CMDS-AD: Cross-Modal Dual-Stream Decoupling for Few-Shot Anomaly Detection

Junhao Cai[1], Deyu Zeng[1,2⋆], Junhao Pang[1], Junyu Chen[2], Qiwei Liang[1,3], Xiaopin Zhong[1], and Zongze Wu[1]

[1] Shenzhen University, Shenzhen, Guangdong 518060, China
caijunhao27@gmail.com, 2500092013@mails.szu.edu.cn, liangqiwei2022@email.szu.edu.cn, {xzhong,zzwu}@szu.edu.cn
[2] Guangzhou Maritime University, Guangzhou, Guangdong 510725, China
{zengdeyu,chenjunyu}@gzmtu.edu.cn
[3] Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong 511453, China

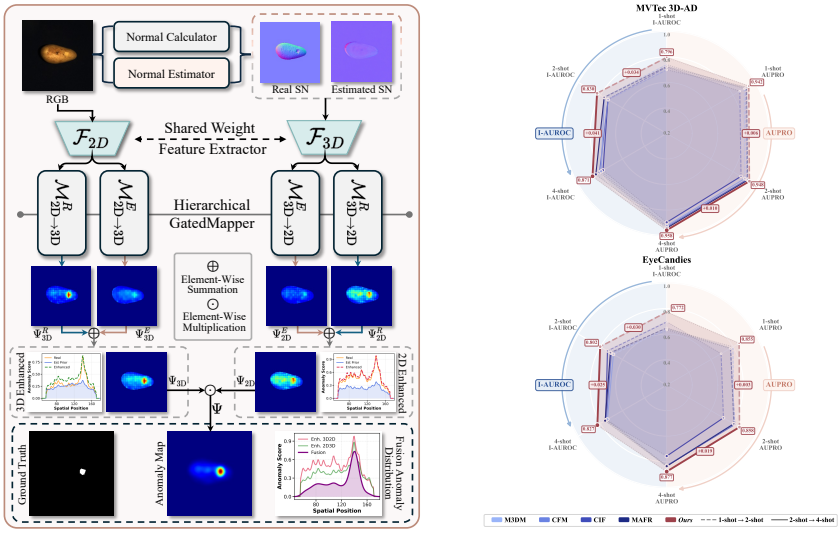
Abstract. Few-shot anomaly detection remains challenging due to limited training data. Multi-modal anomaly detection (MAD) offers a viable solution, leveraging 3D geometric cues to enrich 2D RGB representations and compensate for this scarcity. However, existing MAD methods apply spatially uniform feature processing, conflating stable macroscopic structures with high-frequency localized defect signals, exacerbating cross-modal misalignment and inflating false-positive rates. To overcome this, we present CMDS-AD, a Cross-Modal Dual-Stream Anomaly Detection framework. A LoRA-guided diffusion model generates diverse RGB samples to mitigate extreme data scarcity. For 3D normal augmentation, we employ a pre-trained diffusion model as a normal estimator. Crucially, this estimator inherently acts as a non-linear low-pass filter, directly extracting low-frequency normal representations from RGB inputs. This establishes an auxiliary estimated stream of purely low-frequency information, anchoring robust structural templates and assisting the uncompressed real stream, containing coupled high- and low-frequency components, to precisely isolate micro-defects. A Coordinate-Aware Hierarchical Feature Mapper adaptively aligns cross-modal semantics, while a multiplicative scoring mechanism filters modality-specific noise. Under the extreme 1-shot setting, CMDS-AD achieves absolute performance gains of **5.7%** (I-AUROC) and **2.0%** (AUPRO) on MVTec 3D-AD, alongside **7.7%** and **5.6%** improvements on EyeCandies, establishing a new state-of-the-art. Code is available at Junhaocai27/CMDS-AD.

Keywords: Few-Shot Learning · Multi-Modal Anomaly Detection · Diffusion Models · Dual-Stream Optimization

1 Introduction

Anomaly detection (AD) has become an indispensable component of industrial visual inspection, ensuring product quality and manufacturing stability

⋆ Corresponding author.



(a) Mechanism **(b)** Results

Fig. 1: Few-shot 3D anomaly detection overview. (a) Illustration showing how estimated streams enhance defect representation in real streams, and multi-modal 2D-3D multiplication further localizes defects. (b) Radar plots on MVTec 3D-AD and EyeCandies report few-shot (1, 2, 4-shot) I-AUROC and AUPRO. Our method (red) outperforms baselines (blue) with annotated gains, demonstrating robust improvements.

where manual inspection is inefficient and error-prone. Since anomalous samples (*e.g.* defects, scratches, structural damages) are unpredictable and extremely rare in real-world scenarios, most existing AD methods adopt an unsupervised paradigm, training exclusively on normal, defect-free data to identify anomalies as deviations from nominal distributions [8, 25, 31, 40].

With the rapid advancement of 3D sensors, multi-modal anomaly detection (MAD) has garnered increasing interest [17, 22, 36]. Relying on a single modality (*e.g.* RGB) often fails to capture crucial geometric cues obscured by complex surface textures, whereas 3D information provides highly complementary structural insights [19, 38]. While significant progress has been made in MAD under full-shot settings—ranging from memory-based pattern matching [6, 34] and feature adaptation [28, 33] to diffusion-driven reconstruction [21, 41]—existing approaches rely on a large volume of normal training data. In agile manufacturing, acquiring extensive multi-modal normal samples for every new product line is highly expensive. Consequently, few-shot anomaly detection (FSAD) has emerged as a critical alternative [9, 13, 14, 16, 30]. However, when restricted to severe few-shot scenarios (*e.g.* 1 to 4 samples), the performance of state-of-the-art multi-modal methods drops precipitously (as evidenced in Fig. 1).

The fundamental bottleneck of extending MAD to few-shot scenarios is the profound modality gap and modality interference under data scarcity. Existing methods attempt to bridge this gap through various fusion strategies, but of-

ten fall short. For instance, fusing modalities via input-level concatenation (*e.g.* AST [26]) or multi-level spatial fusion (*e.g.* MMRD [10]) risks severe modality interference. Cross-modal mapping approaches, such as CFM [7], attempt to map features between modalities; however, the inherent modality gap often leads to large reconstruction errors even for normal samples, inherently resulting in high false-positive rates and coarse localization boundaries. Meanwhile, memory-based architectures like M3DM [34] and ShapeGuided [6] struggle because limited normal samples fail to adequately cover the nominal feature patterns. Fundamentally, these methods treat all spatial features uniformly, neglecting the frequency properties of industrial data. By entangling stable low-frequency macroscopic structures with unpredictable high-frequency localized variations (*e.g.* minor texture shifts), they inevitably confuse normal sensor noise with genuine structural defects in data-scarce scenarios.

To overcome these severe limitations, we propose Cross-Modal Dual-Stream Anomaly Detection (CMDS-AD), a novel few-shot MAD framework that explicitly establishes a dual-stream architecture while deeply exploiting generative diffusion priors. Rather than solely utilizing diffusion models as monolithic augmentors via LoRA-guided synthesis to generate paired training data, we innovatively repurpose a pre-trained diffusion-based geometric estimator as a non-linear low-pass filter. This mathematically establishes an estimated stream comprising purely low-frequency information, which serves as an auxiliary anchor for the stable macroscopic structural representation, constructing a robust nominal template even with single-digit samples. Concurrently, the uncompressed real stream, encompassing coupled high- and low-frequency components, is explicitly enhanced by this auxiliary prior to capture localized microscopic details. By synergistically aggregating the anomaly responses from these parallel streams, our framework naturally isolates and amplifies genuine defect signals while effectively suppressing cross-modal interference.

To precisely align these decoupled dual-stream representations without risking feature collapse caused by cross-modal amplitude discrepancies, we design an adaptive Coordinate-Aware Hierarchical Feature Mapper. Unlike previous rigid fusion mechanisms, it dynamically regulates hierarchical feature aggregation within each independent mapping pathway across multiple scales, while rigorously preserving spatial positional priors. Finally, we introduce a Cross-Modal Multiplicative Anomaly Scoring module ($\Psi_{\text{2D}} \odot \Psi_{\text{3D}}$). This acts as a stringent spatial filter that flags defects only given concurrent multi-modal anomalies, dramatically suppressing previously unavoidable modality-specific false alarms triggered by localized sensory artifacts.

Our main contributions are summarized as follows:

1. **Cross-Modal Dual-Stream Framework (CMDS-AD):** We propose the first few-shot MAD architecture driven by a dual-stream perspective. By innovatively repurposing diffusion estimators as non-linear low-pass filters, we establish an auxiliary estimated stream that effectively enhances the real stream to isolate unpredictable micro-defects from stable macroscopic structures.

2. **Adaptive Cross-Modal Feature Alignment:** We design a Coordinate-Aware Hierarchical Feature Mapper coupled with a Decoupled Multi-Scale Mask-Aware Optimization strategy. This adaptively aligns heterogeneous 2D and 3D semantic spaces, fundamentally mitigating the cross-modal gap and modality interference under severe data scarcity.
3. **Synergistic Multiplicative Anomaly Scoring:** We introduce a novel scoring mechanism that aggregates anomaly responses from both the real and estimated streams to filter out isolated modality-specific noise. Extensive experiments on MVTec 3D-AD and EyeCandies demonstrate that our method establishes a new state-of-the-art in few-shot settings.

2 Related Work

Multi-modal Anomaly Detection. With the advent of 3D sensors, multi-modal anomaly detection (MAD) has gained significant attention [6,7,34] as 3D geometric cues effectively complement RGB features. Recent methods typically explore cross-modal feature alignment (*e.g.* CFM [7], CIF [20]), memory-based architectures (*e.g.* ShapeGuided [6], M3DM [34], MAFR [1]), or multi-level feature fusion (*e.g.* MMRD [10]) to localize anomalies. Although these approaches achieve remarkable performance under full-shot settings, they heavily rely on extensive normal training data or perfectly aligned networks. In severe few-shot scenarios, the profound modality gap between 2D images and 3D point clouds critically hinders reliable cross-modal alignment. To address this fundamental limitation, we propose a Coordinate-Aware Hierarchical Feature Mapper that bypasses amplitude discrepancies and adaptively aligns multi-scale contextual features, efficiently closing the modality gap under data scarcity.

Few-shot Anomaly Detection. To alleviate the prohibitive cost of collecting large-scale normal data, few-shot anomaly detection (FSAD) has emerged as a promising alternative [9,13,14,16,30]. Current research focuses on the RGB domain, utilizing vision-language prompts (AnoPLe [16]), feature regression (FastRecon [9]), window-based fusion (WinCLIP [14]), or image registration (RegAD [13]). However, existing single-modality FSAD approaches struggle to capture structural distortions. Extending FSAD to multi-modal scenarios remains challenging because traditional methods uniformly treat all spatial features, exacerbating modality interference when data is scarce. Motivated by this, our method explicitly establishes a dual-stream architecture to process coupled and purely low-frequency representations, effectively mitigating cross-modal interference.

Diffusion Models in Visual Inspection. Diffusion models are increasingly applied in anomaly detection for data augmentation [24,32] or reconstruction [35, 39]. However, existing methods often treat generation as a monolithic module, overlooking the frequency dynamics of progressive denoising [5, 23]. Distinctly, we rethink their role in FSAD. While utilizing LoRA [12,29] for data synthesis to mitigate scarcity, we exploit findings that diffusion estimators inherently act as non-linear low-pass filters [27,37]. This establishes an auxiliary estimated stream

of purely low-frequency information to anchor macroscopic structures, enhancing the sensitive real stream (coupled with high/low frequencies) to capture micro-defects. This explicitly bridges generative priors with dual-stream alignment, resolving few-shot overfitting.

