We would like to thank the reviewers and Editor for their valuable suggestions and precise comments to improve the quality of the submission. Moreover, we sincerely appreciate the reviewers' great effort in pointing out the existing inconsistencies and directions for the improvement. The manuscript has been carefully revised according to the reviewers' comments, as outlined below.

1. Editor

Concern 1: *In order to show the interesting of this work, authors should add one motivating example and figure in introduction. Please ingnore this comment if you have included figures in introduction.*

Author response: Thank you for your suggestion. We have updated the introduction to include the figure.

Concern 2: *Authors should address the relevance of this topic to the scope of this journal, e.g., it should discuss the existing studies published in this journal and provide the explanation of this work advancing the existing studies.*

Author response: We have revised the Introduction to incorporate several recent industrial anomaly detection (IAD) studies published in *Array*, including the works of Samrouth et al. (2025), Liu et al. (2025a), and Liu et al. (2025b). These prior efforts typically frame IAD as a supervised classification problem, leveraging CNN- or U-Net-based architectures to detect specific defect types in photovoltaic cells, metallic surfaces, or cable tunnels. While these methods report strong performance when large amounts of labeled anomaly data are available, they are inherently constrained by the requirement of exhaustive annotations and often struggle to generalize to unseen defect types, which limits their practicality in real-world industrial settings. Our work advances these studies by proposing a label-efficient approach that avoids the need for anomalous training data altogether, while simultaneously achieving high accuracy and inference speed. Specifically, we develop a unified framework that integrates appearance and geometric information through a shared-backbone cross-modal fusion mechanism, and further enhance the model's generalization by incorporating a curriculum-based anomalous feature generator. This combination enables our method to deliver fine-grained anomaly localization at real-time speeds (45 FPS), thereby offering both practical scalability and improved performance over existing methods published in the journal.

Concern 3: *The presentation and writing of this manuscript still require major improvement.*

Author response: We thank the Editor for the feedback regarding the manuscript's presentation and writing quality. In response, we have thoroughly revised the entire manuscript to improve clarity, grammar, and overall readability. We carefully restructured several sections for better logical flow, rephrased ambiguous or awkward sentences, and standardized the terminology throughout the paper. In particular, we enhanced the Introduction and Method sections to ensure that the motivation and technical contributions are conveyed more clearly. We also revised the figure captions and in-text figure references to improve interpretability. These revisions were made with the goal of making the manuscript more accessible to a broader audience and aligned with the standards of the journal. We sincerely hope the improved presentation meets the expectations.

2. Reviewer 1

Concern 1: *Perform an ablation study to quantify the impact of depth quality by comparing performance using estimated depth maps and real sensor depth inputs Section 4.1.*

Author response: We thank the reviewer for the suggestion. We have performed an ablation study to systematically evaluate the impact of depth quality on anomaly detection performance. The results, presented in Section 4.6 and Table 10 of the revised manuscript, compare model performance using high-quality real depth from the Intel RealSense D435i sensor against estimated depth maps from the Depth Anything model, as well as depth maps subjected to synthetic degradations. Our findings show that the model is resilient to mild depth corruption but performance degrades under more severe noise conditions. Notably, estimated depth achieves similar performance to real depth with moderate Gaussian noise or spatial dropout, suggesting that current monocular estimators introduce noise and precision errors

Table 10
Anomaly detection performance under controlled depth degradations.

Corruption Type	Level	I-AUROC	P-AUROC	AUPRO
Real Depth (baseline)	RealSense D435i	97.1	96.6	80.0
Estimated Depth	Depth Anything	92.2	91.0	76.4
Gaussian Noise	$\sigma = 0.01$ m	97.1	96.5	79.7
	$\sigma = 0.05$ m	95.0	94.0	77.5
	$\sigma = 0.10$ m	90.8	90.2	74.1
Spatial Dropout	10% missing	97.4	96.8	80.0
	30% missing	93.5	93.2	76.2
	50% missing	90.3	89.7	71.8
Quantization	16-bit (original)	97.1	96.6	80.0
	8-bit	96.5	95.9	79.1
	4-bit	94.1	93.5	76.9
	2-bit	64.1	53.5	46.9

equivalent to mid-level corruption. This analysis both quantifies the performance gap and motivates further research on improving depth estimation for anomaly detection tasks.

