EXPLORING MODALITY COMPETITION IN UNSUPERVISED MULTIMODAL ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised Multimodal Anomaly Detection (MAD) identifies anomalies by modelling normality from normal multimodal data and leveraging cross-modal complementarity. However, modality competition remain an issue during training: one modality continually dominates the other. This directly hinders normality modelling on the suppressed modality, making it hard to distinguish anomalies on that modality and leading to inaccurate MAD. Worse still, when anomalies in the dominant modality remain undetected due to their similarity to normal features or their small area, the accuracy of MAD further degrades. We approach this problem from a Mixture-of-Experts (MoE) perspective. To better learn multimodal normal features, we propose Hindsight Learning (HL), a novel MoE training paradigm that constructs a post-hoc routing teacher from observable per-expert modal normality reconstruction advantages using Optimal Transport (OT) with expert capacity constraints. The teacher explicitly trains the router to learn which modality token is sent to a particular expert. We then introduce a Debt-aware MoE (DEMO) based on HL to mitigate modality competition by preventing starvation in per-modality normality reconstruction utility. Specifically, each modality maintains a virtual debt queue, where debt accrues when its normality reconstruction utility falls below a minimum service objective in the queue. To reduce debt, the router is trained via debt-aware OT to preferentially serve indebted modality. We theoretically prove a long-run modality-service guarantee. Empirical results show DEMO surpasses state-of-the-art methods by 2.4% AUROC on MAD benchmarks and remains effective in continual learning.

1 INTRODUCTION

Background. Multimodal Anomaly Detection (MAD) exploits RGB and 3D modalities (e.g., depth and normal maps) to recognize anomalous regions in objects, with wide applications across different fields (Rudolph et al., 2023; Xu et al., 2026). Due to the rarity of anomalous samples, this direction generally trains models in an unsupervised manner. In recent years, to enhance practical efficiency (You et al., 2022), researchers have moved beyond the conventional *one-model-one-class* paradigm (Costanzino et al., 2024; Long et al., 2025) toward unified MAD models (Gu et al., 2026) capable of detecting anomalies across multiple classes simultaneously.

Challenges. Existing MAD methods (Wang et al., 2023; Costanzino et al., 2024) align RGB with other modalities to leverage complementary information across modalities during training, leading to promising detection performance. Despite this success, these works overlook a potential challenge: multimodal optimization can become imbalanced, allowing a dominant modality to continually suppress the learning of another (Peng et al., 2022). We refer to this phenomenon as *modality competition*. Figure 1(a) shows its impacts on AST (Rudolph et al., 2023) and CFM (Costanzino et al., 2024). Before intervention, the information acquisition of the depth modality is low. This directly hinders normality modelling on the depth modality, leaving anomaly discrimination on that modality inaccurate and leading to suboptimal MAD performance. Worse still, when the dominant modality RGB fails to detect anomalies due to the similarity between anomalies and normal features (e.g., the holes and chocolates on the *cookies*), the accuracy of MAD further declines. Besides, we also found that this issue still persists empirically during Continual Learning (CL), as shown in Figure 1(b). Thus, even when forgetting is mitigated, modality competition alone can degrade MAD

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.

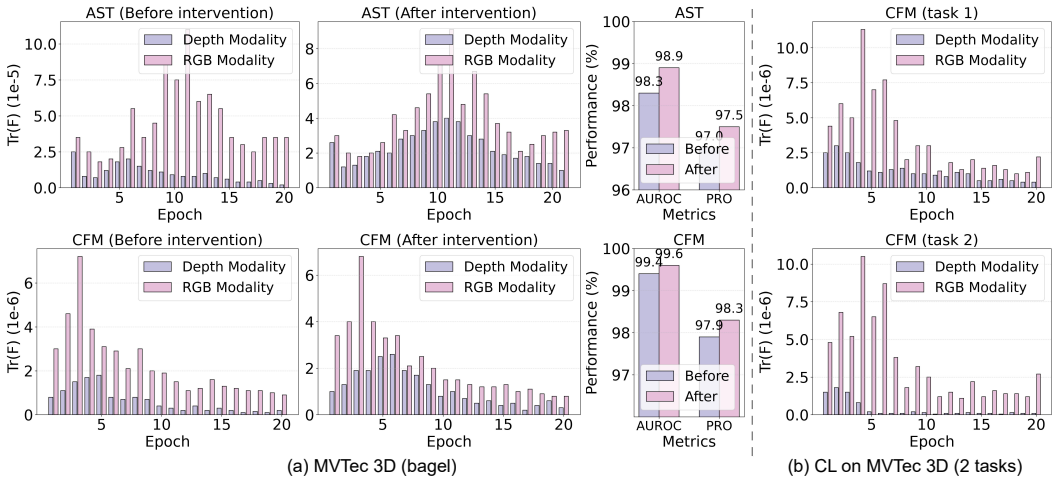


Figure 1: The impact of modality competition on information acquisition and performance. We use trace of Fisher information matrix Tr(F) (Huang et al., 2025) to quantify information acquisition.

performance. These observations motivate the central question of this paper: *how can modality competition be effectively mitigated during multimodal learning for robust MAD performance?*

Key insight. We tackle this problem from a Mixture-of-Experts (MoE) perspective. Following early works (Shazeer et al., 2017), existing MoE-based MAD methods (Gu et al., 2026) implicitly train the router to use the predicted weights for expert assignment via task cost (e.g., reconstruction loss), which we call foresight learning, as shown in Figure 2(a). However, the router in fact fails to tell whether this prediction is reasonable, and thus unsuitable experts may be used. This affects normality modelling (see Figure 2(a)). We propose a novel MoE training paradigm termed hindsight learning. As shown in Figure 2(b), it constructs a target routing teacher from the post-hoc, observable modal normality reconstruction advantages of each expert relative to anchor expert via Optimal Transport (OT), and uses the teacher to explicitly guide the router. This allows the router to clearly learn why this modality token is assigned to a particular expert, enhancing expert specialization in modality-wise normality modelling.