3 Methodology

3.1 Overall Pipeline Architecture

Figure 2 shows our CMDS-AD framework, designed to bridge the 2D-3D semantic gap under few-shot constraints. To overcome data scarcity, a Diffusion-Driven Multimodal Few-Shot Augmentation module (Sec. 3.2) first synthesizes paired RGB-normal training data. The pipeline then employs a bidirectional mapping mechanism (2D $\leftrightarrow$ 3D) comprising parallel real and estimated streams. Given the input modalities, a frozen backbone (e.g., ViT) extracts multi-scale features, which a Coordinate-Aware Hierarchical Feature Mapper (Sec. 3.4) adaptively aggregates via spatial gating. During training, a Decoupled Multi-Scale Mask-Aware Optimization strategy (Sec. 3.5) ensures precise alignment across the heterogeneous domains. During inference, the mappers ($\mathcal{M}_{\text{2D}\to\text{3D}}, \mathcal{M}_{\text{3D}\to\text{2D}}$) generate directional anomaly distance maps for both streams. Finally, a Cross-Modal Multiplicative Anomaly Scoring module (Sec. 3.6) synergistically fuses these directional predictions into a high-precision dense anomaly map Ψ, thereby effectively filtering out modality-specific artifacts to reduce false positives. The complete procedure is summarized in Algorithm 1.

3.2 Diffusion-Driven Multimodal Few-Shot Augmentation

In Few-Shot Anomaly Detection (FSAD), extreme data scarcity severely bottlenecks model generalization. To overcome this, we propose a multimodal data augmentation framework driven by diffusion priors. For the 2D (RGB) modality, we adopt a Low-Rank Adaptation (LoRA)-guided image-to-image generation paradigm. Standard diffusion models learn data distributions by optimizing the denoising objective:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{x_0,\epsilon\sim\mathcal{N}(0,\mathbf{I}),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right] \tag{1}$$

where x_0 is the clean image, ϵ represents the injected Gaussian noise, t is the timestep, and ϵ_θ denotes the denoising network parameterized by θ. Directly fine-tuning θ on limited samples inevitably triggers catastrophic forgetting and severe overfitting. Thus, we freeze the pre-trained projection matrices $W_0 \in \mathbb{R}^{d\times k}$ in the attention layers and inject trainable low-rank matrices $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ (with $r \ll \min(d, k)$). The forward pass update is formulated as:

$$W = W_0 + \Delta W = W_0 + BA \tag{2}$$

where W is the updated weight matrix and ΔW is the weight increment. This parameter-efficient formulation enables the model to extract domain-specific textural features from scarce normal samples, synthesizing structurally consistent

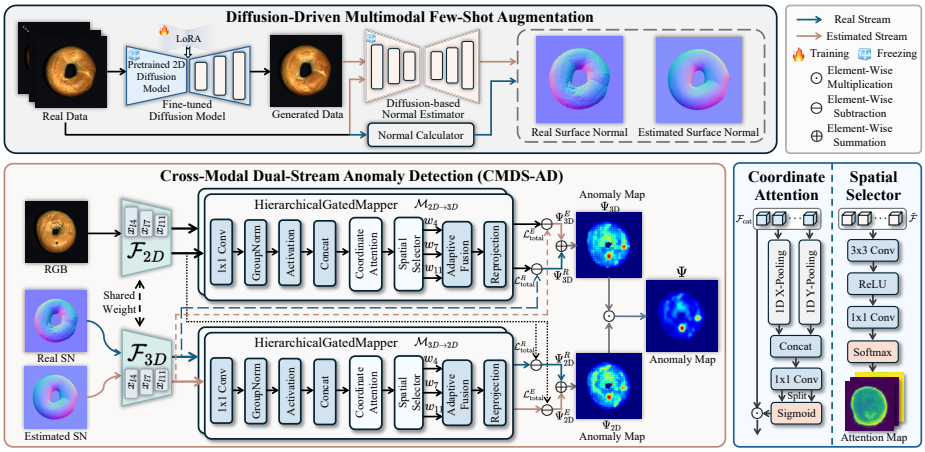


Fig. 2: Proposed Cross-Modal Dual-Stream Anomaly Detection (CMDS-AD) framework. **(Top)** Diffusion-driven augmentation synthesizes paired RGB-normal training data. **(Bottom Left)** Dual-stream pipeline processing Real and Estimated streams, where the purely low-frequency estimated stream acts as an auxiliary guide to enhance the coupled real stream. **(Bottom Right)** Coordinate-Aware Hierarchical Feature Mapper for adaptive multi-scale feature fusion.

yet distributionally diverse RGB representations. For the 3D modality, surface normal maps are more sensitive to subtle geometric discontinuities (*e.g.* scratches, dents) than depth maps. Consequently, we introduce a diffusion-based normal estimator. By processing both the real and LoRA-generated RGB images through this estimator, we obtain their estimated normal maps, completing the real-estimated cross-modal augmented dataset.

3.3 Dual-Stream Decoupled Anomaly Detection

During the pipeline, the network jointly processes the real surface normals and the estimated normals generated by the diffusion estimator Φ_{diff}. Constrained by generative priors and latent space compression, Φ_{diff} acts essentially as a robust non-linear low-pass filter, tending to output overly smooth results that inadvertently discard high-frequency details.

To formalize this theoretically, in the frequency domain, the real surface normal N encompasses a low-frequency structural component N_{low} (global shape) and a high-frequency detail component N_{high} (local micro-textures). For an input image I, the estimator yields a smoothed estimation $\hat{N} = \Phi_{\text{diff}}(I) \approx N_{\text{low}}$. Consequently, mapping these inputs through the backbone Φ yields two complementary feature spaces rather than a single entangled representation:

$$\mathcal{F}_{\text{est}} \approx \Phi(N_{\text{low}}), \quad \mathcal{F}_{\text{real}} \approx \Phi(N_{\text{low}} + N_{\text{high}}) \tag{3}$$

Leveraging this inherent filtering property, we design a synergistic dual-stream architecture. To avoid amplifying cross-modal noise via explicit fea-

Algorithm 1: CMDS-AD: Cross-Modal Dual-Stream Anomaly Detection

Input: Support set $\mathcal{D}_{\text{train}} = \{I_{\text{rgb}}, N, M\}$, Pre-trained backbone Φ, Untrained Mappers $\mathcal{M}_{2\text{D}\to3\text{D}}, \mathcal{M}_{3\text{D}\to2\text{D}}$

Output: Optimized Mappers $\mathcal{M}_{2\text{D}\to3\text{D}}, \mathcal{M}_{3\text{D}\to2\text{D}}$, Final Anomaly Map Ψ

Stage 1: Multimodal Data Augmentation

$\{A, B\} \leftarrow \arg\min_{A,B} \mathcal{L}_{\text{DM}}$; `// LoRA FT via Eq. 1 and 2`

$\hat{I}_{\text{rgb}} \sim \text{DM}_{W_0+BA}(I_{\text{rgb}}), \quad \hat{N} \leftarrow \Phi_{\text{diff}}(\hat{I}_{\text{rgb}})$; `// Generate cross-modal pairs`

$\mathcal{D}_{\text{aug}} \leftarrow \mathcal{D}_{\text{train}} \cup \{(\hat{I}_{\text{rgb}}, \hat{N})\}$;

Stage 2: Hierarchical Feature Mapper Optimization

for *batch* $\mathcal{B} \sim \mathcal{D}_{aug}$ **do**

  $\mathcal{F}^S_{2\text{D}}, \mathcal{F}^S_{3\text{D}} \leftarrow \Phi(\mathcal{B}), \quad S \in \{R, E\}$; `// Extract` $l \in \{4, 7, 11\}$ `layers`

  `// Coordinate-Aware Mapping (Sec. 3.4)`

  $\mathcal{P}^S_{3\text{D}} \leftarrow \mathcal{M}_{2\text{D}\to3\text{D}}(\mathcal{F}^S_{2\text{D}}), \quad S \in \{R, E\}$; `// 2D → 3D direction`

  $\mathcal{P}^S_{2\text{D}} \leftarrow \mathcal{M}_{3\text{D}\to2\text{D}}(\mathcal{F}^S_{3\text{D}}), \quad S \in \{R, E\}$; `// 3D → 2D direction`

  `// Decoupled Mask-Aware Optimization (Sec. 3.5)`

  for *layer* $l \in \{4, 7, 11\}$ **do**

    $\mathcal{L}^R_{\text{L}l} \leftarrow \text{Eq. } 8\big(\mathcal{L}_{\text{align}}(\mathcal{P}^R_{\text{L}l}, \mathcal{F}^R_{\text{L}l}), M\big)$; `// Masked avg for Real`

    $\mathcal{L}^E_{\text{L}l} \leftarrow \text{Eq. } 8\big(\mathcal{L}_{\text{align}}(\mathcal{P}^E_{\text{L}l}, \mathcal{F}^E_{\text{L}l})\big)$; `// Global avg for Est`

  end

  $\mathcal{L}^S_{\text{total}} \leftarrow \alpha\mathcal{L}^S_{\text{L4}} + \beta\mathcal{L}^S_{\text{L7}} + \gamma\mathcal{L}^S_{\text{L11}}, \quad S \in \{R, E\}$; `// Eq. 9`

  $\mathcal{L} \leftarrow \mathcal{L}^R_{\text{total}} + \mathcal{L}^E_{\text{total}}$; `// Overall objective`

  $\{\mathcal{M}_{2\text{D}\to3\text{D}}, \mathcal{M}_{3\text{D}\to2\text{D}}\} \xleftarrow{\nabla} \mathcal{L}$; `// Backpropagation`

end

Stage 3: Cross-Modal Multiplicative Anomaly Scoring

$\mathcal{F}^S_{2\text{D}}, \mathcal{F}^S_{3\text{D}} \leftarrow \Phi(X_{\text{test}}), \quad S \in \{R, E\}$;

$\mathcal{P}^S_{3\text{D}} \leftarrow \mathcal{M}_{2\text{D}\to3\text{D}}(\mathcal{F}^S_{2\text{D}}), \quad \mathcal{P}^S_{2\text{D}} \leftarrow \mathcal{M}_{3\text{D}\to2\text{D}}(\mathcal{F}^S_{3\text{D}})$;

$\Psi^S_{2\text{D}} \leftarrow \mathcal{L}_{\text{align}}(\mathcal{P}^S_{2\text{D}}, \mathcal{F}^S_{2\text{D}}), \quad \Psi^S_{3\text{D}} \leftarrow \mathcal{L}_{\text{align}}(\mathcal{P}^S_{3\text{D}}, \mathcal{F}^S_{3\text{D}})$; `// Eq. 7`

$\Psi_{2\text{D}} \leftarrow \Psi^R_{2\text{D}} + \lambda_1\Psi^E_{2\text{D}}, \quad \Psi_{3\text{D}} \leftarrow \Psi^R_{3\text{D}} + \lambda_2\Psi^E_{3\text{D}}$; `// Eq. 10`

$\Psi \leftarrow \Psi_{2\text{D}} \odot \Psi_{3\text{D}}$; `// Eq. 11`

return Ψ

ture residuals, we process both streams independently. The estimated stream ($\mathcal{F}_{\text{est}}$) acts as a purely low-frequency auxiliary anchor for macroscopic structures. Guided by this reference, the uncompressed real stream ($\mathcal{F}_{\text{real}}$)—which couples high- and low-frequency components—precisely captures localized surface defects. Synergizing the independent anomaly measurements from these distinct sub-spaces ensures comprehensive defect coverage.

