MambaX-AD: Region-Aware Contrastive Learning with Mamba-CNN Cross-Modulation for Multimodal Anomaly Detection

Anonymous authors
Paper under double-blind review

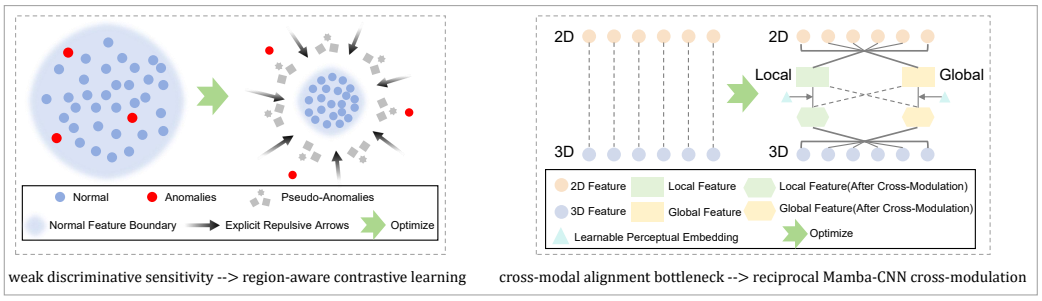


Figure 1: Conceptual illustration of the two proposed mechanisms in MambaX-AD. Region-aware contrastive learning (RA-CL) (Left): targeting the weak discriminative sensitivity. Reciprocal Mamba-CNN cross-modulation (Right): targeting the cross-modal alignment bottleneck.

Abstract

In industrial manufacturing, multimodal anomaly detection combining 2D images and 3D point clouds provides complementary appearance and geometric cues for defect inspection. However, existing unsupervised cross-modal mapping methods typically rely on passive fitting of normal correspondences without explicit negative constraints, which can leave an overly broad normal feature boundary and weak discriminative sensitivity. Moreover, static cross-modal mappings have limited capacity to reconcile the asymmetric local details and global context of the two modalities, resulting in a cross-modal alignment bottleneck. To address these two limitations, we propose MambaX-AD. First, we introduce region-aware contrastive learning (RA-CL), a region-aware, mask-conditioned contrastive strategy that uses the spatial mask of feature-level pseudo-anomalies to impose stronger repulsion on corrupted regions while applying a mild balancing penalty to uncorrupted nominal regions. Second, we introduce reciprocal Mamba-CNN cross-modulation mechanism, in which the local CNN and global Mamba branches generate modulation signals for each other and direction- and depth-aware embeddings condition these interactions. Experiments on MVTec 3D-AD achieve the highest mean image-level I-AUROC among the compared methods while maintaining competitive localization performance, and experiments on Eyecandies achieve the highest results on I-AUROC, P-AUROC, and AUPRO@1% among the compared methods.

1 Introduction

Industrial anomaly detection (IAD), aimed at automatically identifying surface or structural defects during manufacturing processes, is an indispensable core technology in automated quality inspection Liu et al. (2024). Given the extreme rarity and unpredictability of defects in real-world industrial applications Wang et al. (2024); Gu et al. (2024); He et al. (2024), the vast majority of research focuses on unsupervised learning paradigms that rely solely on normal samples for training Costanzino et al. (2024); Wang et al. (2025); Liao et al. (2026). Moreover, traditional IAD primarily relies on single 2D RGB images Wang et al. (2025); however, 2D appearance cues alone can be insufficient

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.

under complex illumination changes or subtle structural defects. With the maturation of 3D sensing technologies, multimodal anomaly detection has therefore garnered widespread attention from both academia and industry Bergmann et al. (2022); Wang et al. (2026); Li et al. (2026a). By combining appearance and geometric information, this paradigm offers a promising route to more reliable inspection in complex industrial scenarios.

Existing methods have conducted extensive explorations. Early memory bank-based methods achieve high accuracy but incur prohibitive memory overhead Gong et al. (2019); Roth et al. (2022); Wang et al. (2023); Yu et al. (2025). To overcome this, cross-modal feature mapping utilizes lightweight networks to learn implicit correlations across modal feature spaces Costanzino et al. (2024); Li et al. (2025). However, under the unsupervised training paradigm, these models rely primarily on fitting normal correspondences. Without explicit negative constraints, the learned normal feature boundary can remain overly broad, so subtle defects may induce insufficient cross-modal deviation and be missed. Furthermore, static mapping architectures do not explicitly account for the semantic asymmetry between 2D and 3D modalities or adaptively modulate local structural details with global context. These limitations give rise to two key challenges: **weak discriminative sensitivity** and the **cross-modal alignment bottleneck** (conceptualized in Figure 1).

To overcome weak discriminative sensitivity, we introduce **Region-Aware Contrastive Learning (RA-CL)**, a region-aware, mask-conditioned contrastive strategy. The spatial corruption mask directly conditions the negative penalty: corrupted locations receive locally amplified, distance-decayed repulsion, whereas uncorrupted nominal locations receive a mild balancing penalty to account for receptive-field leakage. This explicitly constrains the normal-feature boundary and improves sensitivity to subtle defects (Figure 1, Left).

Meanwhile, to address the cross-modal alignment bottleneck, we introduce **Reciprocal Mamba-CNN Cross-Modulation**. Each branch generates a modulation signal that gates the other branch, enabling bidirectional interaction between local structural details and global context. Direction- and depth-aware embeddings further condition these reciprocal gates to account for semantic differences across mapping directions and network depths. Together, these operations enable adaptive local–global feature modulation for cross-modal prediction (Figure 1, Right).

In summary, our main contributions are as follows:

- We propose region-aware contrastive learning (RA-CL), a mask-conditioned contrastive strategy that explicitly differentiates corrupted and uncorrupted nominal regions when repelling feature-level pseudo-anomalies. This directly targets the weak discriminative sensitivity caused by the lack of explicit contrastive constraints, which can leave the normal feature boundary overly broad.
- We propose a reciprocal Mamba-CNN cross-modulation mechanism, in which local and global branches mutually generate modulation signals and are further conditioned by direction- and depth-aware embeddings. This directly targets the cross-modal alignment bottleneck caused by static mapping architectures that overlook the asymmetric local and global semantics of 2D and 3D modalities.
- Extensive experiments achieve the highest mean image-level I-AUROC on MVTec 3D-AD with competitive localization performance, and highest results on multiple metrics on Eyecandies among the compared methods.

2 Related Work

2.1 Multimodal and 3D Anomaly Detection

With the maturation of 3D sensing technologies, recent research has increasingly shifted towards multimodal anomaly detection that integrates 2D images with 3D point clouds Costanzino et al. (2024); Wang et al. (2023); Yu et al. (2025); Li et al. (2025); Horwitz and Hoshen (2023); Rudolph et al. (2023); Tu et al. (2024); Long et al. (2025); Asad et al. (2025); Ali et al. (2025); Zhao et al. (2025); Li et al. (2026b). Among the earliest and most prevalent are memory bank-based methods Wang et al. (2023); Yu et al. (2025); for instance, M3DM Wang et al. (2023) proposes a hybrid fusion scheme, utilizing PointMAE Pang et al. (2022) and Vision Transformers (ViT Dosovitskiy et al. (2021)) to construct massive multimodal memory banks. Similarly, recent works such as Reg2Inv Yu et al. (2025) explore registration-induced memory banks to achieve rotation-invariant representations.

