UNIMRE: MODALITY-SPECIALIZED RETRIEVAL AND REASONING FOR UNIFIED MULTIMODAL ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multi-class anomaly detection has attracted growing interest as an efficient alternative to the conventional one-class-one-model paradigm, reducing training and maintenance costs for industrial deployment. However, real-world industrial defects span both appearance defects, such as scratches and contamination, and geometric defects, such as protrusions and deformations. Most existing unified methods rely primarily on RGB images and therefore lack explicit geometric cues for detecting structural anomalies. In this paper, we propose **Unified Multimodal Retrieval-guided Experts (UniMRE)**, a unified framework that preserves category-specific normality and complementary appearance–geometry evidence through dual-stream retrieval-guided reasoning. UniMRE constructs class-conditional normal memories in separate RGB and joint RGBD feature spaces, where stream-wise retrieval provides stream-specific normal references and matching evidence for anomaly reasoning. Retrieval-guided expert pathways progressively reason over the stream-specific evidence, while lightweight bidirectional cross-stream conditioning enriches feature reasoning without altering the matching volumes. The resulting predictions are integrated by a learnable class-aware pixel fusion conditioned on local expert responses and nearest-normal retrieval cues. Extensive experiments on MVTec-3D, Eyecandies, and MulSen-AD demonstrate improved image-level anomaly detection over the existing unified multimodal baseline under the full-shot setting, together with strong few-shot performance and competitive results against specialized one-class-one-model approaches. Code is available at `https://anonymous.4open.science/r/unimre-iclr-anon-62EA`.

1 INTRODUCTION

Anomaly detection is a core problem in computer vision and a vital tool for industrial quality inspection (Bergmann et al., 2019). Due to the scarcity and diversity of defective samples, conventional methods typically follow a one-class-one-model paradigm (Roth et al., 2022), training a separate detector for each object category. However, this paradigm incurs substantial training and maintenance costs as the number of categories increases. Consequently, recent research has shifted toward *unified multi-class* anomaly detection (You et al., 2022; Luo et al., 2025), which aims to handle multiple categories within a single model.

Under the unified setting, most existing methods still rely primarily on RGB images (You et al., 2022). While effective at detecting appearance-related defects such as scratches and contamination, they lack explicit geometric cues for capturing structural anomalies such as deformations and protrusions. Depth images and 3D point clouds (Bergmann et al., 2021) provide complementary structural information and have therefore been increasingly explored for multimodal anomaly detection (Wang et al., 2023; Zhu et al., 2025; Li et al., 2025a).

However, as illustrated in Fig. 1(a), most existing multimodal approaches follow the one-model-per-class paradigm (Wang et al., 2023; Costanzino et al., 2024; Xiang et al., 2025), requiring a specialized detector for each category and leaving unified multimodal anomaly detection under-explored (Zhao et al., 2026). Developing a unified multimodal detector is non-trivial, as a single

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.

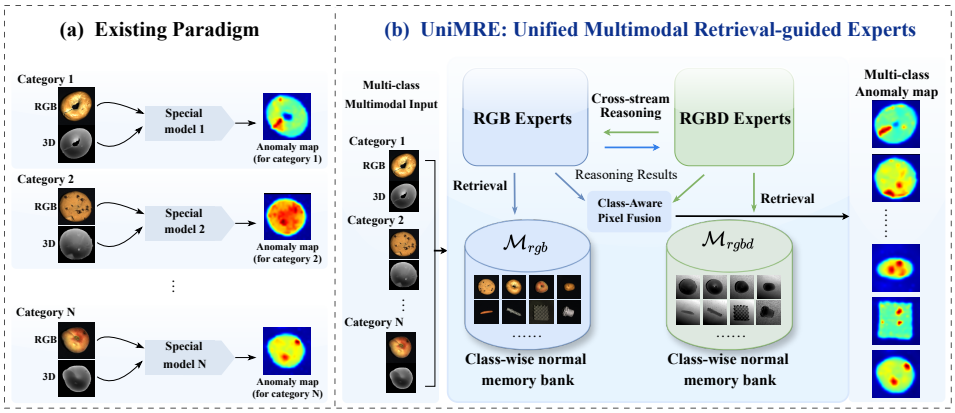


Figure 1: Comparison of multimodal anomaly detection paradigms. (a) Conventional approaches typically require a specialized model for each category. (b) UniMRE handles multiple categories within a single framework through class-conditional retrieval in separate RGB and joint RGBD feature spaces, followed by retrieval-guided expert reasoning and learnable class-aware pixel fusion.

model must capture diverse normal patterns across categories with substantially different appearance and geometric structures, while effectively exploiting complementary multimodal cues for different anomaly types. This requires preserving category-specific normality while flexibly exploiting complementary appearance and geometric information within a unified framework. In particular, directly coupling heterogeneous modalities during normal matching may obscure stream-specific query-to-normal discrepancies, whereas keeping the modalities fully independent limits their complementary use during anomaly reasoning.

To address these challenges, we propose **UniMRE**, a unified multimodal anomaly detection framework based on dual-stream retrieval-guided experts, as illustrated in Fig. 1(b). UniMRE performs class-conditional stream-wise retrieval over separate RGB and joint RGBD normal memories, providing stream-specific normal references and matching evidence for anomaly reasoning. The two retrieval spaces remain independent so that appearance- and geometry-aware matching discrepancies are preserved, while lightweight bidirectional cross-stream conditioning is introduced during expert feature reasoning to exploit complementary information across streams. The resulting RGB and RGBD predictions are then adaptively integrated through learnable class-aware pixel fusion.

Our main contributions are summarized as follows:

- We propose UniMRE, a unified multimodal anomaly detection framework that preserves category-specific normality and complementary appearance–geometry evidence through class-conditional dual-stream retrieval in separate RGB and joint RGBD feature spaces.
- We develop retrieval-guided expert reasoning with lightweight bidirectional cross-stream feature conditioning and learnable class-aware pixel fusion, enabling complementary interaction without contaminating stream-specific matching evidence and adaptive integration of the resulting anomaly predictions.
- Extensive experiments on MVTec-3D, Eyecandies, and MulSen-AD demonstrate consistent improvements in image-level AUROC over the existing unified multimodal baseline, together with strong few-shot performance and competitive results against specialized one-class-one-model approaches.

2 Related Work

Retrieval-Based Industrial Anomaly Detection. Industrial anomaly detection commonly identifies defects by modeling normal feature distributions, retrieving normal references, or learning reconstruction discrepancies. Representative embedding-based methods such as PaDiM (Defard et al., 2021) and PatchCore (Roth et al., 2022) model normal patterns in pretrained feature spaces, while reconstruction- and discriminative-learning approaches such as DRAEM (Zavrtanik et al., 2021)

and EfficientAD (Batzner et al., 2024) learn normality-aware representations for anomaly detection and localization. More recently, CostFilter-AD (Zhang et al., 2025b) constructs matching cost volumes between query and normal features and learns to refine the resulting matching evidence, while RAID (Cai et al., 2026) further combines hierarchical retrieval with guided MoE filtering to reason over retrieved normal templates and matching costs. These retrieval-based methods are predominantly RGB-centric and therefore do not address how complementary appearance and geometric evidence can be preserved across distinct feature spaces within a unified multimodal detector.

Unified Multi-Class Anomaly Detection. To alleviate the overhead of category-specific models, recent studies have explored unified anomaly detection, where a single model (Gu et al., 2025) handles multiple object categories. UniAD (You et al., 2022) establishes a unified reconstruction framework for multi-class anomaly detection, followed by methods that explore discriminative learning (Liu et al., 2023; Fan et al., 2025), feature reconstruction (Guo et al., 2025; Zhang et al., 2025a), diffusion models (He et al., 2024), and vision-language representations (Jeong et al., 2023) to better accommodate diverse normal patterns across categories. While these approaches substantially advance unified anomaly detection, most remain RGB-centric, leaving the preservation and reasoning of complementary geometric evidence under a unified multi-class detector insufficiently explored.

Multimodal Anomaly Detection. Multimodal anomaly detection exploits complementary appearance and geometric information (Mao et al., 2025; Wang et al., 2025; Tao et al., 2025) from RGB images, depth maps, or point clouds to improve the detection of diverse industrial defects. Most existing methods follow the one-model-per-class paradigm. M3DM (Wang et al., 2023) combines feature- and decision-level fusion with modality-specific memory banks, while CFM (Costanzino et al., 2024) learns cross-modal feature mappings between RGB and 3D representations. Subsequent methods explore different multimodal modeling strategies: 3D-ADNAS (Long et al., 2025) searches modality-specific modules and fusion architectures; FIND (Li et al., 2025c) combines intra-modal reverse distillation with cross-modal feature mapping for few-shot anomaly inspection; GPFlow (Li et al., 2026) models normal variations with Gaussian prototypes and integrates intra- and cross-modal reconstruction; and MulSen-TripleAD (Li et al., 2025b) extends multimodal inspection to RGB, point cloud, and infrared data through decision-level fusion. Despite their effectiveness, these methods generally require category-specific models, limiting their scalability to multi-class industrial scenarios.

More recently, UniMMAD (Zhao et al., 2026) extends multimodal anomaly detection to the unified multi-class setting, using MoE-based feature decompression to accommodate heterogeneous modalities and object categories within a single model. Nevertheless, unified multimodal anomaly detection remains relatively underexplored, particularly in preserving category-specific normal patterns and complementary appearance–geometry evidence within a shared detector. UniMRE approaches this problem from a retrieval-based perspective by maintaining separate RGB and joint RGBD memories and matching evidence, while allowing lightweight cross-stream interaction only during expert feature reasoning. The resulting stream predictions are adaptively integrated through learnable class-aware pixel-wise fusion.

3 METHOD

3.1 OVERVIEW

We propose **UniMRE**, a unified multimodal framework for multi-class anomaly detection and localization. The key idea is to preserve category-specific normality and complementary appearance–geometry evidence through stream-specialized retrieval and reasoning. As illustrated in Fig. 2, paired RGB and depth inputs are encoded into an RGB stream and a joint RGBD stream. For each category, separate normal memories are constructed in the two feature spaces, allowing each stream to retrieve its own normal reference and query-to-normal matching evidence.

The retrieved evidence is processed by retrieval-guided experts to produce stream-specific anomaly predictions. The two streams retain separate memories and matching volumes, while lightweight bidirectional cross-stream conditioning is injected only into the expert feature pathway before matching-aware prediction. Finally, a learnable class-aware pixel fusion integrates the RGB and RGBD predictions according to their local prediction and retrieval evidence, yielding the final anomaly map and image-level score.

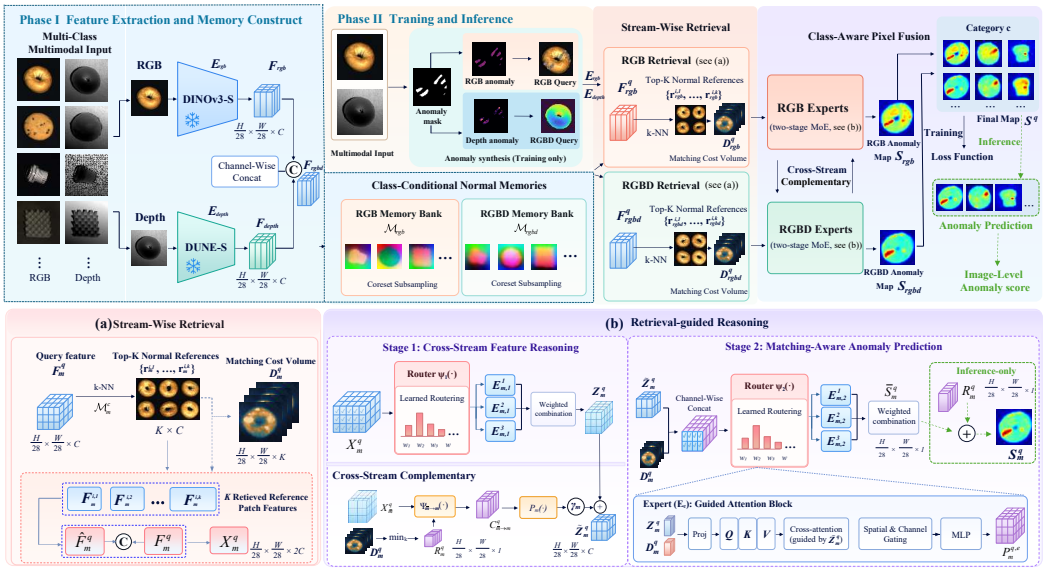


Figure 2: Overview of **UniMRE**. Phase I performs frozen RGB/depth encoding and constructs class-conditional RGB and joint RGBD memories. Phase II conducts stream-wise retrieval and retrieval-guided expert reasoning with lightweight cross-stream conditioning, then integrates the stream predictions by learnable class-aware pixel fusion.

3.2 Class-Conditional Dual-Stream Retrieval

Dual-Stream Representation. Given a paired RGB image I_{rgb} and depth image I_{depth}, frozen encoders E_{rgb} and E_{depth} extract appearance and geometric features, respectively. We retain the RGB representation and construct a joint RGBD representation by channel-wise concatenation:

$$F_{rgb} = E_{rgb}(I_{rgb}), \qquad F_{depth} = E_{depth}(I_{depth}), \qquad F_{rgbd} = [F_{rgb} \,\|\, F_{depth}]. \tag{1}$$

We use $m \in \{rgb, rgbd\}$ to index the two streams.

Class-Conditional Normal Memories. Normal patterns differ substantially across object categories. A single category-agnostic memory may therefore introduce irrelevant normal correspondences. For each category c, we construct separate RGB and RGBD normal memories $\mathcal{M}^c_{rgb}$ and $\mathcal{M}^c_{rgbd}$ from the ℓ_2-normalized features of normal training samples in **Phase I** (frozen encoding and memory construction). The memories are optionally subsampled to reduce redundant normal features while retaining representative patterns.