3.4 Coordinate-Aware Hierarchical Feature Mapper

To align multimodal features across varying receptive fields, we design a feature mapper comprising hierarchical branches. The backbone (*e.g.* ViT) extracts multi-scale representations from the 4-th, 7-th, and 11-th layers, denoted

as $x_l \in \mathbb{R}^{C\times H\times W}$ for $l \in \{4, 7, 11\}$ (where $C = 768$). These represent local edge textures, part-level patterns, and global semantics, respectively. Prior to spatial gating, each extracted feature x_l first passes through a local adaptation block—comprising a 1×1 convolution, Group Normalization (GN), and a GELU activation—to harmonize the heterogeneous semantic spaces across different network depths. The adapted features are subsequently concatenated along the channel dimension to form a unified multi-scale input feature $\mathcal{F}_{\text{cat}} \in \mathbb{R}^{3C\times H\times W}$:

$$\hat{x}_l = \text{GELU}\big(\text{GN}(\text{Conv}_{1\times1}(x_l))\big), \quad \mathcal{F}_{\text{cat}} = \text{Concat}(\hat{x}_4, \hat{x}_7, \hat{x}_{11}) \tag{4}$$

Since anomaly detection requires spatially sensitive dense predictions, traditional Channel Attention is sub-optimal as it collapses spatial dimensions via Global Average Pooling. Instead, we introduce Coordinate Attention (CA), which factorizes 2D spatial pooling into two 1D directional encoding operations. For the c-th channel of $\mathcal{F}_{\text{cat}}$ at coordinates (h, w), the aggregated features along height and width are computed as:

$$z_c^h(h) = \frac{1}{W}\sum_{i=0}^{W-1}\mathcal{F}_{\text{cat}}^{(c)}(h, i), \quad z_c^w(w) = \frac{1}{H}\sum_{j=0}^{H-1}\mathcal{F}_{\text{cat}}^{(c)}(j, w) \tag{5}$$

These direction-aware vectors are concatenated, transformed, and split to generate attention weights, achieving feature refinement without sacrificing positional precision. The refined concatenated feature $\tilde{\mathcal{F}}$ is then fed into a lightweight Spatial Selector. This module generates a mutually exclusive 3D weight map $\mathcal{W} \in \mathbb{R}^{3\times H\times W}$ (where $\mathcal{W} = [\omega_4, \omega_7, \omega_{11}]$) via a pixel-wise Softmax operation:

$$[\omega_4, \omega_7, \omega_{11}] = \text{Softmax}\big(\text{Conv}_{1\times1}(\text{ReLU}(\text{Conv}_{3\times3}(\tilde{\mathcal{F}})))\big) \tag{6}$$

where ω_l indicates the dynamic aggregation weight of the l-th layer at a specific pixel ($\sum_l \omega_l = 1$). The preprocessed single-layer features f_l are then spatially fused and projected to yield the predicted feature $\mathcal{P} = \mathit{\Pi}\Big(\sum_{l\in\{4,7,11\}} \omega_l \odot f_l\Big)$.

3.5 Decoupled Multi-Scale Mask-Aware Optimization

To compel the mapping network to finely align local microscopic textures while robustly matching global macroscopic structures, we propose a Decoupled Multi-Scale Loss governed by a divide-and-conquer strategy.

Feature Decoupling and Alignment. Prior to loss computation, the unified predicted feature $\mathcal{P}$ is decoupled back into three distinct representation sub-spaces along the multi-scale spectrum (ranging from high-frequency local textures to low-frequency global structures, each with 768 dimensions). Given the massive absolute numerical discrepancies between RGB and 3D normal domains, we discard magnitude-sensitive distance metrics and rely exclusively on Cosine Distance as the core alignment objective:

$$\mathcal{L}_{\text{align}}(P, T) = 1 - \frac{P \cdot T}{\|P\|_2\|T\|_2} \tag{7}$$

where $P, T \in \mathbb{R}^{768}$ denote the predicted and target feature vectors at a specific pixel. This compels the network to align strictly based on directional consistency, circumventing gradient collapse caused by cross-modal amplitude variations.

Mask-Aware Regional Focusing. To prevent massive irrelevant backgrounds (*e.g.* conveyors) from dominating the loss, we introduce a flattened binary spatial mask $M \in \{0,1\}^N$ (where $N = H \times W$). For the real stream, adaptive focusing is achieved by computing the masked average ($\mathcal{L}^R_{\text{layer}}$). Conversely, for the estimated stream where precise masks are unavailable, we default to global spatial averaging ($\mathcal{L}^E_{\text{layer}}$):

$$\mathcal{L}^R_{\text{layer}} = \frac{\sum_{i=1}^{N} \left(\mathcal{L}^{(i)}_{\text{align}} \cdot M^{(i)} \right)}{\sum_{i=1}^{N} M^{(i)} + \epsilon}, \quad \mathcal{L}^E_{\text{layer}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}^{(i)}_{\text{align}} \tag{8}$$

where $\epsilon = 10^{-8}$, superscript (i) denotes the i-th pixel index, and the binary mask M is deterministically computed from the raw 3D point cloud to precisely isolate the foreground object.

Hierarchical Weighted Fusion. Deep features exhibit high spatial translation invariance; enforcing strict pixel-wise alignment at semantic levels easily triggers overfitting. Thus, for both the real (R) and estimated (E) streams, we introduce an asymmetric descending penalty strategy based on their respective layer-wise losses:

$$\mathcal{L}^S_{\text{total}} = \alpha \mathcal{L}^S_{\text{L4}} + \beta \mathcal{L}^S_{\text{L7}} + \gamma \mathcal{L}^S_{\text{L11}}, \quad S \in \{R, E\} \tag{9}$$

where weights satisfy $\alpha > \beta > \gamma$. This strategically anchors macroscopic structural comprehension (via a smaller γ for deep layers) while intensely directing focal attention towards critical localized micro-defects (via a larger α for shallow layers). The overall optimization objective of our framework is the sum of both streams: $\mathcal{L} = \mathcal{L}^R_{\text{total}} + \mathcal{L}^E_{\text{total}}$.

3.6 Cross-Modal Multiplicative Anomaly Scoring

During inference, the framework yields four distinct anomaly distance maps: 2D real Ψ^R_{2D}, 2D estimated Ψ^E_{2D}, 3D real Ψ^R_{3D}, and 3D estimated Ψ^E_{3D}. To obtain comprehensive modality-specific responses, we first perform a weighted summation of the real and estimated streams within both the 2D and 3D modalities:

$$\Psi_{\text{2D}} = \Psi^R_{\text{2D}} + \lambda_1 \Psi^E_{\text{2D}}, \quad \Psi_{\text{3D}} = \Psi^R_{\text{3D}} + \lambda_2 \Psi^E_{\text{3D}} \tag{10}$$

where λ_1 and λ_2 are balancing coefficients. Subsequently, the overall anomaly score map Ψ is computed via synergistic cross-modal fusion:

$$\Psi = \Psi_{\text{2D}} \odot \Psi_{\text{3D}} \tag{11}$$

where $\odot$ denotes the Hadamard product. This multiplicative fusion acts as a stringent spatial filter: a region is flagged as defective only if both modalities exhibit high anomalous responses concurrently, thereby significantly suppressing isolated modality-specific noise and minimizing false positives.

4 Experiments

4.1 Implementation Setup

Datasets and Protocol. We evaluate our framework on two standard anomaly detection datasets: MVTec 3D-AD [2] and EyeCandies [3]. Following the standard few-shot protocol, models are trained exclusively on $k \in \{1, 2, 4\}$ anomaly-free samples per category. To ensure a standardized and fully reproducible evaluation, we deterministically select the first k normal images from each class to construct the training subset.

Baselines. We compare our approach against eight representative state-of-the-art methods: (1) **Patchcore+FPFH** [11], concatenating deep 2D and handcrafted 3D features; (2) **CIF** [20], a cross-modal information fusion strategy; (3) **AST** [26], an asymmetric teacher-student architecture; (4) **EasyNet** [4], an efficient network tailored for industrial inspection; (5) **ShapeGuided** [6], explicitly leveraging 3D shape priors; (6) **M3DM** [34], a memory-bank-based multi-modal fusion model; (7) **CFM** [7], a memory-efficient 2D-3D feature alignment approach; and (8) **MAFR** [1], a multi-modal feature representation and alignment network.

Evaluation Metrics. Following standard protocols, we employ Image-level AUROC (I-AUROC) for anomaly classification, alongside Pixel-level AUROC (P-AUROC) and Area Under Per-Region Overlap (AUPRO) for fine-grained localization. In our main baseline comparisons, we report I-AUROC and AUPRO integrated up to a 30% False Positive Rate (FPR). For a comprehensive analysis in the ablation studies, we additionally report P-AUROC and AUPRO evaluated at stricter FPR thresholds of 1%, 5%, and 10%. Higher scores across all metrics denote superior performance.

Implementation Details. Experiments utilize an RTX 5090 (32GB) GPU. For augmentation, Stable Diffusion v2.1 is LoRA-finetuned (rank 16, 1,000 steps), and Marigold [15] serves as the normal estimator Φ_{diff}. Following [7], a frozen DINO ViT-B/8 extracts 2D and 3D representations [18] from 224×224 inputs. Features from layers 4, 7, and 11 are bicubically up-sampled and optimized via AdamW for 3,000 steps (batch size 2/stream, Cosine Annealing LR 10^{-4} to 10^{-6}, weight decay 10^{-4}). Additive Gaussian noise ($\sigma = 0.01$) is injected into RGB inputs to mitigate overfitting. Multi-scale loss weights (Eq. 9) are set to $\alpha = 1.2, \beta = 1.0, \gamma = 0.8$. Inference balancing coefficients (λ_1, λ_2) are 0.1.

4.2 Quantitative Comparison

Performance on MVTec 3D-AD. Table 1 shows our framework consistently outperforms all baselines across few-shot settings. In the 1-shot scenario, it achieves 79.6% I-AUROC and 94.2% AUPRO@30%, substantially exceeding the second-best M3DM. Under the 4-shot setting, performance scales to 87.1% I-AUROC and 95.8% AUPRO@30%, significantly surpassing MAFR. Notably, it demonstrates remarkable robustness on geometrically complex categories (*e.g.*,

Table 1: Few-shot performance comparison on the MVTec 3D-AD dataset. The metrics reported are I-AUROC / AUPRO@30%. The best results are highlighted in **bold**, and the second-best are underlined.

