An Efficient Cross-modal Feature Reconstruction Model for Multimodal Multi-class Anomaly Detection

Zhenghan Wang[1,2] **Zhihui Luo**[3] **Haitao Hu**[1,2] **Wengang Cheng**[1,2]*
[1]North China Electric Power University
[2]Engineering Research Center of Intelligent Computing for Complex Energy Systems
[3]Electric Power Research Institute of Xinjiang Power Grid Corporation

Abstract

Industrial anomaly detection is evolving from category-specific detectors toward unified models capable of handling multiple categories simultaneously. However, in the multi-class setting, greater inter-category diversity not only compels enhanced reconstruction capacity to model normal patterns but also exacerbates the "identical shortcut" problem, wherein anomalies are likewise well reconstructed. Moreover, existing unified frameworks are typically confined to single-modal (RGB) inputs and lack multimodal capability. To address these, we propose an Efficient Cross-modal Feature Reconstruction (ECFR) model for unified multi-class anomaly detection, which harnesses the inherent difficulty of cross-modal reconstruction to alleviate the shortcut issue and amplify anomaly discrimination. The framework is built on two core modules: an Adaptive Feature Interaction and Recalibration (AFIR) module and a Hybrid Attention Convolution (HAC) module. Through feature interaction, fusion, and reconstruction, our model achieves two key outcomes: learning normal patterns of multi-class objects while forcing reconstruction failures for anomalous inputs. Extensive experiments on the MVTec 3D-AD and Eyecandies datasets validate our approach. ECFR achieves the best average detection and localization performance among the compared non-memory-bank methods on both benchmarks, with 44.63 M parameters and 15.76 GFLOPs. Code is available at `https://github.com/wang1393/ECFR`.

1 Introduction

Industrial anomaly detection (IAD) aims to develop models that can pinpoint anomalous regions in products, a capability crucial for reducing losses and enhancing efficiency in smart manufacturing. Due to the scarcity of anomalous samples, IAD is typically framed as an unsupervised problem. As dominant approaches to unsupervised anomaly detection, reconstruction-based methods seek to model normal patterns through image (or feature) reconstruction. Since the model is trained solely on normal data, it reliably reconstructs normal inputs but not anomalous ones. Consequently, anomalies are identified by evaluating the discrepancy between the reconstructed output and the original input. Most existing IAD methods [45, 29, 44, 6, 41] train a separate model for each object category. This single-class paradigm, however, incurs significant overhead and hampers scalability as the number of categories grows. In response, recent multi-class methods [38, 46, 8, 37, 5, 32] aim to construct a single, unified model for multiple categories. Such a transition from category-specific to unified modeling is essential for improving the practical scalability of IAD systems.

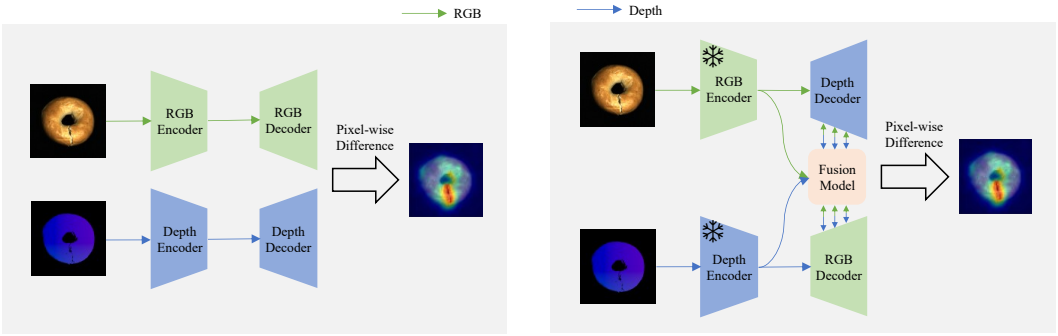
However, a critical challenge that persists and is even amplified in multi-class IAD is the "identical shortcut" problem inherent to reconstruction networks, where anomalous samples are also well reconstructed. Fundamentally, the network tends to learn an over-generalized mapping, which fails to

*Corresponding author: `wgcheng@ncepu.edu.cn`

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

discriminate between normal and anomalous patterns and consequently hinders anomaly identification. While this issue already hampers category-specific models, it worsens in unified multi-class settings. Driven by the greater inter-category diversity, the model is forced to develop enhanced reconstruction capacity to capture more complex normal patterns, which in turn exacerbates the "identical shortcut" problem. The phenomenon is also termed "over-generalization" in [33].

Another limitation of existing multi-class methods is their inability to process multimodal inputs, as they are predominantly designed for single-modality RGB data. This makes them unsuitable for real-world industrial scenarios involving multiple sensor types. Studies have shown that some anomalies are difficult to detect using RGB data alone [17]. While multimodal IAD methods have been developed by fusing diverse data sources, such as RGB images and 3D point clouds, they remain confined to the single-class setting. Consequently, there is a clear and unmet need for a multi-class anomaly detection framework that can effectively leverage and integrate multimodal inputs. However, current multimodal solutions largely constitute superficial adaptations of single-modal paradigms (Fig. 1a). Crucially, cross-modal reconstruction remains underexplored, primarily due to inter-modal heterogeneity: reconstructing one modality from another introduces fundamental challenges that inherently compromise reconstruction accuracy.



(a) Existing reconstruction-based multimodal methods (b) Our method

Figure 1: Different reconstruction-based multimodal anomaly detectors. Existing methods (a) independently reconstruct each modality using duplicated single-modal pipelines, making them susceptible to the identical shortcut problem. Our method (b) performs cross-modal reconstruction to exploit inter-modal heterogeneity, inherently suppressing shortcut and amplifying anomaly signals.

In this research, we take a different perspective: the inherent difficulty of cross-modal reconstruction can be advantageous for multi-class IAD. Specifically, cross-modal heterogeneity prevents learning identical shortcut in homogeneous data spaces and, when properly harnessed, induces discriminatively higher errors for anomalous samples. This mechanism amplifies anomaly detectability, as illustrated in Fig. 1b—transforming heterogeneity from a barrier into a detection asset.

Based on the above considerations, we propose an efficient cross-modal feature reconstruction framework for multi-class anomaly detection. Our framework is architected to fulfill dual objectives: (1) modeling the multimodal feature distributions of normal samples across classes, while (2) amplifying reconstruction failures for anomalous inputs. We design two core modules: the Adaptive Feature Interaction and Recalibration (AFIR), and the Hybrid Attention Convolution (HAC). Specifically, the AFIR module learns normal multimodal fusion patterns through joint modeling of intra-modal dependencies and inter-modal correlations, while deliberately forcing catastrophic propagation of anomalous features—systematically degrading cross-modal representations when single-modality outliers occur. The HAC module integrates Agent Attentions [14] with parallel multi-kernel depth-wise convolutions to holistically model global context and local structures in normal samples. When distributional deviations arise, the HAC-equipped decoder induces high reconstruction errors via targeted spatial-feature disintegration, thereby amplifying anomaly signals.

Our contributions:

- We address multi-class anomaly detection from a multimodal perspective. Through deliberate exploitation of cross-modal reconstruction difficulty, our approach fundamentally

alleviates the "identical shortcut" problem while establishing an anomaly amplification mechanism.

- We propose a simple but efficient cross-modal feature reconstruction framework. We devise two core modules: the Adaptive Feature Interaction and Recalibration (AFIR) module and the Hybrid Attention Convolution (HAC) module. Through feature interaction, fusion, and reconstruction, our framework achieves dual outcomes: learning normal patterns of multi-class objects while forcing reconstruction failures for anomalous inputs.
- ECFR delivers outstanding performance while maintaining high computational efficiency. This is achieved through several efficiency-oriented designs, including an efficient attention mechanism, an optimal stacking configuration in HAC module, and a lightweight yet powerful backbone. These components collectively contribute to the highest mean detection and localization scores among the evaluated methods without memory banks on two 3D-AD datasets, with remarkably low model parameters (44.63M) and computational complexity (15.76GFLOPs).

2 Related Work

We review three closely related fields: unsupervised image anomaly detection, multimodal industrial anomaly detection, and multi-class unsupervised anomaly detection.

Unsupervised Image Industrial Anomaly Detection: Existing mainstream Image IAD methods can be categorized into three types: 1) Embedding-based approaches [45, 29, 20, 24, 9] leverage networks pre-trained on ImageNet [26] to embed normal data into a compact feature space, identifying anomalies as outliers. However, the distribution gap between ImageNet and industrial images often leads to domain mismatch, thereby degrading detection performance. 2) Synthesis-based methods [44, 18, 27, 39, 28] actively construct pseudo-anomalies near the normal distribution, thereby transforming anomaly detection into a refined discriminative learning task. However, limited synthesis rules prevent these methods from exhaustively covering diverse real-world defects or accurately characterizing true anomaly distributions. 3) Reconstruction-based methods [10, 43, 22, 36, 13, 11] detect anomalies by measuring the discrepancy between inputs and their reconstructions. However, early pixel-level models often succumb to the "identical shortcut" driven by pixel fidelity. Although subsequent feature reconstruction incorporates semantic context, this issue persists. Reverse Distillation [7] (RD) effectively mitigates identity mapping through an asymmetric teacher-student architecture with a bottleneck design. Nevertheless, the student network still faces "over-generalization", where it inadvertently reconstructs anomalous features, thereby limiting detection performance.

Multimodal Industrial Anomaly Detection: MIAD typically outperforms single-modality approaches by leveraging heterogeneous inputs. M3DM [34] fuses RGB and point cloud features using multiple memory banks, while G^2SF [31] performs geometry-guided anomaly-score fusion over stored normal features. Such memory-bank methods [34, 30, 31] explicitly retain normal representations for retrieval-based matching. AST [25] employs an asymmetric teacher-student architecture that locates anomalies by comparing the outputs of both networks, yet uses depth only for background removal, discarding valuable structural details. EasyNet [4] avoids pre-trained models through reconstruction and entropy-based fusion, enhancing inference speed, but remains susceptible to shortcut learning. This limitation also extends to 3D-ADNAS [21], despite its use of NAS to optimize fusion strategies. CFMs [6] explores inter-modal correlations through cross-modal mapping, yet non-unique mappings lead to false positives in complex scenarios, and performance remains bounded by pretrained feature expressiveness. 3DSR [40] simulates depth anomalies via affine transformations to obtain suitable extractors, yet its effectiveness is constrained by synthetic anomaly limitations. These methods were originally developed for single-class anomaly detection. In contrast, our method detects anomalies across all object categories with a unified model.

