DELTAAD: VISUAL DELTA MODELING FOR ZERO-SHOT ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot anomaly detection (ZSAD) aims to identify anomalies in unseen domains without access to target-specific training data. To alleviate cross-domain visual discrepancies, existing ZSAD methods typically leverage pretrained vision or vision-language models to characterize visual distributions or learn text prompts from auxiliary data. However, their cross-domain generalization remains limited under substantial visual distribution shifts. To address this challenge, we propose **DeltaAD**, a novel framework that detects anomalies by amplifying anomaly-related differences while suppressing visual variations shared by normal and anomalous responses through *differential similarity*. Our train-free variant, **DeltaAD-TF**, freezes the CLIP image encoder and constructs separate memory banks of normal and anomalous region-pooled features from auxiliary data. During inference, it contrasts retrieval evidence from the two memory banks under shared visual contexts, highlighting relative differences between normal and anomalous responses while suppressing their shared variations. Building on this principle, our training-based variant, **DeltaAD-LE**, learns a residual relation that characterizes anomalies relative to retrieved normal visual explanations. Extensive experiments on 15 industrial and medical benchmarks demonstrate the strong zero-shot generalization of DeltaAD for both image-level anomaly detection and pixel-level anomaly localization under train-free and training-based settings.

1 INTRODUCTION

Anomaly detection (AD) aims to identify anomalous samples or regions that deviate from normal visual patterns, with important applications in industrial inspection and medical imaging (Czimmermann et al., 2020; Bergmann et al., 2019; Gutman et al., 2016). Its practical deployment remains challenging because anomalies are often scarce, highly diverse, and difficult to characterize exhaustively during training (Chandola et al., 2009; Pang et al., 2021; Mou et al., 2023; Xie et al., 2023). This challenge motivates zero-shot anomaly detection (ZSAD), which aims to identify anomalies in unseen domains without access to target-specific training data (Gao et al., 2026; Cao et al., 2024; Zhou et al., 2024).

With the development of large-scale vision–language models, CLIP (Radford et al., 2021) has provided transferable representations that are widely exploited for ZSAD. Recent methods improve anomaly detection through prompt learning, feature adaptation, or visual retrieval, and have demonstrated strong generalization to unseen categories (Jeong et al., 2023; Zhu et al., 2025; Zhou et al., 2024; Cao et al., 2024; Hou et al., 2026; Xu et al., 2026). VisualAD (Hou et al., 2026) further shows that competitive ZSAD can be achieved using only pretrained visual encoders, while MRAD (Xu et al., 2026) demonstrates the effectiveness of exploiting auxiliary visual features through memory retrieval. Despite these advances, their cross-domain generalization remains limited under substantial visual distribution shifts.

Although existing ZSAD approaches exploit pretrained visual knowledge in different ways, normal and anomalous responses may be derived from semantic prompts, memory banks, or visual representations. Under cross-domain transfer, these responses can contain substantial visual variations shared by the two responses, making their relative comparison less reliable (Chen et al., 2023; Zhou et al., 2024). This motivates constructing normal and anomalous evidence under shared visual contexts before contrasting their responses, thereby emphasizing anomaly-related differences while

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.

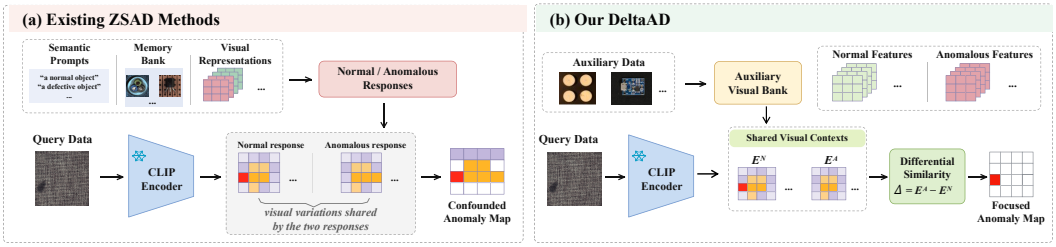


Figure 1: Illustration of the motivation behind DeltaAD. Under cross-domain transfer, normal and anomalous responses may contain substantial visual variations shared by the two responses, making their relative comparison less reliable. DeltaAD constructs normal and anomalous evidence under shared visual contexts and contrasts them through differential similarity.

reducing shared variation. A related differential principle has been explored in sequence modeling, where response differences suppress shared noise while preserving discriminative signals (Ye et al., 2025).

Building on this differential perspective, we propose **DeltaAD**, a visual framework for transferable anomaly detection through *differential similarity*, as illustrated in Fig. 1. Rather than directly relying on normal and anomalous responses that may contain shared visual variations, DeltaAD constructs comparable normal and anomalous evidence under shared visual contexts before differential comparison. Specifically, the train-free model, DeltaAD-TF, realizes differential similarity by contrasting normal and anomalous retrieval evidence under shared visual contexts. Based on DeltaAD-TF, we propose DeltaAD-LE, which further extends this principle to a learnable residual relation that characterizes anomalies relative to retrieved normal visual explanations.

In summary, our contributions are as follows:

- We propose **DeltaAD**, a zero-shot anomaly detection framework that constructs normal and anomalous retrieval evidence under shared visual contexts and contrasts their responses, thereby emphasizing anomaly-related differences while suppressing visual variations shared across the two responses under cross-domain transfer.
- We develop two variants of DeltaAD. **DeltaAD-TF** realizes differential similarity in a train-free manner by contrasting normal and anomalous retrieval evidence, while **DeltaAD-LE** extends this principle to a learnable residual relation that characterizes anomalies relative to retrieved normal visual explanations.
- Extensive experiments on diverse industrial and medical benchmarks demonstrate strong zero-shot generalization for both image-level anomaly detection and pixel-level anomaly localization. Comprehensive analyses further validate the effectiveness of the proposed differential formulation.

2 Related Work

2.1 Conventional Anomaly Detection

Conventional anomaly detection typically models normality from target-domain normal samples. Embedding-based methods such as PaDiM (Defard et al., 2021), PatchCore (Roth et al., 2022), and CS-Flow (Rudolph et al., 2022) estimate normal feature distributions or construct normal feature memories, while synthetic-anomaly approaches such as CutPaste (Li et al., 2021) learn anomaly-sensitive representations from artificial irregularities. DRAEM (Zavrtanik et al., 2021) further combines reconstruction and discriminative learning for anomaly localization, whereas reconstruction-based methods detect anomalies through reconstruction errors. Recent distillation and feature-reconstruction methods, including RD4AD (Deng & Li, 2022), UniAD (You et al., 2022), and EfficientAD (Batzner et al., 2024), model normal representations in feature space. Despite strong performance, these methods generally require target-category normal samples, limiting direct generalization to unseen categories and domains.

2.2 Zero-Shot Anomaly Detection

Zero-shot anomaly detection exploits transferable knowledge from large-scale pretrained models without target-domain training samples. Early training-free methods such as WinCLIP (Jeong et al., 2023) use handcrafted normal and anomalous prompts. Later approaches improve cross-domain anomaly modeling through learnable prompts and lightweight adaptation: AnomalyCLIP (Zhou et al., 2024) learns object-agnostic anomaly prompts, AdaCLIP (Cao et al., 2024) and VCP-CLIP (Qu et al., 2024) introduce adaptive or visually conditioned prompts, while CLIP-AD (Chen et al., 2024), AA-CLIP (Ma et al., 2025), AdaptCLIP (Gao et al., 2026), and Bayes-PFL (Qu et al., 2025) further improve localization and transferability through feature modeling, anomaly-aware anchors, adapters, or probabilistic prompts. Along another direction, VisualAD (Hou et al., 2026) learns anomaly cues purely in visual space, whereas MRAD (Xu et al., 2026) retrieves directly from auxiliary feature–label memories. Different from these approaches, DeltaAD realizes differential similarity by constructing normal and anomalous retrieval evidence under shared visual contexts before contrasting their responses.

3 DeltaAD: Differential Similarity Modeling

3.1 Motivation and Overview

3.1.1 Revisiting Anomaly Evidence

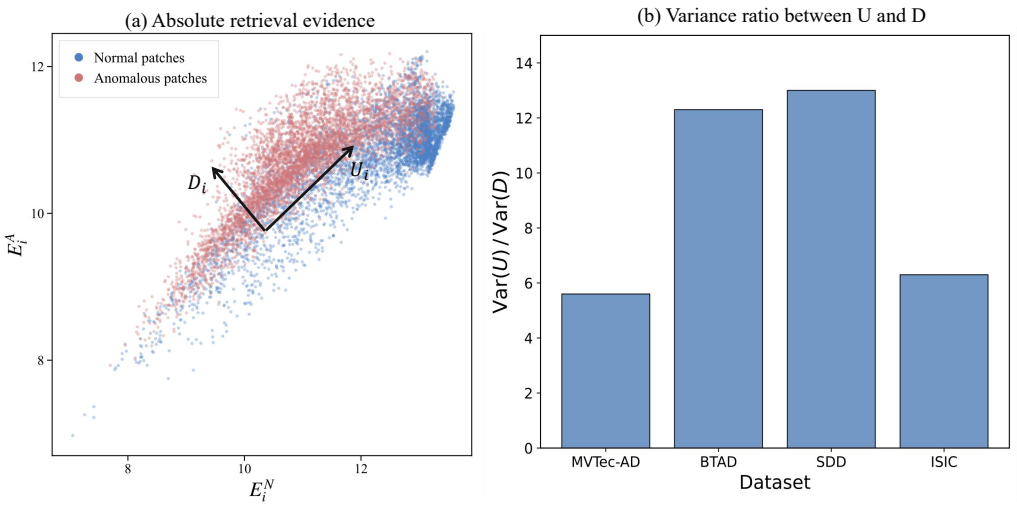


Figure 2: Analysis of normal and anomalous retrieval evidence under shared visual contexts. (a) Normal and anomalous retrieval evidence are visualized in the (E_i^N, E_i^A) space, where $U_i = (E_i^A + E_i^N)/2$ represents the common component and $D_i = (E_i^A - E_i^N)/2$ represents the relative component. (b) The variance ratio between U and D across unseen datasets shows consistently larger variation along the common component.