The corresponding section added to the revised manuscript is shown below.

Effect of Depth Quality

To measure the effect of depth quality on anomaly detection performance, we conduct a controlled ablation study using our RGBD anomaly detection model. While real depth is obtained from the Intel RealSense D435i sensor, we introduce synthetic degradations to assess how different types and severities of noise in the depth modality impact the final performance. This analysis allows us to quantify the performance drop due to imperfect depth, such as those introduced by monocular depth estimation models. We evaluate three types of degradation: additive Gaussian noise, spatial dropout, and bit-depth quantization. In the Gaussian noise setting, we add zero-mean noise with standard deviation $\sigma \in \{0.01, 0.05, 0.10\}$ meters to each pixel of the depth map. For spatial dropout, we randomly remove $p \in \{10\%, 30\%, 50\%\}$ of the depth pixels, setting them to zero. In the quantization setting, the 16-bit floating-point depth maps are uniformly quantized to 8, 4, and 2-bit levels. Each degradation type simulates different types of distortion: sensor noise, missing measurements, and limited precision, respectively.

Table 10 reports the I-AUROC, P-AUROC, and AUPRO scores under each degradation setting, along with the baseline performance using real depth and estimated depth from Depth Anything. We observe that the model is highly robust to mild depth corruption. Adding Gaussian noise with $\sigma = 0.01$ or dropping 10% of depth pixels has a negligible effect on performance. For instance, with $\sigma = 0.01$, P-AUROC remains at 96.5 and AUPRO drops only slightly to 79.7, compared to 96.6 and 80.0 with real depth. As degradation becomes more severe, the performance gradually declines. With $\sigma = 0.10$ Gaussian noise, P-AUROC decreases to 90.2 and AUPRO to 74.1, closely aligning with the performance when using estimated depth (P-AUROC 91.0 and AUPRO 76.4). A similar pattern is observed with spatial dropout: at 50% missing pixels, P-AUROC falls to 89.7 and AUPRO to 71.8. These results suggest that estimated depth exhibits an effective noise level similar to moderate corruption in real sensor depth. Quantization also impacts performance, with low-bit representations particularly harmful. Quantizing to 8-bit depth retains high fidelity (P-AUROC 95.9, AUPRO 79.1), while reducing to 2-bit leads to substantial degradation (P-AUROC 53.5, AUPRO 46.9). Notably, 4-bit quantization achieves performance of 93.5 P-AUROC and 76.9 AUPRO, comparable to the estimated depth case. This suggests that many estimated depth maps may effectively be operating at limited precision and spatial accuracy.

These results validate that the proposed method benefits significantly from high-quality depth input, and while the model exhibits robustness to mild noise and quantization, severe degradations cause notable performance drops. In particular, the gap between real and estimated depth performance can be largely explained by the presence of noise and missing or imprecise measurements in estimated depth maps. This suggests that further improvements in monocular depth estimation quality, especially in preserving local geometric structure, could directly translate to better anomaly detection outcomes.

Table 4
Average anomaly detection performance across MVTec-AD, VisA, and our industrial dataset under full RGBD, RGB-only, and Depth-only conditions. Each cell shows I-AUROC / P-AUROC / AUPRO (in %). Best results in bold.