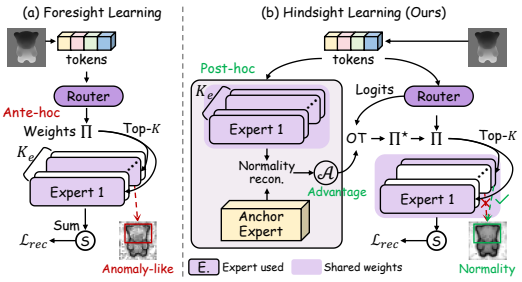


Figure 2: MoE training paradigm comparison.

Building upon this paradigm, we then propose a Debt-aware MoE (DEMO) for modality competition intervention. Its main components include a shared anchor expert, multiple residual experts, a debt-aware router, and one modality-debt queue per modality. During training, DEMO computes the post-hoc normality reconstruction gain of each expert relative to anchor expert and normalizes the realized gain as utility within each modality. This utility is interpreted as virtual service in the debt queue. When a serviceable modality repeatedly receives utility below its prescribed minimum objective, its debt accumulates. To alleviate debt, the router is trained to prioritize assigning useful expert capacity to indebted modality using debt-aware OT with expert capacity constrains. Notably, existing MoE-based MAD methods (Gu et al., 2026) regulate expert usage via load-balancing objectives (Shazeer et al., 2017), but they do not impose a lower bound on effective modality service, potentially leading to modality competition. Conversely, DEMO prioritizes modality-wise minimum service over equal expert traffic. Because we theoretically prove that expert load balancing alone cannot guarantee non-starved modality service. Further, we demonstrate that no serviceable modality is starved in long-run expectation.

Before delving into the details, we highlight the contributions as follows:

- **Novel paradigm.** We propose hindsight learning, a new MoE training paradigm that constructs a post-hoc routing teacher from observable per-expert normality reconstruction advantages for explicit supervision for the router.

- **Simple method.** We introduce DEMO, a method that alleviates modality competition by using debt-aware OT to train a router that preferentially serves indebted modalities and enforces modality-wise minimum reconstruction service.
- **Theoretical guarantee.** We offer theoretical justifications for DEMO, demonstrating that expert load balancing in MoE cannot guarantee effective modality service and further proving that every serviceable modality meets its minimum long-run service objective in expectation under a stationary stochastic routing environment and strict feasibility.
- **Empirical validation.** Large-sacle experimental results reveal that DEMO effectively mitigates modality competition and achieves superior performance across MAD datasets and various CL settings.

2 RELATED WORK

Multimodal Anomaly Detection. Compared to unimodal anomaly detection approaches (Deng & Li, 2022; Wei et al., 2025), MAD methods can use complementary modalities to identify undetected anomalies on a single modality. Existing works mainly focus on cross-modal alignment (Costanzino et al., 2024; Tu et al., 2024) and feature storage (Wang et al., 2023; Chu et al., 2023; Li et al., 2026). For instance, CFM (Costanzino et al., 2024) aligns RGB and 3D by cross-modal feature mapping using simple linear layers. M3DM (Wang et al., 2023) stores RGB, 3D, and fusion features on respective memory bank, while PIRN (Li et al., 2026) introduces a codebook of some discrete prototypes for every modality. Beyond these mainstream approaches, recent some efforts have made to an architecture perspective (Long et al., 2025; Gu et al., 2026) and knowledge distillation (Gu et al., 2024). Despite this significant progress in MAD, existing methods overlook a critical issue, modality competition, which leaves one modality suppressed by the other. This hinders normality modelling on the suppressed modality, leading to inaccurate MAD results. This paper tackles this challenge from an MoE perspective.

Mixture-of-Experts. Inspired by the idea of conditional computation, Jacobs et al. (1991) first introduced MoE architecture. Given its potential in designing lightweight model and enhancing model capacity, this approach is widely applied in different fields (Li et al., 2025a), including AD tasks. For instance, MoEAD (Meng et al., 2024) actives fewer experts through stacked MoE layers but reaches better detection results. Adapted-MoE (Lei et al., 2025) dynamically selects experts using a routing network and adaptively adjusts expert weights to eliminate deviations between test samples and training distributions during inference. Thanks to the expert specialization ability, recent efforts have attempted to use this merit to resolve more challenging tasks, such as multiclass MAD (Gu et al., 2026; Zhao et al., 2026) and zero-shot AD (Park et al., 2026; Wei et al., 2026). However, in existing MoE, it follows a foresight training paradigm, so its router may not learn why these experts are used. Differently, we propose a novel training paradigm for MoE, called hindsight learning, which enables the router to learn the reasons behind it via a post-hoc teacher.

3 METHODOLOGY

3.1 APPROACH OVERVIEW

We propose a novel MoE training paradigm, termed hindsight learning. To better learn multimodal normal features, it explicitly guides the router to learn expert assignment through a post-hoc routing teacher derived from observable per-expert advantages using OT, as illustrated in Figure 2. Grounded in this paradigm, we then introduce DEMO, a method that mitigates modality competition by preventing starvation in modal normality reconstruction utility, as shown in Figure 3. DEMO calculates post-hoc reconstruction gains brought by each expert relative to an anchor expert for each modality, and normalizes the realized gains as utility, where this utility serves as the virtual service in a modality-debt queue. If a serviceable modality receives utility below its minimum service objective in the queue, its debt accumulates. To reduce debt, the router is trained to first allocate useful expert capacity to that modality using debt-aware OT.