Existing ZSAD methods commonly estimate anomalies through correspondence between target features and auxiliary visual references. Under cross-dataset transfer, normal and anomalous responses may contain substantial visual variations shared by the two responses, making their relative comparison less reliable. For a target patch i, we retrieve normal and anomalous visual features from the Auxiliary Visual Bank constructed using the labeled auxiliary dataset. Their similarities are aggregated under shared visual contexts to obtain the normal evidence E_i^N and anomalous evidence E_i^A, respectively. A larger E_i^s indicates stronger compatibility with auxiliary visual features of state $s \in \{N, A\}$.

We further decompose the normal and anomalous evidence into a common component $U_i = (E_i^A + E_i^N)/2$ and a relative component $D_i = (E_i^A - E_i^N)/2$. Fig. 2(a) illustrates this decomposition

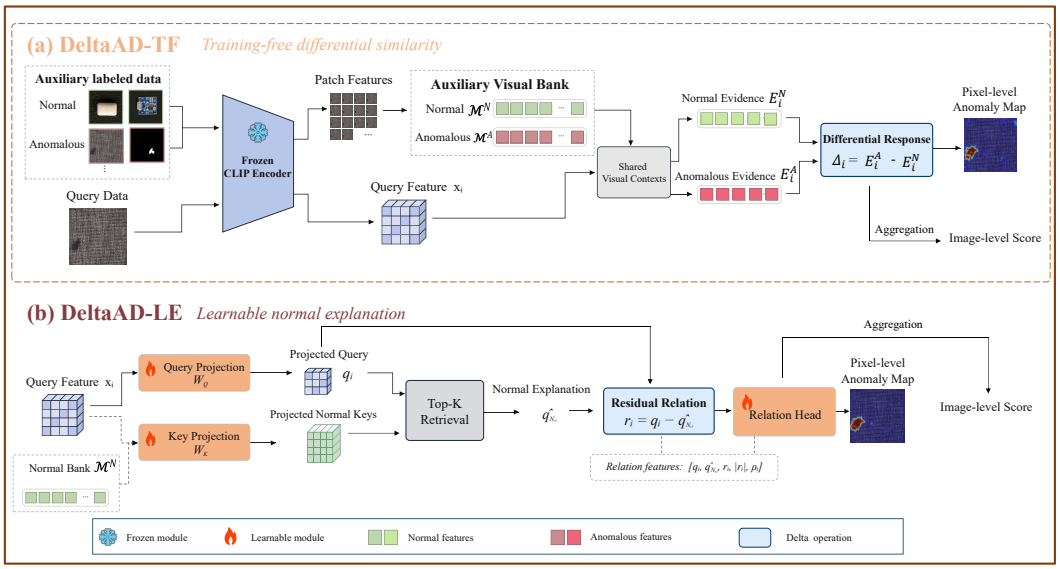


Figure 3: Overall architecture of the proposed DeltaAD framework. **(a) DeltaAD-TF**: query features from the frozen visual encoder retrieve normal and anomalous entries from the Auxiliary Visual Bank under shared visual contexts, and estimate anomaly evidence through differential similarity. **(b) DeltaAD-LE**: learnable query and key projections retrieve a query-conditioned normal visual explanation from normal bank entries, whose residual relation to the query is modeled by a lightweight relation head. Both variants produce pixel-level anomaly maps and image-level anomaly scores.

on MVTec-AD under the VisA-source protocol. Although normal and anomalous patches remain strongly overlapped along either absolute axis, they become more separable along D_i. Fig. 2(b) provides a quantitative view of this phenomenon: $\mathrm{Var}(U)$ is consistently larger than $\mathrm{Var}(D)$ across the evaluated unseen datasets, revealing substantial common variation in the normal and anomalous evidence. Accordingly, we define the differential response as $\Delta_i = E_i^A - E_i^N = 2D_i$, which retains the relative preference between the two states while suppressing their common component. Additional evidence analysis across the remaining target datasets is provided in Appendix A.2.

3.1.2 Approach Overview

As illustrated in Fig. 3, DeltaAD instantiates differential similarity in two complementary forms within the same pretrained visual space. DeltaAD-TF provides a train-free formulation based on differential normal–anomalous retrieval evidence without learnable optimization, whereas DeltaAD-LE introduces lightweight learnable components to characterize the residual relation to a retrieved normal visual explanation.

3.2 DeltaAD-TF: Train-Free Differential Similarity

3.2.1 Auxiliary Visual Bank and Context Routing

Given an input image x, the frozen CLIP visual encoder (Radford et al., 2021) extracts a spatial feature grid $\mathbf{F} \in \mathbb{R}^{P \times d}$ with P patch locations, where P denotes the number of patch locations and d denotes the visual feature dimension. Under our 518×518 input resolution, $P = 1369$ and $d = 1024$. We apply multi-scale average pooling to obtain $\tilde{\mathbf{x}}_i \in \mathbb{R}^d$.

Using the labeled auxiliary dataset, we construct an *Auxiliary Visual Bank* from patch-level visual features. The annotation mask is aligned with the CLIP patch grid to obtain normal and anomalous regions Ω^N and Ω^A. For each non-empty region $s \in \{N, A\}$, the corresponding bank entry is

obtained by masked average pooling followed by normalization:

$$\mathbf{m}^s = \mathrm{Norm}\left(\frac{1}{|\Omega^s|}\sum_{i\in\Omega^s}\tilde{\mathbf{x}}_i\right) \in \mathbb{R}^d, \qquad s \in \{N, A\}. \tag{1}$$

A normal image contributes one normal entry, while an anomalous image may contribute both normal and anomalous entries. All entries form $\mathcal{M} = \mathcal{M}^N \cup \mathcal{M}^A$, where $\mathcal{M}^N$ and $\mathcal{M}^A$ denote the normal and anomalous subsets. Let N_N and N_A denote the numbers of normal and anomalous bank entries, respectively, such that $\mathcal{M}^N \in \mathbb{R}^{N_N\times d}$ and $\mathcal{M}^A \in \mathbb{R}^{N_A\times d}$.

For retrieval, we normalize the query patch feature as $\mathbf{x}_i = \mathrm{Norm}(\tilde{\mathbf{x}}_i)$. To retrieve visually compatible entries, we partition the bank into C visual contexts using K-means (Lloyd, 1982). For each query patch $\mathbf{x}_i$, we select the top-L contexts $\mathcal{C}_i$ according to centroid similarity and normalize the routing weights w_{ic} over the selected contexts. The routed contexts and weights are shared across normal and anomalous evidence aggregation. We denote by $\mathcal{M}_c^s$ the bank entries of state $s \in \{N, A\}$ assigned to context c.

3.2.2 Differential Similarity and Anomaly Detection

For each state $s \in \{N, A\}$, we aggregate retrieval evidence over the same routed contexts:

$$E_i^s = \log \sum_{c\in\mathcal{C}_i} w_{ic} \exp\left(\log \frac{1}{|\mathcal{M}_c^s|}\sum_{\mathbf{k}\in\mathcal{M}_c^s}\exp\left(\frac{\mathbf{x}_i^\top \mathbf{k}}{\tau_p}\right)\right), \qquad s \in \{N, A\}, \tag{2}$$

where τ_p denotes the temperature for patch-level evidence aggregation.

The differential response Δ_i is converted into patch-level anomaly probabilities by applying a sigmoid mapping. The patch-level probabilities are reshaped, upsampled, and smoothed to obtain the pixel-level anomaly map M_{TF}.

For image-level detection, we aggregate retrieval evidence over the same two states before subtraction to preserve their state-specific responses. Specifically, we apply temperature-scaled aggregation:

$$U_x^s(T) = T\left[\log \sum_{i=1}^{P} \exp(E_i^s/T) - \log P\right], \qquad s \in \{N, A\}, \tag{3}$$

where T controls the aggregation sharpness and is fixed to 0.15. The image-level anomaly score is computed as:

$$S_x^{TF} = U_x^A(T) - U_x^N(T). \tag{4}$$

3.3 DeltaAD-LE: Learnable Normal Explanation

3.3.1 Normal Explanation Retrieval

While DeltaAD-TF estimates anomaly evidence by contrasting normal and anomalous entries from the Auxiliary Visual Bank, it requires anomalous visual entries during inference.