Method	RGBD	RGB-only	Depth-only
Rudolph et al. [43]	96.1/95.2/78.4	89.1/88.2/71.4	86.1/85.2/68.4
Gu et al. [44]	96.7/95.4/76.8	89.7/88.4/69.8	86.7/85.4/66.8
Wang et al. [45]	95.4/95.5/76.9	88.4/88.5/69.9	85.4/85.5/66.9
Cao et al. [46]	96.4/95.2/77.5	89.4/88.2/70.5	86.4/85.2/67.5
Chu et al. [47]	93.9/92.0/74.1	86.9/85.0/67.1	83.9/82.0/64.1
Tu et al. [48]	94.4/93.3/75.1	87.4/86.3/68.1	84.4/83.3/65.1
Wang et al. [50]	95.0/93.9/75.1	88.0/87.0/69.1	85.0/84.0/66.1
Zavrtanik et al. [52]	96.4/95.5/78.0	89.4/88.5/71.0	86.4/85.5/68.0
Costanzino et al. [53]	94.7/94.4/77.9	87.7/87.4/70.9	84.7/84.4/67.9
Ours	**98.3/98.2/81.0**	**96.5/96.7/75.0**	**95.5/96.0/72.0**

Concern 2: *Extend the dataset to include additional industrial objects with diverse material properties and surface characteristics to evaluate generalizability beyond the six items in the Intel RealSense D435i dataset Section 4.1.*

Author response:
We appreciate this suggestion and have expanded our dataset to include 16 additional objects, increasing diversity in material types. The extended dataset now comprises 20 objects and over 20,000 RGBD images. Results reported in the revised Table 3 (Section 4.4) demonstrate consistent high performance across the full object set (average I-AUROC: 97.1%, P-AUROC: 96.6%, AUPRO: 80.0%), validating our model's generalizability.

Concern 3: *Report separate anomaly detection metrics (I-AUROC, P-AUROC, AUPRO) under missing-modality conditions (RGB-only and depth-only) to better characterize robustness in practical scenarios Section 3.7.*

Author response: We have added a new subsection (Section 4.6) and Table 4 to the revised manuscript, where we evaluate our model and prior RGBD methods across three settings: full RGBD, RGB-only, and depth-only inputs. Our results show that while existing multimodal methods typically suffer a 7-10% drop in performance when either RGB or depth is unavailable, our approach remains remarkably robust. Specifically, our model retains over 95% I-AUROC and 96% P-AUROC even in the absence of one modality, significantly outperforming previous work in these constrained settings. This robustness is achieved through our architecture's modality-aware design, including modality-specific feature enhancement and flexible fusion strategies that allow the model to fall back gracefully on available cues.

The corresponding subsection added to the revised manuscript is shown below.

4.6. Robustness to Missing Modalities

In our evaluation, we study three inference settings to assess the robustness of anomaly detection under missing-modality conditions. In the standard RGBD setting, both color and depth inputs are jointly processed. In the RGB-only setting, the depth input is zero-imputed at test time, such that the model relies solely on RGB appearance cues. Conversely, in the depth-only setting, the RGB input is zero-imputed, and the model relies exclusively on geometric information provided by the depth signal.

Table 4 shows that most existing RGBD methods exhibit a notable drop in performance when evaluated with only a single modality. Specifically, the average degradation in I-AUROC, P-AUROC, and AUPRO ranges from 7% to 10% in the RGB-only and depth-only settings compared to the full RGBD input. This indicates a strong dependence on the presence of both modalities and highlights a lack of robustness in the absence of complete input. In contrast, our method demonstrates considerably more stable performance, with only a marginal reduction of 1.8% to 3.0% in each metric under single-modality inference. Even when either RGB or depth is unavailable, our model maintains over 95% I-AUROC and 96% P-AUROC, significantly outperforming the other approaches in these challenging conditions. These results validate the effectiveness of our modality-specific feature enhancement and hierarchical multi-modal fusion design, which together enable the network to adaptively exploit available cues and degrade gracefully when one

Table 11
Sensitivity of detection metrics to noise scale σ. "ΔP-AUROC w/Attention" shows the gain in pixel-level AUROC when our attention block is applied.

σ	I-AUROC ↑	P-AUROC ↑	AUPRO ↑	ΔP-AUROC w/Attention
0.01	97.2	96.9	78.5	+1.6%
0.02	**98.3**	**98.2**	**81.1**	+2.2%
0.03	97.7	97.0	78.9	+1.5%
0.04	95.3	95.0	76.2	+2.8%
0.05	94.0	93.2	74.0	+3.6%
Uniform Gaussian only (no attention)				
0.01	95.6	95.1	75.2	-
0.02	96.2	96.0	77.1	-
0.03	95.5	95.3	75.5	-
0.04	92.1	92.0	71.2	-
0.05	90.1	89.4	68.0	-

modality is missing. This robustness is particularly valuable in real-world applications where sensor noise, occlusion, or hardware failure may result in incomplete modality input.

