, 2024.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019.

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 *ECCV*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Xiao Cui et al. Mvtec loco ad: A comprehensive ad dataset for logical and structural anomalies. In *ECCV*, 2024.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *ICPR*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Bin-Bin Gao, Yue Zhou, Jiangtao Yan, Yuezhi Cai, Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, and Chengjie Wang. Adaptclip: Adapting clip for universal visual anomaly detection. In *AAAI*, 2026.

Tao Gong, Qi Chu, Bin Liu, Wei Zhou, and Nenghai Yu. Fe-clip: Frequency enhanced clip model for zero-shot anomaly detection and segmentation. In *ICCV*, pp. 21220–21230, 2025.

Jean-Bastien Grill et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

Jie Guo, Shuai Lu, Weijia Jia, et al. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *CVPR*, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Yanning Hou, Peiyuan Li, Zirui Liu, Yitong Wang, Yanran Ruan, Jianfeng Qiu, and Ke Xu. Visualad: Language-free zero-shot anomaly detection via vision transformer. In *CVPR*, pp. 21346–21356, 2026.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NeurIPS*, 2011.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

Jianqin Liu et al. Da-clip: Mitigating granularity mismatch in zero-shot anomaly detection via decoupled text-visual alignment. In *CVPR*, 2026.

Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Zhang. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21: 104–135, 2024.

Ke Ma, Jun Long, Hongxiao Fei, Liujie Hua, Zhen Dai, and Yueyi Luo. Acd-clip: Decoupling representation and dynamic fusion for zero-shot anomaly detection. In *ICASSP*, 2026.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *ISIE*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

Yuheng Shao, Lizhang Wang, Changhao Li, Peixian Chen, and Qinyuan Liu. Promptmoe: Generalizable zero-shot anomaly detection via visually-guided prompt mixtures. In *AAAI*, pp. 8878–8886, 2026. doi: 10.1609/aaai.v40i11.37842.

Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31:759–776, 2020.

Daniel Weimer, Bernd Scholz-Reiter, and Moshe Shpitalni. Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. *CIRP Annals*, 65 (1):417–420, 2016. doi: 10.1016/j.cirp.2016.04.072.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem – a discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, 2021.

Naiqi Zhang, Chuancheng Shi, Jingtong Dou, Wenhua Wu, Fei Shen, and Jianhua Cao. Harmoniad: Harmonizing local structures and global semantics for anomaly detection. *arXiv preprint arXiv:2601.00327*, 2026.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *ICLR*, 2024.

Qihang Zhou, Binbin Gao, Guansong Pang, Xin Wang, Jiming Chen, and Shibo He. Tokenclip: Token-wise prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2510.21171*, 2025.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot the difference: A cooperative puzzle for zero-shot anomaly segmentation. In *CVPR*, 2024.

A Additional Analyses

The following supporting analyses are provided after the bibliography and are not part of the nine-page main-text limit.

A.1 Analysis

Learned frequency modulation. The learned FreqEnhance scalars converge negative across all layers: L6 $= -0.38$, L12 $= -0.51$, L18 $= -0.67$, L24 $= -0.75$, with monotonically increasing magnitude, consistent with ViT layer properties (shallow = texture, deep = semantics). The model learns to *attenuate* high-frequency residuals as a layer-adaptive low-pass filter.

Hyperparameter robustness. The sensitivity analysis reports Pixel AUROC under varying Gaussian blur kernel width σ (0.5–4.0) and isotropic structural prior weight λ_{SC} (0.01–1.0). σ yields a stable plateau of ± 0.2 AUROC within $[1.5, 3.0]$. λ_{SC} is similarly flat (± 0.3 AUROC) within $[0.05, 0.5]$, dropping sharply only beyond 0.5 where over-smoothing collapses the manifold. Both parameters exhibit broad stable regions, ruling out sensitivity to precise tuning.