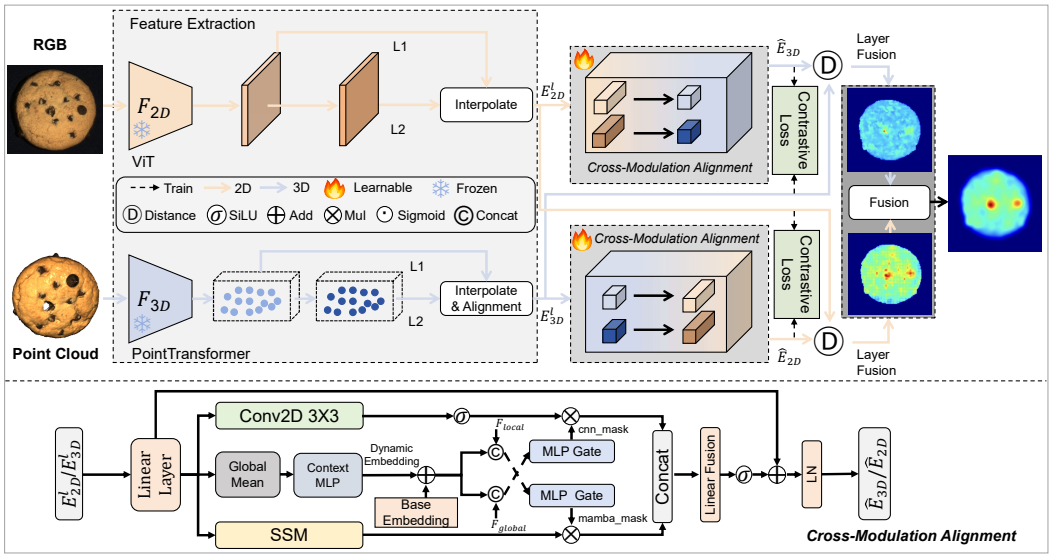


Figure 2: Framework overview. It mainly consists of three stages: Multi-Scale Feature Extraction, Cross-Modulation Alignment, and Region-Aware Contrastive Learning (RA-CL). The lower part details the reciprocal Mamba-CNN cross-modulation mechanism, while RA-CL is detailed in Fig. 3.

Furthermore, MAFR Ali et al. (2025) attempts to localize anomalies through reconstruction errors via cross-modal latent space synthesis and attention-guided restoration.

Despite their high detection accuracy, these methods face severe limitations in practical applications. Memory bank-based approaches require storing millions of high-dimensional feature vectors, resulting in prohibitive memory overhead and unacceptable inference latency, which fail to meet the demands of real-time industrial deployment. On the other hand, reconstruction-based methods Li et al. (2026b); Wu et al. (2026); Cheng et al. (2025); Fang et al. (2025) depend strongly on reconstruction fidelity and may struggle to suppress anomalous content while preserving fine normal structures.

2.2 Cross-Modal Feature Mapping

To overcome the above problems, recent studies have pioneered cross-modal feature mapping Costanzino et al. (2024); Li et al. (2025). CFM Costanzino et al. (2024) was the first to propose learning mapping functions between 2D and 3D feature spaces using normal samples, pinpointing anomalies by measuring the cosine distance between the observed and mapped features. Additionally, FIND Li et al. (2025) employs a dual-student framework integrated with cross-modal feature mapping for few-shot scenarios.

Cross-modal feature mapping provides an efficient alternative to large memory banks, but its normal-only objective and static mappings leave the two challenges identified above insufficiently addressed. Building on this paradigm, we introduce a region-aware, mask-conditioned contrastive objective for constraining the normal-feature boundary and reciprocal Mamba-CNN cross-modulation for adaptive local–global interaction during cross-modal feature mapping.

3 Methodology

3.1 Framework Overview

As illustrated in Figure 2, MambaX-AD first aligns multi-scale RGB and point-cloud features from frozen backbones onto a unified 2D spatial grid. The learnable part is then organized around the two proposed mechanisms: **Region-Aware Contrastive Learning (RA-CL)** for mask-conditioned boundary regularization during training, and **Reciprocal Mamba-CNN Cross-Modulation** for cross-modal prediction.

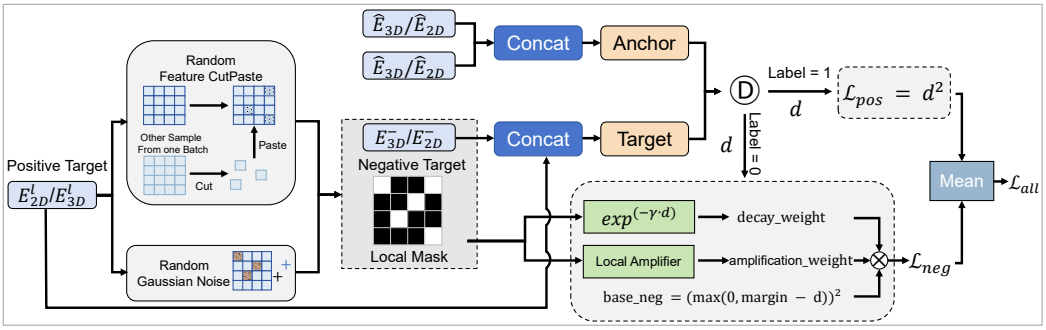


Figure 3: Region-Aware Contrastive Learning (RA-CL).

3.2 Region-Aware Contrastive Learning

To improve discriminative sensitivity to subtle defects, as illustrated in Figure 3, we propose region-aware contrastive learning (RA-CL). Its central design is a mask-conditioned contrastive objective that assigns different repulsive behavior to corrupted and uncorrupted nominal regions, using feature-level pseudo-anomalies as negative targets.

Pseudo-Anomaly Targets and Region Mask. During training, let E^+ denote the clean normal target feature (either E^l_{2D} or E^l_{3D}). We construct a pseudo-anomalous target E^- using Random Feature CutPaste Li et al. (2021) and Random Gaussian Noise injection Liu et al. (2023). The resulting corrupted coordinates define a binary region mask $M_{pa} \in \{0,1\}^{H' \times W'}$, where $M^{(i)}_{pa} = 1$ marks a corrupted spatial location.

Mask-Conditioned Positive and Negative Pairs. Let $\widehat{E}$ denote the feature predicted by the cross-modal mapping network, E^+ the clean normal target, and E^- the pseudo-anomalous target. We explicitly form the positive pair $(\widehat{E}, E^+)$ and the negative pair $(\widehat{E}, E^-)$. Equivalently, an implementation may concatenate them as $E_{anchor} = [\widehat{E}; \widehat{E}]$ and $E_{target} = [E^+; E^-]$.

For spatial location i, define the positive and negative distances as

$$d_i^+ = \|\widehat{E}^{(i)} - E^{+,(i)}\|_2, \qquad d_i^- = \|\widehat{E}^{(i)} - E^{-,(i)}\|_2. \tag{1}$$

The positive term directly aligns the predicted feature with the clean normal target:

$$\mathcal{L}^{(i)}_{pos} = (d_i^+)^2. \tag{2}$$

For the negative pair, we use a margin loss whose weight is conditioned on the pseudo-anomaly region mask:

$$\mathcal{L}^{(i)}_{neg} = \mathcal{W}_i \big(\max(0, m - d_i^-)\big)^2, \tag{3}$$

where m is the safety margin and

$$\mathcal{W}_i = \begin{cases} \alpha_{local} \exp(-\gamma d_i^-), & \text{if } M^{(i)}_{pa} = 1, \\ \alpha_{bf}, & \text{if } M^{(i)}_{pa} = 0. \end{cases} \tag{4}$$

Here α_{local} locally amplifies repulsion on directly corrupted regions, while the distance-decay factor emphasizes hard pseudo-negatives that remain close to the predicted normal feature. Once the negative distance approaches or exceeds the margin, the repulsive effect is attenuated or vanishes, avoiding unnecessary optimization of easy negatives. For uncorrupted nominal regions ($M^{(i)}_{pa} = 0$), the unified normal-region weight α_{bf} provides a mild baseline penalty that mitigates receptive-field leakage Luo et al. (2016); Roth et al. (2022); Wang et al. (2023) without disrupting normal cross-modal alignment.

