HIERARCHICAL PROTOTYPE-CONDITIONED MASKED DIFFUSION FOR MULTI-VIEW ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view anomaly detection leverages complementary observations to identify anomalies that may be occluded or weakly visible from a single viewpoint. However, existing methods often rely on dense patch-level cross-view interactions, which may introduce redundant and spatially inconsistent information. To address this limitation, we propose HPMD-AD, a **H**ierarchical **P**rototype-Conditioned **M**asked **D**iffusion model for multi-view **A**nomaly **D**etection, which models complementary cross-view normal patterns through compact hierarchical prototypes to guide normal feature reconstruction. Specifically, we progressively abstract features from each view into input-dependent view-level prototypes and aggregate them into sample-level prototypes, capturing view-specific normal patterns and cross-view shared structures without explicit spatial alignment. These hierarchical prototypes jointly guide diffusion-based feature reconstruction, while prototype masking and feature masking encourage complementary cross-view reasoning and mitigate identity shortcuts during reconstruction. Furthermore, we introduce a prototype-prior adaptive fusion strategy that jointly considers inter-view score disagreement and prototype reliability for robust multi-view score aggregation. Extensive experiments on Real-IAD and the substantially larger Real-IAD Variety benchmark demonstrate that HPMD-AD achieves state-of-the-art performance, attaining sample-level AUROC scores of 94.6% and 96.5%, respectively.

1 INTRODUCTION

Anomaly detection aims to identify abnormal samples and localize anomalous regions Cao et al. (2024), playing a pivotal role in industrial inspection Bergmann et al. (2019), medical image analysis Guo et al. (2023a), and video surveillance Ramachandra et al. (2020). Despite remarkable progress, most existing methods are developed under single-view settings You et al. (2022); Yin et al. (2023); He et al. (2024b); Mousakhan et al. (2024); Guo et al. (2025); Luo et al. (2025), where the anomaly state of an object is inferred from only one observation. In practical scenarios, however, defects may appear on surfaces invisible from a particular viewpoint Wang et al. (2024); Zhu et al. (2026), be occluded, or exhibit only weak visual cues, making a single view insufficient to capture the complete anomaly evidence. Multi-view anomaly detection addresses this limitation by jointly analyzing multiple observations of the same object to exploit complementary evidence across viewpoints He et al. (2026). Therefore, effective multi-view modeling requires characterizing both view-specific normal patterns and cross-view shared structures while reliably integrating complementary information across observations.

Recent multi-view anomaly detection methods increasingly focus on joint cross-view representation learning to exploit complementary information among different observations Mao et al. (2025); Liu et al. (2025); Zhou et al. (2025); He et al. (2026). According to the dominant mechanism used to establish cross-view relations, existing approaches can be broadly divided into two paradigms. One line of methods relies on data-driven feature interaction, directly propagating or fusing information across views through local feature exchange or attention mechanisms Mao et al. (2025); Zhou et al. (2025); Kruse & Rosenhahn (2025); He et al. (2026). Another line incorporates explicit geometric priors, such as camera calibration and epipolar geometry, to guide cross-view matching and suppress irrelevant interactions Liu et al. (2025). Despite encouraging progress, both paradigms still exhibit notable limitations. Data-driven methods often operate on dense local features, where cross-view interactions may aggregate redundant or spatially mismatched information, particularly under

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.

substantial viewpoint-induced appearance variations and limited spatial overlap Bai et al. (2026). Geometry-guided methods can alleviate such mismatches by imposing explicit correspondence constraints, but typically depend on additional calibration information or geometric priors, which may reduce their flexibility and practicality in real-world industrial deployment Liu et al. (2025). Consequently, a key challenge is to establish reliable cross-view interactions without excessive redundancy or reliance on additional geometric priors.

These limitations motivate a fundamental question in multi-view anomaly detection: *what information should actually be exchanged across views, and how should it contribute to normality modeling?* Rather than relying on dense local correspondence, we argue that effective cross-view modeling should emphasize structured information that is both complementary across views and representative of the underlying normal state. We therefore characterize cross-view relationships through compact hierarchical representations that preserve view-specific normal patterns while progressively abstracting shared structures across observations. Learned from normal training data, this hierarchy provides structured conditioning that enables feature reconstruction in each view to exploit both view-specific patterns and complementary cross-view context, thereby supporting subsequent anomaly assessment.

Guided by this perspective, we propose HPMD-AD, a **H**ierarchical **P**rototype-Conditioned **M**asked **D**iffusion model for multi-view **A**nomaly **D**etection. We first introduce Hierarchical Dynamic Prototype Modeling (HDPM), which abstracts dense features from each view into input-dependent view-level prototypes and aggregates them into compact sample-level prototypes, capturing view-specific normal patterns and cross-view shared structures without explicit spatial alignment. Building on these hierarchical representations, we develop Dual-Masked Prototype-Conditioned Diffusion (DMPCD), in which view- and sample-level prototypes jointly guide diffusion-based feature reconstruction. Prototype masking encourages complementary reasoning across views, while feature masking mitigates identity shortcuts and promotes contextual reconstruction. For sample-level anomaly scoring, we further introduce a Prototype-Prior Adaptive Fusion Strategy (PPAFS), which uses inter-view score disagreement to adaptively balance consensus averaging and salient-view aggregation, while leveraging prototype consistency as a reliability cue to calibrate the contribution of each view. Extensive experiments on Real-IAD and the substantially larger Real-IAD Variety benchmark demonstrate that HPMD-AD consistently achieves state-of-the-art performance under the multi-class multi-view setting, attaining sample-level AUROC scores of 94.6% and 96.5%, respectively. The main contributions of our work are summarized as follows:

- **Hierarchical Dynamic Prototype Modeling.** We propose HDPM to progressively abstract dense view features into input-dependent view-level prototypes and learn sample-level prototypes through cross-view interaction, capturing view-specific normal patterns and cross-view shared structures without explicit spatial alignment.
- **Dual-Masked Prototype-Conditioned Diffusion.** DMPCD integrates hierarchical prototype conditioning with prototype masking and feature masking, promoting complementary cross-view reasoning and mitigating identity shortcuts during feature reconstruction.
- **Prototype-Prior Adaptive Fusion Strategy.** We propose PPAFS for sample-level anomaly scoring, which adaptively balances consensus averaging and salient-view aggregation based on inter-view score disagreement, while leveraging prototype consistency as a reliability cue to calibrate the contribution of each view.

2 RELATED WORK

2.1 ANOMALY DETECTION

Unsupervised anomaly detection learns normal patterns from anomaly-free data to identify deviations at inference. Representative methods include DRAEM Zavrtanik et al. (2021) and SimpleNet Liu et al. (2023) for anomaly synthesis, DiAD He et al. (2024b) for diffusion-based reconstruction, PatchCore Roth et al. (2022) for feature embedding, and RD4AD Deng & Li (2022) for knowledge distillation. Recent methods, including UniAD You et al. (2022), ViTAD Zhang et al. (2023a), MambaAD He et al. (2024a), and Dinomaly Guo et al. (2025), further enable unified multi-class modeling, but primarily operate under single-view settings. Multi-view anomaly detection exploits complementary observations to alleviate limited defect visibility from individual

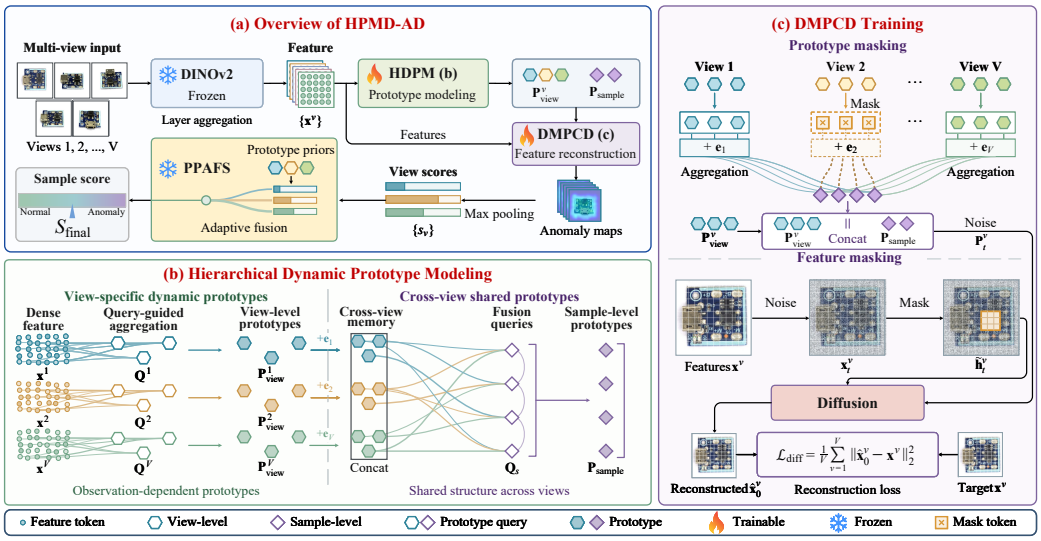

Figure 1: Overview of the HPMD-AD framework. (a) The overall framework, where HDPM learns hierarchical prototypes, DMPCD performs prototype-conditioned diffusion reconstruction, and PPAFS aggregates view-wise anomaly scores. (b) HDPM constructs view- and sample-level prototypes to model view-specific patterns and cross-view shared structures without explicit spatial alignment. (c) Hierarchical prototypes guide diffusion-based feature reconstruction. Feature masking and diffusion operate on DINOv2 feature tokens, while RGB images are shown only for schematic visualization.

viewpoints. MVAD He et al. (2026) performs semantically guided cross-view feature selection and fusion, while IDIF Mao et al. (2025) decouples view-common and view-specific features for implicit voxel-based fusion. MVEAD Liu et al. (2025) incorporates epipolar constraints into cross-view attention to improve correspondence reliability. In contrast, HPMD-AD organizes complementary cross-view information into compact hierarchical prototypes that capture view-specific and shared normal structures to condition feature reconstruction without explicit spatial alignment.

2.2 Prototype-Based Representation Learning

Prototype-based representation learning Snell et al. (2017) provides a compact way to summarize representative normal patterns and has been increasingly explored in anomaly detection. PatchCore Roth et al. (2022) retains representative normal patch features for distance-based anomaly scoring, while FastRef Li et al. (2026) further improves prototype-based detection through test-time refinement. HVQ-Transformer Lu et al. (2023) introduces vector-quantized prototypes into reconstruction-based anomaly detection. These prototype-based detectors primarily model normal representations within individual images or across categories. HPMD-AD constructs an input-dependent hierarchy of view-level and sample-level prototypes, summarizing view-specific patterns and cross-view shared structures. The two prototype levels jointly condition diffusion-based feature reconstruction, with prototype and spatial feature masking encouraging complementary cross-view reasoning and limiting identity shortcuts.

3 Method

We propose HPMD-AD, a Hierarchical Prototype-Conditioned Masked Diffusion model for multi-view anomaly detection. As illustrated in Fig. 1, HDPM progressively abstracts multi-layer features extracted by a frozen DINOv2 Oquab et al. (2023); Darcet et al. (2024) backbone into view-level prototypes and further aggregates them into sample-level prototypes. DMPCD employs a prototype-conditioned Diffusion Transformer (DiT) Peebles & Xie (2023) for normal feature reconstruction, with prototype and feature masking during training. Finally, PPAFS combines score disagreement and prototype-derived reliability to aggregate view-wise anomaly scores into a sample-level score.

3.1 HIERARCHICAL DYNAMIC PROTOTYPE MODELING

Multi-view anomaly detection requires modeling both view-specific appearance patterns and cross-view shared normal structures. However, dense interaction among different views may introduce redundant information and unreliable correspondences, particularly under substantial viewpoint variations. To obtain compact normal representations without explicit spatial alignment, we introduce HDPM, which progressively abstracts multi-view features into view-level and sample-level prototypes. The former characterize normal patterns within individual views, while the latter aggregate complementary information across views to capture shared normal structures.

Multi-view Feature Representation. Given a multi-view sample $\mathcal{X} = \{X^v\}_{v=1}^V$, we employ a frozen DINOv2 backbone to extract multi-layer visual features from each view. The selected layer features are aggregated into latent tokens $\mathbf{x}^v \in \mathbb{R}^{N\times D}$ for subsequent prototype learning.

View-level Prototype Learning. Although dense feature tokens preserve rich local information, they also contain redundant appearance variations and lack compact structural abstraction. We therefore dynamically generate view-level prototypes to summarize representative normal patterns within each observation perspective. Specifically, a set of prototype queries $\mathbf{Q}^v = \{\mathbf{q}_m^v\}_{m=1}^M$ interacts with view-specific tokens through cross-attention:

$$\mathbf{P}_{\text{view}}^v = \text{FFN}\left(\text{CrossAttn}\left(\mathbf{Q}^v, \mathbf{x}^v, \mathbf{x}^v\right)\right), \tag{1}$$

where $\mathbf{P}_{\text{view}}^v = \{\mathbf{p}_m^v\}_{m=1}^M$ denotes the dynamically generated view-level prototypes.

Unlike fixed prototypes shared across samples, the generated prototypes are conditioned on the current observation, allowing them to adapt to view-dependent normal structures. In this way, dense feature tokens are compressed into compact structural representations.

Sample-level Prototype Learning. Although view-level prototypes capture view-specific normal patterns, they primarily characterize individual views and cannot explicitly represent normal structures shared across viewpoints. We therefore aggregate them into a compact set of sample-level prototypes to model cross-view shared information.

Given the view-level prototypes $\mathbf{P}_{\text{view}}^v \in \mathbb{R}^{M\times D}$, we first inject view identity using a learnable embedding and concatenate the resulting prototypes into a cross-view prototype memory:

$$\widetilde{\mathbf{P}}_{\text{view}}^v = \mathbf{P}_{\text{view}}^v + \mathbf{e}_v, \qquad \mathbf{M} = \text{Concat}\left(\widetilde{\mathbf{P}}_{\text{view}}^1, \ldots, \widetilde{\mathbf{P}}_{\text{view}}^V\right), \tag{2}$$

where $\mathbf{e}_v$ denotes the embedding associated with the v-th view.

To distill shared normal structures from $\mathbf{M}$, we introduce K learnable fusion queries $\mathbf{Q}_s$ and process them through a single query interaction block:

$$\widetilde{\mathbf{Q}} = \text{SelfAttn}(\mathbf{Q}_s), \quad \widehat{\mathbf{Q}} = \text{CrossAttn}\left(\widetilde{\mathbf{Q}}, \mathbf{M}, \mathbf{M}\right), \quad \mathbf{Q} = \text{FFN}\left(\widehat{\mathbf{Q}}\right). \tag{3}$$

Self-attention captures dependencies among fusion queries, while cross-attention enables each query to selectively retrieve complementary normal information from different views.