During training, anomalous RGB–Depth pairs are synthesized from normal samples (Zavrtanik et al., 2021) to provide pixel-level supervision. Appearance perturbations are applied to RGB inputs and geometric perturbations to depth inputs. The normal memories are built exclusively from clean samples and remain fixed throughout training.

Stream-Wise Retrieval and Feature Packing. Once the normal memories are constructed, each input is treated as a query for retrieval. During training, the query is a synthesized anomalous sample with a known category, whereas at inference its category is estimated as described in Sec. 3.4. Given the associated category c, each query patch $f^{q,i}_m$ retrieves its K nearest normal neighbors (Roth et al., 2022) from the corresponding stream-specific memory. The retrieved neighbors and their matching costs are:

$$\mathcal{N}_K(f^{q,i}_m) = \{r^{i,1}_m, \ldots, r^{i,K}_m\}, \qquad d^{i,k}_m = 1 - \mathrm{sim}(f^{q,i}_m, r^{i,k}_m), \tag{2}$$

where $r^{i,k}_m$ denotes the k-th retrieved normal feature for query location i, and $d^{i,k}_m$ is the corresponding cosine distance computed on ℓ_2-normalized features. The patch-wise costs are arranged into the matching cost volume (Zhang et al., 2025b; Cai et al., 2026):

$$D^q_m = [d^{i,k}_m]_{i,k}, \tag{3}$$

where the k-th channel records the discrepancy to the k-th retrieved normal neighbor over the spatial query grid, allowing multiple normal correspondences to be retained at each location.

In parallel, the retrieval index assigns each query patch to its nearest IVF centroid, forming a compact spatially aligned normal reference $\widehat{F}_m^q$ for feature-level reasoning. Thus, each stream supplies the subsequent expert module with $(F_m^q, \widehat{F}_m^q, D_m^q)$. Importantly, retrieval and matching are performed independently in the RGB and RGBD feature spaces, preventing cross-stream features from contaminating the matching cost volumes.

3.3 Retrieval-Guided Reasoning and Fusion

Retrieval distances provide useful anomaly evidence, but they remain sensitive to imperfect normal correspondences and offer only a limited description of query-to-normal deviations. In Phase II, we therefore progressively reason over the retrieved normal context and matching evidence to obtain stream-specific anomaly predictions, which are subsequently integrated by a learnable class-aware pixel fusion module.

Cross-Stream Feature Reasoning. The retrieved normal reference provides local normal context for reasoning over the query representation. For each stream, the query feature and its spatially aligned normal reference are jointly routed to a set of feature experts. The query–reference representation is formulated as:

$$X_m^q = [F_m^q \,\|\, \widehat{F}_m^q], \qquad Z_m^q = \sum_{e \in \mathcal{T}_{m,1}^q} \alpha_m^{q,e} E_{m,1}^e(X_m^q), \tag{4}$$

where $\mathcal{T}_{m,1}^q$ denotes the activated experts and $\alpha_m^{q,e}$ is the corresponding normalized routing weight. The resulting Z_m^q forms a reference-aware feature representation for subsequent cross-stream conditioning and anomaly prediction.

To exchange complementary information while preserving stream-specific retrieval evidence, we derive a compact condition from the complementary stream. Let $\bar{m}$ denote the stream complementary to m. We first extract the nearest-normal retrieval response from its matching volume and use it together with the corresponding query–reference representation to construct the cross-stream condition:

$$R_m^{q,i} = \min_k d_m^{i,k}, \qquad C_{\bar{m}\to m}^q = \Psi_{\bar{m}\to m}\left(X_{\bar{m}}^q, R_{\bar{m}}^q\right), \qquad \widetilde{Z}_m^q = Z_m^q + \gamma_m \Pi_m\left(C_{\bar{m}\to m}^q\right), \tag{5}$$

where $R_m^{q,i}$ denotes the nearest-normal retrieval response at spatial location i, $R_m^q = \{R_m^{q,i}\}_i$ denotes the corresponding spatial response map, $\Psi_{\bar{m}\to m}$ denotes the lightweight cross-stream conditioner, $\Pi_m(\cdot)$ projects the condition to the current feature space, and γ_m is a learnable residual scale. The cross-stream condition is injected only into the feature pathway, while the original matching volume D_m^q remains unchanged. This conditioning is applied symmetrically to both streams, enabling bidirectional feature interaction while preserving stream-specific retrieval statistics.

Matching-Aware Anomaly Prediction. The cross-conditioned representation $\widetilde{Z}_m^q$ enriches the reference-aware features with complementary stream information, while the matching cost volume D_m^q retains the original stream-specific query-to-normal discrepancies. The prediction experts jointly exploit these two forms of evidence to generate dense anomaly responses:

$$P_m^{q,e} = E_{m,2}^e(\widetilde{Z}_m^q, D_m^q), \qquad \bar{S}_m^q = \sum_e \beta_m^{q,e} P_m^{q,e}, \tag{6}$$

where $E_{m,2}^e$ denotes the e-th matching-aware prediction expert, $P_m^{q,e}$ is its dense anomaly response, and $\beta_m^{q,e}$ denotes the corresponding expert aggregation weight. The aggregated response $\bar{S}_m^q$ represents the learned stream-level anomaly prediction produced by retrieval-guided expert reasoning and is directly supervised during training.

At inference, we additionally preserve the direct non-parametric retrieval evidence by adding the nearest-normal response:

$$S_m^q = \bar{S}_m^q + R_m^q, \tag{7}$$

where S_m^q denotes the resulting stream-level anomaly map that combines the learned expert prediction with the direct nearest-normal retrieval evidence. The RGB and RGBD streams consequently produce S_{rgb}^q and S_{rgbd}^q for subsequent fusion.

Learnable Class-Aware Pixel Fusion. The relative reliability of RGB and RGBD evidence varies across both object categories and spatial locations. We therefore replace a fixed fusion coefficient with a learnable pixel-wise gate that jointly considers the learned expert responses and direct retrieval evidence from the two streams. Let $\widetilde{D}_m^q$ denote the matching volume resized to the prediction resolution, with the expert prediction maps resized accordingly, and let $\widetilde{R}_m^q(i) = \min_k \widetilde{D}_m^q(i,k)$ denote its nearest-normal retrieval response. The spatial fusion gate is computed as:

$$G^q = \sigma\Big(\Phi_{\text{gate}}\left([\bar{S}_{rgb}^q \,\|\, \bar{S}_{rgbd}^q \,\|\, \widetilde{R}_{rgb}^q \,\|\, \widetilde{R}_{rgbd}^q]\right) + b_{c_q}\Big), \tag{8}$$

where Φ_{gate} denotes the lightweight convolutional fusion network, b_{c_q} is a learnable category-aware bias, and $\sigma(\cdot)$ is the sigmoid function. The resulting G^q represents the spatially varying contribution of the RGB stream. During training, c_q is the known category, whereas at inference it is replaced by the estimated category $\hat{c}$.

At inference, the learned gate is applied to the stream-level anomaly maps to obtain the final fused prediction:

$$S^q = G^q \odot S_{rgb}^q + (1 - G^q) \odot S_{rgbd}^q, \tag{9}$$

where $\odot$ denotes element-wise multiplication. This formulation preserves the direct retrieval evidence of each stream while allowing spatially adaptive fusion according to local expert responses, retrieval discrepancy, and category-level context.

3.4 Training and Inference

Training. The expert prediction maps $\bar{S}_{rgb}^q$ and $\bar{S}_{rgbd}^q$ are optimized with separate focal losses, together with load-balancing regularization for expert routing. The fusion gate is additionally supervised by a focal loss on the fused expert prediction map before adding the direct retrieval residual. The overall training objective is defined as:

$$\mathcal{L} = \sum_{m \in \{rgb, rgbd\}} \lambda_m \left(\mathcal{L}_{\text{foc}}^m + \lambda_{\text{bal}} \mathcal{L}_{\text{bal}}^m\right) + \lambda_{\text{fuse}} \mathcal{L}_{\text{foc}}^{\text{fuse}}, \tag{10}$$

where λ_m and λ_{fuse} control the contributions of the stream-specific and fusion losses, respectively, and λ_{bal} weights the expert load-balancing regularization. The stream predictions and matching volumes are detached before entering the fusion gate, so that $\mathcal{L}_{\text{foc}}^{\text{fuse}}$ updates only the fusion module without inducing cross-stream co-adaptation through the fusion loss.

Inference. The query category is not assumed to be known at test time. For category assignment, we summarize the global RGBD features of normal training samples from each category by a single centroid $\boldsymbol{\mu}_c$. We use the joint RGBD representation for this global routing step so that both appearance and geometric cues contribute to category assignment. Given the global RGBD descriptor $\boldsymbol{z}^q$ of a query, the category is estimated as:

$$\hat{c} = \arg\max_c \text{sim}\left(\boldsymbol{z}^q, \boldsymbol{\mu}_c\right), \tag{11}$$

where $\hat{c}$ determines the corresponding RGB and RGBD memories used for retrieval and the category-aware bias used in pixel fusion.

The fused anomaly map S^q is directly used for pixel-level localization. For image-level detection, we follow Roth et al. (2022) and compute the anomaly score A^q as the average of the highest 1% responses in S^q.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Protocol. We evaluate UniMRE on three multimodal industrial anomaly detection benchmarks: MVTec-3D (Bergmann et al., 2021), Eyecandies (Bonfiglioli et al., 2022), and MulSen-AD (Li et al., 2025b), which provide RGB observations together with geometric information and cover both appearance- and structure-related defects. Unless otherwise specified, we follow the unified multi-class setting and train a single model per dataset to handle all object

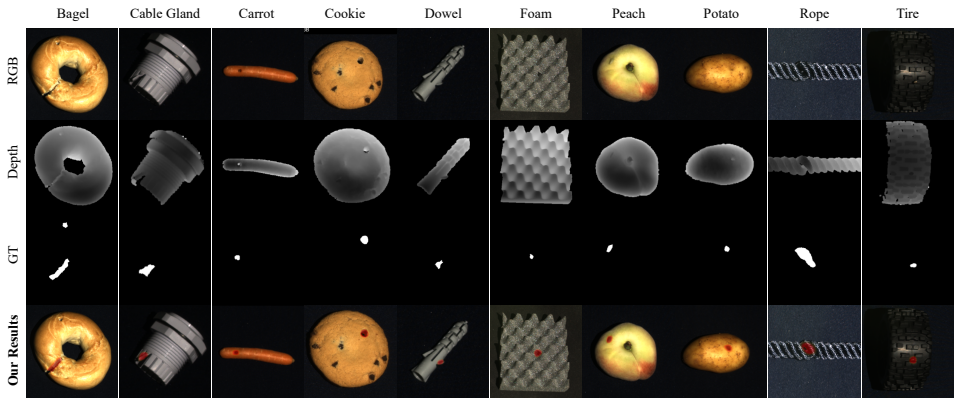


Figure 3: Qualitative anomaly localization results of UniMRE on MVTec-3D AD.

Table 1: Quantitative comparison on MVTec-3D, Eyecandies, and MulSen-AD under the full-shot setting. Best results within each group are in **bold** and second-best are underlined.

Setting	Method	MVTec-3D				Eyecandies				MulSen-AD			
		I-AUROC	I-F1	P-AUROC	AUPRO	I-AUROC	I-F1	P-AUROC	AUPRO	I-AUROC	I-F1	P-AUROC	AUPRO
Specialized multimodal	M3DM (Wang et al., 2023)	94.5	–	99.2	96.4	89.7	–	97.7	88.2	70.5	–	**61.1**	–
	CFM (Costanzino et al., 2024)	95.4	–	99.3	97.1	88.1	–	97.4	88.7	–	–	–	–
	3D-ADNAS (Long et al., 2025)	95.1	–	–	–	94.6	–	–	89.8	–	–	–	–
	FIND (Li et al., 2025c)	**97.8**	–	**99.5**	**98.3**	**95.5**	–	98.3	92.5	–	–	–	–
	GPFlow (Li et al., 2026)	97.0	–	**99.5**	**98.3**	94.2	–	**98.9**	**93.1**	–	–	–	–
	MulSen-TripleAD (Li et al., 2025b)	–	–	–	–	–	–	–	–	**91.9**	**92.8**	–	–
Unified Multimodal and Multi-class	UniMMAD (Zhao et al., 2026)	92.5	94.9	99.0	95.9	85.5	**84.1**	96.9	89.2	85.4	90.2	**97.9**	**92.1**
	UniMRE (ours)	**93.3**	**95.0**	**99.1**	**96.1**	**87.0**	82.9	**97.7**	**90.0**	**91.2**	**94.2**	96.9	90.1

categories. For image-level anomaly detection, we report I-AUROC and I-F1, while pixel-level localization is evaluated by P-AUROC and AUPRO with an FPR integration limit of 0.3 (Bergmann et al., 2019).

Implementation. Frozen DINOv3-S (Siméoni et al., 2025) and DUNE-S (Sarıyıldız et al., 2025) are used as the RGB and geometric feature encoders, respectively, and all inputs are resized to 448×448. UniMRE is trained with Adam for 150 epochs using a batch size of 8 and an initial learning rate of 1×10^{-4}, with same-class mini-batches under the unified multi-class protocol. Class-conditional RGB and RGBD normal memories are constructed from clean training patches and queried by FAISS retrieval to form stream-specific matching evidence; anomalous RGB–depth pairs are synthesized for pixel-level supervision. For the 5/10/50-shot settings, each category retains only N normal samples for both memory construction and training. Additional training, retrieval, anomaly-synthesis, and few-shot details are provided in Appendix A, and dataset-specific settings are described in Appendix B.