Let $N = H'W'$ denote the number of spatial locations. Each location contributes one positive pair and one negative pair, i.e., $2N$ pairwise terms in total. The overall loss is

$$\mathcal{L}_{all} = \frac{1}{2N} \sum_{i=1}^{N} \left(\mathcal{L}^{(i)}_{pos} + \mathcal{L}^{(i)}_{neg} \right). \tag{5}$$

By explicitly conditioning the negative penalty on the corruption mask, RA-CL constrains the normal-feature boundary while preserving contextual alignment, thereby improving discriminative sensitivity to subtle defects.

3.3 RECIPROCAL MAMBA-CNN CROSS-MODULATION

To address the cross-modal alignment bottleneck, we propose a reciprocal Mamba-CNN cross-modulation mechanism for cross-modal feature prediction. Prior to the proposed modulation, we follow CFM Costanzino et al. (2024) to extract frozen 2D and 3D backbone features and spatially align them onto a unified 2D grid. Multi-scale features from intermediate layers yield aligned pairs E^l_{2D} and E^l_{3D}, where l indexes network depth.

We define two mappings with the same architecture but independent inputs and non-shared parameters: $\mathcal{M}^l_{2D\to3D}(\cdot;\Theta_{2D\to3D})$ maps E^l_{2D} to the 3D feature space, whereas $\mathcal{M}^l_{3D\to2D}(\cdot;\Theta_{3D\to2D})$ maps E^l_{3D} to the 2D feature space, with $\Theta_{2D\to3D} \neq \Theta_{3D\to2D}$. For compactness, let $d \in \{2D \to 3D, 3D \to 2D\}$ denote the mapping direction. A direction-specific input projection gives $E^{d,l}_{in}$, which is processed by a local CNN branch and a global bidirectional Mamba Gu and Dao (2024) branch to produce $F^{d,l}_{local}$ and $F^{d,l}_{global}$, respectively.

Direction- and Depth-Aware Embedding. The perceptual embedding conditions the reciprocal interaction on both mapping direction and network depth. It is formed as

$$E^{d,l}_{emb} = \phi_{\theta_d}\big(\mu(E^{d,l}_{in})\big) + E^{d,l}_{base}, \tag{6}$$

where $\mu(\cdot)$ denotes global mean pooling, ϕ_{θ_d} is a direction-specific contextual MLP, and $E^{d,l}_{base}$ is a learnable base embedding dedicated to direction d and depth l.

Reciprocal Cross-Modulation. Rather than directly adding or concatenating the two branches, each branch generates the modulation signal applied to the other branch:

$$\widehat{F}^{d,l}_{local} = F^{d,l}_{local} \odot \sigma\Big(\psi^{g\to l}_{\omega_d}\big([F^{d,l}_{global} \parallel E^{d,l}_{emb}]\big)\Big), \tag{7}$$

$$\widehat{F}^{d,l}_{global} = F^{d,l}_{global} \odot \sigma\Big(\psi^{l\to g}_{\omega_d}\big([F^{d,l}_{local} \parallel E^{d,l}_{emb}]\big)\Big), \tag{8}$$

where $\psi^{g\to l}_{\omega_d}$ and $\psi^{l\to g}_{\omega_d}$ are direction-specific gating networks, $\parallel$ denotes channel-wise concatenation, and σ is the Sigmoid function. This reciprocal design is the key mechanism that allows local structural details and global context to adaptively regulate one another.

After concatenation and linear fusion of $\widehat{F}^{d,l}_{local}$ and $\widehat{F}^{d,l}_{global}$, the direction-specific mapping outputs are

$$\widehat{E}^l_{3D} = \mathcal{M}^l_{2D\to3D}\big(E^l_{2D};\Theta_{2D\to3D}\big), \tag{9}$$

$$\widehat{E}^l_{2D} = \mathcal{M}^l_{3D\to2D}\big(E^l_{3D};\Theta_{3D\to2D}\big), \tag{10}$$

with the two directions parameterized independently and without weight sharing. The Mamba-based global branch Gu and Dao (2024) scales linearly with spatial sequence length, while reciprocal cross-modulation and direction/depth conditioning provide the adaptive alignment mechanism.

Asymmetric Bias Initialization. To provide the reciprocal modulation with a lightweight directional prior, we further adopt asymmetric bias initialization for the gating networks. For $2D \to 3D$, the Mamba and CNN gate biases are initialized to $+1.5$ and -0.5, respectively, favoring global geometric modeling; for $3D \to 2D$, the initialization is reversed to emphasize local appearance details. These biases only provide an initialization prior, while all gating parameters remain fully learnable and are adaptively refined together with the direction- and depth-aware embedding.

Anomaly Scoring. During the inference phase, the contrastive learning module is disabled. Relying on the cross-modulation predictions ($\widehat{E}_{3D}$ and $\widehat{E}_{2D}$), the pixel-level anomaly score $S(i,j)$ is quantified by the alignment error between the original and predicted features via cosine similarity:

$$S^l_{2D\to3D} = 1 - \cos(\widehat{E}^l_{3D}, E^l_{3D}), \quad S^l_{3D\to2D} = 1 - \cos(\widehat{E}^l_{2D}, E^l_{2D}) \tag{11}$$

We fuse the pixel-level anomaly score maps across network depths and both mapping directions through Layer Fusion and Modality Fusion. Gaussian smoothing is then applied to obtain the final high-resolution anomaly localization map ($R^{H_0\times W_0}$), matching the resolution of the input RGB image ($R^{H_0\times W_0\times 3}$). The image-level anomaly score is defined as the maximum value in this localization map.

Table 1: I-AUROC on MVTec 3D-AD Bergmann et al. (2022) across all 10 classes. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Bagel	Cable	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
BTF Horwitz and Hoshen (2023)	0.918	0.748	0.967	0.883	0.932	0.582	0.896	0.912	0.921	0.886	0.865
AST Rudolph et al. (2023)	0.983	0.873	0.976	0.971	0.932	0.885	0.974	**0.981**	**1.000**	0.797	0.937
M3DM Wang et al. (2023)	0.994	0.909	0.972	0.976	0.960	0.942	0.973	0.899	0.972	0.850	0.945
CFM Costanzino et al. (2024)	0.994	0.888	0.984	0.993	0.980	0.888	0.941	0.943	0.980	0.953	0.954
LSFA Tu et al. (2024)	**1.000**	0.939	0.982	0.989	0.961	0.951	0.983	0.962	0.989	0.951	0.971
3D-ADNAS Long et al. (2025)	0.997	**1.000**	0.971	0.986	0.966	0.948	0.897	0.873	**1.000**	0.867	0.951
2M3DF Asad et al. (2025)	–	–	–	–	–	–	–	–	–	–	0.966
MAFR Ali et al. (2025)	0.997	0.934	0.987	0.989	0.983	0.945	0.978	0.960	0.992	0.963	0.972
HFMM-Net Zhao et al. (2025)	0.995	0.920	0.985	0.995	0.974	0.900	0.960	0.933	0.980	0.920	0.956
PIRN Li et al. (2026b)	0.971	0.973	0.941	0.957	0.975	**0.993**	**0.992**	0.950	0.996	0.880	0.963
Ours	0.999	0.941	**0.993**	**0.999**	**0.994**	0.980	0.975	0.951	0.998	**0.990**	**0.982**

Table 2: Mean results on Eyecandies Bonfiglioli et al. (2022) across all 10 classes. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	I-AUROC	P-AUROC	AUPRO@1%
BTF Horwitz and Hoshen (2023)	0.740	0.883	0.261
AST Rudolph et al. (2023)	0.780	0.902	0.149
M3DM Wang et al. (2023)	0.882	0.977	0.331
CFM Costanzino et al. (2024)	0.881	0.974	0.335
3D-ADNAS Long et al. (2025)	–	0.970	–
2M3DF Asad et al. (2025)	0.897	–	–
HFMM-Net Zhao et al. (2025)	0.887	0.966	0.341
Ours	**0.899**	**0.979**	**0.348**

Table 3: Mean localization metrics on MVTec 3D-AD Bergmann et al. (2022) across all 10 classes. The best and second-best reported results are highlighted in **bold** and underlined, respectively.

