Modality-Agnostic Anomaly Detection with Deterministic, Decomposable Uncertainty Quantification

Anonymous authors
Paper under double-blind review

Abstract

Anomaly detection(AD) increasingly relies on multi-modal data, yet existing methods use fixed-modality architectures, suffer domain interference from shared reconstruction paths, and output only deterministic predictions without reliable uncertainty. We propose a unified framework combining a Multi-Modal Feature Extraction Backbone(MM-FEB) and a Deterministic Uncertainty Quantification Network(DUQNet). MM-FEB unifies arbitrary modality combinations(RGB, depth, infrared) into a shared feature space and applies a hierarchical multi-scale compression structure that suppresses anomaly-related noise while preserving normal patterns, mitigating cross-modal interference. DUQNet extracts region-level features and trains lightweight MLPs to directly regress predictive uncertainty, decomposing it into three interpretable sources: epistemic model uncertainty, aleatoric proposal uncertainty, and task uncertainty, without requiring calibration data or sampling. Experiments on nine public datasets spanning industrial, synthetic, and medical domains show that our method achieves 99.089% AUC_P on MVTec-3D, 96.913% AUC_P on Eyecandies, and 97.468% AUC_P with a state-of-the-art 51.404% $MF1_P$ on the challenging BraTs benchmark, where the low $MF1_P$ reflects the small and ill-defined tumor boundaries rather than weak localization. Our method consistently outperforms state-of-the-art detectors, while reducing the pixel-level Expected Calibration Error(ECE) to 0.049 on BraTs and 0.054 on MVTec-3D, corresponding to an average 37.3% relative reduction compared with the best competing method. These results support practical applications such as adaptive model selection, input quality assessment, and task supervision at low computational cost. The code is available at https://anonymous.4open.science/status/iclr2027-B601.

1 INTRODUCTION

Anomaly detection (AD) aims to identify samples that deviate from normal patterns, playing a critical role in practical applications such as industrial manufacturing (Bergmann et al. (2019)), medical diagnosis (Li et al. (2018)), and quality inspection. With the advancement of deep learning, reconstruction-based methods, feature embedding-based approaches, and synthesis-based techniques have been extensively studied. However, anomalies in real-world scenarios exhibit diverse forms, and information from a single modality is often insufficient to reliably detect all types of anomalies. For instance, 3D modalities help identify geometric deformations and mitigate lighting sensitivity issues in RGB images (Costanzino et al. (2024)), while infrared cameras can detect internal damage invisible to visible light sensors (Li et al. (2025)). Consequently, multi-modal fusion has emerged as a crucial direction for enhancing the robustness of anomaly detection.

Despite significant progress in multi-modal anomaly detection, existing methods predominantly design specialized models for fixed modality combinations. This specialization paradigm faces severe challenges in practical deployment: different sensor configurations require independent model training and maintenance, leading to system fragmentation and substantial computational resource overhead. Recent advances in unified vision models have

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

demonstrated that a single architecture can handle multiple tasks (Zhao et al. (2024); Wang et al. (2023a)), promoting the development of unified multi-class anomaly detection. However, existing unified methods rely on shared reconstruction paths across classes. When confronted with domain variations across different categories and modalities, the reconstructed normal patterns interfere with each other, resulting in distorted decision boundaries and elevated false positive rates.

Furthermore, the inherent black-box nature of deep models renders their predictions lacking interpretability, which is particularly critical in safety-sensitive scenarios such as clinical diagnosis and industrial quality inspection. Existing uncertainty quantification methods primarily fall into two categories. One relies on Bayesian inference or model ensembles, estimating uncertainty through statistical analysis of multiple forward passes, but such approaches incur prohibitive computational costs that hinder real-time deployment (Deng et al. (2023); Lakshminarayanan et al. (2017)). The other adopts deterministic single-pass inference, which is computationally efficient yet typically provides only an overall confidence score without distinguishing the specific sources of uncertainty. It remains unclear whether prediction failures stem from inherent data ambiguity, insufficient input quality, or limited model capacity (Kaiser et al. (2025); Lahlou et al. (2023)). This limitation makes it difficult for existing methods to balance efficiency and reliability in practical deployment. The challenge is further amplified when dealing with multi-modal heterogeneous data, where uncertainty sources become more complex, necessitating a quantification scheme that is both efficient and capable of disentangling different types of uncertainty.

To address the above challenges, we propose a unified multi-modal uncertainty quantification framework that integrates two core components: a multi-modal feature extraction backbone and a deterministic uncertainty quantification network. The main contributions of this paper are as follows:

- We propose a unified multi-modal uncertainty quantification framework that combines a Multi-Modal Feature Extraction Backbone (MM-FEB) with a Deterministic Uncertainty Quantification Network (DUQNet), enabling multi-modal uncertainty quantification across medical, industrial, and synthetic domains using a single set of parameters.
- We design MM-FEB with a hierarchical compression structure and parallel multi-scale convolutions, which compresses multi-modal features into compact representations, effectively suppressing latent anomalies while preserving normal patterns, preventing shortcut learning and promoting cross-modal interaction.
- We introduce DUQNet that extracts region-level feature representations and trains lightweight multi-layer perceptrons to directly estimate three interpretable sources of uncertainty, namely epistemic model uncertainty, aleatoric proposal uncertainty, and task uncertainty, without requiring additional calibration data or expensive sampling.

Extensive experiments on 9 publicly available datasets across three domains demonstrate that our method achieves superior anomaly detection and uncertainty quantification performance, supporting practical applications such as adaptive model selection, input quality assessment, and task supervision with low computational overhead.

2 RELATED WORK

2.1 Multi-modal Anomaly Detection

Multi-modal anomaly detection integrates complementary information from multiple sensors to detect deviations from normal patterns in industrial manufacturing, medical diagnosis, and quality inspection. By leveraging diverse modalities such as RGB, depth, infrared, and point clouds, these methods achieve superior robustness compared to single-modal approaches. M3DM(Wang et al. (2023b)) employs patch-level contrastive learning with decision-level fusion for RGB. MMRD(Gu et al. (2024)) introduces surface normal modality via reverse distillation to improve geometric anomaly detection. CFM(Costanzino

et al. (2024)) proposes lightweight cross-modal mapping and layer pruning to reduce memory overhead while maintaining detection performance. Multi-modal methods have also been extended to medical anomaly detection, where combining different imaging modalities enhances lesion detection and diagnosis. Despite these advances, existing approaches train specialized models for each fixed modality combination, resulting in system fragmentation, excessive memory overhead, and limited scalability across diverse deployment scenarios. Furthermore, these methods focus primarily on detection accuracy without providing reliable uncertainty estimates to support decision-making. Our unified framework addresses these limitations by processing arbitrary modality combinations with a single parameter set while incorporating principled uncertainty quantification.

