MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

See More, Detect Less? Taming Information Leakage in Multi-View Anomaly Detection

Chen, Shang-Fu; Peng, Kuan-Chuan; Wu, Jhih-Ciang; Cheng, Wen-Huang; Hua, Kailung

TR2026-171 November 11, 2026

Abstract

In multi-view anomaly detection, more cross-view information can actually hurt. When multiple inspection views are naively fused in a reconstruction-based pipeline, normal cues from intact views propagate to the decoder, which faithfully reconstructs anomalous regions, collapsing the reconstruction gap the detector depends on. We call this failure mode cross-view information leakage and show that effective multi-view fusion must explicitly restrict the information reaching the decoder. Building on this insight, we present GLAD (Global-Local Attention Driven framework), the first framework combining vision foundation model features with local and global cross-view fusion for multi-view anomaly detection. The Multi-view Merging Attention (MMA) module performs local cross-view fusion at linear complexity with learnable view importance weighting and token-wise gating, letting each view selectively incorporate fine-grained evidence from other views at O(N)cost. The Object-Guided Attention (OGA) module captures global context by aggregating class tokens from all views into a single object-level representation and broadcasting it back to patch tokens via temperature-scaled sigmoid gating, replacing the original patch representations rather than adding a residual to preserve the reconstruction gap. Experiments on Real-IAD and MANTA-Tiny show that GLAD outperforms state-of-the-art methods across sample-, image-, and pixel-level metrics, confirming that principled information restriction is key to multi-view anomaly reasoning.

ACM International Conference on Multimedia (ACM MM) 2026

© 2026 ACM. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org or Publications Dept., ACM, Inc., fax +1 (212) 869-0481.

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

See More, Detect Less? Taming Information Leakage in Multi-View Anomaly Detection

Shang-Fu Chen
National Taiwan University
Taipei, Taiwan
chenshangfu@cmlab.csie.ntu.edu.tw

Kuan-Chuan Peng
Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA, USA
kpeng@merl.com

Jhih-Ciang Wu
National Taiwan Normal University
Taipei, Taiwan
jcwu@csie.ntnu.edu.tw

Wen-Huang Cheng
National Taiwan University
Taipei, Taiwan
VinUniversity
Hanoi, Vietnam
wenhuang@csie.ntu.edu.tw

Kai-Lung Hua
Microsoft Taiwan Corporation,
National Taiwan University of Science and Technology
Taipei, Taiwan
kai.hua@microsoft.com

Abstract

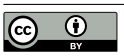
In multi-view anomaly detection, more cross-view information can actually hurt. When multiple inspection views are naively fused in a reconstruction-based pipeline, normal cues from intact views propagate to the decoder, which faithfully reconstructs anomalous regions, collapsing the reconstruction gap the detector depends on. We call this failure mode *cross-view information leakage* and show that effective multi-view fusion must explicitly restrict the information reaching the decoder. Building on this insight, we present GLAD (Global-Local Attention Driven framework), the first framework combining vision foundation model features with local and global cross-view fusion for multi-view anomaly detection. The Multi-view Merging Attention (MMA) module performs local cross-view fusion at linear complexity with learnable view importance weighting and token-wise gating, letting each view selectively incorporate fine-grained evidence from other views at $O(N)$ cost. The Object-Guided Attention (OGA) module captures global context by aggregating class tokens from all views into a single object-level representation and broadcasting it back to patch tokens via temperature-scaled sigmoid gating, replacing the original patch representations rather than adding a residual to preserve the reconstruction gap. Experiments on Real-IAD and MANTA-Tiny show that GLAD outperforms state-of-the-art methods across sample-, image-, and pixel-level metrics, confirming that principled information restriction is key to multi-view anomaly reasoning.

CCS Concepts

• **Computing methodologies** → **Visual inspection**; **Anomaly detection**; *Neural networks.*

Keywords

Multi-View, Anomaly Detection



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835123

ACM Reference Format:
Shang-Fu Chen, Kuan-Chuan Peng, Jhih-Ciang Wu, Wen-Huang Cheng, and Kai-Lung Hua. 2026. See More, Detect Less? Taming Information Leakage in Multi-View Anomaly Detection. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835123

1 Introduction

Industrial inspection systems [11, 32] often capture each object from multiple camera angles to achieve complete surface coverage. This acquisition process naturally produces a multi-view visual signal, in which different viewpoints provide complementary evidence about the object's geometry and appearance. Multi-view anomaly detection (AD) in such a setting is of great practical importance for industrial quality control. However, most existing approaches [13, 14, 43, 44] still treat each view as an independent sample, largely ignoring the cross-view relationships inherent in the data. At the same time, recent progress in single-view anomaly detection [8, 12, 34] has been substantial, particularly with reconstruction-based frameworks [14, 43] built upon Vision Transformer (ViT) architectures [9, 15, 29, 40, 43]. These methods typically freeze a pre-trained encoder and train a lightweight decoder to reconstruct only normal patterns, enabling anomalies to be identified through reconstruction errors. While effective for single-view inspection, this paradigm does not fully suit the multi-view setting, where cross-view consistency is essential for resolving ambiguities in individual viewpoints. In practice, a defect that is difficult to recognize in one view may become evident in another. Motivated by this, recent benchmarks such as Real-IAD [32] and MANTA [11] have brought multi-view anomaly detection to the forefront. Yet existing methods [17, 26] remain limited, as they either fail to model global object-level interactions across views or introduce substantial computational overhead for high-dimensional feature fusion.

Vision foundation models [6, 8, 12, 25, 27] have shown strong performance in single-view AD by providing rich, pre-trained representations. Within reconstruction-based frameworks [9, 14, 23, 43], a frozen ViT encoder extracts these features, while a lightweight

decoder learns to reproduce normal patterns, revealing anomalies through reconstruction discrepancies. However, extending these powerful representations to effective multi-view reasoning remains unexplored. To this end, we propose the Global-Local Attention Driven (GLAD) framework, the first approach that integrates vision foundation model features with both global and local cross-view fusion for multi-view AD. GLAD encodes each view independently using a shared frozen encoder and subsequently fuses the resulting token sequences across views before decoding. This design preserves the strengths of single-view representations while enabling structured inter-view interactions. A central challenge in this setting is computational efficiency, as naive cross-attention over all tokens and views incurs quadratic complexity in both the token count and the number of views. To address this, we design a global-local fusion pipeline composed of two complementary modules. The **Multi-view Merging Attention (MMA)** module performs local cross-view fusion, reducing the cost from $O(N^2)$ to $O(N)$, where N is the number of tokens per view. It further incorporates a learnable view-importance matrix to adaptively weight different viewpoints, along with a token-wise gating mechanism to regulate the contribution of fused features. The **Object-Guided Attention (OGA)** module captures global object-level context by aggregating class tokens from all views into a unified object representation via a learned projection. This global token is then propagated back to patch tokens through cross-attention with temperature-scaled sigmoid gating, enabling each view to be conditioned on holistic object-level information.