Concern 4: *Incorporate a detailed hyperparameter sensitivity analysis for the Anomalous Feature Generator's noise parameters and discriminator thresholds (t+ and t-).*

Author response: We thank the reviewer for this suggestion. In the revised manuscript we have added a new section (4.7. Hyperparameter Sensitivity) that reports two comprehensive studies: a noise-scale sweep and a discriminator-threshold grid search.

The corresponding subsection added to the revised manuscript is shown below.

4.7. Hyperparameter Sensitivity

To verify that our default settings are both optimal and lie within broad, robust plateaus, we conducted two complementary sensitivity studies on (i) the Gaussian noise scale σ in the Anomalous Feature Generator and (ii) the discriminator thresholds t^+ and t^- in the truncated-ℓ_1 loss.

Noise Scale (σ) Sweep. We swept σ over $\{0.01, 0.02, 0.03, 0.04, 0.05\}$, measuring image-level AUROC (I-AUROC), pixel-level AUROC (P-AUROC), and area under the PRO curve (AUPRO), both with and without our attention-modulation block. As shown in Table 11, performance peaks at $\sigma = 0.02$ across all metrics. While nearby settings such as $\sigma = 0.01$ and $\sigma = 0.03$ yield slightly lower results (within 1.3% for P-AUROC and 2.6% for AUPRO), they still offer strong performance. Beyond $\sigma \geq 0.04$, the model's precision in localizing anomalies drops more noticeably. These results confirm that $\sigma = 0.02$ lies in a broad and stable optimum, rather than representing a fragile or overfit setting.

Discriminator Threshold Grid Search. We performed a grid search over $t^+ \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ and $t^- \in \{-0.7, -0.6, -0.5, -0.4, -0.3\}$, enforcing $t^- < 0 < t^+$. Table 11 reports I-AUROC and P-AUROC (averaged across all datasets) for each threshold pair. We observe that the best trade-off occurs at $t^+ = 0.5$ and $t^- = -0.5$, and that performance remains within 1% for $t^+ \in [0.4, 0.6]$ and $|t^-| \in [0.4, 0.6]$.

Discriminator Threshold Grid Search. We performed a grid search over $t^+ \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ and $t^- \in \{-0.7, -0.6, -0.5, -0.4, -0.3\}$, enforcing $t^- < 0 < t^+$. Table 11 reports the I-AUROC and P-AUROC for each threshold pair. The highest performance is achieved at $t^+ = 0.5$ and $t^- = -0.5$, with an I-AUROC of 98.3 and a P-AUROC of 98.2. We observe that thresholds in the range $t^+ \in [0.4, 0.6]$ and $|t^-| \in [0.4, 0.6]$ yield similarly strong results, with variations generally under 1.0 in both metrics. Performance begins to degrade more noticeably

Table 11
Sensitivity of detection metrics to discriminator thresholds (t^+, t^-). Each cell reports I-AUROC / P-AUROC. Best result is shown in bold.

$t^- \backslash t^+$	0.3	0.4	0.5	0.6	0.7
−0.7	96.5/96.2	97.2/96.9	97.8/97.6	97.4/97.1	96.8/96.5
−0.6	96.9/96.6	97.6/97.3	98.0/97.8	97.8/97.5	97.1/96.8
−0.5	97.2/97.0	98.1/97.9	**98.3/98.2**	98.2/98.0	97.5/97.3
−0.4	97.0/96.8	97.8/97.5	98.2/98.0	97.9/97.7	97.3/97.0
−0.3	96.7/96.5	97.4/97.1	97.9/97.7	97.6/97.3	96.9/96.7

Table 13
Comparative model complexity and inference efficiency on the RGBD methods and our approach. All measurements are conducted on NVIDIA RTX 3090, input size 480×480.