Multi-class Unsupervised Anomaly Detection: Current multi-class unsupervised anomaly detection (MUAD) methods [3, 16] predominantly employ reconstruction architectures. UniAD [38] pioneered this domain using a Transformer encoder-decoder with neighbor mask attention and feature noise to mitigate shortcut learning, though not entirely. OmniAL [46] enhances normal pattern learning by injecting synthetic anomalies. However, its efficacy depends on artificial anomaly quality, which struggles to capture the diversity of real-world defects. More recently, UniNet [35] employs a ReContrast-style contrastive reconstruction scheme with an additional frozen teacher to discourage

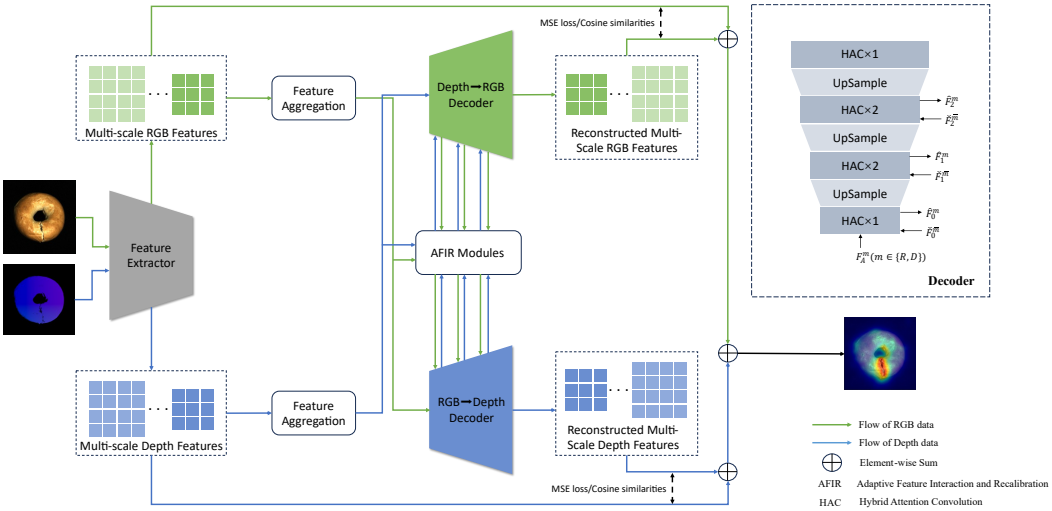


Figure 2: The overall framework of ECFR. The encoder extracts multi-scale RGB and depth features, which are cross-modally reconstructed via AFIR and HAC modules to detect anomalies.

shortcut learning, but its training remains unstable. While effective, these methods remain confined to RGB inputs, inheriting single-modality constraints. In comparison, our method achieves multi-class anomaly detection through a multimodal paradigm. By deliberately exploiting the inherent difficulty of cross-modal feature reconstruction, it fundamentally alleviates the "identical shortcut" problem.

3 Method

3.1 Problem Definition

The proposed method targets the multi-class MIAD setting, in contrast to traditional methods that handle only a single class. We formulate the problem as follows. Let $\mathcal{T} = \{x_1, x_2, \ldots, x_n\}$ be a set of training samples, where each sample x_i is a normal pair consisting of an RGB image $x_i^{\mathrm{R}} \in \mathbb{R}^{H\times W\times C}$ and a depth image $x_i^{\mathrm{D}} \in \mathbb{R}^{H\times W\times C}$, i.e., $x_i =< x_i^{\mathrm{R}}, x_i^{\mathrm{D}} >$. In the single-class MIAD case, $\mathcal{T}$ is drawn from one specific class c_j, $c_j \in \mathcal{C}$, where $\mathcal{C}$ is the set of all object classes. In the multi-class MIAD case, $\mathcal{T}$ contains normal samples from all classes in $\mathcal{C}$. At test time, given a normal or anomalous sample from any class c_j, the IAD model is required to predict the presence of anomalies and, upon detection, localize the specific anomalous region.

3.2 Model Overview

Figure 2 illustrates the overall architecture of our method, which employs a pyramid-structured auto-encoder to reconstruct multi-scale features. Following the reverse-distillation paradigm, we use a pair of frozen pre-trained feature extractors with shared parameters to obtain feature maps from the RGB and depth inputs at three scales ($K = 3$), i.e., $\{< F_i^{\mathrm{R}}, F_i^{\mathrm{D}} >\}_{i=1}^{K} \in \mathbb{R}^{H_i\times W_i\times C_i}$. The Half-FPN bottleneck then combines these features to output aggregated representations $< F_{\mathrm{A}}^{\mathrm{R}}, F_{\mathrm{A}}^{\mathrm{D}} >\in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 8C}$, which are subsequently used for cross-modal feature reconstruction to generate the reconstructed multi-scale features, i.e., $\{< \hat{F}_i^{\mathrm{R}}, \hat{F}_i^{\mathrm{D}} >\}_{i=1}^{K} \in \mathbb{R}^{H_i\times W_i\times C_i}$. The cross-modal reconstruction is built on two core modules: Adaptive Feature Interaction and Recalibration (AFIR) and Hybrid Attention Convolution (HAC). AFIR contains an Intra–inter Modal Interaction (IMI) block and an Adaptive Channel-wise Recalibration (ACR) block. The IMI block integrates RGB and depth features to enhance complementary representations and, in the presence of anomalies, encourages their propagation both within and across modalities. The ACR block reweights channels to emphasize informative features. On the other hand, HAC is designed to capture both global feature distributions and local details. For notational simplicity, we use $m(m \in \{\mathrm{R}, \mathrm{D}\})$ to denote one modality and $\bar{m}$ to denote the other. Each reconstruction branch starts from the aggregated features F_{A}^{m} and progressively reconstructs features of modality $\bar{m}$. During decoding, AFIR uses

complementary information from modality $\bar{m}$ to modulate only the global branch of HAC and guide reconstruction. Each modality-specific decoder has four stages, and each stage is a stack of HAC blocks. The outputs of the last three stages correspond to the three scales of feature maps from the encoder. To balance efficiency and accuracy, we apply AFIR only in the first three decoding stages, where feature maps retain lower spatial resolution and thus incur less computational overhead. At inference time, we compute pixel-wise similarities between $\{< F_i^{\mathrm{R}}, F_i^{\mathrm{D}} >\}_{i=1}^{K}$ and $\{< \hat{F}_i^{\mathrm{R}}, \hat{F}_i^{\mathrm{D}} >\}_{i=1}^{K}$ for anomaly detection.

3.3 Adaptive Feature Interaction and Recalibration

RGB and depth features are inherently different. Directly reconstructing one modality from the other often leads to poor reconstruction. To reduce this gap, we introduce the AFIR module, which has two components: intra–inter modal interaction and adaptive channel-wise recalibration as shown in Figure 3a.

Intra-inter Modal Interaction (IMI). Relevant studies [23] show that inter-modal relationships are crucial when modalities share significant predictive information, whereas intra-modality variations are more effective when cross-modal information is sparse or absent. To make the IMI module effective in both regimes, we explicitly model both inter- and intra-modal dependencies. Specifically, at the i-th stage, the input feature maps $\hat{F}_{i-1}^{m}$ and $\hat{F}_{i-1}^{\bar{m}}$ are first flattened and reshaped into vectors, resulting in two matrices of size $\mathbb{R}^{N\times C}$, where $N = H_i \times W_i$. These matrices are then concatenated and projected via a 1×1 convolution to generate the Query (Q_i), Key (K_i), and Value (V_i) vectors, each of size $\mathbb{R}^{2N\times C'}$, where $C' = C/2$. The conventional approach of computing an attention map by multiplying Q_i and K_i leads to a high computational complexity of $O(N^2C)$. To address this challenge, we adopt Agent Attention [14] instead. Agent attention employs a small group of agent tokens to aggregate and broadcast global information. It integrates Softmax and linear attention, combining high expressiveness with low computational cost. Concretely, the agents $A_i \in \mathbb{R}^{n\times C'}$ are obtained by pooling the Query Q_i, where $n \ll N$. Thus, the computational complexity is reduced to $O(NnC)$. Under this design, the interacted feature $\widetilde{F}_i$ is computed as:

$$\widetilde{F}_i = \mathrm{Conv}\left(\mathrm{softmax}(Q_i A_i^{\top})\ \mathrm{softmax}(A_i K_i^{\top}) V_i\right). \tag{1}$$

Finally, we split the unified RGB–D feature representation $\widetilde{F}_i$ into its RGB and Depth components and reshape them back to their original spatial dimensions, i.e., $\widetilde{F}_i^{m} \in \mathbb{R}^{H_i\times W_i\times C_i}$ and $\widetilde{F}_i^{\bar{m}} \in \mathbb{R}^{H_i\times W_i\times C_i}$.