The output queries $\mathbf{Q}$ are projected to obtain the sample-level prototypes $\mathbf{P}_{\text{sample}} \in \mathbb{R}^{K\times D}$. Together with the view-level prototypes, they form a hierarchical representation that captures view-specific patterns and cross-view shared structures without dense cross-view feature interaction or explicit spatial correspondence.

3.2 DUAL-MASKED PROTOTYPE-CONDITIONED DIFFUSION RECONSTRUCTION

HDPM learns hierarchical prototypes that capture view-specific normal patterns and cross-view shared structures. However, prototype aggregation may over-rely on dominant views, while direct feature reconstruction may exploit local appearance cues instead of hierarchical priors. To address these issues, we propose DMPCD, a dual-masked prototype-conditioned diffusion framework. Prototype masking encourages complementary cross-view aggregation, while feature masking promotes prototype-guided normal feature reconstruction.

Prototype Masking. To improve cross-view prototype aggregation, we randomly select one view and mask all its view-level prototypes before constructing the sample-level prototypes. For view v,

let $M_p^v \in \{0,1\}$ indicate whether it is masked, with $\sum_{v=1}^{V} M_p^v = 1$. The masked prototypes are

$$\overline{\mathbf{P}}_{\text{view}}^{v} = (1 - M_p^v)\mathbf{P}_{\text{view}}^{v} + M_p^v \mathbf{1}_M \mathbf{m}_p^\top, \tag{4}$$

where $\mathbf{m}_p \in \mathbb{R}^D$ is a learnable mask token and $\mathbf{1}_M$ replicates it across prototype positions. The masked prototypes are used for sample-level aggregation, while the original view-level prototypes are retained as view-specific reconstruction conditions. This encourages the fusion module to exploit complementary information from the remaining views rather than relying on a particular viewpoint.

Feature Masking. Given clean feature representations $\mathbf{x}^v$, we first generate noisy features at diffusion timestep t:

$$\mathbf{x}_t^v = \sqrt{\bar{\alpha}_t}\mathbf{x}^v + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_x, \qquad \boldsymbol{\epsilon}_x \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{5}$$

The noisy features are projected into the DiT embedding space with positional embeddings:

$$\mathbf{h}_t^v = \mathrm{Proj}(\mathbf{x}_t^v) + \mathbf{E}_{\text{pos}}.$$

We randomly replace a subset of tokens with a learnable mask token while preserving positional information:

$$\widetilde{\mathbf{h}}_t^v = (1 - \mathbf{M}_f^v) \odot \mathbf{h}_t^v + \mathbf{M}_f^v \odot (\mathbf{m}_f + \mathbf{E}_{\text{pos}}), \tag{6}$$

where $\mathbf{M}_f^v$ denotes the feature mask and $\mathbf{m}_f$ is a learnable mask token. Since masking is applied after diffusion noising, the mask token itself is not perturbed. The clean features remain the reconstruction targets, discouraging local copying and promoting prototype-guided recovery.

Noise-Aware Prototype-Conditioned Diffusion Reconstruction. During training, $\mathbf{P}_{\text{sample}}$ is constructed from the masked view-level prototypes, while the original $\mathbf{P}_{\text{view}}^v$ is retained as the view-specific condition. To align the hierarchical prototype conditions with the corruption level of the noisy feature representation, they are concatenated and perturbed according to the same noise schedule:

$$\mathbf{P}_t^v = \sqrt{\bar{\alpha}_t}\mathrm{Concat}\left(\mathbf{P}_{\text{view}}^v, \mathbf{P}_{\text{sample}}\right) + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_p, \qquad \boldsymbol{\epsilon}_p \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{7}$$

The masked feature tokens, noisy prototype conditions, and timestep embedding are fed into the DiT denoiser to predict clean features:

$$\hat{\mathbf{x}}_0^v = \mathrm{DiT}_\theta\left(\widetilde{\mathbf{h}}_t^v, t, \mathbf{P}_t^v\right). \tag{8}$$

Within DiT, hierarchical prototypes interact with feature tokens through cross-attention and serve as structured normal priors for reconstruction.

Training Objective and Inference. We adopt the x_0-prediction parameterization and optimize

$$\mathcal{L}_{\text{diff}} = \frac{1}{V}\sum_{v=1}^{V} \|\hat{\mathbf{x}}_0^v - \mathbf{x}^v\|_2^2. \tag{9}$$

The loss is computed over all feature tokens, with both masking strategies used only during training. During inference, noisy test features are reconstructed using DDIM Song et al. (2020) conditioned on hierarchical prototypes. The reconstruction errors yield view-wise anomaly maps, whose maximum responses define the scores s_v used by PPAFS.

3.3 Prototype-Prior Adaptive Fusion Strategy

Defect visibility varies across viewpoints, and strong reconstruction responses may reflect either anomaly evidence or view-specific structural mismatch. We propose PPAFS to balance cross-view consensus and discriminative anomaly responses by considering score disagreement and prototype-based structural consistency.

Given nonnegative view-wise anomaly scores $\{s_v\}_{v=1}^{V}$, let μ_s and σ_s denote their mean and standard deviation, respectively. We define the disagreement measure and fusion coefficient as

$$D_s = \frac{\sigma_s}{\mu_s + \sigma_s + \epsilon}, \qquad \alpha = 1 - (1 - D_s)^{V-1}, \tag{10}$$

where ϵ is a small constant for numerical stability. For $V > 1$, larger disagreement increases α, allowing the fusion to depart further from mean aggregation. When all view scores are equal, $\alpha = 0$.

Table 1: Quantitative comparison on Real-IAD and Real-IAD Variety under the multi-class setting.

Dataset	Methods	Sample-level			Image-level			Pixel-level				mAD
		AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO	
Real-IAD (30 categories)	SimpleNet	61.2	77.9	82.1	82.6	79.2	74.1	76.3	3.4	6.9	42.5	58.62
	CFA	65.8	80.2	82.2	58.3	53.9	61.7	82.4	2.2	5.9	51.2	54.38
	CFLOW-AD	80.2	89.8	84.9	75.1	74.1	69.1	95.0	18.1	21.7	81.7	68.97
	PyramidFlow	63.1	79.3	82.0	56.2	52.6	62.5	71.5	1.5	1.7	35.5	50.59
	RD4AD	85.7	92.4	87.7	83.0	79.6	74.4	97.3	25.8	33.5	90.4	74.98
	ViTAD	88.8	94.3	89.3	82.8	80.0	73.7	97.1	23.9	31.8	84.3	74.60
	UniAD	87.3	93.1	88.9	83.0	83.4	74.6	97.4	23.7	31.4	87.3	75.01
	RD++	87.0	93.3	88.2	83.7	80.9	74.9	97.7	25.8	33.5	90.0	75.50
	DeSTSeg	89.0	94.6	89.2	82.3	79.2	73.2	94.6	37.9	41.7	40.6	72.23
	MVAD	90.2	95.3	90.1	86.6	84.8	77.2	97.9	30.3	36.8	91.2	78.04
	HPMD-AD	**94.6**	**97.4**	**93.3**	**89.9**	**87.1**	**81.1**	**99.1**	**38.5**	**42.4**	**92.1**	**81.53**
Real-IAD Variety (160 categories)	SimpleNet	70.3	94.6	94.5	62.1	91.4	93.1	78.1	9.7	14.3	42.1	65.02
	RD4AD	76.7	95.6	95.1	72.0	93.6	93.5	88.8	22.3	28.1	68.8	73.45
	ViTAD	83.6	97.2	95.6	78.3	95.4	93.7	90.9	31.9	36.3	70.3	77.32
	UniAD	67.9	93.7	94.5	67.1	92.3	93.3	87.0	17.9	23.9	62.1	69.97
	MVAD	71.3	94.2	94.7	69.2	92.8	93.3	86.5	18.1	25.3	60.9	70.60
	MambaAD	85.0	97.4	95.6	79.0	95.6	93.7	90.2	30.9	35.5	70.6	77.35
	EILA	91.3	98.6	96.8	87.1	97.5	94.7	93.8	44.5	46.8	78.6	82.97
	Dinomaly	92.5	98.9	97.1	87.0	97.5	94.8	93.5	44.0	47.0	79.5	83.18
	HPMD-AD	**96.5**	**99.4**	**98.2**	**92.1**	**98.4**	**96.1**	**94.3**	**47.7**	**49.3**	**79.6**	**85.17**

To estimate structural consistency, we construct a sample-specific prior from the view-level prototypes learned by HDPM. We first normalize prototype tokens to remove magnitude variations, aggregate them into view centers $\mathbf{c}_v$, and further obtain a cross-view consensus center $\mathbf{c}_s$:

$$\mathbf{c}_v = \mathcal{N}_\delta\left(\frac{1}{M}\sum_{m=1}^{M}\mathcal{N}_\delta(\mathbf{p}_m^v)\right), \qquad \mathbf{c}_s = \mathcal{N}_\delta\left(\frac{1}{V}\sum_{v=1}^{V}\mathbf{c}_v\right), \tag{11}$$

where $\mathcal{N}_\delta(\mathbf{x}) = \mathbf{x}/\max(\|\mathbf{x}\|_2, \delta)$ denotes stabilized ℓ_2 normalization with a small constant δ for numerical stability. The prototype inconsistency of view v is measured by $d_v = 1 - \mathbf{c}_v^\top \mathbf{c}_s$. Let μ_d and σ_d denote the mean and population standard deviation of these inconsistencies. We define a sample-adaptive temperature $\tau_s = \sigma_d + \epsilon$ to scale the structural weighting.

We obtain anomaly-saliency weights q_v from standardized reconstruction scores and structural weights r_v from prototype inconsistencies. Their product determines the final view weights:

$$q_v = \mathrm{softmax}_v\left(\frac{s_v - \mu_s}{\sigma_s + \epsilon}\right), \quad r_v = \mathrm{softmax}_v\left(-\frac{d_v - \mu_d}{\tau_s}\right), \quad \pi_v = \frac{q_v r_v}{\sum_{u=1}^{V} q_u r_u}. \tag{12}$$

Here, $\mathrm{softmax}_v$ denotes normalization over all views. Higher reconstruction scores indicate stronger anomaly saliency, while lower prototype inconsistency suggests stronger agreement with the cross-view prototype prior. Structural consistency serves as an auxiliary prior, as genuine defects may also alter prototype consistency and thus this cue alone does not determine anomaly evidence reliability.

Finally, PPAFS uses α to interpolate between mean aggregation and the saliency- and prototype-weighted response:

$$S_{\text{final}} = (1-\alpha)\mu_s + \alpha\sum_{v=1}^{V}\pi_v s_v. \tag{13}$$

This formulation favors consensus under low disagreement and adaptive weighting under high disagreement. Training and inference pseudocode is provided in Appendix A.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset: We evaluate HPMD-AD in the multi-class multi-view setting on Real-IAD Wang et al. (2024) (30 categories) and Real-IAD Variety Zhu et al. (2026) (160 categories), both providing five

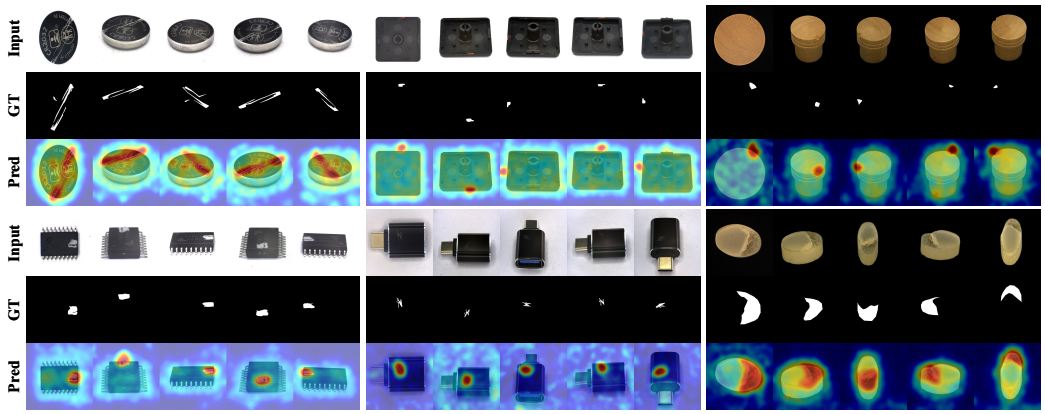

Figure 2: Qualitative anomaly localization results of HPMD-AD on representative multi-view samples from Real-IAD. We show the input image, ground-truth (GT), and predicted anomaly map.

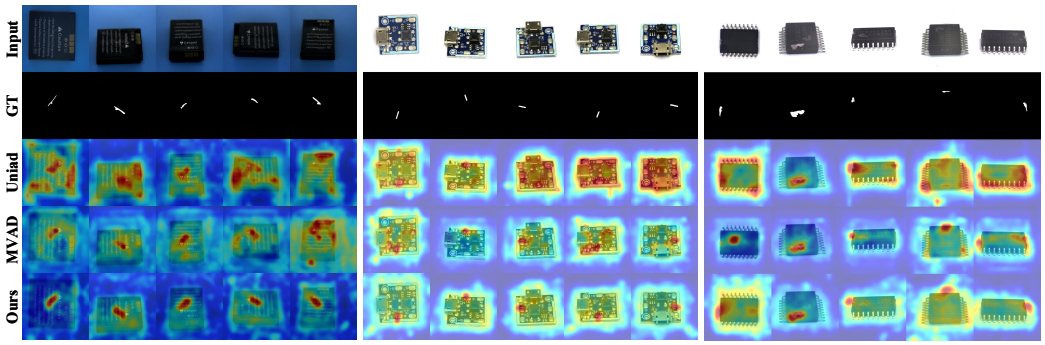

Figure 3: Qualitative comparison of anomaly localization among UniAD, MVAD, and Ours.

views per sample. Real-IAD Variety covers a larger category space with greater intra-class variation. **Evaluation Metrics:** We report AUROC, AP Zavrtanik et al. (2021), and F_1-max Zou et al. (2022) at the sample, image, and pixel levels, together with pixel-level AUPRO Bergmann et al. (2020). Extending the mAD aggregation protocol of Zhang et al. (2023a), we compute mAD as the mean of these ten metrics. **Baselines:** We compare HPMD-AD with representative single-view anomaly detection methods and the dedicated multi-view method MVAD He et al. (2026). **Implementation Details:** We set view-level and sample-level prototype numbers to $M = 8$ and $K = 4$, respectively, with prototype and feature masking ratios of 0.2. Further details are provided in Appendix C.

4.2 Comparison with SoTAs on Different AD Datasets

As shown in Table 1, HPMD-AD outperforms existing methods on both Real-IAD and Real-IAD Variety under the multi-class setting. On Real-IAD, HPMD-AD obtains the best results across sample-, image-, and pixel-level evaluations, achieving an overall mAD of 81.53%. Compared with MVAD, a dedicated multi-view anomaly detection approach, HPMD-AD improves the mAD by 3.49 points. At the sample level, HPMD-AD achieves 94.6/97.4/93.3 in AUROC/AP/F_1-max (%), yielding gains of 4.4/2.1/3.2 points over MVAD, respectively. Similar improvements are observed at the image and pixel levels, where HPMD-AD reaches 89.9% image-level AUROC, 99.1% pixel-level AUROC, and 92.1% AUPRO. Additional overall comparisons with recent methods on Real-IAD and detailed per-category quantitative comparisons are provided in Appendices F and G, respectively.