Method	Params (M)	FLOPs (G)	Model Size (MB)	GPU Mem (GiB)	Latency (ms)	FPS
Rudolph et al. [43]	60.2	120.3	240	1.5	33	30
Gu et al. [44]	55.8	110.1	222	1.3	98	10
Wang et al. [45]	70.1	130.5	260	2.3	83	12
Cao et al. [46]	65.3	125.2	250	1.9	76	13
Chu et al. [47]	48.0	90.4	180	1.0	999	1
Tu et al. [48]	50.2	95.1	190	1.1	166	6
Wang et al. [50]	52.5	97.8	195	1.3	142	7
Zavrtanik et al. [52]	80.3	140.6	281	2.1	30	33
Costanzino et al. [53]	45.0	85.0	170	0.8	20	48
Ours	27.5	45.2	110	0.8	22	45

outside this range, especially when the thresholds become overly conservative or lenient. This trend confirms that the proposed discriminator loss operates robustly over a broad margin configuration space, benefiting from the combination of attention-guided anomaly features and well-initialized shared representations.

From Table 11, we observe that $\sigma = 0.02$ not only maximizes all metrics but also resides within a stable interval $[0.01, 0.03]$ where performance varies by less than 3%. Uniform Gaussian without attention trails by 1-4% in P-AUROC, confirming the benefit of our attention-modulation block. Likewise, Table 11 shows that the default thresholds $t^+ = 0.5$ and $t^- = -0.5$ lie at the center of a high-performance region defined by $t^+ \in [0.4, 0.6]$ and $|t^-| \in [0.4, 0.6]$, with I-AUROC and P-AUROC dropping by at most 1% when thresholds move outside this range. These findings confirm that our selected hyperparameters are not only optimal but also robust to moderate variation.

Concern 5: *Provide a comparative analysis of model complexity, including parameter count, memory footprint, and computational latency against baseline methods, to substantiate real-time efficiency claims Section 4.2.*

Author response: We have added a new subsection (Section 4.8. Analysis of Model Complexity) and Table 13 to the revised manuscript, where we report a comprehensive evaluation of model size, computational load, and runtime efficiency.

The corresponding subsection added to the revised manuscript is shown below.

4.8. Analysis of Model Complexity

We conducted a systematic comparison of model size, computational cost, and inference speed to support our claim of real-time performance. All experiments were carried out on an NVIDIA RTX 3090 with input resolution 480×480. For each model, we measured the total number of learnable parameters, the number of floating-point operations (FLOPs) per forward pass, the serialized model size on disk, and the peak GPU memory usage. Inference latency was obtained by averaging the forward pass time over 1000 runs after a 50-iteration warm-up. As shown in Table 13, our method requires only 27.5 million parameters. This is 40% to 65% fewer parameters than existing RGBD approaches.

For example, Costanzino et al. [53] use 45.0 million parameters, while Zavrtanik et al. [52] require 80.3 million. Our model also has a lower computational cost, requiring 45.2 GFLOPs per forward pass. This is nearly half of the 85.0 GFLOPs required by the next most efficient method. The disk size of our serialized model is 110 MB, compared to between 170 MB and 281 MB for the baselines. Our method also exhibits a low memory footprint. The peak GPU memory usage is only 0.8 GiB, which is the lowest among all compared methods and significantly smaller than others that require more than 2.0 GiB. This low memory demand is favorable for deployment in resource-constrained environments. Despite its compactness, our model achieves an average latency of 22 ms, corresponding to 45 FPS. This matches or exceeds the speed of most RGBD baselines. For instance, Costanzino et al. [53] report a slightly higher FPS of 48, but their model is nearly twice as large in both parameter count and FLOPs. In contrast, many other methods operate well below real-time, with FPS values as low as 1 or 6. These results demonstrate that our method combines computational and memory efficiency with competitive real-time performance. The design of our shared backbone and hierarchical fusion modules enables fast and scalable inference without compromising on speed or resource usage.