Setting	Type	Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
1-shot	TF	Patchcore+FPFH [11]	62.2/92.8	53.4/76.8	54.0/96.7	55.9/92.8	54.7/84.6	63.3/71.9	49.6/95.9	60.7/96.1	88.8/90.8	56.6/84.1	59.9/88.3
		CIF [20]	78.9/85.2	**68.7**/79.5	72.2/95.6	81.2/86.2	62.9/82.0	**72.8**/70.3	83.9/93.3	60.4/94.2	83.5/88.1	55.8/86.4	72.0/86.1
	TB	AST [26]	70.7/75.9	42.2/73.3	54.8/88.0	49.0/60.2	53.8/79.4	46.4/44.0	51.9/84.0	49.7/85.9	72.0/75.8	41.9/74.0	53.2/74.0
		EasyNet [4]	61.4/79.6	21.2/75.1	52.0/91.0	75.9/69.8	56.5/85.8	62.8/49.4	65.7/69.0	63.0/88.1	94.6/71.8	47.7/75.4	60.1/75.5
		ShapeGuided [6]	65.9/95.9	44.4/71.6	62.3/93.5	93.8/**94.1**	59.3/86.4	57.6/63.8	67.6/94.0	42.8/96.3	93.3/88.8	62.9/90.1	65.0/87.4
		M3DM [34]	87.8/95.3	64.1/81.5	78.0/97.2	92.7/90.3	**64.2**/81.6	65.3/82.0	75.5/94.0	**79.8**/94.8	85.8/95.2	45.4/89.8	73.9/90.2
		CFM [7]	88.1/95.4	54.0/74.0	62.5/96.5	96.4/92.4	60.2/**91.4**	68.7/89.3	61.9/93.8	56.0/95.6	90.9/95.7	58.8/88.9	69.8/91.4
		MAFR [1]	86.7/95.0	48.9/78.3	70.1/96.9	**96.7**/92.2	56.9/88.0	59.9/**89.7**	78.4/96.6	67.3/96.8	95.3/96.8	**64.0**/91.8	72.4/92.2
		Ours	**97.0**/**97.0**	63.8/**88.0**	**92.0**/**98.1**	92.3/93.2	59.6/88.5	69.9/89.4	**85.6**/**97.9**	77.4/**98.0**	**97.1**/**97.6**	61.4/**93.8**	**79.6**/**94.2**
2-shot	TF	Patchcore+FPFH [11]	63.2/94.6	47.7/76.5	55.4/96.7	64.5/93.4	58.3/85.0	61.1/67.4	55.0/96.2	56.6/96.6	88.2/91.0	64.4/88.3	61.4/88.6
		CIF [20]	85.3/87.1	62.9/79.5	74.0/95.8	72.2/87.1	64.6/82.1	79.5/75.3	77.6/93.7	66.9/94.9	86.1/88.7	62.7/87.7	73.2/87.2
	TB	AST [26]	71.9/75.9	43.4/74.0	54.5/87.8	50.8/62.2	53.7/79.5	46.1/43.6	51.6/83.7	50.4/85.6	75.8/76.5	40.2/72.8	53.8/74.2
		EasyNet [4]	47.6/77.8	**76.1**/62.9	52.6/92.6	60.2/59.1	31.7/58.8	52.3/57.2	71.9/21.1	76.1/15.2	61.2/43.1	51.2/4.7	58.1/49.3
		ShapeGuided [6]	47.9/96.6	46.0/73.2	60.5/96.5	95.9/**95.5**	55.3/86.5	50.2/71.4	69.7/95.3	41.3/96.3	93.6/89.3	**79.3**/91.3	64.0/89.2
		M3DM [34]	91.8/95.5	57.0/82.9	79.8/97.2	94.5/88.0	61.4/87.0	79.5/79.6	79.2/95.1	75.1/94.2	92.8/95.5	54.1/91.0	76.5/90.6
		CFM [7]	88.9/95.9	53.3/76.2	69.6/97.7	94.5/92.5	70.1/**93.0**	77.6/**92.6**	63.9/94.2	63.8/97.3	91.5/96.3	57.2/90.7	73.0/92.4
		MAFR [1]	91.8/95.7	49.8/79.7	69.8/97.1	**96.9**/92.0	**70.5**/89.8	71.4/92.2	81.2/97.0	72.8/97.1	95.3/96.9	66.9/94.2	76.6/93.2
		Ours	**96.4**/**97.2**	62.3/**89.1**	**93.0**/**98.2**	93.5/93.3	58.8/89.3	**87.7**/**92.6**	**86.3**/**97.8**	**83.5**/**98.1**	**97.6**/**97.7**	70.7/**94.8**	**83.0**/**94.8**
4-shot	TF	Patchcore+FPFH [11]	52.3/95.7	54.0/79.7	59.2/97.2	62.9/95.1	59.4/87.5	58.2/75.3	61.6/96.5	70.5/97.3	91.8/91.2	73.5/88.0	64.3/90.4
		CIF [20]	91.8/92.9	70.0/83.3	77.5/96.9	83.4/86.4	70.2/84.7	79.4/84.1	85.6/95.2	75.0/95.5	89.8/89.3	53.4/88.2	77.6/89.6
	TB	AST [26]	70.1/74.7	42.9/73.7	55.7/87.7	51.8/61.3	54.0/79.6	46.6/41.3	52.0/84.3	49.8/85.9	72.6/75.9	39.8/74.2	53.5/73.9
		EasyNet [4]	67.5/70.7	36.3/13.8	54.7/86.9	69.1/72.0	72.4/39.3	50.2/53.6	74.3/86.2	63.1/90.2	44.9/15.5	53.7/66.3	58.6/56.5
		ShapeGuided [6]	65.4/**97.3**	48.8/78.9	73.1/97.3	96.5/**95.4**	69.8/90.4	59.1/83.6	68.1/95.7	49.9/97.5	92.2/89.6	75.1/92.1	69.8/91.8
		M3DM [34]	**98.6**/96.1	68.5/87.2	83.7/97.3	93.8/90.5	59.6/86.5	**87.8**/86.6	85.3/96.4	70.5/94.6	89.3/95.5	56.1/93.1	79.3/92.4
		CFM [7]	95.6/96.7	62.2/82.9	86.1/98.0	96.7/93.3	**84.3**/**94.2**	74.6/93.2	72.0/96.2	75.8/97.7	95.7/96.8	65.2/93.3	80.8/93.2
		MAFR [1]	94.7/96.7	53.4/84.5	92.5/97.8	**96.9**/91.8	81.2/92.2	82.2/**95.2**	85.3/97.5	77.6/97.3	**97.3**/96.9	**79.4**/**96.1**	84.1/94.1
		Ours	97.4/**97.3**	**72.2**/**92.2**	**94.9**/**98.2**	95.8/93.7	74.6/94.1	85.9/93.0	**93.6**/**98.2**	**91.2**/**98.2**	96.7/**97.6**	68.6/95.9	**87.1**/**95.8**

Table 2: Few-shot performance comparison on the EyeCandies dataset. The metrics reported are I-AUROC / AUPRO@30%. The best results are highlighted in **bold**, and the second-best are underlined.

Setting	Type	Method	Can. C.	Cho. C.	Cho. P.	Conf.	Gum. B.	Haz. T.	Lic. S.	Lollip.	Marsh.	Pep. C.	Mean
1-shot	TF	CIF [20]	38.1/88.4	**81.8**/71.4	70.2/56.3	84.0/86.9	59.6/63.9	59.7/50.6	58.1/55.7	68.6/85.8	89.1/65.7	85.4/67.1	69.5/69.2
	TB	M3DM [34]	36.2/86.8	66.9/82.5	73.1/70.6	84.8/94.2	71.3/74.9	50.2/54.8	57.4/71.0	59.9/84.2	60.3/89.8	81.0/**90.0**	64.1/79.9
		CFM [7]	49.6/90.1	55.8/83.8	61.4/70.1	87.0/90.7	65.4/79.1	73.4/56.4	55.5/61.7	52.2/85.5	67.5/79.0	64.2/78.9	63.2/77.5
		MAFR [1]	**50.1**/91.4	64.8/82.8	76.2/60.9	86.9/**94.6**	38.9/78.3	70.9/53.9	48.6/63.1	58.3/83.9	92.5/88.4	59.4/80.3	64.7/77.8
		Ours	42.1/**92.6**	57.9/**89.3**	**88.8**/**79.9**	**90.6**/90.5	**84.1**/**84.3**	**74.7**/**66.5**	**81.0**/**82.5**	**69.9**/**87.1**	**94.9**/**92.7**	**87.8**/89.4	**77.2**/**85.5**
2-shot	TF	CIF [20]	41.8/87.0	79.0/76.9	78.6/57.9	**94.1**/85.5	71.3/61.0	70.9/52.4	55.7/59.9	66.4/86.8	84.5/71.8	**93.6**/74.3	73.6/71.3
	TB	M3DM [34]	38.9/81.1	67.5/84.4	81.1/68.6	92.3/**97.3**	61.7/76.1	53.6/57.6	59.4/72.8	64.0/85.6	76.5/89.3	86.9/**90.8**	68.2/80.4
		CFM [7]	49.8/91.3	**84.3**/85.2	74.1/74.3	87.7/92.3	77.1/79.6	64.2/**67.7**	59.8/68.5	54.0/83.9	86.6/91.0	63.8/83.4	70.1/81.7
		MAFR [1]	**52.0**/92.0	70.2/84.5	77.1/69.0	87.5/95.2	64.1/80.8	**75.7**/59.1	53.3/65.0	62.3/84.0	91.2/90.9	75.4/86.8	70.9/80.7
		Ours	42.9/**93.3**	80.5/**90.9**	**90.2**/**81.0**	88.0/92.0	**84.8**/**84.8**	75.0/65.8	**81.1**/**80.9**	**72.2**/**88.5**	**98.6**/**93.5**	88.6/87.7	**80.2**/**85.8**
4-shot	TF	CIF [20]	48.2/87.4	81.8/78.1	74.6/58.4	**97.9**/88.9	61.9/68.0	69.8/58.1	73.3/63.9	61.8/88.1	94.2/86.0	87.4/78.8	75.1/75.6
	TB	M3DM [34]	42.4/82.4	74.9/84.9	78.7/72.1	91.4/**96.8**	70.2/80.2	53.4/61.7	80.2/81.6	67.8/88.6	86.6/**94.6**	**90.6**/**92.1**	73.6/83.5
		CFM [7]	47.5/90.9	85.3/85.5	76.5/72.6	90.6/94.4	76.9/80.0	62.6/68.5	65.9/73.5	**70.6**/85.7	92.2/93.2	78.1/87.1	74.6/83.1
		MAFR [1]	**49.9**/92.6	74.7/87.3	81.1/70.8	87.5/96.0	79.3/84.8	77.6/62.4	61.3/73.3	68.6/84.9	96.0/93.1	80.5/90.2	75.7/83.5
		Ours	45.0/**94.9**	**91.7**/**91.3**	**95.0**/**81.7**	88.6/92.6	**80.6**/**86.9**	**85.4**/**71.6**	**85.6**/**85.8**	67.7/**89.0**	**98.1**/94.3	89.1/89.1	**82.7**/**87.7**

Bagel, *Rope*), securing near-perfect localization. These consistent gains highlight the superior cross-modal alignment and generalization enabled by our dual-stream representations.

Performance on EyeCandies. As shown in Table 2, our approach establishes a new state-of-the-art across all shot settings, validating its strong generalizability. In the 1-shot setting, we achieve 77.2% I-AUROC and 85.5% AUPRO@30%, eclipsing the prior best CIF (69.5% and 69.2%, respectively). This superiority persists at 4-shot (82.7% I-AUROC), demonstrating precise fine-grained anomaly

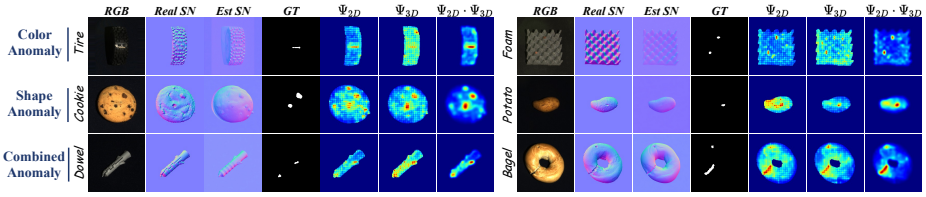


Fig. 3: Qualitative evaluation of the complementarity between 2D and 3D modalities. We visualize the predicted anomaly maps across three distinct defect categories: 2D color-only anomalies, 3D shape-only anomalies, and combined anomalies. Synergistically fusing these modality-specific predictions effectively captures all defect types while suppressing background noise.

localization under severe data scarcity. Overall, results on both benchmarks confirm that our diffusion-enhanced dual-stream architecture effectively addresses the fundamental bottlenecks of few-shot 3D anomaly detection.

4.3 Qualitative Analysis

Complementarity of 2D and 3D Modalities. Figure 3 visualizes predicted anomaly maps across three distinct defect categories: 2D color-only, 3D shape-only (*e.g.* bumps or dents), and combined anomalies. As illustrated, 2D-focused predictions effectively isolate subtle textural variations, such as surface discolorations or contaminations, which lack geometric deformation. Conversely, 3D-focused predictions precisely highlight structural irregularities that are often imperceptible in the RGB domain. By synergistically fusing these modality-specific outputs, our combined anomaly map robustly localizes complex defects while substantially suppressing modality-specific background noise. This underscores the strong complementarity of our learned 2D and 3D representations.