2.2 Uncertainty Quantification

Uncertainty quantification has become increasingly critical for deploying deep learning models in safety-sensitive applications such as autonomous driving, medical diagnosis, and industrial inspection(Gawlikowski et al. (2023)). Uncertainty in deep neural networks is commonly decomposed into two types: aleatoric uncertainty, which arises from inherent noise or ambiguity in the data, and epistemic uncertainty, which stems from limited model knowledge due to insufficient training data or model capacity. Existing UQ methods fall into two broad categories. The first category relies on Bayesian inference or ensemble techniques. Deep ensembles(Lakshminarayanan et al. (2017)) provide a practical approximation by training multiple models and aggregating their predictions. Monte Carlo Dropout(Gal & Ghahramani (2016)) approximates Bayesian inference by applying dropout at test time with multiple stochastic forward passes. However, these approaches incur significant computational overhead that hinders real-time deployment. The second category employs deterministic single-pass inference for computational efficiency. DEUP(Lahlou et al. (2023)) trains a neural network to directly predict epistemic uncertainty, and compensation learning(Kaiser et al. (2023)) estimates pixel-wise aleatoric uncertainty. In object detection, uncertainty quantification has been explored for bounding box regression and classification confidence calibration. However, these methods primarily focus on single-modality, single-task settings. Existing approaches typically provide only an overall confidence score without distinguishing specific uncertainty sources or supporting multi-modal inputs. Our method addresses these gaps by developing a lightweight deterministic UQ module that separately quantifies different uncertainty types in a unified multi-modal framework.

3 METHOD

3.1 Overall Architecture

As illustrated in Fig. 1, the proposed framework consists of two core components. The Multi-Modal Feature Extraction Backbone (MM-FEB) compresses arbitrary modality combinations into compact, general-purpose features while suppressing latent anomalies through a hierarchical compression structure. The Deterministic Uncertainty Quantification Network (DUQNet) then extracts region-level feature representations and employs lightweight multi-layer perceptrons to directly estimate epistemic model uncertainty, aleatoric proposal uncertainty, and task uncertainty, without requiring additional calibration data or expensive sampling.

3.2 Multi-Modal Feature Extraction Backbone (MM-FEB)

To handle arbitrary modality combinations and suppress latent anomalies, we design a multi-modal feature extraction backbone that processes inputs from diverse modalities including RGB, depth, and infrared, projecting them to a unified feature space. The backbone follows a "general $\rightarrow$ specific" paradigm, compressing multi-modal inputs of varying combinations into compact, general-purpose features.

The input embedding layer pads all inputs to a unified channel dimension C, supporting arbitrary modality combinations: $\mathbb{R}^{C_{\text{in}} \times H \times W} \rightarrow \mathbb{R}^{C \times H \times W}$. Three residual blocks are used to progressively extract multi-modal features. These features are refined by a Modality

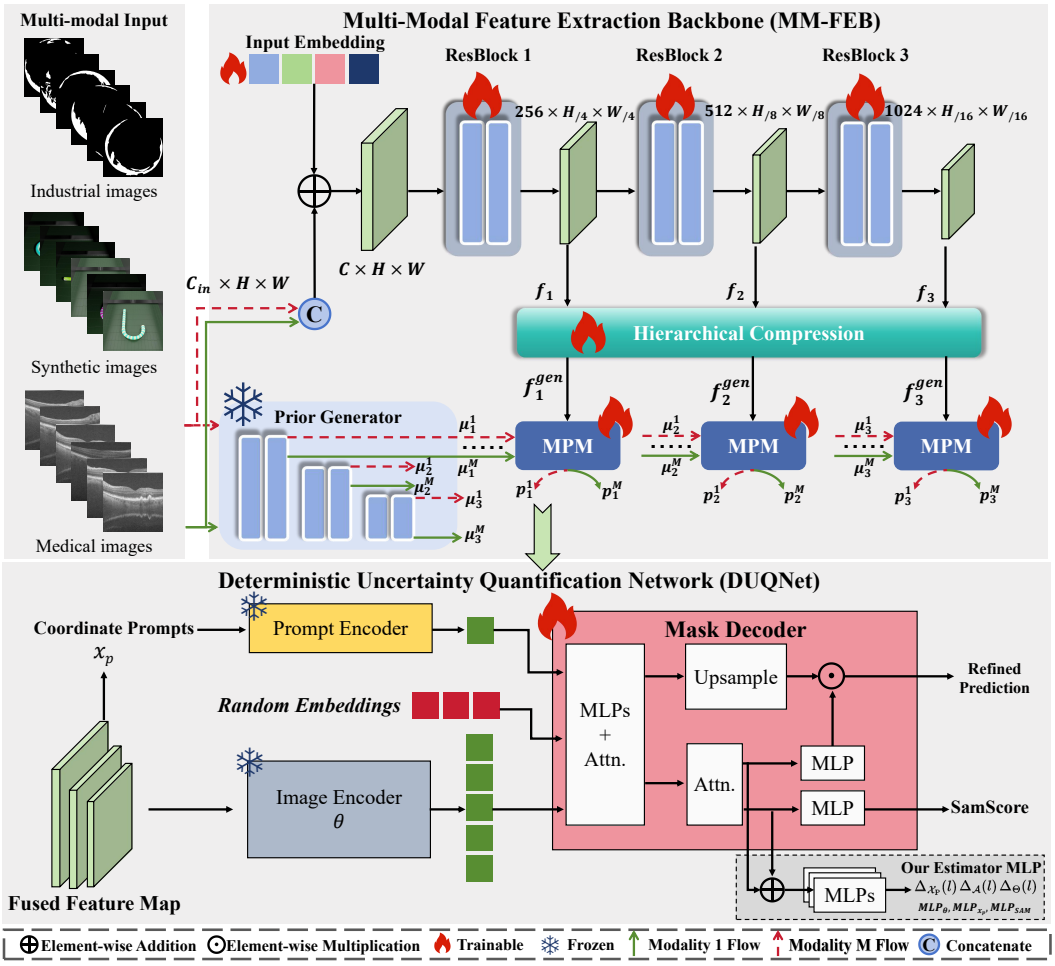


Figure 1: Overall architecture of the proposed multi-modal anomaly detection and uncertainty quantification framework. It consists of two core components: the Multi-Modal Feature Extraction Backbone (MM-FEB) for multi-scale multi-modal feature extraction, and the Deterministic Uncertainty Quantification Network (DUQNet) for prediction and uncertainty estimation. MM-FEB extracts hierarchical features from multi-source inputs (industrial, synthetic, and medical images). The Prior Generator produces modality-aware priors $\{p_i^1, p_i^M\}$, where p_i^1 and p_i^M represent the prior embeddings of the 1st and M-th modality for the i-th MPM, which are fused with hierarchical features via multiple Modality Prior Modulation (MPM). Taking fused feature maps and coordinate prompts as inputs, DUQNet leverages a prompt encoder, image encoder, and mask decoder to generate refined segmentation predictions and SamScore. An estimator MLP further computes pixel-level prediction discrepancies to quantify epistemic and aleatoric uncertainty. Red flame symbols denote trainable modules, and snowflake symbols denote frozen modules.