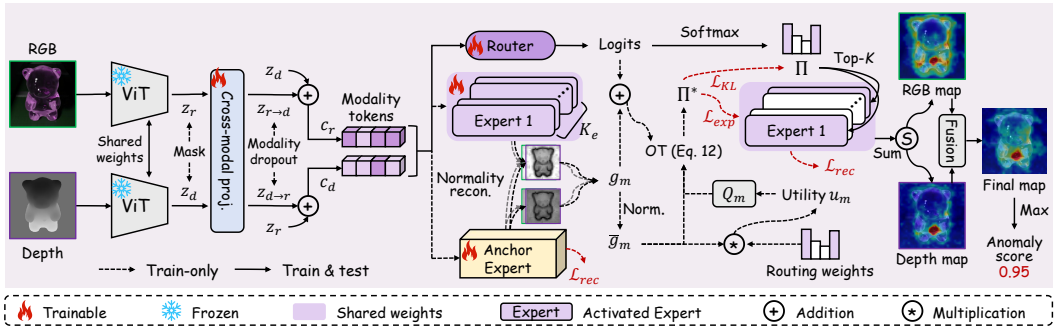


Figure 3: Overall framework of DEMO.

3.2 HINDSIGHT LEARNING

Motivation. Existing MoE-based MAD methods (Gu et al., 2026) follow previous sparse paradigm (Shazeer et al., 2017). Given the input x, the router $\mathcal{R}$ activates K expert(s) within K_e experts $\{E_k\}_{k=1}^{K_e}$ via its predicted routing weights $\hat{\Pi}$, and the output is the weighted sum of selected experts:

$$y_m = \sum\nolimits_{j \in \mathrm{Top}K(\hat{\Pi}_m)} \hat{\Pi}_{mj} E_j(x_m), \quad m \in \mathcal{M} = \{r, d\}, \tag{1}$$

$$\hat{\Pi}_m = \mathrm{Softmax}(\mathcal{R}(x_m)), \quad \hat{\texttt{Logits}}_m = \mathcal{R}(x_m), \tag{2}$$

During backpropagation, the router gradients are derived solely from task cost (e.g., reconstruction loss) propagated through the selected experts' outputs. Thus, the router learns to adjust its predicted weights implicitly and indirectly. We term this foresight learning, as shown in Figure 2(a). However, this paradigm suffers from a critical issue: the router is trained only to decide which expert receives each modality token, without learning why that assignment is beneficial.

Novel paradigm. We propose a novel MoE training paradigm, dubbed hindsight learning. Instead of relying solely on the implicit gradients of the task cost, it exploits the fact that task-specific cost of every expert is observable during training. We can thus construct a post-hoc advantage-derived teacher that evaluates the marginal value of assigning each modality token to every expert.

Concretely, as shown in Figure 2(b), each expert first processes the tokens to compute its task cost (e.g., reconstruction loss) $c(E_k)$ during training. Then, we define the advantage $\mathcal{A}_k$ of k-th expert against an anchor expert E_0 as: $\mathcal{A}_k = c(E_0) - c(E_k)$. Based on the raw logits $\hat{\texttt{Logits}}$ and $\mathcal{A}$, a post-hoc routing teacher $\Pi^\star$ is generated using entropy-regularized optimal transport and the Sinkhorn algorithm (Cuturi, 2013) under expert-capacity constraints:

$$\begin{aligned} \hat{\Pi}^* \in \arg\max_{\hat{\Pi}} & \sum\nolimits_{m,i,k} \hat{\Pi}_{mik} \left(\hat{\texttt{Logits}}_{mik} + \mathcal{A}_{mik} \right) + \tau H(\hat{\Pi}), \\ \text{s.t.} \quad & \sum\nolimits_k \hat{\Pi}_{mik} = 1, \quad \sum\nolimits_{m,i} \hat{\Pi}_{mik} = \frac{|\mathcal{M}|N}{K_e}, \end{aligned} \tag{3}$$

where i indexes the tokens, N is the total number of tokens per modality, $H(\hat{\Pi})$ is entropy regularization, and $\tau > 0$ is the routing balancing coefficient. With $\hat{\Pi}^*$, we train the router to approximate the routing teacher via KL divergence:

$$\mathcal{L}_{\text{route}}(\hat{\Pi}^*, \hat{\Pi}) = \mathrm{KL}\left(\hat{\Pi}^* \parallel \hat{\Pi} \right). \tag{4}$$

By doing so, the router learns to imitate a post-hoc teacher informed by the observed expert advantages for the task. During inference, the target reconstruction is not used. Based on the learned expert advantages, the router activates K ($K = 2$ by default) experts that will be the best in hindsight given its input via Eq. (1). In DEMO, we use the relative normality reconstruction gain (Eq. (9)) to instantiate $\mathcal{A}$.

3.3 DEBT-AWARE MOE (DEMO)

Modal normality reconstruction. Given RGB and depth modalities for input I, we employ a pre-trained Vision Transformer (ViT) as the shared, frozen feature extractor $\mathcal{E}$: $z_m = \mathcal{E}(I_m)$. Then, we

use a shared projection layer ϕ to generate cross-modal complementary features:

$$z_{r\to d} = \phi(z_r), \quad z_{d\to r} = \phi(z_d). \tag{5}$$

At the same time, we also randomly discard cross-modal features with a probability of $p = 0.15$ during training, in order to prevent the network from relying on RGB or depth modality throughout. Besides, to avoid identity mapping issue (You et al., 2022), we apply random masks $\mathbf{M}_{mask}$ to extracted features: $\tilde{z}_m = \mathbf{M}_{m,mask} \odot z_m$. Then, we can obtain the local context of the masked modal features and another modal features as follows:

$$c_r = \tilde{z}_r + z_{d\to r}, \quad c_d = \tilde{z}_d + z_{r\to d}. \tag{6}$$

Based on these multimodal local contexts, they are respectively reconstructed by anchor expert E_0 and K_e residual experts as follows:

$$\hat{z}_m^{(0)} = E_0(c_m), \quad \hat{z}_m^{(k)} = \hat{z}_m^{(0)} + E_k(c_m), \quad k \in \{1, \dots, K_e\}. \tag{7}$$

The corresponding task cost for anchor expert and residual experts can be computed as:

$$\ell_{mi}^{(0)} = d\left(z_{mi}, \hat{z}_{mi}^{(0)}\right), \quad \ell_{mi}^{(k)} = d\left(z_{mi}, \hat{z}_{mi}^{(k)}\right). \tag{8}$$

where $d(:,:)$ denotes cosine distance.