Qualitative Comparison with Baselines. To further validate our approach, Fig. 4 visualizes anomaly score overlays for selected categories from the MVTec 3D-AD and EyeCandies datasets against state-of-the-art methods. While existing baselines frequently trigger false alarms in normal regions or struggle to cleanly delineate subtle defect boundaries, our method yields sharp, highly localized anomaly masks that align tightly with the ground truth. Specifically, the diffusion-enhanced structural priors enable our model to maintain high confidence precisely on defective pixels without bleeding into surrounding normal areas. This superior qualitative performance corroborates our quantitative gains, highlighting the robustness and precision of our framework under challenging few-shot constraints.

4.4 Ablation Study

Effectiveness of Core Components. We decouple the real stream, estimated (est) stream, and Feature Mapper under the 4-shot setting (Tab. 3). Relying

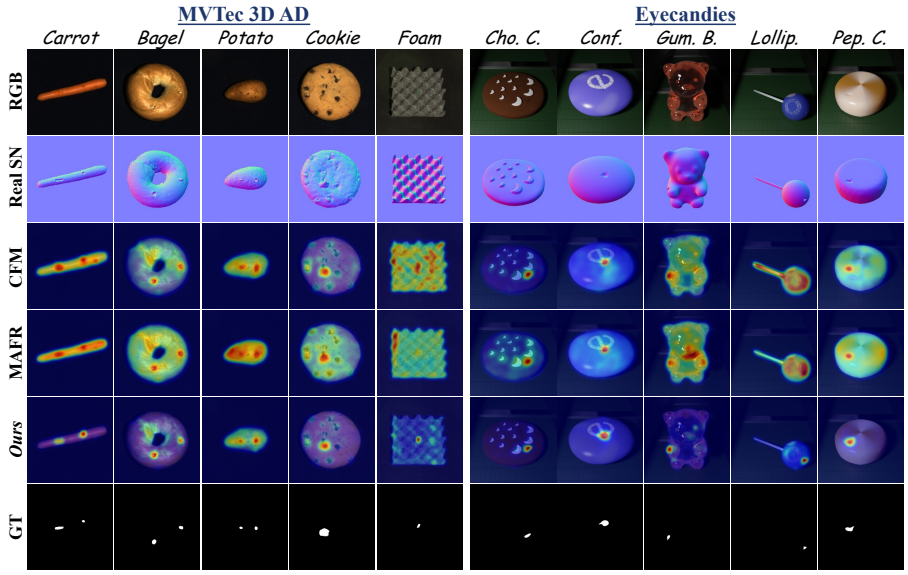


Fig. 4: Qualitative comparison of anomaly score overlays between our proposed method and state-of-the-art baselines on the MVTec 3D-AD and EyeCandies datasets. Under challenging few-shot constraints, our approach effectively suppresses false positives in normal regions and delineates subtle defect boundaries with significantly higher precision than existing methods.

Table 3: Ablation study on the core components of the proposed framework under the **4-shot** setting. The checkmark (✓) indicates the inclusion of a specific component. When the proposed Feature Mapper is omitted, a standard MLP is utilized for fusion. The best results are highlighted in **bold**, and the second-best are underlined.

Real Stream	Est Stream	Feature Mapper	I-AUROC	P-AUROC	AUPRO@30%	AUPRO@10%	AUPRO@5%	AUPRO@1%
✓	✓		0.872	0.988	0.955	0.877	0.792	0.399
✓		✓	**0.873**	**0.989**	0.956	0.880	0.795	0.400
	✓	✓	0.813	0.979	0.932	0.831	0.732	0.350
✓	✓	✓	0.871	**0.989**	**0.958**	**0.886**	**0.806**	**0.410**

on the est stream yields the weakest I-AUROC (0.813), whereas the real stream alone achieves 0.873. Crucially, integrating both unlocks fine-grained localization, boosting AUPRO@1% from 0.400 (real-only) to 0.410, confirming that generative priors supplement subtle defect boundaries. Replacing our Feature Mapper with a standard MLP degrades all localization metrics, highlighting the necessity of our hierarchical gating for aligning heterogeneous 2D-3D semantic spaces.

Influence of Estimated Stream Weights. We vary the estimated stream balancing coefficients (λ_1, λ_2) from 0.0 to 1.0 (Fig. 5). Incorporating a moderate generative prior strictly enhances performance: increasing weights from 0.0 to 0.1 improves the 2-shot I-AUROC on EyeCandies (78.46% to 80.20%) and the 4-shot AUPRO@30% on MVTec 3D-AD (95.63% to 95.83%). Conversely, over-

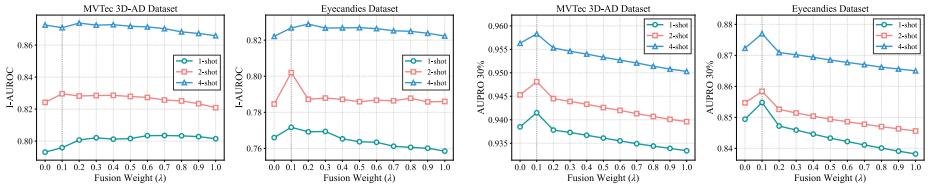


Fig. 5: Parameter sensitivity analysis regarding the weights of the estimated stream (λ_1, λ_2) on MVTec 3D-AD and EyeCandies datasets. The plots illustrate the impact on I-AUROC and AUPRO@30% metrics as the weights vary from 0.0 to 1.0 with a step size of 0.1. From left to right: (a) I-AUROC on MVTec 3D-AD, (b) I-AUROC on EyeCandies, (c) AUPRO@30% on MVTec 3D-AD, and (d) AUPRO@30% on EyeCandies.

Table 4: Ablation study on the fusion strategy of anomaly maps under the **4-shot** setting. The best results are highlighted in **bold**, and the second-best are underlined.

Anomaly Map	I-AUROC	P-AUROC	AUPRO@30%	AUPRO@10%	AUPRO@5%	AUPRO@1%
Ψ_{2D}	0.807	0.985	0.948	0.864	0.773	0.372
Ψ_{3D}	0.848	0.984	0.945	0.855	0.766	0.381
$\Psi_{2D} + \Psi_{3D}$	**0.875**	0.986	0.952	0.874	0.791	0.401
$\max(\Psi_{2D}, \Psi_{3D})$	0.861	0.985	0.950	0.869	0.785	0.394
$\Psi_{2D} \odot \Psi_{3D}$	0.871	**0.989**	**0.958**	**0.886**	**0.806**	**0.410**

reliance on estimated features introduces generative noise (*e.g.* at a weight of 1.0, 1-shot I-AUROC on EyeCandies drops to 75.85%). Thus, an empirical optimal balance is robustly achieved at 0.1 across diverse few-shot settings.

Anomaly Map Fusion Strategy. We evaluate fusion operations for anomaly maps (Ψ_{2D} and Ψ_{3D}) in Tab. 4. Unimodal detection proves insufficient. Among multi-modal strategies, element-wise multiplication ($\Psi_{2D} \odot \Psi_{3D}$) delivers the best localization (0.989 P-AUROC, 0.958 AUPRO@30%). Multiplication acts as a strict spatial filter, suppressing isolated modality-specific false positives by retaining high scores only when both modalities indicate anomalies. Although addition ($\Psi_{2D} + \Psi_{3D}$) yields a marginally higher I-AUROC (0.875), multiplication ensures significantly more robust pixel-level localization under strict false-positive constraints (*e.g.* AUPRO@1% and 5%).

5 Conclusion

We propose CMDS-AD, a Cross-Modal Dual-Stream framework for few-shot anomaly detection. Repurposing a diffusion estimator as a non-linear low-pass filter establishes a low-frequency auxiliary stream that enhances the real stream, isolating micro-defects from macroscopic structures. A Coordinate-Aware Hierarchical Feature Mapper adaptively bridges cross-modal semantic gaps, and a multiplicative scoring mechanism ($\Psi_{2D} \odot \Psi_{3D}$) suppresses modality-specific noise. Experiments on MVTec 3D-AD and EyeCandies show CMDS-AD overcomes data scarcity, achieving state-of-the-art 1- to 4-shot anomaly localization.

Acknowledgements

This research work was financially supported in part by the Guangdong Major Project of Basic Research under Grant 2023B0303000009, in part by the NSFC Youth Fund Project under Grant 62403326, in part by the Shenzhen Fundamental Research Fund under Grant JCYJ20230808105212023, in part by the Research Team Cultivation Program of ShenZhen University under Grant 2023JCT004, and in part by the Shenzhen University 2035 Program for Excellent Research under Grant 00000224.

References

1. Ali, U., Zia, A., Rehman, A., Ramzan, U., Hassan, Z., Sattar, T., Wang, J., Xiang, W.: 2d-3d feature fusion via cross-modal latent synthesis and attention-guided restoration for industrial anomaly detection. In: 2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1–8. IEEE (2025) 4, 10, 11
2. Bergmann, P., Jin, X., Sattlegger, D., Steger, C.: The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. arXiv preprint arXiv:2112.09045 (2021) 10
3. Bonfiglioli, L., Toschi, M., Silvestri, D., Fioraio, N., De Gregorio, D.: The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In: Proceedings of the Asian Conference on Computer Vision. pp. 3586–3602 (2022) 10
4. Chen, R., Xie, G., Liu, J., Wang, J., Luo, Z., Wang, J., Zheng, F.: Easynet: An easy network for 3d industrial anomaly detection. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7038–7046 (2023) 10, 11
5. Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S.: Perception prioritized training of diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11472–11481 (2022) 4
6. Chu, Y.M., Chieh, L., Hsieh, T.I., Chen, H.T., Liu, T.L.: Shape-guided dual-memory learning for 3d anomaly detection (2023) 2, 3, 4, 10, 11
7. Costanzino, A., Ramirez, P.Z., Lisanti, G., Di Stefano, L.: Multimodal industrial anomaly detection by crossmodal feature mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17234–17243 (2024) 3, 4, 10, 11
8. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9737–9746 (2022) 2
9. Fang, Z., Wang, X., Li, H., Liu, J., Hu, Q., Xiao, J.: Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17481–17490 (2023) 2, 4
10. Gu, Z., Zhang, J., Liu, L., Chen, X., Peng, J., Gan, Z., Jiang, G., Shu, A., Wang, Y., Ma, L.: Rethinking reverse distillation for multi-modal anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 8445–8453 (2024) 3, 4