Adaptive Channel-wise Recalibration (ACR). Feature recalibration focuses on adaptively weighting the channels in each modality to highlight the informative channels with high discriminability. Specifically, at the i-th stage, $\widetilde{F}_i^{m}$ is passed through 1×1 convolutions, global average pooling and a sigmoid activation to produce the channel-wise weight matrices W_i^{m}:

$$W_i^{m} = \mathrm{Sigmoid}\left(\mathrm{GAP}\left(\mathrm{Conv}\left(\widetilde{F}_i^{m}\right)\right)\right). \tag{2}$$

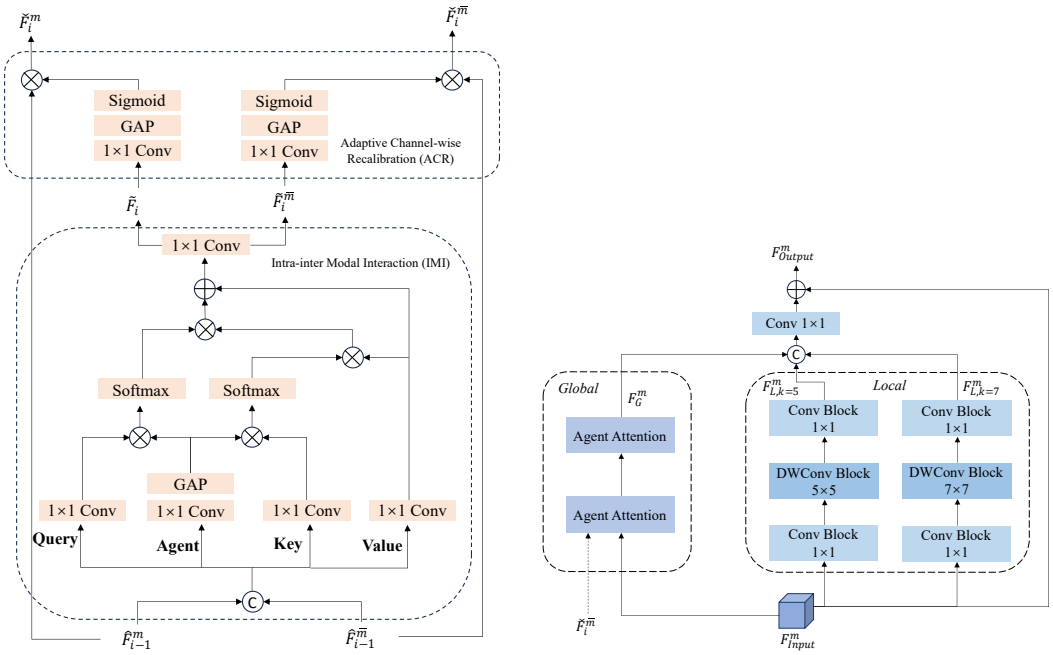
We then multiply W_i^{m} with $\hat{F}_{i-1}^{m}$ to obtain the recalibrated features $\check{F}_i^{m}$:

$$\check{F}_i^{m} = W_i^{m}\ \hat{F}_{i-1}^{m}. \tag{3}$$

3.4 Hybrid Attention Convolution Module

In feature reconstruction, capturing normal patterns within complex multi-class scenarios requires jointly modeling global contextual dependencies and fine-grained local features. Such integration establishes a comprehensive pattern representation and stable inter-modal correlations. To this end, we propose the HAC module (Figure 3b), which combines cascaded Agent Attention for global modeling with parallel multi-kernel depth-wise separable convolutions for local information capture. In the global branch, the input feature map $F_{\mathrm{Input}}^{m} \in \mathbb{R}^{H_i\times W_i\times C_i}$ passes through two Agent Attention blocks to produce the global feature F_{G}^{m}:

$$F_{\mathrm{G}}^{m} = \mathrm{Attn}_2\left(\mathrm{Attn}_1\left(F_{\mathrm{Input}}^{m}\right)\right). \tag{4}$$



(a) Overview of the proposed AFIR.

(b) Overview of the proposed HAC.

Figure 3: Architectural details of the two core modules. AFIR (a) jointly models intra- and inter-modal dependencies to bridge the cross-modal gap while propagating anomalous signals across modalities. HAC (b) combines Agent Attention with multi-kernel depth-wise convolutions to capture both global context and local structures for robust reconstruction.

The agents A are obtained in the same way as in AFIR. In the first three decoder stages, $\check{F}_i^{\bar{m}}$ is also fed into the first HAC block to assist global modeling, and the global feature F_{G}^m is computed as:

$$F_{\mathrm{G}}^m = \mathrm{Attn}_2\left(\mathrm{Attn}_1\left(F_{\mathrm{Input}}^m,\, \check{F}_i^{\bar{m}}\right)\right), \tag{5}$$

where $\mathrm{Attn}_1(\cdot)$ acts as cross-attention: $\check{F}_i^{\bar{m}}$ is the query Q, and F_{Input}^m provides the Key K and Value V. In the local branch, F_{Input}^m is processed by two parallel depth-wise Conv blocks. Each block consists of a 1×1 convolution, a $k \times k$ depth-wise convolution, and another 1×1 convolution.

$$F_L^m = \mathrm{ConvB}_{1\times1}\Big(\mathrm{DWConvB}_{k\times k}\big(\mathrm{ConvB}_{1\times1}(F_{\mathrm{Input}}^m)\big)\Big), \tag{6}$$

where k is the kernel size for the depth-wise convolution. Our experiments show the best performance when $k = 5$ and $k = 7$. Each convolutional block comprises a convolution layer, an instance normalization layer, and a SiLU activation. Local and global features are concatenated along the channel dimension, and a final 1×1 convolution plus a residual connection with F_{Input}^m produce the output F_{Output}^m:

$$F_{\mathrm{Output}}^m = \mathrm{Conv}_{1\times1}\Big(\mathrm{Concat}\big(F_{\mathrm{G}}^m,\, F_{\mathrm{L},\,k=5}^m,\, F_{\mathrm{L},\,k=7}^m\big)\Big) + F_{\mathrm{Input}}^m. \tag{7}$$

3.5 Loss Function

Through cross-modal feature reconstruction, we obtain the multi-scale feature maps for the reconstructed modalities $\hat{F}_i^m$. The final loss function of each branch is the sum of the mean squared errors (MSE) between $\hat{F}_i^m$ and the corresponding original feature maps F_i^m at three scales.

$$\mathcal{L}^m = \sum_{i=1}^{K} \frac{\left\| F_i^m - \hat{F}_i^m \right\|_2^2}{H_i W_i}. \tag{8}$$

The total loss is

$$\mathcal{L} = \mathcal{L}^m + \mathcal{L}^{\bar{m}}, \tag{9}$$

where $\mathcal{L}$ denotes the overall loss. After training on the dataset that contains only normal samples, the learned network parameters capture the normal patterns.

3.6 Inference and Scoring Function

During inference, we first extract the multi-scale feature maps F_i^m from the test samples. We then perform cross-modal reconstruction to obtain the predicted features $\hat{F}_i^m$, which represent the expected normal response of modality m under cross-modal consistency. Next, we compute the cosine similarity between F_i^m and $\hat{F}_i^m$ and use it to construct an anomaly score map: regions with high similarity (good cross-modal consistency) receive low anomaly scores, whereas regions with low similarity, indicating broken consistency, are assigned higher anomaly scores. The anomaly score map for modality m is computed as:

$$S^m = \frac{1}{K}\sum_{i=1}^{K} \text{Upsample}\left(1 - \frac{\langle F_i^m,\, \hat{F}_i^m \rangle}{\|F_i^m\|_2 \cdot \|\hat{F}_i^m\|_2}\right). \tag{10}$$

The final score map is

$$S = S^m + S^{\bar{m}}, \tag{11}$$

where $S \in \mathbb{R}^{H\times W\times C}$ is the overall anomaly score map. For each feature level i, we upsample the corresponding score map to the original image size and then average the scores over all levels to obtain the score map for each modality. Finally, we sum the score maps from the two modalities to produce the final anomaly score map.

4 Experiments

4.1 Experimental Setup

Datasets. To validate our method, we conduct extensive experiments on two multimodal benchmarks: MVTec 3D-AD [1] and Eyecandies [2]. More details on both datasets are provided in Appendix E.1. Beyond these, we further validate ECFR on additional multimodal datasets (Appendix E.7).

Metrics. Following prior work [34, 4, 40], we adopt I-AUROC for image-level detection, and P-AUROC and AUPRO for localization. Definitions are provided in Appendix E.2.

Implementation. All inputs are resized to 256×256. A frozen pre-trained ResNet-34 [15] with shared parameters serves as the feature extractor. The agent token count is set to [49, 49, 16, 9] across decoder stages (Appendix E.9). Training and inference details are provided in Appendix E.3.

4.2 Comparison with SOTAs on Different AD datasets

Following the standard multi-class protocol of UniAD [38], all baselines are retrained by merging all training categories into a single unified class, using their released code without modification. Our experimental evaluation covers three complementary comparisons. (1) We adapt four representative state-of-the-art multimodal anomaly detection methods to the multi-class setting: the teacher–student model AST [25], the reconstruction-based methods EasyNet [4] and 3DSR [40], and the cross-modal mapping method CFMs [6]. We separately present a comparison of ECFR with the memory-bank methods M3DM [34] and G^{2}SF [31] in Appendix E.4, as their additional feature-bank storage requirements could place them at an unfair disadvantage in the main comparison. (2) We compare ECFR against these methods under the single-class multimodal setting to verify its competitiveness (Appendix E.6). (3) We also include comparisons with multi-class single-modal methods for completeness (Appendix E.8). For comparison (1), results for MVTec 3D-AD are reported in Table 1, Eyecandies in Table 2, and efficiency comparisons in Table 3.

Anomaly Detection. Our method achieves state-of-the-art image-level anomaly detection performance (I-AUROC) on both MVTec 3D-AD and Eyecandies within the primary comparison. Compared with the reconstruction-based methods 3DSR and EasyNet, our approach improves I-AUROC by 4.08% and 13.16% on MVTec 3D-AD, and by 12.86% and 1.49% on Eyecandies, respectively.

Table 1: Quantitative results with SOTA methods on the MVTec 3D-AD benchmark.

Metric	Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
	AST	82.85	73.56	92.45	89.60	84.99	74.94	94.01	**94.81**	93.89	52.69	83.38
	EasyNet	91.16	69.62	83.75	75.34	88.49	88.93	68.03	61.61	89.85	70.11	78.69
I-AUC	3DSR	97.31	68.25	**93.43**	86.75	87.86	**92.00**	94.21	90.61	**97.75**	69.56	87.77
	CFMs	94.36	83.14	93.08	96.12	91.41	82.73	91.55	91.02	92.39	86.71	90.25
	Ours	**97.31**	**94.75**	89.56	**96.85**	**91.86**	83.81	**97.57**	85.67	94.16	**86.94**	**91.85**
	AST	97.43	94.45	**99.13**	94.81	91.69	96.02	99.46	**99.73**	97.26	93.64	96.36
	EasyNet	93.10	95.47	97.82	89.99	64.78	**99.87**	97.69	96.77	89.93	98.67	92.41
P-AUC	3DSR	95.95	95.16	97.49	89.97	94.67	97.86	95.57	95.93	97.46	95.27	95.53
	CFMs	99.05	**97.72**	98.83	96.98	97.69	98.23	99.13	98.81	99.18	99.02	98.46
	Ours	**99.35**	97.30	98.99	**98.47**	**98.90**	95.50	**99.60**	99.15	**99.64**	**99.12**	**98.60**
	AST	94.02	81.92	**97.69**	91.99	87.59	85.93	97.73	**98.14**	90.92	72.76	89.87
	EasyNet	80.66	85.76	95.82	69.49	70.54	82.57	88.52	91.12	90.11	89.61	84.42
PRO	3DSR	92.93	83.59	96.31	76.43	88.78	91.82	95.20	94.03	97.11	84.20	90.04
	CFMs	96.85	**94.67**	96.51	93.13	91.19	**93.07**	96.93	97.26	95.83	96.06	95.15
	Ours	**97.62**	92.33	96.92	**95.72**	**95.50**	85.78	**98.62**	97.52	**97.41**	**96.23**	**95.37**

Table 2: Quantitative results with SOTA methods on the Eyecandies benchmark.