4.2 Comparison with Existing Methods

Tables 1 and 2 compare UniMRE with specialized one-class-one-model multimodal detectors and unified multimodal multi-class methods under full-shot and few-shot settings, respectively. Results reported by prior methods are taken from their original papers when available.

Full-Shot Results. As shown in Table 1, UniMRE consistently improves I-AUROC over the existing unified multimodal baseline UniMMAD across all three benchmarks. Specifically, I-AUROC improves from 92.5% to 93.3% on MVTec-3D, from 85.5% to 87.0% on Eyecandies, and from 85.4% to 91.2% on MulSen-AD, with the largest gain of 5.8 points on the latter. UniMRE also achieves strong localization performance, reaching 99.1% P-AUROC and 96.1% AUPRO on MVTec-3D. As illustrated in Fig. 3, the predicted anomaly maps align closely with the ground-truth masks across diverse appearance and geometric defects, while suppressing most background responses.

Compared with specialized multimodal detectors that train a separate model for each category, UniMRE achieves competitive performance on several metrics while using a single unified model for all categories. For example, on MulSen-AD, UniMRE achieves 91.2% I-AUROC, approach-

Table 2: Few-shot comparison on MVTec-3D and Eyecandies (5/10/50-shot). Best results within each group are in **bold** and second-best are underlined.

k-Shot	Classification	Method	MVTec-3D			Eyecandies		
			I-AUROC	P-AUROC	AUPRO	I-AUROC	P-AUROC	AUPRO
5	Specialized Multimodal Models	M3DM (Wang et al., 2023)	82.2	98.4	93.7	76.4	87.1	80.7
		CFM (Costanzino et al., 2024)	81.1	98.6	94.9	79.5	87.9	80.1
		3D-ADNAS (Long et al., 2025)	82.6	–	–	77.5	87.5	–
		FIND (Li et al., 2025c)	89.9	99.1	**96.3**	84.0	95.1	86.2
		GPFlow (Li et al., 2026)	**91.2**	**99.2**	**96.3**	**91.3**	**98.2**	**90.5**
		MulSen-TripleAD (Li et al., 2025b)	–	–	–	–	–	–
	Unified Multimodal and Multi-class	UniMMAD* (Zhao et al., 2026)	72.5	96.9	89.7	70.7	96.3	86.2
		UniMRE (ours)	**85.2**	**98.3**	**93.9**	**78.4**	**96.6**	**85.5**
10	Specialized Multimodal Models	M3DM (Wang et al., 2023)	84.5	98.6	94.3	82.4	89.0	81.2
		CFM (Costanzino et al., 2024)	84.5	98.7	95.4	83.8	90.3	82.5
		3D-ADNAS (Long et al., 2025)	84.8	–	–	80.7	86.9	–
		FIND (Li et al., 2025c)	92.1	99.3	96.7	86.8	95.9	88.9
		GPFlow (Li et al., 2026)	**93.5**	**99.4**	**96.8**	**92.3**	**98.3**	**90.8**
		MulSen-TripleAD (Li et al., 2025b)	–	–	–	–	–	–
	Unified Multimodal and Multi-class	UniMMAD* (Zhao et al., 2026)	76.5	97.5	91.3	73.0	**97.0**	88.1
		UniMRE (ours)	**88.1**	**98.7**	**95.1**	**80.3**	96.9	**86.4**
50	Specialized Multimodal Models	M3DM (Wang et al., 2023)	90.7	98.9	95.5	83.6	93.3	84.6
		CFM (Costanzino et al., 2024)	90.6	99.1	96.5	85.2	92.6	85.1
		3D-ADNAS (Long et al., 2025)	89.0	–	–	86.8	91.2	–
		FIND (Li et al., 2025c)	95.2	**99.5**	**97.3**	89.7	97.6	90.7
		GPFlow (Li et al., 2026)	**95.5**	**99.5**	97.1	**92.5**	**98.5**	**91.3**
		MulSen-TripleAD (Li et al., 2025b)	–	–	–	–	–	–
	Unified Multimodal and Multi-class	UniMMAD* (Zhao et al., 2026)	88.8	**99.0**	**95.9**	81.4	**98.0**	**91.2**
		UniMRE (ours)	**91.4**	**99.0**	95.8	**83.6**	97.8	89.9

ing the 91.9% of the specialized MulSen-TripleAD model (Li et al., 2025b), while substantially outperforming UniMMAD (Zhao et al., 2026) under the same unified setting. These results show that preserving category-specific normal evidence in separate RGB and RGBD retrieval spaces is effective for unified multimodal anomaly detection.

Few-Shot Results. We further evaluate UniMRE under 5-, 10-, and 50-shot settings in Table 2. Since UniMMAD (Zhao et al., 2026) does not report results under the same few-shot protocol, we reproduce its official implementation under our few-shot setting and denote the resulting baseline as UniMMAD*; further reproduction details are provided in Appendix A.

The I-AUROC advantage of UniMRE over the unified baseline becomes particularly pronounced when normal training data are scarce. On MVTec-3D, UniMRE improves I-AUROC over UniMMAD* by 12.7, 11.6, and 2.6 points under the 5-, 10-, and 50-shot settings, respectively. The corresponding gains on Eyecandies are 7.7, 7.3, and 2.2 points. The larger margins in the lowest-shot regimes indicate that class-conditional retrieval can provide informative normal references even from a limited normal training set. Meanwhile, UniMRE maintains competitive pixel-level localization and approaches its full-shot performance as additional normal samples become available.

Specialized few-shot multimodal methods such as FIND (Li et al., 2025c) and GPFlow (Li et al., 2026) remain stronger on some metrics, but require category-specific models. In contrast, UniMRE retains a single unified model across all categories, demonstrating a favorable trade-off between data efficiency and unified deployment.

4.3 Ablation Studies

Expert Pathway Configuration. We first study how different expert pathways (modality branches) affect anomaly detection and localization on MVTec-3D. As shown in Table 3, using either RGB or Depth alone yields 88.1% and 84.5% I-AUROC, respectively, indicating that a single source of evidence is insufficient for the unified multimodal setting. Combining separate RGB and Depth pathways improves I-AUROC to 92.2% and AUPRO to 95.8%, confirming the benefit of complementary appearance and geometric information. Replacing the pure Depth pathway with the joint RGBD pathway further improves performance to 93.3% I-AUROC and 96.1% AUPRO. This com-

Table 3: Ablation of expert pathway configurations.

Setting	I-AUROC	I-F1	P-AUROC	AUPRO
RGB	88.1	92.9	98.5	94.2
Depth	84.5	93.2	97.3	90.5
RGB+Depth	92.2	94.4	99.0	95.8
RGB+RGBD (UniMRE)	**93.3**	**95.0**	**99.1**	**96.1**

Table 4: Ablation of different key modules.

Setting	I-AUROC	I-F1	P-AUROC	AUPRO
w/o Class-Cond. Retrieval	92.2	94.8	**99.1**	**96.1**
w/o Cross-Stream Cond.	91.8	94.3	99.0	96.0
Retrieval Only	92.2	**95.0**	99.0	95.4
Avg. Fusion	91.1	93.9	98.9	95.4
Full UniMRE	**93.3**	**95.0**	**99.1**	**96.1**

parison supports our dual-stream design, where the RGB pathway preserves appearance-oriented evidence while the joint RGBD pathway provides a richer appearance–geometry representation.

Component Analysis. We further examine the contribution of the key components of UniMRE in Table 4. Replacing class-conditional retrieval with a shared class-agnostic retrieval space decreases I-AUROC from 93.3% to 92.2%, indicating the importance of preserving category-specific normal patterns when performing retrieval across heterogeneous object categories. Moreover, the shared memory enlarges the per-query retrieval search space by mixing normal features from all categories, whereas class-conditional memories restrict retrieval to the relevant category and thereby reduce unnecessary cross-category search. Removing cross-stream conditioning further reduces I-AUROC to 91.8% and I-F1 to 94.3%, showing that interaction between the RGB and RGBD streams during feature reasoning provides complementary information beyond independent stream processing. Using only direct retrieval evidence without retrieval-guided expert reasoning also decreases I-AUROC to 92.2% and reduces AUPRO from 96.1% to 95.4%, suggesting that the retrieved normal context benefits from learned reasoning beyond direct matching evidence. The largest image-level degradation occurs when the learnable class-aware pixel fusion is replaced with simple average fusion, for which I-AUROC drops to 91.1% and I-F1 to 93.9%. This result shows that uniformly combining the two streams is insufficient to account for their different contributions to anomaly prediction. Overall, these ablations support the complementary roles of class-conditional retrieval, cross-stream feature conditioning, retrieval-guided expert reasoning, and learnable class-aware pixel fusion in UniMRE.

4.4 Efficiency Analysis

We further evaluate the efficiency of UniMRE against representative retrieval-based multimodal detectors. As shown in Table 5, UniMRE requires 5.2 GiB of peak GPU memory and 0.185 s per image on average. For comparison, the efficiency statistics of M3DM (Wang et al., 2023) and MulSen-TripleAD (Li et al., 2025b) are taken from UniMMAD (Zhao et al., 2026). UniMRE achieves lower reported memory consumption and inference latency while maintaining a comparable parameter budget. Despite using separate RGB and RGBD memories, the proposed class-conditional dual-stream retrieval therefore introduces limited deployment overhead. These results indicate that UniMRE supports unified multimodal modeling with practical inference cost.

Table 5: Efficiency comparison.

Method	Mem. (GiB)	Params (M)	Lat. (s)
M3DM (Wang et al., 2023)	72.0	122.9	2.564
MulSen-TripleAD (Li et al., 2025b)	10.1	107.0	0.840
UniMRE	**5.2**	**100.1**	**0.185**

5 Conclusion

We presented **UniMRE**, a unified multimodal framework for multi-class industrial anomaly detection and localization. By preserving category-specific normality through class-conditional RGB and joint RGBD retrieval, while enabling complementary interaction through retrieval-guided expert reasoning and class-aware pixel-wise fusion, UniMRE effectively integrates appearance and geometric anomaly evidence within a single model. Experiments on MVTec-3D, Eyecandies, and MulSen-AD demonstrate consistent improvements in image-level AUROC over the existing unified multimodal baseline, strong few-shot performance, and competitive results against specialized one-class-one-model approaches. The favorable memory footprint and inference efficiency further support the practicality of UniMRE for unified multimodal industrial inspection.

AI USE STATEMENT

Generative AI tools were used only to assist with language polishing, grammatical correction, and improving the clarity and presentation of the manuscript. They were not used to generate research ideas, design the proposed method, conduct experiments, analyze experimental results, or produce the reported findings. All AI-assisted revisions were reviewed and verified by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work focuses on multimodal industrial anomaly detection using publicly available benchmark datasets and does not involve human subjects, personally identifiable information, or sensitive personal data. The proposed method is intended for industrial inspection and quality-control applications. We are not aware of specific ethical concerns arising from the methodology or experimental protocol beyond the general risks associated with automated decision-support systems. All experiments and results are reported to the best of our knowledge in accordance with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have made several efforts to facilitate reproducibility of UniMRE. The method is specified in Sec. 3, including dual-stream representation and class-conditional retrieval (Sec. 3.2), retrieval-guided expert reasoning with cross-stream conditioning and class-aware pixel fusion (Sec. 3.3), and the training objective together with the inference-time category assignment and top-1% image scoring protocol (Sec. 3.4). Experimental settings, evaluation metrics, and the few-shot protocol are summarized in Sec. 4.1, while the main comparisons and ablation studies are reported in Sec. 4. Additional implementation details, including optimization hyperparameters, memory construction, retrieval settings, anomaly synthesis, and dataset preprocessing, are provided in Appendix A and Appendix B. Per-category and few-shot results are included in Appendix C and Appendix D. Anonymous source code is available at `https://anonymous.4open.science/r/unimre-iclr-anon-62EA`.

REFERENCES

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.

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.

Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. *arXiv preprint arXiv:2112.09045*, 2021.

Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *Asian Conference on Computer Vision*, pp. 459–475. Springer, 2022.

Mingxiu Cai, Zhe Zhang, Gaochang Wu, Tianyou Chai, and Xiatian Zhu. Raid: Retrieval-augmented anomaly detection. *arXiv preprint arXiv:2602.19611*, 2026.

Alex Costanzino, Pierluigi Zama Ramirez, Giuseppe Lisanti, and Luigi Di Stefano. Multimodal industrial anomaly detection by crossmodal feature mapping. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17234–17243. IEEE, 2024.

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.

Lei Fan, Junjie Huang, Donglin Di, Anyang Su, Tianyou Song, Maurice Pagnucco, and Yang Song. Salvaging the overlooked: Leveraging class-aware contrastive learning for multi-class anomaly detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21419–21428. IEEE, 2025.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Univad: A training-free unified model for few-shot visual anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15194–15203. IEEE, 2025.

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

Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 8472–8480, 2024.

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.

Min Li, Jinghui He, Jiachen Li, Delong Han, Jin Wan, and Gang Li. Hgcf: Hierarchical geometry-color fusion for multimodal industrial anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1023–1031, 2025a.

Wenqiao Li, Bozhong Zheng, Xiaohao Xu, Jinye Gan, Fading Lu, Xiang Li, Na Ni, Zheng Tian, Xiaonan Huang, Shenghua Gao, et al. Multi-sensor object anomaly detection: Unifying appearance, geometry, and internal properties. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9984–9993. IEEE, 2025b.

Yiting Li, Fayao Liu, Jingyi Liao, Sichao Tian, Chuan-Sheng Foo, and Xulei Yang. Find: Few-shot anomaly inspection with normal-only multi-modal data. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025c.