However, anomalous appearances are highly diverse and difficult to exhaustively cover across unseen categories. DeltaAD-LE therefore constructs a normal visual explanation for each query patch and identifies anomalies through its residual relation to this explanation (Wang et al., 2025).

DeltaAD-LE reuses the normal subset $\mathcal{M}^N$ of the Auxiliary Visual Bank defined in Sec. 3.2. These bank entries remain fixed, while two learnable projections establish a query–normal retrieval space. Given a query patch feature $\mathbf{x}_i$ and a normal bank entry $\mathbf{m}_j^N$, we obtain

$$\mathbf{q}_i = \mathrm{Norm}(W_Q\mathbf{x}_i), \qquad \mathbf{k}_j^N = \mathrm{Norm}(W_K\mathbf{m}_j^N), \tag{5}$$

where $W_Q, W_K \in \mathbb{R}^{d_r\times d}$ denote independently learned linear projections, and $\mathbf{q}_i, \mathbf{k}_j^N \in \mathbb{R}^{d_r}$, with d_r denoting the retrieval feature dimension.

We compute cosine similarities $s_{ij} = \mathbf{q}_i^\top \mathbf{k}_j^N$ and retrieve the top-K compatible normal bank entries, denoted by $\mathcal{T}_i$. The normal explanation is obtained by softmax-weighted aggregation:

$$\hat{\mathbf{q}}_{N,i} = \sum_{j \in \mathcal{T}_i} \frac{\exp(s_{ij}/\tau_N)}{\sum_{\ell \in \mathcal{T}_i} \exp(s_{i\ell}/\tau_N)} \mathbf{k}_j^N. \tag{6}$$

where τ_N denotes the temperature for normal explanation retrieval.

The resulting $\hat{\mathbf{q}}_{N,i} \in \mathbb{R}^{d_r}$ represents the normal visual explanation of query patch i.

3.3.2 Residual Relation Learning

Given the normal explanation, we define the residual relation between the query and its normal explanation as

$$\mathbf{r}_i = \mathbf{q}_i - \hat{\mathbf{q}}_{N,i}. \tag{7}$$

We further collect retrieval statistics $\boldsymbol{\rho}_i \in \mathbb{R}^4$, consisting of the maximum retrieval similarity, softmax-weighted similarity, residual norm, and retrieval entropy. The anomaly logit is predicted by a lightweight relation head:

$$z_i = H\left([\mathbf{q}_i, \hat{\mathbf{q}}_{N,i}, \mathbf{r}_i, |\mathbf{r}_i|, \boldsymbol{\rho}_i]\right), \tag{8}$$

where $|\mathbf{r}_i|$ denotes the element-wise absolute residual, yielding a 3076-dimensional relation representation. The resulting logits are converted into patch-level anomaly probabilities for localization.

DeltaAD-LE is trained on labeled auxiliary data, where all optimization is performed only on auxiliary samples and no target-domain data are involved. Its objective combines patch-level anomaly supervision with a residual separation constraint:

$$\mathcal{L}_{\mathrm{LE}} = \mathcal{L}_{\mathrm{WBCE}} + \mathcal{L}_{\mathrm{Dice}} + \lambda_{\mathrm{sep}} \mathcal{L}_{\mathrm{sep}}. \tag{9}$$

where

$$\mathcal{L}_{\mathrm{sep}} = \mathbb{E}_{y_i=0} \|\mathbf{r}_i\|_2^2 + \lambda_m \mathbb{E}_{y_i=1} \left[\max(0, m - \|\mathbf{r}_i\|_2)\right]^2. \tag{10}$$

Here, $\mathcal{L}_{\mathrm{WBCE}}$ and $\mathcal{L}_{\mathrm{Dice}}$ supervise patch-level anomaly prediction, while λ_{sep} weights the residual separation term. λ_m weights the anomalous margin term, $y_i \in \{0, 1\}$ denotes the auxiliary patch label, and m is the residual margin for anomalous patches.

For pixel-level localization, the patch probabilities are reshaped and upsampled to obtain the anomaly map M_{LE}. For image-level detection, we aggregate the raw patch-level logits before sigmoid conversion. Let $\mathcal{I}_{1\%}$ denote the indices of the top 1% logits among $\{z_i\}_{i=1}^P$. The image-level anomaly score is computed as

$$S_x^{\mathrm{LE}} = \frac{1}{|\mathcal{I}_{1\%}|} \sum_{i \in \mathcal{I}_{1\%}} z_i. \tag{11}$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate DeltaAD on 15 benchmarks spanning industrial and medical domains. The industrial datasets include MVTec-AD (Bergmann et al., 2019), VisA (Zou et al., 2022), BTAD (Mishra et al., 2021), DTD-Synthetic (Aota et al., 2023), SDD (Tabernik et al., 2020), and KSDD2 (Božič et al., 2021). The medical datasets include HeadCT (Salehi et al., 2021), Br35H (Hamada, 2020), and HIS (Bejnordi et al., 2017) for image-level anomaly detection, and ISIC (Codella et al., 2018), Endo (Hicks et al., 2021), CVC-ColonDB (Tajbakhsh et al., 2015), CVC-ClinicDB (Bernal et al., 2015), Kvasir (Jha et al., 2020), and CVC-300 (Vázquez et al., 2017) for pixel-level anomaly localization. Following the cross-dataset zero-shot setting, VisA serves as the auxiliary dataset for all target datasets except VisA, for which MVTec-AD is used instead. DeltaAD-TF constructs the Auxiliary Visual Bank, while DeltaAD-LE optimizes its learnable components exclusively on the corresponding auxiliary dataset. More dataset statistics and evaluation details are provided in Appendix B.

Table 1: ZSAD results on industrial and medical datasets. Top: pixel-level results (P-AUROC% / PRO%). Bottom: image-level results (I-AUROC% / I-AP%). Higher is better. **Bold** denotes the best result within each method group (*train-free* or *training-based*) for each dataset/metric.

(a) Pixel-level ZSAD (P-AUROC% / PRO%).

Method →	TRAIN-FREE METHODS			TRAINING-BASED METHODS			
Dataset ↓	WinCLIP (CVPR'23)	MRAD-TF (ICLR'26)	**DeltaAD-TF (Ours)**	AnomalyCLIP (ICLR'24)	MRAD-CLIP (ICLR'26)	VisualAD-CLIP (CVPR'26)	**DeltaAD-LE (Ours)**
Industrial							
MVTec-AD	(85.1, 64.6)	(86.7, 63.5)	(**88.0**, **73.2**)	(91.1, 81.4)	(93.0, 86.8)	(90.8, **87.5**)	(**93.1**, 87.2)
VisA	(79.6, 56.8)	(91.0, 71.0)	(**92.1**, **76.5**)	(95.5, 87.0)	(**95.9**, 88.0)	(95.8, **91.0**)	(95.6, 87.9)
BTAD	(71.4, 32.8)	(**80.5**, **36.3**)	(78.8, 35.5)	(93.3, 69.3)	(95.4, 72.8)	(91.1, 80.4)	(**97.4**, **81.3**)
DTD-Synthetic	(78.4, 51.0)	(90.7, 73.8)	(**93.7**, **82.2**)	(97.6, 88.3)	(98.1, 89.8)	(98.1, **94.8**)	(**98.3**, 94.5)
SDD	(55.9, 14.7)	(85.0, 63.0)	(**93.8**, **68.7**)	(90.1, 62.9)	(93.0, 72.0)	(94.6, 68.6)	(**97.1**, **73.4**)
KSDD2	(75.4, 69.2)	(94.9, 87.1)	(**96.9**, **88.8**)	(97.9, 94.9)	(98.9, 95.6)	(98.5, **98.5**)	(**99.3**, 97.1)
Medical							
ISIC	(83.5, 55.1)	(84.4, 72.1)	(**86.7**, **78.2**)	(88.4, 78.1)	(91.3, 83.4)	(84.0, 73.2)	(**93.6**, **87.6**)
Endo	(68.2, 28.3)	(82.7, 60.0)	(**84.9**, **71.2**)	(86.3, 67.3)	(88.3, 71.6)	(84.9, 66.3)	(**89.9**, **75.2**)
CVC-ColonDB	(64.8, 28.4)	(77.6, 59.6)	(**79.2**, **68.0**)	(81.9, 71.4)	(84.7, 73.9)	(76.9, 72.4)	(**86.8**, **77.3**)
CVC-ClinicDB	(70.7, 32.5)	(81.6, 59.0)	(**83.8**, **70.6**)	(85.9, 69.6)	(87.3, 73.9)	(85.2, 70.6)	(**88.4**, **76.5**)
Kvasir	(69.8, 31.0)	(77.1, 52.3)	(**79.4**, **52.5**)	(81.9, 45.7)	(84.3, 52.7)	(80.3, 50.8)	(**87.0**, **54.5**)
CVC-300	(50.4, 12.2)	(91.4, 84.8)	(**93.0**, **89.7**)	(95.8, 56.4)	(96.9, 94.8)	(93.9, 90.2)	(**97.6**, **95.6**)
Average	(71.1, 39.7)	(85.3, 65.2)	(**87.5**, **71.3**)	(90.5, 72.7)	(92.3, 79.6)	(89.5, 78.7)	(**93.7**, **82.3**)

(b) Image-level ZSAD (I-AUROC% / I-AP%).