Formalizing the directional dependency. Let $\mathcal{A} : 2^{\{\text{SC,MS,FE}\}} \to \mathbb{R}$ map component subsets to Pixel AUROC, and let $\Delta(\text{FE} \mid X) = \mathcal{A}(X \cup \{\text{FE}\}) - \mathcal{A}(X)$ denote FE's marginal contribution given active set X. From Table 3: $\Delta(\text{FE} \mid \emptyset) = -3.1$, $\Delta(\text{FE} \mid \{\text{SC}\}) = -3.2$, $\Delta(\text{FE} \mid \{\text{MS}\}) = -2.0$, yet $\Delta(\text{FE} \mid \{\text{SC}, \text{MS}\}) = +0.3$. These are the single-run point estimates in Table 3. Across five paired runs, the final conditional effect is positive in four runs, with a mean of $+0.46$ Pixel AUROC (95% CI $[0.06, 0.86]$). The change in sign across conditioning sets, rather than the absolute size of the final increment, indicates that FE's contribution is conditional.

Spectral overlap and the cable boundary. As previewed in Section 3.1, the global scalar w_ℓ (the per-layer FreqEnhance weight from Eq. 3) cannot distinguish periodic texture from defect signatures when their spatial frequency bands overlap. The cable class (Sample AUROC -3.1) is the only category where this overlap is severe. Let $\Phi_{\text{normal}}(f)$ and $\Phi_{\text{defect}}(f)$ denote the expected radial power spectra of normal and defect regions:

$$\Omega = \int_0^\infty \Phi_{\text{normal}}(f)\, \Phi_{\text{defect}}(f)\, df \Big/ \Big(\big(\int \Phi_{\text{normal}}^2\big)^{1/2} \big(\int \Phi_{\text{defect}}^2\big)^{1/2} \Big) \tag{10}$$

We compute Ω from the MVTec-AD test images. For most MVTec categories, $\Omega < 0.3$: texture and defect power spectra occupy distinct frequency bands. For cable, $\Omega = 0.71$, more than double the next-highest class (transistor, $\Omega = 0.34$). The cable's periodic wire spacing produces a dominant spectral peak near $f \approx 0.15$ cyc/px, which coincides exactly with the typical spatial scale of strand breaks and insulation cuts. A single global w_ℓ therefore cannot attenuate one without suppressing the other. Figure 5 visualizes this diagnosis.

The cable degradation is consistent with a limitation of global frequency modulation: a single scalar cannot selectively suppress normal texture when its spectrum overlaps the defect signal. Because the current data contain only one extreme-overlap category, Ω is best interpreted as an exploratory diagnostic rather than a validated deployment threshold. The observation motivates a spatially adaptive mask in which w_ℓ varies with local texture periodicity, which we leave as future work.

Manifold distortion and per-region precision. We next test whether the observed SC–FE coupling is consistent with how differential frequency scaling interacts with cosine projection.

Consider two adjacent patches with features $\mathbf{f}_i, \mathbf{f}_j$ in a normal image. Their displacement $\delta = \mathbf{f}_i - \mathbf{f}_j$ decomposes into low- and high-frequency components relative to the Gaussian cutoff $f_c \approx \sigma^{-1}$ (where the filter attenuates by $e^{-2\pi^2}$): $\delta = \delta_{\text{L}} + \delta_{\text{H}}$. FE (Eq. 3) scales these components unequally:

$$\tilde{\delta} = \delta_{\text{L}} + (1 + w_\ell)\, \delta_{\text{H}}, \qquad w_\ell < 0 \;\Rightarrow\; |1 + w_\ell| < 1 \tag{11}$$

Two consequences follow. First, the magnitude shrinks ($\|\tilde{\delta}\| < \|\delta\|$), which is benign. Second, differential scaling *rotates* the displacement vector: δ_{L} and δ_{H} point in different directions in $\mathbb{R}^D$, and FE attenuates the latter while preserving the former. The result is $\angle(\tilde{\delta}, \delta) > 0$.