Modal utility and debt. For each token and residual expert, we then define marginal noramlity reconstruction gains that k-th expert brings to token(m, i) relative to anchor expert as:

$$g_{mik}^t = \frac{[\ell_{mi}^{(0)} - \ell_{mi}^{(k)}]_+}{\ell_{mi}^{(0)} + \varepsilon}. \tag{9}$$

If reconstruction loss is non-negative, $g_{mik}^t \in [0, 1]$. Superscript t is t-th batch of continual training. g_{mik}^t measures how much additional normality reconstruction gains the residual expert brings to modality m. However, marginal reconstruction gains of different modalities vary greatly, due to the inherent disparities in reconstruction difficulty. Thus, we introduce modal reconstruction utility to measure the proportion of the modality m that has obtained its unconstrained per-token upper bound of reconstruction gains, independent of absolute reconstruction difficulty. Formally, let a_m^t denote whether batch t is serviceable and $\Pi = \text{Softmax}(\mathcal{R}(c))$, we define modal reconstruction utility as:

$$u_m^t(\Pi) = \sum\nolimits_{i,k} \Pi_{mik} \bar{g}_{mik}^t, \quad \bar{g}_{mik}^t = \frac{a_m^t g_{mik}^t}{\max\{G_m^t, \epsilon_G\}}, \quad G_m^t = \sum\nolimits_i \max_k g_{mik}^t, \tag{10}$$

where $a_m^t = \mathbf{1}\{G_m^t > \epsilon_G\}$ and $\epsilon_G = 1e\text{-}6 > 0$ is a small demand threshold. For a serviceable batch, utility $u_m^t(\Pi)$ is the realized fraction of the modality's available gain and lies in $[0, 1]$. If $G_m^t \le \epsilon_G$, then $a_m^t = 0$, $\bar{g}_{mik}^t = 0$, and $u_m^t = 0$. This indicates that the batch provides no actionable expert demand. To measure whether modal reconstruction utility is effective, inspired by virtual queues and Lyapunov optimization (Tassiulas & Ephremides, 1992; Neely, 2010), we introduce a virtual modality-debt queue Q for each modality, which can be defined as:

$$Q_m(t+1) = \left[Q_m(t) + a_m^t \gamma_m - u_m^t(\Pi^t)\right]_+, \qquad Q_m(0) \ge 0, \tag{11}$$

where $\gamma_m \in (0, 1)$ denotes the minimum service objective on serviceable batches and $u_m^t(\Pi^t)$ serves as virtual modality service in the queue Q_m. If the service is less than γ_m, the debt increases. To be precise, the debt grows only when serviceable demand exists and the realized service falls below γ_m. When on unserviceable batches ($G_m^t \le \epsilon_G$), we skip debt accrual and repayment because no actionable gain is available. Besides, note that Q_m is not cleared at the task boundary if the model involves CL. The debt will be carried over to subsequent training tasks.

Modality-debt routing. Based on the queue state, we train the router to learn debt-aware allocation to repay debt using debt-aware OT. It constructs the debt-aware teacher by modifying Eq. (3) as:

$$\begin{aligned} \Pi_t^* \in \arg\max_{\Pi \in \mathcal{P}_t} &\left\{ \sum\nolimits_{m,i,k} \Pi_{mik} b_{mik}^t + \beta \sum\nolimits_{m,i,k} Q_m(t) \Pi_{mik} \bar{g}_{mik}^t + \tau H(\Pi) \right\}, \\ \mathcal{P}_t = &\left\{ \sum\nolimits_k \Pi_{mik} = 1,\ \sum\nolimits_{m,i} \Pi_{mik} = \frac{|\mathcal{M}|N}{K_e},\ \Pi_{mik} \ge 0 \right\}. \end{aligned} \tag{12}$$

Here $b^t_{mik} = \texttt{Logits}^t_{mik} + g^t_{mik}$ is the bounded base hindsight score, where the logits $\texttt{Logits}_m = \mathcal{R}(c_m)$ and are clipped for a fixed range. $\beta > 0$ controls the trade-off between base score and debt. Correspondingly, Eq. (4) can be rewritten as $\mathcal{L}^t_{\text{route}}(\Pi^*_t, \Pi^t)$. In Appendix B.2, we further show that the above Eq. (12) is algebraically equivalent to Eq. (24), a drift-plus-penalty (DPP) objective (Neely, 2010), and state the assumptions used for the starvation-free guarantee on modality service.

Extension to CL. As empirically discussed in Section 1, modality competition persists during CL. To adapt DEMO to CL, we freeze ϕ and the anchor expert E_0 to serve as a stable normality anchor point, while the residual experts remain trainable during CL. We then propose a latent capsule $\mathcal{C}$ with a limited memory of M to mitigate forgetting. During l-th training step/task ($l = 1$ denotes the base task), we randomly store some current data in $\mathcal{C}$ and sample a small amount of previous data from it for replay, mixing current and replay data to train DEMO via $\mathcal{L}_{\text{replay}}$ (Eq. (20) in Appendix A). More details are presented in Appendix A.