Yiting Li, Xulei Yang, Jingyi Liao, Jing Zhang, and Fayao Liu. Gpflow: Gaussian prototype probability flow for unsupervised multi-modal anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43103–43112, 2026.

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

Kaifang Long, Guoyang Xie, Lianbo Ma, Jiaqi Liu, and Zhichao Lu. Revisiting multimodal fusion for 3d anomaly detection from an architectural perspective. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 12273–12281, 2025.

Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9974–9983. IEEE, 2025.

Kai Mao, Ping Wei, Yiyang Lian, Yangyang Wang, and Nanning Zheng. Beyond single-modal boundary: Cross-modal anomaly detection through visual prototype and harmonization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9964–9973. IEEE, 2025.

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. IEEE, 2022.

Mert Bülent Sarıyıldız, Philippe Weinzaepfel, Thomas Lucas, Pau De Jorge, Diane Larlus, and Yannis Kalantidis. Dune: Distilling a universal encoder from heterogeneous 2d and 3d teachers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30084–30094. IEEE, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Chengyu Tao, Xuanming Cao, and Juan Du. G 2 sf: Geometry-guided score fusion for multimodal industrial anomaly detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20551–20560. IEEE, 2025.

Chengjie Wang, Haokun Zhu, Jinlong Peng, Yue Wang, Ran Yi, Yunsheng Wu, Lizhuang Ma, and Jiangning Zhang. M3dm-nr: Rgb-3d noisy-resistant industrial anomaly detection via multimodal denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8032–8041. IEEE, 2023.

An Xiang, Zixuan Huang, Xitong Gao, Kejiang Ye, and Cheng-zhong Xu. Bridgenet: a unified multimodal framework for bridging 2d and 3d industrial anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1579–1587, 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.

Qiao Zhang, Mingwen Shao, Xinyuan Chen, Xiang Lv, and Kai Xu. Wave-mambaad: Wavelet-driven state space model for multi-class unsupervised anomaly detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20868–20877. IEEE, 2025a.

Zhe Zhang, Mingxiu Cai, Hanxiao Wang, Gaochang Wu, Tianyou Chai, and Xiatian Zhu. Costfilter-ad: Enhancing anomaly detection through matching cost filtering. *arXiv preprint arXiv:2505.01476*, 2025b.

Yuan Zhao, Youwei Pang, Lihe Zhang, Hanqi Liu, Jiaming Zuo, Huchuan Lu, and Xiaoqi Zhao. Unimmad: Unified multi-modal and multi-class anomaly detection via moe-driven feature decompression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28502–28511, 2026.

Wenbing Zhu, Lidong Wang, Ziqing Zhou, Chengjie Wang, Yurui Pan, Ruoyi Zhang, Zhuhao Chen, Linjie Cheng, Bin-Bin Gao, Jiangning Zhang, et al. Real-iad d 3: a real-world 2d/pseudo-3d/3d dataset for industrial anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15214–15223. IEEE, 2025.

UNIMRE: MODALITY-SPECIALIZED RETRIEVAL AND REASONING FOR UNIFIED MULTIMODAL ANOMALY DETECTION
SUPPLEMENTARY MATERIAL

A IMPLEMENTATION DETAILS

Training Setup. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24 GB memory. The implementation is based on PyTorch, with FAISS used for approximate nearest-neighbor retrieval. Unless otherwise stated, a fixed training seed is used for each experiment. For the few-shot robustness analysis, we additionally evaluate three independently sampled normal subsets using sampling seeds 0, 1, and 2. UniMRE is trained with Adam for 150 epochs using a batch size of 8, an initial learning rate of 1×10^{-4}, cosine decay to a minimum learning rate of 1×10^{-6}, and 5 warm-up epochs. Gradient clipping with a maximum norm of 1.0 is applied. Each mini-batch contains samples from a single category, with categories visited in a round-robin schedule to balance exposure under the unified multi-class protocol. The RGB and RGBD expert maps are supervised by focal losses with equal stream weights $\lambda_{rgb} = \lambda_{rgbd} = 1$, while expert routing is regularized by a load-balancing term with weight $\lambda_{\text{bal}} = 0.005$. The class-aware pixel fusion is supervised by an additional focal loss on the fused expert prediction before adding the nearest-normal residual, as in Eq. equation 10. Stream predictions and matching volumes are detached before entering the fusion module.

Memory Construction and Retrieval. For each object category, separate RGB and RGBD normal memories are constructed from ℓ_2-normalized patch features of clean training samples. To reduce redundancy, patch features are randomly subsampled with a keep ratio of 0.1 on MVTec-3D and MulSen-AD and 0.2 on Eyecandies, whose training sets are substantially larger. For few-shot experiments, the keep ratio is set to 1.0 so that all available normal patches are retained. Retrieval is performed using FAISS IVF indexes with $K{=}150$ nearest neighbors per query patch. The spatially aligned normal reference provided to the experts is the nearest IVF quantizer centroid assigned to each query patch, while the matching cost volume is formed from cosine distances to the K retrieved normal neighbors, as in Eqs. equation 2–equation 3.

Anomaly Synthesis. Since UniMRE follows the standard unsupervised anomaly detection setting, only normal RGB–depth pairs are available during training. To provide pixel-level supervision for learning stream-specific anomaly predictions and the fusion module, we synthesize anomalous queries from normal samples following the general idea of DRAEM-style anomaly injection (Zavrtanik et al., 2021). The synthesis procedure introduces appearance and geometric perturbations into normal observations while retaining an explicit anomaly mask as pixel-level supervision.

Figure 4 illustrates the synthesis procedure. Starting from a normal RGB image and its corresponding depth map, we generate an irregular anomaly mask using Perlin noise and restrict the perturbed region to the foreground object. The resulting mask determines the spatial support of the synthetic defect and is used as pixel-level supervision during training. For appearance perturbations, the RGB observation is locally corrupted within the masked region to simulate deviations in visual appearance. For geometric perturbations, local deformations are applied to the depth observation to simulate structural deviations from the normal surface. Representative geometric perturbations include local bump, hole, and tilt patterns. As illustrated in Figure 4, these operations produce paired anomalous RGB and depth observations together with their corresponding spatial anomaly mask.

During training, different forms of synthetic queries are randomly generated, including appearance perturbations, geometric perturbations, joint RGB–depth perturbations, and unchanged normal samples. The first two provide modality-specific abnormal cues, whereas joint perturbations expose the model to anomalies that are simultaneously observable from appearance and geometry. Normal samples are also retained during training so that the model continues to observe unperturbed normal patterns. By exposing the two streams to diverse combinations of appearance and geometric deviations, anomaly synthesis provides the supervision required for retrieval-guided expert reasoning and for integrating complementary anomaly evidence from the RGB and RGBD streams.

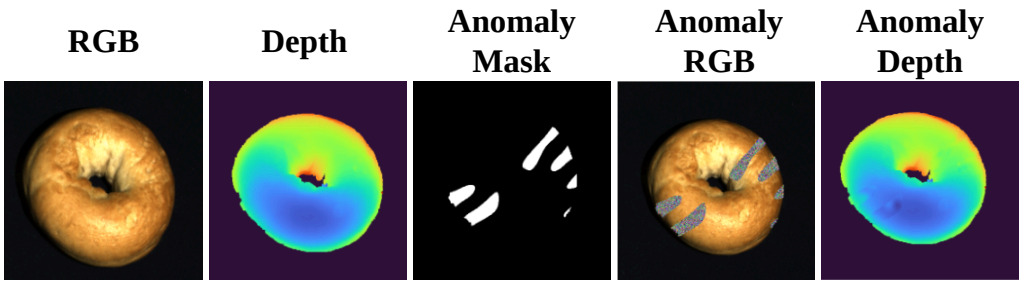


Figure 4: Illustration of the anomaly synthesis procedure. From left to right: the normal RGB image, the corresponding depth map, the generated anomaly mask, the synthesized anomalous RGB image, and the synthesized anomalous depth map. Appearance and geometric perturbations are introduced within the masked object region to construct anomalous RGB–depth queries with pixel-level supervision.

Class-Aware Pixel Fusion Details. Figure 5 provides a detailed illustration of the proposed class-aware pixel fusion module. The fusion module combines the learned anomaly responses and the direct retrieval evidence from the RGB and RGBD streams, while incorporating category-level information through a learnable bias.

Given the learned expert predictions $\bar{S}^q_{rgb}$ and $\bar{S}^q_{rgbd}$, the corresponding matching cost volumes D^q_{rgb} and D^q_{rgbd} are first resized to the prediction resolution. For each stream, the nearest-normal retrieval response is obtained by taking the minimum matching cost along the neighbor dimension:

$$\widetilde{R}^q_m(i) = \min_k \widetilde{D}^q_m(i,k), \qquad m \in \{rgb, rgbd\}, \tag{12}$$

where $\widetilde{D}^q_m$ denotes the resized matching volume. The expert prediction maps are resized to the same spatial resolution when necessary. We then concatenate the two learned predictions and the two nearest-normal retrieval responses to form the input to the fusion network:

$$U^q = [\bar{S}^q_{rgb} \,\|\, \bar{S}^q_{rgbd} \,\|\, \widetilde{R}^q_{rgb} \,\|\, \widetilde{R}^q_{rgbd}]. \tag{13}$$

Thus, the fusion network receives both the learned anomaly evidence and the non-parametric query-to-normal discrepancy from each stream.

The concatenated representation is processed by the lightweight convolutional fusion network Φ_{gate}. A learnable category-aware bias b_{c_q} is then incorporated before the sigmoid activation:

$$G^q = \sigma\left(\Phi_{\text{gate}}(U^q) + b_{c_q}\right). \tag{14}$$

The resulting gate $G^q \in [0,1]^{H\times W}$ provides a spatially varying fusion weight. A larger value of $G^q(i)$ assigns greater contribution to the RGB stream at location i, whereas a smaller value assigns greater contribution to the RGBD stream. The category-aware bias provides a lightweight category-level prior without replacing the spatially varying evidence inferred from the two streams. During training, c_q is obtained from the ground-truth object category, whereas the inferred category $\hat{c}$ is used at test time.

Importantly, the fusion module is optimized without propagating the fusion loss back into the stream-specific expert and retrieval pathways. Specifically, the expert prediction maps and matching evidence are detached before being provided to the fusion module. During training, the gate is supervised using the fused learned expert prediction

$$\bar{S}^q_{\text{fuse}} = G^q \odot \bar{S}^q_{rgb} + (1 - G^q) \odot \bar{S}^q_{rgbd}, \tag{15}$$

and a focal loss is applied to $\bar{S}^q_{\text{fuse}}$ using the synthetic anomaly mask. The direct nearest-normal residual is not added to the stream predictions when optimizing the fusion gate. This design allows the gate to learn how to combine the two expert responses without inducing cross-stream co-adaptation through the fusion loss.

At inference, the direct retrieval evidence is first restored in each stream:

$$S_m^q = \bar{S}_m^q + R_m^q, \qquad m \in \{rgb, rgbd\}, \tag{16}$$

where R_m^q denotes the nearest-normal retrieval response defined from the stream-specific matching volume. The same learned spatial gate is then applied to the residual-augmented stream-level anomaly maps:

$$S^q = G^q \odot S_{rgb}^q + (1 - G^q) \odot S_{rgbd}^q. \tag{17}$$

Consequently, the final anomaly map preserves the direct non-parametric retrieval evidence while allowing the relative contribution of appearance and appearance–geometry information to vary across spatial locations and object categories.

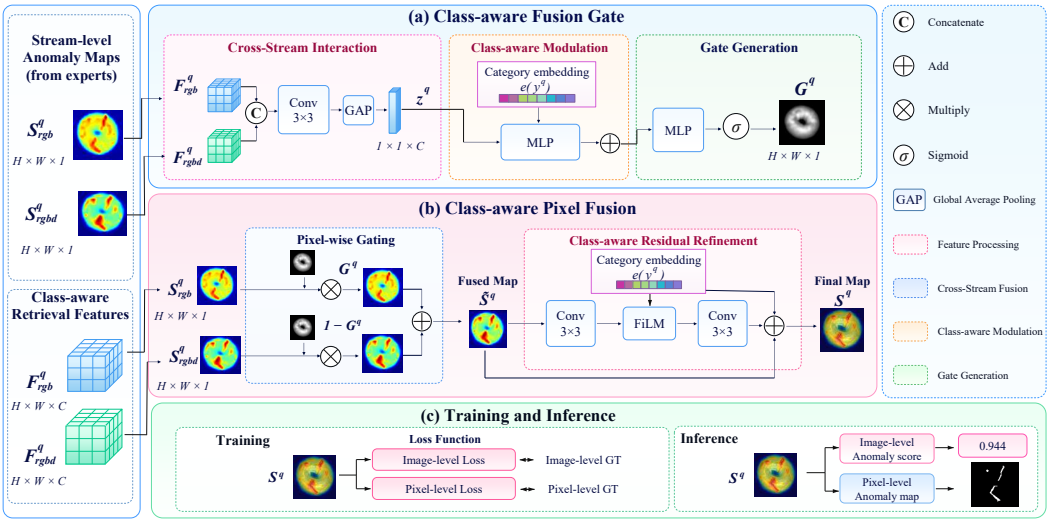


Figure 5: Detailed implementation of the learnable class-aware pixel fusion module. The learned RGB and RGBD expert predictions are combined with their nearest-normal retrieval responses to generate a spatial fusion gate. A learnable category-aware bias provides category-level modulation. During training, the gate is supervised on the fused expert prediction, with the stream predictions and retrieval evidence detached from the fusion loss. At inference, the same gate combines the residual-augmented RGB and RGBD anomaly maps to obtain the final prediction.