Metric	Method	Candy	Choc-C	Choc-P	Conf	Gummy	Truffle	Licorice	Lolli	Marsh	Pepper	Mean
	AST	49.76	64.96	60.64	**96.48**	82.69	47.52	70.08	81.80	97.92	85.76	73.76
	EasyNet	64.64	91.84	**89.28**	96.32	67.30	**78.40**	85.28	89.15	89.28	**99.96**	85.15
I-AUC	3DSR	57.76	74.56	69.60	85.04	81.44	62.72	71.36	66.18	81.12	88.00	73.78
	CFMs	46.63	89.33	87.01	89.00	80.41	75.49	**87.44**	82.45	97.88	84.34	82.00
	Ours	**65.12**	**93.60**	85.28	92.80	**89.42**	68.36	84.80	**89.71**	**99.68**	97.60	**86.64**
	AST	96.96	86.86	85.93	89.26	45.48	77.25	61.84	95.62	93.48	90.58	82.33
	EasyNet	95.62	93.24	**95.46**	94.29	91.98	92.36	**97.99**	96.67	96.76	70.98	92.54
P-AUC	3DSR	87.90	92.48	90.20	93.19	89.23	87.88	88.55	84.00	83.90	91.66	88.90
	CFMs	**97.71**	98.26	95.20	98.65	94.73	**93.52**	96.18	98.19	99.25	97.38	96.91
	Ours	97.25	**98.92**	93.99	**99.47**	**94.96**	92.47	97.26	**98.21**	**99.49**	**99.32**	**97.13**
	AST	90.19	59.78	54.63	64.06	35.25	37.13	34.95	86.44	78.47	69.53	61.04
	EasyNet	87.79	84.08	**83.58**	87.94	81.31	66.39	90.35	84.40	89.26	**97.43**	85.25
PRO	3DSR	70.16	75.44	75.66	81.51	66.94	59.67	63.29	46.94	68.42	79.05	68.71
	CFMs	**92.18**	89.61	80.52	95.70	86.87	**75.86**	77.50	90.12	93.41	92.80	87.46
	Ours	89.98	**95.16**	81.79	**98.07**	**88.45**	70.27	**91.27**	**92.93**	**96.85**	97.28	**90.21**

These single-class methods show clear performance degradation in the multi-class setting, indicating that they struggle to learn shared normal patterns across categories. Memory-bank methods, by comparison, explicitly retain normal features across categories and may thus be less sensitive to inter-category diversity. ECFR remains competitive in image-level detection on MVTec 3D-AD, achieving 91.85 I-AUROC compared with 90.25 for G^2SF and 93.33 for M3DM.

Anomaly Localization. Our method also achieves the highest mean pixel-level localization performance (P-AUROC and AUPRO) within the primary comparison on both datasets (Tables 1 and 2). Although 3DSR and EasyNet are pixel-level reconstruction and detection models, which should favour localization, our method still surpasses them by 3.07%/5.33% and 6.19%/10.95% on MVTec 3D-AD (P-AUROC/AUPRO), respectively. On Eyecandies, ECFR improves P-AUROC and AUPRO over CFMs by 0.22% and 2.75%, respectively. This confirms the effectiveness of our localization strategy. In the separate memory-bank comparison, ECFR achieves slightly lower localization performance than both methods [34, 31] on MVTec 3D-AD. Appendix E.4 details these differences.

Qualitative Results. We further provide qualitative results on MVTec 3D-AD and Eyecandies to illustrate the segmentation quality. Figure 4 visualizes the performance of ECFR on MVTec 3D-AD, demonstrating that our method possesses more precise anomaly segmentation capabilities. Additional visualizations on Eyecandies are given in Appendix E.5.

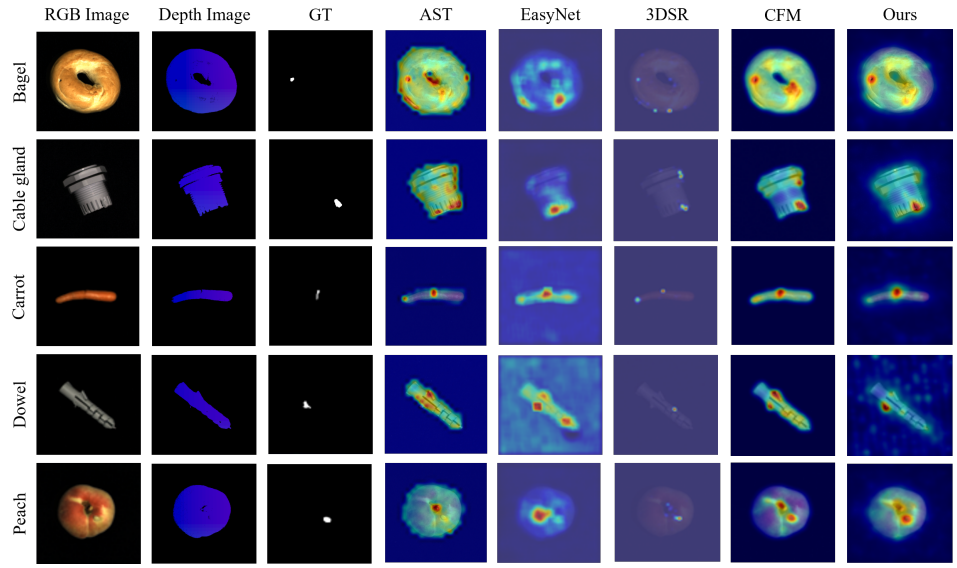


Figure 4: Qualitative visualization for pixel-level anomaly segmentation on MVTec 3D-AD.

Table 3: Efficiency comparison of SOTA methods.

Method	Params(M)	FLOPs(G)	FPS
AST	147.271	96.068	4.743
3DSR	60.754	252.606	20.487
EasyNet	140.599	465.023	20.522
CFMs	113.168	440.836	14.513
Ours	**44.627**	**15.762**	**21.356**

Table 4: Ablation studies on each component.

Cross Rec.	AFIR	Loc.	I-AUC	P-AUC	PRO
-	-	√	87.16	97.37	92.08
√	-	√	91.18	98.26	93.81
-	√	√	89.83	98.53	95.18
√	√	-	89.59	98.43	93.37
√	√	√	**91.85**	**98.60**	**95.37**

4.3 Efficiency Evaluation

Computational efficiency is critical for deploying industrial anomaly detection systems. We compare our method with four SOTA baselines in terms of parameter count, floating-point operations (FLOPs), and inference speed (FPS), as summarized in Table 3. All experiments are conducted on an NVIDIA A800 80GB GPU using PyTorch 1.13. FPS is averaged over the MVTec 3D-AD test set. Inference time covers the full pipeline—from input pre-processing to anomaly score computation—with GPU thread synchronization enforced before timing. Memory usage is approximated by the total number of model parameters. Our method is more efficient than all baselines, with significantly fewer parameters and lower computation cost. Compared with CFMs, our model uses $2.5\times$ fewer parameters, $28\times$ fewer FLOPs, and achieves 47.2% higher FPS. AST is inefficient due to its two feed-forward networks and the Normalizing Flow module. 3DSR and EasyNet report higher FPS but rely on dual encoder-decoder architectures, which increases complexity. CFMs uses two large pretrained backbones to extract high-resolution features, which are mapped by an MLP from very long vectors, reducing efficiency. In contrast, we adopt a parameter-sharing ResNet-34 as a compact yet expressive backbone. Our Half-FPN aggregates multi-scale features while avoiding oversized decoder inputs. Additionally, we replace standard attention with the more efficient Agent Attention and reduce decoder-side attention computation, which further improves speed and scalability.

4.4 Ablation and Analysis

Ablations of Each Component. We perform ablation studies on the unified MVTec 3D-AD setting to evaluate the contribution of each proposed module, as reported in Table 4. "Cross Rec." denotes cross-modal reconstruction, "AFIR" denotes adaptive feature interaction and recalibration, and "Loc." denotes the local branch of the HAC module. Disabling AFIR leaves each decoder to rely solely on the aggregated features of its input modality, without guidance from the other modality to the global HAC

Table 5: Ablation studies on different numbers of HAC.

HAC Number	I-AUC	P-AUC	PRO
[1, 1, 1, 1]	90.39	98.45	94.86
[1, 2, 2, 1]	**91.85**	**98.60**	**95.37**
[2, 2, 2, 2]	90.25	98.52	95.08

Table 6: Ablation studies on HAC's designs.

Kernel Size	I-AUC	P-AUC	PRO
k = 3, 5	90.77	98.57	95.14
k = 5, 7	**91.85**	98.60	**95.37**
k = 3, 5, 7	91.25	**98.67**	95.32

Table 7: Contribution of individual modalities on MVTec 3D-AD.

Input	I-AUROC	P-AUROC	AUPRO
RGB only	87.75	**98.89**	**95.76**
Depth only	83.11	95.51	84.75
RGB + Depth	**91.85**	98.60	95.37

branch. In this setting, cross-modal reconstruction improves I-AUC, P-AUC, and PRO over single-modal reconstruction by 4.02, 0.89, and 1.73 percentage points, respectively. These gains demonstrate that cross-modal reconstruction itself improves anomaly discrimination even without incorporating additional information from the other modality through AFIR. Appendix C corroborates this finding through reconstruction-error analysis. In addition, AFIR consistently improves performance with or without cross-modal reconstruction, indicating that it effectively facilitates feature reconstruction. Finally, enabling the local branch in HAC brings a further performance gain, confirming the usefulness of local modeling.