Method →	TRAIN-FREE METHODS			TRAINING-BASED METHODS			
Dataset ↓	WinCLIP (CVPR'23)	MRAD-TF (ICLR'26)	**DeltaAD-TF (Ours)**	AnomalyCLIP (ICLR'24)	MRAD-CLIP (ICLR'26)	VisualAD-CLIP (CVPR'26)	**DeltaAD-LE (Ours)**
Industrial							
MVTec-AD	(**91.8**, **95.1**)	(79.0, 90.0)	(88.5, 94.3)	(91.5, 96.2)	(94.0, 97.4)	(92.2, 96.7)	(**94.3**, **97.8**)
VisA	(78.1, 77.5)	(75.0, 79.1)	(**81.0**, **83.1**)	(82.1, 85.4)	(**85.7**, **88.3**)	(84.7, 87.6)	(85.5, 88.2)
BTAD	(83.3, 84.1)	(**86.3**, **90.3**)	(81.0, 85.0)	(89.1, 91.1)	(92.8, 94.2)	(94.9, 97.0)	(**95.3**, **97.1**)
DTD-Synthetic	(63.8, 83.5)	(79.1, 89.6)	(**93.5**, **97.3**)	(94.5, 97.7)	(96.0, 98.4)	(**97.5**, **99.1**)	(96.5, 98.7)
SDD	(54.8, 37.7)	(80.0, 66.3)	(**86.7**, **68.4**)	(82.7, 73.8)	(83.9, 76.2)	(83.4, 31.7)	(**95.1**, **81.8**)
KSDD2	(77.5, 43.9)	(90.4, 82.6)	(**90.6**, **83.4**)	(90.8, 76.5)	(95.1, 88.9)	(98.0, **98.3**)	(**98.4**, 94.6)
Medical							
HeadCT	(83.7, 81.6)	(69.3, 73.6)	(**89.9**, **94.4**)	(95.3, 95.2)	(97.1, 97.6)	(89.3, 84.8)	(**98.7**, **99.2**)
Br35H	(80.5, 82.2)	(91.0, 92.1)	(**91.9**, **92.5**)	(97.3, 96.1)	(97.9, 97.6)	(94.4, 91.6)	(**98.4**, **97.9**)
HIS	(57.6, 59.1)	(63.0, 60.1)	(**65.0**, **61.4**)	(55.2, 56.1)	(62.9, 56.2)	(60.1, 55.9)	(**64.9**, **58.0**)
Average	(74.6, 71.6)	(79.2, 80.4)	(**85.3**, **84.4**)	(86.5, 85.3)	(89.5, 88.3)	(88.3, 82.5)	(**91.9**, **90.4**)

Evaluation Metrics. For image-level anomaly detection, we report the Area Under the Receiver Operating Characteristic Curve (AUROC) and Average Precision (AP). For pixel-level anomaly localization, we report pixel-level AUROC and Per-Region Overlap (PRO).

Implementation Details. We adopt CLIP ViT-L/14@336px (Radford et al., 2021) as the frozen visual backbone and resize all images to 518×518. Features from the last transformer layer (L24) construct the Auxiliary Visual Bank. DeltaAD-TF uses 128 visual contexts with the top 16 retained per query patch, while DeltaAD-LE retrieves the top 64 normal bank entries from $\mathcal{M}^N$. Its learnable components are optimized with Adam (Kingma & Ba, 2014) for one epoch using a learning rate of 4×10^{-4} and batch size 8. Results are averaged over target categories, and all experiments are conducted on a single NVIDIA RTX 3090 (24GB). Further implementation details are provided in Appendix A.1.

4.2 PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS

We compare DeltaAD with representative ZSAD methods under train-free and training-based settings. Train-free baselines include WinCLIP (Jeong et al., 2023) and MRAD-TF (Xu et al., 2026), while training-based baselines include AnomalyCLIP (Zhou et al., 2024), MRAD-CLIP (Xu et al., 2026), and VisualAD-CLIP (Hou et al., 2026). For fair comparison, all training-based methods use

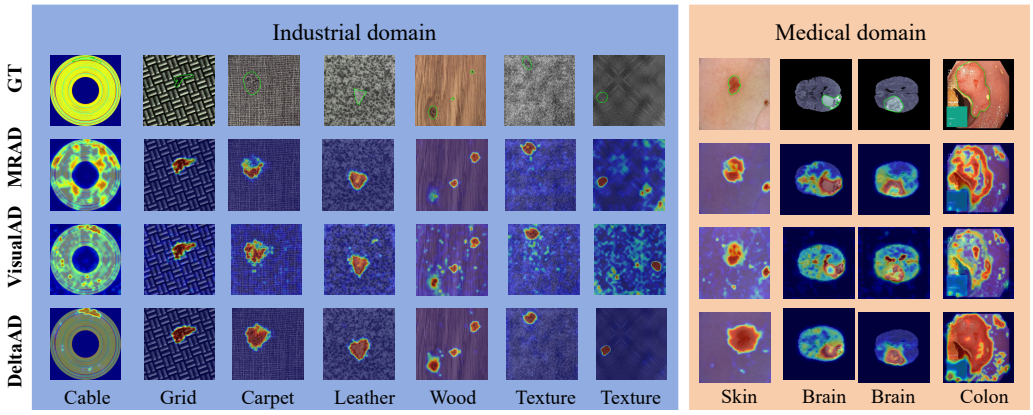


Figure 4: Qualitative comparison of anomaly localization, where DeltaAD produces more focused anomaly responses than MRAD and VisualAD across industrial and medical samples.

CLIP ViT-L/14@336px as the visual encoder. Table 1 reports image- and pixel-level results on industrial and medical benchmarks. Detailed category-level results are provided in Appendix C.

DeltaAD-TF achieves the best average performance among train-free methods at both image and pixel levels, while DeltaAD-LE achieves the best average results among training-based methods across all four metrics. These results demonstrate strong zero-shot generalization across industrial and medical benchmarks under both settings. Efficiency, trainable parameters, and model size are compared in Appendix D.

Fig. 4 presents qualitative localization results on representative industrial and medical samples. Compared with MRAD and VisualAD, DeltaAD generally produces more focused anomaly responses around anomalous regions with fewer irrelevant background activations, consistent with the quantitative results. Additional representation visualizations are provided in Appendix F.

4.3 Ablation Study

Effect of Differential Response. We investigate the effect of representing anomaly evidence through the differential response rather than an absolute retrieval response. Keeping the Auxiliary Visual Bank and context-routing protocol fixed, we compare anomaly-only evidence E^A, normal rejection $-E_N$, the common component $(E^A + E^N)/2$, and the proposed differential response $E^A - E^N$. As shown in Table 2, the differential response achieves the best performance on seven of the eight applicable metrics across the three target datasets, consistently outperforming either anomaly-only or normal-only evidence on the main detection and localization metrics. In contrast, the common component exhibits inconsistent performance across pixel- and image-level evaluation. Overall, the differential response provides more consistent and effective anomaly evidence than either individual normal or anomalous evidence or their common component. Additional ablations, robustness analyses, and hyperparameter sensitivity studies are provided in Appendix E.

Table 2: Ablation of anomaly-evidence formulation in DeltaAD-TF. Results are reported in terms of pixel-level (AUROC, PRO) and image-level (AUROC, AP) metrics. The best performance in each column is shown in bold. "–" indicates an undefined PRO score due to a degenerate anomaly map.

Evidence	MVTec-AD		ISIC	HeadCT
	Pixel	Image	Pixel	Image
E^A	(82.9, **74.9**)	(86.8, 93.7)	(78.0, 43.0)	(86.3, 91.3)
$-E^N$	(86.1, 52.2)	(43.5, 72.1)	(84.5, 65.4)	(62.5, 73.1)
$(E^A + E^N)/2$	(17.6, –)	(86.6, 93.4)	(17.6, –)	(81.8, 89.7)
$E^A - E^N$	(**88.0**, 73.2)	**(88.5, 94.3)**	**(86.7, 78.2)**	**(89.9, 94.4)**

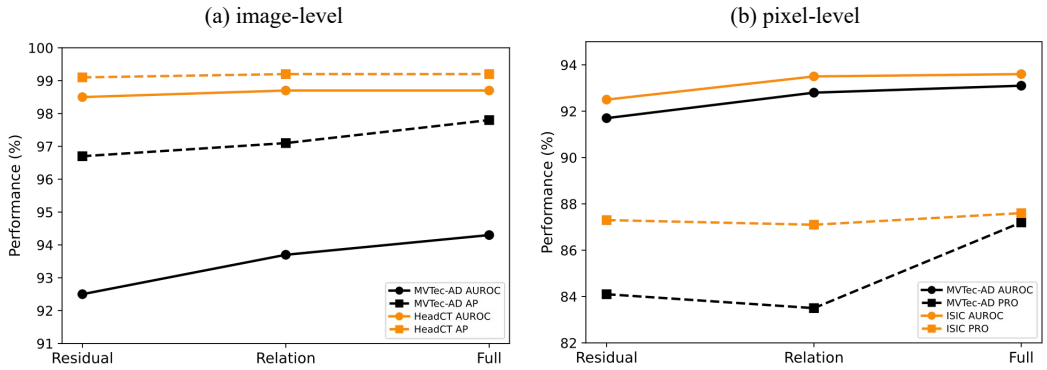


Figure 5: Ablation of normal explanation relation modeling in DeltaAD-LE. *Residual* denotes anomaly scoring based on the residual magnitude $\|\mathbf{r}\|_2$; *Relation* uses the query and its retrieved normal explanation $[\mathbf{q}, \hat{\mathbf{q}}_N]$; and *Full* denotes the complete relation representation adopted in DeltaAD-LE. Panel (a) reports image-level AUROC and AP on MVTec-AD and HeadCT, while panel (b) reports pixel-level AUROC and PRO on MVTec-AD and ISIC.

