Appendix

UniMMAD: Unified Multi-Modal and Multi-Class Anomaly Detection via MoE-driven Feature Decompression

A. Overview

The appendix includes the following sections to complement the main manuscript:

- Sec. B: Additional Experimental Details
- Sec. C: Dataset Details
- Sec. D: Methods and Baselines for Comparison
- Sec. E: More Quantitative Comparison
- Sec. F: Ablation on Number of Activated Experts
- Sec. G: Influence of Resolution and Backbone
- Sec. H: Unified vs. Specialized Training
- Sec. I: Per-Class Results
- Sec. J: More Qualitative Comparison

B. Additional Experimental Details

Experimental Details. All experiments were conducted over 300 epochs with a batch size of 10 on one NVIDIA RTX 4090 GPU. The number of input channels C to the general multi-modal encoder is set to 4. The Adam optimizer was used to optimize network parameters, with the model's learning rate set to 1×10^{-3}.

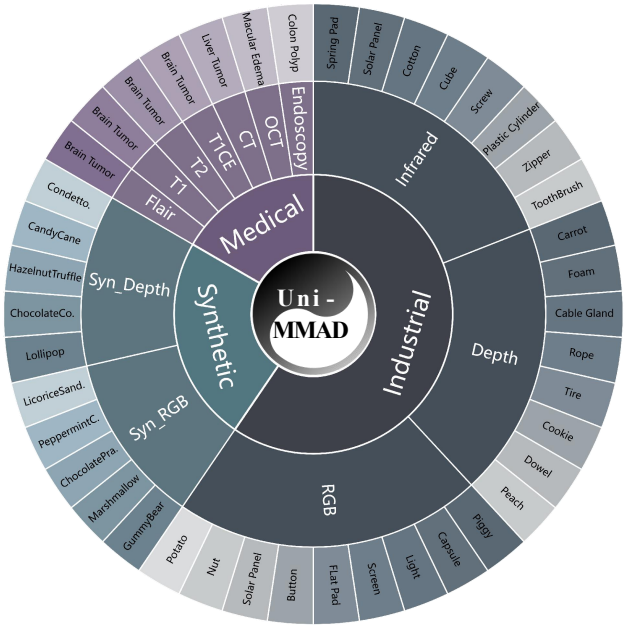

Figure S1. Dataset details for the multi-scene, multi-modal, multi-class benchmark. From the inner ring to the outer ring, the rings represent fields, modalities, and classes, respectively.

C. Dataset Details

Fig. S1 provides an overview of involved datasets, and the number of images in each dataset is listed in Tab.S2. We split the BraTS 2020 dataset into training and test sets, including both normal and abnormal brain images. The enhancing tumor, peritumoral edema, and necrotic/non-enhancing tumor core are collectively treated as the abnormal region. For the Retinal OCT and Liver CT datasets, we followed the BMAD protocol [1] and retained only the images with clearer tissue structures. Hyper-Kvasir, Retinal OCT, and Liver CT were grouped into a unified three-class medical dataset (UniMed). In MulSen-AD [15], only the infrared images were retained because the multi-modal images lacked proper calibration. All other datasets remain unchanged.

D. Methods and Baselines for Comparison

As shown in Tab. 1, We evaluate UniMMAD against SOTA specialized and generalist models across diverse domains under the same resolution, post-processing, and evaluation metrics. Specialized methods include M3DM, CFM, MulSen-TripleAD, and P+MMRD (PatchCore [20] + parameter-free fusion [10]), with WideResNet50 as the RGB encoder for memory-based models. Uni-modal baselines include RD, SimpleNet, as well as several multi-class methods such as UniAD, ViTAD, and MambaAD. All specialized models are tailored to specific modality combinations and evaluated on their respective datasets under a multi-class setting. We also evaluate generalist models AdaCLIP, MVFA, and AA-CLIP, which predict each modality independently and aggregate the results. Among the specialized models, we compared four representative multi-modal approaches—M3DM, CFM, MulSen-AD, and PatchCore+MMRD (PatchCore [20] + the parameter-free fusion strategy from MMRD [10]). For fairness, the RGB encoders of memory bank–based methods are kept consistent with our model, and both are implemented using WideResNet50. For uni-modal methods, we consider several widely adopted baselines, including RD, SimpleNet, PatchCore, UniAD, ViTAD, MambaAD, and INP-Former. These specialized methods are tailored to specific modality combinations and are evaluated independently on their corresponding datasets at a resolution of 256×256, whereas ViT-like architectures use a resolution of 252×252. We also evaluate recent generalist models, including AdaCLIP, MVFA, and AA-CLIP, which predict results for each modality separately and aggregate the outputs.

Table S1. More comprehensive comparison on the MVTec-AD and VisA datasets under the super–multi-class setting with a resolution of 256×256, evaluated using image-level metrics (AUC_I/AP_I/$MF1_I$) and pixel-level metrics (AUC_P/$MF1_P$/AUPRO).

Datasets	Publication	MVTec-AD		VisA	
Method↓	-	Image-level	Pixel-level	Image-level	Pixel-level
RD [7]	CVPR2022	95.8/97.8/95.0	95.1/51.5/90.7	88.4/89.9/86.7	96.8/38.7/87.0
UniAD [23]	NeurIPS2022	95.0/97.7/94.5	95.8/46.8/90.0	90.7/92.2/86.9	98.3/38.1/88.8
ViTAD [24]	CVIU2025	97.7/98.9/96.5	97.1/57.0/90.9	89.1/90.4/85.4	98.0/39.8/84.2
MambaAD [11]	NeurIPS2024	94.7/97.7/94.5	96.3/51.5/90.5	90.2/91.4/87.5	97.7/39.4/87.9
INP-Former [17]	CVPR2025	99.2/99.5/98.6	**98.2/60.7/93.8**	95.2/95.7/91.9	98.8/44.4/**91.5**
CCL [8]	ICCV2025	98.2/99.2/97.0	97.3/57.1/92.6	93.6/94.3/90.3	98.4/45.4/91.1
UniMMAD	Ours	**99.4/99.6/98.7**	98.1/60.2/93.0	**95.5/96.2/92.4**	**98.9/47.2**/91.3

E. More Quantitative Comparison

In this section, we compare additional metrics and comparison methods. As shown in Tab. S1, we compare with the latest method, CCL [8], using the same WideResNet50 backbone, and our approach achieves overall superior results.

F. Ablation on Number of Activated Experts

We conduct an ablation study on the number of activated routed experts. As shown in Fig. S2, as the number of activated experts increases, the model's image and pixel-level metrics show a slight improvement. However, each additional expert also increases the number of dynamic filtering operations, reducing inference efficiency. To balance inference efficiency and performance, we set the number of activated experts to 2. This is a commonly used activation number, adopted by many MoE-based methods [9, 14].

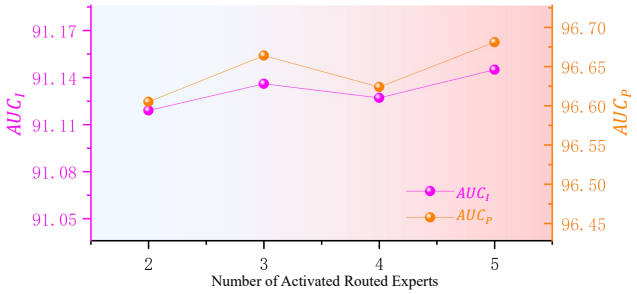

Figure S2. Impact of the number of activated routed experts on performance.

G. Influence of Resolution and Backbone

As shown in Fig .S3, we conducted an ablation study to assess the impact of input resolution and backbone architecture on model performance. Overall, larger image sizes and more powerful pre-trained models resulted in improved performance, demonstrating the scalability of our model in these aspects. This highlights the model's scalability in both input resolution and backbone architecture. Specifically, larger image sizes result in more significant improvements for the RGB-D datasets Eyecandies and MVTec-3D, likely due to the correlation with anomaly size and point cloud accuracy. Higher resolutions help the model capture finer defect localization. Larger backbones improve anomaly localization for medical lesions, likely due to the enhanced generalization ability of larger models in medical image tasks.