Method	P-AUROC	AUPRO@30%	AUPRO@1%
BTF Horwitz and Hoshen (2023)	0.992	0.959	0.383
AST Rudolph et al. (2023)	0.976	0.944	0.398
M3DM Wang et al. (2023)	0.992	0.964	0.394
LSFA Tu et al. (2024)	**0.993**	0.968	–
MAFR Ali et al. (2025)	**0.993**	**0.970**	–
Ours	0.992	0.968	**0.456**

4 Experiments

4.1 Experimental Setup

Datasets. MVTec 3D-AD Bergmann et al. (2022): This industrial dataset comprises over 4,147 high-resolution 3D point clouds and corresponding RGB images across 10 categories, covering diverse geometric and textural anomalies caused by manufacturing defects. **Eyecandies Bonfiglioli et al. (2022)**: This dataset contains 10 categories of candy objects rendered under complex lighting and highly challenging backgrounds, specifically designed to test model robustness in noisy environments.

Metrics. We employ four standard evaluation metrics: the Area Under the Receiver Operating Characteristic curve for image-level anomaly detection (I-AUROC); the pixel-level AUROC for anomaly localization (P-AUROC); the Area Under the Per-Region Overlap curve up to 30% and a stringent 1% false positive rate (AUPRO@30% and AUPRO@1%).

Training Protocol. All reported MambaX-AD results follow a category-specific training protocol: one model is trained independently for each object category using only the normal samples of that category.

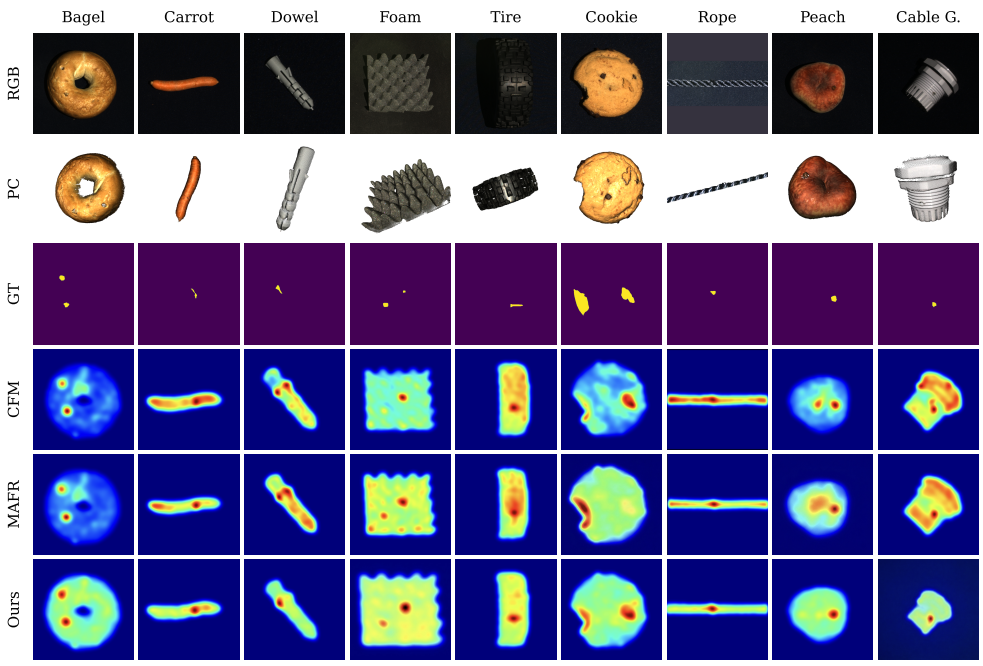


Figure 4: MVTec 3D-AD Bergmann et al. (2022) qualitative comparison on identical test samples. From top to bottom: RGB images, point clouds, ground-truth, anomaly maps from CFM Costanzino et al. (2024), anomaly maps from MAFR Ali et al. (2025), and anomaly maps from MambaX-AD.

Implementation Details. For the contrastive learning, we use the absolute hyperparameter values $\alpha_{local} = 3$, $\gamma = 2$, and $\alpha_{bf} = 1$. Furthermore, we adopt dataset-specific training schedules tailored to their respective convergence behaviors. On the MVTec 3D-AD dataset, the model is trained for 250 epochs with an initial learning rate of 5×10^{-4} and evaluated every 50 epochs. For the Eyecandies dataset, the training process is set to 100 epochs with a lower initial learning rate of 1×10^{-4}, and the validation interval is 10 epochs.

4.2 Quantitative Results

MVTec 3D-AD Bergmann et al. (2022). As shown in Table 1, our method achieves the highest mean image-level I-AUROC among the compared methods (0.982), outperforming recent strong baselines such as MAFR (0.972). Table 3 further shows that MambaX-AD maintains competitive localization performance, achieving 0.992 P-AUROC, 0.968 AUPRO@30%, and 0.456 AUPRO@1%. The full five-seed evaluation in Appendix C further confirms the stability of the image-level improvement. Overall, these results show that MambaX-AD strengthens image-level anomaly discrimination while preserving competitive localization quality, consistent with RA-CL constraining the normal-feature boundary and reciprocal cross-modulation improving cross-modal prediction.

Eyecandies Bonfiglioli et al. (2022). Table 2 evaluates robustness under complex backgrounds and variable lighting. Among the compared methods, MambaX-AD achieves the strongest I-AUROC (0.899), P-AUROC (0.979), and AUPRO@1% (0.348). The consistent performance under substantially different appearance and geometry further supports the robustness of RA-CL's region-aware discrimination and reciprocal cross-modulation's adaptive local–global alignment.

4.3 Qualitative Analysis

As visualized in Figure 4, we compare against both CFM Costanzino et al. (2024), the closest cross-modal mapping baseline, and the stronger recent MAFR method Ali et al. (2025) on identical test samples. Relative to both baselines, MambaX-AD produces anomaly responses that are more concentrated around the ground-truth defect regions with less diffuse background activation. These

Table 4: Macro ablation on MVTec 3D-AD Bergmann et al. (2022). MS: multi-scale features; Std-CL/RA-CL: standard/region-aware contrastive learning; M-CNN: Mamba-CNN branches; CM: cross-modulation; AE: aware embedding. Best results are in **bold**.

Components						Metrics	
MS	Std-CL	RA-CL	M-CNN	CM	AE	I-AUROC (↑)	AUPRO@1% (↑)
						0.954	0.455
✓						0.962	0.448
✓	✓					0.972	0.455
✓	✓		✓			0.977	0.455
✓		✓	✓			0.978	0.449
✓		✓	✓	✓		0.980	0.454
✓		✓	✓	✓	✓	**0.982**	**0.456**

Table 5: Ablation of pseudo-anomaly targets on MVTec 3D-AD Bergmann et al. (2022). The first row uses no contrastive learning; best results are in **bold**.

Pseudo-Anomaly Strategy		Metrics	
Gaussian Noise	CutPaste	I-AUROC (↑)	AUPRO@1% (↑)
		0.962	0.448
✓		0.966	0.449
	✓	0.964	0.449
✓	✓	**0.972**	**0.455**

Table 6: Ablation of cross-modulation on MVTec 3D-AD Bergmann et al. (2022). $L \rightarrow G$ denotes local-to-global modulation and $G \rightarrow L$ the reverse; best results are in **bold**.

Strategy	Modulation		Aware Embed.	Metrics		
	$L \rightarrow G$	$G \rightarrow L$		I-AUROC	AUPRO@30%	AUPRO@1%
Direct Addition				0.978	0.965	0.449
Uni-Modulation	✓			0.979	0.966	0.453
Bi-Modulation	✓	✓		0.980	0.967	0.454
Bi-Mod. + Embed.	✓	✓	✓	**0.982**	**0.968**	**0.456**

sharper and cleaner responses qualitatively support the intended benefits of RA-CL for defect-sensitive discrimination and reciprocal cross-modulation for cross-modal prediction.