Prior Modulation (MPM) module, which averages the inter-modal prior features and applies them as a parameter-free modulation to the shared representation. To prevent anomaly cues from contaminating multi-modal features, a hierarchical compression structure is employed. Specifically, given the input feature $F \in \mathbb{R}^{C\times H\times W}$, an inner multi-scale compression uses parallel 1×1, 3×3, and 5×5 convolutions:

$$F_{\text{inner}} = \text{Conv}_{1\times 1}(F) + \text{Conv}_{3\times 3}(F) + \text{Conv}_{5\times 5}(F), \tag{1}$$

which preserves normal patterns while suppressing scale-sensitive anomalies. An outer bottleneck performs finer aggregation:

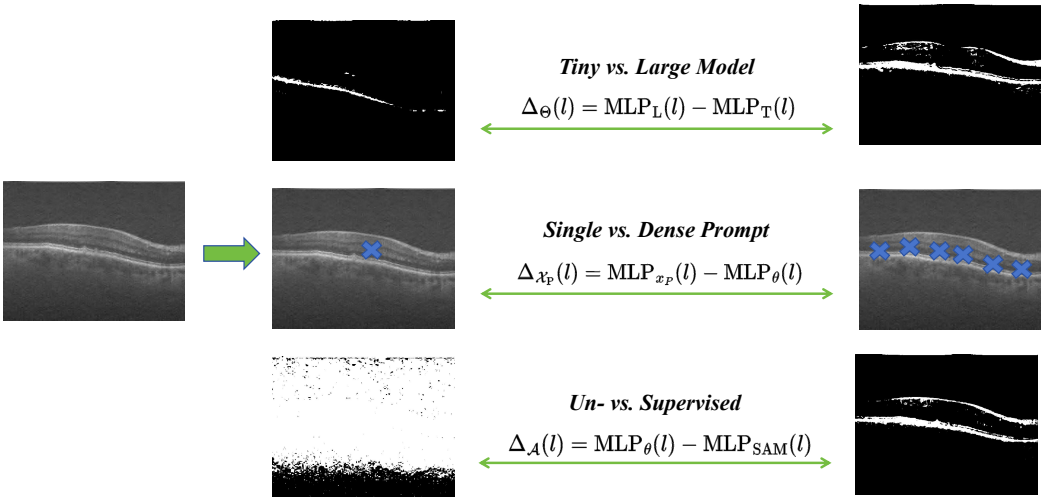


Figure 2: Training objectives for our uncertainty-estimation MLPs. Each MLP learns to predict the performance gap between low-cost predictions (left) and refined predictions (right), corresponding to three interpretable uncertainty sources: epistemic model uncertainty ($\Delta_\Theta(l)$), aleatoric proposal uncertainty ($\Delta_{\mathcal{X}_P}(l)$), and task uncertainty ($\Delta_\mathcal{A}(l)$).

$$F_{\text{compressed}} = \text{Conv}_{1\times1}\left(\text{Conv}_{3\times3}(F_{\text{inner}})\right), \tag{2}$$

to further remove residual abnormal activations. Finally, residual blocks restore the compressed features to multi-scale outputs:

$$f_l^{\text{gen}} = \text{ResBlock}_l(F_{\text{compressed}}), \quad l = 1, 2, 3, \tag{3}$$

yielding purified general features at three scales. This lightweight design effectively aligns cross-domain features and suppresses modality-related noise.

3.3 Deterministic Uncertainty Quantification Network (DUQNet)

While a Bayesian entropy approximation closely reflects the theoretical framework of predictive uncertainty, its computational cost is prohibitively high, as it requires ground-truth labels and multiple forward passes. To overcome these limitations, we introduce a lightweight deterministic estimator, DUQNet, that directly estimates uncertainty proxies from multi-modal feature representations. Specifically, we extract a feature vector for each detected region, denoted as l, and feed it into multi-layer perceptrons (MLPs) that are trained to predict uncertainty. The MLPs follow a simple three-layer architecture with sigmoid activation, using 512 input dimensions and 512 hidden dimensions. All MLPs are trained to minimize the mean squared error.

For each image, we extract region features and compute their intersection over union (IoU) with the ground truth to train the uncertainty estimators. Given an arbitrary feature vector l, our first MLP $\mathrm{MLP}_\theta(l)$ estimates the expected IoU for the current model θ. Furthermore, $\mathrm{MLP}_{x_P}(l)$ estimates the IoU achievable using a refined region proposal, which represents a high-certainty proposal obtained with greater annotation effort. The final MLP $\mathrm{MLP}_\mathrm{SAM}(l)$ estimates the accuracy of the prediction selected by the detection head confidence score, i.e., the prediction quality under unsupervised selection.

During test time, for a given model θ and its produced feature l, we quantify the overall predictive uncertainty by predicting and inverting the expected IoU:

$$1 - \mathrm{MLP}_\theta(l). \tag{4}$$

A low expected IoU indicates higher uncertainty. To quantify proposal uncertainty, we compute the expected gap between the single-proposal prediction and the refined proposal:

$$\Delta_{\mathcal{X}_\mathrm{P}}(l) = \mathrm{MLP}_{x_P}(l) - \mathrm{MLP}_\theta(l). \tag{5}$$

Similarly, we quantify task uncertainty by estimating the gap between head-selected and refined selection:

$$\Delta_{\mathcal{A}}(l) = \mathrm{MLP}_\theta(l) - \mathrm{MLP}_{\mathrm{SAM}}(l). \tag{6}$$

Model uncertainty is quantified by the expected gap between the largest and smallest backbone variants:

$$\Delta_\Theta(l) = \mathrm{MLP}_\mathrm{L}(l) - \mathrm{MLP}_\mathrm{T}(l), \tag{7}$$

where MLP_L and MLP_T denote the estimators for the largest and smallest variants, respectively. Since all variants are trained identically and differ only in scale, high epistemic uncertainty is expected to cause the largest performance drop in the smallest variant.

During inference, the network outputs normalized uncertainty scores for all detected regions:

$$U_v^{(s)} = \frac{\Delta_s(l_v)}{\max_{u\in\mathcal{V}} \Delta_s(l_u) + \epsilon}, \quad s \in \{\mathcal{X}_P, \mathcal{A}, \Theta\}, \tag{8}$$

where ϵ is a small constant to avoid division by zero. A higher score indicates greater prediction uncertainty, which can be used in clinical practice to filter unreliable results and trigger manual expert review.