On the larger Real-IAD Variety benchmark with 160 categories, HPMD-AD achieves an overall mAD of 85.17%, outperforming the best competing method by 1.99 points. It achieves the highest performance across all ten evaluation metrics, including 96.5% sample-level AUROC and 92.1% image-level AUROC, which improve over the corresponding best baselines by 4.0 and 5.0 points, respectively. These results demonstrate the effectiveness of hierarchical prototype modeling in cap-

Table 2: **Multi-view context gain under partial observations on Real-IAD.** Reference and Missing share identical retained views and ground truth, while Reference additionally uses V_{add} context views for multi-view inference. Gain is computed as Reference minus Missing using the unrounded results. Reference and Missing are reported to one decimal place, and Gain to two decimal places.

V_{add}	Method	Setting	Sample-level			Image-level			Pixel-level				mAD
			AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO	
1	**HPMD-AD**	Reference	93.0	95.9	91.3	89.3	86.5	80.4	98.9	37.1	41.7	91.9	80.6
		Missing	92.6	95.4	91.2	89.0	85.7	80.1	98.9	36.2	40.7	91.7	80.1
		Gain	0.48	0.57	0.12	0.28	0.84	0.24	0.03	0.96	0.95	0.15	**0.46**
	MVAD	Reference	89.3	94.0	88.4	86.6	84.7	77.3	97.7	31.2	37.4	91.2	77.8
		Missing	89.3	94.0	88.4	86.6	84.7	77.2	97.7	31.2	37.4	91.1	77.8
		Gain	0.00	0.01	0.00	0.02	0.01	0.06	0.00	0.03	0.04	0.01	0.02
2	**HPMD-AD**	Reference	91.9	94.2	89.1	89.2	86.6	80.5	98.9	37.0	41.6	91.8	80.1
		Missing	91.2	93.3	88.8	88.7	85.5	80.1	98.8	35.9	40.5	91.5	79.4
		Gain	0.70	0.85	0.23	0.50	1.08	0.42	0.07	1.11	1.06	0.33	**0.63**
	MVAD	Reference	88.5	92.3	86.3	86.6	84.7	77.3	97.5	31.0	37.2	91.1	77.3
		Missing	88.4	92.2	86.3	86.6	84.7	77.4	97.5	31.0	37.1	91.1	77.2
		Gain	0.04	0.05	-0.01	0.04	0.04	-0.05	0.00	0.06	0.14	0.04	0.04
3	**HPMD-AD**	Reference	90.7	91.3	85.8	89.1	86.2	80.3	99.0	36.8	41.4	91.8	79.2
		Missing	89.8	90.0	85.3	88.4	84.9	79.6	98.9	35.5	40.1	91.2	78.4
		Gain	0.90	1.26	0.49	0.75	1.31	0.67	0.10	1.31	1.26	0.65	**0.87**
	MVAD	Reference	87.6	89.5	83.1	86.5	84.4	77.2	98.4	31.3	37.3	91.1	76.6
		Missing	87.4	89.3	82.9	86.4	84.3	77.1	98.4	31.1	37.1	91.0	76.5
		Gain	0.21	0.16	0.21	0.14	0.14	0.12	0.02	0.18	0.22	0.10	0.15

turing both view-specific normal patterns and cross-view shared structures, yielding robust representations despite the substantially expanded category space and increased intra-class variation. Detailed per-category quantitative results on Real-IAD Variety are reported in Appendix H.

4.3 Qualitative Results of HPMD-AD

To qualitatively evaluate anomaly localization, we visualize representative multi-view samples from Real-IAD. As shown in Fig. 2, HPMD-AD accurately localizes anomalous regions across viewpoints despite variations in appearance, orientation, and defect visibility, while suppressing responses in normal regions. It captures both small defects and large structural abnormalities, demonstrating effective use of complementary multi-view information. More results on Real-IAD and Real-IAD Variety are provided in Appendices I and J, respectively.

Figure 3 presents a qualitative comparison among UniAD, MVAD, and HPMD-AD. UniAD tends to generate relatively scattered responses over normal object regions, while MVAD benefits from multi-view information but may still respond to irrelevant structures in some cases. In contrast, HPMD-AD produces more compact anomaly responses concentrated around defective regions and better suppresses false activations across different viewpoints. This observation is consistent with our design motivation: hierarchical prototypes provide compact normal priors for modeling view-specific and cross-view structures, while prototype-conditioned reconstruction enables more accurate discrimination between normal variations and anomalous patterns.

4.4 Multi-View Context Gain under Partial Observations

We further investigate how effectively HPMD-AD exploits additional cross-view context under partial observations, without retraining or fine-tuning. Let $V_{\mathrm{add}} \in \{1, 2, 3\}$ denote the number of additional views available to Reference but not to Missing. Both settings use identical retained views and annotations, differing only in the availability of these additional views during multi-view inference. Accordingly, *Gain*, computed as Reference minus Missing, measures the benefit of the additional cross-view context. Further details of the paired evaluation protocol are provided in Appendix E.

As shown in Table 2, HPMD-AD benefits increasingly from additional cross-view observations. Specifically, its mAD *Gain* increases from 0.46 to 0.63 and 0.87 points as V_{add} grows from 1 to 3, indicating that HPMD-AD progressively exploits complementary information as more views become available. Under the same settings, HPMD-AD consistently achieves larger mAD gains than

Table 3: Ablation study on Real-IAD. Exp. (a) denotes the diffusion reconstruction baseline without prototype conditioning. VP and SP denote the view-level and sample-level prototypes, respectively. PM and SM denote prototype masking and feature masking, respectively.

Exp.	VP	SP	PM	SM	PPAFS	Sample-level			Image-level			Pixel-level			
						AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
(a)						90.8	95.5	90.5	85.9	83.2	76.4	98.3	35.6	40.3	89.6
(b)	✓					92.4	96.2	91.6	88.0	85.6	78.8	98.8	35.7	40.6	89.8
(c)	✓	✓				93.7	96.9	92.6	89.1	86.2	80.2	99.0	35.6	40.6	90.9
(d)	✓	✓	✓			93.8	97.0	92.5	89.5	86.9	80.5	99.0	37.4	41.8	91.6
(e)	✓	✓		✓		93.9	96.9	92.6	89.6	86.8	80.6	99.0	37.3	41.7	91.9
(f)	✓	✓	✓	✓		94.1	97.1	93.1	**89.9**	**87.1**	**81.1**	**99.1**	**38.5**	**42.4**	**92.1**
(g)	✓	✓	✓	✓	✓	**94.6**	**97.4**	**93.3**	**89.9**	**87.1**	**81.1**	**99.1**	**38.5**	**42.4**	**92.1**

MVAD. Moreover, HPMD-AD maintains higher absolute performance than MVAD under both Reference and Missing settings, including when only partial views are available. These results indicate that HPMD-AD not only performs strongly under incomplete observations, but also derives greater benefit from additional cross-view context, providing further evidence that hierarchical prototype modeling effectively captures and utilizes complementary information across views.

4.5 Ablation Studies and Analysis

Table 3 presents ablation studies of the key components in HPMD-AD on Real-IAD. We evaluate the contributions of HDPM, DMPCD, and PPAFS, with additional ablations provided in Appendix D.

Effect of HDPM. We first investigate the effect of hierarchical prototype modeling. Exp. (a) removes prototype conditioning and performs diffusion-based feature reconstruction without structured prototype priors. Introducing view-level prototypes in Exp. (b) improves sample-level AUROC from 90.8% to 92.4% and image-level AUROC from 85.9% to 88.0%, demonstrating the benefit of modeling view-specific normal patterns. Further incorporating sample-level prototypes in Exp. (c) increases the two metrics to 93.7% and 89.1%, respectively, indicating that modeling cross-view shared structures provides complementary information beyond individual views.

Effect of DMPCD. We then evaluate the dual masking strategy in DMPCD. Compared with Exp. (c), prototype masking in Exp. (d) improves image-level AUROC and pixel-level AUPRO (%) from 89.1/90.9 to 89.5/91.6, while feature masking in Exp. (e) increases them to 89.6/91.9. Combining both masking mechanisms in Exp. (f) further improves sample-level AUROC to 94.1% and pixel-level AUPRO to 92.1%, demonstrating the complementary benefits of the two masking mechanisms for robust prototype-conditioned reconstruction.

Effect of PPAFS. Compared with mean score fusion in Exp. (f), PPAFS improves sample-level AUROC/AP/F_1-max (%) from 94.1/97.1/93.1 to 94.6/97.4/93.3, demonstrating the benefit of jointly modeling inter-view score disagreement and prototype reliability for multi-view score aggregation.

5 Conclusion

In this work, we study multi-view anomaly detection from the perspective of structured cross-view normal modeling and propose HPMD-AD, a Hierarchical Prototype-Conditioned Masked Diffusion model. Rather than relying on dense cross-view feature interactions or explicit spatial correspondence, HPMD-AD organizes complementary observations into hierarchical prototypes that capture both view-specific normal patterns and cross-view shared structures. Specifically, HDPM, DMPCD, and PPAFS jointly leverage complementary information across views to enable robust multi-view anomaly detection. The resulting framework integrates cross-view representation learning, normal feature reconstruction, and score aggregation within a unified prototype-guided formulation. Extensive experiments on Real-IAD and the larger-scale Real-IAD Variety benchmark show that HPMD-AD achieves state-of-the-art performance in the multi-class multi-view setting. Moreover, our multi-view context analysis reveals progressively larger gains as additional views become available, providing further evidence that HPMD-AD effectively exploits complementary cross-view information.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, translation, polishing, drafting portions of the manuscript, and writing or refining parts of the experimental code. All AI-assisted text and code were carefully reviewed, revised, verified, and tested by the authors. The authors take full responsibility for the accuracy, originality, integrity, and correctness of the submitted work.

REFERENCES

Letian Bai, Chengyu Tao, and Juan Du. Sganet: Semantic and geometric alignment for multimodal multi-view anomaly detection. *arXiv preprint arXiv:2604.05632*, 2026.

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

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4182–4191. IEEE, 2020.

Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. A survey on visual anomaly detection: Challenge, approach, and prospect. *arXiv preprint arXiv:2401.16402*, 2024.

Shang-Fu Chen, Jhih-Ciang Wu, Kuan-Chuan Peng, Wen-Huang Cheng, and Kai-Lung Hua. Geomad: Geometry-aware multi-view anomaly detection via deformable fusion and distributional alignment. *arXiv preprint arXiv:2608.26724*, 2026.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9727–9736. IEEE, 2022.

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *2022 IEEE/CVF winter conference on applications of computer vision (WACV)*, pp. 1819–1828. IEEE, 2022.

Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. Encoder-decoder contrast for unsupervised anomaly detection in medical images. *IEEE transactions on medical imaging*, 43(3): 1102–1112, 2023a.

Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. *Advances in Neural Information Processing Systems*, 36: 10721–10740, 2023b.

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20405–20415. IEEE, 2025.

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. *Advances in Neural Information Processing Systems*, 37:71162–71187, 2024a.

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, pp. 8472–8480, 2024b.

Haoyang He, Jiangning Zhang, Guanzhong Tian, Chengjie Wang, and Lei Xie. Learning multi-view anomaly detection with efficient adaptive selection. *IEEE Transactions on Multimedia*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Duy Hoang Khuong, Tri Nguyen Minh, and Ngu Huynh Cong Viet. Borad: Bootstrap your own representations for multi-class anomaly detection. *arXiv preprint arXiv:2606.14129*, 2026.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Mathis Kruse and Bodo Rosenhahn. Multi-flow: Multi-view-enriched normalizing flows for industrial anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3933–3944. IEEE, 2025.

Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10:78446–78454, 2022.

Jiarui Lei, Xiaobo Hu, Yue Wang, and Dong Liu. Pyramidflow: High-resolution defect contrastive localization using pyramid normalizing flow. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14143–14152. IEEE, 2023.

Yufei Li, Long Tian, Yuyang Dai, Wenchao Chen, Liang Bao, and Xiyang Liu. Fastref: Fast prototype refinement for few-shot industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43040–43049, 2026.

Tongtong Liu, Hongxia Gao, Yuxuan Tan, Jinpeng Li, and Jinhui Zhao. Entropy-increasing linear attention for multi-class unsupervised anomaly detection. *Pattern Recognition*, pp. 113089, 2026.

Yifan Liu, Xun Xu, Shijie Li, Jingyi Liao, and Xulei Yang. Multi-view industrial anomaly detection with epipolar constrained cross-view fusion. *arXiv preprint arXiv:2503.11088*, 2025.

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

Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 36:8487–8500, 2023.

Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9974–9983. IEEE, 2025.

Kai Mao, Yiyang Lian, Yangyang Wang, Meiqin Liu, Nanning Zheng, and Ping Wei. Unveiling multi-view anomaly detection: Intra-view decoupling and inter-view fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12381–12389, 2025.

Arian Mousakhan, Thomas Brox, and Jawad Tayyub. Anomaly detection with conditioned denoising diffusion models. In *DAGM German Conference on Pattern Recognition*, pp. 181–195. Springer, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(5): 2293–2312, 2020.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14298–14308. IEEE, 2022.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24511–24520, 2023.

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 *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 22883–22892. IEEE, 2024.

Haonan Yin, Guanlong Jiao, Qianhui Wu, Borje F Karlsson, Biqing Huang, and Chin Yew Lin. Lafite: Latent diffusion model with feature editing for unsupervised multi-class anomaly detection. *arXiv preprint arXiv:2307.08059*, 2023.

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.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dræm–a discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8310–8319. IEEE, 2021.

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*, 2023a.

Jiangning Zhang, Chengjie Wang, Xiangtai Li, Guanzhong Tian, Zhucun Xue, Yong Liu, Guansong Pang, and Dacheng Tao. Learning feature inversion for multi-class anomaly detection under general-purpose coco-ad benchmark. *International Journal of Computer Vision*, 134(5):235, 2026. doi: 10.1007/s11263-026-02809-z.

Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3914–3923. IEEE, 2023b.

Jingyu Zhou, Min Liu, Yunfeng Ma, Shuai Jiang, and Yaonan Wang. Multi-view attention guided feature learning for unsupervised surface defect detection. *IEEE/ASME Transactions on Mechatronics*, 30(4):2844–2852, 2025.

Wenbing Zhu, Chengjie Wang, Bin-Bin Gao, Jiangning Zhang, Guannan Jiang, Jie Hu, Zhenye Gan, Lidong Wang, Ziqing Zhou, Jianghui Zhang, et al. Real-iad variety: Pushing industrial anomaly detection dataset to a modern era. *Pattern Recognition*, pp. 113354, 2026.