11. Horwitz, E., Hoshen, Y.: Back to the feature: classical 3d features are (almost) all you need for 3d anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2968–2977 (2023) 10, 11
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022) 4
13. Huang, C., Guan, H., Jiang, A., Zhang, Y., Spratling, M., Wang, Y.F.: Registration based few-shot anomaly detection. In: European conference on computer vision. pp. 303–319. Springer (2022) 2, 4
14. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19606–19616 (2023) 2, 4
15. Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A., Schindler, K.: Marigold: Affordable adaptation of diffusion-based image generators for image analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 10
16. Lee, Y., Jang, S., Yoon, H.: Anople: Few-shot anomaly detection via bi-directional prompt learning with only normal samples. arXiv e-prints pp. arXiv–2408 (2024) 2, 4
17. Li, W., Xu, X., Gu, Y., Zheng, B., Gao, S., Wu, Y.: Towards scalable 3d anomaly detection and localization: A benchmark via 3d anomaly synthesis and a self-supervised learning network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22207–22216 (2024) 2
18. Li, Y., Liu, F., Liao, J., Tian, S., Foo, C.S., Yang, X.: Find: Few-shot anomaly inspection with normal-only multi-modal data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23290–23299 (2025) 10
19. Liang, H., Xie, G., Hou, C., Wang, B., Gao, C., Wang, J.: Look inside for more: Internal spatial modality perception for 3d anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5146–5154 (2025) 2
20. Lin, Y., Yan, H., Tong, X., Chang, Y., Wang, H., Zhou, Z., Gao, S., Wang, Y., Zhang, W.: Commonality in few: Few-shot multimodal anomaly detection via hypergraph-enhanced memory. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 7015–7023 (2026) 4, 10, 11
21. Liu, C., Chu, Y.M., Hsieh, T.I., Chen, H.T., Liu, T.L.: Learning diffusion models for multi-view anomaly detection. In: European Conference on Computer Vision. pp. 328–345. Springer (2024) 2
22. Long, K., Xie, G., Ma, L., Liu, J., Lu, Z.: Revisiting multimodal fusion for 3d anomaly detection from an architectural perspective. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 12273–12281 (2025) 2
23. Luo, X., Xie, Y., Qu, Y., Fu, Y.: Skipdiff: Adaptive skip diffusion model for high-fidelity perceptual image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4017–4025 (2024) 4
24. Lyu, S., Mo, D., keung Wong, W.: Reb: Reducing biases in representation for industrial anomaly detection. Knowledge-Based Systems **290**, 111563 (2024) 4
25. Ristea, N.C., Madan, N., Ionescu, R.T., Nasrollahi, K., Khan, F.S., Moeslund, T.B., Shah, M.: Self-supervised predictive convolutional attentive block for anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13576–13586 (2022) 2
26. Rudolph, M., Wehrbein, T., Rosenhahn, B., Wandt, B.: Asymmetric student-teacher networks for industrial anomaly detection. In: Proceedings of the

IEEE/CVF winter conference on applications of computer vision. pp. 2592–2602 (2023) 3, 10, 11
27. Si, C., Huang, Z., Jiang, Y., Liu, Z.: Freeu: Free lunch in diffusion u-net. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4733–4743 (2024) 4
28. Sui, W., Lichau, D., Lefèvre, J., Phelippeau, H.: Incomplete multimodal industrial anomaly detection via cross-modal distillation. Information Fusion p. 103572 (2025) 2
29. Tang, N., Luo, X., Cheng, Z., Zhou, L., Zhang, D., Qu, Y.: Diffusion once and done: Degradation-aware lora for all-in-one image restoration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 9448–9456 (2026) 4
30. Tian, L., Zhao, H., Lu, R., Wang, R., Wu, Y., Wang, L., He, X., Liu, X.: Foct: Few-shot industrial anomaly detection with foreground-aware online conditional transport. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 6241–6249 (2024) 2, 4
31. Tien, T.D., Nguyen, A.T., Tran, N.H., Huy, T.D., Duong, S., Nguyen, C.D.T., Truong, S.Q.: Revisiting reverse distillation for anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24511–24520 (2023) 2
32. Trabucco, B., Doherty, K., Gurinas, M., Salakhutdinov, R.: Effective data augmentation with diffusion models. In: International Conference on Learning Representations. vol. 2024, pp. 14590–14612 (2024) 4
33. Tu, Y., Zhang, B., Liu, L., Li, Y., Zhang, J., Wang, Y., Wang, C., Zhao, C.: Self-supervised feature adaptation for 3d industrial anomaly detection. In: European conference on computer vision. pp. 75–91. Springer (2024) 2
34. Wang, Y., Peng, J., Zhang, J., Yi, R., Wang, Y., Wang, C.: Multimodal industrial anomaly detection via hybrid fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8032–8041 (2023) 2, 3, 4, 10, 11
35. Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 650–656 (2022) 4
36. Yan, X., Zhan, H., Zheng, C., Gao, J., Zhang, R., Cui, S., Li, Z.: Let images give you more: Point cloud cross-modal training for shape analysis. Advances in Neural Information Processing Systems **35**, 32398–32411 (2022) 2
37. Yang, X., Zhou, D., Feng, J., Wang, X.: Diffusion probabilistic model made slim. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 22552–22562 (2023) 4
38. Zavrtanik, V., Kristan, M., Skočaj, D.: Cheating depth: Enhancing 3d surface anomaly detection via depth simulation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2164–2172 (2024) 2
39. Zhang, H., Wang, Z., Zeng, D., Wu, Z., Jiang, Y.G.: Diffusionad: Norm-guided one-step denoising diffusion for anomaly detection. IEEE transactions on pattern analysis and machine intelligence (2025) 4
40. Zhang, X., Li, S., Li, X., Huang, P., Shan, J., Chen, T.: Destseg: Segmentation guided denoising student-teacher for anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3914–3923 (2023) 2
41. Zhou, Z., Wang, L., Fang, N., Wang, Z., Qiu, L., Zhang, S.: R3d-ad: Reconstruction via diffusion for 3d anomaly detection. In: European conference on computer vision. pp. 91–107. Springer (2024) 2

Supplementary Material for CMDS-AD: Cross-Modal Dual-Stream Decoupling for Few-Shot Anomaly Detection

A Extended Analysis on Strict FPR Thresholds

A.1 Evaluation under Strict Few-Shot Settings

In industrial anomaly detection, achieving precise defect localization with a low False Positive Rate (FPR) is crucial. However, current few-shot multi-modal anomaly detection literature predominantly reports the Area Under the Per-Region-Overlap curve up to a 30% FPR (AUPRO@30%). While this metric provides a general sense of localization capability, it may not fully reflect a model's reliability in real-world deployments. Specifically, in high-throughput manufacturing environments, allowing up to a 30% false positive rate would result in an unacceptable volume of normal items being flagged for review, incurring significant manual re-inspection costs and disrupting production efficiency.

To better align with practical requirements, it is necessary to evaluate models under stricter FPR thresholds. Evaluating at these constrained limits effectively isolates a model's ability to suppress background noise and handle ambiguous structural variations without triggering false alarms. In Table 5, we present the class-wise performance of our CMDS-AD framework on the MVTec 3D-AD dataset across 1-shot, 2-shot, and 4-shot settings, detailing AUPRO at 10%, 5%, and 1%. By providing these detailed metrics, we aim to offer a more practical reference baseline for future research focused on robust, low-FPR defect localization.

A.2 PRO Curve Dynamics: The Superiority of Multiplicative Fusion

To further elucidate the mechanisms behind our model's robustness under strict industrial constraints, we analyze the continuous PRO curve dynamics across different fusion strategies in Figure 6. We select four geometrically and texturally challenging classes (*Dowel*, *Rope*, *Bagel*, and *Carrot*) as representative cases, as they encompass a diverse spectrum of complex local curvatures, highly variable surface reflectivity, and non-deterministic structural boundaries.

As observed in the macroscopic view, at looser constraints approaching 30% FPR, all tested fusion strategies, including additive ($\Psi_{2D} + \Psi_{3D}$), maximum ($\max(\Psi_{2D}, \Psi_{3D})$), and multiplicative ($\Psi_{2D} \odot \Psi_{3D}$), yield ostensibly comparable and near-saturated performance. However, as the FPR threshold is restricted to the extreme low end (the highlighted $\leq 5\%$ zone), a distinct bifurcation occurs. Modality-specific noise, such as specular highlights in 2D (e.g., on the *Bagel* surface) or inherent sensor artifacts in 3D (e.g., on the thin structure of *Dowel*), causes the single-modality branches to experience a steep, cliff-like drop in precision. Without a mechanism to critically cross-reference conflicting signals, these

Table 5: Comprehensive performance of CMDS-AD under strict FPR thresholds on MVTec 3D-AD. We report AUPRO evaluated at integration limits of 10%, 5%, and 1% to demonstrate robustness in strict industrial scenarios. For each threshold across different shots, the best results are highlighted in **bold**, and the second-best are underlined.

Class	1-shot			2-shot			4-shot		
	AUPRO@10%	AUPRO@5%	AUPRO@1%	AUPRO@10%	AUPRO@5%	AUPRO@1%	AUPRO@10%	AUPRO@5%	AUPRO@1%
Bagel	0.915	0.842	0.427	0.920	0.851	0.436	**0.922**	**0.855**	**0.442**
Cable Gland	0.679	0.507	0.179	0.702	0.535	0.192	**0.776**	**0.636**	**0.268**
Carrot	0.942	0.886	0.472	0.944	0.890	0.478	**0.947**	**0.894**	**0.485**
Cookie	0.869	0.801	0.416	0.873	0.806	0.419	**0.877**	**0.810**	**0.429**
Dowel	0.657	0.461	0.150	0.681	0.497	0.166	**0.824**	**0.705**	**0.312**
Foam	0.734	0.624	0.285	0.802	0.704	0.343	**0.815**	**0.722**	**0.359**
Peach	0.938	0.876	0.455	0.935	0.871	0.444	**0.944**	**0.889**	**0.477**
Potato	0.939	0.879	0.451	0.944	0.888	0.475	**0.946**	**0.893**	**0.482**
Rope	0.929	0.868	0.455	**0.931**	**0.872**	**0.459**	0.929	0.867	0.456
Tire	0.818	0.692	0.301	0.848	0.733	0.330	**0.878**	**0.784**	**0.384**
MEAN	0.842	0.743	0.359	0.858	0.765	0.374	**0.886**	**0.805**	**0.409**

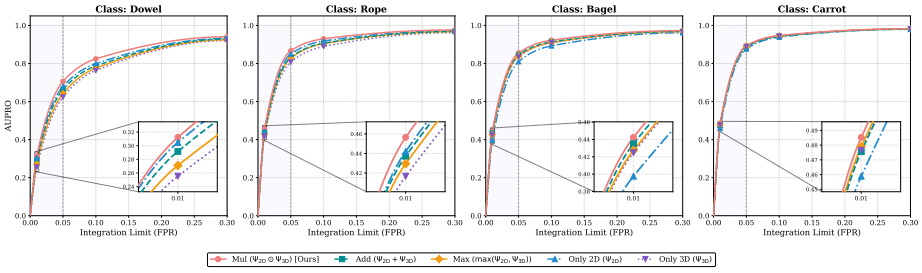


Fig. 6: PRO curve dynamics across different integration limits (FPR). We select four representative classes (*Dowel*, *Rope*, *Bagel*, *Carrot*) to illustrate the varying degradation behaviors of different fusion strategies. The inset panels provide a 10× magnified view specifically anchored at the strict 1% FPR threshold. Multiplicative fusion ($\Psi_{2D} \odot \Psi_{3D}$) consistently maintains top-tier stability, effectively resisting the unilateral noise that severely degrades single-stream, Additive, and Max fusion strategies.

traditional strategies are easily overwhelmed by localized but entirely normal variations.

Crucially, as highlighted in the inset zoom panels at exactly 1% FPR, both Additive and Max fusion strategies fail to suppress this degradation, as they are highly susceptible to confident false positives originating from any single noisy modality. In stark contrast, our Multiplicative fusion acts as a stringent logical spatial *AND* filter. A localized anomaly is only strongly activated if corroborated by both the 2D texture variation and the 3D structural deviation simultaneously. This explicit cross-modal verification effectively neutralizes unilateral noise, allowing CMDS-AD to maintain a resilient and dominant PRO curve trajectory even at ultra-low FPRs, ensuring that the model exclusively flags regions with genuine multimodal defect evidence rather than transient background artifacts.

Table 6: Comprehensive class-wise ablation of Core Components (4-shot, MVTec 3D-AD). Corresponding to Table 3 in the main text. The best results are highlighted in **bold**, and the second-best are underlined. Note the stable and saturated P-AUROC values across all configurations.