4.4 Ablation Study

To isolate and quantify the contributions of individual modules, we conduct a series of ablation analyses on the MVTec 3D-AD Bergmann et al. (2022) dataset.

Macro Ablation. Table 4 traces the contribution of each stage of the framework. Multi-scale features and standard contrastive learning establish a 0.972 I-AUROC reference, and adding the local/global Mamba-CNN branches raises it to 0.977. Replacing the standard objective with RA-CL improves I-AUROC to 0.978. Reciprocal cross-modulation then raises the result to 0.980/0.454 (I-AUROC/AUPRO@1%), and direction- and depth-aware conditioning further improves the complete model to 0.982/0.456. The consistent progression directly supports the effectiveness of the proposed mask-conditioned contrastive formulation and reciprocal cross-modulation.

Pseudo-Anomaly Targets. Table 5 examines the negative targets used by RA-CL. Gaussian noise and Feature CutPaste provide complementary feature-space perturbations, and their combination yields the strongest result in this comparison. This combination therefore provides a stronger set of negative targets for the region-aware, mask-conditioned boundary regularization in RA-CL.

Reciprocal Cross-Modulation. Table 6 isolates the proposed information-flow mechanism. Directly combining local and global branches is weaker than learned modulation; unidirectional modulation improves the result, while reciprocal modulation ($L \rightarrow G$ and $G \rightarrow L$) improves it further. Direction-

Table 7: Ablation of asymmetric bias and aware embedding on MVTec 3D-AD.

Cross-mod.	Asym. Bias	Aware Embed.	I-AUROC	AUPRO@1%
✓	×	×	0.979	0.453
✓	✓	×	0.980	0.454
✓	✓	✓	**0.982**	**0.456**

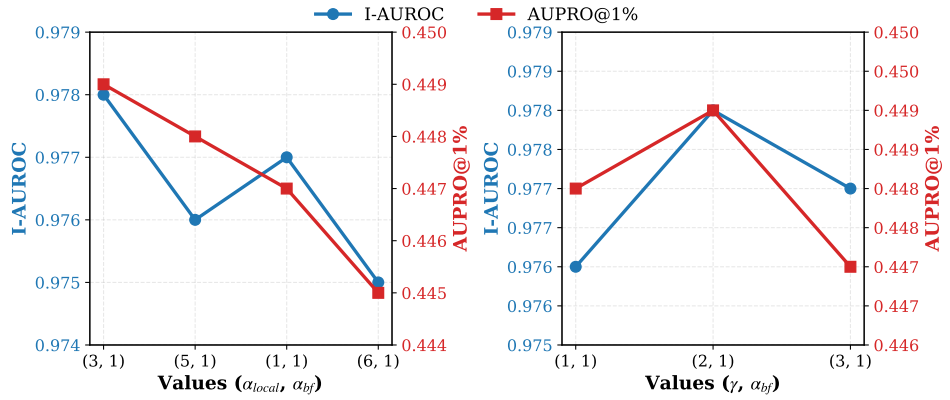


Figure 5: Coupled hyperparameter sensitivity analysis on MVTec 3D-AD Bergmann et al. (2022). **Left:** Effect of varying the local amplification coefficient α_{local} while fixing $\gamma = 2$ and $\alpha_{bf} = 1$. **Right:** Effect of varying the distance-decay coefficient γ while fixing $\alpha_{local} = 3$ and $\alpha_{bf} = 1$.

and depth-aware embeddings provide an additional gain by conditioning the reciprocal gates on the specific cross-modal mapping context. This progression verifies that reciprocal local–global modulation is more effective than direct or one-way interaction for cross-modal alignment.

Asymmetric Bias vs. Aware Embedding. In addition to the learnable aware embedding, our reciprocal cross-modulation employs a lightweight asymmetric bias initialization to inject a direction-specific prior into the modulation gates. Table 7 disentangles the contributions of these two designs under the same reciprocal cross-modulation architecture. Starting from reciprocal cross-modulation with symmetric initialization, the asymmetric bias improves I-AUROC from 0.979 to 0.980 and AUPRO@1% from 0.453 to 0.454. Incorporating the aware embedding further improves the performance to 0.982 I-AUROC and 0.456 AUPRO@1%. These results show that the two designs are complementary: the asymmetric bias provides a useful directional initialization prior, while the aware embedding enables adaptive direction- and depth-aware conditioning. Together, they strengthen the proposed reciprocal cross-modulation by better exploiting the asymmetric local and global representations of 2D and 3D modalities.

Hyperparameter Sensitivity. Figure 5 evaluates the absolute penalty values with $\alpha_{bf} = 1$. Increasing α_{local} initially improves sensitivity, while overly large values (e.g., 5 or 6) can disrupt normal alignment. A moderate decay $\gamma = 2$ best emphasizes hard pseudo-negatives without over-penalizing easy ones, giving the selected setting $\alpha_{local} = 3$, $\gamma = 2$, $\alpha_{bf} = 1$. The smooth behavior around this point supports the robustness of RA-CL's mask-conditioned dynamic penalty.

5 Conclusion

We presented MambaX-AD to address weak discriminative sensitivity and the cross-modal alignment bottleneck in normal-only cross-modal feature mapping. RA-CL uses region-aware, mask-conditioned repulsion to constrain the normal-feature boundary, while reciprocal Mamba-CNN cross-modulation adaptively couples local and global representations under direction- and depth-aware conditioning. MambaX-AD achieves the highest mean image-level I-AUROC on MVTec 3D-AD with competitive localization performance, and the highest results on multiple metrics on Eyecandies, supporting the effectiveness of both mechanisms.

AI USE STATEMENT

Large language models were used mainly to assist with language polishing, grammatical correction and format layout. The research problem, method design, implementation, experiments, data analysis, and scientific conclusions were developed and verified by the authors. All AI-assisted content was manually reviewed, and the authors take full responsibility for the final content of this work.

REFERENCES

Usman Ali, Ali Zia, Abdul Rehman, Umer Ramzan, Zohaib Hassan, Talha Sattar, Jing Wang, and Wei Xiang. 2D-3D feature fusion via cross-modal latent synthesis and attention-guided restoration for industrial anomaly detection. In *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, 2025. doi: 10.1109/DICTA68720.2025.11302453.

Mujtaba Asad, Waqar Azeem, He Jiang, Hafiz Tayyab Mustafa, Jie Yang, and Wei Liu. 2M3DF: Advancing 3D industrial defect detection with multi-perspective multimodal fusion network. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(7):6803–6815, 2025. doi: 10.1109/TCSVT.2025.3536475.

Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The MVTec 3D-AD dataset for unsupervised 3D anomaly detection and localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, volume 5, pages 202–213, 2022. doi: 10.5220/0010865000003124.

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 (ACCV)*, pages 3586–3602, 2022.

Jiayi Cheng, Can Gao, Jie Zhou, Jiajun Wen, Tao Dai, and Jinbao Wang. MC3D-AD: A unified geometry-aware reconstruction model for multi-category 3D anomaly detection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 837–845, 2025. doi: 10.24963/ijcai.2025/94.

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)*, pages 17234–17243, 2024. doi: 10.1109/CVPR52733.2024.01631.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

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

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1705–1714, 2019.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024. URL `https://openreview.net/forum?id=tEYskw1VY2`.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. AnomalyGPT: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1932–1940, 2024. doi: 10.1609/aaai.v38i3.27963.

Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8472–8480, 2024. doi: 10.1609/aaai.v38i8.28690.

Eliahu Horwitz and Yedid Hoshen. 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 Workshops (CVPRW)*, pages 2968–2977, 2023. doi: 10.1109/CVPRW59228.2023.00298.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. CutPaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9664–9674, 2021. doi: 10.1109/CVPR46437.2021.00954.