Training and inference. Overall losses for DEMO can be expressed as:

$$\mathcal{L}^t = \mathcal{L}^t_{\text{rec}} + \tau \mathcal{L}^t_{\text{route}}(\Pi^*_t, \Pi^t) + \lambda_1 \mathcal{L}^t_{\text{exp}} + \lambda_2 \mathcal{L}^t_{\text{replay}}, \tag{13}$$

$$\mathcal{L}^t_{\text{rec}} = \sum\nolimits_m (\ell^{(0)}_m + d(z^t_m, \Pi^t_{mk^*_t} \cdot z^{(k^*_t)}_m)), \quad k^*_t = \arg\max_k \Pi^t_{mk}, \tag{14}$$

$$\mathcal{L}^t_{\text{exp}} = \frac{1}{|\mathcal{M}|N} \sum\nolimits_{m,i,k} \Pi^*_{tmik} \cdot \ell^{(k)}_{mi}. \tag{15}$$

Here $\ell^{(0)}_m = 0$ if $l > 1$. λ_1 and $\lambda_2 > 0$ are hyper-parameters for balancing the losses. For reconstruction objective ((14) and the first term of Eq. (20)), we employ a hard routing strategy to train the expert to be more specialized. The gradient is propagated back to the router via a straight-through estimator. Meanwhile, to ensure that other experts also receive supervisory signals and to enforce the global capacity constraints derived from OT Π^*_t, we introduce an expert-alignment loss (Eq. (15)). Algorithm 1 in Appendix provides the training process of DEMO. During inference, modality-wise anomaly map is defined as: $\text{map}_m = d(z_m, \sum_{j \in \text{Top}K(\Pi_m)} \Pi_{mj} \hat{z}^{(j)}_m)$. Both maps are upsampled to the input resolution and the final multimodal anomaly map is defined as: $\text{map}_{\text{fusion}} = \text{map}_r + \text{map}_d$. The maximum value in $\text{map}_{\text{fusion}}$ is the image-level anomaly score.

3.4 THEORETICAL ANALYSIS

We first distinguish expert load balancing from modality service, and then establish a minimum-service guarantee for DEMO under a stationary stochastic routing environment. The queue-stability argument follows the Max-Weight and virtual-queue DPP frameworks (Tassiulas & Ephremides, 1992; Neely, 2010). We instantiate service as normalized modality-specific marginal normality reconstruction utility under expert-capacity constraints.

In conventional sparse MoE (Shazeer et al., 2017; Fedus et al., 2022), considering Top-1 routing, load-balancing objectives encourage $\text{Load}(E_k) \approx N/K_e$ to prevent expert collapse. However, balanced expert traffic does not imply non-zero modality service.

Proposition 1 (Non-entailment)**.** *Even if each expert E_k receives an identical number of tokens, it cannot be inferred that each modality m obtains non-zero effective normality reconstruction utility u_m. See proof in Appendix B.1.*

We next show that normalized reconstruction service cannot vanish on serviceable batches under the stationary assumptions.

Theorem 2 (Starvation-free modality service)**.** *Assume a stationary stochastic routing environment, bounded normalized utility and routing cost, a feasible δ-approximate debt-aware DPP router, and a strictly feasible stationary policy. If serviceable batches occur with non-vanishing frequency, then the modality-debt queues are strongly stable and mean-rate stable, and every modality satisfies*

$$\liminf_{T\to\infty} \frac{\sum_{t=0}^{T-1} \mathbb{E}[u^t_m(\Pi^t)]}{\sum_{t=0}^{T-1} \mathbb{E}[a^t_m]} \geq \gamma_m > 0. \tag{16}$$

When every batch is serviceable ($a^t_m = 1$), this reduces to $\liminf_{T\to\infty} T^{-1} \sum_{t<T} \mathbb{E}[u^t_m] \geq \gamma_m$.

In Eq. (16), γ_m is the minimal service objective in the queue in Eq. (11). Theorem 2 guarantees that no modality can receive reconstruction service approaching zero over long-term training, thereby mitigating modality competition. See Appendix B.2 for proof.

Table 1: Results in different settings of two datasets. Average image-level AUROC and PRO (↑) and their respective FM (↓) are listed. †: memory-limited. **Best** and second-best results are highlighted.

(a) MVTec 3D dataset

Method	Venue	setting 1		setting 2		setting 3		setting 4	
		Average	FM	Average	FM	Average	FM	Average	FM
CFM	CVPR24	92.1 / **93.0**	– / –	84.9 / 89.7	8.5 / 6.9	79.3 / 82.0	9.3 / 8.4	65.7 / 74.6	16.8 / 12.3
AnomalyMoE†	AAAI26	93.0 / 92.5	– / –	84.0 / 90.4	8.8 / 5.3	78.4 / 83.1	10.2 / 6.7	63.4 / 75.4	19.1 / 13.0
IUF	ECCV24	90.4 / 91.3	– / –	82.9 / 85.2	9.4 / 7.0	77.7 / 81.2	14.7 / 8.0	63.8 / 75.0	14.4 / 10.5
CDAD	CVPR25	81.5 / 89.8	– / –	72.6 / 79.9	9.8 / 8.6	70.6 / 76.4	8.9 / 7.5	64.2 / 68.1	15.5 / 12.2
IB-IUMAD	CVPR26	91.5 / 91.7	– / –	84.4 / 87.1	7.9 / 7.1	80.7 / 84.0	8.4 / 7.2	67.6 / 76.4	10.1 / 9.9
DEMO	**Ours**	**93.5** / 92.7	– / –	**86.7 / 94.6**	**7.9 / 1.8**	**85.0 / 93.6**	**3.4 / 1.2**	**71.1 / 90.3**	**6.6 / 2.1**