Class	w/o Est Stream			w/o Real Stream			w/o Feature Mapper			**Ours (Dual-Stream)**		
	PRO@30%	PRO@1%	P-AUC	PRO@30%	PRO@1%	P-AUC	PRO@30%	PRO@1%	P-AUC	PRO@30%	PRO@1%	P-AUC
Bagel	0.968	0.404	**0.994**	0.964	0.421	0.990	0.968	0.409	0.993	**0.972**	**0.442**	**0.994**
Cable Gland	0.916	0.254	0.972	0.907	0.259	0.969	0.907	0.251	0.969	**0.922**	**0.268**	**0.974**
Carrot	**0.982**	**0.485**	**0.998**	0.980	0.470	0.997	**0.982**	**0.485**	**0.998**	**0.982**	**0.485**	**0.998**
Cookie	**0.939**	**0.430**	**0.971**	0.876	0.290	0.948	0.930	0.417	0.968	0.936	0.429	0.970
Dowel	0.937	0.297	0.984	0.910	0.226	0.976	0.934	0.296	0.984	**0.940**	**0.312**	**0.985**
Foam	0.924	0.346	0.981	0.829	0.289	0.943	**0.929**	0.353	**0.982**	**0.929**	**0.359**	**0.982**
Peach	**0.981**	**0.477**	**0.998**	0.979	0.467	0.997	**0.981**	0.473	**0.998**	**0.981**	**0.477**	**0.998**
Potato	**0.982**	0.479	**0.998**	0.971	0.370	0.992	**0.982**	0.481	**0.998**	**0.982**	**0.482**	**0.998**
Rope	**0.976**	**0.458**	**0.997**	0.970	0.440	0.996	0.975	0.455	**0.997**	**0.976**	0.456	**0.997**
Tire	0.955	0.367	0.990	0.925	0.260	0.979	0.955	0.366	0.989	**0.958**	**0.384**	**0.991**
MEAN	0.956	0.400	0.988	0.931	0.349	0.979	0.954	0.399	0.988	**0.958**	**0.409**	**0.989**

A.3 Why Single-Stream Architecture Collapses at Low FPRs

The divergence in performance under strict FPRs can be deeply understood through the lens of our dual-stream architecture (Real Stream vs. Estimation Stream). As demonstrated in our ablation studies, relying exclusively on the Real Stream (which inherently contains both high and low-frequency spatial variations) allows the model to perform adequately at AUPRO@30% (0.9563). However, under the extreme pressure of AUPRO@1%, its performance collapses to 0.4002.

The underlying mechanism for this failure is the model's over-sensitivity to normal, localized high-frequency texture variations, which it misclassifies as micro-defects when forced to minimize false positives. By introducing the Estimation Stream—where the diffusion model acts as an advanced low-pass filter—we extract stabilized low-frequency priors. These priors serve as structural anchors. The explicit decoupling of high-frequency details (Real) and macro-structural semantics (Estimation) is the decisive factor that inhibits false positives and overcomes the data-scarcity bottleneck inherent in few-shot settings.

B Comprehensive Class-wise Ablation and P-AUROC Results

Due to space constraints in the main manuscript, Tables 3 and 4 reported the mean metrics across all classes. To provide a completely transparent view of our model's behavior, Tables 6 and 7 present the detailed class-wise breakdown for the core component and fusion strategy ablations, respectively (evaluated at 4-shot on MVTec 3D-AD).

Furthermore, these tables incorporate the pixel-level anomaly detection metric (P-AUROC). As observed in recent state-of-the-art literature, P-AUROC has largely saturated (consistently hovering between 98% and 99% across most configurations). Thus, while it is no longer a discriminative metric for assess-

Table 7: Comprehensive class-wise ablation of Fusion Strategies (4-shot, MVTec 3D-AD). Corresponding to Table 4 in the main text. The best results are highlighted in **bold**, and the second-best are underlined.

Class	Only 2D			Only 3D			Max			Add			**Mul (Ours)**		
	PRO@30	PRO@1	P-AUC	PRO@30	PRO@1	P-AUC	PRO@30	PRO@1	P-AUC	PRO@30	PRO@1	P-AUC	PRO@30	PRO@1	P-AUC
Bagel	0.961	0.397	0.991	0.966	0.424	0.990	0.967	0.427	0.991	0.968	0.434	0.992	**0.972**	**0.442**	**0.994**
Cable Gland	**0.931**	**0.288**	**0.977**	0.885	0.216	0.963	0.913	0.245	0.972	0.914	0.258	0.972	0.922	0.268	0.974
Carrot	0.979	0.458	0.996	0.981	0.476	**0.998**	0.981	0.479	**0.998**	0.981	0.475	0.997	**0.982**	**0.485**	**0.998**
Cookie	0.897	0.327	0.951	0.910	0.419	0.953	0.907	0.413	0.951	0.911	0.428	0.954	**0.936**	**0.429**	**0.970**
Dowel	0.932	0.304	0.982	0.920	0.255	0.979	0.924	0.271	0.980	0.928	0.291	0.981	**0.940**	**0.312**	**0.985**
Foam	0.912	0.326	0.975	0.909	0.289	0.977	0.920	0.348	0.978	**0.929**	0.355	0.981	**0.929**	**0.359**	**0.982**
Peach	0.975	0.423	0.995	**0.981**	0.476	**0.998**	0.980	0.471	**0.998**	**0.981**	0.474	**0.998**	**0.981**	**0.477**	**0.998**
Potato	0.974	0.410	0.994	**0.982**	**0.484**	**0.998**	**0.982**	0.481	**0.998**	0.981	0.476	0.997	**0.982**	0.482	**0.998**
Rope	0.972	0.442	**0.997**	0.961	0.416	0.995	0.967	0.429	0.996	0.968	0.437	0.996	**0.976**	**0.456**	**0.997**
Tire	0.948	0.341	0.987	0.945	0.354	0.987	0.951	0.372	0.988	0.955	0.380	0.989	**0.958**	**0.384**	**0.991**
MEAN	0.948	0.372	0.985	0.944	0.381	0.984	0.949	0.394	0.985	0.951	0.401	0.986	**0.958**	**0.409**	**0.989**

ing fine-grained localization superiority, we report it fully in this appendix to facilitate alignment with earlier literature.

More importantly, the granular data in Table 6 explicitly reveals the critical role of the dual-branch decoupling mechanism. For instance, when omitting the Estimation Stream (*w/o Est Stream*), the model still achieves a near-perfect P-AUROC (0.988) and a competitive AUPRO@30% (0.956), creating an illusion of high performance. However, its AUPRO@1% plummets to 0.400 across all classes. This validates our core hypothesis: without the low-frequency structural anchor provided by the diffusion model, the network inevitably overfits to normal high-frequency texture variations during few-shot learning, leading to a catastrophic collapse in precise localization when false positives are strictly penalized. The detailed data unequivocally demonstrates that our full CMDS-AD framework maintains exceptional stability across all diverse and geometrically challenging object categories.

C Mechanistic Validation and Robustness Analysis

C.1 Spectral Evidence for Frequency-Decoupled Priors

To directly substantiate the frequency-decoupling interpretation, Figure 7 visualizes the FFT spectra of real and estimated normals, while Table 8 reports the corresponding normalized spectral energy statistics. The estimated normals clearly suppress high-frequency responses and retain coarse structural components. Quantitatively, the High/Low energy ratio drops from 0.1251 to 0.0221 on normal samples and from 0.1293 to 0.0239 on anomalous samples. Together, these results support our view that the diffusion-based normal estimator acts as a non-linear low-pass prior: it preserves stable global structure while suppressing modality-specific high-frequency noise that would otherwise trigger false positives.

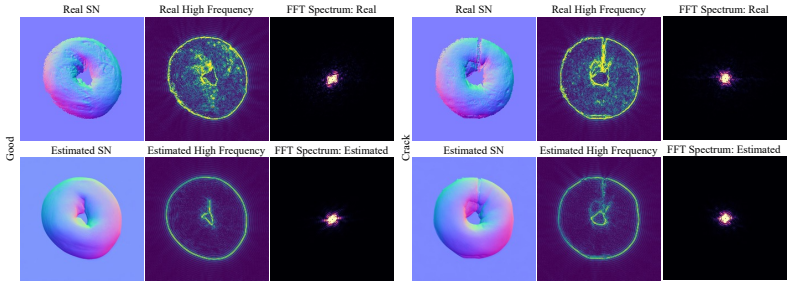


Fig. 7: FFT-based frequency analysis of real and estimated normals. Estimated normals suppress high-frequency responses while preserving dominant structural components.

Table 8: Spectral energy distribution of real and estimated normals. Estimated normals exhibit substantially lower high-frequency energy and a markedly smaller High/Low ratio.

Method	Normal Samples			Anomaly Samples		
	Low E.	High E.	H/L Ratio	Low E.	High E.	H/L Ratio
Real Normal	0.8888	0.1112	0.1251	0.8855	0.1145	0.1293
Estimated Normal	0.9784	0.0216	0.0221	0.9766	0.0234	0.0239

Table 9: Ablation isolating the architectural contribution under identical diffusion priors on MVTec 3D-AD (4-shot). "MLP" uses the same Stable Diffusion v2.1 augmentation and Marigold normal estimation as our method.

Method	I-AUROC	P-AUROC	PRO@30%	PRO@10%	PRO@5%	PRO@1%
CFM	80.8	98.5	94.2	84.3	73.9	35.2
MAFR	84.1	98.4	94.6	85.6	76.5	37.7
MLP	**87.2**	98.8	95.5	87.7	79.2	39.9
Ours	87.1	**98.9**	**95.8**	**88.6**	**80.6**	**41.0**

C.2 Isolating Architectural Gains Beyond Shared Diffusion Priors

The component ablations above validate the role of each module within CMDS-AD, but they do not fully separate architectural gains from the strength of the shared diffusion priors. To disentangle these factors, Table 9 compares our method against an "MLP" baseline that uses the same Stable Diffusion v2.1 augmentation and Marigold-estimated normals, but replaces the dual-stream design with a shallow multimodal predictor. Even under identical priors, CMDS-AD still delivers consistent gains, especially at strict operating points such as PRO@1% (+1.1 points over MLP). This shows that the benefit is not merely inherited from better priors; it is amplified by the proposed dual-stream architecture and its cross-modal interaction design.

Table 10: Sensitivity to the shared estimated-stream balancing weight w, where $\lambda_1 = \lambda_2 = w$. Best and second-best are highlighted in **bold** and underlined, respectively.

w	MVTec 3D-AD				EyeCandies			
	I-AUROC	P-AUROC	PRO@30%	PRO@1%	I-AUROC	P-AUROC	PRO@30%	PRO@1%
0.1	87.1	**98.9**	**95.8**	**41.0**	**82.7**	**97.4**	**87.7**	**30.4**
0.3	**87.3**	98.8	95.5	40.1	**82.7**	97.3	87.0	29.4
0.5	87.2	98.7	95.3	39.9	**82.7**	97.2	86.9	29.3
0.7	87.0	98.7	95.2	39.6	82.5	97.2	86.7	29.2
0.9	86.7	98.6	95.1	39.4	82.4	97.2	86.6	29.0

C.3 Sensitivity to Estimated-Stream Balancing Weights

To complement the sensitivity curves in the main text, Table 10 summarizes the mean performance when the two estimated-stream balancing coefficients are tied as $\lambda_1 = \lambda_2 = w$. Across $w \in [0.1, 0.9]$, P-AUROC remains nearly constant, while PRO@1% varies within only about 1.6 points on MVTec 3D-AD and 1.4 points on EyeCandies. This indicates that the method does not require delicate tuning, and the default choice $w = 0.1$ used in the main manuscript remains a robust operating point across both benchmarks.

C.4 Cross-Shot Stability and Synthetic-Count Sensitivity

Table 11 extends the strict-FPR analysis by aggregating baseline comparisons across 1-shot, 2-shot, and 4-shot settings, together with 5-seed statistics and an ablation on the number of synthetic samples N generated per real image. CMDS-AD consistently improves over CFM and MAFR, with the largest margin appearing at the strictest PRO@1% criterion. Moreover, the 4-shot 5-seed results exhibit low variance ($41.3_{\pm 0.8}$ at PRO@1%), and varying N from 1 to 3 only causes marginal shifts. These observations indicate that the reported gains are stable rather than driven by a favorable seed or a highly sensitive augmentation count.