In standard multi-view learning [10, 33, 37, 41], unconstrained cross-view information flow is typically considered beneficial. However, we find that reconstruction-based AD breaks this intuition. When multiple views are naively fused, normal cues from intact viewpoints can propagate to the decoder, enabling it to faithfully reconstruct even anomalous regions. This phenomenon, which we define as *cross-view information leakage*, effectively collapses the "reconstruction gap" upon which AD relies. As shown in in Figure 1, naïve fusion obscures the anomaly signal and collapses the normal-abnormal reconstruction gap, while our information restricted design restores it. Our analysis reveals a critical insight: effective multi-view fusion in this setting demands explicit information restriction rather than unconstrained sharing. This insight directly informs the design of the OGA module. Unlike conventional fusion [31] that adds features as a residual, OGA replaces original patch representations with globally conditioned outputs, thereby limiting the decoder's access to view-specific normal details that might otherwise facilitate "leaked" reconstructions. Furthermore, we employ temperature-scaled sigmoid gating instead of standard softmax normalization to avoid forcing competition among normal tokens, using a sharpened temperature to concentrate attention on a sparse set of relevant positions. These deliberate constraints ensure that the decoder remains conditioned on holistic object context rather than local anomalous cues, thereby preserving the reconstruction gap.

We evaluate GLAD on the MANTA [11] and Real-IAD [32] benchmarks, showing consistent improvements at the sample, image, and pixel levels over both single-view baselines and the state-of-the-art methods. Our main contributions are summarized as follows:

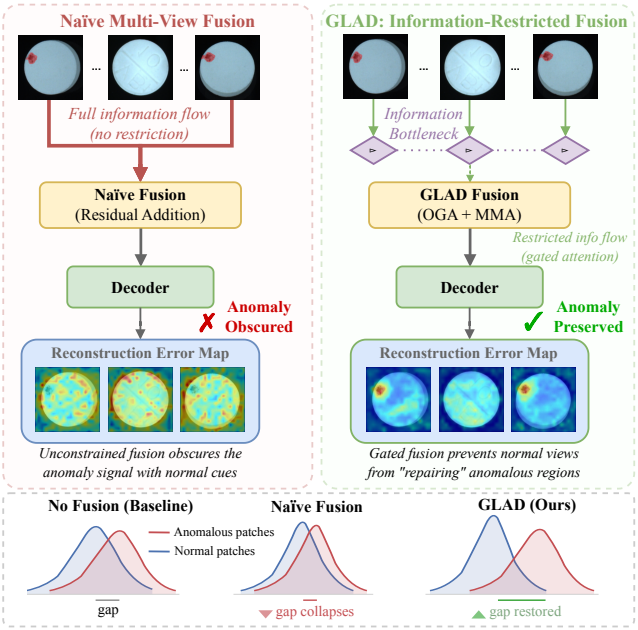


Figure 1: Cross-view information leakage in multi-view anomaly detection. Unconstrained fusion obscures anomalies by collapsing the normal–abnormal reconstruction gap. GLAD restores this gap through information-restricted fusion.

- We propose GLAD, a global-local multi-view anomaly detection framework built on vision foundation models that integrates object-level global context with token-level local correspondences to achieve state-of-the-art performance.
- We introduce the Multi-view Merging Attention (MMA) module, a linear-complexity cross-view mechanism that enables scalable local fusion through learnable view-importance weighting and adaptive token-wise gating.
- We design the Object-Guided Attention (OGA) module, which constructs a holistic object representation and employs an information-restricted broadcast mechanism to prevent the collapse of the reconstruction gap.
- We identify and analyze cross-view information leakage, a fundamental challenge in multi-view reconstruction, and demonstrate that strategic information restriction is essential for robust defect localization.

2 Related Work

Anomaly Detection (AD). Unsupervised AD methods identify defects using only normal training samples. Table 1 categorizes existing single-view AD methods into three groups according to their modeling scope and backbone choice: separate methods that train one model per class ($\mathcal{G}_0$), unified methods built on conventional backbones ($\mathcal{G}_1$), and unified methods that leverage vision foundation models ($\mathcal{G}_2$). Category $\mathcal{G}_2$ [14, 43] adopt frozen foundation models as encoders to obtain rich, generalizable features. ViTAD [43] performs plain ViT feature regression on DINO [27] features, and Dinomaly [14] reconstructs foundation-model features with a noisy bottleneck. These approaches achieve strong single-view results, yet they still treat each view independently and resort to voting ensembles for

Table 1: Characteristic comparison between GLAD (ours) and prior anomaly detection (AD) works ($\mathcal{G}_0$: [4, 7, 9, 18, 21, 24, 28, 42], $\mathcal{G}_1$: [15, 16, 20, 35, 40, 44], $\mathcal{G}_2$: [14, 43]).

AD Method Conditions	AD Categories					
	$\mathcal{G}_0$	$\mathcal{G}_1$	$\mathcal{G}_2$	IDIF [26]	MVAD [17]	GLAD
Single model for all classes	✗	✓	✓	✗	✓	✓
Multi-view input	✗	✗	✗	✓	✓	✓
Vision foundation model	✗	✗	✓	✗	✗	✓
Token-Level cross-view fusion	✗	✗	✗	✓	✓	✓
Object-Level cross-view fusion	✗	✗	✗	✗	✗	✓

sample-level decisions, discarding valuable cross-view correlations. Category $\mathcal{G}_1$ [15, 16, 20, 35, 40, 44] handle all object classes in a single model. DiffusionAD [16] formulates AD as a denoising diffusion process, ROADS [20] leverages reverse distillation with an online learning strategy, and UniNet [35] fuses multi-layer features within a unified transformer. MambaAD [15] and WaveMambaAD [44] explore state-space models as decoders for efficient long-range modeling, while retaining pre-trained CNN encoders for feature extraction. Although these methods reduce deployment overhead, they rely on conventional backbones whose limited representational capacity struggles with the diverse appearances that arise in multi-view settings. Category $\mathcal{G}_0$ [4, 7, 9, 18, 21, 24, 28, 42] dedicate one model to each object class, yielding competitive accuracy at the cost of scalability. Within this group, synthesis-based approaches [18, 21, 24] generate artificial defects for training but face a synthetic-to-real gap that intensifies when views vary in pose and illumination. Embedding-based approaches [7, 28] compare test features against a stored normal distribution; yet features extracted independently per view lack spatial correspondence across viewpoints. Reconstruction-based approaches [4, 9, 42] learn to reproduce normal patterns and flag deviations, but without cross-view constraints they can reconstruct subtle anomalies visible only from certain angles, causing missed detections. In contrast, GLAD builds on foundation-model features while introducing explicit multi-view fusion through linear-complexity local cross-view attention (MMA) and information-restricted global object-level aggregation (OGA), where cross-view information flow is carefully controlled to preserve the reconstruction gap that AD relies on.