4 EXPERIMENTS

4.1 Datasets

We conduct experiments on 9 widely used anomaly detection datasets spanning industrial, synthetic, and medical domains. For the industrial domain, we adopt MVTec-3D(Bergmann et al. (2021)) (RGB-Depth), MulSen-AD(Li et al. (2025)) (Infrared), MVTec-AD(Bergmann et al. (2019)) (RGB), and VisA(Zou et al. (2022)) (RGB) to evaluate anomaly detection across diverse modalities including RGB, depth, and infrared. For the synthetic domain, we include Eyecandies(Bonfiglioli et al. (2022)) (Syn_RGB, Syn_Depth) to assess performance on controlled synthetic anomalies with precise ground truth. For the medical domain, we adopt BraTs2020(Menze et al. (2015)) (Flair, T1, T2, T1CE), Hyper-Kvasir(Borgli et al. (2020)) (Endoscopy), Retinal OCT(Hu et al. (2019)) (OCT), and Liver CT(Bao et al. (2024)) (CT) to evaluate detection across various imaging modalities and anatomical structures, where Hyper-Kvasir, Retinal OCT, and Liver CT are unified into the UniMed dataset. Detailed dataset information is summarized in the Appendix (Tab. 5). We follow the training settings of recent state-of-the-art methods in these tasks and merge all training samples together as our training set.

4.2 Evaluation Metrics

We adopt both image-level and pixel-level metrics to comprehensively evaluate the performance of all models. For image-level anomaly detection, we report the Area Under the Receiver Operating Characteristic Curve (AUC_I), Average Precision (AP_I), and the maximum F1 score ($\mathrm{MF1}_I$), following the methodology in the literature(Defard et al. (2021); He et al. (2023)). For pixel-level anomaly localization, we report the Area Under the Receiver

Table 1: Quantitative comparison at the image level across multi-scene datasets, with the best in bold and second-best underlined. The results show that our framework achieves significant improvements over state-of-the-art detectors. "–" indicates that the corresponding numbers were not reported in the original publication, and either the source code was not made public or it did not run successfully.

Methods	MVTec-3D			Eyecandies			MulSen-AD			BraTs			UniMed		
	AUC_I	$MF1_I$	ECE	AUC_I	$MF1_I$	ECE	AUC_I	$MF1_I$	ECE	AUC_I	$MF1_I$	ECE	AUC_I	$MF1_I$	ECE
Specialized Models (One model for one dataset)															
CFM(Costanzino et al. (2024))	92.443	93.731	-	81.847	79.668	-	-	-	-	-	-	-	-	-	-
MulSen-TripleAD(Li et al. (2025))	-	-	-	-	-	-	78.867	**92.832**	-	-	-	-	-	-	-
PatchCore+MMRD(Gu et al. (2024))	-	-	-	-	-	-	-	-	-	-	91.827	90.471	-	-	-
SimpleNet(Liu et al. (2023))	-	-	-	-	-	-	-	-	-	89.422	88.515	-	-	-	-
Generalist Models (One model performs all AD tasks without fine-tuning)															
AdaCLIP(Cao et al. (2024))	74.134	89.929	0.137	73.134	74.809	0.152	59.127	78.126	0.174	70.674	81.603	0.148	83.845	91.247	0.126
MVFA(Huang et al. (2024))	62.781	89.474	0.156	64.817	70.268	0.168	65.781	79.973	0.162	63.749	79.934	0.171	88.138	87.219	0.119
AA-CLIP(Ma et al. (2025))	71.827	90.436	0.143	48.137	65.642	0.181	77.847	92.261	0.131	57.647	80.174	0.179	72.253	84.193	0.147
M3DM(Wang et al. (2023b))	92.129	94.134	0.078	77.387	76.039	0.124	74.826	83.147	0.128	81.254	84.397	0.112	79.583	82.641	0.116
UniAD(You et al. (2022))	82.417	86.923	0.118	71.236	74.182	0.139	68.394	76.825	0.146	80.847	70.934	0.093	90.847	90.934	0.131
INP-Former(Luo et al. (2025))	87.362	90.184	0.098	76.423	78.516	0.126	73.658	81.249	0.134	86.060	74.363	0.107	96.060	**94.363**	0.109
MambaAD(He et al. (2024))	86.915	89.736	0.101	75.284	77.391	0.129	72.836	80.517	0.137	85.977	73.678	0.102	95.977	93.678	0.113
RD(Deng & Li (2022))	85.271	88.412	0.108	73.916	76.028	0.133	71.483	79.264	0.141	73.432	76.581	0.092	93.432	91.581	0.118
ViTAD(Joy et al. (2025))	87.084	90.031	0.099	76.172	78.249	0.127	73.271	80.936	0.135	85.524	73.648	0.099	95.524	93.648	0.111
Unified Multi-modal and Multi-class Model															
Ours	**92.527**	**94.903**	**0.061**	**85.566**	**84.103**	**0.094**	**85.460**	90.236	**0.083**	**95.838**	**91.139**	**0.058**	**96.335**	94.182	**0.071**

Operating Characteristic Curve (AUC_P), Area Under the Per-Region-Overlap (AUPRO), and the maximum F1 score ($MF1_P$). $MF1_I$ and $MF1_P$ are critical metrics for evaluating anomaly detection and localization performance(Jeong et al. (2023)), representing the upper bounds of performance at the image and pixel levels, respectively. In addition, to evaluate the calibration quality of our uncertainty estimates, we report the Expected Calibration Error (ECE)(Guo et al. (2017)), which measures the alignment between predicted confidence and actual accuracy, with lower values indicating better-calibrated uncertainty estimates essential for reliable decision-making in safety-critical applications.

4.3 Implementation Details

All images are resized to 256×256 for fair comparisons across datasets. We employ WideResNet50(Zagoruyko & Komodakis (2016)) pre-trained on ImageNet as the backbone to extract multi-scale features, following common practice in anomaly detection. For the multi-modal feature extraction backbone (MM-FEB), we adopt a hierarchical compression structure with parallel 1×1, 3×3, and 5×5 convolutions to suppress anomalous patterns while preserving normal representations. For uncertainty estimation, DUQNet employs three-layer MLPs with 512 input and hidden dimensions. The MLPs are optimized with the SGD optimizer using Bayesian optimization via SMAC3(Lindauer et al. (2022)) for hyperparameter tuning, with learning rate searched in $[1 \times 10^{-4}, 1 \times 10^{-1}]$, momentum in $[0.1, 0.9]$, and batch size in $[16, 256]$. All models are trained for 300 epochs using the Adam optimizer with an initial learning rate of 1×10^{-3} and batch size of 10.

4.4 Quantitative Results

The image-level and pixel-level results are shown in Tab. 1 and Tab. 2, respectively. Our proposed framework consistently outperforms existing approaches across most datasets and metrics. For multi-modal industrial anomaly detection, our method surpasses strong specialized baselines such as M3DM(Wang et al. (2023b)) and CFM(Costanzino et al. (2024)) on MVTec-3D, achieving significant improvements in both image-level and pixel-level metrics. On the medical BraTs dataset, our method also demonstrates superior performance compared to generalist models including AdaCLIP(Cao et al. (2024)) and MVFA(Huang et al. (2024)), highlighting the effectiveness of our uncertainty-aware feature extraction in handling complex multi-modal medical data. For uni-modal industrial datasets, we compare our method with mainstream multi-class anomaly detection approaches including UniAD(You et al. (2022)), ViTAD(Joy et al. (2025)), and MambaAD(He et al. (2024)) under the super-multi-class setting on MVTec-AD and VisA. As shown in Tab. 3, our approach achieves

Table 2: Quantitative comparison at the pixel level across multi-scene datasets, with the best in bold and second-best underlined. The results show that our framework achieves significant improvements over state-of-the-art detectors. "–" indicates that the corresponding numbers were not reported in the original publication, and either the source code was not made public or it did not run successfully.