D Extended Qualitative Results and Failure Cases

D.1 Additional Qualitative Results

To complement the visual demonstrations provided in the main text, we present an extended gallery of qualitative results covering the remaining categories of both the MVTec 3D-AD and EyeCandies datasets. Notably, all visual results presented in this section are generated under the 4-shot training setting. Specifically, Figures 9 and 10 display the anomaly localization results for the MVTec 3D-AD dataset, which features diverse industrial materials and complex 3D topologies. Meanwhile, Figures 11 and 12 demonstrate our performance on the EyeCandies dataset, which introduces challenging scenarios with intricate, highly reflective textures.

Table 11: Cross-shot strict-FPR comparison on MVTec 3D-AD, including 4-shot stability over 5 random seeds and an ablation on the number of synthetic samples N generated per real image.

Shot	Method	I-AUC	P-AUC	30%	10%	5%	1%
1	CFM	69.8	97.5	91.3	76.8	63.4	26.2
	MAFR	72.4	97.7	92.2	79.4	67.6	30.0
	Ours	**79.6**	**98.3**	**94.2**	**84.3**	**74.4**	**36.0**
2	CFM	73.0	98.0	92.6	80.3	68.5	30.6
	MAFR	76.6	98.0	93.2	81.8	71.6	33.6
	Ours	**83.0**	**98.5**	**94.8**	**85.9**	**76.5**	**37.5**
4	CFM	80.8	98.5	94.2	84.3	73.9	35.2
	MAFR	84.1	98.4	94.6	85.6	76.5	37.7
	Ours	**87.1**	**98.9**	**95.8**	**88.6**	**80.6**	**41.0**
	5 seeds	$86.8_{\pm 1.2}$	$98.9_{\pm 0.1}$	$95.9_{\pm 0.3}$	$88.8_{\pm 0.7}$	$80.9_{\pm 0.9}$	$41.3_{\pm 0.8}$
	$N = 1$	86.5	98.9	95.8	88.5	80.3	40.4
	$N = 3$	86.2	98.9	95.9	88.7	80.5	40.6

As illustrated in these figures, the visualizations are arranged from top to bottom, comprising the input RGB image, the Real Normal map, the Surface Normal map, the Ground Truth mask, the anomaly map predicted by MAFR, and our model's final dense prediction output. This comprehensive layout allows for a direct visual correlation between the raw multi-modal inputs, the distinct geometric cues provided by different normal representations, and the final localization performance.

Compared to the MAFR baseline, which frequently struggles with over-segmentation and is easily distracted by benign high-frequency texture variations or sensor noise, our proposed CMDS-AD exhibits remarkable robustness. Benefiting from the dual-branch decoupling mechanism, our method effectively cross-verifies structural deviations and texture anomalies. Consequently, CMDS-AD not only accurately localizes subtle defects but also demonstrates exceptional precision in delineating complex anomaly boundaries across highly varied material surfaces, consistently suppressing spurious background activations and yielding predictions that closely align with the Ground Truth.

D.2 Failure Cases Analysis

In Figure 8, we highlight some typical failure cases of this approach under strictly constrained settings. For instance, in the first left row, we note that our method cannot accurately highlight the missing left part of the *cookie*. This limitation fundamentally stems from the physical absence of the region itself: since the network cannot assign anomaly scores to non-existent pixels (or empty background), the anomaly signal is inevitably forced to manifest on the remaining broken edge adjacent to the defect. In the second left row, the *potato* presents a tiny defect on its body, while the anomaly map—although covering the defect correctly—predicts a much broader anomaly region. This over-segmentation phenomenon can be attributed to the low-pass filtering nature of the estimation

stream, where the highly localized defect signal bleeds into the smoothed surrounding features during the multi-scale feature comparison. In the first and second right rows, categories such as the *candy cane* and the *hazelnut truffle* present hyper-complex, high-frequency 2D or 3D patterns that produce higher anomaly scores compared to the real defects. Such false positives highlight a fundamental challenge in extreme few-shot scenarios: when the limited normal reference set fails to capture the full manifold of benign texture variations, the fusion mechanism may occasionally misinterpret these unseen but normal high-frequency patterns as structural anomalies.

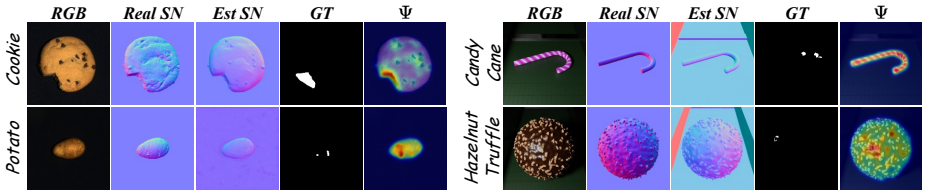


Fig. 8: Typical failure cases of CMDS-AD. We highlight instances where the model misses specific structural defects or predicts broader false positive regions due to hyper-complex high-frequency patterns.

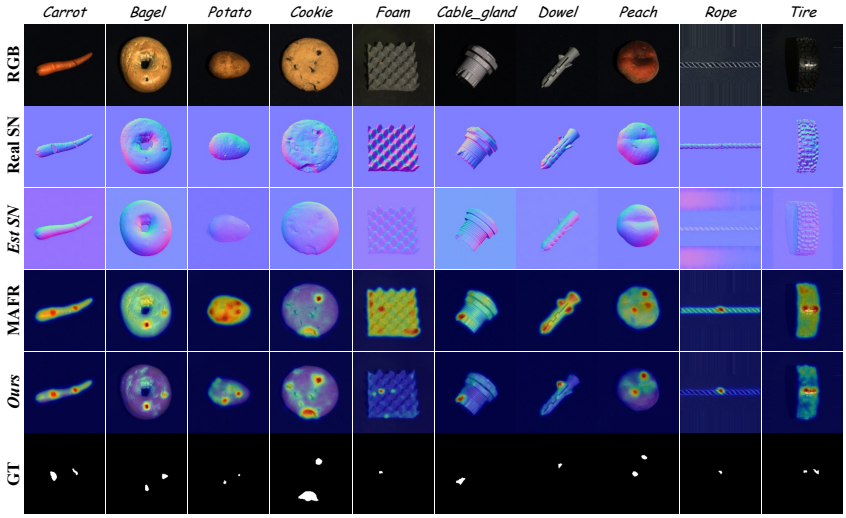


Fig. 9: Extended qualitative results on the MVTec 3D-AD dataset (Part 1) under the 4-shot training setting. From top to bottom: RGB image, Real Normal map, Surface Normal map, Ground Truth mask, MAFR prediction, and our CMDS-AD predicted anomaly map.

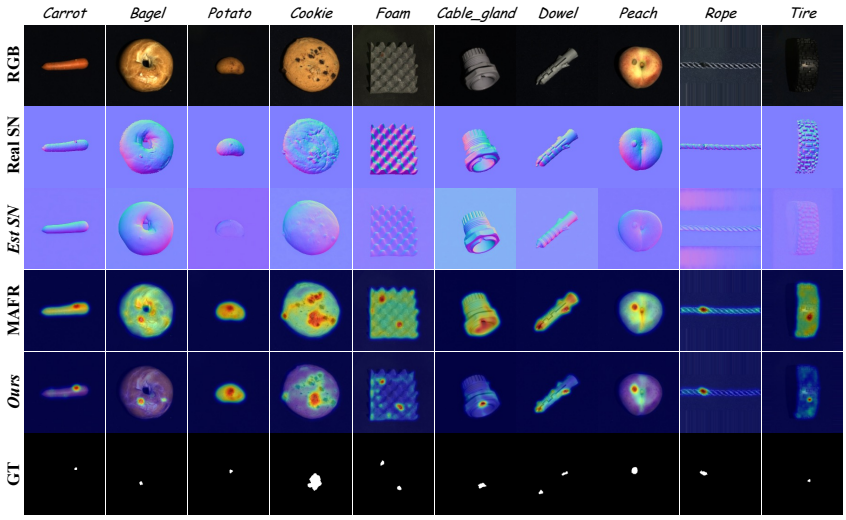


Fig. 10: Extended qualitative results on the MVTec 3D-AD dataset (Part 2) under the 4-shot training setting. The layout of the visualizations is identical to Figure 9.

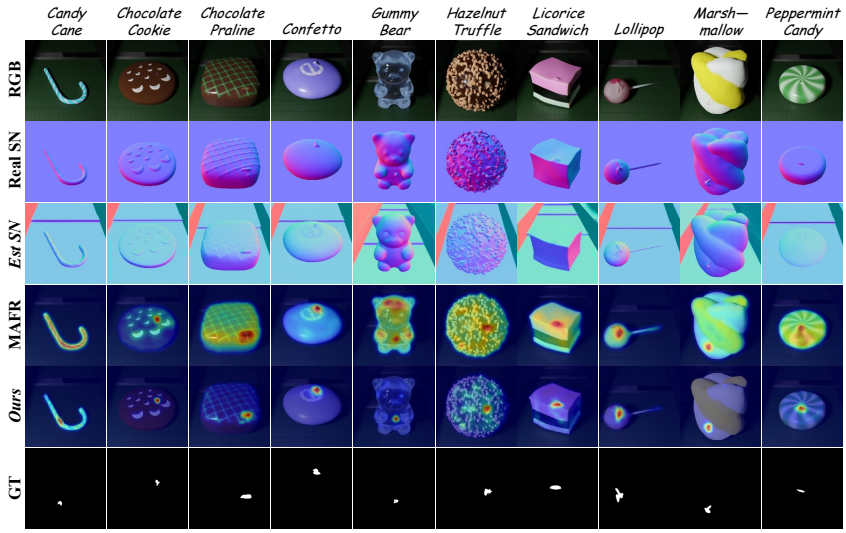


Fig. 11: Extended qualitative results on the EyeCandies dataset (Part 1) under the 4-shot training setting. From top to bottom: RGB image, Real Normal map, Surface Normal map, Ground Truth mask, MAFR prediction, and our CMDS-AD predicted anomaly map.

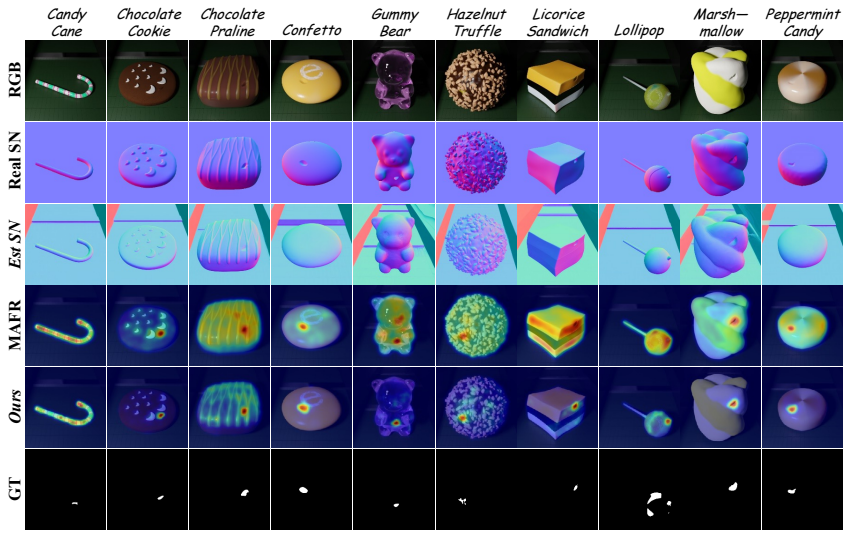


Fig. 12: Extended qualitative results on the EyeCandies dataset (Part 2) under the 4-shot training setting. The layout of the visualizations is identical to Figure 11.