Effect of Local Context Routing. We further investigate the role of local context routing before computing the differential response. We compare global retrieval, independent routing for normal and anomalous evidence, and the proposed shared weighted routing. As shown in Table 3, removing local routing leads to a consistent degradation across detection and localization metrics, indicating that restricting retrieval to visually compatible contexts is important before contrasting normal and anomalous evidence. Independent routing can improve individual metrics, but its performance varies more noticeably across tasks. In contrast, shared weighted routing yields the strongest overall performance, supporting the use of shared visual contexts for differential response estimation.

Table 3: Ablation of local context routing in DeltaAD-TF. Results are reported as Pixel-level (AUROC, PRO) and Image-level (AUROC, AP). The best performance in each column is shown in bold.

Routing	**MVTec-AD**		**ISIC**	**HeadCT**
	Pixel	Image	Pixel	Image
Global delta	(87.1, 64.9)	(72.7, 87.0)	(85.1, 73.7)	(73.9, 87.0)
Independent routing	(87.6, 67.2)	(66.9, 83.0)	**(87.5, 78.3)**	(65.3, 79.8)
Shared weighted routing	**(88.0, 73.2)**	**(88.5, 94.3)**	(86.7, 78.2)	**(89.9, 94.4)**

Effect of Normal Explanation Relation Modeling. We further investigate how the retrieved normal explanation is exploited in DeltaAD-LE. We compare three progressively enriched residual relations, ranging from residual-magnitude scoring to explicit query–normal relation modeling and the complete relation representation used in DeltaAD-LE. As shown in Fig. 5, incorporating the retrieved normal explanation generally improves anomaly discrimination over using residual magnitude alone. The complete residual relation further achieves the strongest overall performance across image-level detection and pixel-level localization, indicating that anomaly evidence is better characterized by the structured relation to a normal visual explanation than by residual magnitude alone.

5 Conclusion

We explore differential similarity for zero-shot anomaly detection. Building on this idea, we propose DeltaAD, a visual framework that characterizes anomalies through relative visual differences with two variants. DeltaAD-TF constructs normal and anomalous evidence under shared visual contexts, while DeltaAD-LE models the residual relation between a query and its retrieved normal visual explanation. Extensive experiments across 15 industrial and medical benchmarks demonstrate strong zero-shot generalization for both image-level detection and pixel-level localization using only the pretrained CLIP visual branch. These results highlight differential similarity as an effective transferable anomaly signal under cross-domain transfer.

Reproducibility Statement

To facilitate reproducibility, we provide detailed implementation settings, evaluation protocols, and additional analyses in the Appendix. Appendix A.1 reports the implementation and hyperparameter settings of DeltaAD, with additional mechanism analysis in Appendix A.2. Appendix B describes the datasets and evaluation protocols, while Appendix C provides category-level results. Additional SOTA comparisons and ablations, including robustness, random-seed, and sensitivity analyses, are provided in Appendices D and E, respectively. Appendix F presents additional visualizations. Generative AI usage is described in the corresponding statement.

References

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

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

Babak Ehteshami Bejnordi, Mitko Veta, and Paul Johannes van Diest. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318:2199–2210, 2017.

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. IEEE, 2019.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

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

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European conference on computer vision*, pp. 55–72. Springer, 2024.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, 2009. doi: 10.1145/1541880.1541882.

Xuhai Chen, Yue Han, and Jiangning Zhang. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023.

Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, and Yong Liu. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. In *International Joint Conference on Artificial Intelligence*, pp. 17–33. Springer, 2024.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168–172. IEEE, 2018.

Tamás Czimmermann, Gastone Ciuti, Mario Milazzo, Marcello Chiurazzi, Stefano Roccella, Calogero Maria Oddo, and Paolo Dario. Visual-based defect detection and classification approaches for industrial applications—a survey. *Sensors*, 20(5), 2020. ISSN 1424-8220. URL https://www.mdpi.com/1424-8220/20/5/1459.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, pp. 475–489. Springer, 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. IEEE, 2022.

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

David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016.

Ahmed Hamada. Br35h: Brain tumor detection 2020. Kaggle Dataset, 2020. Available online: `https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection`.

Steven A Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo L Hammer, and Michael A Riegler. The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In *International Conference on Pattern Recognition*, pp. 263–274. Springer, 2021.

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

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

Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-seg: A segmented polyp dataset. In Yong Man Ro, Wen-Huang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve (eds.), *MultiMedia Modeling*, pp. 451–462, Cham, 2020. Springer International Publishing.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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. IEEE, 2021.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4744–4754. IEEE, 2025.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pp. 01–06. IEEE, 2021.

Shancong Mou, Xiaoyi Gu, Meng Cao, Haoping Bai, Ping Huang, Jiulong Shan, and Jianjun Shi. Rgi: robust gan-inversion for mask-free image inpainting and unsupervised pixel-wise anomaly detection. In *The Eleventh International Conference on Learning Representations*, 2023.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

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

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

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 *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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.

Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1829–1838. IEEE, 2022.

Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14902–14912, 2021.

Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020.

Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017(1):4037190, 2017.

Yuexin Wang, Xiaolei Wang, Yizheng Gong, and Jimin Xiao. Normal-abnormal guided generalist anomaly detection. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025.

Guoyang Xie, Jinbao Wang, Jiaqi Liu, Yaochu Jin, and Feng Zheng. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

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

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 144–164, 2025.

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

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dræm–a discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8310–8319. IEEE, 2021.

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

Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22241–22251. IEEE, 2025.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pp. 392–408. Springer, 2022.

A Appendix

A.1 Implementation Details

We implement DeltaAD using the frozen visual encoder of CLIP ViT-L/14@336px. All images are resized to 518×518. We use patch features from the last transformer layer (L24) with multi-scale aggregation at scales $\{1, 3, 5\}$, where the visual feature dimension is $d = 1024$. We apply Gaussian smoothing with $\sigma = 4$ to the pixel-level anomaly maps. Following the cross-dataset setting, VisA is used as the auxiliary dataset for all target datasets except VisA, for which MVTec-AD is used. No target-domain data are used for training or hyperparameter selection.

DeltaAD-TF. We partition the Auxiliary Visual Bank into $C = 128$ visual contexts and retain the top $L = 16$ contexts for each query patch. The context-routing temperature is set to $\tau_c = 0.07$, the patch-level retrieval temperature to $\tau_p = 0.07$, and the image-level aggregation temperature to $T = 0.15$.

DeltaAD-LE. The query and key projections map features to a retrieval space of dimension $d_r = 768$. For each query patch, we retrieve the top $K = 64$ normal bank entries with temperature $\tau_N = 0.07$. The retrieval-statistics vector $\boldsymbol{\rho}_i \in \mathbb{R}^4$ contains the maximum retrieval similarity, softmax-weighted similarity, residual norm, and retrieval entropy. The relation head takes the 3076-dimensional relation representation as input and consists of LayerNorm followed by linear layers of dimensions $3076 \to 768 \to 384 \to 1$, with GELU activations between the linear layers. We set $\lambda_{\text{sep}} = 0.02$, $\lambda_m = 0.1$, and $m = 1.3$. The learnable projections and relation head are optimized on the auxiliary dataset using Adam with a learning rate of 4×10^{-4}, batch size 8, and one training epoch, while the CLIP visual encoder and Auxiliary Visual Bank remain frozen.

A.2 Mechanism Behind Cross-Domain Retrieval Generalization

DeltaAD models anomaly evidence through differential similarity between normal and anomalous responses. By emphasizing their relative difference, the model reduces the influence of visual variations shared by the two responses.

For DeltaAD-TF, normal and anomalous retrieval evidence is evaluated under shared visual contexts, and their difference emphasizes the relative preference between the two states while suppressing the common component shared by the two responses.

For DeltaAD-LE, a query-conditioned normal visual explanation is retrieved from the normal bank, and anomaly evidence is derived from its residual relation to the query. Both formulations therefore characterize anomalies through relative visual differences under cross-domain transfer.