(b) Eyecandies dataset

Method	Venue	setting 1		setting 2		setting 3		setting 4	
		Average	FM	Average	FM	Average	FM	Average	FM
CFM	CVPR24	84.3 / 86.7	– / –	72.5 / 75.9	10.1 / 8.8	70.4 / 74.2	9.2 / 6.5	59.5 / 68.2	17.1 / 14.3
AnomalyMoE†	AAAI26	88.7 / 88.5	– / –	78.8 / 80.2	7.2 / 5.7	70.0 / 76.6	10.1 / 9.8	62.5 / 69.1	20.4 / 15.2
IUF	ECCV24	82.9 / 87.8	– / –	67.3 / 71.4	15.6 / 10.2	67.7 / 73.0	13.0 / 9.4	61.4 / 68.0	18.1 / 12.8
CDAD	CVPR25	81.2 / 84.7	– / –	68.1 / 73.0	10.5 / 7.6	69.5 / 74.1	9.0 / 7.2	60.8 / 67.3	16.4 / 13.0
IB-IUMAD	CVPR26	83.3 / 88.0	– / –	70.4 / 75.7	9.8 / 7.7	72.2 / 77.0	8.8 / 6.9	64.7 / 70.6	14.5 / 11.1
DEMO	**Ours**	**91.1 / 90.1**	– / –	**85.9 / 88.0**	**4.1 / 1.5**	**79.9 / 84.9**	**4.1 / 2.6**	**71.3 / 74.4**	**6.4 / 5.7**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets & evaluation metrics. Two MAD datasets were introduced: MVTec 3D (Bergmann et al., 2022) and Eyecandies (Bonfiglioli et al., 2022). We introduced different MAD settings, including multiclass MAD (setting 1) and three continual MAD (CMAD) settings: 6 – 4 with 1 step (setting 2), 6 – 1 with 4 steps (setting 3), and 1 – 1 with 9 steps (setting 4). We adopted two widely used metrics: image-level AUROC and PRO. To evaluate CL ability, the forgetting measure (FM) (Tang et al., 2024) was considered.

Implementation details & baselines. We employed ViT-B/14 pre-trained by DINOv2 (Oquab et al., 2023) as the feature extractor. The main hyper-parameters, K_e, γ, and M, were set to 4, 0.8, and 32, respectively. The balancing hyper-parameters, λ_1, λ_2, τ, and β, were set to 0.5, 0.05, 0.2, and 2, respectively. We compared DEMO against several representative methods, including IUF (Tang et al., 2024), CDAD (Li et al., 2025b), IB-IUMAD (Long et al., 2026)), CFM (Costanzino et al., 2024), and AnomalyMoE (Gu et al., 2026). Both CFM and AnomalyMoE do not have the ability of CL, so we introduced EWC (Kirkpatrick et al., 2017) into them to mitigate forgetting. For fair comparison, the encoder or feature extractor in baselines was uniformly replaced with DINOv2. Complete experimental setup has been presented in Appendix C.

4.2 MAIN RESULTS

Results on MAD. We first assess the validity of DEMO in multiclass MAD (setting 1) of both MVTec 3D and Eyecandies datasets, as reported in Table 1. It demonstrates that DEMO achieves superior performance compared to competing methods, particularly in terms of image-level AUROC metric, with an improvement of 0.5% and 2.4% on the respective dataset. DEMO also achieves comparable and even better results in terms of PRO metric on datasets. These results are mainly attributed to our intervention in modality competition, which renders the information of every modality available. This, in turn, enables effective modality-wise normality modelling and leads to optimal MAD performance.

Results on CMAD. In Table 1, we further compare DEMO with competing methods in CL tasks. As recorded, DEMO consistently outperforms all baselines across different CL settings on two datasets, with marked gains of at least 1.8% and lower forgetting rate. In particular, it improves state-of-the-art method by 7.1% Average image-level AUROC in setting 2 of Eyecandies dataset and 13.9%

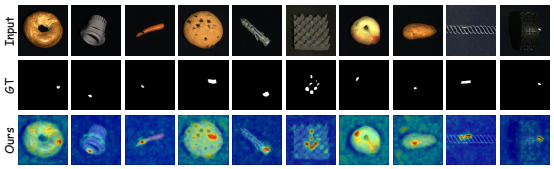


Figure 4: Qualitative results in setting 4 of MVTec 3D dataset. Best viewed in colours.

Table 2: Ablation studies on core modules in both setting 2 and 3 of MVTec 3D dataset.

	Setting 2 (2 tasks)		Setting 3 (5 tasks)	
	Average	FM	Average	FM
w/o HL	**86.9** / 89.4	7.9 / 1.9	84.1 / 92.5	4.2 / 2.6
w/o Q	85.1 / 92.7	**7.7** / **0.5**	82.2 / 92.9	4.4 / 0.5
w/o $\mathcal{R}$	85.9 / 93.2	8.4 / 2.3	83.0 / 93.0	5.3 / **0.4**
DEMO	86.7 / **94.6**	7.9 / 1.8	**85.0** / **93.6**	**3.4** / 1.2

Average PRO in setting 4 of MVTec 3D dataset. These results highlight the effectiveness of DEMO in mitigating forgetting and preventing modality competition. To more intuitively show the effectiveness of DEMO, Figure 4 illustrates the visualization results in setting 4 of MVTec 3D dataset. More visualization results have been provided in Appendix D.2.

4.3 ABLATION STUDY

The core modules in DEMO include hindsight learning (denoted as HL), modality-debt queue Q, and debt-aware router $\mathcal{R}$. To evaluate their importance, we provide a series of experimental results in both setting 2 and 3 of MVTec 3D dataset.

Study on hindsight learning. One of the elements in DEMO that differs the most from existing MoE-based methods is HL. To assess its necessity, we replace it with original MoE training paradigm, namely foresight learning (FL). Table 2 shows the results in detail. Compared to HL, using FL hardly breaks the bottleneck of detection performance and even achieves decreased results, particularly Average PRO metric ($94.6\% \rightarrow 89.4\%$) in setting 2. Because FL trains the router implicitly and indirectly. Conversely, HL explicitly guides the router via per-expert advantage feedback to adjust its predicted weights, thereby achieving better results.

Study on modality-debt queue. We then evaluate the contribution of Q in DEMO, which enforces a lower bound on the virtual reconstruction service for each modality. Upon removing Q, Eq. (12) degenerates into Eq. (3) and Π^* in Eq. (15) becomes Π. The router $\mathcal{R}$ loses its capacity of debt-aware allocation. Table 2 presents the resulting impact on performance. As expected, the performance is not promising. The more tasks there are, the worse the performance becomes ($86.7\% \rightarrow 85.1\%$ in setting 2 and $85.0\% \rightarrow 82.2\%$ in setting 3). As modality competition persists during CL, this results in degraded performance in MAD. To clearly show the intervention on modality competition, we provide visualization results in setting 1 of MVTec 3D dataset in Figure 5. As visualized, DEMO effectively mitigates modality competition, which is intuitively manifested as the growth of the information acquisition of the depth modality. This guarantees better normality modelling on that modality.