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*, pp. 392–408. Springer, 2022.

A Training and Inference Algorithms

Algorithms 1 and 2 summarize the training and inference procedures of HPMD-AD, respectively. During training, hierarchical prototypes guide feature reconstruction, while prototype masking and feature masking encourage the model to recover normal structures from incomplete information. During inference, both masking strategies are disabled, and DDIM reconstructs clean features from noise-corrupted test features conditioned on the hierarchical prototypes.

Algorithm 1 HPMD-AD training

Input: Normal data $\mathcal{D}$, frozen F, learning rate η.
Output: Trained parameters θ.

Initialize HDPM and DMPCD parameters θ.
repeat
  Draw $\mathcal{X} \sim \mathcal{D}$ and timestep t.
  $\boldsymbol{x}^v \leftarrow F(X^v), \forall v.$
  $\boldsymbol{U}^v \leftarrow \mathrm{CrossAttn}(\boldsymbol{Q}^v, \boldsymbol{x}^v, \boldsymbol{x}^v), \forall v.$
  $\boldsymbol{P}^v_{\mathrm{view}} \leftarrow \mathrm{FFN}(\boldsymbol{U}^v), \forall v.$
  Sample $M_p^v \in \{0,1\}, \sum_{v=1}^V M_p^v = 1.$
  $\overline{\boldsymbol{P}}^v_{\mathrm{view}} \leftarrow (1 - M_p^v)\boldsymbol{P}^v_{\mathrm{view}} + M_p^v \mathbf{1}_M \boldsymbol{m}_p^\top.$
  $\boldsymbol{M} \leftarrow \mathrm{Concat}_v(\overline{\boldsymbol{P}}^v_{\mathrm{view}} + \boldsymbol{e}_v).$
  $\boldsymbol{P}_{\mathrm{sample}} \leftarrow F_s(\boldsymbol{Q}_s, \boldsymbol{M})$, where F_s applies Eq. 3 followed by the output projection.
  for $v = 1, \ldots, V$ **do**
    Draw $\boldsymbol{\epsilon}_x^v \sim \mathcal{N}(0, \boldsymbol{I})$.
    $\boldsymbol{x}_t^v \leftarrow \sqrt{\bar{\alpha}_t}\boldsymbol{x}^v + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_x^v.$
    $\boldsymbol{h}_t^v \leftarrow \mathrm{Proj}(\boldsymbol{x}_t^v) + \boldsymbol{E}_{\mathrm{pos}}.$
    Sample feature mask $\boldsymbol{M}_f^v$.
    $\widetilde{\boldsymbol{h}}_t^v \leftarrow (1 - \boldsymbol{M}_f^v) \odot \boldsymbol{h}_t^v + \boldsymbol{M}_f^v \odot (\boldsymbol{m}_f + \boldsymbol{E}_{\mathrm{pos}}).$
    Draw $\boldsymbol{\epsilon}_p^v \sim \mathcal{N}(0, \boldsymbol{I})$.
    $\boldsymbol{P}_t^v \leftarrow \sqrt{\bar{\alpha}_t}\,\mathrm{Concat}(\boldsymbol{P}^v_{\mathrm{view}}, \boldsymbol{P}_{\mathrm{sample}}) + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_p^v.$
    $\hat{\boldsymbol{x}}_0^v \leftarrow \mathrm{DiT}_\theta(\widetilde{\boldsymbol{h}}_t^v, t, \boldsymbol{P}_t^v).$
  end for
  $\mathcal{L}_{\mathrm{diff}} \leftarrow \frac{1}{V}\sum_{v=1}^V \|\hat{\boldsymbol{x}}_0^v - \boldsymbol{x}^v\|_2^2.$
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\mathrm{diff}}.$
until convergence

Algorithm 2 HPMD-AD inference

Input: Test sample $\mathcal{X}$, F, θ, DDIM schedule $\mathcal{T}$ from t_0.
Output: Anomaly maps $\{\boldsymbol{A}^v\}$ and score S_{final}.

Disable both masks.
Extract $\boldsymbol{x}^v$ and view prototypes $\boldsymbol{P}^v_{\mathrm{view}}, \forall v.$
$\boldsymbol{M} \leftarrow \mathrm{Concat}_v(\boldsymbol{P}^v_{\mathrm{view}} + \boldsymbol{e}_v)$, $\boldsymbol{P}_{\mathrm{sample}} \leftarrow F_s(\boldsymbol{Q}_s, \boldsymbol{M})$.
for $v = 1, \ldots, V$ **do**
  Draw $\boldsymbol{\epsilon}_x^v \sim \mathcal{N}(0, \boldsymbol{I})$.
  $\boldsymbol{x}_{t_0}^v \leftarrow \sqrt{\bar{\alpha}_{t_0}}\boldsymbol{x}^v + \sqrt{1-\bar{\alpha}_{t_0}}\boldsymbol{\epsilon}_x^v.$
  Draw $\boldsymbol{\epsilon}_p^v \sim \mathcal{N}(0, \boldsymbol{I})$.
  $\boldsymbol{P}_t^v \leftarrow \sqrt{\bar{\alpha}_t}\,\mathrm{Concat}(\boldsymbol{P}^v_{\mathrm{view}}, \boldsymbol{P}_{\mathrm{sample}}) + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_p^v, \quad \forall t \in \mathcal{T}.$
  $\hat{\boldsymbol{x}}_0^v \leftarrow \mathrm{DDIM}_\theta(\boldsymbol{x}_{t_0}^v, \{\boldsymbol{P}_t^v\}_{t\in\mathcal{T}}; \mathcal{T}).$
  $a_n^v \leftarrow \|\boldsymbol{x}_n^v - \hat{\boldsymbol{x}}_{0,n}^v\|_2^2, \forall n.$
  $\boldsymbol{A}^v \leftarrow \mathrm{Up}_{\mathrm{bilinear}}(\mathrm{Reshape}(\boldsymbol{a}^v)).$
  $s_v \leftarrow \max(\boldsymbol{A}^v).$
end for
$(\mu_s, \sigma_s) \leftarrow \mathrm{MeanStd}_{\mathrm{pop}}(\{s_v\}).$
$D_s \leftarrow \frac{\sigma_s}{\mu_s + \sigma_s + \epsilon}$; $\alpha \leftarrow 1 - (1 - D_s)^{V-1}.$
$q_v \leftarrow \mathrm{Softmax}_v((s_v - \mu_s)/(\sigma_s + \epsilon)).$
$\boldsymbol{c}_v \leftarrow \mathcal{N}_\delta(\mathrm{Mean}_m(\mathcal{N}_\delta(\boldsymbol{p}_m^v))).$
$\boldsymbol{c}_s \leftarrow \mathcal{N}_\delta(\mathrm{Mean}_v(\boldsymbol{c}_v)).$
$d_v \leftarrow 1 - \boldsymbol{c}_v^\top \boldsymbol{c}_s, \forall v.$
$(\mu_d, \sigma_d) \leftarrow \mathrm{MeanStd}_{\mathrm{pop}}(\{d_v\})$; $\tau_s \leftarrow \sigma_d + \epsilon.$
$r_v \leftarrow \mathrm{Softmax}_v(-(d_v - \mu_d)/\tau_s).$
$\pi_v \leftarrow \frac{q_v r_v}{\sum_{u=1}^V q_u r_u}.$
$S_{\mathrm{final}} \leftarrow (1-\alpha)\mu_s + \alpha \sum_v \pi_v s_v.$

B Diffusion Preliminaries

We briefly review the diffusion formulation used in our method, following the standard DDPM formulation Ho et al. (2020). Given a clean representation $\mathbf{x}_0$, the forward diffusion process gradually perturbs it with Gaussian noise according to

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{14}$$

which allows direct sampling as

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{15}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$.

Prediction Parameterization. The denoising model can be parameterized to predict either the injected noise $\boldsymbol{\epsilon}$ or the clean representation $\mathbf{x}_0$. Under noise prediction, the model is trained as

$$\hat{\boldsymbol{\epsilon}}_t = f_\theta(\mathbf{x}_t, t), \qquad \mathcal{L}_\epsilon = \|\hat{\boldsymbol{\epsilon}}_t - \boldsymbol{\epsilon}\|_2^2. \tag{16}$$

Alternatively, under the x_0-prediction parameterization, the model directly predicts the clean representation:

$$\hat{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t, t), \qquad \mathcal{L}_{x_0} = \|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|_2^2. \tag{17}$$

For a given $\mathbf{x}_t$ and diffusion schedule, the two prediction parameterizations are analytically interconvertible. An x_0 prediction gives the corresponding noise estimate as

$$\hat{\boldsymbol{\epsilon}}_t = \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0}{\sqrt{1-\bar{\alpha}_t}}, \tag{18}$$

while a noise prediction yields

$$\hat{\mathbf{x}}_0 = \frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_t}{\sqrt{\bar{\alpha}_t}}. \tag{19}$$

DDIM Sampling. During inference, we employ DDIM Song et al. (2020) sampling for the reverse process. Let t' denote the previous timestep in the selected DDIM sampling schedule. Given $\hat{\mathbf{x}}_0$ and the corresponding noise estimate $\hat{\boldsymbol{\epsilon}}_t$, the update from timestep t to t' is

$$\mathbf{x}_{t'} = \sqrt{\bar{\alpha}_{t'}}\hat{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_{t'}-\sigma_t^2}\hat{\boldsymbol{\epsilon}}_t + \sigma_t \mathbf{z}, \tag{20}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and

$$\sigma_t = \eta\sqrt{\frac{1-\bar{\alpha}_{t'}}{1-\bar{\alpha}_t}}\sqrt{1-\frac{\bar{\alpha}_t}{\bar{\alpha}_{t'}}}. \tag{21}$$

When $\eta = 0$, DDIM becomes deterministic.

In HPMD-AD, we adopt the x_0-prediction parameterization to directly reconstruct normal feature representations. Diffusion is performed in the DINOv2 feature space, while the denoising network is additionally conditioned on the hierarchical prototypes introduced in the main paper. The formulation above describes the underlying diffusion process, with prototype conditioning incorporated into the denoising model.

C Additional Experimental Details

Dataset: Real-IAD Wang et al. (2024) is a large-scale real-world benchmark designed for multi-view industrial anomaly detection. It contains 151,050 high-resolution images from 30 industrial object categories, including 99,721 normal images and 51,329 anomalous images. Each object is captured from five different viewpoints, with image resolutions ranging from approximately 2K to 5K, providing complementary visual observations across views. The dataset covers diverse industrial defects, including pit, deformation, abrasion, scratch, damage, missing parts, foreign objects, and contamination, reflecting realistic challenges in industrial inspection scenarios. In addition, Real-IAD provides pixel-level anomaly masks and supports evaluation at sample, image, and pixel levels, enabling comprehensive assessment of multi-view anomaly detection methods. We adopt Real-IAD as the primary benchmark to evaluate the effectiveness of HPMD-AD in multi-view industrial anomaly detection.

Real-IAD Variety Zhu et al. (2026) is a recently proposed large-scale multi-view industrial anomaly detection benchmark with substantially expanded category diversity and visual complexity. It contains 198,950 high-resolution images from 160 object categories, covering 28 industrial domains, 24 material types, 22 color variations, and 27 defect types. Each sample is captured from five distinct viewpoints with high-quality pixel-level annotations. Compared with Real-IAD, Real-IAD Variety introduces a much larger category space and more diverse industrial scenarios, materials, appearances, and defect patterns, making it suitable for evaluating the scalability and generalization capability of multi-class multi-view anomaly detection methods. We further evaluate HPMD-AD on Real-IAD Variety to assess its robustness when scaling to a substantially larger and more heterogeneous category space.

Evaluation Metrics: We evaluate anomaly detection performance at the sample, image, and pixel levels. For the single-view baseline methods, we follow the evaluation setting of MVAD He et al. (2026), where each view is processed independently, the maximum response of the anomaly map

Table 4: Ablation study on the components of PPAFS. DG and PR denote disagreement gating and prototype reliability, respectively.

Fusion Strategy	DG	PR	Sample-level		
			AUROC	AP	F_1-max
Mean			94.1	97.1	93.1
Max			93.1	96.6	92.0
Saliency			93.7	96.9	92.5
Adaptive Mean–Saliency	✓		94.2	97.2	93.1
Prototype-Calibrated Saliency		✓	94.1	97.1	92.7
PPAFS	✓	✓	**94.6**	**97.4**	**93.3**

is used as the image-level score, and the maximum image-level score across the five views is used as the sample-level score. At each level, we report the Area Under the Receiver Operating Characteristic Curve (AUROC), Average Precision (AP) Zavrtanik et al. (2021), and maximum F1-score (F_1-max) Zou et al. (2022). AUROC measures the overall ranking ability across different decision thresholds, while AP is more sensitive to performance under class imbalance. F_1-max reports the best trade-off between precision and recall over all possible thresholds. For pixel-level anomaly localization, we additionally report the Area Under the Per-Region-Overlap curve (AUPRO) Bergmann et al. (2020), which evaluates region-level localization quality and is less dominated by large anomalous regions than pixel-wise metrics alone. To provide a comprehensive assessment across sample-level detection, image-level detection, and pixel-level localization, we further compute the mean of all ten reported metrics, denoted as mAD Zhang et al. (2023a). A higher value indicates better performance for all metrics.

Baselines: We compare HPMD-AD with representative anomaly detection methods spanning embedding-, reconstruction-, flow-, distillation-, and transformer-based paradigms. On Real-IAD, we include the single-view methods SimpleNet Liu et al. (2023), CFA Lee et al. (2022), CFlow-AD Gudovskiy et al. (2022), PyramidFlow Lei et al. (2023), RD4AD Deng & Li (2022), ViTAD Zhang et al. (2023a), UniAD You et al. (2022), RD++ Tien et al. (2023), and DeSTSeg Zhang et al. (2023b), together with MVAD He et al. (2026), a dedicated multi-view anomaly detection method. For the larger-scale Real-IAD Variety benchmark, we compare against SimpleNet, RD4AD, ViTAD, UniAD, MVAD, and recent multi-class methods including MambaAD He et al. (2024a), EILA Liu et al. (2026), and Dinomaly Guo et al. (2025), providing a diverse set of baselines for evaluating performance under substantially increased category diversity.

Implementation Details: We employ a pretrained DINOv2-B Oquab et al. (2023); Darcet et al. (2024) as the frozen feature extractor and aggregate intermediate features from transformer layers 6–11 for subsequent prototype modeling and reconstruction. In HDPM, the numbers of view-level and sample-level prototypes are set to $M = 8$ and $K = 4$, respectively. In DMPCD, both the prototype masking ratio and feature masking ratio are set to 0.2. Both masking strategies are applied only during training and disabled at inference. For diffusion training, we adopt $T = 1000$ timesteps with a linear noise schedule ranging from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$, following the standard DDPM setting Ho et al. (2020). During inference, we start the reverse process from timestep $t_0 = 860$ and employ DDIM Song et al. (2020) sampling with only 3 denoising steps for efficient feature reconstruction. The input images are first resized to 512×512 and then center-cropped to 448×448. The model is optimized using Adam Kingma & Ba (2014) with a learning rate of 1×10^{-4}. The batch size is set to 30 multi-view samples, corresponding to 150 images in total since each sample contains five views. All experiments are implemented using PyTorch 2.1.0 Paszke et al. (2019) and conducted on a single NVIDIA H100 GPU with 80 GB of memory.