Effectiveness comparison of different numbers of HAC. As shown in Table 5, we study how different HAC configurations affect performance while keeping the backbone fixed. To keep the model lightweight, we evaluate three configurations with small numbers of HAC modules per decoder stage: [1, 1, 1, 1], [1, 2, 2, 1], and [2, 2, 2, 2]. The results show that the configuration [1, 2, 2, 1] yields the best overall performance. Therefore, we adopt [1, 2, 2, 1] as the default HAC configuration in our method.

Effectiveness comparison of different HAC designs. The HAC module design primarily involves three factors: the number of cascaded Agent Attention layers, the configuration of the parallel multi-kernel convolutions, and the choice of depth-wise convolution kernel sizes. As adding more Agent Attention layers significantly increases computational cost, we employ the minimal cascade of two Agent Attention blocks for a lightweight design. Thus, the remaining design choices focus on the latter two factors, as summarized in Table 6. We evaluate three depth-wise convolution configurations with kernel sizes $k = 3$, $k = 5$ and $k = 7$. The results show that $k = 5$ and $k = 7$ are more effective for capturing local structures and modeling local dependencies. Therefore, we adopt parallel depth-wise convolution blocks with kernel sizes $k = 5$ and $k = 7$ in our HAC module.

Contribution of Individual Modalities. We compare RGB-only, depth-only, and RGB–depth inputs with the architecture and settings unchanged. For each single-modality configuration, the same input is supplied to both branches during training and inference. As shown in Table 7, adding depth to RGB improves mean I-AUROC by 4.10 percentage points, while P-AUROC and AUPRO decrease by 0.29 and 0.39 points, respectively. These results demonstrate the benefit of complementary modalities for image-level detection, with small changes in average localization performance.

5 Conclusion

This paper presents Efficient Cross-modal Feature Reconstruction (ECFR) for multi-class multimodal anomaly detection. By leveraging cross-modal reconstruction, ECFR fundamentally alleviates the "identical shortcut" problem. Central to this framework are two synergistic modules: the AFIR alleviates modal discrepancies, while the HAC captures discriminative normal patterns. Together, they enable robust modeling of multi-class normality while forcing catastrophic reconstruction failures for anomalies. Extensive qualitative and quantitative experiments demonstrate that the proposed method achieves the best average detection and localization performance among the compared non-memory-bank methods across two multi-class multimodal benchmarks, while maintaining low model parameters and computational complexity.

Acknowledgments and Disclosure of Funding

We thank all area chairs and anonymous reviewers for their time and valuable feedback. This work received no external funding, and the authors declare no competing interests.

References

[1] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The MVTec 3D-AD dataset for unsupervised 3d anomaly detection and localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 202–213. INSTICC, SciTePress, 2022. ISBN 978-989-758-555-5. doi: 10.5220/0010865000003124.

[2] Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *Proceedings of the Asian Conference on Computer Vision*, pages 3586–3602, 2022.

[3] Qiyu Chen, Huiyuan Luo, Haiming Yao, Wei Luo, Zhen Qu, Chengkan Lv, and Zhengtao Zhang. Center-aware residual anomaly synthesis for multiclass industrial anomaly detection. *IEEE Transactions on Industrial Informatics*, 21(9):7276–7286, 2025. doi: 10.1109/TII.2025.3575122.

[4] Ruitao Chen, Guoyang Xie, Jiaqi Liu, Jinbao Wang, Ziqi Luo, Jinfan Wang, and Feng Zheng. Easynet: An easy network for 3d industrial anomaly detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7038–7046, 2023.

[5] Jiayi Cheng, Can Gao, Jie Zhou, Jiajun Wen, Tao Dai, and Jinbao Wang. MC3D-AD: A unified geometry-aware reconstruction model for multi-category 3d anomaly detection. In James Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 837–845. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/94. URL `https://doi.org/10.24963/ijcai.2025/94`. Main Track.

[6] Alex Costanzino, Pierluigi Zama Ramirez, Giuseppe Lisanti, and Luigi Di Stefano. Multimodal industrial anomaly detection by crossmodal feature mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17234–17243, 2024.

[7] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9737–9746, 2022.

[8] Hanqiu Deng and Xingyu Li. Structural teacher-student normality learning for multi-class anomaly detection and localization. *arXiv preprint arXiv:2402.17091*, 2024.

[9] André Luiz Vieira e Silva, Francisco Simões, Danny Kowerko, Tobias Schlosser, Felipe Battisti, and Veronica Teichrieb. Attention modules improve image-level anomaly detection for industrial inspection: A DifferNet case study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8246–8255, January 2024.

[10] Lei Fan, Yiwen Ding, Maurice Pagnucco, and Yang Song. Patch-wise augmentation for anomaly detection and localization. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5425–5429. IEEE, 2024.

[11] Qingqing Fang, Qinliang Su, Wenxi Lv, Wenchao Xu, and Jianxing Yu. Boosting fine-grained visual anomaly detection with coarse-knowledge-aware adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16532–16540, 2025.

[12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[13] Matic Fučka, Vitjan Zavrtanik, and Danijel Skočaj. Transfusion–a transparency-based diffusion model for anomaly detection. In *European conference on computer vision*, pages 91–108. Springer, 2024.

[14] Dongchen Han, Tianzhu Ye, Yizeng Han, Zhuofan Xia, Siyuan Pan, Pengfei Wan, Shiji Song, and Gao Huang. Agent attention: On the integration of softmax and linear attention. In *European conference on computer vision*, pages 124–140. Springer, 2024.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[16] Liren He, Zhengkai Jiang, Jinlong Peng, Wenbing Zhu, Liang Liu, Qiangang Du, Xiaobin Hu, Mingmin Chi, Yabiao Wang, and Chengjie Wang. Learning unified reference representation for unsupervised multi-class anomaly detection. In *European Conference on Computer Vision*, pages 216–232. Springer, 2024.

[17] Eliahu Horwitz and Yedid Hoshen. Back to the feature: classical 3d features are (almost) all you need for 3d anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2968–2977. IEEE, 2023.

[18] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021.

[19] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9984–9993, 2025.

[20] Xinyue Liu, Jianyuan Wang, Biao Leng, and Shuo Zhang. Dual-modeling decouple distillation for unsupervised anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pages 5035–5044, 2024.

[21] 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, pages 12273–12281, 2025.

[22] Wei Luo, Haiming Yao, Wenyong Yu, and Zhengyong Li. AMI-Net: Adaptive mask inpainting network for industrial anomaly detection and localization. *IEEE Transactions on Automation Science and Engineering*, 22:1591–1605, 2025. doi: 10.1109/TASE.2024.3368142.

[23] Divyam Madaan, Taro Makino, Sumit Chopra, and Kyunghyun Cho. Jointly modeling inter- & intra-modality dependencies for multi-modal learning. *Advances in Neural Information Processing Systems*, 37:116084–116105, 2024.

[24] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1088–1097, 2022.

[25] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2592–2602, 2023.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[27] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, pages 474–489. Springer, 2022.

[28] Alessandro Simoni and Francesco Pelosin. Bounding box-guided diffusion for synthesizing industrial images and segmentation maps. In *Synthetic Data for Computer Vision Workshop@ CVPR 2025*, 2025.

[29] Luc PJ Sträter, Mohammadreza Salehi, Efstratios Gavves, Cees GM Snoek, and Yuki M Asano. Generalad: Anomaly detection across domains by attending to distorted features. In *European conference on computer vision*, pages 448–465. Springer, 2024.

[30] Wenbo Sui, Daniel Lichau, Josselin Lefèvre, and Harold Phelippeau. Incomplete multimodal industrial anomaly detection via cross-modal distillation. *Information Fusion*, 126:103572, 2026. ISSN 1566-2535. doi: 10.1016/j.inffus.2025.103572. URL https://www.sciencedirect.com/science/article/pii/S156625352500644X.

[31] Chengyu Tao, Xuanming Cao, and Juan Du. G^2SF: Geometry-guided score fusion for multimodal industrial anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20551–20560, 2025.

[32] Jiaxiang Wang, Haote Xu, Xiaolu Chen, Haodi Xu, Yue Huang, Xinghao Ding, and Xiaotong Tu. Exploiting point-language models with dual-prompts for 3d anomaly detection. *Expert Systems with Applications*, 298:129758, 2026. ISSN 0957-4174. doi: 10.1016/j.eswa.2025.129758. URL https://www.sciencedirect.com/science/article/pii/S0957417425033731.

[33] Xiaolei Wang, Xiaoyang Wang, Huihui Bai, Eng Gee Lim, and Jimin Xiao. Cnc: Cross-modal normality constraint for unsupervised multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7943–7951, 2025.

[34] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8041, 2023.

[35] Shun Wei, Jielin Jiang, and Xiaolong Xu. Uninet: A contrastive learning-guided unified framework with feature selection for anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9994–10003, 2025.

[36] Hang Yao, Ming Liu, Zhicun Yin, Zifei Yan, Xiaopeng Hong, and Wangmeng Zuo. Glad: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024.

[37] Xincheng Yao, Ruoqi Li, Zefeng Qian, Lu Wang, and Chongyang Zhang. Hierarchical gaussian mixture normalizing flow modeling for unified anomaly detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 92–108, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73411-3.

[38] 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.

[39] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021.

[40] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Cheating depth: Enhancing 3d surface anomaly detection via depth simulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2164–2172, 2024.

[41] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Keep dræming: discriminative 3d anomaly detection through anomaly simulation. *Pattern Recognition Letters*, 181:113–119, 2024.

[42] Jiangning Zhang, Haoyang He, Zhenye Gan, Qingdong He, Yuxuan Cai, Zhucun Xue, Yabiao Wang, Chengjie Wang, Lei Xie, and Yong Liu. A comprehensive library for benchmarking multi-class visual anomaly detection. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3574–3583, 2025. doi: 10.1109/ICCVW69036.2025.00373.

[43] Ruifan Zhang, Hao Wang, Mingyao Feng, Yikun Liu, and Gongping Yang. Dual-constraint autoencoder and adaptive weighted similarity spatial attention for unsupervised anomaly detection. *IEEE Transactions on Industrial Informatics*, 20(7):9393–9403, 2024. doi: 10.1109/TII.2024.3384583.

[44] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16699–16708, 2024.

[45] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3914–3923, June 2023.

[46] Ying Zhao. Omnial: A unified cnn framework for unsupervised anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3924–3933, June 2023.

[47] 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 (CVPR)*, pages 28502–28511, June 2026.