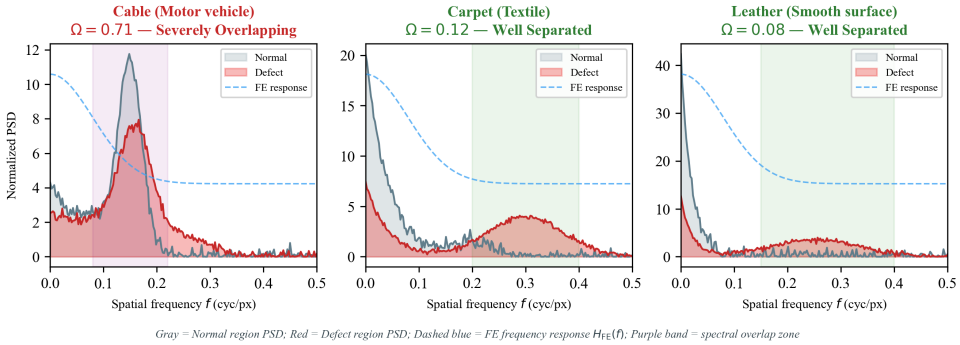


Figure 5: **Spectral overlap analysis.** Normalized radial power spectral density (PSD) of normal and defect regions for cable (left), carpet (center), and leather (right). Dashed blue: FE frequency response $H_{\text{FE}}(f)$. Purple band: spectral overlap zone. Cable exhibits severe overlap ($\Omega = 0.71$) around the dominant periodic peak at $f \approx 0.15$ cyc/px, where FE suppresses both normal texture and defect signatures indiscriminately. Carpet and leather have well-separated spectra ($\Omega < 0.15$), allowing FE to attenuate high-frequency noise without damaging defect-relevant signals.

This rotation is dangerous because cosine similarity projection amplifies it. The Lipschitz constant $\gamma(\mathbf{f}) = \|\nabla_{\mathbf{f}} \cos(\mathbf{f}, \mathbf{t}_a)\| = \sin\theta \,/\, \|\mathbf{f}\|$, where $\theta = \angle(\mathbf{f}, \mathbf{t}_a)$, measures how sensitive the scalar score is to feature perturbations at $\mathbf{f}$. The post-cosine distance between neighbors satisfies:

$$|\cos(\tilde{\mathbf{f}}_i, \mathbf{t}_a) - \cos(\tilde{\mathbf{f}}_j, \mathbf{t}_a)| \leq \gamma_{\max} \cdot \|\tilde{\delta}\|, \qquad \gamma_{\max} = \max_{\alpha \in [0,1]} \gamma(\tilde{\mathbf{f}}_j + \alpha\tilde{\delta}) \tag{12}$$

The directional effect enters through γ: $\gamma(\mathbf{f})$ varies with $\angle(\mathbf{f}, \mathbf{t}_a)$, so rotating $\tilde{\delta}$ can increase $\gamma_{\max}$ even when $\|\tilde{\delta}\|$ decreases. FE therefore admits two competing effects on $\tilde{d}_{ij}$ — magnitude attenuation (beneficial) and orientation-dependent γ-amplification (harmful). When the latter dominates, the post-cosine distance *increases*, producing isolated false positives on homogeneous surfaces.

Direct measurements follow this predicted chain (Table 7b). FE reduces normalized Dirichlet energy to 0.86 but raises $\gamma_{\max}$ to 1.38; the resulting bound and adjacent-score variation increase to 1.28 and 1.20, respectively, while normal-region FPR rises from 5.8% to 8.6%. Adding SC lowers the bound to 0.84, score variation to 0.83, and FPR to 5.1%. These measurements support the proposed SC–FE coupling, although the bound does not by itself establish the role of MS or prove the full three-way sign reversal.

SC (Eq. 5) resolves this geometry directly. The Dirichlet energy $\sum_{(i,j)} \|\delta_{ij}\|^2$ penalizes large displacements regardless of direction. More importantly, on normal images the data manifold exhibits low-dimensional intrinsic structure: smooth regions dominate, and displacement vectors naturally align with directions of slow γ variation (the tangent space of the manifold). Minimizing $\mathcal{L}_{\text{spatial}}$ reinforces this alignment by suppressing large-$\|\delta\|$ pairs, which are precisely the pairs where differential frequency scaling would cause the largest rotation. SC constrains the magnitude of local displacements and encourages preservation of neighborhood relationships. Geometrically, FE applies anisotropic scaling before the nonlinear cosine map, while SC penalizes large local displacements (Figure 6). This offers one explanation for the observed coupling without asserting that the bound uniquely determines the three-way interaction. The per-region results are consistent with this account: PRO improves from 87.5 to 88.5 on MVTec-AD (+1.0) and from 91.0 to 92.1 on VisA (+1.1) when SC is active. Together, the bound and direct measurements support a mechanism in which graph regularization limits the local score variation introduced by frequency modulation.