D Additional Ablation Studies

To provide a more comprehensive analysis of HPMD-AD, we conduct additional ablation studies to further examine the effectiveness and robustness of its key design choices. These experiments complement the analysis in the main paper and provide further insights into the proposed framework.

Table 5: Effect of different inference timesteps t on Real-IAD. † denotes the default setting, and the best result in each row is highlighted in bold.

t	10	50	100	200	300	400	500	600	700	800	860†	900	950	990
Sample AUROC	79.2	79.5	79.9	81.1	83.2	85.8	88.3	90.4	92.2	93.9	94.6	**94.7**	94.0	91.8
Image AUROC	73.1	73.4	73.8	75.3	77.6	80.5	83.5	86.1	88.1	89.6	**89.9**	89.4	87.6	85.4
Pixel AUROC	96.4	96.4	96.5	96.7	97.1	97.6	98.1	98.5	98.8	99.0	**99.1**	**99.1**	98.9	98.6

Effect of PPAFS. We provide a more detailed analysis of the score fusion mechanism in Table 4. Mean fusion serves as a consensus-based baseline and achieves 94.1% AUROC. In contrast, directly emphasizing highly responsive views through max or saliency fusion reduces AUROC to 93.1% and 93.7%, respectively, suggesting that a strong view-wise response is not always reliable anomaly evidence. Introducing disagreement gating (DG) allows the fusion strategy to adapt between cross-view consensus and salient responses according to the degree of score disagreement, improving AUROC to 94.2%. Prototype reliability (PR) provides an additional structural cue derived from cross-view prototype consistency. Although PR alone does not improve AUROC over mean fusion, combining it with DG further improves the sample-level AUROC/AP/F_1-max (%) to 94.6/97.4/93.3. This indicates that prototype reliability is most effective as a complementary calibration signal: DG determines when salient-view evidence should receive more attention, while PR helps regulate how strongly individual views contribute. Their combination enables PPAFS to adaptively exploit complementary anomaly evidence while reducing the influence of structurally inconsistent views.

Effect of the inference timestep.

We further investigate the sensitivity of diffusion reconstruction to the starting timestep t in Table 5. In the high-timestep regime, HPMD-AD maintains stable performance from $t = 800$ to 900, achieving sample-/image-/pixel-level AUROC (%) of 93.9/89.6/99.0, 94.6/89.9/99.1, and 94.7/89.4/99.1 at $t = 800$, 860, and 900, respectively. Across these settings, the sample-, image-, and pixel-level AUROC values vary by only 0.8, 0.5, and 0.1 percentage points, respectively, indicating that the method remains effective over a relatively broad range of starting timesteps. A small t introduces only limited perturbation to the input features, which may make reconstruction overly dependent on the observed features and weaken the resulting anomaly discrepancy. Increasing t provides stronger corruption and can encourage greater reliance on the learned normal prototype priors. However, excessive corruption may remove useful structural information and hinder normal feature recovery, consistent with the performance degradation observed at $t = 950$ and 990. Unless otherwise specified, we use $t = 860$ as a fixed default setting across all experiments.

E Details of Multi-View Context Analysis

We provide additional details of the paired evaluation protocol used in Table 2. For each test sample, we generate a deterministic sample-specific permutation of the five views using a fixed seed and a stable sample identifier composed of the category, relative parent directory, and sample ID. Taking the first k entries of the same permutation produces nested retained-view subsets. We consider $k \in \{2, 3, 4\}$, corresponding to $V_{\text{add}} = 5 - k$ additional context views. The same retained subsets are used across all compared methods.

For each method, sample, and k, Reference and Missing use exactly the same retained views and evaluation annotations. Reference receives all five views during multi-view inference, whereas Missing receives only the retained k views. For HPMD-AD, the additional views participate in hierarchical prototype construction and feature reconstruction, while they are completely excluded from the inference input of Missing. These additional views serve only as inference context and do not contribute separate scores or annotations to metric computation.

Both settings are evaluated exclusively on the same retained views. Image- and pixel-level metrics are computed from the retained observations and their corresponding annotations. Sample-level scores are likewise recomputed using only the retained views, and the sample-level ground-truth label is defined as the logical OR of their view labels. Thus, for every sample at a fixed k, Reference and Missing are evaluated on the same retained observations using identical ground-truth annota-

Table 6: Comparison with recent state-of-the-art methods on Real-IAD under the multi-class setting. Methods are grouped into single-view and multi-view approaches. "–" denotes results not reported in the original papers under the same setting.

Setting	Method	Sample-level	Image-level	Pixel-level
Single-view	BoRAD	–	87.4	98.2
	ReContrast	–	86.4	97.8
	DRAEM	–	50.9	44.0
	DiAD	–	75.6	88.0
	InvAD	–	89.4	98.4
	MambaAD	–	87.0	98.6
	Dinomaly	–	89.3	98.8
Multi-view	MVAG-FL	–	89.7	98.9
	MVEAD	**94.6**	87.8	–
	GeoMAD	91.8	87.5	97.8
	HPMD-AD	**94.6**	**89.9**	**99.1**

tions. Sample-level labels may differ across values of k because the retained subsets themselves change.

For each method, all settings use the same checkpoint trained with complete five-view samples, without retraining or fine-tuning. For HPMD-AD, stochastic diffusion noise is additionally paired between Reference and Missing for each retained view by deterministically deriving the random seed from the sample identifier, original view ID, noise type, and diffusion timestep. This ensures matched diffusion-noise realizations for corresponding retained views in the paired comparison. Metrics are averaged over the 30 categories. Gain is computed as Reference minus Missing from the unrounded results and reported in percentage points. Consequently, Gain isolates the effect of providing additional cross-view inference context while keeping the evaluated observations fixed.

F Additional Comparison with State-of-the-Art Methods on Real-IAD

We further compare HPMD-AD with additional representative anomaly detection methods on Real-IAD under the multi-class setting. Specifically, the single-view methods include BoRAD Khuong et al. (2026), ReContrast Guo et al. (2023b), DRAEM Zavrtanik et al. (2021), DiAD He et al. (2024b), InvAD Zhang et al. (2026), MambaAD He et al. (2024a), and Dinomaly Guo et al. (2025), while the multi-view methods include MVAG-FL Zhou et al. (2025), MVEAD Liu et al. (2025) and GeoMAD Chen et al. (2026). Since the original papers do not consistently report the full set of ten evaluation metrics used in our main comparison, we provide an additional comparison based on AUROC, a widely adopted and representative metric for anomaly detection, using the available results at the sample, image, and pixel levels. As shown in Table 6, HPMD-AD achieves the highest reported image- and pixel-level AUROC, while matching the highest reported sample-level AUROC. Results not reported in the original papers under the same multi-class setting are denoted by "–".

G Per-Category Quantitative Comparison on Real-IAD

To provide a more fine-grained evaluation, we report category-wise quantitative results on Real-IAD at the sample, image, and pixel levels in Table 7. For UniAD and MVAD, the category-wise results are taken from an earlier version of MVAD, since the final TMM version reports only aggregate results under the multi-class setting. Therefore, their category-wise averages may differ slightly from the aggregate results reported in Table 1. Across the 30 categories, HPMD-AD maintains consistently strong performance, indicating that the overall improvements are not driven by a small subset of object categories but extend across diverse object appearances and anomaly patterns. These results complement the aggregate comparison in the main paper with a more detailed category-level analysis.

Table 7: Per-category quantitative comparison among UniAD, MVAD, and HPMD-AD on Real-IAD under the multi-class multi-view setting. Results are reported at the sample, image, and pixel levels, with the best results highlighted in bold.

	Sample-level AUROC/AP/F_1-max			**Image-level** AUROC/AP/F_1-max			**Pixel-level** AUROC/AP/F_1-max/AUPRO		
Category	UniAD	MVAD	**Ours**	UniAD	MVAD	**Ours**	UniAD	MVAD	**Ours**
audiojack	90.3/95.2/88.3	91.2/94.9/90.7	**93.1/96.5/91.3**	81.4/76.6/64.9	80.3/72.8/62.8	**91.1/85.0/77.3**	97.6/20.0/31.0/83.7	97.1/**25.9/36.2**/81.9	**98.9**/20.6/33.7/**85.4**
bottle_cap	97.2/98.7/93.6	97.8/**98.9**/95.3	**97.9/98.9/96.1**	92.5/91.7/81.7	92.4/90.9/81.5	**93.3/92.1/83.9**	99.5/19.4/29.6/96.0	**99.6**/19.5/27.2/**96.8**	99.5/**34.2/38.3**/95.5
button_battery	**95.5/95.1/95.0**	80.0/90.4/83.4	92.0/95.0/91.1	75.9/81.6/76.3	76.6/83.2/75.6	**86.7/91.1/80.8**	96.7/28.5/34.4/77.5	97.9/44.8/47.8/84.2	**98.7/65.4/61.0/88.9**
end_cap	87.7/**94.2**/86.4	86.6/93.0/87.4	**89.2**/93.6/**89.7**	80.9/**86.1**/78.0	79.4/84.6/77.4	**84.6**/84.1/**80.8**	95.8/8.8/17.4/85.4	96.7/11.9/19.9/89.0	**99.1/12.8/22.7/95.0**
eraser	90.1/95.3/87.9	87.2/94.3/85.2	**92.2/96.4/89.4**	90.3/89.2/80.2	88.6/87.2/77.8	**93.2/91.8/83.0**	99.3/24.4/30.9/94.1	99.2/25.3/30.7/93.8	**99.5/36.9/39.2/96.3**
fire_hood	85.9/92.9/85.5	81.8/90.7/82.1	**88.0/93.2/87.8**	80.6/74.8/66.4	78.6/71.8/64.1	**84.2/78.0/71.4**	98.6/23.4/32.2/85.3	98.7/21.3/28.6/84.5	**99.4/44.5/44.5/88.6**
mint	66.5/89.7/89.7	68.1/91.0/89.7	**83.2/96.0/89.8**	67.0/66.6/64.6	68.9/70.2/64.5	**75.1/74.1/70.2**	94.4/7.7/18.1/62.3	95.9/14.4/26.5/**71.8**	**97.9/19.8/34.9**/69.8
mounts	98.3/99.2/95.5	98.8/99.5/96.1	**99.6/99.8/98.5**	87.6/77.3/**77.2**	89.5/**81.5**/76.5	**90.2**/78.4/76.0	99.4/28.0/32.8/95.2	99.2/26.3/32.9/91.0	**99.7/41.5/44.4/97.1**
pcb	83.8/91.6/85.2	90.0/95.3/87.8	**93.1/96.9/89.6**	81.0/88.2/79.1	87.7/92.5/83.1	**91.4/95.3/87.1**	97.0/18.5/28.1/81.6	98.8/34.6/42.1/91.4	**99.4/53.6/56.0/93.8**
phone_battery	85.4/93.1/85.3	92.7/96.7/90.5	**95.7/98.1/92.7**	83.6/80.0/71.6	**90.6/89.1/81.5**	90.2/88.9/80.0	85.5/11.2/21.6/88.5	80.8/22.0/31.4/93.9	**99.6/40.8/44.2/95.0**
plastic_nut	84.7/89.5/87.6	90.8/94.7/89.1	**91.3/94.9/91.3**	80.0/69.2/63.7	84.9/77.2/69.5	**89.1/83.7/75.9**	98.4/20.6/27.1/88.9	99.2/24.6/29.9/94.5	**99.5/37.0/37.7/96.9**
plastic_plug	80.2/90.3/82.9	89.1/94.6/87.6	**94.5/97.6/91.1**	81.4/75.9/67.6	85.2/80.1/71.6	**89.1/84.3/78.7**	98.6/17.4/26.1/90.3	**99.1/21.3/28.3/93.5**	98.9/17.4/27.3/92.9
porcelain_doll	90.1/91.9/90.4	94.9/97.3/93.6	**97.4/98.8/95.2**	85.1/75.2/69.3	**89.2/83.4/76.4**	83.5/63.9/68.9	98.7/14.1/24.5/93.2	**99.2/27.9/34.3/95.7**	99.1/8.2/16.6/92.7
regulator	65.7/80.6/82.1	73.6/86.8/81.2	**97.0/98.5/93.8**	56.9/41.5/44.5	66.6/55.4/47.2	**84.7/76.5/68.5**	95.5/9.1/17.4/76.1	97.1/10.2/23.5/85.3	**99.0/39.0/42.4/89.2**
rolled_strip_base	98.6/99.3/97.5	97.7/98.7/95.4	**99.4/99.6/98.8**	98.7/99.3/96.5	96.9/98.2/94.0	**99.2/99.6/97.7**	99.6/20.7/32.2/97.8	99.5/20.4/29.4/97.8	**99.7/30.4/37.2/98.0**
sim_card_set	87.3/90.8/87.0	96.2/97.9/95.0	**98.3/99.1/96.8**	89.7/90.3/83.2	94.1/94.9/87.6	**96.9/97.3/91.5**	97.9/31.6/39.8/85.0	98.9/45.5/46.1/89.8	**99.7/56.9/55.9/95.4**
switch	91.5/96.2/88.7	94.9/97.6/91.4	**99.2/99.7/97.2**	85.5/88.6/78.4	89.1/91.6/81.8	**97.1/97.8/91.8**	98.1/33.8/40.6/90.7	98.7/34.5/41.9/93.2	**99.7/76.8/69.0/93.6**
tape	98.2/99.0/96.0	98.4/99.3/96.0	**99.2/99.6/97.6**	**97.2/96.2**/89.4	96.8/96.1/**89.8**	97.1/95.7/88.6	99.7/29.2/36.9/97.5	99.7/35.2/41.8/98.1	**99.8/39.8/45.1/98.6**
terminalblock	95.7/98.2/93.1	96.8/98.7/94.2	**98.5/99.5/97.8**	87.5/89.1/81.0	93.5/94.4/87.3	**98.5/99.0/96.1**	99.2/23.1/30.5/94.4	99.6/29.7/35.5/97.1	**99.8/43.6/43.2/98.6**
toothbrush	89.8/92.9/90.0	87.0/93.1/87.5	**98.4/98.7/97.2**	78.4/80.1/75.6	84.8/86.7/79.7	**89.5/90.0/84.4**	95.7/16.4/25.3/84.3	**97.1/27.3**/35.1/89.6	97.0/25.5/**36.8/89.7**
toy	75.4/85.7/85.0	86.4/92.3/86.5	**94.6/97.2/91.9**	68.4/75.1/74.8	79.1/84.2/78.0	**85.4/88.0/80.4**	93.4/4.6/12.4/70.5	95.9/12.5/21.8/83.7	**97.2/16.3/23.0/92.6**
toy_brick	78.6/84.3/84.2	69.4/81.7/80.8	**83.7/91.0/86.0**	77.0/71.1/66.2	66.4/58.8/60.6	**78.6/76.3/68.6**	97.4/17.1/27.6/**81.3**	96.0/13.4/21.4/71.7	**97.9/43.0/45.3**/80.9
transistor1	98.4/99.2/95.7	98.8/99.5/96.3	**99.4/99.7/98.5**	93.7/95.9/88.9	94.3/96.0/89.2	**96.9/97.8/92.6**	98.9/25.6/33.2/94.3	99.3/37.6/39.6/96.0	**99.6/44.4/44.6/97.4**
u_block	94.8/97.0/91.2	93.0/96.1/90.0	**98.3/99.1/95.7**	88.8/**84.2/75.5**	89.1/84.0/74.2	**90.0**/83.8/75.3	99.3/22.3/29.6/94.3	**99.5**/28.3/37.2/**96.5**	99.4/**40.2/45.5**/92.3
usb	79.7/88.4/82.7	92.1/95.9/89.0	**95.4/97.6/93.5**	78.7/79.4/69.1	90.1/90.5/81.9	**94.8/93.2/85.7**	97.9/20.6/31.7/85.3	99.2/33.6/39.9/**94.4**	**99.7/36.7/44.2**/94.2
usb_adaptor	82.9/90.3/86.3	91.0/95.5/90.4	**95.1/97.9/93.5**	76.8/71.3/64.9	78.1/**72.4**/66.1	**81.8**/71.9/**68.6**	96.6/10.5/19.0/78.4	97.1/14.4/22.3/80.9	**99.5/22.1/24.5/93.8**
vcpill	80.7/89.9/83.6	88.2/94.2/85.3	**97.4/98.7/94.6**	87.1/84.0/74.7	83.7/80.9/70.5	**91.5/89.8/81.9**	99.1/40.7/43.0/**91.3**	98.6/40.9/47.4/88.0	**99.3/67.7/63.5**/87.2
wooden_beads	77.3/90.1/86.2	82.2/92.7/86.3	**91.8/96.9/90.5**	78.4/77.2/67.8	84.3/83.1/73.1	**88.6/87.0/78.9**	97.6/16.5/23.6/84.6	98.1/26.5/35.1/**86.8**	**99.2/46.5/48.6/86.8**
woodstick	84.0/90.7/**85.1**	76.8/87.5/81.6	**84.1/92.2**/83.3	80.8/72.6/63.6	78.0/65.3/59.8	**85.9/79.3/73.1**	94.0/36.2/44.3/77.2	97.4/35.0/40.7/81.5	**99.1/57.0/56.9/90.1**
zipper	97.8/98.6/96.8	99.9/99.9/98.7	**100.0/100.0/99.7**	98.2/98.9/95.3	**98.9/99.4/95.7**	97.6/97.8/94.8	98.4/32.5/36.1/95.1	**99.1/45.4/51.3/96.6**	98.8/36.9/45.4/95.3
Mean	87.1/92.9/88.8	89.0/94.6/89.5	**94.6/97.4/93.3**	83.0/80.9/74.3	85.2/83.2/76.0	**89.9/87.1/81.1**	97.3/21.1/29.2/86.7	97.7/27.0/34.2/89.6	**99.1/38.5/42.4/92.1**