Multi-view AD. Multi-view AD is widely adopted in industrial quality control, yet existing single-view methods discard cross-view consistency cues and therefore cannot exploit the complementary information captured from different viewpoints. MVAD [17] is among the first to jointly process multiple views via patch-level cross-attention with an adaptive view selection mechanism. However, its token-level interactions incur quadratic complexity with respect to the top-k related patches across views, which, although more efficient than full attention, still incurs considerable memory cost that limits scalability. Moreover, MVAD lacks a global, object-level representation, so sample-level decisions rely solely on aggregating local patch scores without holistic object context. IDIF [26] takes a different route by decoupling intra-view features and projecting them into a shared voxel space for geometric cross-view alignment. While this provides explicit spatial correspondence, the voxel-based design is inherently class-specific, restricting IDIF to the separate (per-class) setting and preventing it from scaling to unified multi-class deployment. Recent benchmarks such as Real-IAD [32] and MANTA [11] provide multi-view evaluation protocols that further expose these limitations, revealing a clear gap between single-view performance and the demands of holistic multi-view reasoning. As summarized in Table 1, GLAD is the first method that simultaneously satisfies all five desirable properties: it adopts a vision foundation model to get rich representation, operates in the unified setting, accepts multi-view input, and performs both local pixel-level fusion via MMA and global object-level fusion via OGA.

Global-to-Local Attention. Using a global or class-level token to refine local features via cross-attention has proven effective across vision tasks. CaiT [30] separates class-attention layers from patch self-attention, and CrossViT [3] performs multi-scale fusion via class token cross-attention. This idea is also used in dense prediction, where MCTformer [38] and CTI [39] use class tokens to generate class-specific attention maps for weakly supervised segmentation. In a broader sense, learned query tokens that cross-attend over input features have become a general design pattern, as seen in Perceiver [19] for modality-agnostic perception, DETR [2] for object detection, and Q-Former [22] for vision-language alignment. However, these methods operate on single images and do not address how to construct a meaningful global token that captures information shared across multiple views of the same object. Our OGA module fills this gap by aggregating CLS tokens from all views into a unified global object token and cross-attending it over each view's patch tokens with temperature-scaled sigmoid gating. Unlike standard cross-attention, OGA replaces rather than adding a residual to prevent cross-view information leakage. Beyond global fusion, GLAD further pairs OGA with a local token-level cross-view mechanism (MMA), enabling the model to capture both object-level consistency and fine-grained spatial correspondences across views.

3 Our Proposed Method – GLAD

3.1 Overview

Problem Setup. Given an object captured from V viewpoints, the input consists of V images $\{\mathbf{x}_v\}_{v=1}^{V}$ with $\mathbf{x}_v \in \mathbb{R}^{H\times W\times 3}$. The goal is to determine whether the object is anomalous at the sample level and, for each view, to produce a pixel-level anomaly map that localizes defective regions. Only normal images are available during training. As shown in Figure 2, GLAD follows the encoder-bottleneck-decoder reconstruction paradigm [14, 43], where a shared, frozen self-supervised Vision Transformer encodes each view independently and a lightweight Transformer decoder is trained to reconstruct the encoder features. For view v, the encoder produces a class token $\mathbf{z}_v^{\text{cls}} \in \mathbb{R}^D$ that summarizes global content and a set of patch tokens $\mathbf{Z}_v \in \mathbb{R}^{N\times D}$ that encodes local spatial information, where N is the number of patches and D is the feature dimension. We extract features from multiple intermediate encoder layers and average them to form a single token sequence per view. The anomaly map for each view is computed as the cosine distance between the encoder output and the decoder reconstruction at every spatial position, and the sample-level score aggregates evidence across all V views. Following [14], we compute a global cosine similarity loss L_{GC} between the encoder and decoder feature maps across all spatial patches. To focus optimization on informative regions, we apply a hard-mining strategy during backpropagation, which

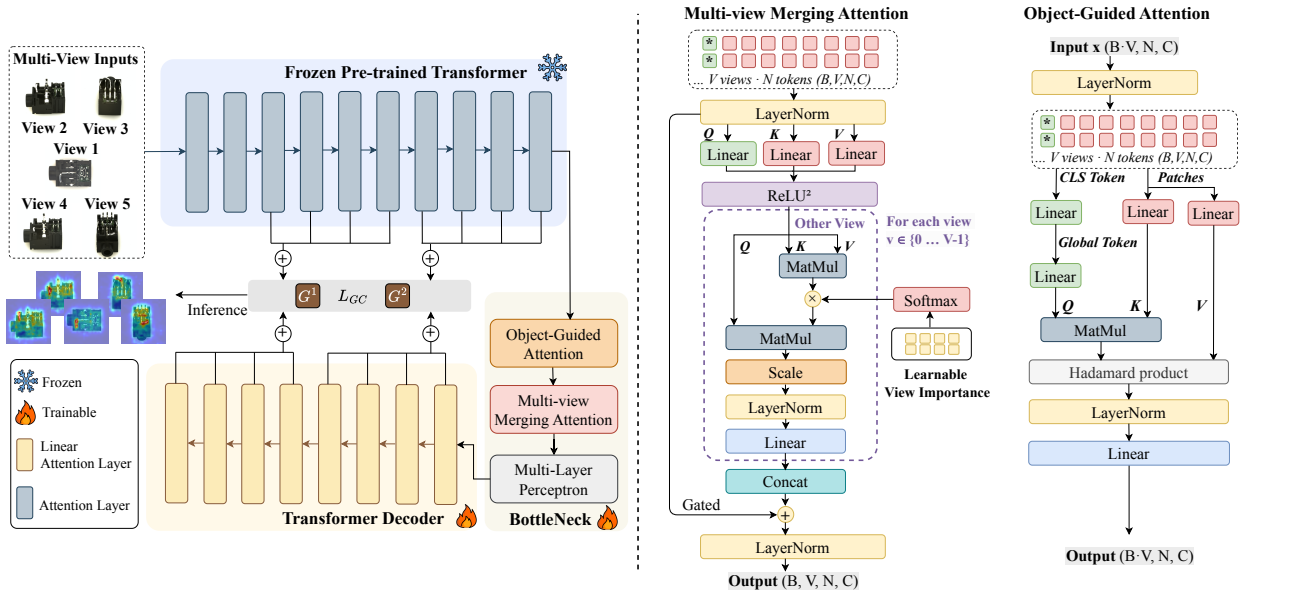


Figure 2: Overview of GLAD. Multi-view images are encoded by a frozen pre-trained Transformer. The resulting tokens pass through the Object-Guided Attention (OGA) module for global object-level fusion and the Multi-view Merging Attention (MMA) module for local token-level cross-view fusion, before entering the MLP bottleneck and Transformer decoder for reconstruction. Right panels detail the MMA (middle) and OGA (right) architectures.

patches are ranked by reconstruction error and partitioned into an easy group G^1 (top p% with lowest error) and a hard group G^2 (remaining $(1-p)$%). The gradients of G^1 are scaled by a near-zero factor, so that parameter updates are predominantly driven by the hard-to-reconstruct patches in G^2.