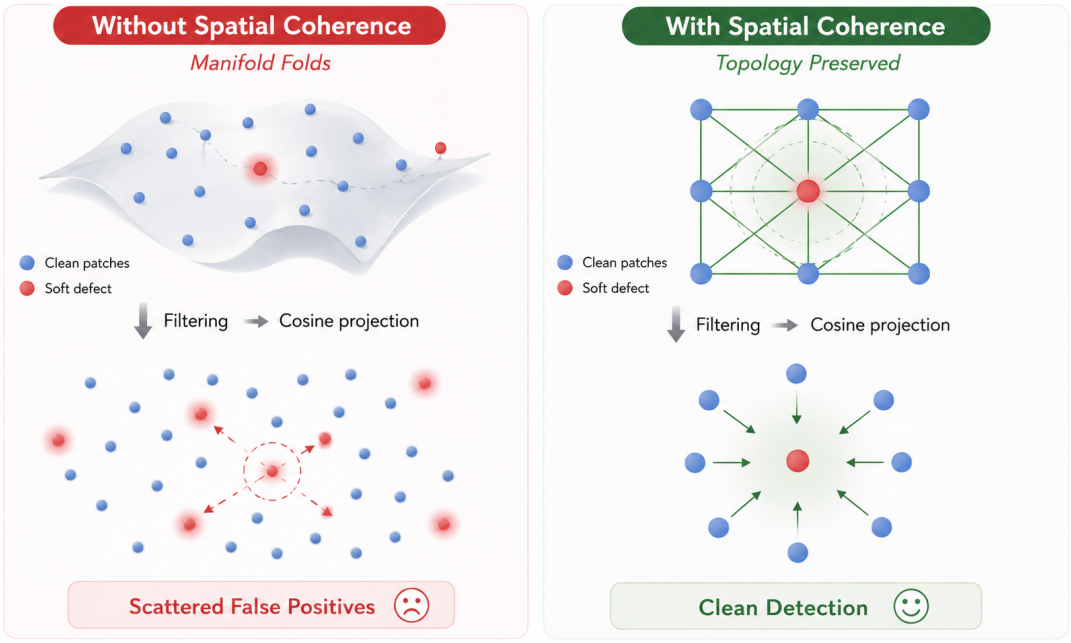


Figure 6: Geometric mechanism of the SC-FE coupling. SC preserves local topology under frequency modulation and cosine projection.

Ruling out optimization artifacts. Freezing baseline tokens and training only FE scalars yields 87.8 Pixel AUROC, statistically identical to joint training (87.7). Three seeds give 88.2 ± 0.5, confirming that the FE-only degradation is not caused by token-feature interference.

Table 8: Controlled FE-only tests.

Configuration	Pixel AUROC	Sample AUROC
Baseline (VisualAD)	90.8	92.2
FE-only	87.7	89.9
FE-only (frozen tokens)	87.8	89.8
FE-only (3 seeds, mean±std)	88.2±0.5	90.0±0.1

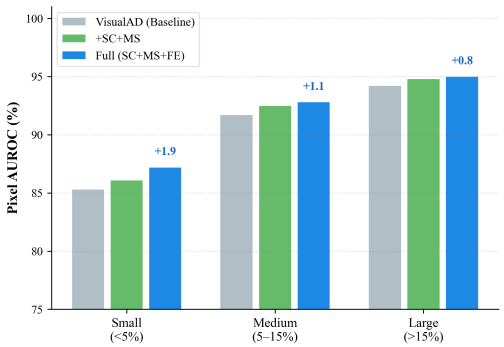


Figure 7: Pixel AUROC by defect size. FE provides its largest gain for small defects, where high-frequency detail is most diagnostic.