To complement the representative analysis in Fig. 2(b), Table 4 reports the variance decomposition of the normal and anomalous evidence over the remaining target datasets. Across these datasets, $\mathrm{Var}(U)$ is consistently larger than $\mathrm{Var}(D)$, supporting the presence of substantial common variation in the normal and anomalous evidence.

Table 4: Variance decomposition of normal and anomalous retrieval evidence over normal target patches. VisA is evaluated with MVTec-AD as the auxiliary source; all other targets use VisA as the auxiliary source.

Dataset	Patches	$\mathrm{Var}(E_A)$	$\mathrm{Var}(E_N)$	$\mathrm{Var}(U)$	$\mathrm{Var}(D)$	$\mathrm{Var}(U)/\mathrm{Var}(D)$
MVTec-AD	2,285,969	0.45	1.45	0.81	0.14	5.62
VisA	2,943,339	0.24	1.10	0.48	0.19	2.49
BTAD	993,024	0.67	1.79	1.13	0.09	12.31
DTD-Synthetic	1,767,159	0.54	0.94	0.70	0.04	16.46
SDD	545,921	0.34	0.77	0.52	0.04	13.02
KSDD2	1,369,879	0.09	0.33	0.18	0.03	5.72
ISIC	373,126	0.32	1.01	0.57	0.09	6.28
Endo	234,156	0.76	2.03	1.27	0.13	10.05
CVC-ColonDB	481,304	0.56	1.65	0.99	0.11	8.91
CVC-ClinicDB	759,965	0.79	2.21	1.36	0.14	10.00
Kvasir	1,154,311	0.79	1.94	1.25	0.11	11.41
CVC-300	79,396	0.47	1.38	0.83	0.09	9.00

Table 5: Statistics of the evaluation splits used in our experiments.

Dataset	**Categories**	**Normal**	**Anomalous**	**Total**
Industrial				
MVTec-AD	15	467	1,258	1,725
VisA	12	962	1,200	2,162
BTAD	3	451	290	741
DTD-Synthetic	12	357	947	1,304
SDD	1	181	74	255
KSDD2	1	894	110	1,004
Medical				
ISIC	1	0	379	379
Endo	1	0	200	200
CVC-ColonDB	1	0	380	380
CVC-ClinicDB	1	0	612	612
Kvasir	1	0	1,000	1,000
CVC-300	1	0	60	60
HeadCT	1	100	100	200
Br35H	1	1,500	1,500	3,000
HIS	1	1,003	994	1,997

B Dataset Details

Our evaluation covers 15 benchmarks spanning industrial inspection and medical imaging. The industrial benchmarks contain multiple object and texture categories with both normal and anomalous test images, whereas the medical benchmarks include both image-level classification datasets and lesion localization datasets. Table 5 summarizes the evaluation split used for each benchmark, including the number of categories and image-level normal/anomalous samples.

The six industrial benchmarks cover diverse defect types across objects, textures, and surface inspection scenarios. MVTec-AD and VisA provide multi-category evaluations, while BTAD, DTD-Synthetic, SDD, and KSDD2 further extend the evaluation to different industrial distributions and defect characteristics.

The medical benchmarks cover both image-level anomaly detection and pixel-level anomaly localization. HeadCT, Br35H, and HIS are evaluated at the image level, whereas ISIC, Endo, CVC-ColonDB, CVC-ClinicDB, Kvasir, and CVC-300 are used for lesion localization. As shown in Table 5, all images in these pixel-only medical benchmarks contain annotated anomalous regions; therefore, they do not define a separate image-level normal/anomalous evaluation split and are evaluated only with pixel-level metrics.

Table 6: Category-level results (%) on MVTec-AD.

Category	DeltaAD-TF				DeltaAD-LE			
	P-AUROC	PRO	I-AUROC	I-AP	P-AUROC	PRO	I-AUROC	I-AP
carpet	99.0	96.0	99.8	99.9	99.5	93.0	100.0	100.0
bottle	85.1	72.9	94.0	98.1	92.9	85.7	93.3	98.3
hazelnut	93.1	84.4	93.2	96.6	97.1	92.6	95.5	98.1
leather	99.2	99.1	100.0	100.0	99.3	97.9	100.0	100.0
cable	70.7	34.7	77.4	86.1	80.3	74.6	93.5	97.3
capsule	92.2	66.9	95.3	99.0	97.1	93.8	97.1	99.6
grid	95.0	83.8	98.5	99.5	98.1	83.5	99.2	99.9
pill	83.4	71.6	85.3	97.1	94.0	92.0	85.6	97.9
transistor	63.3	53.5	65.1	64.4	73.3	59.4	88.0	85.8
metal_nut	74.5	60.8	64.6	91.6	74.7	76.1	78.8	95.9
screw	93.4	77.1	76.0	89.9	99.1	92.3	86.5	95.5
toothbrush	90.1	53.4	91.7	96.9	97.3	93.6	98.1	99.4
zipper	88.2	59.9	87.8	96.9	98.0	87.7	99.8	99.9
tile	95.5	90.9	98.3	99.1	97.5	90.5	99.2	99.7
wood	96.8	93.5	99.9	100.0	98.3	95.3	99.9	99.7
Average	**88.0**	**73.2**	**88.5**	**94.3**	**93.1**	**87.2**	**94.3**	**97.8**

Table 7: Category-level results (%) on VisA.

Category	DeltaAD-TF				DeltaAD-LE			
	P-AUROC	PRO	I-AUROC	I-AP	P-AUROC	PRO	I-AUROC	I-AP
candle	96.5	93.1	86.4	89.7	98.8	93.5	86.0	89.9
capsules	88.4	62.7	83.9	90.5	96.6	90.6	85.4	92.3
cashew	91.6	77.2	82.7	91.7	94.1	91.0	86.6	93.6
chewinggum	98.4	82.7	97.5	98.8	99.5	92.9	97.6	99.0
fryum	93.2	76.4	90.3	95.7	94.0	88.6	92.3	96.5
macaroni1	93.5	76.9	77.6	77.3	98.8	93.5	80.0	81.7
macaroni2	97.0	85.0	67.4	65.6	96.5	81.7	76.2	77.8
pcb1	84.1	64.2	78.4	77.9	93.8	86.0	83.2	84.4
pcb2	87.7	65.4	64.0	61.8	91.9	79.0	75.3	75.9
pcb3	86.1	67.8	63.1	62.7	88.5	72.5	71.6	72.1
pcb4	93.0	82.2	92.4	91.8	97.1	91.8	98.7	98.3
pipe_fryum	95.7	84.4	88.3	93.7	97.6	93.7	93.1	96.9
Average	**92.1**	**76.5**	**81.0**	**83.1**	**95.6**	**87.9**	**85.5**	**88.2**

C Category-Level Results

We report category-level ZSAD performance on the four multi-category benchmarks, MVTec-AD, VisA, BTAD, and DTD-Synthetic. Detailed results of DeltaAD-TF and DeltaAD-LE are summarized in Tables 6–9. The remaining benchmarks contain a single category, so their category-level results are identical to the dataset-level results reported in the main paper.

D Additional Comparison with State-of-the-Art Methods

We further compare DeltaAD with recent ZSAD methods in terms of average detection performance and computational efficiency. The reported I-AUROC and P-AUROC are the average results over the image-level and pixel-level benchmarks in Table 1, respectively. For WinCLIP, MRAD-TF, AnomalyCLIP, and MRAD-CLIP, the efficiency statistics follow the RTX 3090 results reported by MRAD. For VisualAD-CLIP, DeltaAD-TF, and DeltaAD-LE, the efficiency statistics are obtained from our full-MVTec native evaluations using VisA as the auxiliary dataset, with batch size one and 1,725 MVTec test images. Inference time is measured after model and cache loading and before final metric aggregation, including preprocessing, forward or retrieval computation, anomaly-map post-processing, and image-level score calculation. Params denotes trainable parameters and excludes the frozen CLIP backbone.

Table 8: Category-level results (%) on BTAD.

Category	DeltaAD-TF				DeltaAD-LE			
	P-AUROC	PRO	I-AUROC	I-AP	P-AUROC	PRO	I-AUROC	I-AP
01	72.8	21.9	88.8	95.0	97.4	82.4	98.6	99.5
02	84.4	46.7	72.4	95.5	96.3	65.0	89.7	97.1
03	79.2	37.8	81.7	64.6	98.5	96.5	97.6	94.7
Average	**78.8**	**35.5**	**81.0**	**85.0**	**97.4**	**81.3**	**95.3**	**97.1**

Table 9: Category-level results (%) on DTD-Synthetic.