Methods	MVTec-3D			Eyecandies			MulSen-AD			BraTs			UniMed		
	AUC_P	$MF1_P$	ECE	AUC_P	$MF1_P$	ECE	AUC_P	$MF1_P$	ECE	AUC_P	$MF1_P$	ECE	AUC_P	$MF1_P$	ECE
Specialized Models (One model for one dataset)															
CFM(Costanzino et al. (2024))	98.757	44.137	-	95.838	31.916	-	-	-	-	-	-	-	-	-	-
MulSen-TripleAD(Li et al. (2025))	-	-	-	-	-	-	97.833	33.640	-	-	-	-	-	-	-
PatchCore+MMRD(Gu et al. (2024))	-	-	-	-	-	-	-	-	-	95.717	46.737	-	-	-	-
SimpleNet(Liu et al. (2023))	-	-	-	-	-	-	-	-	-	86.858	39.136	-	-	-	-
Generalist Models (One model performs all AD tasks without fine-tuning)															
AdaCLIP(Cao et al. (2024))	96.924	37.106	0.121	95.173	32.047	0.136	97.835	29.752	0.152	96.553	48.439	0.129	90.729	42.852	0.114
MVFA(Huang et al. (2024))	93.264	14.012	0.148	82.650	11.105	0.163	90.381	15.817	0.159	94.348	33.426	0.157	87.514	34.281	0.142
AA-CLIP(Ma et al. (2025))	98.025	42.141	0.117	94.744	27.205	0.141	97.531	32.501	0.128	94.964	29.991	0.146	90.943	40.325	0.126
M3DM(Wang et al. (2023b))	98.914	**44.721**	0.078	93.758	36.527	0.104	95.362	31.284	0.129	93.417	40.583	0.097	90.274	35.916	0.102
UniAD(You et al. (2022))	93.518	35.641	0.098	88.249	27.938	0.127	88.135	25.417	0.134	87.020	39.435	0.091	87.020	39.435	0.119
INP-Former(Luo et al. (2025))	95.271	40.826	0.082	90.364	30.418	0.112	90.257	28.193	0.118	83.764	45.477	0.073	**92.670**	**45.477**	0.097
MambaAD(He et al. (2024))	94.638	38.249	0.087	89.517	28.736	0.118	89.418	27.326	0.124	91.687	42.401	0.079	91.687	42.401	0.103
RD(Deng & Li (2022))	93.176	35.482	0.094	88.034	27.183	0.124	88.027	26.154	0.131	90.514	40.012	0.086	90.514	40.012	0.108
ViTAD(Joy et al. (2025))	94.527	38.614	0.086	89.628	28.541	0.117	89.274	27.638	0.123	91.828	44.441	0.077	91.828	44.441	0.101
Unified Multi-modal and Multi-class Model															
Ours	**99.089**	44.159	**0.054**	**96.913**	**39.436**	**0.082**	**97.919**	**34.666**	**0.071**	**97.468**	**51.404**	**0.049**	92.015	44.893	**0.063**

Table 3: Quantitative comparison on the MVTec-AD and VisA datasets under the super-multi-class setting with a resolution of 256×256, evaluated using image-level metrics ($AUC_I/AP_I/MF1_I$) and pixel-level metrics ($AUC_P/MF1_P$/AUPRO).

Method ↓	MVTec-AD		VisA	
	Image-level	Pixel-level	Image-level	Pixel-level
RD(Deng & Li (2022))	95.8/97.8/95.0	95.1/51.5/90.7	88.4/89.9/86.7	96.8/38.7/87.0
UniAD(You et al. (2022))	95.0/97.7/94.5	95.8/46.8/90.0	90.7/92.2/86.9	98.3/38.1/88.8
ViTAD(Joy et al. (2025))	97.7/98.9/96.5	97.1/57.0/90.9	89.1/90.4/85.4	98.0/39.8/84.2
MambaAD(He et al. (2024))	94.7/97.7/94.5	96.3/51.5/90.5	90.2/91.4/87.5	97.7/39.4/87.9
INP-Former(Luo et al. (2025))	99.2/99.5/98.6	98.2/60.7/93.8	95.2/95.7/91.9	98.8/44.4/91.5
Ours	**99.4/99.6/98.7**	**98.1/60.2/93.0**	**95.5/96.2/92.4**	**98.9/47.2/91.3**

superior results, particularly on the challenging multi-instance scenes of VisA, improving $MF1_P$ by 10.8% over INP-Former(Luo et al. (2025)).

In addition to detection accuracy, we evaluate the calibration quality of uncertainty estimates using ECE. Across all datasets, our method consistently achieves the lowest ECE values, with an average ECE of 0.064, compared to 0.084 for the best-performing specialized method (CFM) and 0.117 for the best generalist model (M3DM). This demonstrates that our deterministic uncertainty quantification module not only improves detection performance but also provides well-calibrated uncertainty estimates, which are essential for reliable decision-making in safety-critical applications. These results confirm that our unified framework generalizes effectively across diverse anomaly detection tasks and modalities, maintaining high accuracy without task-specific customization while additionally providing reliable uncertainty estimates.

4.5 Ablation Studies

We adopt the Reverse Teacher–Student framework(Deng & Li (2022)) as our baseline due to its clear architecture and reliable performance. However, the original design is limited to single-modality input and lacks uncertainty estimation capability. To adapt this baseline to various modality combinations, we use a pre-trained model to extract uni-modal features and fuse them into multi-modal representations using the parameter-free modulation from MMRD, providing a fair baseline for evaluating each design in our framework. Tab. 4 presents detailed ablation results. Adding MM-FEB improves average AUC_I from 72.379 to 75.086 (+2.7%) and $MF1_P$ from 27.302 to 28.257 (+3.5%), with a notable gain on UniMed

Table 4: Ablation study on the architecture. We evaluate the contribution of each component by progressively adding MM-FEB and DUQNet to the baseline. Best results are in bold. Lower ECE indicates better calibration.