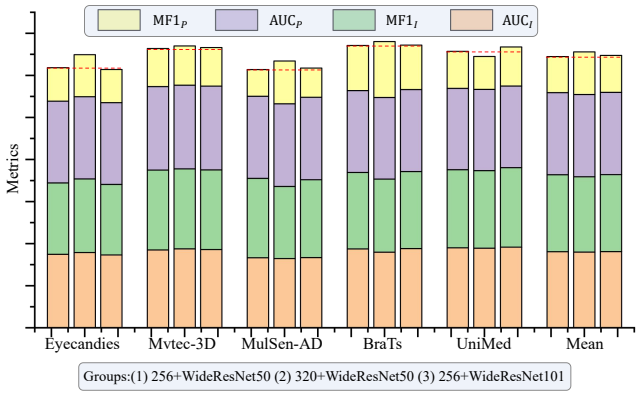

Figure S3. Impact of input resolution and backbone. Each group is divided into three settings: (1) 256×256 resolution + WideResNet50, (2) 320×320 resolution + WideResNet50, and (3) 256×256 resolution + WideResNet101.

H. Unified vs. Specialized Training

Unlike baselines suffering from parameter interference, UniMMAD maintains exceptional stability (Tab. S6) by effectively disentangling domain conflicts via adaptive routing. Additionally, unified training yields the gains on data-scarce infrared modalities (MulSen-AD) and complex 4-modality BraTs.

I. Per-Class Results

This section presents a class-wise comparison between the proposed method and recent state-of-the-art baselines.

Table S2. Overview of the Datasets.

Datasets		Fields	Modality	#Classes	Train		Test	
					#Items	#Images	#Items	#Images
MVTec-3D [3]		Industrial	RGB, Depth	10	2656	5312	1197	2394
Eyecandies [4]		Synthetic	Syn_RGB, Syn_Depth	10	8751	17502	500	1000
MulSen-AD [15]		Industrial	Infrared	15	1391	1391	644	644
MVTec-AD [2]		Industrial	RGB	15	3629	3629	2180	2180
VisA [25]		Industrial	RGB	12	8659	8659	2162	2162
Brats2020		Medical	Flair, T1, T2, T1CE	1	1183	4732	167	668
UniMed	Hyper-Kvasir [21]	Medical	Endoscopy	1	2020	2020	184	184
	Retinal OCT [12]	Medical	OCT	1	1009	1009	270	270
	Liver CT [1]	Medical	CT	1	404	404	158	158

Table S3. Per-Class comparison on the BraTs2020 Dataset.

Methods	Publication	$AUC_I(\%)$	$AP_I(\%)$	$MF1_I$ (%)	AUC_P (%)	$MF1_P$ (%)	AUPRO (%)
Specialized Models (One model for one dataset)							
PatchCore [20]+MMRD [10]	-	91.8	95.7	90.5	95.7	46.7	82.6
Generalist Models (One model performs all AD tasks without fine-tuning)							
AdaCLIP [5]	ECCV'24	70.7	81.7	81.6	96.6	48.4	78.2
MVFA [13]	CVPR'24	63.7	78.7	79.9	94.3	33.4	67.5
AA-CLIP [18]	CVPR'25	57.6	69.0	80.2	95.0	30.0	77.2
Unified Multi-modal and Multi-class Model							
UniMMAD	Ours	**95.8**	**98.0**	**91.1**	**97.5**	**51.4**	**84.4**

Table S4. Per-Class comparison on the MVTec-AD [2] Dataset for image-level metrics: AUC_I/AP_I / $MF1_I$.

Method→ Class↓	RD [7] CVPR'22	UniAD [23] NeurIPS'22	ViTAD [24] CVIU'25	MambaAD [11] NeurIPS'24	INP-Former [17] CVPR'25	UniMMAD Ours
Capsule	91.4/98.0/93.9	85.8/96.6/91.9	94.0/98.6/95.4	84.7/96.2/93.3	97.4/98.2/96.7	**97.6/99.3/97.2**
Carpet	97.9/99.4/96.0	99.9/99.9/99.4	99.4/99.8/99.4	96.7/99.0/95.4	**100./100./100.**	99.3/99.7/98.3
Grid	92.3/97.2/93.2	97.2/99.1/96.4	99.3/99.7/99.1	96.8/99.0/97.2	**100./100./100.**	**100.**/99.7/**100.**
Leather	**100./100./100.**	**100./100./100.**	**100./100./100.**	**100./100./100.**	**100./100./100.**	**100.**/99.8/**100.**
Tile	98.5/99.4/96.4	99.1/99.6/97.0	**100./100./100.**	99.4/99.7/97.6	**100./100./100.**	**100.**/99.8/**100.**
Wood	99.2/99.7/**98.3**	98.2/99.4/97.5	98.6/99.5/96.7	**99.5/99.8/98.3**	99.2/99.1/98.1	98.8/99.4/97.6
Bottle	99.7/99.9/98.4	99.9/99.9/99.2	**100./100./100.**	99.4/99.8/98.4	**100./100./100.**	**100.**/99.8/**100.**
Cable	78.9/86.8/80.5	84.5/91.9/81.2	97.3/98.4/93.3	85.4/91.2/83.7	99.5/**99.9/98.4**	**99.6**/99.5/97.3
Hazelnut	99.8/99.9/98.5	99.9/99.9/99.2	99.1/99.5/96.4	99.9/99.9/99.2	**100./100./100.**	**100.**/99.7/**100.**
Metal nut	99.1/99.7/97.3	99.3/99.8/97.8	99.3/99.8/98.4	97.5/99.4/95.7	**100./100./100.**	**100.**/99.8/**100.**
Pill	95.7/99.2/95.7	94.4/98.9/94.5	96.9/99.4/96.1	88.2/97.6/93.1	**98.3**/99.1/96.8	97.9/**99.6/97.1**
Screw	96.1/98.7/94.9	83.1/93.0/88.0	87.1/95.5/88.2	89.0/96.0/90.1	95.0/98.2/94.0	**98.2/99.4/97.0**
Toothbrush	99.1/99.6/98.3	91.9/96.4/93.7	99.7/99.8/98.3	97.7/99.1/95.2	**100./100./100.**	99.4/99.3/98.3
Transistor	88.8/88.3/85.0	93.7/91.0/84.4	95.9/94.2/88.6	89.2/88.8/83.3	99.0/98.3/95.3	**99.6/98.7/98.7**
Zipper	99.3/99.8/98.7	97.6/99.3/97.1	97.6/99.2/97.5	96.5/98.9/96.7	**100./100./100.**	99.6/99.8/99.1
Mean	95.7/97.7/95.0	95.0/97.7/94.5	97.7/98.9/96.5	94.7/97.6/94.5	99.2/99.5/98.6	**99.4/99.6/98.7**

Tabs. S4–S8 illustrate that our method outperforms mainstream multi-class approaches on the widely used datasets MVTec-AD and VisA. Tab. S9 and Tab.S10 summarize image-level and pixel-level results on MVTec-3D [3]. The proposed approach attains the best scores on categories dominated by geometric defects (e.g., Potato) and on anomalies that benefit from multi-modal cues (e.g., Carrot). On the synthetic Eyecandies benchmark (Tabs. S11–S12), our model achieves high accuracy on the ChocolateCookie and PeppermintCandy classes. These results indicate that the method effectively exploits complementary RGB–Depth information to pinpoint fine-grained anomalies. Tab. S14 and S15 report infrared-based results on MulSen-AD, highlighting the model's strength in detecting subsurface defects. In addi-

Table S5. Per-Class comparison on the MVTec-AD [2] dataset for pixel-level metrics: AUC_P/$MF1_P$ / AUPRO.