Few-Shot Evaluation. Under the 5/10/50-shot settings, each category retains only N normal training samples, which are used both to construct the class-conditional memories and to train the experts and class-aware pixel fusion; the test set remains unchanged. Since UniMMAD (Zhao et al., 2026) does not report results under the same few-shot protocol, we reproduce its official implementation as UniMMAD*. UniMRE and UniMMAD* use the same few-shot subsets and are both evaluated at 448×448 resolution. Following the official training recipe, UniMMAD* is trained for 200 epochs with batch size 4, while its architecture and optimization settings are otherwise retained.

B Dataset Details

We evaluate UniMRE on three multimodal industrial benchmarks that provide RGB observations together with geometric information. All experiments follow the unified multi-class protocol: a single model is trained per dataset to cover all categories.

B.1 MVTec-3D

MVTec-3D (Bergmann et al., 2021) is a real-world industrial anomaly detection benchmark with paired RGB images and 3D scans. It contains 10 object categories: bagel, cable gland, carrot, cookie, dowel, foam, peach, potato, rope, and tire. Each category provides anomaly-free training samples and a test set with both normal and anomalous instances, accompanied by pixel-level ground-truth

masks. Following common practice for RGB–depth industrial anomaly detection, we convert the provided 3D geometry into a depth map aligned with the RGB view and use the resulting RGB–depth pairs as model inputs. The dataset covers both appearance defects and geometric defects, making it a standard benchmark for multimodal industrial inspection.

B.2 EYECANDIES

Eyecandies (Bonfiglioli et al., 2022) is a synthetic multimodal anomaly detection dataset rendered under controlled lighting. It contains 10 candy categories: Candy Cane, Chocolate Cookie, Chocolate Praline, Confetto, Gummy Bear, Hazelnut Truffle, Licorice Sandwich, Lollipop, Marshmallow, and Peppermint Candy. For each sample, RGB images are provided together with depth and surface-normal maps. We use the RGB view and the corresponding depth map as inputs to UniMRE. Anomaly-free samples are used for training, while anomalous samples with pixel-level annotations are reserved for testing. Compared with MVTec-3D, Eyecandies exhibits larger intra-class appearance variation induced by procedural rendering, providing a complementary evaluation of unified multimodal detection.

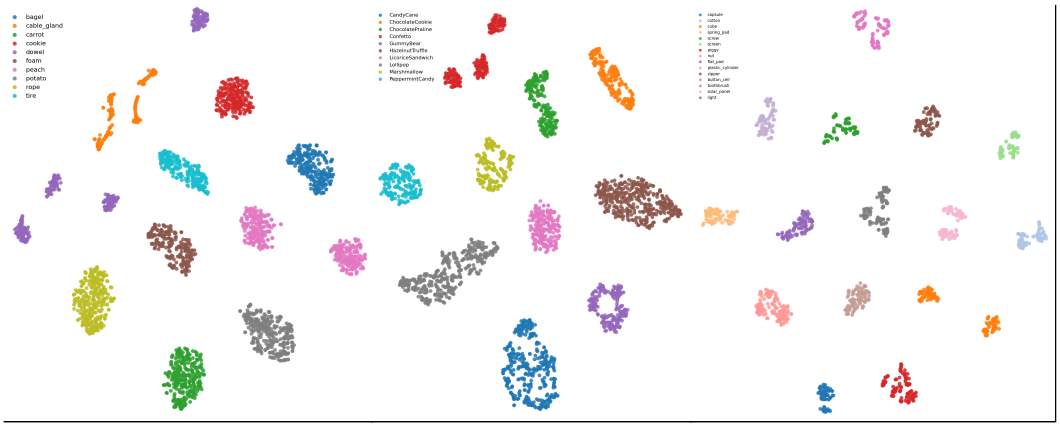
B.3 MULSEN-AD

MulSen-AD (Li et al., 2025b) is a multi-sensor industrial anomaly detection benchmark that jointly captures appearance, geometry, and internal properties. It contains 15 object categories: capsule, cotton, cube, spring pad, screw, screen, piggy, nut, flat pad, plastic cylinder, zipper, button cell, toothbrush, solar panel, and light. We use the RGB modality together with geometric information derived from the provided 3D scans. As in the other datasets, only normal samples are used for training under the unified multi-class setting, and evaluation is performed on the official test split with image- and pixel-level annotations when available. MulSen-AD is particularly informative for assessing whether a unified multimodal detector can approach specialized multi-sensor methods while retaining a single shared model.

C DETAILED RESULTS ON INDIVIDUAL CATEGORIES

To provide a more detailed view of the performance across heterogeneous object categories, we report per-category comparisons between UniMRE and UniMMAD on MVTec-3D, Eyecandies, and MulSen-AD. The results include both image-level detection metrics (I-AUROC and I-F1) and pixel-level localization metrics (P-AUROC and AUPRO). These category-wise results complement the aggregate comparisons in the main paper and further illustrate how the two methods behave across objects with different appearance and geometric characteristics. Overall, UniMRE shows clear improvements in image-level anomaly detection on many categories, while the relative performance in pixel-level localization is more category dependent.

Figure 6 provides a t-SNE visualization of the category-level feature distributions on the three datasets. Despite the substantial diversity of object appearance and geometry, samples from the same category tend to form compact clusters, while different categories remain reasonably well separated. This observation is consistent with the motivation of UniMRE, which preserves category-specific normality through class-conditional retrieval and performs reasoning on top of category-aware normal references. At the same time, the varying cluster compactness and inter-category margins also suggest that some categories are intrinsically more difficult than others, which helps explain the category-dependent behavior observed in the detailed results reported below.



(a) MVTec-3D (b) Eyecandies (c) MulSen-AD

Figure 6: t-SNE visualization of category-level feature distributions on MVTec-3D (left), Eyecandies (middle), and MulSen-AD (right). Each color denotes one object category. Samples from the same category generally form compact clusters, while different categories remain reasonably separated, indicating that the feature space preserves category-specific structure across all three datasets.

C.1 Results on MVTec-3D

As shown in Table 6, UniMRE improves the mean I-AUROC from 92.5% to 93.3% and I-F1 from 94.9% to 95.0%, while also slightly improving P-AUROC and AUPRO from 99.0% and 95.9% to 99.1% and 96.1%, respectively. At the category level, the improvement is particularly evident on Bagel, Peach, Rope, and Tire. For example, the I-AUROC on Tire increases from 70.4% to 83.2%, while Rope reaches 99.7% I-AUROC and 98.6% I-F1. UniMRE also improves localization on several categories, including Foam, where AUPRO increases from 90.3% to 94.2%. At the same time, UniMMAD remains stronger on several categories such as Cable Gland, Carrot, and Dowel at the image level. These results indicate that the overall improvement of UniMRE does not arise from uniformly improving every object category, but from better handling a subset of categories for which complementary RGB and geometric evidence is particularly beneficial.

Table 6: Per-category comparison with UniMMAD on MVTec-3D.

Category	I-AUROC		I-F1		P-AUROC		AUPRO	
	UniMMAD	UniMRE	UniMMAD	UniMRE	UniMMAD	UniMRE	UniMMAD	UniMRE
Bagel	94.4	**96.8**	94.2	**96.1**	98.8	**99.5**	94.2	**97.6**
Cable Gland	**94.8**	89.1	**96.0**	93.3	99.2	**99.3**	**97.1**	97.0
Carrot	**99.4**	94.9	**98.8**	95.3	99.6	**99.7**	**97.9**	97.8
Cookie	94.2	**98.8**	93.4	**97.6**	**99.3**	96.7	**96.9**	93.2
Dowel	**95.9**	93.5	**96.2**	95.6	**98.9**	98.6	**96.0**	93.5
Foam	**89.1**	89.0	**91.6**	91.2	97.4	**98.6**	90.3	**94.2**
Peach	95.8	**98.3**	96.3	**97.7**	99.5	**99.9**	97.5	**98.3**
Potato	**93.3**	89.2	**96.2**	95.3	99.6	**99.9**	97.8	**98.3**
Rope	97.5	**99.7**	97.1	**98.6**	99.4	**99.5**	96.4	**97.0**
Tire	70.4	**83.2**	88.7	**89.2**	98.6	**98.9**	**94.6**	94.5
Mean	92.5	**93.3**	94.9	**95.0**	99.0	**99.1**	95.9	**96.1**

C.2 Results on Eyecandies

Table 7 presents the detailed results on Eyecandies. UniMRE improves the mean I-AUROC from 85.5% to 87.0%, together with improvements in P-AUROC from 96.9% to 97.7% and AUPRO from 89.2% to 90.0%. The mean I-F1, however, decreases from 84.1% to 82.9%, showing that the gain in ranking-based image-level detection does not translate uniformly to the maximum-F1 operating point.

The category-wise results exhibit substantial variation. UniMRE obtains notable I-AUROC gains on Candy Cane, Chocolate Praline, and Hazelnut Truffle, improving from 57.4% to 70.6%, from 82.7% to 93.1%, and from 61.9% to 77.6%, respectively. Chocolate Praline and Hazelnut Truffle also show clear improvements in pixel-level localization, with AUPRO increasing from 83.0% to 93.3% and from 65.2% to 73.5%. In contrast, UniMMAD performs better on categories such as Chocolate Cookie, Confetto, and Marshmallow. The mixed category-wise behavior suggests that the benefit of multimodal retrieval and reasoning depends on the anomaly characteristics of each object, while the overall results favor UniMRE in I-AUROC and pixel-level localization.

Table 7: Per-category comparison with UniMMAD on Eyecandies.

Category	I-AUROC		I-F1		P-AUROC		AUPRO	
	UniMMAD	UniMRE	UniMMAD	UniMRE	UniMMAD	UniMRE	UniMMAD	UniMRE
Candy Cane	57.4	**70.6**	69.5	**73.3**	95.4	**95.8**	86.9	85.2
Chocolate Cookie	**97.9**	87.5	**96.0**	81.0	98.5	97.1	**93.1**	88.6
Chocolate Praline	82.7	**93.1**	76.1	**89.4**	93.2	**98.5**	83.0	**93.3**
Confetto	**97.2**	92.5	**92.0**	86.8	**99.6**	99.3	**98.0**	96.6
Gummy Bear	**90.7**	83.5	**89.7**	81.8	95.9	**97.0**	88.6	**89.7**
Hazelnut Truffle	61.9	**77.6**	69.6	**74.5**	90.6	**95.0**	65.2	**73.5**
Licorice Sandwich	**90.0**	86.4	**86.9**	82.4	**98.4**	98.4	**93.5**	93.5
Lollipop	85.1	**87.3**	**78.7**	77.4	98.0	**98.3**	90.3	**93.5**
Marshmallow	**99.8**	97.0	**98.0**	93.6	**99.5**	98.4	**96.6**	88.9
Peppermint Candy	92.6	**94.9**	84.0	**88.9**	**99.4**	**99.4**	**97.2**	97.0
Mean	85.5	**87.0**	**84.1**	82.9	96.9	**97.7**	89.2	**90.0**

C.3 Results on MulSen-AD

The per-category results on MulSen-AD are reported in Table 8. UniMRE achieves a substantially higher mean I-AUROC of 91.2%, compared with 85.4% for UniMMAD, and improves mean I-F1 from 90.2% to 94.2%. The improvement is especially pronounced for several challenging categories. For Screen, I-AUROC increases from 45.3% to 90.3%, while Nut improves from 75.0% to 95.2%, Button Cell from 75.2% to 100.0%, and Toothbrush from 81.4% to 99.2%. These gains account for a substantial part of the overall improvement in image-level anomaly detection.

The localization results are more mixed. UniMRE improves AUPRO on categories including Cotton, Spring Pad, Nut, Toothbrush, and Solar Panel, with particularly clear gains on Nut ($89.6\% \rightarrow 96.2\%$) and Solar Panel ($88.3\% \rightarrow 97.9\%$). However, UniMMAD remains stronger on several other categories, resulting in higher mean P-AUROC and AUPRO overall. Specifically, the mean P-AUROC changes from 97.9% for UniMMAD to 96.9% for UniMRE, while AUPRO changes from 92.1% to 90.1%. These results show that the main advantage of UniMRE on MulSen-AD lies in image-level anomaly discrimination, whereas pixel-level localization remains more sensitive to individual object categories and modality characteristics.

Table 8: Per-category comparison with UniMMAD on MulSen-AD.

Category	I-AUROC		I-F1		P-AUROC		AUPRO	
	UniMMAD	UniMRE	UniMMAD	UniMRE	UniMMAD	UniMRE	UniMMAD	UniMRE
Capsule	94.2	**99.8**	95.8	**99.0**	98.9	**99.2**	**95.4**	89.1
Cotton	**99.7**	89.0	**98.7**	92.7	**96.8**	96.3	**89.1**	92.8
Cube	98.1	**100.0**	93.3	**100.0**	99.7	99.3	**92.3**	91.0
Spring Pad	**91.6**	84.6	**86.6**	85.7	**94.8**	97.9	**88.0**	92.9
Screw	75.0	**83.2**	82.4	**87.3**	**99.2**	95.3	**96.1**	83.8
Screen	45.3	**90.3**	79.8	**93.3**	**90.4**	87.3	**79.0**	65.2
Piggy	97.9	**100.0**	94.9	**100.0**	**98.5**	98.6	91.0	**91.7**
Nut	75.0	**95.2**	79.9	**94.9**	**98.8**	98.6	89.6	**96.2**
Flat Pad	97.6	91.7	**93.8**	91.2	**99.7**	93.2	**98.2**	83.3
Plastic Cylinder	99.3	**100.0**	97.9	**100.0**	99.3	**99.4**	**97.0**	**97.2**
Zipper	**100.0**	93.2	**100.0**	94.7	**98.8**	94.9	**95.8**	87.6
Button Cell	75.2	**100.0**	83.9	**100.0**	**99.7**	97.9	**98.3**	96.5
Toothbrush	81.4	**99.2**	85.2	**98.0**	**98.2**	97.6	89.2	**92.7**
Solar Panel	71.0	**81.5**	**91.9**	88.6	96.9	**99.4**	88.3	**97.9**
Light	**80.1**	60.3	**88.8**	87.8	**98.5**	97.9	**94.7**	93.1
Mean	85.4	**91.2**	90.2	**94.2**	**97.9**	96.9	**92.1**	90.1

D Detailed Few-Shot Results

We further provide detailed few-shot results on MVTec-3D, Eyecandies, and MulSen-AD under the 5-, 10-, and 50-shot settings. For each category, only N normal training samples are retained and used for both model training and normal-memory construction, while the original test set is kept unchanged. Since UniMMAD (Zhao et al., 2026) does not report results under the same few-shot protocol, we reproduce its official implementation and denote the resulting baseline as UniMMAD*. For a fair comparison, UniMRE and UniMMAD* use the same few-shot subsets for each category and are evaluated under the same input resolution. The following subsections report the detailed results for the three datasets across all few-shot settings, including both image-level detection and pixel-level localization metrics.