Kaiqiang Li, Gang Li, Mingle Zhou, Min Li, Delong Han, and Jin Wan. Back to point: Exploring point-language models for zero-shot 3D anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14167–14177, 2026a.

Yiting Li, Fayao Liu, Jingyi Liao, Sichao Tian, Chuan-Sheng Foo, and Xulei Yang. FIND: Few-shot anomaly inspection with normal-only multi-modal data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23290–23299, 2025. doi: 10.1109/ICCV51701.2025.02162.

Yiting Li, Xulei Yang, Jing Zhang, Sichao Tian, Jingyi Liao, and Fayao Liu. PIRN: Prototypical-based intra-modal reconstruction with normality communication for multi-modal anomaly detection. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026b. URL https://openreview.net/forum?id=7L7kmHHfgf.

Kailun Liao, Jianfeng Yang, Tao Tao, Wenfei Wu, Jiaming Jiang, and Jinsheng Xiao. Multi-prototype compactness and boundary-aware synthesis for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28524–28533, 2026.

Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21:104–135, 2024. doi: 10.1007/s11633-023-1459-z.

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 (CVPR)*, pages 20402–20411, 2023. doi: 10.1109/CVPR52729.2023.01954.

Kaifang Long, Guoyang Xie, Lianbo Ma, Jiaqi Liu, and Zhichao Lu. Revisiting multimodal fusion for 3D anomaly detection from an architectural perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12273–12281, 2025. doi: 10.1609/aaai.v39i12.33337.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision (ECCV)*, pages 604–621. Springer, 2022.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14318–14328, 2022. doi: 10.1109/CVPR52688.2022.01392.

Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2592–2602, 2023.

Yuanpeng Tu, Boshen Zhang, Liang Liu, Yuxi Li, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Cairong Zhao. Self-supervised feature adaptation for 3D industrial anomaly detection. In *European Conference on Computer Vision (ECCV)*, pages 75–91, 2024. doi: 10.1007/978-3-031-72627-9_5.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22883–22892, 2024.

Gang Wang, Yisheng Zou, Songlin He, Yakun Wang, and Ruihong Dai. Anomaly detection and localization via reverse distillation with latent anomaly suppression. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(10):9592–9607, 2025. doi: 10.1109/TCSVT.2025.3562258.

Jiaxiang Wang, Haote Xu, Xiaolu Chen, Haodi Xu, Yue Huang, Xinghao Ding, and Xiaotong Tu. Exploiting point-language models with dual-prompts for 3D anomaly detection. *Expert Systems with Applications*, 298:129758, 2026. doi: 10.1016/j.eswa.2025.129758.

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)*, pages 8032–8041, 2023. doi: 10.1109/CVPR52729.2023.00776.

Rongcheng Wu, Hao Zhu, Shiying Zhang, Mingzhe Wang, Zhidong Li, Hui Li, Jianlong Zhou, Jiangtao Cui, Fang Chen, Pingyang Sun, Qiyu Liao, and Ye Lin. RcAE: Recursive reconstruction framework for unsupervised industrial anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10736–10744, 2026. doi: 10.1609/aaai.v40i13.38048.

Yuyang Yu, Zhengwei Chen, Xuemiao Xu, Lei Zhang, Haoxin Yang, Yongwei Nie, and Shengfeng He. Registration is a powerful rotation-invariance learner for 3D anomaly detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pages 5806–5832, 2025. doi: 10.52202/085713-0180.

Guo Zhao, Liang Tan, Musong He, and Qi Wu. HFMM-Net: A hybrid fusion Mamba network for efficient multimodal industrial defect detection. *Information*, 16(12):1018, 2025. doi: 10.3390/info16121018.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16259–16268, 2021.

Appendix

A Compute Resources and Dataset Preprocessing

A.1 Compute Resources

All experiments were conducted on a workstation equipped with two NVIDIA GeForce RTX 4090 GPUs (24GB VRAM each). The framework is implemented using PyTorch 2.0.1 and CUDA 12.4, utilizing Distributed Data Parallel (DDP) to accelerate the training process. On average, the training time for each individual category is approximately 1.5 hours.

A.2 Dataset Preprocessing

Due to the inherent noise and complex backgrounds in raw 3D datasets, we apply specialized preprocessing pipelines before feature extraction. For the MVTec 3D-AD dataset, we employ the RANSAC algorithm with a strict distance threshold of 0.005 to accurately estimate and remove background planes. Subsequently, to eliminate isolated noise and outliers, we perform connected component cleaning using DBSCAN clustering, configured with an ϵ of 0.006 and a minimum cluster size of 30 points.

For the Eyecandies dataset, the raw depth maps are first converted into 3D point clouds utilizing the established camera intrinsics with a focal length of 711.11. We then apply a predefined 3D rotation matrix and targeted spatial coordinate thresholding to explicitly strip away the complex background environments.

B Additional Architecture Details

Feature Extraction Backbones. Following the standard practice in cross-modal mapping Costanzino et al. (2024), we employ the ViT-Base/8 architecture Dosovitskiy et al. (2021) pre-trained on ImageNet as our 2D backbone. For the 3D modality, we utilize a pre-trained Point Transformer Zhao et al. (2021). Both backbones are strictly frozen during the entire training process. We extract intermediate features from multi-scale layers to construct the unified semantic representations.

Bidirectional Mamba and Perceptual Embedding. In the global branch, the State Space Models are implemented as bidirectional Mamba blocks, configured with a state dimension of $d_{state} = 16$, a convolution kernel size of $d_{conv} = 4$, and an expansion factor of 2. The dynamic perceptual embedding (E_{emb}) is computed by extracting the global spatial context via mean pooling, which is then processed by a contextual MLP comprising two linear layers separated by a SiLU activation. This dynamic vector is subsequently added to a static base embedding to form the final direction- and depth-aware learnable embeddings.

Gating Networks (ψ_ω). The global-to-local and local-to-global gating networks are implemented as lightweight Multi-Layer Perceptrons (MLPs). The $2D \rightarrow 3D$ and $3D \rightarrow 2D$ mappings use independent, non-shared gate parameters, consistent with the direction-specific mappings defined in the main text. Each gating network receives a concatenated input of feature representations and perceptual embeddings, processes it through a hidden bottleneck dimension ($C/2$), and outputs spatial modulation masks via the Sigmoid function.

Crucially, to explicitly model the semantic asymmetry between the 2D and 3D modalities, we introduce a direction-aware prior bias initialization for the final linear layers of these gates:

- For the $\mathcal{M}_{2D\rightarrow 3D}$ mapping, we initialize the bias of the Mamba gate to $+1.5$ to favor global geometric structures, while the CNN gate bias is suppressed to -0.5.
- Conversely, for the $\mathcal{M}_{3D\rightarrow 2D}$ mapping, the local CNN gate is strongly favored with a bias of $+1.5$ to emphasize fine-grained 2D textures, while the Mamba gate bias is suppressed to -0.5.

This asymmetrical initialization serves as a direction-aware optimization prior; Table 7 shows that the learnable aware embedding provides further gains with this prior.

Table C.1: Five-seed results on all ten MVTec 3D-AD categories. Values are mean ± standard deviation over five independent runs.