Category	DeltaAD-TF				DeltaAD-LE			
	P-AUROC	PRO	I-AUROC	I-AP	P-AUROC	PRO	I-AUROC	I-AP
Woven_001	99.6	96.0	98.1	99.4	99.8	96.7	100.0	100.0
Woven_127	82.0	62.6	90.9	93.5	94.2	94.9	93.9	95.0
Woven_104	86.7	71.8	96.3	99.1	97.7	97.3	99.3	100.0
Stratified_154	99.1	91.6	97.0	99.4	99.7	93.8	98.9	100.0
Blotchy_099	99.4	95.8	96.5	99.0	99.2	98.8	95.1	99.9
Woven_068	96.3	87.5	86.9	91.1	98.8	94.8	90.6	94.6
Woven_125	99.2	96.2	97.0	99.3	99.5	95.3	100.0	100.0
Marbled_078	98.6	95.8	98.2	99.6	99.3	98.6	99.8	100.0
Perforated_037	82.2	55.0	87.1	96.8	95.2	83.5	93.9	99.6
Mesh_114	84.5	57.1	78.4	91.7	96.6	86.4	87.2	95.3
Fibrous_183	97.7	92.4	98.9	99.7	99.5	99.4	100.0	100.0
Matted_069	98.9	83.8	97.0	99.3	99.5	94.5	99.3	100.0
Average	**93.7**	**82.2**	**93.5**	**97.3**	**98.3**	**94.5**	**96.5**	**98.7**

Table 10: Comparison with recent SOTA methods in average detection performance and computational efficiency. I-AUROC and P-AUROC are averaged over the image-level and pixel-level benchmarks in Table 1, respectively. Lower is better for inference time, trainable parameters, and model size. Red and blue denote the best and second-best results for each metric, respectively.

	Accuracy & Inference			Training Cost	
Method	**I-AUROC**	**P-AUROC**	**Inf. (ms/img)**	**Params**	**Model Size (MB)**
WinCLIP	74.6	71.1	840.3	–	–
MRAD-TF	79.2	85.3	198.3	–	–
DeltaAD-TF (Ours)	85.3	87.5	197.8	–	–
AnomalyCLIP	86.5	90.5	177.6	5,555,200	22
MRAD-CLIP	89.5	92.3	203.0	9,491,968	54
VisualAD-CLIP	88.3	89.5	146.1	31,522,836	126.12
DeltaAD-LE (Ours)	91.9	93.7	226.7	6,994,953	27.99

As shown in Table 10, DeltaAD-LE achieves the highest average I-AUROC and P-AUROC while maintaining a relatively lightweight trainable model. DeltaAD-TF introduces little additional inference cost over MRAD-TF while providing substantially stronger average detection and localization performance. Overall, the results show that DeltaAD improves zero-shot anomaly detection performance without requiring a large increase in computational cost.

E Additional Ablation Results

We provide additional ablations on the relation representation of DeltaAD-LE and sensitivity analyses of the principal retrieval hyperparameters in DeltaAD-TF and DeltaAD-LE. Pixel-level results are reported as (P-AUROC, PRO), while image-level results are reported as (I-AUROC, I-AP). All values are in percentage, and the best result for each metric is shown in **bold**.

Table 11: Ablation of relation inputs in DeltaAD-LE. Results are reported as pixel-level (P-AUROC, PRO) and image-level (I-AUROC, I-AP). The best result for each metric is shown in bold.

Relation input	MVTec Pixel	MVTec Image	ISIC Pixel	HeadCT Image
Query-only $[\mathbf{q}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}]$	(91.6, 82.2)	(92.4, 96.1)	(92.3, 86.2)	(97.6, 97.1)
$\|\mathbf{r}\|_2$	(91.7, 84.1)	(92.5, 96.7)	(92.5, 87.3)	(98.5, 99.3)
$[\mathbf{q}, \hat{\mathbf{q}}_N]$	(92.8, 83.5)	(93.7, 97.1)	(93.5, 87.1)	(98.7, **99.4**)
$[\mathbf{q}, \hat{\mathbf{q}}_N, \mathbf{r}]$	(92.8, 83.4)	(93.4, 97.0)	(**93.6**, **87.6**)	(98.8, **99.4**)
$[\mathbf{q}, \hat{\mathbf{q}}_N, \mathbf{r}, \|\mathbf{r}\|]$	(92.8, 83.5)	(93.5, 97.0)	(**93.6**, **87.6**)	(**98.9**, 99.1)
Full $[\mathbf{q}, \hat{\mathbf{q}}_N, \mathbf{r}, \|\mathbf{r}\|, \boldsymbol{\rho}]$	(**93.1**, **87.2**)	(**94.3**, **97.8**)	(**93.6**, **87.6**)	(98.7, 99.2)

Table 12: Effect of the auxiliary dataset on MVTec-AD. Results are reported as P-AUROC / I-AUROC (%).

Auxiliary dataset	DeltaAD-TF	DeltaAD-LE
VisA	88.0 / 88.5	**93.1 / 94.3**
BTAD	85.5 / 72.4	92.0 / 93.2
DTD-Synthetic	**88.6 / 87.0**	91.8 / 92.9

E.1 Relation Representation in DeltaAD-LE

We examine the contribution of different inputs to the relation head in DeltaAD-LE. We include a query-only control that retains only the query feature, while all remaining relation inputs are zero-filled. All learnable variants use the same relation-head capacity and training protocol, with removed inputs zero-filled such that the head architecture remains unchanged. Table 11 summarizes the results.

As shown in Table 11, the query-only control is weaker than the full relation representation across all reported metrics, showing that the additional relation inputs contribute beyond the query feature alone. Incorporating the retrieved normal visual explanation improves over the query-only control on nearly all metrics. The complete representation achieves the best performance on all four MVTec-AD metrics and matches the best pixel-level performance on ISIC. We therefore adopt the full relation representation in DeltaAD-LE.

E.2 Effect of Auxiliary Source Dataset

We further examine the effect of the auxiliary dataset while fixing MVTec-AD as the target benchmark. We compare VisA, BTAD, and DTD-Synthetic as labeled auxiliary datasets. For each source, DeltaAD-TF reconstructs the Auxiliary Visual Bank and visual contexts, while DeltaAD-LE is trained independently on the corresponding auxiliary dataset. All other settings remain unchanged, and no target-domain data are used. As shown in Table 12, auxiliary datasets with broader category coverage, such as VisA and DTD-Synthetic, generally provide more favorable transfer performance, particularly for DeltaAD-TF, suggesting that diverse auxiliary visual patterns can benefit cross-domain anomaly modeling.

E.3 Stability Across Random Seeds

We further evaluate the stability of DeltaAD on MVTec-AD using three random seeds, $\{110, 111, 112\}$. For DeltaAD-TF, the seed controls the K-means initialization for visual-context construction. For DeltaAD-LE, each run is trained independently using the same auxiliary-only training protocol. As shown in Table 13, both variants exhibit small variation across random seeds, indicating stable performance under different visual-context initializations and independent training runs.

Table 13: Stability across random seeds on MVTec-AD. We report the mean P-AUROC and I-AUROC over three runs, together with the maximum deviation from the corresponding mean.

Method	P-AUROC	I-AUROC	Max deviation
DeltaAD-TF	88.2	88.6	≤ 0.3
DeltaAD-LE	93.1	94.2	≤ 0.2

E.4 Robustness to Appearance Perturbations

We further evaluate the robustness of DeltaAD-TF to appearance variations at test time. We apply a fixed set of perturbations to the target test images, including brightness scaling with factors 0.8 and 1.2, contrast scaling with factors 0.8 and 1.2, and Gaussian blur with $\sigma = 1.0$. The model, Auxiliary Visual Bank, and inference settings remain unchanged, and no target-domain training or adaptation is performed.

Table 14 reports performance on the original test set and the average performance over the five perturbed test sets. DeltaAD-TF exhibits only small performance fluctuations under these appearance perturbations and consistently outperforms MRAD-TF on the evaluated datasets. Overall, its performance is largely preserved under mild appearance variations.

Table 14: Robustness to test-time appearance perturbations. We report P-AUROC and I-AUROC (%) on the original test set and the average over five perturbed test sets.

		Original		Perturbed	
Dataset	**Method**	P-AUROC	I-AUROC	P-AUROC	I-AUROC
MVTec-AD	MRAD-TF	86.7	79.0	86.6	78.7
	DeltaAD-TF	**88.0**	**88.5**	**88.4**	**88.3**
ISIC	MRAD-TF	84.4	–	84.3	–
	DeltaAD-TF	**86.7**	–	**86.9**	–

Detailed perturbation results. Table 15 further reports the individual results for each appearance perturbation. All values use the same frozen visual encoder, Auxiliary Visual Bank, context partition, and inference settings as the original evaluation; only the raw target image is perturbed before the unchanged preprocessing transform.

Table 15: Detailed robustness results (%) under each test-time appearance perturbation. P and I denote P-AUROC and I-AUROC, respectively. ISIC provides pixel-level evaluation only.

	MVTec-AD				ISIC	
Perturbation	MRAD-TF		DeltaAD-TF		MRAD-TF	DeltaAD-TF
	P	I	P	I	P	P
Brightness (0.8)	86.8	79.0	**88.5**	**88.4**	84.2	**86.7**
Brightness (1.2)	86.7	77.4	**88.4**	**88.5**	85.3	**87.4**
Contrast (0.8)	86.6	79.2	**88.3**	**88.3**	83.7	**86.2**
Contrast (1.2)	86.7	79.1	**88.3**	**88.7**	84.6	**87.5**
Gaussian blur ($\sigma = 1.0$)	86.5	78.7	**88.3**	**87.8**	83.8	**86.8**

As shown in Table 15, mild appearance perturbations do not uniformly decrease AUROC and can induce small positive or negative fluctuations depending on the perturbation. Nevertheless, DeltaAD-TF remains consistently stronger than MRAD-TF under every evaluated perturbation, indicating that its anomaly detection performance is largely preserved under these appearance variations.