Multi-view Fusion. Existing single-view methods pass the encoder tokens directly through an MLP bottleneck and into the decoder, treating each view in isolation. GLAD instead inserts a multi-view fusion stage between the encoder and the bottleneck, enabling cross-view reasoning before reconstruction (Figure 2). This stage comprises two complementary modules. First, Object-Guided Attention (OGA, Sec. 3.2) aggregates the class tokens $\{\mathbf{z}_v^{\text{cls}}\}_{v=1}^{V}$ from all views into a single global object token via a learned projection, then cross-attends this token over patch tokens across all views and broadcasts the result back as a gated residual, injecting holistic object-level context into every view. Second, Multi-view Merging Attention (MMA, Sec. 3.3) performs token-level cross-view fusion with linear complexity, where each view attends to patch tokens from the remaining views using an ReLU^2 kernel, weighted by a learnable view importance matrix and modulated by a token-wise gate. By combining global (OGA) and local (MMA) fusion, GLAD enables the decoder to reconstruct features informed by both object-level consistency and fine-grained cross-view correspondences.

3.2 Object-Guided Attention (OGA)

A patch token knows what it sees, but not what it is part of. When multiple views capture the same object, each patch encodes a local surface fragment without awareness of the object's overall identity or how the other views perceive it. Before fusing token-level details across views, GLAD therefore first establishes a holistic understanding of the object. OGA fills this gap by constructing a single global object token from all views and using it to modulate every patch token through cross-attention, giving each local feature a sense of the whole object before reconstruction begins.

Global Object Token. To obtain a compact, object-level representation, OGA concatenates the class tokens from all V views and projects them through a learned linear layer:

$$\mathbf{g} = \mathbf{W}_g \left[\mathbf{z}_1^{\text{cls}}; \mathbf{z}_2^{\text{cls}}; \ldots; \mathbf{z}_V^{\text{cls}}\right], \tag{1}$$

where $\mathbf{W}_g \in \mathbb{R}^{D \times VD}$ is a learnable projection and $[\cdot;\cdot]$ denotes concatenation. Each class token $\mathbf{z}_v^{\text{cls}}$ encodes the global semantics of a single view; by fusing all of them, $\mathbf{g} \in \mathbb{R}^D$ captures an object-level summary that is aware of all viewpoints simultaneously.

Cross-Attention over Multi-view Patches. The global object token is used as a query to cross-attend over the patch tokens from all views. Specifically, all V sets of patch tokens are concatenated into a single sequence $\mathbf{P} = [\mathbf{Z}_1; \mathbf{Z}_2; \ldots; \mathbf{Z}_V] \in \mathbb{R}^{VN \times D}$, and multi-head projections are applied:

$$\mathbf{q} = \mathbf{g}\,\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{P}\,\mathbf{W}^K, \quad \mathbf{V} = \mathbf{P}\,\mathbf{W}^V, \tag{2}$$

where the query $\mathbf{q} \in \mathbb{R}^{N_h \times 1 \times d}$ and the keys and values $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{N_h \times VN \times d}$ are reshaped into N_h heads of dimension $d = D/N_h$. The attention weights are computed with a temperature-scaled sigmoid activation:

$$\mathbf{A} = \sigma\left(\frac{\mathbf{q}\,\mathbf{K}^\top}{\tau\sqrt{d}}\right), \tag{3}$$

where σ denotes the sigmoid function and τ is a temperature hyperparameter (set to 0.7). Each entry $A_i \in [0, 1]$ acts as an independent gate that controls how much information from the i-th patch flows into the output. The per-patch output is then obtained by element-wise gating rather than a weighted sum:

$$\hat{\mathbf{z}}_i = VN \cdot A_i \cdot \mathbf{v}_i, \tag{4}$$

where $\mathbf{v}_i$ is the value vector of the i-th patch and the scaling factor VN compensates the reduced magnitude due to sigmoid gating (unlike softmax, does not guarantee that the weights sum to one).

Information-Restricted Broadcast. In reconstruction-based AD, the model relies on a gap between reconstruction errors for normal and anomalous inputs. A central challenge in designing OGA is that the global token should amplify the difference between normal and abnormal regions rather than smooth everything toward a normal template. If the cross-attention output is naively added as a residual, the global token effectively produces a normal template, because the vast majority of training patches are normal. At test time, this template would push anomalous regions closer to the normal distribution, collapsing the reconstruction gap that the detector depends on. We refer to this failure mode as *cross-view information leakage* and analyze it empirically in Sec. 4. To prevent leakage, OGA operates in a *token replacement* mode, where the cross-attention output directly replaces the original patch tokens instead of being added as a residual. By discarding the original features entirely, the model avoids accumulating a normal template that would dilute the anomaly signal:

$$\hat{\mathbf{Z}}_v = \mathrm{LN}(\mathrm{OGA}(\mathbf{Z}_v)) \,, \tag{5}$$

where LN denotes the layer normalization [1]. The attention mechanism itself is also designed to sharpen the normal-abnormal distinction. OGA uses sigmoid rather than softmax in Eq. (3), because softmax normalizes across all patches, forcing normal patches to compete with each other for attention weight and reducing the mechanism to a normal-vs-normal comparison that is insensitive to anomalies. Sigmoid gates each patch independently, allowing abnormal patches to receive distinctly different activation levels without being diluted by the normal majority. A temperature $\tau = 0.7 < 1$ further sharpens the distribution, widening the gap between the activation levels assigned to normal and abnormal patches. Finally, the global object token $\mathbf{g}$ is *not* added back to the per-view class tokens after cross-attention, preserving the single-view nature of the class token for downstream anomaly scoring. These choices ensure that OGA injects global object context into each view's patch tokens in a way that amplifies the normal-abnormal distinction, rather than blurring it, while preserving the reconstruction bottleneck that is essential for AD.

3.3 Multi-view Merging Attention (MMA)

While OGA provides each patch token with global object-level context, it does not enable direct communication between individual patches across views. Consider a scratch on an object that is clearly visible in views 2 and 4 but occluded in views 1, 3, and 5. OGA can inform all patches that the object has a certain global profile, but it cannot propagate the fine-grained evidence of the scratch from the views where it is visible to the views where it is occluded. MMA addresses this by letting each view selectively query the other views at the token level, so that spatially corresponding or semantically related patches can exchange information before reconstruction. This cross-view exchange is what enables the model to produce view-consistent anomaly maps and more reliable sample-level decisions.

Concretely, for each view v, MMA computes queries from v's own tokens and aggregates key-value information from the remaining $V-1$ views, weighted by learned per-view importance scores. A straightforward implementation would concatenate all N patch tokens from all V views into a single sequence of length VN and apply standard softmax attention, but this yields $O(V^2N^2)$ complexity for the attention matrix alone, which quickly becomes prohibitive for high-resolution ViT features. MMA sidesteps this bottleneck by leveraging the associativity of linear attention. Instead of materializing the full $VN \times VN$ attention matrix, it decomposes the cross-view computation into compact per-view statistics that can be accumulated in $O(Vd^2)$ space (where $d{=}D/N_h$ is the per-head dimension defined below).