D.1 Few-Shot Results on MVTec-3D

The detailed results on MVTec-3D are reported in Tables 9–11. UniMRE consistently improves image-level anomaly detection over UniMMAD* across all three few-shot settings. The mean I-AUROC increases from 72.5% to 85.2% under 5-shot, from 76.5% to 88.1% under 10-shot, and from 88.8% to 91.4% under 50-shot supervision. The largest gains therefore occur in the more data-scarce regimes, indicating that the retrieval-based use of category-specific normal references remains effective when only a small number of normal training samples are available.

The category-wise results show that these improvements are distributed across a broad range of objects rather than being driven by a single category. Under the 5-shot setting, particularly clear I-AUROC gains are observed on Tire, Foam, Cookie, Peach, and Potato, while the advantage remains pronounced for Tire, Foam, Peach, and Potato in the 10-shot setting. With 50 training samples, the gap becomes smaller as UniMMAD* benefits from the increased amount of training data, but UniMRE still achieves higher I-AUROC on most categories, including Carrot, Cookie, Foam, Peach, and Potato.

Pixel-level localization shows a similar advantage in the lower-shot regimes. UniMRE improves the mean P-AUROC/AUPRO from $96.9\%/89.7\%$ to $98.3\%/93.9\%$ in the 5-shot setting and from $97.5\%/91.3\%$ to $98.7\%/95.1\%$ in the 10-shot setting. At 50 shots, the two methods become comparable in localization, with the same rounded mean P-AUROC of 99.0% and AUPRO values of 95.8% for UniMRE and 95.9% for UniMMAD*. These results suggest that the main benefit of UniMRE under limited normal supervision is its stronger image-level discrimination, while the localization gap gradually narrows as more normal samples become available.

Table 9: Per-category comparison with UniMMAD* under the 5-shot setting on MVTec-3D.

Category	I-AUROC		P-AUROC		AUPRO	
	UniMMAD*	UniMRE	UniMMAD*	UniMRE	UniMMAD*	UniMRE
Bagel	91.5	**94.4**	**99.4**	99.3	**97.2**	97.1
Cable Gland	47.0	**57.1**	94.2	**96.0**	82.2	**87.7**
Carrot	84.4	**86.6**	**99.7**	99.6	97.7	**97.7**
Cookie	79.7	**97.2**	**98.2**	95.5	**95.0**	91.1
Dowel	72.1	**80.4**	**98.3**	97.5	**93.3**	89.9
Foam	65.8	**81.2**	88.6	**98.0**	68.1	**92.0**
Peach	76.6	**95.3**	99.4	**99.8**	97.2	**98.1**
Potato	67.8	**80.3**	99.0	**99.7**	95.9	**97.8**
Rope	91.2	**98.0**	99.2	**99.4**	96.0	**96.0**
Tire	48.9	**81.7**	93.2	**98.1**	74.0	**91.1**
Mean	72.5	**85.2**	96.9	**98.3**	89.7	**93.9**

Table 10: Per-category comparison with UniMMAD* under the 10-shot setting on MVTec-3D.

Category	I-AUROC		P-AUROC		AUPRO	
	UniMMAD*	UniMRE	UniMMAD*	UniMRE	UniMMAD*	UniMRE
Bagel	91.9	**96.4**	**99.4**	**99.4**	97.0	**97.4**
Cable Gland	62.5	**67.8**	96.4	**97.7**	89.2	**92.4**
Carrot	83.8	**93.1**	**99.7**	**99.7**	**98.0**	97.8
Cookie	85.0	**95.9**	**98.8**	95.8	**96.3**	91.4
Dowel	**81.1**	80.8	**99.0**	98.0	**95.6**	91.7
Foam	66.3	**81.7**	90.8	**98.2**	74.9	**92.8**
Peach	82.5	**97.5**	99.5	**99.9**	97.4	**98.3**
Potato	68.3	**84.4**	99.0	**99.9**	95.7	**98.2**
Rope	95.5	**99.0**	**99.6**	**99.6**	**97.8**	97.0
Tire	47.8	**84.6**	92.3	**98.9**	71.6	**94.5**
Mean	76.5	**88.1**	97.5	**98.7**	91.3	**95.1**

Table 11: Per-category comparison with UniMMAD* under the 50-shot setting on MVTec-3D.

Category	I-AUROC		P-AUROC		AUPRO	
	UniMMAD*	UniMRE	UniMMAD*	UniMRE	UniMMAD*	UniMRE
Bagel	94.1	**96.8**	99.5	**99.5**	96.9	**97.7**
Cable Gland	**86.3**	75.8	**98.9**	98.8	**96.3**	95.8
Carrot	91.6	**97.8**	**99.8**	99.7	**98.1**	98.0
Cookie	92.7	**99.5**	**98.9**	97.4	**96.4**	94.4
Dowel	**95.6**	89.5	**99.7**	98.5	**97.9**	93.0
Foam	80.3	**87.1**	94.4	**97.8**	84.3	**91.6**
Peach	93.4	**98.6**	99.8	**99.9**	98.0	**98.3**
Potato	78.7	**91.2**	99.4	**99.9**	96.6	**98.3**
Rope	97.5	**98.9**	**99.7**	99.5	**98.1**	97.2
Tire	77.6	**79.2**	**99.4**	98.8	**96.5**	93.7
Mean	88.8	**91.4**	99.0	**99.0**	**95.9**	95.8

D.2 Few-Shot Results on Eyecandies

The detailed few-shot results on Eyecandies are reported in Tables 12–14. UniMRE consistently improves image-level anomaly detection over UniMMAD* across all three shot settings. The mean I-AUROC increases from 70.7% to 78.4% under 5-shot supervision, from 73.0% to 80.3% under 10-shot supervision, and from 81.4% to 83.6% under 50-shot supervision. As on MVTec-3D, the improvement is more pronounced when fewer normal samples are available, with gains of 7.7 and 7.3 percentage points under the 5-shot and 10-shot settings, respectively.

The category-wise results show that the improvement is mainly contributed by several categories that remain difficult for UniMMAD* under limited supervision. For example, Chocolate Praline improves from 69.0% to 89.4% I-AUROC in the 5-shot setting and from 72.5% to 94.1% in the 10-shot setting. Hazelnut Truffle and Lollipop also show clear image-level improvements in the low-shot regimes, while Peppermint Candy reaches 94.9% I-AUROC with 10 training samples. With 50 shots, the difference becomes smaller, although UniMRE still improves several categories, including Chocolate Praline, Licorice Sandwich, Lollipop, and Peppermint Candy.

In contrast to the image-level results, pixel-level localization is more mixed on Eyecandies. Under the 5-shot setting, UniMRE slightly improves mean P-AUROC from 96.3% to 96.6%, while its mean AUPRO is slightly lower (85.5% versus 86.2%). For the 10-shot and 50-shot settings, UniMMAD* retains slightly higher mean localization performance, with P-AUROC/AUPRO values of 97.0%/88.1% and 98.0%/91.2%, compared with 96.9%/86.4% and 97.8%/89.9% for UniMRE. This category-dependent behavior shows that the main benefit of UniMRE on Eyecandies lies in image-level anomaly discrimination, particularly when normal training data are scarce, while improvements in pixel-level localization are not uniform across categories.

Table 12: Per-category comparison with UniMMAD* under the 5-shot setting on Eyecandies.

Category	I-AUROC		P-AUROC		AUPRO	
	UniMMAD*	UniMRE	UniMMAD*	UniMRE	UniMMAD*	UniMRE
Candy Cane	45.1	**50.7**	**96.6**	92.8	**87.3**	73.1
Chocolate Cookie	**81.3**	73.8	95.4	**96.0**	85.3	**85.6**
Chocolate Praline	69.0	**89.4**	94.0	**97.3**	81.0	**88.9**
Confetto	79.7	**91.5**	99.3	**99.6**	97.0	**97.6**
Gummy Bear	**78.5**	76.3	**97.7**	96.8	**90.6**	86.5
Hazelnut Truffle	64.5	**77.1**	89.7	**94.1**	59.1	**69.0**
Licorice Sandwich	64.6	**77.0**	**97.4**	97.1	**91.0**	88.8
Lollipop	52.6	**64.9**	95.3	**96.6**	80.8	**85.5**
Marshmallow	**95.0**	93.8	**99.2**	98.3	**94.7**	89.6
Peppermint Candy	76.8	**89.3**	**98.8**	97.7	**95.3**	90.1
Mean	70.7	**78.4**	96.3	**96.6**	**86.2**	85.5

Table 13: Per-category comparison with UniMMAD* under the 10-shot setting on Eyecandies.

Category	I-AUROC		P-AUROC		AUPRO	
	UniMMAD*	UniMRE	UniMMAD*	UniMRE	UniMMAD*	UniMRE
Candy Cane	**47.4**	43.5	**96.8**	91.9	**88.9**	73.6
Chocolate Cookie	**71.8**	69.6	95.8	**96.0**	**86.4**	83.8
Chocolate Praline	72.5	**94.1**	95.0	**97.9**	84.0	**91.3**
Confetto	82.6	**91.0**	**99.4**	98.9	**97.3**	95.4
Gummy Bear	77.9	**83.2**	**97.6**	96.5	**90.5**	85.5
Hazelnut Truffle	70.7	**80.6**	92.0	**96.2**	63.4	**76.2**
Licorice Sandwich	67.2	**82.2**	**98.2**	96.8	**93.1**	87.3
Lollipop	59.4	**69.1**	96.9	**97.8**	86.8	**87.9**
Marshmallow	**96.0**	95.0	**99.2**	98.3	**94.5**	88.6
Peppermint Candy	85.0	**94.9**	**99.2**	98.7	**96.3**	93.8
Mean	73.0	**80.3**	**97.0**	96.9	**88.1**	86.4

Table 14: Per-category comparison with UniMMAD* under the 50-shot setting on Eyecandies.

Category	I-AUROC		P-AUROC		AUPRO	
	UniMMAD*	UniMRE	UniMMAD*	UniMRE	UniMMAD*	UniMRE
Candy Cane	**50.2**	45.6	**97.2**	93.8	**90.7**	79.5
Chocolate Cookie	88.5	**89.3**	**97.4**	97.3	89.7	**89.7**
Chocolate Praline	89.6	**91.8**	97.3	**98.3**	92.0	**93.1**
Confetto	**97.0**	95.2	**99.6**	**99.6**	**97.9**	97.6
Gummy Bear	82.4	**84.8**	**97.9**	97.4	**91.4**	89.0
Hazelnut Truffle	**76.3**	74.7	94.2	**96.3**	71.7	**76.7**
Licorice Sandwich	76.5	**83.5**	**98.6**	98.0	**94.4**	92.4
Lollipop	68.8	**78.7**	98.1	**98.7**	89.6	**92.8**
Marshmallow	**95.7**	95.0	**99.6**	99.0	**96.6**	92.3
Peppermint Candy	89.3	**97.4**	**99.6**	99.2	**97.5**	96.2
Mean	81.4	**83.6**	**98.0**	97.8	**91.2**	89.9

D.3 Few-Shot Results on MulSen-AD

Although the main paper focuses its few-shot comparison on MVTec-3D and Eyecandies, we additionally report detailed results on MulSen-AD here to provide a more complete evaluation under limited normal supervision. Tables 15–17 present the category-wise comparisons between UniMRE and UniMMAD* under the 5-, 10-, and 50-shot settings.

The advantage of UniMRE is most pronounced in image-level anomaly detection when only a small number of normal samples are available. Under the 5-shot setting, the mean I-AUROC increases from 73.5% for UniMMAD* to 88.4% for UniMRE, corresponding to a gain of 14.9 percentage points. With 10 training samples, UniMRE achieves 89.2% I-AUROC compared with 81.9% for UniMMAD*. The category-wise results show particularly clear improvements on Screen, Screw, Button Cell, Toothbrush, and Light in the low-shot regimes. For example, under 5-shot supervision, the I-AUROC increases from 45.3% to 83.7% on Screen, from 46.7% to 87.4% on Screw, and from 53.9% to 96.2% on Toothbrush.

Table 15: Per-category comparison with UniMMAD* under the 5-shot setting on MulSen-AD.