Method→ Class↓	RD [7] CVPR'22	UniAD [23] NeurIPS'22	ViTAD [24] CVIU'25	MambaAD [11] NeurIPS'24	INP-Former [17] CVPR'25	UniMMAD Ours
Capsule	98.5/47.5/95.0	98.5/48.3/92.5	97.9/47.9/92.2	98.6/44.4/94.3	**99.0/52.1/95.6**	98.9/50.2/92.5
Carpet	98.9/60.2/95.4	98.5/56.1/**95.6**	98.8/**62.2**/94.2	98.5/56.4/93.8	**99.1**/60.8/95.3	99.0/**62.2**/95.1
Grid	98.8/43.3/**96.1**	96.2/28.4/90.8	98.4/36.5/95.3	98.8/46.8/96.0	98.4/40.4/94.6	**99.0/47.2**/95.5
Leather	99.3/47.0/97.7	98.9/43.5/**98.1**	**99.5/53.5**/97.9	99.1/43.7/97.6	99.2/42.9/96.1	99.4/52.8/97.4
Tile	94.4/57.1/84.0	92.4/51.6/81.1	96.4/**68.1**/87.4	92.5/53.5/79.0	**97.5**/67.5/**91.0**	96.0/63.7/89.0
Wood	95.2/50.6/91.0	92.8/43.1/89.6	95.9/56.0/87.6	94.3/44.7/89.6	**96.4/59.5/91.3**	94.9/50.7/88.0
Bottle	97.3/66.2/93.5	97.4/64.9/92.7	98.7/73.9/94.6	97.2/64.8/93.2	**99.0/78.6/95.6**	98.7/74.4/93.4
Cable	77.4/30.8/73.5	91.4/32.1/77.0	92.4/41.4/85.4	88.2/33.1/78.9	**98.7/67.5/94.4**	97.8/60.8/91.9
Hazelnut	98.5/59.1/**95.6**	97.9/56.0/94.6	98.7/61.3/94.7	98.5/59.3/95.3	**99.2/68.7**/95.3	98.9/64.0/94.8
Metal nut	91.3/57.1/89.1	92.2/57.5/87.2	95.5/74.7/92.1	93.9/62.7/90.4	96.6/79.2/93.3	**97.4/81.8/94.1**
Pill	96.6/58.7/95.6	94.5/45.5/95.4	**98.6/75.4**/95.6	97.4/64.0/95.4	97.2/65.1/**96.5**	98.1/69.3/95.5
Screw	99.0/41.0/95.8	98.6/33.9/93.7	98.6/38.7/92.6	99.0/44.4/95.7	**99.4/46.9/96.0**	**99.4**/45.7/**96.0**
Toothbrush	98.7/55.7/90.8	97.9/45.0/85.1	**99.1/64.9**/90.9	98.8/57.6/89.6	99.0/58.8/**93.5**	**99.1**/62.4/91.7
Transistor	83.4/39.5/71.0	92.1/45.4/82.8	92.0/51.4/74.0	90.3/41.5/73.4	**96.0/61.9**/84.2	95.7/61.2/**85.5**
Zipper	**98.2**/58.1/**94.7**	97.1/49.3/92.4	95.8/49.8/89.4	98.0/55.5/94.0	98.1/**59.6**/94.2	97.8/55.5/93.2
Mean	95.0/51.5/90.6	95.8/46.7/89.9	97.1/57.0/90.9	96.2/51.5/90.4	**98.2/60.6/93.8**	98.1/60.2/93.0

Table S6. Comparison (AUC_I/AUC_P/$MF1_P$) with Gains (Δ) between Specialized (Speci.) and Unified Training for our UniMMAD and Baseline.

Methods	Mean	MVTec-3D	Eyecandies	MulSen-AD	BraTs	UniMed
Baseline/Speci.	81.2/89.9/33.0	87.2/99.1/42.4	61.4/65.5/11.0	77.8/98.1/32.8	88.3/96.0/36.0	91.6/90.8/43.1
Baseline/Unified	75.6/86.6/28.4	82.7/98.8/39.2	58.1/53.8/10.6	79.9/98.0/34.0	69.2/91.8/19.7	88.0/90.5/38.5
Δ Speci.-Unified	↓5.6/↓3.3/↓4.6	↓4.5/↓0.3/↓3.2	↓3.3/↓11.7/↓0.4	↑2.1/↓0.1/↑1.2	↓19.1/↓4.2/↓16.3	↓3.6/↓0.3/↓4.6
Ours/Speci.	91.2/96.7/43.2	92.9/99.1/45.6	86.7/97.1/40.8	83.4/97.5/34.3	95.0/97.2/49.9	97.7/92.6/45.5
Ours/Unified	91.1/96.6/42.9	92.5/99.0/44.1	85.5/96.9/39.4	85.4/97.9/34.6	95.8/97.4/51.4	96.3/92.0/44.8
Δ Speci.-Unified	↓0.1/↓0.1/↓0.3	↓0.4/↓0.1/↓1.5	↓1.2/↓0.2/↓1.4	↑2.0/↑0.4/↑0.3	↑0.8/↑0.2/↑1.5	↓1.4/↓0.6/↓0.7

tion, Tab. S13 and S3 list results on UniMed and BraTS 2020 datasets, showing that the unified model remains competitive despite large domain shifts. Overall, the class-wise analysis confirms that our unified framework generalizes across modalities and defect types, consistently outperforming prior work.

J. More Qualitative Comparison

Fig. S4, Fig. S6, Fig. S7, and Fig. S5 show more qualitative comparison on MVTec-3D (industrial RGB-D), Eyecandies (synthetic RGB-D), BraTs2020 (Brain Tumor), UniMed (three medical datasets) and MulSen-AD (industrial infrared [19]). It demonstrates proposed method can accurately segment defect and lesion for a wide range of classes. In RGB-D datasets such as Cookie, our heatmaps highlight elongated defects along fiber directions. The model also emphasizes depth-based holes (e.g., Peach) and RGB-based color contamination (e.g., Candy Cane). In medical datasets such as UniMed, our method effectively separates boundary-blurred lesions from normal tissues, enabling high-confidence localization of conditions like colon polyps and brain tumors. Some generalist models like MVFA and AdaCLIP also show confident segmentation for colon polyps, possibly due to overlaps with their supervised training data. However, they incorrectly focus on normal regions in MulSen-AD, Liver Tumor, and Macular Edema. In contrast, UniMMAD consistently delivers accurate defect localization across diverse scenes, modalities, and object categories. These results highlight the effectiveness of the proposed "general → specific" paradigm and domain-specific reconstruction pathways in C-MoE, which mitigate cross-domain reconstruction interference. UniMMAD demonstrates precise defect localization and sharper boundaries.

Table S7. Per-Class comparison on the VisA [25] dataset for image-level metrics: AUC_I/AP_I / $MF1_I$.

Method→ Class↓	RD [7] CVPR'22	UniAD [23] NeurIPS'22	ViTAD [24] CVIU'25	MambaAD [11] NeurIPS'24	INP-Former [17] CVPR'25	UniMMAD Ours
Candle	89.0/89.3/82.2	93.8/94.4/86.2	87.3/88.2/82.1	90.3/91.2/84.2	**97.4/97.2/92.6**	95.3/94.8/90.3
Capsules	75.2/86.0/78.4	74.8/85.2/78.6	78.6/87.0/78.1	76.8/84.1/80.8	**90.3/94.8/88.5**	88.5/93.6/84.6
Cashew	87.8/93.6/88.1	91.1/95.6/88.6	82.2/91.1/86.0	88.5/94.2/88.1	95.5/97.6/**93.9**	**95.7/98.0**/93.2
Chewinggum	91.1/95.5/88.8	97.4/98.8/95.7	92.1/96.3/88.3	89.8/95.1/87.0	98.3/99.3/96.4	**98.8/99.5/97.9**
Fryum	94.7/97.6/90.8	92.3/96.3/91.2	93.3/97.0/89.6	94.8/97.8/92.0	**98.0/99.0/94.8**	96.2/98.3/92.8
Macaroni1	92.2/89.9/86.3	87.5/83.1/79.6	84.5/81.2/77.9	88.9/87.7/82.2	94.2/93.3/88.6	**97.2/97.2/91.5**
Macaroni2	80.7/76.6/75.7	76.2/76.0/69.2	78.3/71.8/73.9	69.6/60.9/71.0	82.5/79.8/78.5	**83.3/82.4/79.4**
Pcb1	95.1/94.9/91.0	96.5/96.0/92.7	95.3/93.7/90.4	96.1/95.9/90.6	**97.3/96.7/95.1**	97.1/96.5/95.0
Pcb2	**97.4/97.6**/93.1	92.4/92.8/86.2	90.8/89.4/85.7	95.6/96.1/90.9	96.7/96.3/91.5	97.2/97.1/**94.0**
Pcb3	60.0/58.9/70.4	88.4/88.6/81.2	90.8/91.4/83.5	94.5/94.6/88.8	95.1/95.7/87.8	**97.1/97.3/93.0**
Pcb4	**99.9/99.9/99.0**	99.4/99.4/96.4	98.9/98.7/96.1	99.7/99.7/97.5	99.7/99.8/97.5	99.8/99.6/98.0
Pipe fryum	97.3/98.6/95.5	97.9/99.1/96.0	95.9/98.0/92.8	97.7/98.7/96.5	97.3/98.8/97.5	**99.8/99.7/99.0**
Mean	88.4/89.9/86.6	90.6/92.1/86.8	89.0/90.3/85.4	90.2/91.3/87.5	95.2/95.7/91.9	**95.5/96.2/92.4**

Table S8. Per-Class comparison on the VisA [25] dataset for pixel-level metrics: AUC_P/$MF1_P$ / AUPRO.