Linear Attention Kernel. All operations in MMA are performed in a multi-head fashion with N_h heads of dimension $d = D/N_h$. Given the token sequence $\mathbf{Z}_v$ of view v after layer normalization, we project it into queries, keys, and values through separate linear layers and reshape the result into N_h by

$$\mathbf{Q}_v \leftarrow \frac{\phi(\mathbf{Z}_v\mathbf{W}^Q)}{\sqrt{d}}, \quad \mathbf{K}_v \leftarrow \frac{\phi(\mathbf{Z}_v\mathbf{W}^K)}{\sqrt{d}}, \quad \mathbf{V}_v \leftarrow \mathbf{Z}_v\mathbf{W}^V, \tag{6}$$

where $\phi(\cdot) = \mathrm{ReLU}(\cdot)^2$ is a non-negative kernel map [47] applied to the queries and keys. This square ReLU kernel ensures that all attention weights remain non-negative, analogous to the softmax function, while being much cheaper to evaluate. The parameters in (6), *i.e.* , $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{D\times D}$, are learnable projection matrices. Compared to alternative non-negative kernels such as plain ReLU or ELU [5], the squared form provides stronger expressiveness by amplifying large activations and suppressing near-zero ones.

View-Weighted Aggregation. In multi-view imagery, a given surface patch is typically visible in only a subset of views, so not all views contribute equally to each query view. To consider this, MMA learns a view importance matrix $\mathbf{W}_{\text{view}} \in \mathbb{R}^{V\times V}$ whose rows are normalized via softmax (excluding the diagonal, *i.e.* , self-attention):

$$w_{vj} = \mathrm{softmax}_{j\neq v}(\mathbf{W}_{\text{view}}[v,:]) \,, \tag{7}$$

where w_{vj} indicates the learned importance of view j to view v. Rather than concatenating all other views' tokens into a single large attention matrix, MMA accumulates compact key-value statistics from each view independently. For each view v, the cross-view output is computed as:

$$\mathrm{MMA}(\mathbf{Q}_v) = \frac{\mathbf{Q}_v \sum_{j\neq v} w_{vj}\, \mathbf{K}_j^\top \mathbf{V}_j}{\mathbf{Q}_v \sum_{j\neq v} w_{vj}\, \mathbf{K}_j^\top \mathbf{1} + \epsilon} \,, \tag{8}$$

where $\mathbf{K}_j^\top \mathbf{V}_j \in \mathbb{R}^{d\times d}$ is the outer-product summary of view j and $\mathbf{K}_j^\top \mathbf{1} \in \mathbb{R}^d$ is the normalization term. The key insight is that these per-view statistics can be computed and accumulated *before* multiplying with $\mathbf{Q}_v$, reducing the per-view complexity from $O(N^2d)$ in standard attention to $O(Nd^2)$ (linear in the number of tokens N). The attention output is then layer-normalized and projected back to D dimensions: $\mathrm{LN}(\mathrm{MMA}(\mathbf{Q}_v))\,\mathbf{W}^O$, where $\mathbf{W}^O \in \mathbb{R}^{D\times D}$.

Token-wise Gating. Not every token benefits equally from cross-view fusion. For instance, a patch depicting a featureless background carries little useful information for other views, while a patch near a defect boundary may gain significantly from complementary viewpoints. A uniform residual connection would blend cross-view signals indiscriminately, potentially introducing noise

Table 2: Average performance comparison on Real-IAD dataset. The best and second-best results are marked in bold and underlined. All methods except MVAD are reproduced under the multi-view setting.

Method	Venue	S-AUROC	S-AP	S-F1	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	P-AUPRO
RD4AD [9]	CVPR'22	85.7	92.4	87.7	83.0	79.6	74.4	97.3	25.8	33.5	90.4
UniAD [40]	NeurIPS'22	87.3	93.1	88.9	83.0	83.4	74.6	97.4	23.7	31.4	87.3
RD++ [29]	CVPR'23	87.0	93.3	88.2	83.7	80.9	74.9	97.7	25.8	33.5	90.0
SimpleNet [24]	CVPR'23	61.2	77.9	82.1	82.6	79.2	74.1	76.3	3.4	6.9	42.5
DeSTSeg [45]	CVPR'23	89.0	94.6	89.2	82.3	79.2	73.2	94.6	37.9	41.7	40.6
RealNet [46]	CVPR'24	72.0	87.1	82.5	65.2	65.4	62.7	60.5	27.2	24.9	28.7
ViTAD [43]	CVIU'25	88.8	94.3	89.3	82.8	80.0	73.7	97.1	23.9	31.8	84.3
MVAD [17]	TMM'25	90.2	95.3	90.1	86.6	84.8	77.2	97.9	30.3	36.8	91.2
Dinomaly [14]	CVPR'25	92.3	96.2	**91.7**	89.4	86.8	**80.3**	**98.9**	43.1	47.3	94.0
GLAD (ours)	-	**92.4**	**96.3**	91.4	**89.7**	**87.2**	**80.3**	**98.9**	**44.7**	**48.2**	**94.9**

Table 3: Average performance comparison on MANTA dataset. The best and second-best results are marked in bold and underlined. All methods except MVAD are reproduced under the multi-view setting.

Method	Venue	S-AUROC	S-AP	S-F1	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	P-AUPRO
RD4AD [9]	CVPR'22	78.3	63.7	68.3	80.5	59.5	62.4	91.8	29.1	34.7	78.8
UniAD [40]	NeurIPS'22	74.9	59.7	65.2	82.7	64.3	65.1	90.7	26.1	32.4	79.6
DeSTSeg [45]	CVPR'23	62.4	49.3	56.5	62.3	37.8	45.5	54.8	12.0	9.0	23.4
RealNet [46]	CVPR'24	55.1	40.9	51.7	57.3	31.8	41.3	50.0	10.3	4.4	15.1
ViTAD [43]	CVIU'25	87.3	80.5	75.9	87.7	75.3	71.0	94.7	39.4	43.1	80.2
MVAD [17]	TMM'25	85.0	76.5	73.8	85.2	69.9	67.9	93.7	32.7	38.5	80.7
Dinomaly [14]	CVPR'25	92.3	87.9	83.4	91.3	82.3	77.7	94.8	44.8	48.3	84.2
GLAD (ours)	-	**92.7**	**88.5**	**83.6**	**92.0**	**83.5**	**78.1**	**95.2**	**47.6**	**49.6**	**85.9**

from uninformative views. To address this, we employ a token-wise gate that lets each token independently control its fusion ratio:

$$\hat{\mathbf{Z}}_v = (1 - \mathbf{G}_v) \odot \mathbf{Z}_v + \mathbf{G}_v \odot \mathrm{MMA}(\mathbf{Q}_v), \tag{9}$$

where $\odot$ is element-wise multiplication, $\mathbf{G}_v = \sigma(\mathrm{MLP}(\mathrm{MMA}(\mathbf{Q}_v))) \in [0,1]^{N\times D}$ and σ denotes the sigmoid function. The gate is predicted from the fused features themselves, allowing the network to learn which tokens contain informative cross-view evidence and should be integrated, and which should retain their original single-view representation. A final layer normalization is applied to $\hat{\mathbf{Z}}_v$ before passing it to the subsequent stage.