H Per-Category Quantitative Results on Real-IAD Variety

We further report detailed category-wise quantitative results on Real-IAD Variety, a substantially more challenging benchmark with 160 object categories, broader category diversity, and increased intra-class variation. As shown in Table 8, HPMD-AD maintains strong performance across diverse categories at the sample, image, and pixel levels. This suggests that the hierarchical prototype modeling remains effective under substantially increased category diversity, providing robust representations of both view-specific and cross-view shared normal patterns even in a much larger category space. The category-wise results complement the aggregate comparison in the main paper and provide a more fine-grained evaluation of HPMD-AD under the substantially enlarged category space and more challenging variations in object appearance and anomaly patterns.

I Additional Qualitative Results on Real-IAD

To provide a more comprehensive qualitative evaluation, we present anomaly localization results of HPMD-AD across all 30 object categories in Real-IAD. As shown in Figs. 4 and 5, the visualizations cover diverse object appearances, defect patterns, and viewpoints, with the input images, ground-truth masks (GT), and predicted anomaly maps shown for each multi-view sample. Despite substantial variations in object appearance, orientation, and defect visibility across viewpoints, HPMD-AD consistently produces pronounced responses around the annotated anomalous regions, covering both small localized defects and larger structural abnormalities. These results complement the quantitative evaluation in the main paper and further demonstrate the robustness of HPMD-AD across diverse industrial objects and anomaly patterns.

Table 8: Per-category quantitative results of HPMD-AD on Real-IAD Variety under the multi-class multi-view setting. We report sample-level and image-level AUROC/AP/F_1-max, and pixel-level AUROC/AP/F_1-max/AUPRO.

No.	Category	Sample-level			Image-level			Pixel-level				No.	Category	Sample-level			Image-level			Pixel-level			
1	2pin_block_plug	99.9	100	99.5	99.0	99.9	98.6	92.1	59.6	56.1	77.1	2	3_adapter	97.6	99.7	98.3	92.7	99.0	94.5	96.7	24.5	38.3	78.5
3	3pin_aviation_connector	91.4	98.9	95.3	91.0	98.6	94.5	97.1	64.8	61.0	85.4	4	4_wire_stepping_motor	86.6	98.1	94.1	62.5	91.8	93.3	65.9	2.1	7.2	36.5
5	DVD_switch	95.3	99.1	99.5	99.5	99.9	99.3	88.3	54.6	56.7	73.2	6	D_sub_connector	100	100	100	99.7	99.9	98.9	97.3	54.7	53.5	90.2
7	LED_indicator	99.8	100	99.2	93.3	99.0	95.4	85.0	6.1	13.9	60.5	8	PLCC_socket	100	100	100	98.9	99.9	98.3	99.0	79.8	71.6	89.8
9	VR_joystick	98.1	99.7	99.0	97.9	99.6	97.0	98.0	60.3	55.8	89.2	10	access_card	96.1	99.4	98.5	79.1	93.5	96.2	63.9	15.3	24.8	75.8
11	accurate_detection_switch	97.1	99.6	97.5	91.8	99.1	96.2	93.6	59.3	55.9	77.2	12	aircraft_model_head	100	100	100	100	100	100	97.6	38.2	43.1	87.2
13	angled_toggle_switch	100	100	100	93.5	99.0	97.6	84.3	11.0	18.6	57.5	14	audio_jack_socket	93.7	99.2	95.8	80.1	94.2	94.0	91.1	7.7	13.6	59.2
15	bag_buckle	100	100	100	99.5	99.9	99.6	99.5	72.2	65.9	96.9	16	ball_pin	100	100	100	99.8	100	99.3	98.8	61.4	59.0	95.9
17	balun_transformer	97.7	99.7	97.2	94.7	99.0	95.2	95.0	60.9	61.8	65.6	18	battery	76.6	96.0	94.3	74.1	94.7	95.7	86.0	13.2	20.2	53.2
19	battery_holder_connector	99.4	99.9	99.2	96.3	99.5	96.8	90.6	55.2	57.0	86.5	20	battery_socket_connector	89.8	98.7	95.0	82.2	97.2	93.6	92.6	47.5	52.7	63.4
21	bend_connector	95.6	99.4	97.3	89.5	98.6	95.2	90.5	31.1	31.3	70.8	22	blade_switch	99.8	100	99.5	98.2	99.8	97.5	95.2	32.5	45.1	79.1
23	blue_light_switch	86.7	97.9	95.2	79.3	96.0	91.9	86.7	25.8	33.0	62.5	24	bluetooth_module	100	100	99.7	98.6	99.8	98.2	97.5	70.9	63.3	87.5
25	boost_converter_module	98.2	99.8	98.5	90.8	98.3	94.1	96.6	70.2	67.5	75.3	26	bread_model	89.0	98.6	94.3	83.6	97.0	91.5	97.9	71.4	64.9	74.1
27	brooch_clasp_accessory	100	100	100	97.4	99.6	97.1	97.1	69.1	65.8	82.4	28	button_battery_holder	99.7	100	99.7	99.9	100	99.6	97.1	48.9	51.8	83.2
29	button_motor	98.3	99.8	99.0	91.1	98.5	96.9	98.8	70.8	65.3	91.7	30	button_switch	97.8	99.6	99.5	80.8	97.4	95.7	99.5	84.3	76.6	85.8
31	car_door_lock_switch	96.7	99.5	98.3	85.6	97.7	94.6	97.1	62.8	62.1	88.6	32	ceramic_fuse	97.9	99.7	98.0	91.6	98.8	93.6	90.3	29.8	40.7	78.4
33	ceramic_wave_filter	100	100	99.7	99.9	100	99.6	94.4	39.1	47.8	84.4	34	charging_port	99.5	99.9	99.2	94.8	99.3	94.6	94.8	45.9	48.6	81.9
35	chip_inductor	98.2	99.8	98.0	96.8	99.5	96.5	94.1	47.4	50.5	78.9	36	circuit_breaker	100	100	100	99.1	99.9	98.6	98.7	63.3	60.7	95.1
37	circular_aviation_connector	90.7	98.7	95.8	84.1	94.8	89.1	82.9	26.9	31.9	58.0	38	common_mode_choke	92.5	98.9	96.6	87.3	97.3	91.2	96.2	52.0	51.7	82.6
39	common_mode_filter	99.4	99.9	99.0	98.6	99.8	97.8	91.9	43.4	48.1	71.2	40	connector	99.6	100	99.2	97.3	99.6	96.8	85.1	48.5	47.4	72.0
41	connector_housing_female	99.1	99.9	99.8	99.8	100	99.7	89.4	57.9	60.5	85.4	42	console_switch	100	100	100	99.2	99.9	99.0	82.1	14.9	24.9	76.8
43	crimp_st_cable_mount_box	96.6	99.5	98.3	83.8	97.8	95.3	99.1	75.3	70.5	88.6	44	dc_jack	98.4	99.8	98.8	86.7	97.5	94.2	86.9	47.0	42.8	76.2
45	dc_power_connector	100	100	100	99.4	99.9	99.0	93.5	40.0	41.9	83.7	46	detection_switch	97.5	99.6	98.0	96.0	99.3	96.0	93.7	55.2	54.8	75.3
47	duckbill_circuit_breaker	97.6	99.7	98.2	84.1	97.5	93.3	96.4	37.8	41.1	80.0	48	earphone_audio_unit	87.5	98.3	94.2	73.7	95.6	94.0	94.4	45.4	46.3	62.1
49	effect_transistor	90.6	98.6	96.3	92.5	98.4	93.1	98.9	76.2	71.7	87.2	50	electronic_watch_movement	79.9	97.2	94.3	79.9	96.6	94.0	97.7	61.6	57.1	72.2
51	ethernet_connector	98.5	99.8	98.0	94.0	99.2	96.5	95.9	51.8	59.8	77.4	52	ferrite_bead	97.9	99.7	98.0	95.5	99.3	96.4	99.3	72.4	69.2	92.6
53	ffc_connector_plug	99.2	99.9	99.5	98.0	99.6	96.8	97.8	45.6	49.3	81.6	54	flow_control_valve	98.3	99.8	97.7	83.1	97.3	94.0	98.4	50.1	50.8	89.0
55	flower_copper_shape	100	100	99.7	99.1	99.9	98.2	94.0	55.7	52.3	82.1	56	flower_velvet_fabric	98.5	99.8	98.8	98.1	99.8	98.3	91.2	62.5	60.5	74.5
57	fork_crimp_terminal	99.7	100	99.5	86.2	96.4	95.9	98.2	45.0	44.5	90.8	58	fuse_cover	86.3	98.2	96.2	77.4	95.7	92.3	91.2	47.3	51.4	76.6
59	fuse_holder	100	100	100	100	100	100	93.8	18.2	25.2	65.9	60	gear	100	100	100	100	100	100	98.7	71.8	67.2	92.2
61	gear_motor	72.1	93.6	97.3	63.2	89.8	94.3	87.2	6.1	13.5	65.8	62	green_ring_filter	97.8	99.7	98.5	87.9	96.1	90.6	93.8	27.7	36.0	70.4
63	hairdryer_switch	99.7	100	99.0	99.1	99.9	98.6	98.7	68.9	62.5	80.4	64	hall_effect_sensor	98.1	99.7	99.7	100	100	100	97.3	50.5	57.2	85.2
65	headphone_jack_female	100	100	100	100	100	100	95.2	61.4	62.3	84.2	66	headphone_jack_socket	89.2	98.6	94.5	87.2	98.1	94.0	96.7	59.5	56.8	86.7
67	hex_plug	100	100	100	97.1	99.6	96.9	94.9	59.4	58.0	80.4	68	humidity_sensor	99.5	99.9	99.7	94.4	98.9	98.3	96.9	32.5	32.0	88.4
69	ingot_buckle	99.7	100	99.5	98.1	99.8	97.5	94.3	33.7	45.5	78.9	70	insect_metal_parts	95.7	99.5	96.5	84.4	97.9	95.2	93.5	18.1	23.5	76.4
71	inverter_connector	99.9	100	99.7	99.3	99.9	98.6	99.1	59.5	58.5	95.7	72	jam_jar_model	80.5	97.5	94.1	76.3	95.0	91.8	85.9	43.6	47.3	77.4
73	joystick_switch	100	100	100	99.8	100	99.6	99.3	54.9	52.0	97.8	74	kfc_push_key_switch	99.8	100	99.7	96.3	99.5	95.9	94.4	44.0	47.9	74.2
75	knob_cap	96.2	99.4	98.8	98.3	99.7	97.8	93.7	26.9	31.0	68.8	76	laser_diode	86.5	97.4	95.4	54.0	85.4	90.3	79.7	4.6	9.3	66.5
77	lattice_block_plug	97.9	99.7	97.5	79.3	96.5	93.4	96.1	17.1	29.3	78.1	78	lego_pin_connector_plate	100	100	100	100	100	100	97.9	52.5	59.4	85.4
79	lego_propeller	99.7	100	99.3	96.4	99.5	97.3	96.5	38.8	42.1	79.9	80	lego_reel	100	100	100	100	100	100	93.9	56.9	57.8	89.0
81	lego_technical_gear	100	100	100	97.9	99.7	97.4	99.1	76.7	70.8	92.5	82	lego_turbine	99.2	99.9	98.5	97.4	99.1	96.8	97.4	63.8	62.0	80.4
83	lighting_connector	86.0	96.7	94.1	90.2	97.8	91.6	97.1	65.6	59.8	74.0	84	lilypad_led	99.2	99.9	98.5	88.5	98.2	94.5	94.9	41.3	45.2	91.0
85	limit_switch	99.9	100	99.5	89.4	93.7	94.5	95.5	32.2	40.7	81.5	86	lithium_battery_plug	96.3	99.5	97.8	94.3	99.0	96.1	97.7	70.2	67.7	80.1
87	littel_fuse	100	100	100	96.3	99.5	96.4	89.7	35.7	46.3	78.4	88	little_cow_model	81.5	96.5	95.2	83.2	96.6	92.8	97.5	61.3	59.0	81.1
89	lock	99.8	100	99.5	95.2	99.0	97.8	90.3	44.0	43.5	75.3	90	long_zipper	97.7	99.7	97.7	96.4	99.5	96.4	90.1	18.8	26.7	80.3
91	meteor_hammer_arrowhead	99.8	100	99.5	87.8	97.7	93.1	97.1	39.6	44.3	83.5	92	miniature_laser_module	97.6	99.7	98.7	98.0	99.8	98.2	98.4	55.5	52.3	93.8
93	miniature_lifting_motor	92.2	98.9	96.9	89.0	98.7	96.0	98.0	72.9	67.6	61.9	94	miniature_motor	96.8	99.4	99.0	93.0	99.1	96.6	98.3	62.8	58.8	77.9
95	miniature_stepper_motor	96.8	99.6	96.8	97.4	99.8	98.2	90.5	39.7	44.8	57.5	96	mobile_charging_connector	99.6	99.9	99.5	98.9	99.9	99.0	97.7	72.0	66.8	84.7
97	model_steering_module	93.3	99.2	95.2	93.9	98.8	94.8	98.3	52.0	54.7	76.5	98	monitor_socket	99.9	100	99.7	98.8	99.8	98.6	98.8	73.6	66.0	88.5
99	motor_bracket	96.7	99.6	97.0	92.9	99.2	95.9	98.6	79.8	74.1	86.5	100	motor_gear_reducer	92.0	99.0	95.6	92.5	99.1	94.7	87.5	31.5	39.2	73.2
101	motor_plug	99.8	100	99.7	100	100	100	98.8	80.8	76.4	92.6	102	mouse_socket	88.7	98.5	94.2	88.9	97.8	91.2	98.5	46.1	50.5	68.5
103	multi_function_switch	100	100	100	99.1	99.9	98.3	89.1	43.3	47.3	79.6	104	nylon_ball_head	99.7	100	99.5	93.9	99.3	98.0	93.1	49.0	50.8	79.3
105	optical_fiber_outlet	99.5	99.9	99.3	100	100	99.6	99.1	66.7	61.4	93.5	106	pencil_sharpener	98.7	99.8	97.8	89.4	98.0	93.3	97.9	43.9	40.9	72.3
107	pinboard_connector	95.3	99.4	97.8	94.6	99.1	97.2	96.1	43.1	39.5	91.2	108	pitch_connector	99.8	100	99.2	96.4	99.4	95.9	94.9	68.7	65.1	88.8
109	pneumatic_elbow	93.7	99.2	96.1	89.7	98.4	94.6	92.0	50.4	51.8	64.2	110	pot_core	99.5	99.9	99.2	95.7	99.3	97.5	92.2	41.0	45.5	54.5
111	potentiometer	100	100	100	99.4	99.9	98.9	99.4	61.5	62.5	94.4	112	power_bank_module	99.1	99.9	99.0	98.5	99.7	99.6	94.6	46.9	47.3	77.2
113	power_inductor	99.0	99.9	98.5	93.0	99.3	96.3	91.6	25.1	32.4	67.2	114	power_jack	88.8	98.4	95.2	85.4	98.3	95.6	92.8	4.2	9.4	67.8
115	power_strip_socket	99.9	100	99.5	92.8	98.1	92.4	99.2	72.2	68.5	92.0	116	pulse_transformer	97.2	99.6	98.0	85.4	95.7	94.9	88.1	38.6	45.8	79.0
117	purple_clay_pot	96.7	99.3	99.5	99.8	100	99.6	98.5	68.9	65.7	86.6	118	push_button_switch	100	100	100	92.9	98.2	97.6	97.3	47.0	54.2	77.9
119	push_in_terminal	98.8	99.9	98.5	96.9	99.7	96.5	99.1	50.9	51.2	92.8	120	recorder_switch	91.8	98.9	95.0	84.6	97.6	93.9	98.6	44.9	46.1	78.1
121	rectangular_connector_accessories	99.8	100	99.5	93.3	97.5	93.0	95.9	38.7	43.5	85.1	122	red_terminal	97.8	99.7	98.0	87.2	97.9	93.6	98.1	66.5	71.4	86.4
123	retaining_ring	99.0	99.9	98.5	93.4	98.8	95.7	80.2	42.0	41.4	62.2	124	rheostat	97.2	99.6	97.3	89.1	98.7	96.6	97.6	40.4	40.2	88.6
125	rotary_position_sensor	99.9	100	99.5	92.2	98.7	99.0	93.6	45.3	41.6	82.1	126	round_twist_switch	88.0	98.4	94.3	86.5	98.3	94.7	98.0	48.7	51.9	83.7
127	self_lock_switch	99.7	100	99.3	97.1	99.6	96.8	95.8	10.4	18.8	82.1	128	shelf_support_module	95.6	99.2	98.5	90.9	98.4	95.4	88.3	34.8	42.6	69.8
129	side_press_switch	99.8	100	99.7	98.6	99.8	97.9	98.6	66.5	66.9	89.8	130	side_release_buckle	98.6	99.8	99.0	96.1	99.5	96.0	97.2	39.5	42.7	93.6
131	silicon_cell_sensor	98.5	99.8	98.8	87.3	97.9	95.9	96.9	16.6	30.7	86.8	132	sim_card_reader	96.3	99.5	96.7	93.4	99.0	95.4	95.0	62.7	60.3	83.0
133	single_pole_potentiometer	95.9	99.5	97.0	84.3	98.0	94.7	95.2	19.0	30.2	66.5	134	single_switch	98.2	99.8	97.8	92.7	99.0	94.4	98.6	44.3	58.1	74.9
135	slide_switch	100	100	100	95.5	99.3	96.4	95.9	36.5	39.8	73.0	136	slipper_model	99.3	99.9	98.8	91.2	98.7	94.7	96.9	53.8	53.4	87.5
137	small_leaf	94.5	99.2	96.4	82.5	95.4	91.9	96.9	41.9	45.8	82.3	138	smd_receiver_module	98.3	99.8	98.5	91.1	97.8	91.9	95.1	62.4	59.5	78.7
139	solderless_adapter	97.8	99.7	97.8	94.0	99.2	95.4	96.3	36.0	36.5	90.2	140	spherical_airstone	99.5	99.9	98.7	88.6	98.5	95.7	98.3	59.9	56.8	83.3
141	spring_antenna	98.2	99.8	97.3	95.9	99.2	94.8	96.6	54.8	55.6	91.1	142	square_terminal	92.0	98.5	96.8	88.0	98.2	94.0	82.0	46.4	47.8	74.1
143	steering_T_head	100	100	100	100	100	99.7	99.1	76.9	71.5	89.5	144	steering_wheel	89.1	98.4	95.5	90.8	99.0	95.4	97.6	70.5	65.3	68.0
145	suction_cup	100	100	100	94.7	99.3	95.5	98.2	66.3	64.3	84.2	146	telephone_spring_switch	97.0	99.6	97.3	97.5	99.7	97.9	98.4	70.8	63.5	92.1
147	tension_snap_release_clip	99.9	100	99.5	99.5	99.9	98.9	95.8	59.1	60.3	86.3	148	thumbtack	84.6	93.3	98.3	73.5	89.1	94.4	98.1	30.3	39.2	88.0
149	thyristor	100	100	100	100	100	100	91.6	24.0	25.5	80.5	150	toy_tire	99.5	99.9	99.0	95.5	99.3	95.4	96.6	64.1	61.5	80.7
151	traceless_hair_clip	99.4	99.9	98.5	98.4	99.8	97.8	98.9	68.8	65.9	84.0	152	travel_switch	97.8	99.7	97.5	89.0	98.3	95.5	80.5	9.1	14.9	45.9
153	travel_switch_green	99.9	100	99.5	98.5	99.8	97.8	97.6	46.9	47.8	92.0	154	tubular_key_switch	97.7	99.7	97.1	92.3	98.9	94.1	84.1	19.7	26.0	67.4
155	vacuum_switch	84.1	97.7	95.4	85.3	97.7	94.3	97.3	33.9	44.5	75.9	156	vehicle_harness_conductor	97.5	99.7	97.8	86.6	97.3	94.8	97.0	40.4	43.7	92.4
157	vertical_adjustable_resistor	98.3	99.8	98.5	90.6	97.2	91.2	97.0	48.4	48.8	85.5	158	vibration_motor	99.4	99.9	99.5	83.5	96.6	93.3	95.7	19.4	29.9	77.5
159	volume_potentiometer	89.0	97.3	97.1	91.9	98.5	95.4	96.3	72.2	69.5	77.2	160	wireless_receiver_module	99.2	99.9	99.5	98.3	99.8	98.6	87.9	36.5	42.3	76.8