Method→ Class↓	RD [7] CVPR'22	UniAD [23] NeurIPS'22	ViTAD [24] CVIU'25	MambaAD [11] NeurIPS'24	INP-Former [17] CVPR'25	UniMMAD Ours
Candle	98.8/34.4/94.3	99.2/33.4/94.9	95.4/26.1/83.7	99.0/30.0/**95.9**	**99.4/37.9**/95.2	99.3/37.0/94.7
Capsules	99.0/56.2/90.9	97.4/40.6/77.6	98.1/41.9/74.7	98.4/46.2/88.0	99.4/56.5/91.4	**99.6/65.8/91.9**
Cashew	87.3/42.5/68.3	97.5/47.7/90.6	97.4/58.6/69.1	91.2/45.2/70.3	97.6/60.5/89.3	**98.8/62.3/90.7**
Chewinggum	97.3/52.7/68.8	**99.3/61.9/87.3**	97.2/57.3/68.8	96.6/52.9/62.0	98.6/58.8/82.9	98.4/59.9/81.9
Fryum	96.8/50.3/**92.6**	96.9/50.4/86.5	**97.3**/51.2/88.9	96.9/51.5/92.2	97.1/48.6/89.9	97.0/**52.5**/88.4
Macaroni1	**99.7**/32.2/**96.4**	99.1/15.4/93.8	98.3/15.3/89.6	99.5/26.4/95.9	98.8/21.1/91.9	99.3/**32.8**/92.4
Macaroni2	**99.4/17.6/95.2**	97.7/10.4/89.6	98.1/10.1/87.3	98.5/ 8.6/91.1	99.0/11.7/93.5	98.9/16.9/92.1
Pcb1	99.2/54.0/95.1	99.2/57.0/91.2	99.3/56.8/88.7	99.5/65.8/**95.2**	**99.7**/68.5/94.9	**99.7/72.9**/94.8
Pcb2	97.7/29.4/90.9	98.1/19.4/86.1	98.0/21.1/83.4	98.5/24.8/**92.5**	**98.9/31.5**/89.9	98.8/26.5/**92.5**
Pcb3	89.7/ 3.5/67.7	98.3/28.2/86.7	98.0/28.5/88.0	98.5/29.5/**93.9**	**99.1/30.1**/91.7	**99.1**/29.0/91.1
Pcb4	97.3/34.6/87.3	97.5/39.6/86.1	**98.9**/44.7/**92.7**	96.1/35.8/81.7	98.7/**45.9**/92.1	98.4/43.1/90.9
Pipe fryum	98.9/56.7/**95.7**	98.7/52.3/94.1	**99.4**/66.0/94.3	99.0/55.2/95.2	99.2/60.9/94.9	99.3/**67.5**/94.2
Mean	96.8/38.7/86.9	98.2/38.0/88.7	97.9/39.8/84.1	97.6/39.3/87.8	98.8/44.3/**91.5**	**98.9/47.2**/91.3

Table S9. Per-Class comparison on the MVTec-3D [3] dataset for image-level metrics: AUC_I/AP_I / $MF1_I$.

Method→ Class↓	M3DM [22] CVPR'23	CFM [6] CVPR'24	AdaCLIP [5] ECCV'24	MVFA [13] CVPR'24	AACLIP [18] CVPR'25	UniMMAD Ours
Bagel	98.7/**99.9/98.1**	**99.2**/99.8/97.8	85.3/95.6/92.1	62.1/87.4/89.3	78.5/93.8/89.2	94.4/98.5/94.2
Cable gland	84.7/96.1/92.3	81.8/95.1/89.8	61.6/87.8/90.2	57.7/82.6/89.2	49.5/82.3/89.7	**94.8/98.5/96.0**
Carrot	94.6/99.0/95.6	97.8/99.5/97.3	80.4/95.8/91.7	71.3/92.5/91.8	64.5/89.5/91.9	**99.4/99.8/98.8**
Potato	**95.8/99.4/96.2**	88.4/96.8/93.3	63.3/86.7/90.2	48.4/80.1/89.3	87.5/96.7/92.1	93.3/97.9/**96.2**
Rope	91.6/96.8/89.1	93.8/97.5/90.9	82.5/90.9/85.5	91.6/96.6/91.5	81.3/91.8/84.1	**97.5/99.0/97.1**
Foam	**91.4/98.0/94.3**	**91.4**/97.8/92.0	69.5/91.2/88.9	53.6/86.9/88.9	49.7/81.8/88.9	89.1/97.3/91.6
Dowel	89.2/95.7/92.3	92.7/98.2/92.4	73.1/92.7/89.7	55.4/84.0/88.9	46.3/79.5/87.9	**95.9/98.8/96.2**
Tire	**87.3/96.0/92.1**	86.6/95.8/90.4	61.6/85.8/87.4	46.9/76.8/87.4	74.4/88.2/90.5	70.4/88.0/88.7
Peach	92.3/97.8/95.1	93.4/98.2/95.0	77.4/92.7/92.0	67.0/86.9/90.0	87.6/96.1/91.2	**95.8/98.8/96.3**
Cookie	95.6/98.9/96.1	98.9/99.7/**98.5**	86.4/95.5/91.8	73.7/89.9/88.4	**99.1/99.8/98.5**	94.2/98.3/93.4
Mean	92.1/97.7/94.1	92.4/**97.8**/93.7	74.1/91.4/89.9	62.7/86.3/89.4	71.8/89.9/90.4	**92.5**/97.5/**94.9**

Table S10. Per-Class comparison on the MVTec-3D [3] dataset for pixel-level metrics: AUC_P/$MF1_P$ / AUPRO.

Method→ Class↓	M3DM [22] CVPR'23	CFM [6] CVPR'24	AdaCLIP [5] ECCV'24	MVFA [13] CVPR'24	AACLIP [18] CVPR'25	UniMMAD Ours
Bagel	**99.4**/51.4/96.4	99.3/**52.8**/**96.9**	98.6/49.4/92.0	88.2/ 6.0/50.4	99.0/44.3/94.1	98.8/48.0/94.2
Cable gland	**99.3**/**47.5**/97.0	98.0/28.2/93.1	96.1/30.6/85.6	91.7/ 4.0/72.9	96.1/29.6/86.7	99.2/44.6/**97.1**
Carrot	99.6/41.4/97.7	**99.8**/**53.7**/**97.9**	98.7/27.6/95.2	97.3/13.6/91.3	99.6/44.1/97.6	99.6/43.6/**97.9**
Potato	99.4/40.4/96.8	99.5/34.0/97.4	99.4/31.1/96.8	97.7/10.6/90.5	**99.7**/**51.5**/**97.8**	99.6/42.1/**97.8**
Rope	99.6/49.4/**96.5**	**99.7**/**63.8**/96.3	98.6/46.1/93.4	97.3/29.1/87.7	99.5/53.3/95.4	99.4/39.5/96.4
Foam	98.5/43.5/93.7	**98.7**/**48.1**/**94.1**	90.7/42.7/70.3	91.9/28.0/73.0	95.1/ 9.9/84.6	97.4/38.9/90.3
Dowel	96.8/33.7/91.1	97.6/35.0/91.6	94.3/10.2/83.1	89.8/ 7.5/72.6	98.4/**37.1**/93.2	**98.9**/35.1/**96.0**
Tire	**99.3**/**41.2**/**96.3**	98.9/17.1/95.1	97.2/40.7/87.6	93.1/ 7.8/76.8	98.9/26.7/95.9	98.6/36.8/94.6
Peach	**99.5**/42.4/97.4	99.2/43.0/96.4	98.3/31.1/93.3	95.1/ 8.8/81.6	97.6/**59.6**/94.7	**99.5**/48.7/**97.5**
Cookie	97.8/56.2/92.4	96.8/**65.5**/93.9	97.2/61.1/93.3	90.3/24.3/72.7	96.8/64.8/94.0	**99.3**/63.8/**96.9**
Mean	98.9/**44.7**/95.5	98.7/44.1/95.3	96.9/37.1/89.1	93.2/14.0/76.9	98.0/42.1/93.4	**99.0**/44.1/**95.9**

Table S11. Per-Class comparison on the Eyecandies [4] dataset for image-level metrics: AUC_I/AP_I / $MF1_I$.