In summary, MMA enriches each view's token representation with fine-grained evidence from other views through three cooperating mechanisms: a linear attention kernel for efficient cross-view matching, view-level importance weighting for selective aggregation, and token-wise gating for adaptive fusion. Because the decoder is trained to reconstruct *normal* features, tokens that have been enriched with cross-view context produce a larger reconstruction gap when an anomaly is present across views, strengthening both localization and detection.

4 Experiments

4.1 Implementation Details

We adopt a frozen DINOv2 ViT-Base/16 [27] as the encoder backbone and extract intermediate features from the layers 2 through 9, which are averaged into a single token sequence per view. The decoder consists of 8 Transformer blocks with 12 attention heads and an MLP ratio of 4. A single-layer MLP bottleneck with a dropout rate of 0.4 connects the encoder features to the decoder. For multi-view fusion, MMA employs linear attention with a ReLU^2 kernel and token-wise gating, while OGA uses cosine-similarity attention with a temperature of 0.7. We train with StableAdamW [36] using a learning rate of 2×10^{-3}, weight decay of 10^{-4}, and a warm-cosine schedule that decays the learning rate to 2×10^{-4} over 50K iterations after 100 warm-up steps. The images are resized to 448×448 (crop 392×392) for Real-IAD and 280×280 (crop 266×266) for MANTA-Tiny. The batch size is 4 for both datasets.

4.2 Datasets

We evaluate GLAD on two multi-view AD benchmarks that differ in object scale and domain coverage. **Real-IAD** [32] is a large-scale multi-view industrial AD benchmark covering 30 object classes with materials ranging from metal and plastic to wood and ceramics. It provides approximately 150K high-resolution images (99,721 normal, 51,329 abnormal) captured from five fixed viewpoints per object. Training uses only normal samples, while testing includes both normal and abnormal ones. Eight defect categories are annotated (pit, deformation, scratch, *etc.*), and defect areas span 0.1% to 6.75% of the image, posing a significant localization challenge. **MANTA** [11] targets tiny-object AD across 38 classes in five domains (mechanics, medicine, electronics, agriculture, and food), offering over 137.3K object instances with 8.6K abnormal samples annotated at the pixel level. We use the official MANTA-Tiny split, which contains 800 instances per class (4,000 images per class over 5 views, 152,000 images in total). The full MANTA dataset contains images of varying sizes that cannot be evenly partitioned into five fixed views, whereas MANTA-Tiny provides uniformly sized 256×256 images suitable for multi-view evaluation. Because GLAD relies solely on visual information, we discard the textual annotations included in the original MANTA release.

Table 4: Ablation study on the contribution of each proposed module. MMA primarily improves sample-level detection, while OGA further boosts pixel-level localization. The best results per dataset are marked in bold.

Dataset	MMA	OGA	S-AUROC	S-AP	S-F1	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	P-AUPRO
Real-IAD	✗	✗	90.3	96.2	**91.7**	89.4	86.8	80.3	**98.9**	43.1	47.3	94.0
	✓	✗	**92.5**	**96.3**	91.6	**89.8**	**87.2**	**80.8**	**98.9**	42.7	47.0	94.4
	✓	✓	92.4	**96.3**	91.4	89.7	**87.2**	80.3	**98.9**	**44.7**	**48.2**	**94.9**
MANTA	✗	✗	92.3	87.9	83.4	91.3	82.3	77.7	94.8	44.8	48.3	84.2
	✓	✗	**92.8**	88.3	83.4	91.8	82.8	77.9	94.9	44.6	47.6	85.1
	✓	✓	92.7	**88.5**	**83.6**	**92.0**	**83.5**	**78.1**	**95.2**	**47.6**	**49.6**	**85.9**

Table 5: Efficiency comparison of cross-view fusion alternatives in the MMA module.

Approach	Params (M)	Fwd (MB)	Fwd+Bwd (MB)	Time (ms)
Deformable Attention	**2.37**	2208.1	3091.1	17.38
MVAS Block [17]	3.86	1318.1	9276.0	47.51
Linear Attention (ours)	2.66	**528.5**	**1312.9**	**12.04**

Table 6: Accuracy comparison of cross-view fusion alternatives in the MMA module.

Approach	S-AUROC	I-AUROC	P-AUROC	P-AP	P-F1	P-AUPRO
Deformable Attention	92.4	89.7	**98.9**	**43.3**	46.7	**94.4**
Linear Attention (ours)	**92.5**	**89.8**	**98.9**	42.7	**47.0**	**94.4**

Table 7: Ablation study on OGA information-restriction mechanisms. All variants include MMA and OGA. TR.: token replacement; SG.: sigmoid gating; NoG.: no global token feedback to per-view CLS tokens; Temp.: sharpened temperature. The best results are marked in bold.

TR.	SG.	NoG.	Temp.	S-AUROC	I-AUROC	P-AUROC	P-AP	P-F1	P-AUPRO
✗	✗	✗	✗	92.2	91.3	94.7	43.7	46.9	84.0
✓	✗	✗	✗	92.4	91.7	95.0	47.0	49.4	85.4
✓	✓	✗	✗	92.3	91.8	95.1	46.8	49.3	85.4
✓	✗	✓	✗	92.4	91.7	95.1	47.7	50.0	86.0
✓	✓	✓	✗	92.6	91.9	**95.3**	47.7	49.8	85.7
✓	✓	✗	✓	92.4	91.8	95.2	**47.8**	**50.1**	**85.9**
✓	✓	✓	✓	**92.7**	**92.0**	95.2	47.6	49.6	**85.9**

4.3 Metrics

We adopt the evaluation protocol of [17] and report metrics at three granularities. *Sample-level* metrics (S-AUROC, S-AP, S-F1) judge whether an object instance is anomalous by combining evidence from all its views. *Image-level* metrics (I-AUROC, I-AP, I-F1) measure per-image detection accuracy independently of multi-view aggregation. *Pixel-level* metrics (P-AUROC, P-AP, P-F1, P-AUPRO) quantify how precisely the model localizes defective regions, where P-AUPRO (Per-Region-Overlap) penalizes large false-positive areas. In ablation studies, we report four key indicators (S-AUROC, I-AUROC, P-AUROC, P-AUPRO) for brevity.

4.4 Comparison with the State-of-the-Art

We compare GLAD with a range of AD methods, including unified models [40], reconstruction-based approaches [9, 14, 29], feature-embedding methods [24, 46], segmentation-based methods [45], and the multi-view method MVAD [17]. Since most existing methods are designed for single-view evaluation, we reproduce them under the multi-view setting for a fair comparison. MVAD is evaluated directly as a multi-view method.

Results on Real-IAD. Table 2 summarizes the results on Real-IAD. GLAD performs the best on 9 out of 10 individual metrics, producing sample-level scores of 92.4/96.3/91.4, image-level scores of 89.7/87.2/80.3, and pixel-level scores of 98.9/44.7/48.2/94.9. Compared with Dinomaly, the strongest single-view baseline, the most notable gains appear at the pixel level, where MMA and OGA provide spatial cues from complementary viewpoints that help the model produce tighter anomaly maps: P-AP improves by +1.6 (43.1→44.7) and P-F1 by +0.9 (47.3→48.2), indicating sharper anomaly localization, while P-AUPRO rises by +0.9 (94.0→94.9), confirming that region-level coverage improves as well. Compared with MVAD, the only other multi-view method, GLAD achieves substantially stronger results across all metrics. The localization gap is particularly striking: +14.4 on P-AP and +11.4 on P-F1, demonstrating that our global-local fusion design produces more precise anomaly maps than its cross-view attention mechanism.