Category	I-AUROC	P-AUROC	AUPRO@30%	AUPRO@1%
Cookie	0.9992 ± 0.0001	0.9632 ± 0.0002	0.9294 ± 0.0003	0.3734 ± 0.0002
Tire	0.9901 ± 0.0002	0.9982 ± 0.0001	0.9825 ± 0.0004	0.4463 ± 0.0003
Bagel	0.9992 ± 0.0004	0.9925 ± 0.0001	0.9685 ± 0.0001	0.4461 ± 0.0002
Cable Gland	0.9415 ± 0.0004	0.9961 ± 0.0003	0.9752 ± 0.0004	0.4663 ± 0.0003
Carrot	0.9931 ± 0.0001	0.9982 ± 0.0006	0.9824 ± 0.0005	0.4792 ± 0.0006
Dowel	0.9942 ± 0.0002	0.9831 ± 0.0007	0.9382 ± 0.0006	0.4523 ± 0.0003
Foam	0.9801 ± 0.0005	0.9962 ± 0.0001	0.9695 ± 0.0008	0.4757 ± 0.0005
Peach	0.9754 ± 0.0002	0.9983 ± 0.0001	0.9815 ± 0.0001	0.4821 ± 0.0003
Potato	0.9511 ± 0.0003	0.9982 ± 0.0004	0.9825 ± 0.0002	0.4521 ± 0.0006
Rope	0.9983 ± 0.0005	0.9982 ± 0.0003	0.9723 ± 0.0004	0.4845 ± 0.0005
Mean	$\mathbf{0.9822 \pm 0.0003}$	$\mathbf{0.9922 \pm 0.0003}$	$\mathbf{0.9682 \pm 0.0004}$	$\mathbf{0.4558 \pm 0.0002}$

C Additional Reliability, Sensitivity, and Efficiency Analyses

C.1 Five-Seed Reliability on MVTec 3D-AD

To assess whether the reported image-level improvement is robust to stochastic initialization and pseudo-anomaly sampling, we train MambaX-AD with five independent random seeds for all ten MVTec 3D-AD categories. For each seed, we first aggregate performance across all categories and then compute the mean and standard deviation over the five aggregate scores. As reported in Table C.1, MambaX-AD achieves I-AUROC 0.9822 ± 0.0003, P-AUROC 0.9922 ± 0.0003, AUPRO@30% 0.9682 ± 0.0004, and AUPRO@1% 0.4558 ± 0.0002. The very small standard deviations show that the overall performance is stable across different random seeds, and that the image-level gain is substantially larger than the measured seed-to-seed variation. This stability further indicates that the improvements brought by the region-aware contrastive formulation and reciprocal cross-modulation are systematic rather than artifacts of a particular initialization or pseudo-anomaly sampling sequence.

C.2 Details of RA-CL

We further evaluate the sensitivity of the RA-CL margin m while keeping all other settings fixed. Table C.2 shows that varying m from 0.3 to 0.5 changes I-AUROC only from 0.9817 to 0.9824, while AUPRO@1% remains within the narrow range of 0.4563–0.4565. The limited variation confirms that RA-CL does not rely on a finely tuned margin value within this reasonable range. This robustness supports the effectiveness of the proposed mask-conditioned contrastive formulation as a stable mechanism for shaping the normal-feature boundary rather than a gain driven by sensitive hyperparameter tuning.

Table C.2: Sensitivity to the RA-CL margin m on MVTec 3D-AD.

m	I-AUROC	AUPRO@1%
0.3	0.9817	0.4563
0.4	0.9821	0.4565
0.5	0.9824	0.4564

Optimization Behavior. We further characterize the repulsive behavior of the RA-CL negative term. For a corrupted location with $M_{\mathrm{pa}}^{(i)} = 1$ and $d_i^- < m$, Eq. (3)–(4) reduces to

$$\mathcal{L}_{\mathrm{neg}}^{(i)} = \alpha_{\mathrm{local}} \exp(-\gamma d_i^-)(m - d_i^-)^2. \tag{12}$$

Its derivative with respect to the negative distance is

$$\frac{\partial \mathcal{L}_{\mathrm{neg}}^{(i)}}{\partial d_i^-} = -\alpha_{\mathrm{local}} \exp(-\gamma d_i^-) \left[\gamma(m - d_i^-)^2 + 2(m - d_i^-)\right] < 0, \qquad 0 \le d_i^- < m. \tag{13}$$

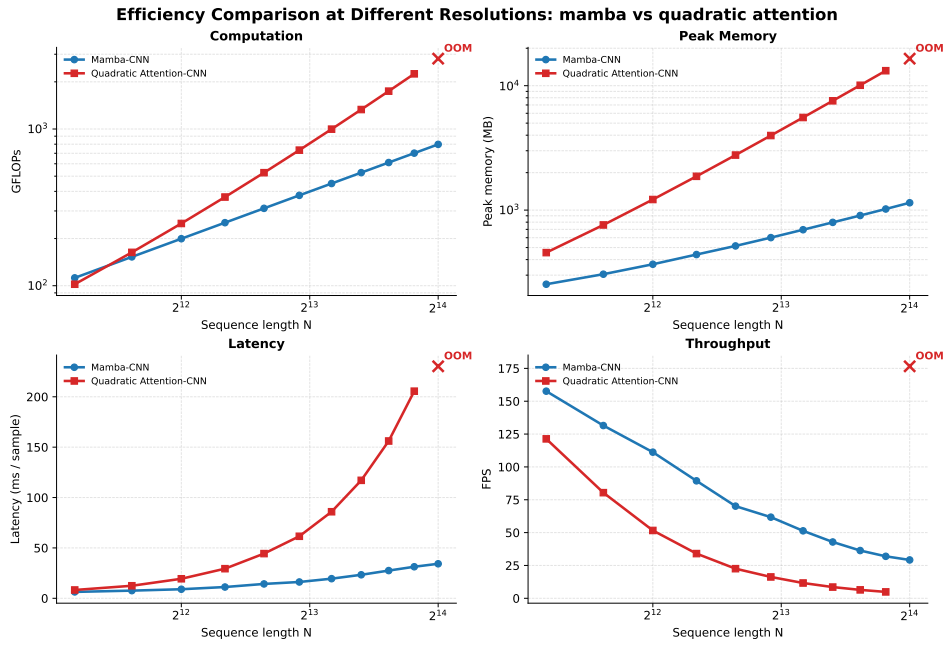


Figure C.1: Controlled scalability comparison between Mamba-CNN and a quadratic attention-CNN counterpart as the cross-modal mapping sequence length $N = H'W'$ increases. The four panels report GFLOPs, peak GPU memory, latency, and throughput under the same implementation settings. OOM denotes out-of-memory on a 24GB RTX 4090 GPU.

Therefore, minimizing the loss explicitly increases the distance between the cross-modal prediction $\widehat{E}^{(i)}$ and the pseudo-anomalous target $E^{-,(i)}$. Moreover, the repulsive gradient is strongest for hard pseudo-negatives close to the predicted normal feature and progressively decreases as d_i^- approaches m; once $d_i^- \geq m$, the hinge term becomes zero and no further repulsion is applied. This behavior prevents unnecessary separation of already well-separated negatives while focusing optimization on ambiguous pseudo-anomalies near the learned normal-feature boundary, providing a principled explanation for the distance-decayed, mask-conditioned design of RA-CL.

C.3 Efficiency and Scalability

We evaluate the computational role of the Mamba-based global branch with a controlled replacement experiment. The CNN branch, reciprocal cross-modulation structure, feature dimensions, batch size, numerical precision, hardware, and raw input resolution are kept fixed, while the global Mamba branch is replaced by standard quadratic self-attention. All measurements cover the complete inference pipeline from preprocessing and feature extraction to cross-modal mapping and anomaly scoring, using FP32, batch size 1, and a single RTX 4090 GPU. The raw input resolution is fixed at 224×224, while the horizontal axis in Figure C.1 denotes the sequence length $N = H'W'$ of the feature map entering the cross-modal mapping module.

Figure C.1 provides a direct view of the scaling behavior across increasing sequence lengths. As N grows, the quadratic-attention counterpart exhibits substantially steeper increases in computation, peak memory, and latency, together with a rapid throughput drop, whereas Mamba-CNN scales much more smoothly. At 120×120 feature resolution ($N = 14{,}400$), Mamba-CNN requires 701.41 GFLOPs, 1,020.43 MB peak memory, and 31.30 ms per sample, achieving 31.95 FPS; the attention counterpart requires 2,249.57 GFLOPs, 13,196.62 MB, and 205.57 ms, with only 4.86 FPS. At 128×128 ($N = 16{,}384$), the attention variant runs out of memory, while Mamba-CNN remains feasible at 798.05 GFLOPs, 1,145.59 MB, 34.25 ms, and 29.19 FPS. These trends show that the Mamba-based global branch provides the reciprocal cross-modulation architecture with substantially better scalability for dense cross-modal feature maps, enabling global modeling without incurring the quadratic cost of self-attention.