Method→ Class↓	M3DM [22] CVPR'23	CFM [6] CVPR'24	AdaCLIP [5] ECCV'24	MVFA [13] CVPR'24	AACLIP [18] CVPR'25	UniMMAD Ours
Gummybear	75.2/81.3/73.2	79.3/**84.7**/77.3	58.2/62.2/66.7	55.0/50.4/66.7	52.2/55.5/67.6	**90.7**/84.3/**89.7**
Lollipop	78.7/61.7/69.0	80.7/**76.6**/71.0	70.0/52.6/59.1	63.6/43.8/56.2	42.5/26.5/49.2	**85.1**/76.1/**78.7**
Marshmallow	91.7/92.2/84.7	97.8/98.3/95.8	82.3/85.3/80.0	82.1/83.8/79.2	60.2/67.1/66.7	**99.8**/**98.9**/**98.0**
Licoricesandwich	79.8/80.1/80.0	88.6/91.8/84.3	63.4/58.9/69.3	69.0/65.3/69.6	47.5/49.4/67.6	**90.0**/**92.6**/**86.9**
Chocolatepraline	75.8/80.1/72.3	**88.0**/**91.3**/82.6	86.6/80.0/**86.8**	72.6/72.7/76.9	54.2/55.0/69.6	82.7/87.2/76.1
Chocolatecookie	74.4/70.6/74.1	88.2/90.9/83.7	78.7/85.3/76.2	58.1/61.6/69.3	55.8/55.6/66.7	**97.9**/**98.1**/**96.0**
Peppermintcandy	**95.3**/**96.0**/**90.6**	85.8/89.3/79.2	85.9/85.7/81.5	76.3/78.7/76.7	27.7/37.0/66.7	92.6/93.1/84.0
Hazelnuttruffle	53.4/55.8/66.7	**71.7**/**80.1**/69.6	37.6/40.3/66.7	60.0/57.7/**70.0**	42.6/42.4/67.6	61.9/64.6/69.6
Confetto	89.1/90.0/83.0	88.2/91.8/83.7	87.8/89.1/83.0	66.6/69.6/71.2	61.3/65.5/67.7	**97.2**/**96.8**/**92.0**
Candycane	60.3/62.2/66.7	49.6/47.5/68.6	**80.6**/**77.0**/**79.3**	44.8/50.7/66.7	37.8/43.4/66.7	57.4/57.2/69.5
Mean	77.3/77.0/76.0	81.8/84.2/79.6	73.1/71.6/74.8	64.8/63.4/70.2	48.1/49.7/65.6	**85.5**/**84.9**/**84.1**

Table S12. Per-Class comparison on Eyecandies [4] Dataset for pixel-level metrics: AUC_P/$MF1_P$ / AUPRO.

Method→ Class↓	M3DM [22] CVPR'23	CFM [6] CVPR'24	AdaCLIP [5] ECCV'24	MVFA [13] CVPR'24	AACLIP [18] CVPR'25	UniMMAD Ours
Gummybear	94.8/37.4/78.5	93.1/19.3/78.9	**96.6**/25.2/87.6	88.6/ 4.7/53.7	88.3/ 6.6/64.8	95.9/**41.5**/**88.6**
Lollipop	96.5/28.6/77.5	**98.3**/22.2/**93.0**	97.7/11.4/87.8	95.1/ 6.1/78.4	97.3/13.6/88.9	98.0/**30.2**/90.3
Marshmallow	98.9/64.5/93.7	99.2/**66.5**/93.0	97.8/56.7/89.1	90.8/23.5/62.2	95.3/30.7/81.4	**99.5**/58.4/**96.6**
Licoricesandwich	96.8/**43.4**/85.4	96.0/28.1/81.7	95.6/31.5/80.3	82.8/11.2/47.6	95.6/36.4/81.4	**98.4**/39.4/**93.5**
Chocolatepraline	95.6/**51.7**/78.4	93.3/46.1/71.4	86.0/46.2/78.3	76.4/19.5/49.0	**98.1**/47.0/**91.9**	93.2/44.0/83.0
Chocolatecookie	89.4/23.5/71.4	97.5/35.5/86.9	95.8/**46.3**/86.8	50.0/12.8/19.5	94.3/45.1/84.0	**98.5**/43.5/**93.1**
Peppermintcandy	94.3/44.3/84.7	95.8/19.0/86.0	97.3/**49.4**/91.1	85.1/12.7/56.0	94.3/33.8/61.6	**99.4**/48.0/**97.2**
Hazelnuttruffle	89.2/19.0/61.0	**90.8**/**38.2**/61.6	87.6/ 7.1/53.5	71.6/ 4.3/42.1	87.8/ 5.6/**96.2**	90.6/23.4/65.2
Confetto	95.6/45.9/85.2	97.5/37.3/90.1	98.9/33.1/96.1	90.2/13.0/70.0	99.1/43.5/93.1	**99.6**/**56.1**/**98.0**
Candycane	85.7/ 6.4/70.8	97.2/ 6.6/90.6	**98.0**/**13.1**/**93.3**	95.7/ 2.7/87.3	97.5/ 9.3/70.8	95.4/ 9.4/86.9
Mean	93.7/36.5/78.7	95.8/31.9/83.3	95.1/32.0/84.4	82.6/11.1/56.6	94.7/27.2/81.4	**96.9**/**39.4**/**89.2**

Table S13. Per-Class comparison across UniMed datasets: Reinal OCT / Liver CT / Hyper-Kvasir / Mean performance.

Methods	**AUC$_I$(%)**	**AP$_I$(%)**	**MF1$_I$ (%)**	**AUC$_P$ (%)**	AUPRO	**MF1$_P$ (%)**
Specialized Models (One model for one dataset)						
UniAD [23]	90.4/83.7/98.4/90.8	95.9/90.8/99.9/95.5	91.0/83.6/98.1/90.9	95.1/97.2/68.6/87.0	81.5/90.4/32.9/68.2	63.3/30.1/24.9/39.4
RD [7]	93.3/90.7/96.3/93.4	97.3/94.6/97.0/96.3	91.5/90.3/92.9/91.6	96.2/96.3/79.0/90.5	86.1/91.1/46.5/74.5	66.0/23.3/30.7/40.0
ViTAD [24]	95.3/92.9/98.3/95.5	98.1/95.5/98.7/97.4	93.8/92.6/94.4/93.6	94.5/97.4/83.5/91.8	80.3/93.4/52.6/75.4	63.4/35.1/34.9/44.4
MambaAD [11]	96.2/94.6/97.1/96.0	98.6/97.4/98.4/98.1	94.1/91.9/94.7/93.6	96.6/96.5/82.0/91.7	86.7/91.5/55.8/78.0	68.4/25.0/33.8/42.4
SimpleNet [16]	90.4/80.2/97.6/89.4	96.4/91.3/97.4/95.0	88.7/80.5/96.3/88.5	92.0/92.5/76.1/86.9	72.3/64.3/36.7/57.8	55.8/34.4/27.2/39.1
INP-Former [17]	97.4/91.0/99.8/96.1	99.0/95.5/99.1/97.9	94.9/90.1/98.1/94.4	95.2/97.2/85.6/92.7	81.6/90.3/57.5/76.5	67.2/33.2/36.0/45.5
Generalist Models (One model performs all AD tasks without fine-tuning)						
AdaCLIP [5]	82.4/70.8/98.2/83.8	92.6/82.2/98.8/91.2	85.1/84.8/95.9/88.6	91.2/92.2/88.8/90.7	78.2/87.7/67.5/77.8	51.9/18.0/58.7/42.9
MVFA [13]	88.6/89.2/88.5/88.8	94.7/95.1/91.7/93.8	91.0/87.3/83.3/87.2	89.2/96.3/77.1/87.5	65.3/87.5/54.7/69.2	43.2/22.4/37.3/34.3
AA-CLIP [18]	71.3/60.4/85.1/72.3	82.7/70.1/91.7/81.5	83.7/83.4/85.4/84.2	92.8/93.4/86.6/90.9	79.3/86.9/63.8/76.7	53.3/16.6/51.1/40.3
Unified Multi-scene, Multi-modal and Multi-class Model						
Ours	96.4/95.9/96.8/96.3	98.6/97.9/96.8/97.7	93.0/94.5/95.0/94.2	96.5/97.1/82.5/92.0	85.4/92.3/48.0/75.2	68.0/31.1/35.6/44.9