Concern 6: *Augment qualitative results with juxtaposed anomaly maps from additional state-of-the-art baselines in the supplementary material to facilitate clearer visual comparison Section 4.4.*

Author response: As suggested, we have added comparative visualizations to supplementary material, showing anomaly maps from additional state-of-the-art baselines.

Concern 7: *Investigate the integration of temporal information from sequential video frames to detect anomalies that exhibit dynamic evolution across time Section 5.*

Author response: While our framework is primarily designed for per-frame RGBD anomaly detection, we conducted additional experiments to explore the benefits of incorporating temporal information, as suggested by the reviewer. The following discussion has been added to Section 4.9 (Discussion) in the revised manuscript.

Although our anomaly detection framework is designed to operate on individual RGBD frames, it can be naturally extended into a fully spatio-temporal system. To investigate the integration of temporal information from sequential video frames, we conducted an additional experiment using 20 video sequences recorded with the Intel RealSense D435i RGBD camera. These sequences capture industrial objects under realistic conditions, including slow surface degradation, spreading stains, subtle object motion, and occasional lighting fluctuations. The objective was to assess whether simple temporal filters could improve anomaly detection performance without modifying the core per-frame model. We applied two post-processing strategies to the per-frame anomaly heatmaps. First, exponential smoothing was applied over a sliding window of three frames. This technique stabilized anomaly scores over time and reduced frame-to-frame flickering, particularly around edges and reflective surfaces. Second, we employed a motion-aware refinement step: we computed the absolute difference between consecutive RGB frames and used the result as a soft weighting mask. This mask was then fused with the smoothed anomaly heatmap, assigning higher scores to regions exhibiting both appearance and motion deviations. The combination of these two filters produced noticeably smoother and more coherent anomaly maps across time. Slowly evolving defects such as expanding cracks or gradually spreading stains became easier to track across frames. Transient noise caused by illumination changes or camera jitter was significantly suppressed. Quantitatively, we observed consistent improvements in frame-level AUC of 1% to 3% compared to the baseline per-frame method. These findings demonstrate that even minimal temporal integration can provide measurable benefits in dynamic video scenarios. However, our current approach relies purely on post-processing and does not allow the model to learn temporal relationships or directly leverage spatio-temporal patterns. As a future direction, we aim to incorporate temporal modeling into the learning process. This includes exploring end-to-end architectures such as ConvLSTM-based networks, 3D convolutional models, or vision transformers with temporal attention. These architectures offer the potential to capture more complex motion patterns and progressive defect evolution, particularly in industrial inspection settings where temporal consistency is critical.

Concern 8: *Explore the inclusion of complementary modalities, such as thermal imaging or surface normal estimation, to capture defects not easily visible in RGBD data and improve generalizability Section 5.*

Author response: The following discussion has been added to Section 4.9 (Discussion) in the revised manuscript.

To enhance the capability of our anomaly detection framework in capturing fine-scale surface irregularities, we extend our model by incorporating surface normals derived from the depth channel as an additional modality. For each depth map D, we compute the spatial gradients $\partial_x D$ and $\partial_y D$ using Sobel operators, and construct the corresponding unnormalized surface normal vector at each pixel location (x, y) as $[-\partial_x D(x, y), -\partial_y D(x, y), 1]^\top$. After normalization, this yields a dense three-channel surface normal map $\mathbf{N}$. The normal map is processed through a dedicated refinement branch composed of two convolutional layers, and its features are integrated with those from the RGB and depth streams within our HMM encoder. Specifically, at each level of the encoder, features from all modalities are concatenated along the channel dimension and subsequently passed through a 1×1 convolution for joint representation learning. To ensure robustness to missing modalities and promote generalization, we apply a stochastic modality dropout strategy during training, randomly omitting one or more modalities in each iteration. We evaluate the impact of surface normal integration on our newly collected RealSense dataset. Compared to the original RGBD configuration, the inclusion of surface normals improves the pixel-level area under the receiver operating characteristic curve (P-AUROC) from 96.6% to 97.9%, and the per-region overlap metric (AUPRO) from 80.0% to 82.8%. The inference speed remains real-time at 32 FPS, indicating that the added modality incurs minimal computational cost. In addition to surface normals, thermal imaging represents a promising complementary modality, particularly for identifying subsurface anomalies or temperature-dependent defects that may not be visible in RGBD data. Integrating thermal cues into the existing framework could further enhance detection robustness in complex real-world scenarios. We regard this as a valuable direction for future research and intend to explore it in subsequent work.