Category	I-AUROC		P-AUROC		AUPRO	
	UniMMAD*	UniMRE	UniMMAD*	UniMRE	UniMMAD*	UniMRE
Capsule	76.7	**87.7**	99.0	**99.2**	**97.1**	91.5
Cotton	86.0	**95.0**	93.2	**94.1**	86.7	**91.7**
Cube	97.6	**100.0**	**99.9**	98.6	**97.8**	88.6
Spring Pad	**81.0**	78.3	**99.7**	99.1	**98.0**	96.8
Screw	46.7	**87.4**	**98.5**	97.3	**95.2**	90.6
Screen	45.3	**83.7**	**90.1**	84.4	**80.0**	53.6
Piggy	77.7	**94.7**	97.3	**97.8**	86.9	**87.7**
Nut	55.6	**76.6**	**99.8**	98.5	**97.9**	95.5
Flat Pad	**96.4**	75.3	**99.7**	95.7	**98.0**	91.3
Plastic Cylinder	**99.7**	96.4	**99.6**	98.3	**97.8**	94.7
Zipper	**94.3**	81.0	**97.0**	90.3	**94.6**	81.8
Button Cell	56.8	**100.0**	**99.6**	99.5	**98.1**	98.0
Toothbrush	53.9	**96.2**	**98.8**	93.7	**94.7**	77.4
Solar Panel	74.9	**85.9**	91.7	**99.0**	80.6	**97.0**
Light	59.8	**87.5**	96.0	**97.2**	**92.7**	91.3
Mean	73.5	**88.4**	**97.3**	96.2	**93.1**	88.5

Table 16: Per-category comparison with UniMMAD* under the 10-shot setting on MulSen-AD.

Category	I-AUROC		P-AUROC		AUPRO	
	UniMMAD*	UniMRE	UniMMAD*	UniMRE	UniMMAD*	UniMRE
Capsule	84.0	**96.3**	**99.2**	98.8	**97.3**	82.3
Cotton	**92.8**	81.2	**96.6**	95.2	**92.5**	92.2
Cube	95.3	**100.0**	**99.9**	98.4	**97.9**	86.6
Spring Pad	85.1	**87.5**	**99.7**	99.0	**98.0**	95.8
Screw	57.2	**90.0**	**98.6**	97.5	**95.8**	91.8
Screen	54.7	**74.1**	**92.8**	83.5	**84.8**	58.6
Piggy	94.6	**95.3**	97.8	**98.3**	88.0	**90.3**
Nut	68.3	**90.3**	**99.8**	95.6	**98.1**	90.6
Flat Pad	**96.6**	94.3	**99.8**	92.7	**98.1**	85.0
Plastic Cylinder	**100.0**	**100.0**	**99.6**	98.0	**97.8**	93.6
Zipper	**97.6**	71.1	**96.8**	90.3	**94.2**	79.7
Button Cell	70.3	**98.7**	**99.7**	97.4	**98.1**	95.8
Toothbrush	73.5	**90.5**	**99.1**	96.2	**95.3**	84.6
Solar Panel	**90.3**	79.7	96.5	**97.4**	92.0	**95.3**
Light	67.6	**88.3**	**97.3**	97.1	**94.8**	91.1
Mean	81.9	**89.2**	**98.2**	95.7	**94.9**	87.6

Table 17: Per-category comparison with UniMMAD* under the 50-shot setting on MulSen-AD.

Category	I-AUROC		P-AUROC		AUPRO	
	UniMMAD*	UniMRE	UniMMAD*	UniMRE	UniMMAD*	UniMRE
Capsule	96.5	**99.4**	**99.2**	98.8	**97.4**	82.1
Cotton	**100.0**	69.0	**96.2**	**96.2**	92.3	**93.8**
Cube	98.0	**100.0**	**99.9**	98.8	**97.6**	89.5
Spring Pad	**92.0**	77.5	98.6	**98.7**	**95.1**	93.6
Screw	73.8	**93.2**	**99.4**	95.3	**97.6**	87.3
Screen	59.3	**79.7**	**92.9**	84.4	**87.1**	57.9
Piggy	96.7	**100.0**	**98.8**	98.6	90.8	**91.8**
Nut	78.4	**97.9**	**99.9**	96.3	**98.3**	92.2
Flat Pad	**97.4**	94.3	**99.8**	92.4	**98.1**	74.4
Plastic Cylinder	99.7	**100.0**	**99.5**	99.2	**97.7**	96.8
Zipper	**99.4**	87.2	**98.7**	94.5	**96.5**	86.9
Button Cell	85.0	**100.0**	**99.7**	99.2	**98.2**	97.8
Toothbrush	84.6	**95.8**	**99.2**	95.5	**94.3**	89.4
Solar Panel	**94.9**	79.2	97.4	**98.9**	93.4	**97.2**
Light	**81.5**	65.3	**98.5**	97.9	**96.1**	90.9
Mean	89.1	**89.2**	**98.5**	96.3	**95.4**	88.1

As the number of normal training samples increases, the image-level gap gradually narrows. At 50 shots, UniMRE and UniMMAD* achieve nearly identical mean I-AUROC values of 89.2% and 89.1%, respectively, although their category-wise behavior remains different. This trend further indicates that the benefit of retrieval-guided reasoning is strongest when the available normal training data are particularly limited.

Pixel-level localization exhibits a different trend. UniMMAD* obtains higher mean P-AUROC and AUPRO under all three few-shot settings, while UniMRE remains competitive on several individual categories such as Cotton, Piggy, Solar Panel, and Light depending on the shot setting. The contrast between image-level and pixel-level results suggests that the main few-shot advantage of UniMRE on MulSen-AD lies in anomaly discrimination at the image level, whereas localization performance remains more sensitive to individual categories.

E Additional Experimental Analysis

We provide several additional analyses to examine factors that may affect the interpretation of the main results. First, we repeat the few-shot experiments on MVTec-3D using different random selections of normal training samples to evaluate robustness to the particular few-shot subset. Second, because UniMRE uses a higher input resolution than the original UniMMAD implementation, we reproduce UniMMAD at the same 448×448 resolution to isolate the effect of input size. Together,

these experiments provide additional evidence that the observed performance trends are not specific to a particular few-shot split or caused by the difference in input resolution. Third, we evaluate the accuracy of inference-time category assignment, which determines the class-conditional memories and the category-aware fusion bias used by UniMRE. We further provide a qualitative analysis of the complementary roles of the dual streams. Together, these experiments clarify the reliability of the reported trends under alternative few-shot splits, resolution-matched comparison, and the category-routing step required by class-conditional retrieval, and further illustrate the complementary roles of the dual streams.

E.1 Few-Shot Results with Different Random Seeds

For the few-shot robustness analysis, we evaluate three independently sampled normal subsets using sampling seeds 0, 1, and 2, where Seed 0 corresponds to the split used in the main paper. For each category, we independently resample the normal training subset with `fewshot_seed` $= 1$ and `fewshot_seed` $= 2$, while keeping the model configuration, training procedure, input resolution, and number of samples per category unchanged. The selected samples are used for both memory construction and few-shot training, following the same protocol as the experiments reported in the main paper.

Table 18 reports the results obtained from the original split used in the main paper (Seed 0) together with the two additional randomly sampled subsets (Seeds 1 and 2). Across the three subsets, UniMRE preserves the same overall performance trend as the number of available normal samples increases. For image-level detection, the mean I-AUROC reaches 84.1%, 86.9%, and 91.7% under the 5-, 10-, and 50-shot settings, respectively. The corresponding standard deviations are 0.8, 1.0, and 0.4 percentage points, indicating moderate sensitivity to support-set composition in the lowest-data regimes and reduced variation as more normal samples become available.

Pixel-level localization is considerably more consistent across the three subsets. P-AUROC remains within 0.1 percentage points across all evaluated seeds, while AUPRO varies by at most 0.6 percentage points. In particular, the mean P-AUROC values are 98.3%, 98.6%, and 99.0% for 5, 10, and 50 shots, respectively, with corresponding mean AUPRO values of 93.7%, 94.7%, and 95.8%.

These results show that, although image-level detection exhibits some expected variation under extremely limited normal supervision, the main few-shot trend remains consistent across independently sampled normal subsets. The additional evaluations therefore provide evidence that the few-shot performance reported in the main paper is not specific to a single favorable selection of normal training samples.

Table 18: Few-shot robustness of UniMRE on MVTec-3D under three random selections of normal training samples. Seed 0 corresponds to the few-shot split reported in the main paper, while Seeds 1 and 2 use independently resampled normal subsets. Mean and standard deviation are computed over the three sampling seeds.

Shots	Sampling Seed	I-AUROC	P-AUROC	AUPRO
5	0	85.2	98.3	93.9
	1	83.8	98.3	93.6
	2	83.3	98.3	93.6
	Mean $\pm$ Std	**84.1** $\pm$ **0.8**	**98.3** $\pm$ **0.0**	**93.7** $\pm$ **0.1**
10	0	88.1	98.7	95.1
	1	85.7	98.6	94.5
	2	86.9	98.6	94.5
	Mean $\pm$ Std	**86.9** $\pm$ **1.0**	**98.6** $\pm$ **0.0**	**94.7** $\pm$ **0.3**
50	0	91.4	99.0	95.8
	1	92.2	99.0	95.9
	2	91.5	98.9	95.6
	Mean $\pm$ Std	**91.7** $\pm$ **0.4**	**99.0** $\pm$ **0.0**	**95.8** $\pm$ **0.1**

E.2 UniMMAD Results at 448×448 Resolution

The original UniMMAD implementation evaluates MVTec-3D using an input resolution of 256×256, whereas UniMRE adopts 448×448 throughout our experiments. To examine whether the performance difference is mainly caused by the higher input resolution, we additionally reproduce UniMMAD using its official code with the input resolution increased to 448×448. All other settings are kept consistent with our UniMMAD reproduction.

Table 19: Comparison with UniMMAD on MVTec-3D under different input resolutions. UniMMAD* denotes our reproduction using the official implementation.

Method	Resolution	I-AUROC	I-F1	P-AUROC	AUPRO
UniMMAD	256×256	92.5	94.9	99.1	95.9
UniMMAD*	448×448	92.0	93.7	**99.1**	**96.2**
UniMRE	448×448	**93.3**	**95.0**	**99.1**	96.1

As shown in Table 19, increasing the input resolution of UniMMAD to 448×448 does not account for the image-level improvement of UniMRE. The reproduced UniMMAD* obtains 92.0% I-AUROC and 93.7% I-F1 at the higher resolution, compared with 93.3% and 95.0% for UniMRE under the same input resolution. In particular, UniMRE retains a 1.3 percentage-point advantage in both I-AUROC and I-F1 after controlling for input resolution.

The effect on pixel-level localization is considerably smaller. Both methods achieve the same rounded P-AUROC of 99.1%, while UniMMAD* obtains a slightly higher AUPRO of 96.2% compared with 96.1% for UniMRE. The reproduced 448×448 UniMMAD also remains close to its originally reported 256×256 result, indicating that increasing the resolution alone does not lead to a corresponding improvement in image-level anomaly detection. These results show that the image-level advantage of UniMRE on MVTec-3D persists under a resolution-matched comparison rather than arising from the difference in input size.

E.3 Class Routing Accuracy

UniMRE relies on class-conditional normal memories and a category-aware bias in pixel fusion. At inference time, the query category is therefore not assumed to be known a priori. Instead, as described in Sec. 3.4, each query is assigned to a category by nearest-centroid matching in the joint RGBD feature space (Eq. equation 11): the global RGBD descriptor of the query is compared with category prototypes computed from normal training samples, and the nearest prototype determines the estimated category $\hat{c}$. The estimated category selects the corresponding RGB and RGBD memories for retrieval and the category-aware fusion bias. Because incorrect routing would retrieve normal references from the wrong category, the reliability of this assignment is important for interpreting the unified multi-class results. We therefore evaluate class routing accuracy on the full test sets of MVTec-3D, Eyecandies, and MulSen-AD under the same encoders and 448×448 input resolution as in the main experiments.

For each dataset, category prototypes are constructed exclusively from anomaly-free training samples using global RGBD features. Every test image is then assigned to the nearest prototype under cosine similarity. We report overall routing accuracy together with the accuracies restricted to normal and anomalous test samples. This split is informative because anomalous queries may deviate from the normal prototypes used for category assignment, and a substantial drop on anomalous samples would indicate that routing becomes unreliable precisely when anomaly evidence is present.

As shown in Table 20, routing is perfect on MVTec-3D and Eyecandies: all 1197 and 500 test images, respectively, are assigned to the correct category, including both normal and anomalous samples. These results indicate that the joint RGBD global descriptors form well-separated category centroids on the two datasets, which is consistent with the compact and separated category clusters observed in the t-SNE visualization in Fig. 6. Consequently, class-conditional retrieval on MVTec-3D and Eyecandies operates under correctly selected normal memories throughout the test evaluation.

Table 20: Inference-time class routing accuracy (%) on MVTec-3D, Eyecandies, and MulSen-AD. Routing follows the nearest-centroid rule in Eq. equation 11 using global RGBD features.

Dataset	#Test	Overall	Normal	Anomaly
MVTec-3D	1197	**100.0**	**100.0**	**100.0**
Eyecandies	500	**100.0**	**100.0**	**100.0**
MulSen-AD	644	95.3	91.5	96.5

On MulSen-AD, overall routing accuracy remains high at 95.3%. Interestingly, anomalous samples are routed more accurately (96.5%) than normal samples (91.5%), suggesting that the residual errors are not primarily caused by anomaly-induced deviations from the normal prototypes. Instead, most mistakes concentrate on a small number of categories with similar appearance or geometry. Table 21 reports the per-category routing accuracy on MulSen-AD. Eleven of the fifteen categories achieve perfect assignment. The remaining errors are largely associated with `spring_pad` (50.0%), `flat_pad` (87.5%), and `screen` (88.1%), while `plastic_cylinder` and `toothbrush` remain above 94%. In particular, `spring_pad` is difficult already in the normal split (10.0% routing accuracy among its normal test samples), indicating category confusion rather than anomaly-specific degradation.