[48] 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 d3: A real-world 2d/pseudo-3d/3d dataset for industrial anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15214–15223, 2025.

Appendix

A Overview

This supplementary material provides additional theoretical analysis, extended experimental results, and discussions of the limitations of the proposed Efficient Cross-modal Feature Reconstruction (ECFR) framework.

–**Sec. B** summarizes recent unsupervised anomaly detection methods along category and modality axes.

–**Sec. C** formally supports the core hypothesis of cross-modal reconstruction through two statistical inequalities with empirical validation.

–**Sec. D** analyzes cross-modal reconstruction from two perspectives: inference-time anomaly response and preservation of modality-specific information.

–**Sec. E** provides extended quantitative results, implementation details, and additional benchmarks.

–**Sec. F** discusses the limitations.

B Summary of Related Work

Existing anomaly detection methods can be broadly organized along two dimensions: category scope and modality scope, as illustrated in Fig. 5. Earlier studies mainly focused on either extending anomaly detection from single-class to unified multi-class modeling, or incorporating multiple modalities under category-specific settings. For instance, methods like UniAD pioneered multi-class modeling within a unified framework, while representative multimodal approaches such as CFM explored cross-modal feature correspondence for specific categories. Building upon these parallel advancements, recent studies such as UniMMAD [47] further extend anomaly detection to jointly handle multiple categories and modalities. Together, these developments reflect a gradual convergence of multi-class modeling and multimodal anomaly detection.

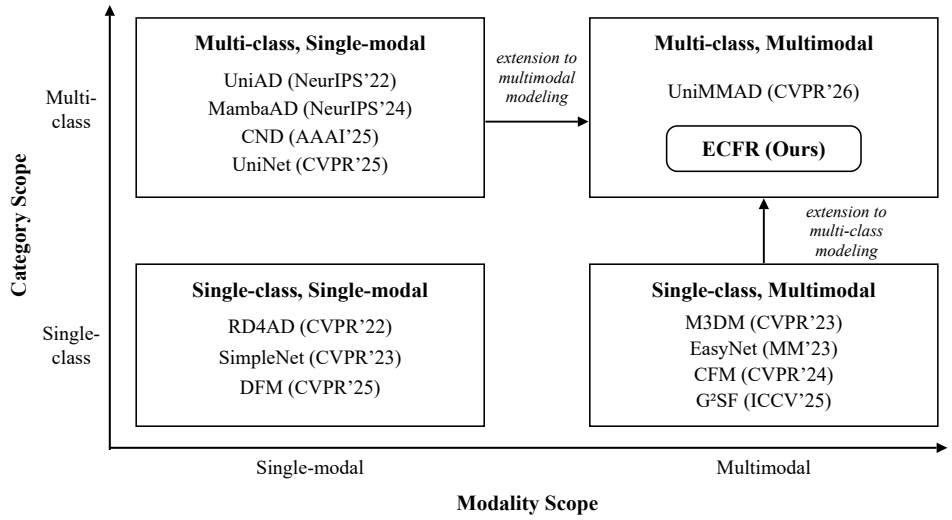


Figure 5: Representative anomaly detection methods organized by category and modality scope. Methods are grouped according to their original training settings.

ECFR contributes to the emerging direction of multi-class multimodal anomaly detection through a distinct cross-modal reconstruction perspective. It jointly models cross-modal interactions and spatial structures across categories, while explicitly exploiting RGB–Depth heterogeneity as an anomaly-discriminative constraint. In this way, the intrinsic difficulty of cross-modal reconstruction is used to suppress shortcut reconstruction and amplify reconstruction discrepancies for anomalous inputs, providing a complementary mechanism for unified multi-class multimodal anomaly detection.

C Empirical Validations on Cross-modal Reconstruction

C.1 Motivation

To formalize the theoretical advantages of cross-modal reconstruction, we establish two statistical inequalities that underpin our core hypotheses: cross-modal heterogeneity inherently suppresses identical shortcut (Inequality 1) and amplifies anomaly detectability (Inequality 2).

Let $< x^m, x^{\bar{m}} >$ denote paired samples from two distinct modalities m and $\bar{m}$ (e.g., RGB/Depth), where $x^m \in \mathbb{R}^d$ and $x^{\bar{m}} \in \mathbb{R}^k$ are random variables drawn from a joint distribution $\mathcal{P}$. $\mathcal{F}_{\text{cross}}$ denotes a cross-modal autoencoder, and $\mathcal{F}_{\text{single}}$ a single-modal autoencoder. The expectation operator $E_{\mathcal{P}}$ denotes the mean over samples from $\mathcal{P}$, where $\mathcal{P}_{\text{norm}}$ and $\mathcal{P}_{\text{anom}}$ refer to the normal and anomalous data distributions, respectively. Under three mild regularity conditions on the joint distribution $\mathcal{P}$ described later in this section, the following holds:

Inequality 1 (Cross-modal Difficulty)

Due to inter-modal heterogeneity, reconstructing one modality from another introduces fundamental challenges. Thus, it is reasonable that cross-modal reconstruction exhibits intrinsically higher complexity than single-modal reconstruction, even for normal inputs:

$$E_{\mathcal{P}_{\text{norm}}}\left[\left\|\mathcal{F}_{\text{cross}}(x^{\bar{m}}) - x^m\right\|_2^2\right] \geq E_{\mathcal{P}_{\text{norm}}}\left[\left\|\mathcal{F}_{\text{single}}(x^m) - x^m\right\|_2^2\right] + \Delta_{\text{hetero}}, \tag{12}$$

where $\Delta_{\text{hetero}} > 0$ quantifies the inherent heterogeneity penalty.

Inequality 2 (Anomaly Amplification)

As the model is trained exclusively on normal samples, the model yields statistically larger reconstruction errors for anomalous inputs than for normal ones:

$$E_{\mathcal{P}_{\text{anom}}}\left[\left\|\mathcal{F}_{\text{cross}}(x^m) - x^{\bar{m}}\right\|_2^2\right] \geq \kappa \cdot E_{\mathcal{P}_{\text{norm}}}\left[\left\|\mathcal{F}_{\text{cross}}(x^m) - x^{\bar{m}}\right\|_2^2\right], \tag{13}$$

with $\kappa > 1$.

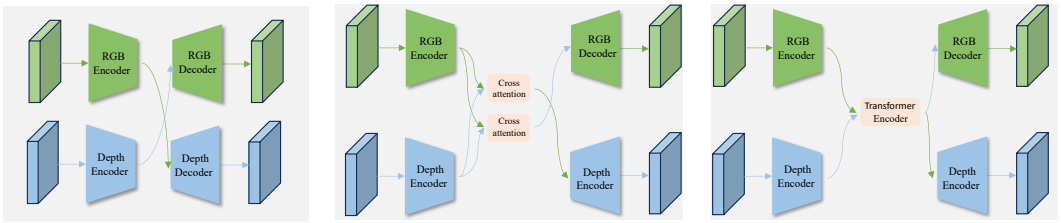
In industrial anomaly detection, three conditions are intrinsically satisfied by domain-specific physical constraints: (1) Modal Correlation is guaranteed through sensor co-capture of physical objects, where RGB texture and depth geometry exhibit deterministic correspondence; (2) Lipschitz Continuity arises from the bounded noise characteristics of industrial imaging systems; (3) Anomaly Shift is fundamentally satisfied as defects (scratches/dents) induce localized feature deviations violating stationarity assumptions in homogeneous normal textures. These properties derive from industrial environments' structured nature—spatially regular objects and statistically anomalous defects—axiomatically justifying Inequalities (12) and (13) for cross-modal reconstruction.

While theoretically grounded in industrial settings, applying Inequalities (12) and (13) to other domains requires explicit verification of the above three conditions. We will present deep analysis in our future work.

C.2 Experimental Setup

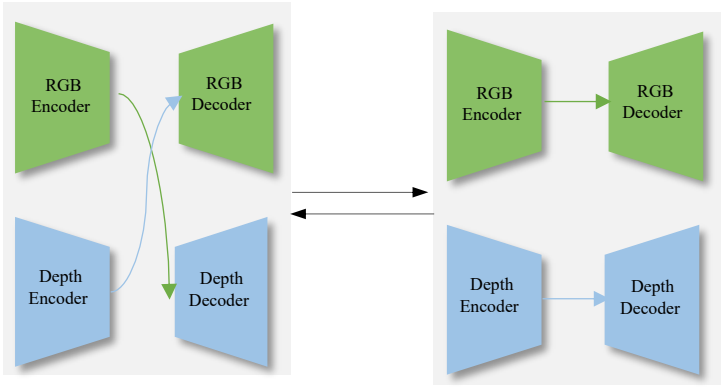
To empirically validate Inequalities (12) and (13) and examine whether cross-modal reconstruction amplifies anomaly signals over dual-branch single-modal reconstruction, we design three deliberately weak baseline architectures. All baselines adopt the same pretrained ResNet-34 backbone to extract three-scale features from both modalities, which are then aggregated and fed into a simple autoencoder for feature reconstruction. The three cross-modal variants differ only in the interaction layer: Baseline (a) employs a non-interactive convolutional autoencoder (Fig. 6a); Baseline (b) inserts cross-attention blocks for modality fusion (Fig. 6b); and Baseline (c) uses a Transformer encoder for interaction (Fig. 6c). For a fair comparison, we also construct dual-branch single-modal variants for each architecture (Fig. 7). These variants share the same backbone and aggregation scheme, differing only in data flow: each modality is reconstructed from its own features. Finally, we evaluate all models on the MVTec 3D-AD dataset under the unified multi-class setting.

Consistent with Inequality (12), we quantified the heterogeneity penalty (Δ_{hetero}) as the reconstruction error gap between cross-modal and dual-branch single-modal models. Specifically, Δ_{hetero}



(a) Independent AE (b) Cross-attention fusion (c) Transformer interaction

Figure 6: Architecture of cross-modal reconstruction baselines. (a) A non-interactive convolutional autoencoder; (b) an architecture with cross-attention blocks for modality fusion; and (c) a Transformer-based interaction model.



cross-modal dual-branch single-modal

Figure 7: A cross-modal reconstruction baseline and its paired dual-branch single-modal reconstruction model.

is computed by subtracting the mean reconstruction error of single-modal models from that of cross-modal models across all normal samples in the dataset.