Results on MANTA-Tiny. Table 3 reports results on MANTA-Tiny. GLAD performs the best on all 10 individual metrics with sample-level scores of 92.7/88.5/83.6, image-level scores of 92.0/83.5/78.1, and pixel-level scores of 95.2/47.6/49.6/85.9. The pixel-level gains over Dinomaly are especially noteworthy: P-AP improves by +2.8 (44.8→47.6) and P-F1 by +1.3 (48.3→49.6), representing the largest improvements among all metric groups. Cross-view fusion is particularly beneficial for tiny-object localization, where single-view features may be ambiguous due to the limited spatial extent of defects; complementary viewpoints supply the additional spatial context needed to resolve such ambiguities. P-AUPRO also rises by +1.7 (84.2→85.9), confirming consistent region-level improvement. Overall, GLAD consistently outperforms all compared methods on both datasets, validating the benefit of incorporating multi-view information through the proposed MMA and OGA modules.

Qualitative Results. Figure 3 shows representative anomaly maps from both datasets. Consistent with the quantitative gains, GLAD hass noticeably fewer false positives and cleaner anomaly maps that focus activation on real defective regions. On Real-IAD *Button Battery* class, the defect occupies only a small region in one view, yet GLAD localizes it with a focused hot spot, while MVAD and Dinomaly spread activations across the entire surface. For *PCB*, whose complex on-board components easily mislead other methods into diffuse responses, GLAD suppresses background noise and assigns distinctly higher scores at ground-truth locations, reflecting the cross-view stability from OGA and MMA. On MANTA-Tiny *Gear* and *Flat Nut* classes, GLAD yields the most saturated predictions aligned with the ground truth, with fewer missed detections and fewer false alarms, even without observing all defect types during training.

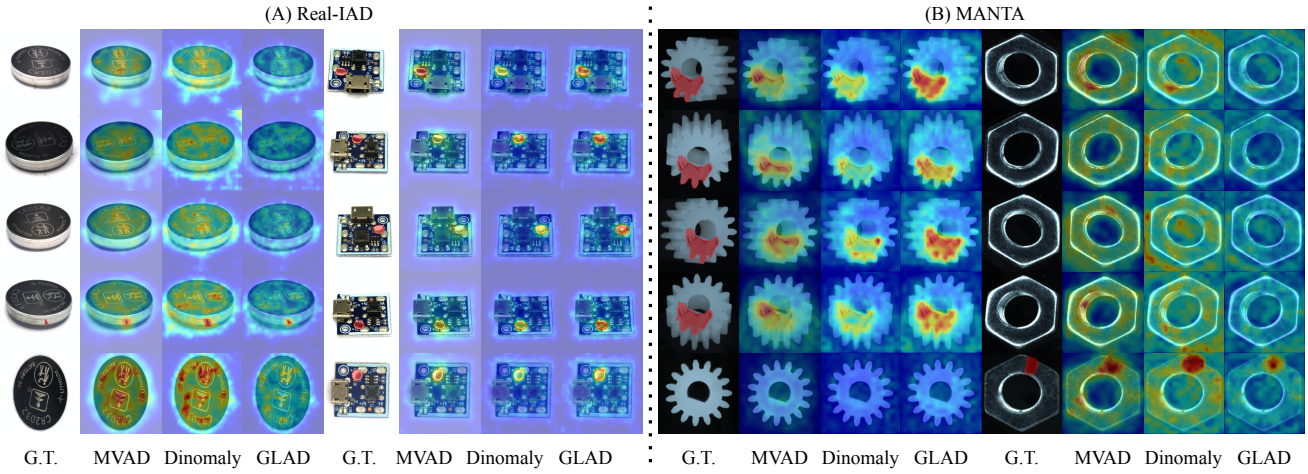


Figure 3: Qualitative comparison of anomaly maps on (A) Real-IAD and (B) MANTA. We show the ground-truth (G.T.) annotation and the anomaly maps produced by MVAD, Dinomaly, and GLAD (ours). Warmer colors indicate higher anomaly scores.

4.5 Ablation Study

Component Analysis. Table 4 reports the contribution of each proposed module. MMA and OGA address different aspects of the multi-view AD problem. Adding MMA alone primarily benefits sample-level detection, with S-AUROC improving by +2.2 on Real-IAD, as cross-view token fusion provides direct evidence for object-level decisions. However, MMA alone slightly lowers P-AP. Because MMA operates at the patch token level, when the anomalous region occupies only a small portion of the image, the fused cross-view features can dilute the localized anomaly signal. OGA tackles this from a different angle. By aggregating class tokens into a global object representation, it widens the gap between normal and abnormal feature distributions, which in turn sharpens pixel-level localization. Adding OGA yields substantial P-AP gains: on Real-IAD, P-AP rises from 42.7 to 44.7 (+2.0) and P-AUPRO from 94.4 to 94.9 (+0.5); on MANTA-Tiny, P-AP jumps from 44.6 to 47.6 (+3.0) and P-AUPRO from 85.1 to 85.9 (+0.8). Overall, the full model improves P-AUPRO by +0.9 on Real-IAD and +1.7 on MANTA-Tiny over the baseline without any multi-view module.

Cross-View Fusion Mechanism. We compare three alternatives for the cross-view attention in MMA: our linear attention variant, deformable attention, and the MVAS block from MVAD [17]. Table 5 shows that linear attention is the parameter-efficient (2.66M) and the fastest (12.04 ms per forward pass). MVAS requires 3.86M parameters, 47.51 ms, and 9,276 MB during training (over 7× the memory of linear attention, 1,312.9 MB); deformable attention is more compact (2.37M) but still costs 3,091.1 MB. Table 6 shows that the two mechanisms achieve comparable pixel-level detection of 98.9/42.7/47.0/94.4 and 98.9/43.3/46.7/94.4, respectively. Given the large efficiency advantage at nearly identical accuracy, linear attention is the preferred choice for MMA.

Information Restriction in OGA. In standard multi-view learning, richer cross-view information exchange is generally beneficial. Multi-view AD, however, presents a counter-intuitive challenge that we term as *cross-view information leakage*. Reconstruction-based methods rely on the gap between normal and anomalous reconstruction errors. When the global attention module passes too many normal cues across views, the decoder can faithfully reconstruct even anomalous regions, collapsing this gap. As shown in Table 7, we introduce four complementary restrictions into OGA. The most impactful is *token replacement*, which directly replaces the original patch tokens with the cross-attention output instead of adding a residual. All subsequent variants build on token replacement. Second, we replace softmax with *sigmoid* gating, which permits sparse, selective activation rather than forcing a normalized distribution that spreads information uniformly across views. Third, the *NoGlobal* strategy withholds the aggregated global object token from per-view CLS tokens, preventing multi-view information from leaking into the per-view classification signal. Fourth, a sharpened *temperature* (τ=0.7) concentrates the attention distribution to further limit diffuse information spread. Table 7 shows that these restrictions yield consistent improvements on MANTA-Tiny. Notably, sample-level metrics remain stable or even improve alongside the pixel-level gains, confirming that the restrictions successfully prevent information leakage without sacrificing detection capability. This finding suggests that information restriction is a fundamental design principle for multi-view AD, distinct from the "more information is better" paradigm that prevails in general multi-view representation learning.