Table S14. Per-Class comparison on the MulSen-AD [15] dataset for image-level metrics: AUC$_I$/AP$_I$ / MF1$_I$.

Method→ Class↓	MulSen-TripleAD [15] CVPR'25	AdaCLIP [5] ECCV'24	MVFA [13] CVPR'24	AA-CLIP [18] CVPR'25	UniMMAD Ours
Cotton	**100./100./100.**	47.3/81.1/87.6	54.3/83.5/87.6	47.3/81.1/87.6	99.7/99.6/98.7
Cube	**100./100./100.**	96.6/96.3/91.3	90.3/84.5/85.7	96.6/96.3/91.3	98.1/96.8/93.3
Zipper	99.1/99.6/97.2	69.8/88.5/82.3	**100./100./100.**	69.8/88.5/82.3	**100.**/99.4/**100.**
Toothbrush	91.3/96.3/89.8	34.0/37.6/66.7	44.4/49.4/65.4	**98.4/99.0/96.2**	81.4/80.5/85.2
Spring pad	67.5/86.2/85.7	74.4/70.1/76.9	91.0/91.0/84.8	85.8/92.9/**90.5**	**91.6/93.2**/86.6
Piggy	**99.7/99.9/98.4**	57.9/75.1/82.3	53.0/72.8/84.3	98.3/99.3/98.3	97.9/98.7/94.9
Capsule	**96.8/99.3/96.0**	52.4/83.8/89.1	74.9/91.5/88.5	87.0/97.2/91.4	94.2/98.1/95.8
Light	61.1/85.0/**89.2**	38.0/51.7/74.0	49.7/54.3/75.0	45.2/79.9/87.8	**80.1/86.8**/88.8
Plastic cylinder	99.6/**99.9**/98.2	88.3/94.6/88.9	71.7/85.1/80.8	**100.**/99.5/**100.**	99.3/99.1/97.9
Screen	20.0/60.5/**86.5**	48.6/57.6/73.0	**60.0**/55.6/78.6	15.3/**61.3**/86.4	45.3/60.7/79.8
Screw	92.9/97.3/96.8	56.2/49.3/53.8	65.9/51.0/62.2	**100./99.6/100.**	75.0/60.6/82.4
Flat pad	95.3/**98.6**/93.5	84.6/86.8/88.9	51.3/64.0/76.7	95.6/98.4/**95.0**	**97.6**/97.9/93.8
Nut	33.4/71.8/85.3	43.2/29.7/53.1	61.5/42.6/61.9	46.3/**78.2/86.9**	**75.0**/72.9/79.9
Button cell	80.0/93.4/87.5	37.2/52.7/75.8	64.3/73.2/79.4	**95.0/98.3/95.2**	75.2/87.1/83.9
Solar panel	46.2/81.2/88.6	57.7/86.7/88.6	54.1/85.5/88.6	**86.7/96.1/94.1**	71.0/91.8/91.9
Mean	78.8/**91.2/92.8**	59.1/69.4/78.1	65.7/72.2/79.9	77.8/91.0/92.2	**85.4**/88.2/90.2

Table S15. Per-Class comparison on the MulSen-AD [15] dataset for pixel-level metrics: AUC$_P$/MF1$_P$ / AUPRO.

Method→ Class↓	MulSen-TripleAD [15] CVPR'25	AdaCLIP [5] ECCV'24	MVFA [13] CVPR'24	AA-CLIP [18] CVPR'25	UniMMAD Ours
Cotton	92.6/**30.8**/77.1	94.1/30.7/85.1	82.8/21.3/56.0	94.1/30.7/85.1	**96.8**/25.6/**89.1**
Cube	**99.9/72.5/96.4**	99.8/65.2/92.2	90.1/39.8/52.6	99.8/65.2/92.2	99.7/66.1/92.3
Zipper	**99.1**/43.0/96.4	**99.1/46.6/96.6**	77.0/25.1/52.0	**99.1/46.6/96.6**	98.8/36.3/95.8
Toothbrush	99.2/**39.3**/94.8	**99.3**/20.8/**95.9**	63.9/ 4.6/36.3	98.2/14.1/91.5	98.2/27.0/89.2
Spring pad	**99.7**/34.6/**98.0**	99.0/25.9/96.6	98.3/ 3.3/94.2	93.1/34.1/84.3	94.8/**35.2**/88.0
Piggy	97.0/29.4/90.4	95.8/22.1/85.3	92.1/10.0/69.7	98.1/32.5/87.4	**98.5/33.8/91.0**
Capsule	**99.3/30.0/97.0**	99.1/25.6/**97.0**	97.3/26.2/90.4	98.8/28.8/95.5	98.9/26.6/95.4
Light	96.7/16.9/91.7	97.3/20.0/93.9	95.8/**25.2**/76.8	97.2/19.5/92.7	**98.5**/25.0/**94.7**
Plastic cylinder	**99.5**/47.4/**97.2**	99.2/46.4/94.7	89.1/ 3.2/60.9	99.4/**50.3**/97.0	99.3/40.9/97.0
Screen	91.4/ 7.6/78.0	**93.3**/11.2/**82.3**	84.1/10.6/36.5	92.3/**14.1**/79.1	90.4/11.7/79.0
Screw	99.0/11.9/**96.5**	98.6/11.2/95.6	97.3/16.8/88.8	98.9/12.7/**96.5**	**99.2/44.9**/96.1
Flat pad	99.6/**45.1**/97.7	99.2/35.9/97.0	97.8/ 6.4/93.0	99.4/43.4/97.4	**99.7**/37.8/**98.2**
Nut	**99.6**/36.8/**97.1**	98.9/16.7/96.0	98.0/ 2.2/94.0	98.5/34.1/92.4	98.8/**45.3**/89.6
Button cell	**99.7**/38.9/98.0	**99.7**/40.8/98.2	99.6/28.1/98.2	99.5/41.6/97.9	**99.7/42.7/98.3**
Solar panel	95.1/19.7/85.4	95.7/**25.6**/75.1	91.0/13.8/66.8	96.5/19.8/87.9	**96.9**/20.5/**88.3**
Mean	97.8/33.6/**92.8**	97.8/29.7/92.1	90.3/15.8/71.1	97.5/32.5/91.6	**97.9/34.6**/92.1