Based on Inequality (13), we define the Relative Error Increasement (REI) between anomalous and normal inputs as:

$$\mathrm{REI} = \frac{E_{\mathcal{P}_{\mathrm{anom}}}\left[\|\mathcal{F}_{\mathrm{cross}}(x^m) - x^{\bar{m}}\|_2^2\right]}{E_{\mathcal{P}_{\mathrm{norm}}}\left[\|\mathcal{F}_{\mathrm{cross}}(x^m) - x^{\bar{m}}\|_2^2\right]} - 1. \tag{14}$$

C.3 Results and Analysis

Figure 8(a) shows that cross-modal reconstruction is consistently more challenging than dual-branch single-modal reconstruction. As shown in Figure 8(b), our method achieves the highest REI among all models, and even its single-modal variant outperforms cross-modal baselines in amplifying anomaly signals, confirming the effectiveness of HAC and AFIR modules.

The statistical guarantees of inequalities (12) and (13) provide a theoretical foundation for AFIR-HAC's efficacy: Inequality (12) explains why cross-modal reconstruction prevents shortcut (higher Δ_{hetero}). Inequality (13) gives a formal basis for the anomaly amplification effect. When an anomaly appears in either modality, the contamination propagation illustrated in Fig. 1 corrupts the corresponding locations in the opposite modality. Both cross-modal reconstruction branches then produce heightened anomaly responses, rather than letting the anomaly remain confined to a single branch.

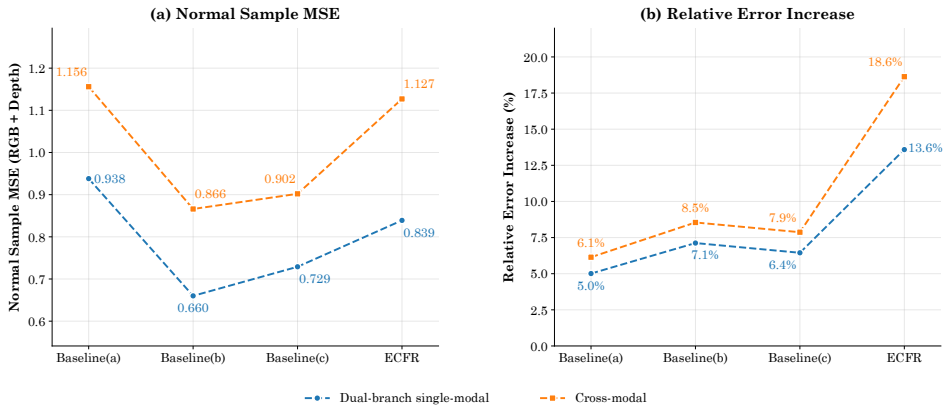


Figure 8: Statistical Evaluation of Reconstruction Errors.

D Analysis of Cross-Modal Reconstruction Mechanisms

Complementing the main text's reconstruction-centric perspective, this section analyzes inference behaviors under anomalous conditions, revealing how AFIR-HAC synergism triggers failure amplification through cross-modal feature interaction and spatial disintegration.

D.1 AFIR: Cross-Modal Propagation of Anomalies

The AFIR module enables cross-modal feature fusion and refinement. Its core component, the Intra-inter Modal Interaction (IMI) block, orchestrates this process by jointly modeling intra-modal dependencies and inter-modal correlations based on normal patterns, thereby providing a basis for discriminative anomaly detection. During inference, anomalous features, e.g., $< F^{m}_{\text{anom}}, F^{\bar{m}}_{\text{anom}} >$ or $< F^{m}_{\text{norm}}, F^{\bar{m}}_{\text{anom}} >$, trigger contamination propagation: anomalies in one modality spread through the attention-based interactions within IMI. Here, the superscript $m \in \{\text{R}, \text{D}\}$ denotes one modality (RGB or depth), and $\bar{m}$ indicates its counterpart, while the subscripts anom and norm represent anomalous and normal features, respectively. Specifically, the attention computation within IMI can be conceptually simplified as a Query-Key-Value (QKV) mechanism for illustrative purposes (Fig. 9). Anomalous features inherently deviate from the learned normal feature distributions. This divergence distorts the softmax-normalized attention weights, causing anomalous features to contaminate the fused representations across both modalities. Consequently, even initially normal features (e.g., F'_{n1} in Fig. 9) are transformed into corrupted outputs ($\widetilde{F_{n1}}$, highlighted purple in Fig. 9), which violate the model's learned normality patterns. The subsequent Adaptive Channel-wise Recalibration (ACR) block further amplifies the influence of these corrupted activations by channel-wise reweighting.

D.2 HAC: Spatial Disintegration in Global and Local Contexts

The Hybrid Attention Convolution (HAC) module integrates cascaded Agent Attention for global modeling and parallel multi-kernel depth-wise separable convolutions for local information capture.

Upon processing the corrupted representation (derived from the AFIR-fused features $\check{F}^{m}_{i}$), HAC exhibits divergent responses to distributional deviations in both global and local contexts. (1) Global Pathway (Agent Attention). Deviated query vectors in the attention mechanism alter affinity structures learned from normal data. Even with normal keys, anomalous queries generate misaligned attention maps, disrupting global contextual coherence. (2) Local Pathway (Multi-kernel Convolutions) Under anomalous inputs, multi-kernel convolutions fail to preserve spatial continuity learned from normal samples. This fractures local structures, further disintegrating the spatial composition of the output features.

Thus, global attention collapse and local structural disintegration jointly trigger spatial incoherence in outputs, amplifying reconstruction errors. This mechanism intrinsically suppresses "identical shortcut" by forcing anomaly-specific error propagation.

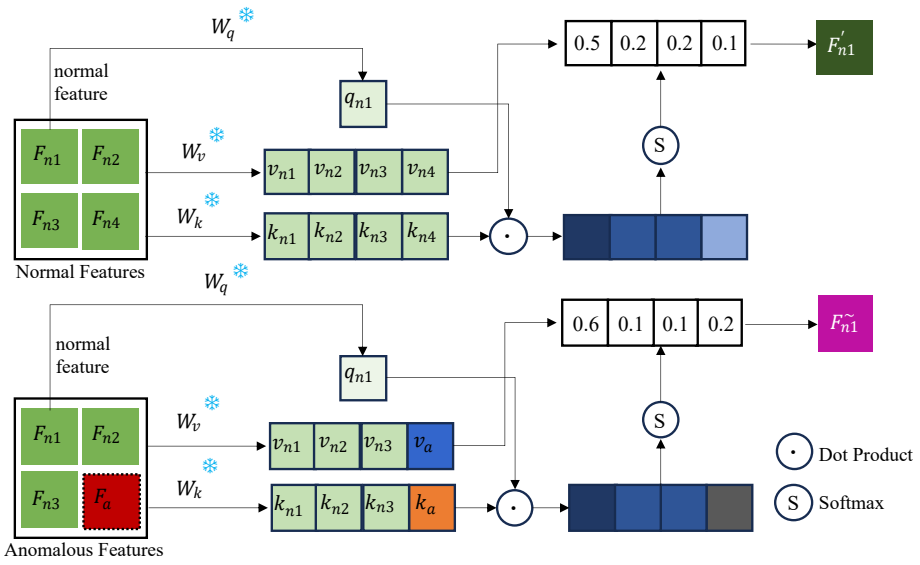


Figure 9: Illustration of Anomaly Propagation Mechanism.

D.3 Preservation of modality-specific Information

A potential risk of cross-modal reconstruction is that it may homogenize features and dilute modality-specific information. ECFR mitigates this through its architectural design and reconstruction objective. AFIR performs adaptive feature modulation, selectively integrating information from the counterpart modality via channel-wise recalibration, rather than uniformly fusing all channels. HAC ensures that features at every scale—from local details to global structures—are captured through dedicated pathways. Moreover, the reconstruction loss $\mathcal{L}^m$ acts as a strong optimization constraint: minimizing it requires the model to capture from the source modality all information necessary to reconstruct the target, including its distinctive attributes. Any loss of modality-specific information would directly increase reconstruction error, contradicting the optimization objective. These mechanisms collectively ensure that cross-modal complementarity is preserved throughout reconstruction. As a result, anomalies present in either modality reliably trigger the propagation and amplification effects described in Sections D.1 and D.2.

E Extended Quantitative Results

E.1 Datasets

MVTec 3D-AD is a comprehensive multimodal industrial anomaly detection dataset that includes 10 real-world industrial products. It contains 2,656 training samples, 294 validation samples, and 1,197 test samples. The training and validation data are defect-free, while the test data consists of 249 normal samples and 948 anomalous samples. The anomalous test samples contain approximately 4 to 5 different types of defects within each category. A key distinction of MVTec 3D-AD from the earlier 2D-AD datasets is that each sample provides a high-resolution point cloud along with the corresponding RGB image.

Eyecandies [2] is a novel synthetic dataset comprising ten categories of candy samples rendered under controlled environments. These ten types of candy exhibit various challenging characteristics, including complex textures, self-occlusion, and specular reflections. By precisely controlling lighting conditions and procedural rendering pipeline parameters within the modeling software, the dataset authors created a dataset featuring a wide range of complex scenes. The dataset includes 10,000 training samples, 1,000 validation samples, and 4,000 test samples, with RGB images and pixel-aligned depth maps provided for each sample. Similar to MVTec 3D-AD, the training set of this dataset contains only normal samples, while the test set includes both normal and anomalous samples.

E.2 I-AUROC, P-AUROC and AUPRO

I-AUROC denotes image-level AUROC, and P-AUROC denotes pixel-level AUROC. Image-level AUROC is computed as the area under the ROC curve based on an entire image, where the x-axis represents the false positive rate (FPR) and the y-axis represents the true positive rate (TPR). Each point on the curve corresponds to the model's performance under a different classification threshold. By adjusting the image-level classification threshold, different points are obtained to plot the complete ROC curve. Image-level AUROC evaluates the model's ability to classify entire images. Pixel-level AUROC, on the other hand, is calculated based on the area under the ROC curve for each pixel, reflecting the model's segmentation accuracy at the pixel level.

AUPRO is defined as the average relative overlap between the binary prediction P and each connected component C_r in the ground-truth mask. This process is repeated across multiple thresholds, and the resulting AUPRO values are plotted against the corresponding false positive rates (FPR) to form a curve. The final metric is obtained by integrating this curve up to a predefined FPR limit (typically 0.3) and normalizing to [0, 1]. This is a standard evaluation metric for unsupervised anomaly localization tasks and is particularly effective when the sizes of anomaly regions vary significantly.