5 Conclusion

We presented GLAD, a global-local multi-view anomaly detection framework that integrates cross-view learning through two modules. Through this work, we identified cross-view information leakage as a fundamental challenge in reconstruction-based multi-view anomaly detection, where naive fusion collapses the reconstruction gap by passing normal cues across views. OGA addresses this through deliberate information restriction, preserving the reconstruction gap essential for detection, while MMA enables efficient token-level cross-view fusion with linear complexity. Our experiments on Real-IAD and MANTA confirm that GLAD outperforms state-of-the-art methods. We believe that the principle of information restriction offers a useful perspective for future work on multi-view reasoning in reconstruction-based settings beyond anomaly detection.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision (ECCV)*.

[3] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. 2021. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[4] Zining Chen, Xingshuang Luo, Weiqiu Wang, Zhicheng Zhao, Fei Su, and Aidong Men. 2025. Filter or compensate: Towards invariant representation from distribution shift for anomaly detection. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[5] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations (ICLR)*.

[6] Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. 2025. AnomalyDINO: Boosting Patch-based Few-shot Anomaly Detection with DINOv2. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

[7] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. 2021. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition (ICPR)*.

[8] Chenghao Deng, Haote Xu, Xiaolu Chen, Haodi Xu, Xiaotong Tu, Xinghao Ding, and Yue Huang. 2024. Simclip: Refining image-text alignment with simple prompts for zero-/few-shot anomaly detection. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

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

[10] Zhibin Dong, Meng Liu, Siwei Wang, Ke Liang, Yi Zhang, Suyuan Liu, Jiaqi Jin, Xinwang Liu, and En Zhu. 2025. Enhanced then Progressive Fusion with View Graph for Multi-View Clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[11] Lei Fan, Dongdong Fan, Zhiguang Hu, Yiwen Ding, Donglin Di, Kai Yi, Maurice Pagnucco, and Yang Song. 2025. MANTA: A Large-Scale Multi-View and Visual-Text Anomaly Detection Dataset for Tiny Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[12] Qingqing Fang, Wenxi Lv, and Qinliang Su. 2025. Af-clip: Zero-shot anomaly detection via anomaly-focused clip adaptation. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

[13] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. 2025. Univad: A training-free unified model for few-shot visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

[15] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. 2024. MambaAD: Exploring state space models for multi-class unsupervised anomaly detection. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[16] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. 2024. A diffusion-based framework for multi-class anomaly detection. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[17] Haoyang He, Jiangning Zhang, Guanzhong Tian, Chengjie Wang, and Lei Xie. 2025. Learning Multi-view Anomaly Detection with Efficient Adaptive Selection. *IEEE Transactions on Multimedia (TMM)* (2025).

[18] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. 2024. AnomalyDiffusion: Few-shot anomaly image generation with diffusion model. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[19] Andrew Jaegle, Felix Gimeno, Andrew Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General Perception with Iterative Attention. In *International Conference on Machine Learning (ICML)*.

[20] Hossein Kashiani, Niloufar Alipour Talemi, and Fatemeh Afghah. 2025. ROADS: Robust prompt-driven multi-class anomaly detection under domain shift. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

[21] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. 2021. CutPaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning (ICML)*.

[23] Xinyue Liu, Jianyuan Wang, Biao Leng, and Shuo Zhang. 2025. Unlocking the potential of reverse distillation for anomaly detection. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[24] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. 2023. SimpleNet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[25] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S. Kevin Zhou. 2025. AA-CLIP: Enhancing Zero-shot Anomaly Detection via Anomaly-Aware CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] Kai Mao, Yiyang Lian, Yangyang Wang, Meiqin Liu, Nanning Zheng, and Ping Wei. 2025. Unveiling Multi-View Anomaly Detection: Intra-view Decoupling and Inter-view Fusion. In *AAAI Conference on Artificial Intelligence (AAAI)*.

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

[28] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[29] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. 2023. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[30] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. 2021. Going Deeper with Image Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).

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

[33] Ren Wang, Haoliang Sun, Yuxiu Lin, Chuanhui Zuo, Yongshun Gong, Yilong Yin, and Wenjia Meng. 2025. SeqMvRL: A Sequential Fusion Framework for Multi-view Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[34] Xiaodong Wang, Hongmin Hu, Fei Yan, Junwen Lu, Zhiqiang Zeng, Weidong Hong, and Zhedong Zheng. 2025. Uniad: Integrating geometric and semantic cues for unified anomaly detection. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

[35] Shun Wei, Jielin Jiang, and Xiaolong Xu. 2025. UniNet: A Contrastive Learning-guided Unified Framework with Feature Selection for Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[36] Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. 2023. Stable and low-precision training for large-scale vision-language models. *Advances in Neural Information Processing Systems* 36 (2023), 10271–10298.

[37] Jie Xu, Na Zhao, Gang Niu, Masashi Sugiyama, and Xiaofeng Zhu. 2025. Robust Multi-View Learning via Representation Fusion of Sample-Level Attention and Alignment of Simulated Perturbation. *arXiv preprint arXiv:2503.04151* (2025).

[38] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. 2022. Multi-class Token Transformer for Weakly Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[39] Sung-Hoon Yoon, Hyein Kim, Keon-Hee Jeong, Suhyeon Park, and Nojun Kwak. 2024. Class Tokens Infusion for Weakly Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[40] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. 2022. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[41] Zhiwen Yu, Ziyang Dong, Chenchen Yu, Kaixiang Yang, Ziwei Fan, and C. L. Philip Chen. 2024. A Review on Multi-View Learning. *Frontiers of Computer Science* 19, 7 (2024).

[42] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. 2021. DRAEM-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[43] Jiangning Zhang, Xuhai Chen, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, Ming-Hsuan Yang, and Dacheng Tao. 2025. Exploring plain ViT features for multi-class unsupervised visual anomaly detection. *Computer Vision and Image Understanding (CVIU)* 253 (2025), 104308.

[44] Qiao Zhang, Mingwen Shao, Xinyuan Chen, Xiang Lv, and Kai Xu. 2025. Wave-MambaAD: Wavelet-driven State Space Model for Multi-class Unsupervised

Anomaly Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20868–20877.

[45] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. 2023. DeSTSeg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[46] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. 2024. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[47] Zhengyan Zhang, Yixin Song, Guanghui Yu, Xu Han, Yankai Lin, Chaojun Xiao, Chenyang Song, Zhiyuan Liu, Zeyu Mi, and Maosong Sun. 2024. ReLU^2 Wins: Discovering Efficient Activation Functions for Sparse LLMs. *arXiv preprint arXiv:2402.03804* (2024).