J Additional Qualitative Results on Real-IAD Variety

To further assess the scalability of HPMD-AD in large-scale multi-class anomaly detection, we conduct additional qualitative evaluation on Real-IAD Variety. The dataset contains 160 object categories and exhibits substantially greater category diversity and intra-class variation than Real-IAD. Due to the large number of categories and limited space, we visualize 54 categories and present the results in three parts, aiming to provide broad coverage of the diverse category space of Real-IAD Variety. The visualized categories span a wide range of industrial scenarios and object types, including electronic components, mechanical parts, connectors, switches, sensors, gears, buckles, and

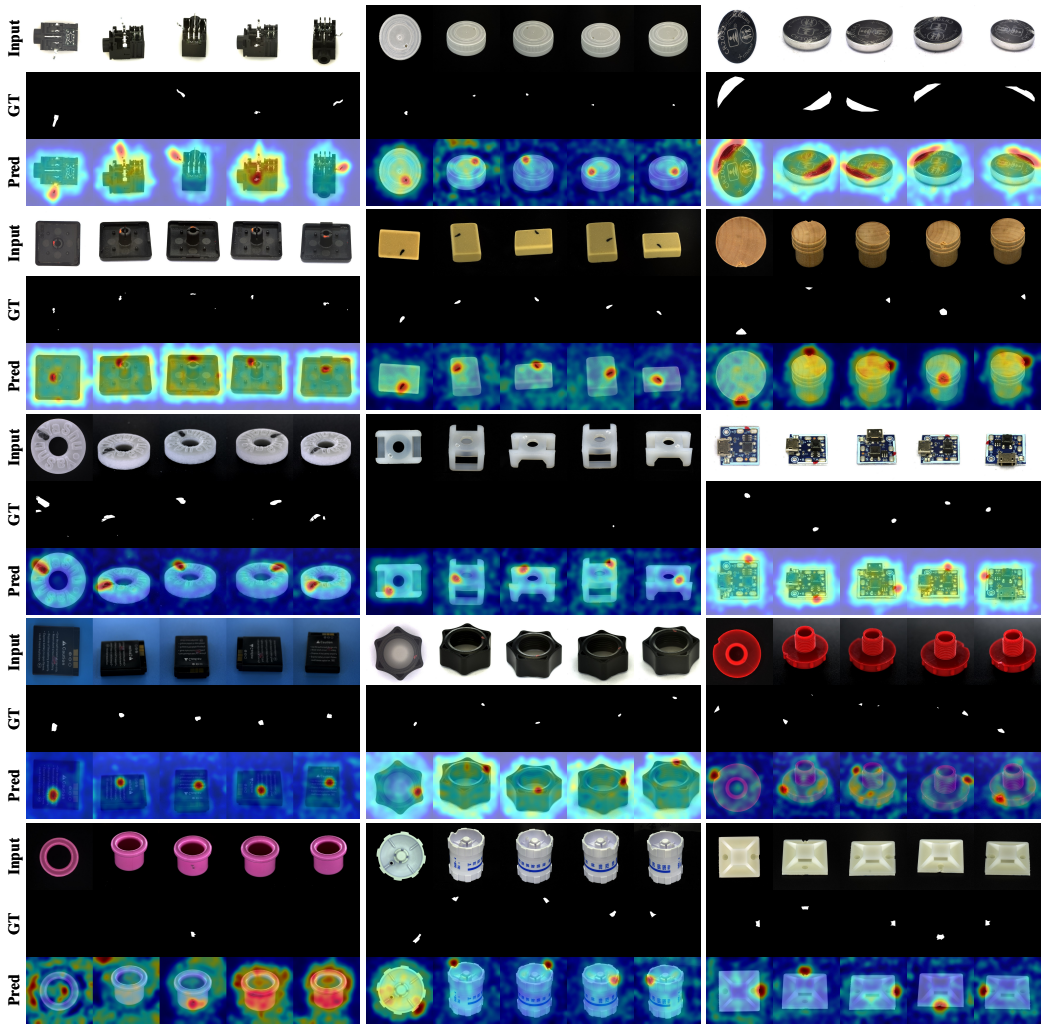

Figure 4: Qualitative anomaly localization results of HPMD-AD on Real-IAD (Part 1 of 2). The visualization covers all 30 object categories together with Fig. 5. For each multi-view sample, input views, ground-truth masks (GT), and predicted anomaly maps (Pred) are shown.

other manufactured objects. They also cover diverse materials, shapes, textures, object scales, and structural complexities, allowing the qualitative evaluation to reflect the diversity of the benchmark as comprehensively as possible.

As shown in Figs. 6, 7, and 8, HPMD-AD consistently produces pronounced responses around the annotated anomalous regions across these diverse categories. The predicted anomaly maps remain well aligned with the ground-truth masks for both small localized defects and larger structural abnormalities, despite substantial variations in object appearance, geometry, texture, and anomaly visibility. Together with the quantitative evaluation conducted on all 160 categories, these qualitative results further demonstrate the scalability and generalization capability of HPMD-AD under the substantially increased category diversity of Real-IAD Variety.