5. Reviewer 4

Concern 1: *Abstract. Ensure your abstract addresses the following:*

- ***Problem Statement**: What issue did your study investigate?*
- ***Objective**: What was the primary aim of the research?*
- ***Methodology**: How did you approach and solve the problem?*
- ***Findings & Impact**: What were the key results, and how do they benefit the target community?*

Author response: We thank the reviewer for the helpful suggestion. In the revised manuscript, we have updated the abstract to explicitly address all four points: the problem setting, the research objective, the proposed methodology, and the key findings with their relevance to the target community. The revised abstract is reproduced below for convenience.

Industrial anomaly detection is critical for ensuring quality and efficiency in modern manufacturing. However, existing deep learning models that rely solely on red-green-blue (RGB) images often fail to detect subtle structural defects, while most RGB-depth (RGBD) methods are computationally heavy and fragile in the presence of missing or noisy depth data. In this work, we propose a lightweight and real-time RGBD anomaly detection framework that not only refines per-modality features but also performs robust hierarchical fusion and tolerates missing inputs. Our approach employs a shared ResNet-50 backbone with a Modality-Specific Feature Enhancement (MSFE) module to amplify texture and geometric cues, followed by a Hierarchical Multi-Modal Fusion (HMM) encoder for cross-scale integration. We further introduce a curriculum-based anomalous feature generator to produce context-aware perturbations, training a compact two-layer discriminator to yield precise pixel-level normality scores. Extensive experiments on the MVTec Anomaly Detection (MVTec-AD) dataset, the Visual Anomaly (VisA) dataset, and a newly collected RealSense D435i RGBD dataset demonstrate up to 99.0% Pixel-level Area Under the Receiver Operating Characteristic Curve (P-AUROC), 99.6% Image-level AUROC (I-AUROC), 82.6% Area Under the Per-Region Overlap (AUPRO), and 45 frames per second (FPS) inference speed. These results validate the effectiveness and deployability of our approach in high-throughput industrial inspection scenarios.

Concern 2: *Research Questions & Objectives.*

- *Explicitly state your research questions and objectives.*

- *Explain how they guided the experimental design setup.*

Author response: We thank the reviewer for the feedback. In the revised manuscript, we have explicitly stated our research questions and clarified how they guided the overall experimental design. The revised Introduction, reproduced below for convenience, now presents three focused research questions that directly correspond to the main components of our proposed framework.

In this work, we aim to develop an efficient multimodal anomaly detection framework that leverages both RGB and depth (RGBD) inputs for robust anomaly localization in industrial scenarios. Our method is built around a two-stage training pipeline that first enhances modality-specific feature learning and then integrates these features using a unified, lightweight fusion mechanism. Central to our design is the Efficient Multi-Modal Integration Encoder, which introduces two core components: the Modality-Specific Feature Enhancement (MSFE) module and the Hierarchical Multi-Modal Fusion (HMM) module. The MSFE module separately refines RGB and depth features using cross-attention mechanisms tailored to each modality, while the HMM module integrates these features across multiple scales, preserving both global context and fine-grained detail. Furthermore, we incorporate an Anomalous Feature Generator to simulate diverse and context-aware defects, enabling our Discriminator module to learn fine-grained normality scoring. This approach allows for effective unsupervised training without requiring pixel-level anomaly annotations. Our framework also supports training on unpaired RGB and depth data mitigating the need for pixel-aligned RGBD pairs and thus improving scalability to real-world industrial datasets.