E.3 Implementation Details

Data pre-processing. As in [6, 34], we estimate the background plane from the point cloud using RANSAC [12] and remove points within a distance of 0.005. The corresponding pixels in the RGB image are set to zero accordingly.

Train and Test Details. The training approach in the multi-class setting differs from the traditional single-class training in that we do not distinguish between categories in the training set. Instead, all training data is treated as a single class during training, following the same experimental protocols of existing multi-class methods such as UniAD and RLR [16]. We use the trained unified model to test each class in the test set. All input images are resized to a uniform size of 256×256. A pre-trained ResNet-34 acts as the feature extractor. The AdamW optimizer is employed with a learning rate of 0.005 and a decay rate of 1×10^{-4}. The model undergoes a training period of 500 epochs for the multi-class setting, conducted on a single NVIDIA A800 80GB GPU. During training, the sum of MSE across different scales is employed as the loss function. In the testing phase, the sum of cosine similarities at various scales is used to generate the anomaly maps.

E.4 Comparison with Memory-bank Methods

Table 8 compares ECFR with M3DM [34] and G^{2}SF [31] under the unified multi-class setting on MVTec 3D-AD. Unlike reconstruction-based models that must encode diverse normal patterns within a shared reconstruction mapping, memory-bank methods explicitly store normal features and retrieve them as references during inference. Consequently, inter-category diversity can be represented directly in the memory bank, making these methods less sensitive to the increased modeling difficulty of the unified multi-class setting. This advantage, however, introduces additional costs for feature storage, bank construction, and retrieval.

Table 8: Memory-bank comparison on MVTec 3D-AD.

Method	I-AUROC	P-AUROC	AUPRO	Params (M)	FPS	Peak bank-construction footprint (GiB)
G^{2}SF	90.25	**99.53**	**97.28**	109.798	1.740	**20.967**
M3DM	**93.33**	99.31	96.48	123.474	0.130	113.472
ECFR	91.85	98.60	95.37	**44.627**	**21.356**	None

None: no feature-bank construction.

Coreset selection reduces these costs by retaining only representative normal features, but may also reduce the coverage of normal patterns. G^{2}SF uses a 10% coreset per modality and retains 1.35 GiB of normal features. In comparison, M3DM has a substantially larger bank-construction footprint and achieves higher I-AUROC. The larger retained feature set may partly contribute to its stronger detection performance, although the two methods also differ in their algorithms. Overall, memory-

bank methods preserve normal patterns explicitly and achieve strong localization performance, whereas ECFR avoids feature-bank storage and retrieval and provides a different accuracy-efficiency trade-off for unified multi-class deployment.

E.5 Visualizations on Eyecandies

Figure 10 illustrates the visualization results of ECFR on the Eyecandies dataset, demonstrating its superior anomaly segmentation accuracy and enhanced robustness compared to existing methods.

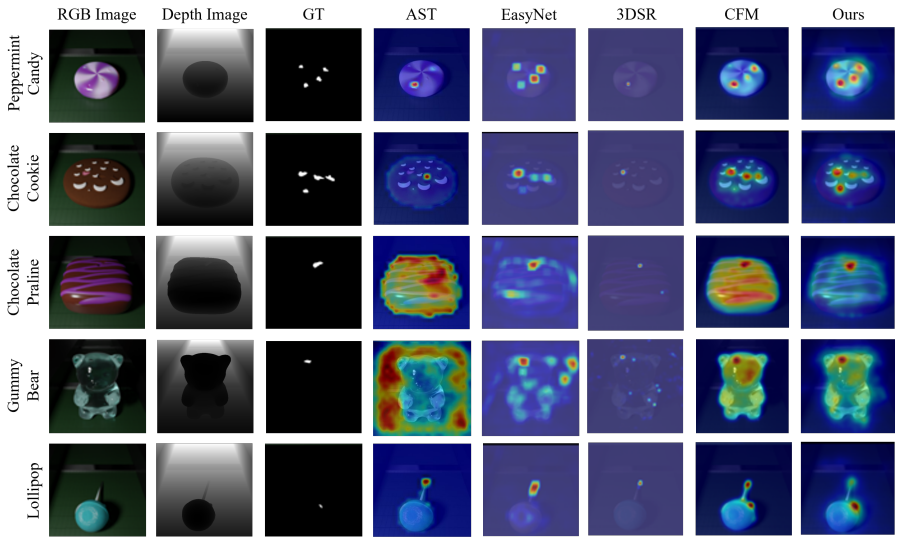


Figure 10: Qualitative visualization for pixel-level anomaly segmentation on Eyecandies.

E.6 Comparison with Single-Class MultiModal AD Methods

As stated in the main paper, we re-trained several state-of-the-art single-class multimodal industrial anomaly detection methods in a multi-class setting by merging all training categories into one, using their released code without modifications. ECFR is designed for multi-class detection, but for comprehensive experimental validation, we also compared it with these methods in the single-class setting for fairness. We evaluated ECFR on the MVTec 3D-AD and Eyecandies datasets (see Table 9). ECFR remained competitive with leading single-class methods and outperformed EasyNet, which also targets efficiency, on multiple metrics. Results for compared methods are taken directly from their original papers. "–" indicates metrics not reported in the original work.

Table 9: Comparison with Single-Class MultiModal AD Methods.

Datasets	Method	I-AUROC	P-AUROC	PRO
MVTec 3D-AD	AST	93.7	97.6	94.4
	EasyNet	92.6	–	82.1
	3DSR	**97.8**	**99.5**	**97.2**
	CFMs	<u>95.4</u>	99.3	<u>97.1</u>
	Ours	93.0	98.6	95.3
Eyecandies	AST	75.8	90.2	87.8
	EasyNet	86.9	–	–
	3DSR	**90.9**	–	–
	CFMs	<u>88.1</u>	**97.4**	**88.7**
	Ours	86.6	<u>96.6</u>	<u>88.1</u>

E.7 Evaluations on more datasets

To further assess ECFR, we evaluate it on two additional benchmarks: MulSen AD [19] (RGB + infrared) to examine applicability to a different modality combination, and Real-IAD D^3 [48] (20 categories) to test performance in a more complex multi-class scenario. Results are compared with those reported in their original papers (Table 10). ECFR achieves competitive performance on both datasets, demonstrating that cross-modal feature reconstruction generalizes to different modality combinations and larger category counts.

Table 10: Results on Additional Datasets Real-IAD and MulSen AD.

Dataset	I-AUROC	P-AUROC	PRO
MulSen AD	90.52	93.34	83.71
Real-IAD D^3	83.19	91.53	76.79

E.8 Comparison with Multi-Class Single-Modal AD Methods

Existing multi-class single-modal methods typically use single-modal datasets (e.g., MVTec AD, VisA), while our method employs multimodal datasets like MVTec 3D-AD. Notably, these single-modal methods show significant performance drops on MVTec 3D-AD, as reported by ADer [42]. We attribute this in part to the nature of multimodal datasets, where certain anomalies are detectable only in specific modalities, putting single-modal approaches at a disadvantage. To provide a more comprehensive evaluation, we include their RGB-only results from ADer alongside ECFR using RGB and depth in Table 11. This comparison reflects both the detection methods and the information available from their input modalities.

Table 11: Comparison with Multi-Class Single-Modal AD Methods.

Method	I-AUROC	P-AUROC	PRO
UniAD (NeurIPS'22)	77.0	<u>96.8</u>	89.4
DiAD (AAAI'24)	78.8	91.8	75.1
MambaAD (NeurIPS'24)	<u>85.5</u>	**98.6**	<u>94.1</u>
Ours	**91.9**	**98.6**	**95.4**

E.9 Ablation Studies on the Agent-Attention Token Count

We conduct ablation studies on the Agent-Attention token count n, as summarized in Table 12. Our method employs pooling to obtain Agent tokens, with varying n values across different stages. Consistent with insights from Agent Attention research [14], reducing the number of proxy tokens in shallower layers does not degrade performance. Following Agent Attention findings, we designed four configurations. Configuration 1 keeps $A = Q$ (no pooling, baseline). Results show the [49, 49, 16, 9] configuration delivers optimal performance.

Table 12: Ablation studies on the agent-attention token count.

ID	Stage1	Stage2	Stage3	Stage4	Params(M)	FLOPs(G)	I-AUROC
1	64	256	1024	4096	60.518	44.251	91.35
2	64	64	16	9	44.870	15.847	91.73
3	49	49	16	9	44.627	15.762	**91.85**
4	36	36	16	9	**44.416**	**15.689**	91.31

F Limitations

Existing multimodal industrial anomaly detection methods adopt datasets that provide well-aligned RGB–Depth pairs, and ECFR follows the same practice. However, in real-world industrial environments, sensor misalignment, calibration drift, and partial modality absence may arise. Our

single-modality evaluation (Table 7) demonstrates that ECFR can operate with either RGB or depth when trained with the corresponding input, while handling missing, noisy, or misaligned modalities remains an open challenge. Additionally, ECFR is designed as a dual-branch architecture and cannot directly accommodate more than two modalities. Extending cross-modal reconstruction to incomplete, misaligned, or multi-source settings is a promising direction for future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our contributions are clearly stated in the Abstract and Sec. 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper presents an empirical approach to multi-class multimodal anomaly detection. Inequalities 12 and 13 are formulated as hypotheses and evaluated experimentally in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The architecture is detailed in Sec. 3, and the code is available at the link in the Abstract.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: As mentioned in the Abstract, our code is publicly available. The datasets used are publicly available benchmarks.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Please see the implementation details in Sec. 4.1 and Appendix E.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report error bars or confidence intervals. This is consistent with the evaluation protocol of prior work in this area, which reports single-run results on standard benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see the implementation details in Sec. E.3 and the efficiency evaluation in Sec. 4.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research conforms with these ethical standards.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work focuses on industrial anomaly detection using publicly available benchmark datasets, and does not have a direct path to negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We think that our paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets and models utilized in this paper are all publicly available, and we have cited the corresponding papers accordingly.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our paper introduces a new framework without releasing new datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The LLM is used only for writing and editing, and the core method development in this paper does not involve LLMs as any important component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.