Controlled brightness analysis. We further examine the behavior of the differential response and context routing under a fixed brightness perturbation with a scaling factor of 0.8. All controlled evaluations use the same frozen visual encoder, VisA-derived Auxiliary Visual Bank, context partition with seed 111, and inference hyperparameters as the main experiments. No target-domain adaptation or parameter selection is performed.

Table 16: Context routing under a controlled brightness perturbation (scaling factor 0.8). The same routing variants as in Table 3 are evaluated with all other settings unchanged. Best results are shown in bold.

Dataset / Metric	Global	Independent	Shared weighted
MVTec P-AUROC	87.1	88.1	**88.5**
MVTec PRO	65.3	69.1	**75.4**
MVTec I-AUROC	73.3	63.8	**88.4**
MVTec I-AP	87.0	81.8	**94.0**
ISIC P-AUROC	84.9	**87.4**	86.7
ISIC PRO	73.7	78.0	**78.6**
HeadCT I-AUROC	74.8	66.0	**89.7**
HeadCT I-AP	87.5	81.1	**94.4**

Table 17: Evidence formulation under a controlled brightness perturbation (scaling factor 0.8). The same formulations as in Table 2 are evaluated without modifying the Auxiliary Visual Bank or inference settings. Best results are shown in bold.

Dataset / Metric	E^A	$-E^N$	$(E^A + E^N)/2$	$E^A - E^N$
MVTec P-AUROC	84.1	86.8	17.4	**88.5**
MVTec PRO	**77.1**	52.1	–	75.4
MVTec I-AUROC	88.2	42.4	87.6	**88.4**
MVTec I-AP	**94.0**	71.5	93.8	94.0
ISIC P-AUROC	77.0	85.7	17.3	**86.7**
ISIC PRO	–	65.8	–	**78.6**
HeadCT I-AUROC	88.1	62.7	82.7	**89.7**
HeadCT I-AP	93.4	73.7	90.2	**94.4**

As shown in Table 17, the differential response achieves the best result on six of the eight reported metrics, including both ISIC and both HeadCT metrics. On MVTec-AD, anomalous evidence alone achieves the best PRO and a marginally higher unrounded I-AP, whereas the differential response remains best for P-AUROC and I-AUROC. Overall, the relative advantage of the differential response is largely preserved under the controlled brightness variation, consistent with the main ablation in Table 2.

Table 16 shows that shared weighted routing achieves the best result on all MVTec-AD and HeadCT metrics, as well as ISIC PRO, under the brightness perturbation. Independent routing is slightly better only for ISIC P-AUROC. Overall, shared weighted routing achieves the best result on seven of the eight reported metrics, supporting the use of shared visual contexts before contrasting normal and anomalous evidence.

E.5 Sensitivity of DeltaAD-TF

We next study the sensitivity of DeltaAD-TF to the number of visual contexts C and the number of routed contexts L. One parameter is varied at a time while fixing the other to its default value.

Number of visual contexts. Table 18 reports the results obtained by varying C with $L = 16$.

Table 18: Sensitivity of DeltaAD-TF to the number of visual contexts C with $L = 16$. Pixel and Image denote (P-AUROC, PRO) and (I-AUROC, I-AP), respectively. The best result for each metric is shown in bold.

C	MVTec Pixel	MVTec Image	ISIC Pixel	HeadCT Image
64	(**88.3**, 72.5)	(83.2, 92.0)	(86.9, 77.5)	(74.8, 87.0)
128	(88.0, 73.2)	(**88.5**, **94.3**)	(86.7, 78.2)	(**89.9**, **94.4**)
256	(87.7, **75.1**)	(88.2, **94.3**)	(**87.4**, **79.9**)	(88.8, 93.7)

Table 18 shows that using a very small number of visual contexts can noticeably degrade image-level performance. Once C is sufficiently large, further increasing the number of contexts leads to relatively small and metric-dependent variations. Overall, DeltaAD-TF shows limited sensitivity to C within the evaluated range beyond the smallest setting.

Number of routed contexts. We further vary the routing size L while fixing $C = 128$. The results are reported in Table 19.

Table 19: Sensitivity of DeltaAD-TF to the number of routed contexts L with $C = 128$. Pixel and Image denote (P-AUROC, PRO) and (I-AUROC, I-AP), respectively. The best result for each metric is shown in bold.

L	MVTec Pixel	MVTec Image	ISIC Pixel	HeadCT Image
4	(87.4, **76.6**)	(86.2, 93.1)	(86.5, **79.9**)	(81.1, 89.9)
8	(87.4, 75.6)	(87.6, 94.1)	(86.8, 79.2)	(88.1, 93.6)
16	(**88.0**, 73.2)	(**88.5**, **94.3**)	(86.7, 78.2)	(**89.9**, **94.4**)
32	(**88.0**, 71.3)	(86.6, 93.7)	(**86.9**, 77.2)	(80.5, 91.1)

Table 19 shows that different routing sizes favor different localization and detection metrics. The default setting $L = 16$ provides the strongest overall performance across the evaluated datasets.

E.6 Sensitivity of DeltaAD-LE

Finally, we study the number of retrieved normal bank entries K in DeltaAD-LE. Each configuration is trained independently using the same auxiliary-only training protocol. Table 20 summarizes the results.

Table 20: Sensitivity of DeltaAD-LE to the number of retrieved normal bank entries K. Pixel and Image denote (P-AUROC, PRO) and (I-AUROC, I-AP), respectively. The best result for each metric is shown in bold.

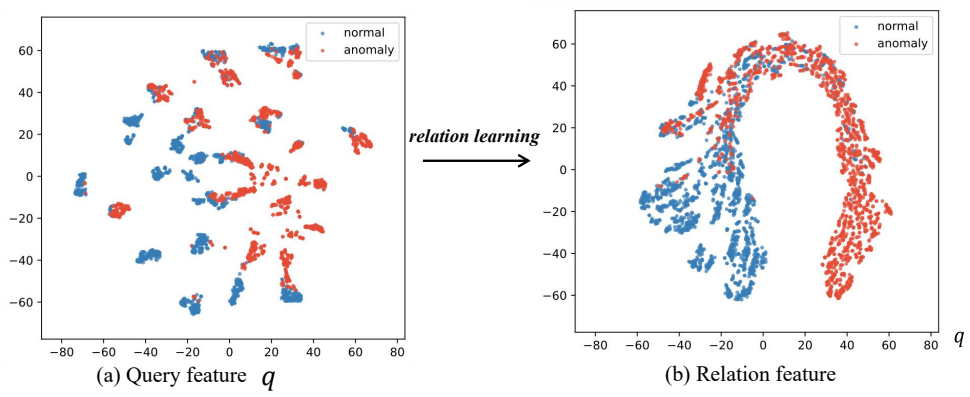
K	MVTec Pixel	MVTec Image	ISIC Pixel	HeadCT Image
16	(92.8, 83.8)	(93.5, 97.0)	(**93.6**, **87.6**)	(**98.8**, 99.4)
32	(92.7, 86.5)	(93.5, 97.0)	(93.5, 87.5)	(**98.8**, **99.5**)
64	(**93.1**, **87.2**)	(**94.3**, **97.8**)	(**93.6**, **87.6**)	(98.7, 99.2)
128	(92.8, 83.7)	(93.6, 97.0)	(**93.6**, 87.5)	(98.7, 99.4)

Table 20 shows that the performance varies only slightly over $K \in \{16, 32, 64, 128\}$, indicating that DeltaAD-LE is not sensitive to the number of retrieved normal bank entries within this range. The default setting used in the main experiments is $K = 64$.

F VISUALIZATION

We further analyze the representation learned by DeltaAD-LE on MVTec-AD using t-SNE (Van der Maaten & Hinton, 2008). Specifically, we visualize the query feature $\mathbf{q}$ and the relation feature produced by the relation head to examine how relation learning transforms the anomaly representation.

As illustrated in Fig. 6, the query feature space exhibits strong category-dependent variation, with normal and anomalous patches remaining partially entangled. After relation learning, the resulting relation features show substantially clearer separation between normal and anomalous patches. This indicates that modeling the relation between the query and its retrieved normal visual explanation produces a more discriminative anomaly representation than the query feature alone.



(a) Query feature q (b) Relation feature

Figure 6: t-SNE visualization of the representation learned by DeltaAD-LE on MVTec-AD. Blue and red points denote normal and anomalous patches, respectively. (a) Query feature $\mathbf{q}$ before relation learning. (b) Relation feature produced by the relation head. Relation learning leads to clearer separation between normal and anomalous patches.

G GENERATIVE AI USE STATEMENT

In preparing this manuscript, we used generative AI tools in a limited capacity for translation, grammar correction, and minor language polishing. These tools were not used for research ideation, hypothesis formulation, experimental design, method development, data analysis, result interpretation, or substantive generation of the scientific content.