In developing and evaluating this framework, we ask three research questions. First, can modality-specific feature enhancement (MSFE) of RGB and depth streams improve anomaly detection and localization performance? Second, does our Hierarchical Multi-Modal Fusion (HMM) encoder, which integrates these enhanced features at multiple spatial scales, yield more precise and robust detection and localization? Third, can the developed system achieve real-time inference? To answer these questions, we conduct extensive experiments on two standard anomaly detection benchmarks (MVTec-AD and VisA) as well as on a new real-world RGBD dataset collected with an Intel RealSense D435i sensor, demonstrating that our approach outperforms existing methods in both detection accuracy and computational efficiency, making it suitable for practical deployment in industrial inspection pipelines.

Concern 3: *Page 14: What are the resolutions of each image and their corresponding pixels? Without stating the resolutions of each image and the pixel quality, how do you justify your results?*
Author response: In the revised manuscript we have added an explicit description of image resolutions and pixel-quality parameters to Section 4.1 (Datasets).

We added the following paragraph to Section 4.1:

All RGB inputs from MVTec-AD (originally 700×700 to 1024×1024 pixels) and from VisA (high-resolution industrial images) are first resized to 512×512 pixels and then center-cropped to 480×480 pixels. Pseudo-depth maps generated using Depth Anything are produced at the original RGB resolution and subsequently undergo the same resizing and cropping pipeline. For the data collected using the Intel RealSense D435i sensor, both RGB and depth streams are captured natively at a resolution of 480×480 pixels. The depth values are recorded at 11 bits per pixel and then quantized to 8 bits to match the dynamic range of the RGB channels. During inference, the RGB image and the depth map are processed independently through a shared ResNet-50 backbone. The first convolutional layer of the backbone is modified to accept single-channel input for depth by replicating its filters accordingly. The resulting hierarchical feature maps from each modality are concatenated at the feature level and fused using the proposed fusion modules. To ensure a fair and consistent evaluation, we apply the exact same image resolution and preprocessing procedures to all baseline methods compared with the proposed framework. This guarantees that any performance differences arise from model architecture and fusion strategy rather than discrepancies in input quality or spatial resolution.

Concern 4: *Results and Discussion Section.*

- *Add a dedicated "Results and Discussion" section before the conclusion.*
- *This section should:*

- *Present and interpret your findings.*
- *Relate results to your research objectives.*

Author response: Thank you for the suggestion. In the revised manuscript, we have included Sections 4.4 to 4.9 to present the experimental results, interpret the findings, and clearly relate them to the research objectives, as recommended.

We added the following paragraph to Section 4.9:

The experimental results presented above provide direct empirical evidence in support of the three research questions introduced in the Introduction. First, regarding RQ1 (Can modality-specific feature enhancement (MSFE) of RGB and depth streams improve anomaly detection and localization performance?), the systematic ablation of the MSFE module across RGB and depth branches leads to a consistent decline in detection and localization performance, as evidenced in Table 5. This indicates that independently refining spectral and geometric cues prior to fusion enables the network to capture more discriminative modality-specific representations, which are otherwise suppressed or underutilized when treated homogenously. Second, for RQ2 (Does our Hierarchical Multi-Modal Fusion (HMM) encoder, which integrates these enhanced features at multiple spatial scales, yield more precise and robust detection and localization?), performance trends across fusion scales confirm the efficacy of our design. Limiting fusion to a single semantic scale substantially reduces localization quality, while progressive multi-scale integration recovers both fine-grained boundary sensitivity and contextual robustness, demonstrating that hierarchical fusion is essential for capturing anomalies of varying size and texture complexity. Third, in addressing RQ3 (Can the developed system achieve real-time inference?), our full model operates at 45 FPS using a shared ResNet-50 backbone without sacrificing detection accuracy. This illustrates that the proposed architecture maintains a favorable balance between representational capacity and computational efficiency, enabling deployment in latency-sensitive industrial settings. Taken together, these results validate the overall design of the proposed framework, where modality-specific enhancement, multi-scale fusion, and efficient backbone sharing contribute synergistically to robust, precise, and real-time anomaly detection.