Table C.3: Inference efficiency comparison with representative multimodal anomaly-detection paradigms using RTX 4090 measurements for the compared implementations.

Method	I-AUROC	Inference memory (MB)	FPS
M3DM	0.945	6526.12	0.514
CFM	0.954	437.91	21.755
MambaX-AD	0.982	1145.59	29.19

Table D.1: Macro ablation on Eyecandies using the same component settings and absolute RA-CL hyperparameters ($\alpha_{local} = 3$, $\gamma = 2$, $\alpha_{bf} = 1$) as MVTec 3D-AD.

Configuration	I-AUROC	P-AUROC	AUPRO@1%
Baseline	0.881	0.974	0.335
+ RA-CL	0.889	0.978	0.339
+ reciprocal Mamba-CNN cross-modulation	0.895	0.979	0.347
+ aware embedding	**0.899**	**0.979**	**0.348**

To complement the controlled branch-level comparison, Table C.3 compares MambaX-AD with representative multimodal anomaly-detection paradigms. M3DM requires 6,526.12 MB inference memory and runs at 0.514 FPS, whereas MambaX-AD uses 1,145.59 MB and reaches 29.19 FPS. Compared with the lightweight CFM baseline, MambaX-AD improves throughput from 21.755 to 29.19 FPS while increasing memory from 437.91 MB to 1,145.59 MB, together with a higher I-AUROC of 0.982 versus 0.954. Therefore, MambaX-AD offers a favorable accuracy–efficiency trade-off: its memory-bank-free reciprocal cross-modal mapping remains considerably lighter than memory-bank methods while preserving the representational capacity needed for strong anomaly discrimination.

D Additional Results and Cross-Dataset Analysis on Eyecandies Bonfiglioli et al. (2022)

In the main manuscript (Table 2), we report the average anomaly detection and localization performance on Eyecandies Bonfiglioli et al. (2022). This section further examines cross-dataset component transfer, per-class behavior, and qualitative localization to provide a more complete evaluation of MambaX-AD under irregular geometries, non-uniform textures, complex illumination, and background clutter.

D.1 Cross-Dataset Macro Ablation

The component configuration and the absolute RA-CL hyperparameters $\alpha_{local} = 3$, $\gamma = 2$, and $\alpha_{bf} = 1$ are transferred directly from MVTec 3D-AD to Eyecandies without dataset-specific retuning. As shown in Table D.1, RA-CL improves the baseline I-AUROC from 0.881 to 0.889 and AUPRO@1% from 0.335 to 0.339. Adding reciprocal Mamba-CNN cross-modulation further raises I-AUROC to 0.895 and AUPRO@1% to 0.347, while the aware embedding yields the final 0.899 I-AUROC and 0.348 AUPRO@1%. The monotonic gains under unchanged hyperparameters demonstrate that both the region-aware contrastive formulation and reciprocal cross-modulation transfer consistently across datasets, with the aware embedding further strengthening direction- and depth-adaptive modulation.

D.2 Per-Class Quantitative Metrics

Table D.2 reports the detailed performance of MambaX-AD across all ten Eyecandies categories. The model achieves a mean I-AUROC of 0.899, P-AUROC of 0.979, AUPRO@30% of 0.889, and AUPRO@1% of 0.348. Although image-level difficulty varies across categories, pixel-level detection remains consistently strong, with P-AUROC ranging from 0.961 to 0.998; several categories also exhibit high image-level discrimination, including Marshmallow (1.000), Confetto (0.979), and Chocolate Cookie (0.966) in I-AUROC. These results indicate that the combination of region-aware

Table D.2: Detailed per-class results on the Eyecandies dataset Bonfiglioli et al. (2022). Performance of MambaX-AD evaluated using I-AUROC, P-AUROC, AUPRO@30%, and AUPRO@1%.

Category	I-AUROC (↑)	P-AUROC (↑)	AUPRO@30% (↑)	AUPRO@1% (↑)
Candy Cane	0.693	0.983	0.942	0.211
Chocolate Cookie	0.966	0.984	0.904	0.410
Chocolate Praline	0.935	0.963	0.813	0.339
Confetto	0.979	0.998	0.981	0.471
Gummy Bear	0.873	0.961	0.858	0.328
Hazelnut Truffle	0.843	0.963	0.800	0.251
Licorice Sandwich	0.931	0.966	0.777	0.337
Lollipop	0.829	0.980	0.902	0.218
Marshmallow	1.000	0.994	0.940	0.452
Peppermint Candy	0.936	0.995	0.976	0.458
Mean (Ours)	**0.899**	**0.979**	**0.889**	**0.348**

contrastive learning and reciprocal local–global cross-modulation preserves strong localization across diverse appearances and geometries, while the fused multimodal representation remains effective under substantial category variation.

D.3 Qualitative Visualizations

Figure D.1 presents representative anomaly localization results on Eyecandies Bonfiglioli et al. (2022), including the RGB input, 3D point cloud, ground-truth mask, and the 2D, 3D, and fused anomaly maps. Despite irregular object geometries, non-uniform textures, complex lighting, and background interference, the fused maps concentrate responses around defective regions while suppressing most background activations. This qualitative behavior is consistent with the complementary roles of our two core designs: reciprocal Mamba-CNN cross-modulation captures local structural details together with long-range context, while RA-CL sharpens sensitivity around the learned normal-feature boundary, jointly enabling precise multimodal anomaly localization.

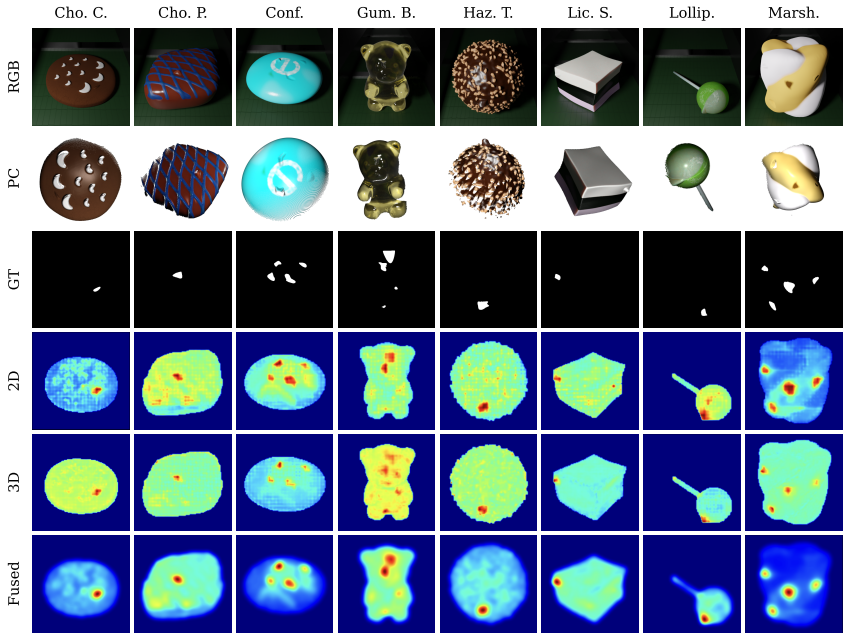


Figure D.1: Qualitative results on the Eyecandies dataset Bonfiglioli et al. (2022). We show RGB images, 3D point clouds, ground-truth segmentation masks, and the 2D, 3D, and fused anomaly maps generated by MambaX-AD under challenging lighting and background conditions.