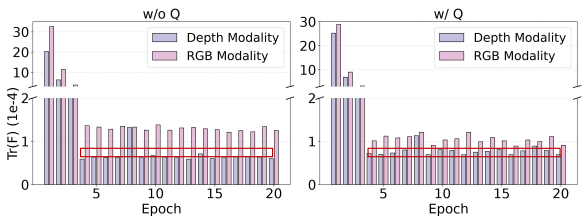


Figure 5: Modality competition analysis w/o and w/ Q.

Study on debt-aware router. We eventually investigate the impact of $\mathcal{R}$ on performance. Although Q guarantees the lower bound, $\mathcal{R}$ fails to perform debt-aware allocation after removing it, leaving modality competition still unresolved. Thus, the performance also declines, as recorded in Table 2.

4.4 ANALYSIS

Computational efficiency analysis. To examine the computational efficiency of DEMO against baselines, we report memory storage and inference speed (fps) in setting 4 of MVTec 3D dataset. During CL, IUF and IB-IUMAD store the old weights along with the channel space obtained by SVD of the old features, whereas CDAD stores the gradient information.

Table 3: Computational efficiency analysis on an NVIDIA GeForce RTX 5090 32 GB GPB.

Method	Memory usage	Frame rate	Average	FM
IUF	55.9 MB	35.74	63.8 / 75.0	14.4 / 10.5
CDAD	2.00 GB	3.11	64.2 / 68.1	15.5 / 12.2
IB-IUMAD	64.5 MB	32.53	67.6 / 76.4	10.1 / 9.9
DEMO	**37.0** MB	**46.99**	**71.1** / **90.3**	**6.6** / **2.1**

Our DEMO stores some data in a latent capsule with limited capacity. Table 3 shows that DEMO requires the lowest memory storage (37 MB) and achieves the highest inference speed (46.99 fps). IUF and IB-IUMAD demand additional memory storage and inference time compared with ours, while CDAD exhibits the worst results in terms of both memory (2 GB) and speed (3.11 fps). Overall, DEMO strikes a favourable balance between computational efficiency and performance. **Limitation.** However, due to hindsight learning, each expert involves forward propagation during training. Thus, our complexity is $O(K_e)$ rather than $O(K)$ (Shazeer et al., 2017). This overhead is negligible at the training phase in our method, but it may be not suited to large models with over hundreds of experts.

Analysis on modality service and expert usage. In DEMO, we prioritize modality service over equal expert traffic. As shown in Figure 6, the normality reconstruction utility for each modality satisfies Eq. 16, namely, every serviceable modality meets its minimum long-run service objective in expectation (e.g., $0.939 > 0.8$). Besides, it is also observed that while depth modality may incur multiple debts in the early stage of training, both modality services approach balance in the later stage. This means no modality remains in a long-run suppressed state, thereby mitigating modality competition. For expert usage, DEMO also ensures relative load balancing, especially when $K = 2$. Overall, unlike existing MoE methods (Shazeer et al., 2017; Gu et al., 2026), our method may sacrifice a slight degree of expert usage balance but achieves better-balanced modality service.

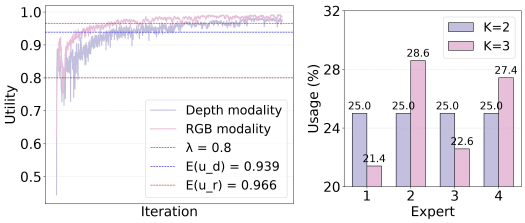


Figure 6: Analysis on modality service and expert usage in setting 1 of MVTec 3D dataset.

Table 4: Analysis on hyper-parameters K_e, γ, and M in setting 3 of the MVTec 3D dataset.

(a) Analysis on the number of experts

K_e	Average	FM
3	83.7 / 92.6	6.6 / 0.4
4	**85.0 / 93.6**	**3.4 / 0.1**
5	81.9 / 93.0	8.1 / **0.1**
6	83.6 / 92.4	4.7 / 1.6

(b) Analysis on minimal service

γ	Average	FM
0.6	83.5 / 92.5	5.0 / 1.9
0.7	84.9 / 93.0	3.8 / 1.6
0.8	**85.0 / 93.6**	**3.4 / 1.2**
0.9	83.9 / 93.3	4.8 / **1.2**

(c) Analysis on memory size

M	Average	FM	Memory usage
16	83.2 / 93.1	4.5 / 1.0	18.5 MB
32	84.5 / 93.6	3.8 / **0.9**	37 MB
64	**85.0** / 93.6	**3.4** / 1.2	74 MB
128	84.5 / **93.9**	4.0 / 1.5	148 MB

Hyper-parameter analysis. The main hyper-parameters in DEMO include K_e, γ, and M. Table 4 presents the detailed results. The number of experts K_e is crucial for MoE, as it affects the degree of expert specialization, training stability, *etc*. As reported in Table 4(a), using more experts does not necessarily improve performance and may even perform worse than using fewer experts. The best results are achieved when $K_e = 4$. As for γ, it controls the lower bound. Table 4(b) shows that both overly low or overly high lower bounds can degrade performance. The former leads to poor modality service since the objective is easily satisfied, while the latter results from an excessively high minimal service objective for both modalities. As recorded, setting γ to 0.8 is the most reasonable. Finally, M decides the memory size. A large size may facilitate the learning of old knowledge. However, it also increases memory overhead. As shown in Table 4(c), when $M = 32$, DEMO achieves comparable detection results while maintaining lower memory overhead than $M \in \{64, 128\}$. More analyses have been provided in Appendix D.1.