Table 21: Per-category class routing accuracy (%) on MulSen-AD. N, N_n, and N_a denote the numbers of all / normal / anomalous test samples in each category.

Category	N	Acc.	N_n	Acc$_n$	N_a	Acc$_a$
button_cell	41	100.0	10	100.0	31	100.0
capsule	58	100.0	10	100.0	48	100.0
cotton	50	100.0	10	100.0	40	100.0
cube	51	100.0	10	100.0	41	100.0
flat_pad	40	87.5	10	100.0	30	83.3
light	46	100.0	10	100.0	36	100.0
nut	39	100.0	10	100.0	29	100.0
piggy	40	100.0	10	100.0	30	100.0
plastic_cylinder	38	94.7	10	100.0	28	92.9
screen	42	88.1	10	70.0	32	93.8
screw	41	100.0	10	100.0	31	100.0
solar_panel	49	100.0	10	100.0	39	100.0
spring_pad	34	50.0	10	10.0	24	66.7
toothbrush	35	97.1	11	90.9	24	100.0
zipper	40	100.0	12	100.0	28	100.0
Mean / Overall	644	95.3	153	91.5	491	96.5

Overall, these results show that the inference-time category assignment used by UniMRE is highly reliable under our evaluation protocol. Routing is exact on MVTec-3D and Eyecandies, and remains accurate on MulSen-AD despite a few visually confusable categories. Therefore, category-routing errors are unlikely to be a major confounding factor in the reported results. For completeness, we note that all routing evaluations use the same frozen DINOv3-S and DUNE-S encoders as the main experiments and do not involve the trainable Phase II experts.

E.4 Complementary Roles of the Dual Streams

The dual-stream design of UniMRE is motivated by the observation that appearance and geometric cues are complementary for industrial defects: some anomalies are primarily visible in RGB, whereas others are more reliably revealed by depth-related structure. To make this complementarity concrete, Figure 7 compares the RGB-stream prediction, the RGBD-stream prediction, and the fused prediction on two representative MVTec-3D examples.

In the first example, the defect is a local geometric indentation on a foam surface. The RGB stream activates around the defective region but produces a comparatively diffuse response, reflecting the

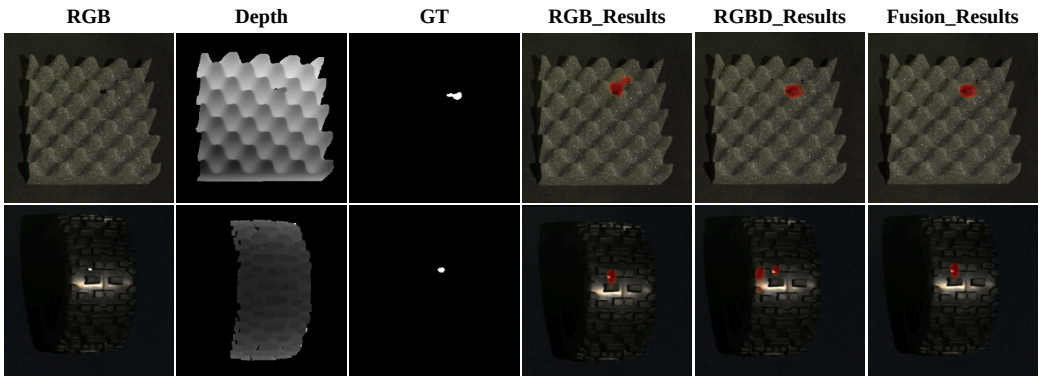


Figure 7: Qualitative comparison of dual-stream predictions on MVTec-3D. From left to right: RGB image, depth map, ground-truth mask, RGB-stream prediction, RGBD-stream prediction, and fused prediction. Top: a geometry-dominated foam defect, where the RGBD stream provides sharper localization than the RGB stream. Bottom: an appearance-dominated tire defect, where the RGB stream is more accurate and fusion suppresses a false activation from the RGBD stream.

limited appearance contrast of this structural change. By contrast, the RGBD stream yields a sharper localization that better matches the ground-truth mask, indicating that joint appearance–geometry evidence is more discriminative for this geometry-dominated anomaly. The fused prediction preserves the more precise RGBD response and suppresses the surrounding ambiguity of the RGB map.

In the second example, the defect is a small appearance mark on a tire, with little geometric deformation. Here the RGB stream localizes the anomaly accurately, whereas the RGBD stream introduces an additional false activation away from the true defect. After class-aware pixel fusion, the false response is suppressed and the final map remains concentrated on the appearance anomaly highlighted by the RGB stream.

These two cases illustrate complementary failure modes of the individual streams: the RGB stream is more reliable for appearance-dominated defects, while the RGBD stream is more reliable for geometry-dominated defects. Learnable fusion integrates the two evidence sources in a spatially adaptive manner, producing a more balanced localization than either stream alone. This qualitative analysis is consistent with the stream-configuration ablation in Sec. 4.3, where combining complementary streams outperforms using a single modality.

F Additional Qualitative Comparisons

To complement the quantitative comparisons, we provide additional qualitative results on MVTec-3D AD and Eyecandies in Figures 8 and 9. These examples visualize the predicted anomaly maps together with the corresponding RGB observations, depth information, and ground-truth anomaly regions. The qualitative comparisons demonstrate the capability of UniMRE to localize diverse anomaly patterns under different appearance and geometric conditions. In particular, UniMRE consistently produces concentrated responses around the actual defective regions while suppressing irrelevant background activations, indicating that the proposed retrieval-guided dual-stream reasoning effectively integrates complementary appearance and geometric evidence.

F.1 Qualitative Results on MVTec-3D AD

Figure 8 presents qualitative comparisons on MVTec-3D AD. The selected examples cover objects with diverse visual appearances and geometric structures, including both appearance-dominated and geometry-sensitive anomaly types. Compared with existing methods, UniMRE generates more localized anomaly responses that better align with the ground-truth defect regions. For objects where anomalies mainly appear as surface changes, the RGB stream provides informative appearance cues, while structural deviations are further captured through the RGBD stream. The final fused prediction combines these complementary signals and produces consistent localization results across different object categories.

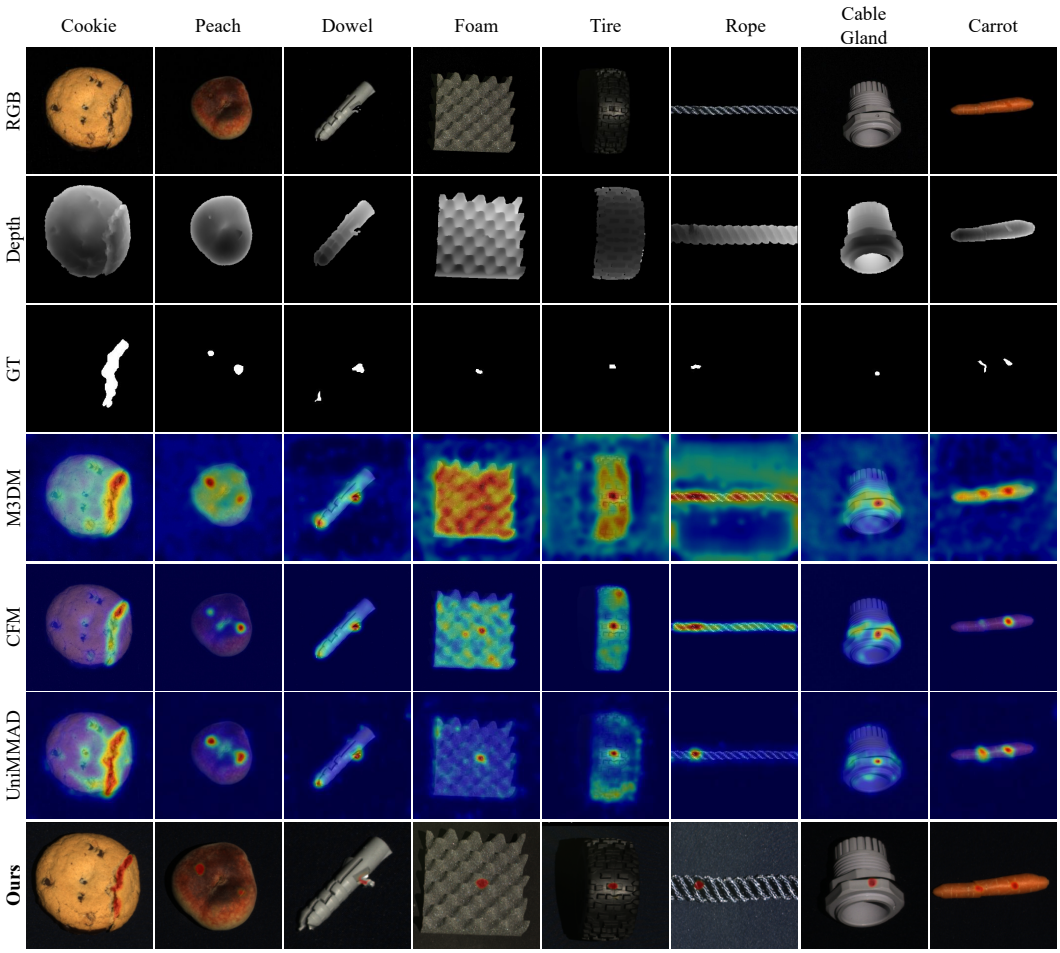


Figure 8: Qualitative comparison on MVTec-3D AD. Each example shows the input RGB image, depth observation, ground-truth anomaly mask, and anomaly maps produced by different methods. UniMRE produces concentrated responses around defective regions while preserving object structures and reducing irrelevant activations.

F.2 Qualitative Results on Eyecandies

Figure 9 provides qualitative comparisons on the Eyecandies dataset. Compared with MVTec-3D AD, Eyecandies contains more diverse object appearances and fine-grained defect patterns, making accurate localization more challenging. UniMRE maintains reliable anomaly responses across different categories, with predicted high-response regions generally corresponding to the actual defective areas. These results further demonstrate that the proposed class-conditional retrieval and dual-stream reasoning framework can adapt to heterogeneous objects and exploit both appearance and geometric information for anomaly localization.

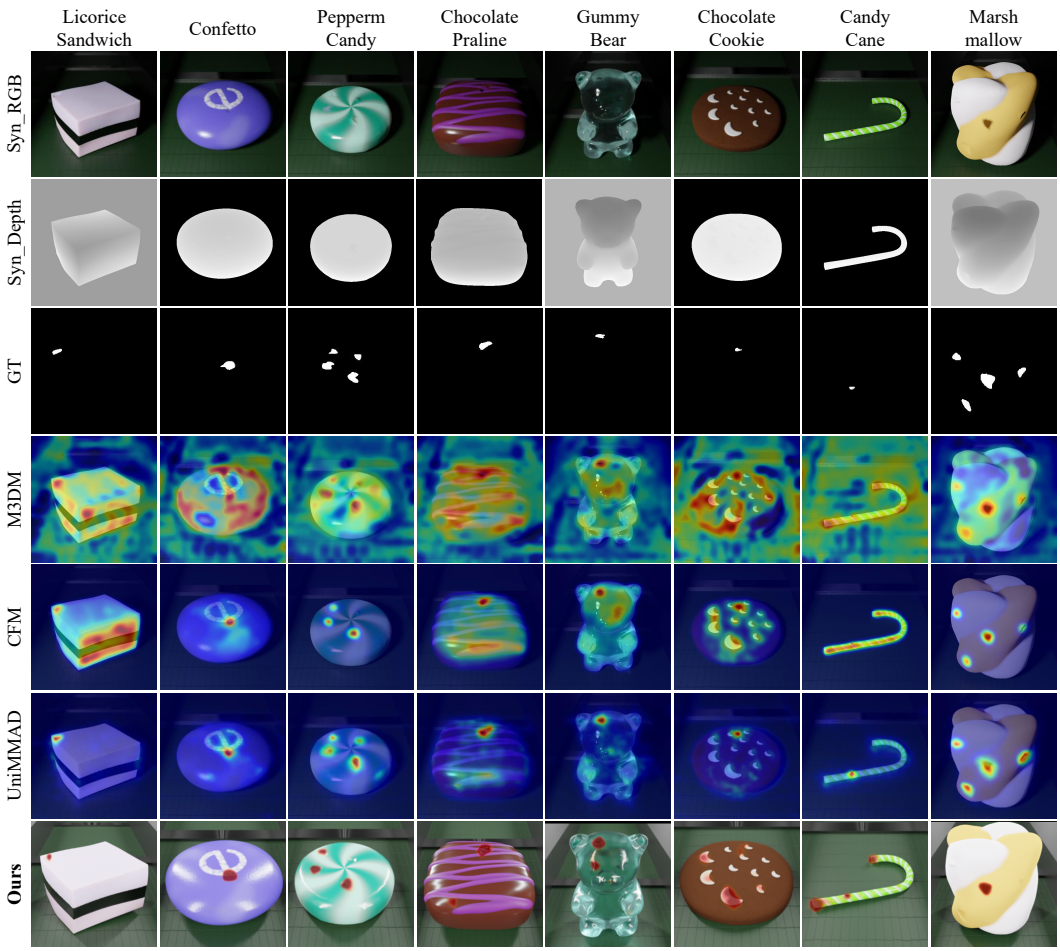


Figure 9: Qualitative comparison on Eyecandies. Each row presents the input observations, ground-truth anomaly mask, and anomaly maps generated by different approaches. UniMRE consistently identifies defective regions with compact anomaly responses across diverse object categories, demonstrating the robustness of the proposed multimodal anomaly reasoning framework.