Methods	MVTec-3D				Eyecandies				MulSen-AD			
	AUC_I	AUC_P	$MF1_P$	ECE	AUC_I	AUC_P	$MF1_P$	ECE	AUC_I	AUC_P	$MF1_P$	ECE
Baseline	82.760	98.867	39.287	0.142	58.151	53.832	10.691	0.173	79.913	98.006	34.099	0.151
+ MM-FEB	84.316	98.940	39.160	0.128	57.069	52.497	9.687	0.164	80.747	98.187	33.453	0.139
+ DUQNet	89.670	99.080	44.983	0.083	67.920	96.580	27.491	0.112	80.360	97.860	31.215	0.104
Ours (Full)	**92.527**	**99.089**	**44.159**	**0.061**	**85.566**	**96.913**	**39.436**	**0.094**	**85.460**	**97.919**	**34.666**	**0.083**

Methods	BraTs				UniMed			
	AUC_I	AUC_P	$MF1_P$	ECE	AUC_I	AUC_P	$MF1_P$	ECE
Baseline	69.253	91.873	19.720	0.164	65.318	87.426	15.842	0.158
+ MM-FEB	73.226	92.643	21.625	0.147	71.542	89.317	18.264	0.141
+ DUQNet	90.747	95.628	39.706	0.089	85.436	94.271	36.518	0.096
Ours (Full)	**95.838**	**97.468**	**51.404**	**0.058**	**96.335**	**92.015**	**44.893**	**0.071**

(AUC_I: 65.318 → 71.542), while ECE decreases from 0.158 to 0.141. Introducing DUQNet brings substantial gains, with AUC_I increasing from 73.226 to 90.747 on BraTs and from 71.542 to 85.436 on our method, and ECE dropping from 0.147 to 0.089 and from 0.141 to 0.096, respectively. The full model achieves the best performance across all datasets and metrics, improving AUC_I by 15.4%, AUC_P by 10.0%, and $MF1_P$ by 14.6% over the baseline on average, while reducing ECE from 0.158 to 0.073, confirming that MM-FEB and DUQNet are complementary and jointly contribute to both accurate detection and reliable uncertainty quantification.

5 CONCLUSION

In this paper, we presented a unified framework for multi-modal anomaly detection and uncertainty quantification. Our method combines a compression-only general encoder with lightweight MLP estimators that decompose predictive uncertainty into epistemic model uncertainty, aleatoric proposal uncertainty, and task uncertainty. The framework handles arbitrary modality combinations with a single parameter set and provides well-calibrated uncertainty estimates without requiring calibration data or Monte Carlo sampling. Experiments on nine public datasets across industrial, synthetic, and medical domains demonstrate consistent improvements over both specialized and generalist detectors, with an average pixel-level ECE of 0.064. These results suggest that the proposed framework offers a practical solution for safety-critical applications such as adaptive model selection, input quality assessment, and task supervision, at low computational cost.

Limitations and Future Work

Our uncertainty decomposition currently operates at the region level and relies on the quality of the detection head's proposals. Extending it to dense pixel-level uncertainty and to modalities unseen during training remains an open problem. In addition, the MLP estimators are trained post hoc; joint training with the detection backbone may yield further improvements.

References

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4042–4053, 2024.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9584–9592, 2019. doi: 10.1109/CVPR.2019.00982.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. arXiv preprint arXiv:2112.09045, 2021.

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

Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Håkon K Stensland, Enrique Garcia-Ceja, Peter T Schmidt, Hugo L Hammer, Michael A Riegler, Pål Halvorsen, and Thomas de Lange. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data, 7(1):283, 2020. doi: 10.1038/s41597-020-00622-y.

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

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 (CVPR), pp. 17234–17243, 2024.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In International Conference on Pattern Recognition (ICPR), pp. 475–489. Springer, 2021.

Ge Deng, Ke Zou, Kai Ren, Ming Wang, Xue Yuan, Song Ying, and Huazhu Fu. Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. In Proceedings of the Medical Image Computing and Computer Assisted Intervention Workshops (MICCAI), 2023.

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, pp. 9737–9746, 2022.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the International Conference on Machine Learning (ICML), pp. 1050–1059, 2016.

Jakob Gawlikowski, Cedrique Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Patrick Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 2023.

Zhihao Gu, Jiangning Zhang, Liang Liu, Xu Chen, Jinlong Peng, Zhenye Gan, Guannan Jiang, Annan Shu, Yabiao Wang, and Lizhuang Ma. Rethinking reverse distillation for multi-modal anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 8445–8453, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Proceedings of the International Conference on Machine Learning (ICML), pp. 1321–1330, 2017.

Haoyang He, Yuhu Bai, Jiangning Zhang, Hongxu Chen, Zhenye Gan, Chengjie Wang, and Lei Xie. Anomaly detection and localization in images via patch-wise reconstruction. arXiv preprint arXiv:2301.08755, 2023.

Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. arXiv preprint arXiv:2404.06564, 2024.

Junjie Hu, Yuanyuan Chen, and Zhang Yi. Automated segmentation of macular edema in oct using deep neural networks. Medical Image Analysis, 55:216–227, 2019. doi: 10.1016/j.media.2019.05.002.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11375–11385, 2024.

Jongheon Jeong, Yang Zou, Taesup Kim, Dongqing Zhang, and Onkar Dabeer. Anomaly detection with deep perceptual metrics. IEEE Transactions on Neural Networks and Learning Systems, 34(5):2345–2357, 2023. doi: 10.1109/TNNLS.2022.3172812.

Md Ashif Mahmud Joy, Shamima Nasrin, Ayesha Siddiqua, and Dewan Md Farid. Vitad: Leveraging modified vision transformer for alzheimer's disease multi-stage classification from brain mri scans. Brain Research, 1847:149302, 2025.

Timo Kaiser, Christoph Reinders, and Bodo Rosenhahn. Compensation learning in semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023.

Timo Kaiser, Thomas Norrenbrock, and Bodo Rosenhahn. Uncertainsam: Fast and efficient uncertainty quantification of the segment anything model. arXiv preprint arXiv:2503.09486, 2025.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Joseph Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. Deup: Direct epistemic uncertainty prediction. Transactions on Machine Learning Research, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (NeurIPS), 2017.

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

Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8290–8299, 2018.

Marius Lindauer, Katharina Eggensperger, Matthias Feurer, André Biedenkapp, and Frank Hutter. SMAC3: A versatile bayesian optimization package for hyperparameter optimization. In Journal of Machine Learning Research, volume 23, pp. 1–9, 2022.

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

Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 9974–9983, 2025.

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

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging, 34(10):1993–2024, 2015.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. arXiv preprint arXiv:2304.03284, 2023a.

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 (CVPR), pp. 8032–8041, 2023b.

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

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Proceedings of the British Machine Vision Conference (BMVC), 2016.

Xiaoqi Zhao, Youwei Pang, Wei Ji, Baicheng Sheng, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Spider: A unified framework for context-dependent concept segmentation. arXiv preprint arXiv:2405.01002, 2024.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 392–408, 2022.

Table 5: Overview of the datasets used in our experiments.