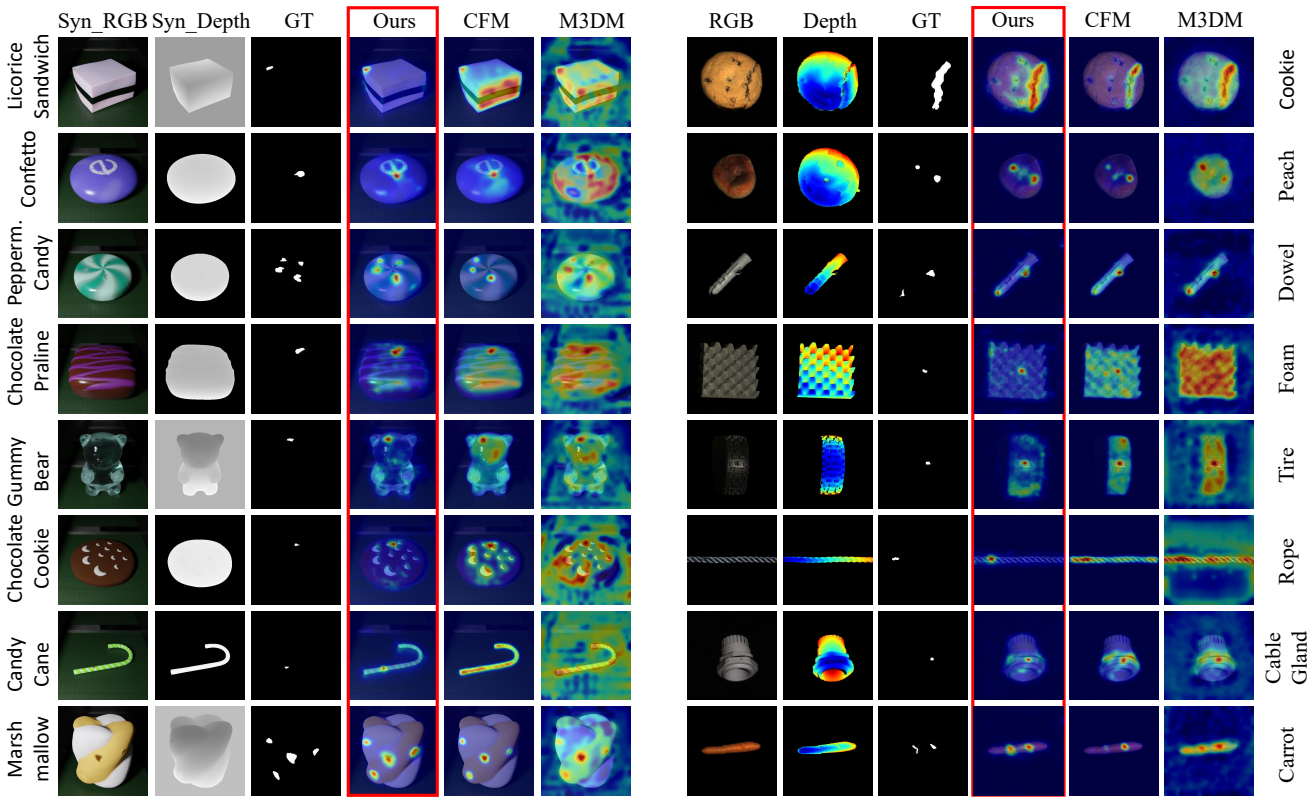

Figure S4. Qualitative comparison on MVTec-3D and Eyecandies, with our method highlighted in the red box.

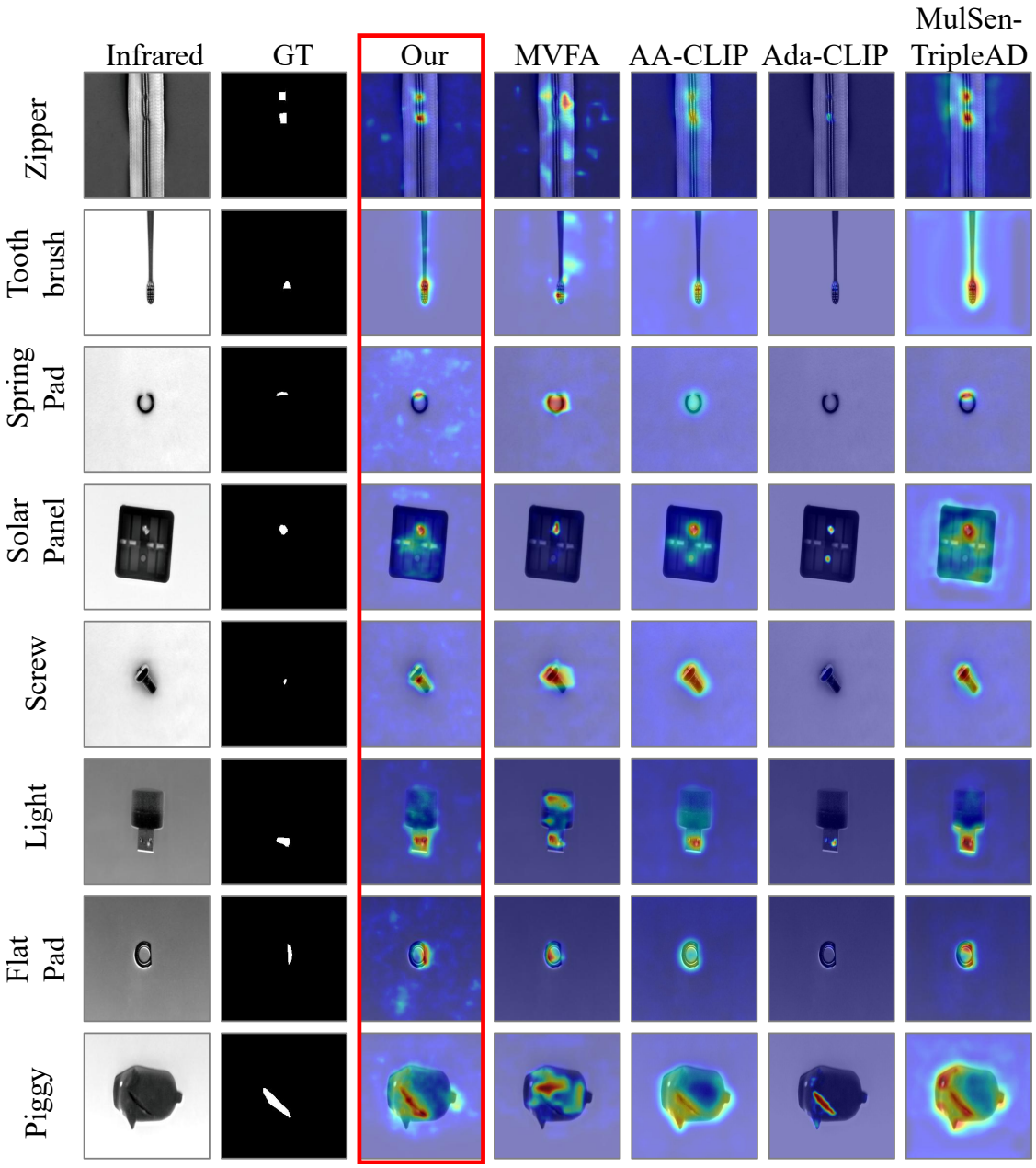

Figure S5. Qualitative comparison on MulSen-AD, with our method highlighted in the red box.