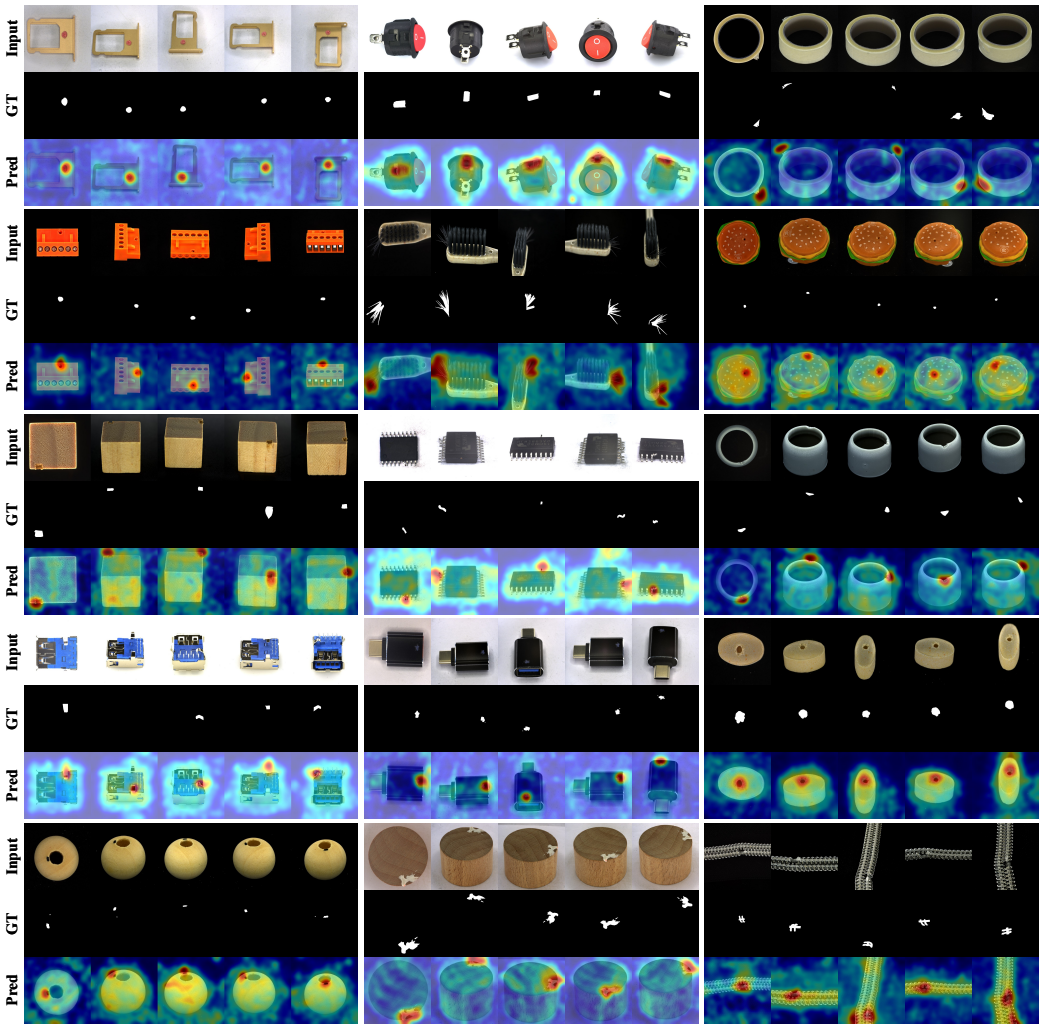

Figure 5: Qualitative anomaly localization results of HPMD-AD on Real-IAD (Part 2 of 2). The visualization covers all 30 object categories together with Fig. 4. For each multi-view sample, input views, ground-truth masks (GT), and predicted anomaly maps (Pred) are shown.

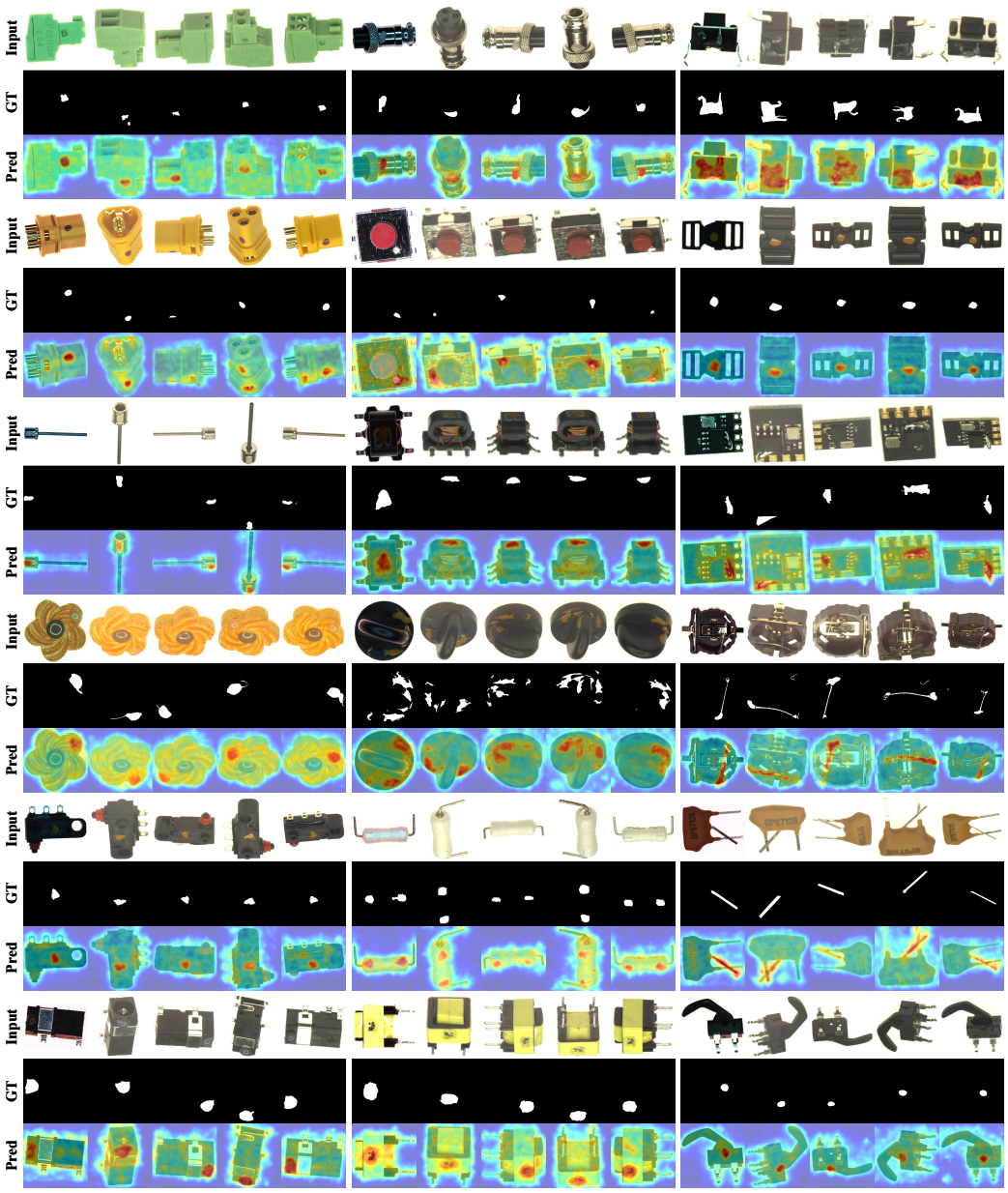

Figure 6: Qualitative anomaly localization results of HPMD-AD on Real-IAD Variety (Part 1 of 3). The visualization covers diverse categories from the 160-category benchmark, showing input views, ground-truth masks (GT), and predicted anomaly maps (Pred).

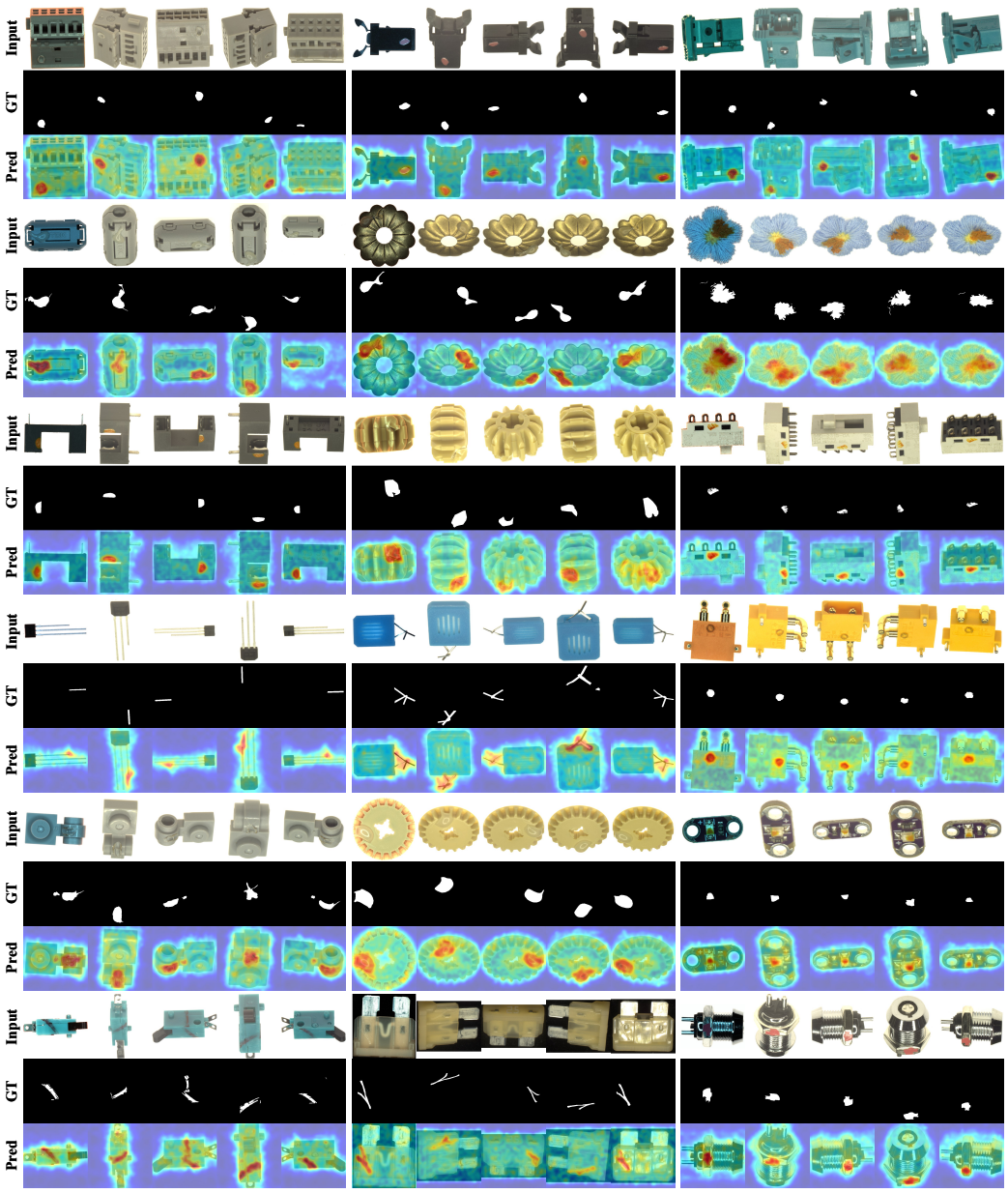

Figure 7: Qualitative anomaly localization results of HPMD-AD on Real-IAD Variety (Part 2 of 3). The visualization covers diverse categories from the 160-category benchmark, showing input views, ground-truth masks (GT), and predicted anomaly maps (Pred).

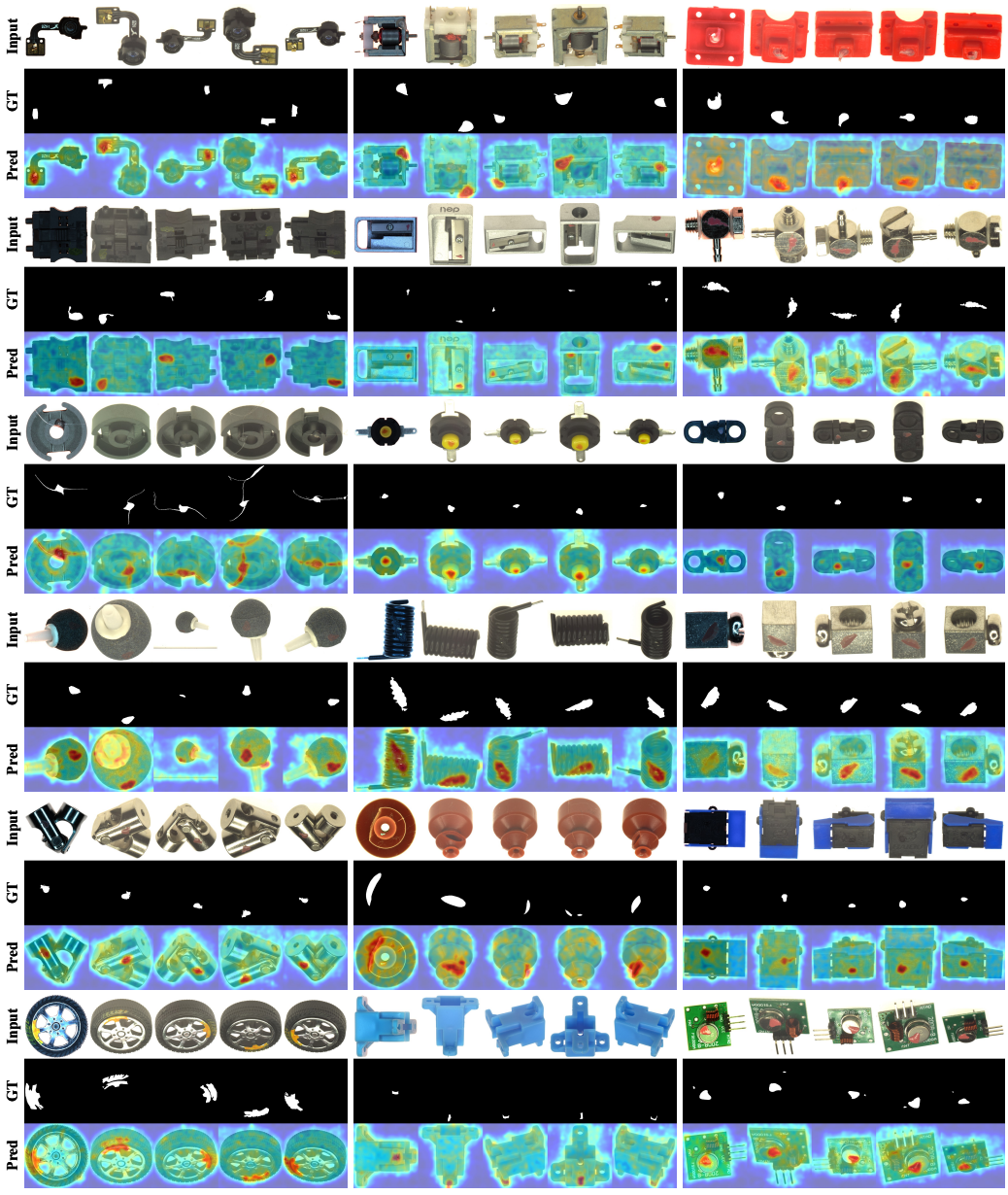

Figure 8: Qualitative anomaly localization results of HPMD-AD on Real-IAD Variety (Part 3 of 3). The visualization covers diverse categories from the 160-category benchmark, showing input views, ground-truth masks (GT), and predicted anomaly maps (Pred).