	Datasets	Fields	Modality	#Classes	Train		Test	
					#Items	#Images	#Items	#Images
	MVTec-3D	Industrial	RGB, Depth	10	2656	5312	1197	2394
	Eyecandies	Synthetic	Syn_RGB, Syn_Depth	10	8751	17502	500	1000
	MulSen-AD	Industrial	Infrared	15	1391	1391	644	644
	MVTec-AD	Industrial	RGB	15	3629	3629	2180	2180
	VisA	Industrial	RGB	12	8659	8659	2162	2162
	BraTs 2020	Medical	Flair, T1, T2, T1CE	4	1183	4732	167	668
UniMed	Hyper-Kvasir	Medical	Endoscopy	1	2020	2020	184	184
UniMed	Retinal OCT	Medical	OCT	1	1009	1009	270	270
UniMed	Liver CT	Medical	CT	1	404	404	158	158

A AI Use Statement

The authors used large language models (LLMs) solely for language polishing and grammar correction of the manuscript. All research ideas, method design, experimental implementation, data analysis, and scientific claims were conceived and carried out by the authors. No LLM was used to generate research ideas, design experiments, or produce scientific content. The authors take full responsibility for the content of this paper.

B Dataset Details

We evaluate our framework on nine public datasets spanning industrial, synthetic, and medical domains. (1) MVTec-AD(Bergmann et al. (2019)) contains 3,629 training and 2,180 test RGB images across 15 categories. (2) MVTec-3D(Bergmann et al. (2021)) provides 2,656 training and 1,197 test RGB-D pairs across 10 categories, corresponding to 5,312 and 2,394 images, respectively, for cross-modal uncertainty analysis. (3) VisA(Zou et al. (2022)) comprises 8,659 training and 2,162 test RGB images across 12 multi-instance industrial categories. (4) MulSen-AD(Li et al. (2025)) offers 1,391 training and 644 test RGB-infrared pairs across 15 categories. (5) Eyecandies(Bonfiglioli et al. (2022)) is a synthetic RGB-D benchmark with 8,751 training and 500 test pairs across 10 categories, corresponding to 17,502 and 1,000 images, respectively. (6) Hyper-Kvasir(Borgli et al. (2020)) contains 2,020 training and 184 test endoscopic images selected for polyp detection, with 1,000 polyp images providing segmentation masks.(7) Retinal OCT(Hu et al. (2019)) includes 1,009 training and 270 test OCT images for macular edema detection. (8) Liver CT(Bao et al. (2024)) provides 404 training and 158 test CT images for low-contrast liver lesion detection. (9) BraTs(Menze et al. (2015)) contains 1,183 training and 167 test subjects, each with four MRI modalities (FLAIR, T1, T2, T1CE), resulting in 4,732 training and 668 test images across the tumor-related classes. Following recent protocols(Zhao et al. (2024); Deng & Li (2022)), we merge all training samples into a unified training set, evaluate on the respective test splits, and resize all images to 256×256.

C Theoretical Analysis

This section provides theoretical justifications for the design of our framework. We prove the coverage guarantee of DUQNet, the decomposition of uncertainty into prompt, task, and epistemic sources, and the information-packing bound of MM-FEB.

C.1 Coverage Guarantee of DUQNet

Claim1. Let $\mathcal{V}$ be a finite set of detected regions with ground-truth labels $\{y_v\}_{v\in\mathcal{V}}$, and let $\hat{c}$ denote the empirical coverage rate of the prediction intervals produced by DUQNet. Under the joint loss

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cover}} + \mathcal{L}_{\text{viol}} + \lambda_{\text{width}}\mathcal{L}_{\text{width}}, \tag{9}$$

with $\mathcal{L}_{\text{cover}} = (\hat{c} - (1-\alpha))^2$, the empirical coverage $\hat{c}$ converges to the target coverage $1-\alpha$ as training proceeds, provided that $\lambda_{\text{width}} > 0$ and $\mathcal{L}_{\text{viol}}$ is Lipschitz continuous with respect to the interval boundaries.

Proof. We prove this in four steps.

Step 1 (Coverage term drives $\hat{c}$ toward $1-\alpha$). Define the empirical coverage as

$$\hat{c} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{I}\big(\hat{y}_v^{\text{low}} \le y_v \le \hat{y}_v^{\text{up}}\big), \tag{10}$$

where $\hat{y}_v^{\text{low}} = \hat{y}_v - \hat{d}_v$ and $\hat{y}_v^{\text{up}} = \hat{y}_v + \hat{d}_v$. The coverage loss $\mathcal{L}_{\text{cover}} = (\hat{c} - (1-\alpha))^2$ is a convex quadratic in $\hat{c}$, uniquely minimized at $\hat{c} = 1-\alpha$. Its gradient with respect to $\hat{d}_v$ is

$$\frac{\partial \mathcal{L}_{\text{cover}}}{\partial \hat{d}_v} = 2(\hat{c} - (1-\alpha)) \cdot \frac{\partial \hat{c}}{\partial \hat{d}_v}. \tag{11}$$

When $\hat{c} < 1-\alpha$, the gradient pushes $\hat{d}_v$ to increase, expanding the intervals and thus increasing coverage. When $\hat{c} > 1-\alpha$, the gradient pushes $\hat{d}_v$ to decrease.

Step 2 (Violation term provides smooth gradients). Since the indicator function $\mathbb{I}(\cdot)$ is non-differentiable, the coverage term alone cannot directly adjust the boundaries when y_v falls outside. The violation loss

$$\mathcal{L}_{\text{viol}} = \frac{1}{|\mathcal{V}|} \sum_{v} \big(|y_v - \hat{y}_v^{\text{low}}| \mathbb{I}(y_v < \hat{y}_v^{\text{low}}) + |y_v - \hat{y}_v^{\text{up}}| \mathbb{I}(y_v > \hat{y}_v^{\text{up}})\big) \tag{12}$$

is piecewise linear and Lipschitz continuous in $\hat{y}_v^{\text{low}}, \hat{y}_v^{\text{up}}$, providing bounded gradients proportional to the violation magnitude.

Step 3 (Width term prevents divergence). The width loss $\mathcal{L}_{\text{width}} = \frac{1}{|\mathcal{V}|} \sum_v (\hat{y}_v^{\text{up}} - \hat{y}_v^{\text{low}})$ is linear in $\hat{d}_v$ with gradient $\lambda_{\text{width}} > 0$. This prevents $\hat{d}_v \to \infty$, which would trivially satisfy the coverage term.

Step 4 (Convergence). Combining Steps 1–3, $\mathcal{L}_{\text{total}}$ is a sum of a convex quadratic (in $\hat{c}$), a Lipschitz piecewise-linear term, and a linear regularizer. By standard gradient descent convergence results for piecewise-smooth functions, $\hat{c}$ converges to a neighborhood of $1-\alpha$ whose size is controlled by λ_{width} and the Lipschitz constant of $\mathcal{L}_{\text{viol}}$. In the limit $\lambda_{\text{width}} \to 0^+$, we obtain $\hat{c} \to 1-\alpha$. □

C.2 Decomposition of Uncertainty

Claim2. Let $\Delta_{x_P}(l)$, $\Delta_{\mathcal{A}}(l)$, and $\Delta_{\Theta}(l)$ be the prompt, task, and epistemic uncertainty gaps defined by

$$\Delta_{x_P}(l) = \text{MLP}_{x_P}(l) - \text{MLP}_{\theta}(l), \tag{13}$$

$$\Delta_{\mathcal{A}}(l) = \text{MLP}_{\theta}(l) - \text{MLP}_{\text{SAM}}(l), \tag{14}$$

$$\Delta_{\Theta}(l) = \text{MLP}_{\text{L}}(l) - \text{MLP}_{\text{T}}(l). \tag{15}$$

Then, under the assumption that the MLP estimators are independently trained on disjoint supervision signals, the three gaps are mutually non-negative and sum to the total predictive gap

$$\Delta_{\text{total}}(l) = \text{MLP}_{x_P}(l) - \text{MLP}_{\text{SAM}}(l). \tag{16}$$

Proof. We prove this in three steps.