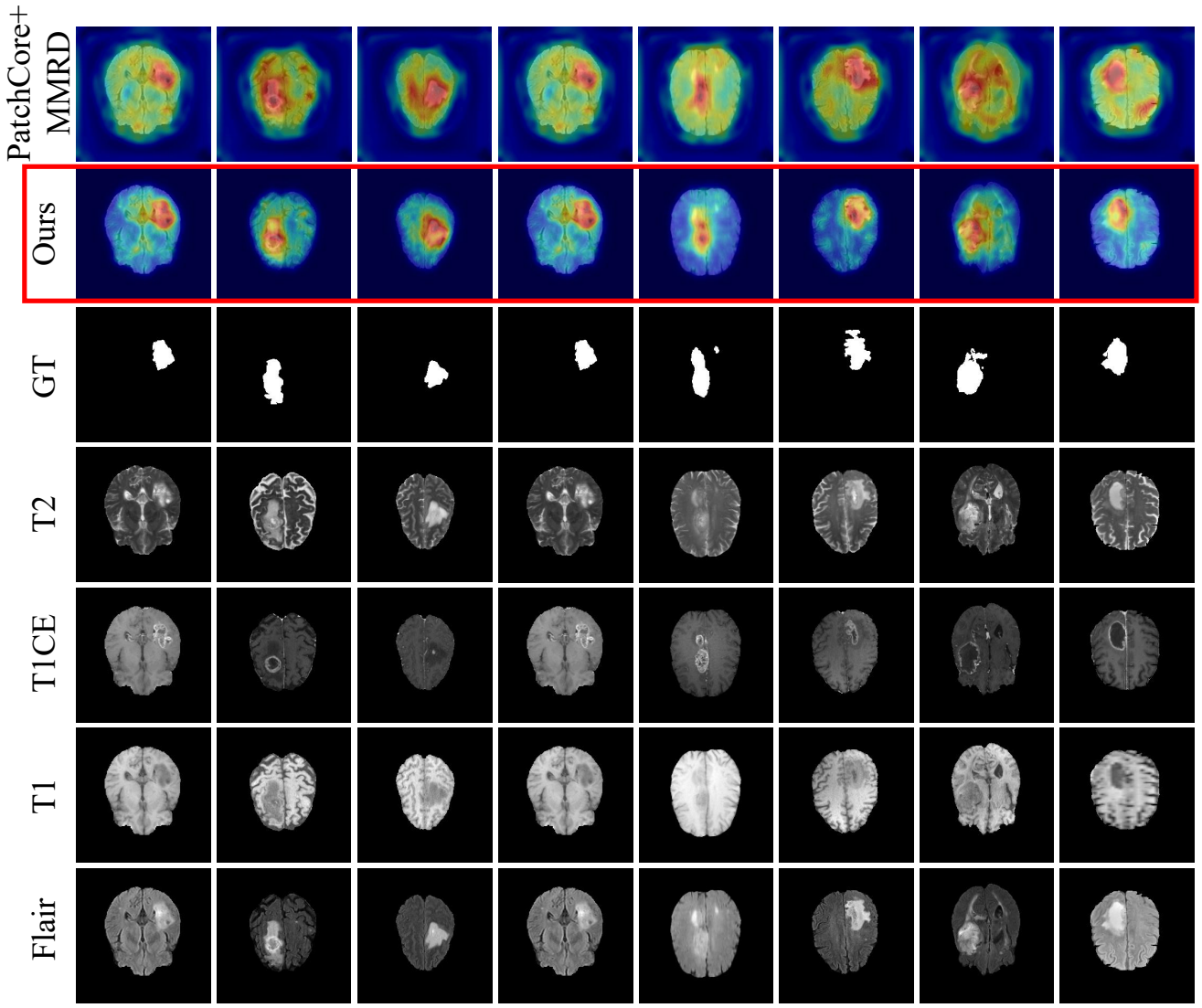

Figure S6. Qualitative comparison on BraTs2020, with our method highlighted in the red box.

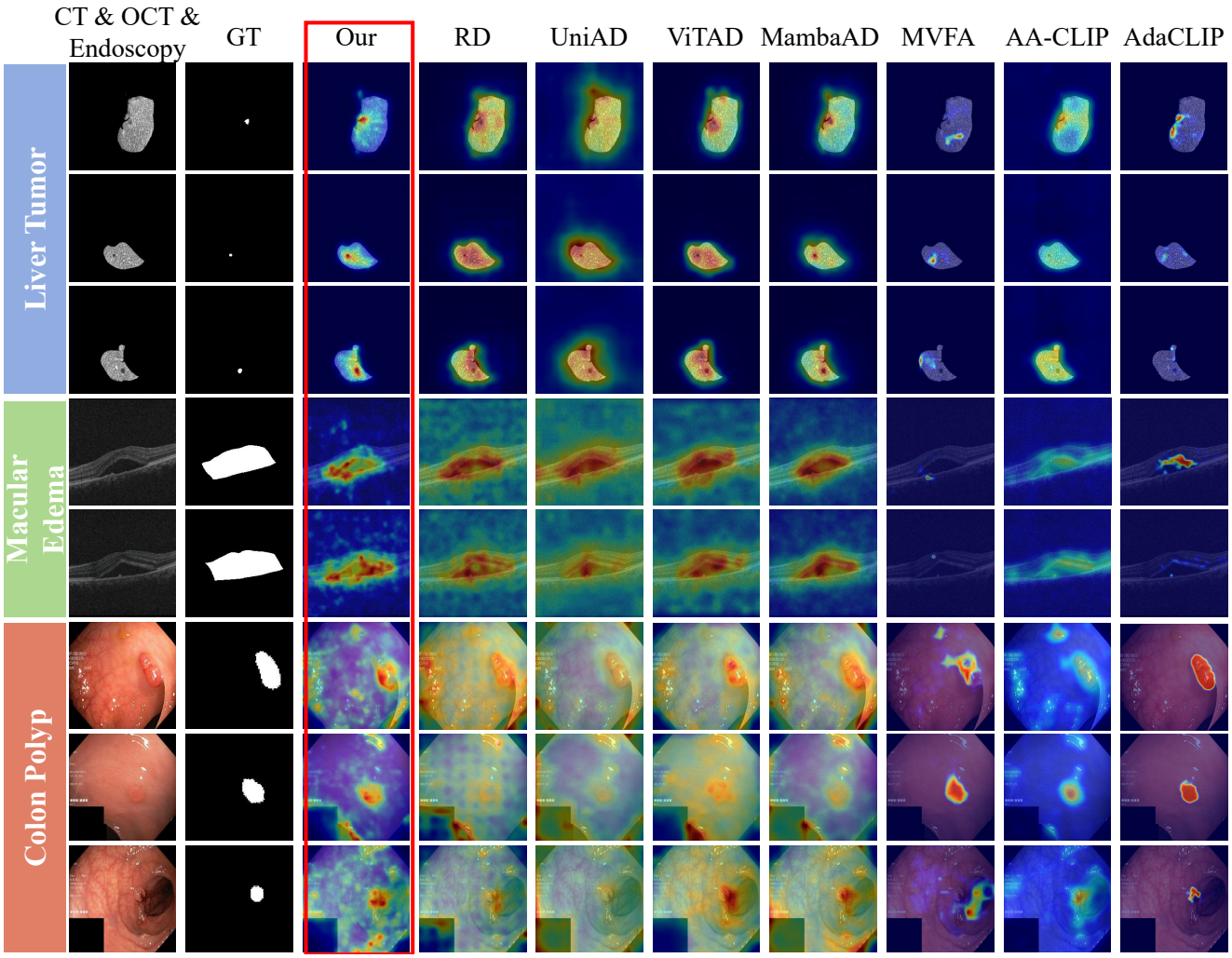

Figure S7. Qualitative comparison on UniMed, with our method highlighted in the red box.

References

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

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9592–9600, 2019.

[3] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. arXiv preprint arXiv:2112.09045, 2021.

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

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

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

[7] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9737–9746, 2022.

[8] Lei Fan, Junjie Huang, Donglin Di, Anyang Su, Tianyou Song, Maurice Pagnucco, and Yang Song. Salvaging the overlooked: Leveraging class-aware contrastive learning for multi-class anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 21419–21428, 2025.

[9] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(120):1–39, 2022.

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

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

[12] Junjie Hu, Yuanyuan Chen, and Zhang Yi. Automated segmentation of macular edema in oct using deep neural networks. Medical image analysis, 55:216–227, 2019.

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

[14] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668, 2020.

[15] Wenqiao Li, Bozhong Zheng, Xiaohao Xu, Jinye Gan, Fading Lu, Xiang Li, Na Ni, Zheng Tian, Xiaonan Huang, Shenghua Gao, et al. Multi-sensor object anomaly detection: Unifying appearance, geometry, and internal properties. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 9984–9993, 2025.

[16] 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, pages 20402–20411, 2023.

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

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

[19] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, Huchuan Lu, Georges El Fakhri, Xiaofeng Liu, and Shijian Lu. Rethinking evaluation of infrared small target detection. In NeurIPS, 2025.

[20] 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, pages 14318–14328, 2022.

[21] Yu Tian, Guansong Pang, Fengbei Liu, Yuanhong Chen, Seon Ho Shin, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro. Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, pages 128–140. Springer, 2021.

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

[23] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. Advances in Neural Information Processing Systems, 35:4571–4584, 2022.

[24] Jiangning Zhang, Xuhai Chen, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, Ming-Hsuan Yang, and Dacheng Tao. Exploring plain vit reconstruction for multi-class unsupervised anomaly detection. arXiv preprint arXiv:2312.07495, 2023.

[25] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In European Conference on Computer Vision, pages 392–408. Springer, 2022.