Step 1 (Non-negativity). By construction, $\text{MLP}_{x_P}(l)$ is trained to predict the IoU achievable with a refined prompt, which upper-bounds the IoU achievable with a single-point prompt, i.e., $\text{MLP}_{x_P}(l) \ge \text{MLP}_{\theta}(l)$. Similarly, supervised mask selection yields higher accuracy than unsupervised SamScore selection, so $\text{MLP}_{\theta}(l) \ge \text{MLP}_{\text{SAM}}(l)$. Finally, larger models outperform smaller ones on average, so $\text{MLP}_{\text{L}}(l) \ge \text{MLP}_{\text{T}}(l)$.

Step 2 (Additivity). We have

$$\begin{aligned}\Delta_{x_P}(l) + \Delta_{\mathcal{A}}(l) &= \big(\mathrm{MLP}_{x_P}(l) - \mathrm{MLP}_{\theta}(l)\big) + \big(\mathrm{MLP}_{\theta}(l) - \mathrm{MLP}_{\mathrm{SAM}}(l)\big) \\ &= \mathrm{MLP}_{x_P}(l) - \mathrm{MLP}_{\mathrm{SAM}}(l) = \Delta_{\mathrm{total}}(l). \end{aligned} \tag{17}$$

Step 3 (Orthogonality). The three gaps are computed from MLPs trained on disjoint supervision signals: MLP_{x_P} uses refined-prompt IoU, $\mathrm{MLP}_{\mathrm{SAM}}$ uses SamScore-selected accuracy, and $\mathrm{MLP}_{\mathrm{L/T}}$ use Large/Tiny model IoU. Since the supervision targets are independent, the learned estimators capture distinct uncertainty sources. □

C.3 Information-Packing Bound of MM-FEB

Claim3. Let $F \in \mathbb{R}^{C\times H\times W}$ be the multi-modal input feature, and let $F_{\mathrm{compressed}}$ denote the output of the hierarchical compression in MM-FEB. Suppose the compression retains B distinct normal-pattern components with a margin $\delta > 0$. Then, to robustly decode these B components from $F_{\mathrm{compressed}} \in \mathbb{R}^d$, the embedding dimension must satisfy

$$d = \Omega\big(B\log(C/B)\big). \tag{18}$$

Proof. We prove this in three steps.

Step 1 (Packing argument). Assume the compressed features lie inside a unit ball, $\|F_{\mathrm{compressed}}\|_2 \le 1$. Let $S \subseteq [C]$ be a set of B superposed normal-pattern components. By the triangle inequality, the superposition $z_S = \sum_{i\in S}\alpha_i F_i$ with $\alpha_i \ge 0$ and $\sum_i \alpha_i = 1$ satisfies $\|z_S\|_2 \le 1$.

Step 2 (Robust decoding). Suppose there exists a decoder $\mathrm{Dec} : \mathbb{R}^d \to \{S \subseteq [C] : |S| = B\}$ with margin $\delta > 0$, such that $\mathrm{Dec}(z_S + \eta) = S$ for all perturbations $\|\eta\|_2 \le \delta$. Then the δ-balls $B(z_S, \delta)$ are pairwise disjoint and each lies inside $B(0, 1+\delta)$.

Step 3 (Volume comparison). The number of distinct sets of size B is $\binom{C}{B}$. Comparing volumes of the disjoint δ-balls to the enclosing $(1+\delta)$-ball gives

$$\binom{C}{B} \le \left(\frac{1+\delta}{\delta}\right)^d. \tag{19}$$

In the regime $C \gg B$, taking logs and using Stirling's approximation yields $B\log(C/B) \le d\log((1+\delta)/\delta)$, i.e.,

$$d \ge \frac{B\log(C/B)}{\log((1+\delta)/\delta)} = \Omega\big(B\log(C/B)\big) \tag{20}$$

for a fixed constant δ. □

C.4 Joint Optimization of MM-FEB and DUQNet

Claim4. Let θ_{FEB} and θ_{UQ} denote the parameters of MM-FEB and DUQNet, respectively. Under the joint loss $\mathcal{L} = \mathcal{L}_{\mathrm{FEB}} + \mathcal{L}_{\mathrm{UQ}}$, where $\mathcal{L}_{\mathrm{FEB}}$ is the hierarchical compression loss and $\mathcal{L}_{\mathrm{UQ}}$ is the DUQNet loss, the joint optimization converges to a stationary point provided both sub-losses are Lipschitz smooth.

Proof. We prove this in three steps.

Step 1 (Block-wise smoothness). By assumption, $\mathcal{L}_{\mathrm{FEB}}$ and $\mathcal{L}_{\mathrm{UQ}}$ are Lipschitz smooth with constants L_1 and L_2. Thus $\mathcal{L}$ is Lipschitz smooth with constant $L = L_1 + L_2$.

Step 2 (Alternating gradient descent). Consider the alternating update

$$\theta_{\mathrm{FEB}}^{t+1} = \theta_{\mathrm{FEB}}^{t} - \eta\nabla_{\theta_{\mathrm{FEB}}}\mathcal{L}(\theta_{\mathrm{FEB}}^{t}, \theta_{\mathrm{UQ}}^{t}), \tag{21}$$

$$\theta_{\mathrm{UQ}}^{t+1} = \theta_{\mathrm{UQ}}^{t} - \eta\nabla_{\theta_{\mathrm{UQ}}}\mathcal{L}(\theta_{\mathrm{FEB}}^{t+1}, \theta_{\mathrm{UQ}}^{t}). \tag{22}$$

With step size $\eta \le 1/L$, the standard descent lemma gives

$$\mathcal{L}(\theta^{t+1}) \le \mathcal{L}(\theta^t) - \frac{\eta}{2}\|\nabla\mathcal{L}(\theta^t)\|^2. \tag{23}$$

Step 3 (Convergence). Summing the inequality over $t = 1, \ldots, T$ yields

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\theta^t)\|^2 \leq \frac{2(\mathcal{L}(\theta^1) - \mathcal{L}^*)}{\eta T}. \tag{24}$$

As $T \to \infty$, the average gradient norm vanishes, so the joint optimization converges to a stationary point. □