5 CONCLUSION

In this paper, we investigated modality competition, a frequently overlooked issue in Multimodal Anomaly Detection (MAD) that persists even in continual learning. We tackled this problem from a Mixture-of-Experts (MoE) perspective and proposed a novel training paradigm termed hindsight learning. It explicitly trains the router to learn token assignment to a particular expert via a post-hoc routing teacher constructed from observable per-expert advantages using Optimal Transport (OT). Based on this paradigm, we proposed a debt-aware MoE (DEMO) to mitigate modality competition. In DEMO, each modality maintains a virtual debt queue. The debts grows when the modality service falls below the minimum service objective. To reduce debt, the router learns debt-aware allocation via debt-aware OT. We provided comprehensive theoretical analyses and empirical results showing the effectiveness of DEMO. Our research revealed that modality competition has a severe impact on MAD, paving the way for future works.

AI USE STATEMENT

We used generative AI tools solely for grammar checking, paper polishing, code bug checking and optimization, and rigorousness verification of theoretical proofs. Apart from these uses, the ideas, the proposed method together with the corresponding original code, the theoretical contributions, and the others (e.g., the overall architecture of the paper and the method framework paradigm) presented in this paper are entirely our own. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All experiments conducted in this paper are reproducible. We used publicly available datasets and pre-trained models, and provided the complete experimental setup (including datasets, protocols, implementation details, *etc*.) in Appendix C. Besides, we provide Algorithm 1 in Appendix C to facilitate understanding of the training process of DEMO. For the theoretical analysis, detailed proofs are provided in Appendices B.1 and B.2. The code will be made publicly available upon acceptance.

REFERENCES

Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pp. 202–213, 2022.

Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *Asian Conference on Computer Vision*, pp. 459–475. Springer, 2022.

Yu-Min Chu, Chieh Liu, Ting-I Hsieh, Hwann-Tzong Chen, and Tyng-Luh Liu. Shape-guided dual-memory learning for 3D anomaly detection. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 6185–6194. PMLR, 2023.

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

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9727–9736. IEEE, 2022.

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

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Wei Ge, Ming Tang, and Jinqiao Wang. Anomalymoe: Towards a language-free generalist model for unified visual anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4348–4356, 2026.

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

Chengxiang Huang, Yake Wei, Zequn Yang, and Di Hu. Adaptive unimodal regulation for balanced multimodal information acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25854–25863. IEEE, 2025.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Tianwu Lei, Silin Chen, Bohan Wang, Zhengkai Jiang, and Ningmu Zou. Adapted-moe: Mixture of experts with test-time adaption for anomaly detection. In *International Conference on Intelligent Computing*, pp. 427–441. Springer, 2025.

Ruocheng Li, Lei Wu, Jingjun Gu, Qi Xu, Wanyi Chen, Xiaoxu Cai, and Jiajun Bu. Moe-sam: Enhancing sam for medical image segmentation with mixture-of-experts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 367–377. Springer, 2025a.

Xiaofan Li, Xin Tan, Zhuo Chen, Zhizhong Zhang, Ruixin Zhang, Rizen Guo, Guannan Jiang, Yulong Chen, Yanyun Qu, Lizhuang Ma, and Yuan Xie. One-for-more: Continual diffusion model for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4766–4775. IEEE, 2025b.

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

Kaifang Long, Guoyang Xie, Lianbo Ma, Jiaqi Liu, and Zhichao Lu. Revisiting multimodal fusion for 3d anomaly detection from an architectural perspective. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 12273–12281, 2025.

Kaifang Long, Lianbo Ma, Jiaqi Liu, Liming Liu, and Guoyang Xie. Towards an incremental unified multimodal anomaly detection: Augmenting multimodal denoising from an information bottleneck perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14116–14125, 2026.

Shiyuan Meng, Wenchao Meng, Qihang Zhou, Shizhong Li, Weiye Hou, and Shibo He. Moead: A parameter-efficient model for multi-class anomaly detection. In *European Conference on Computer Vision*, pp. 345–361. Springer, 2024.

Michael J. Neely. *Stochastic network optimization with application to communication and queueing systems*. Morgan & Claypool Publishers, 2010. URL https://doi.org/10.2200/S00271ED1V01Y201006CNT007.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://doi.org/10.48550/arXiv.2304.07193.

Jun Yeong Park, JunYoung Seo, Minji Kang, and Yu Rang Park. Moeclip: patch-specialized experts for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35534–35544, 2026.

Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 6443–6451, 2025.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8237. IEEE, 2022.

Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2591–2601. IEEE, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Jiaqi Tang, Hao Lu, Xiaogang Xu, Ruizheng Wu, Sixing Hu, Tong Zhang, Tsz Wa Cheng, Ming Ge, Ying-Cong Chen, and Fugee Tsung. An incremental unified framework for small defect inspection. In *European Conference on Computer Vision*, pp. 307–324. Springer, 2024.

L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 37(12):1936–1948, 1992.

Yuanpeng Tu, Boshen Zhang, Liang Liu, Yuxi Li, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Cairong Zhao. Self-supervised feature adaptation for 3d industrial anomaly detection. In *European conference on computer vision*, pp. 75–91. Springer, 2024.

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

Shun Wei, Jielin Jiang, and Xiaolong Xu. 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*, pp. 9994–10003. IEEE, 2025.

Shun Wei, Jielin Jiang, and Xiaolong Xu. Exploring high-order-aware prompt learning for zero-shot anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10566–10574, 2026.

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

Muhao Xu, Lili Qu, Zihan Nie, Feng Li, Zhuangzhuang Chen, Qi Liao, Yi Wan, Sijie Niu, Runmin Cong, Xin Li, and Weiye Song. Towards reliable medical unsupervised anomaly detection: A benchmark and dataset for PET/CT with cross-modality knowledge distillation network. *Information Fusion*, 133:104341, 2026.

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.

Yuan Zhao, Youwei Pang, Lihe Zhang, Hanqi Liu, Jiaming Zuo, Huchuan Lu, and Xiaoqi Zhao. Unimmad: Unified multi-modal and multi-class anomaly detection via moe-driven feature decompression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28502–28511, 2026.

A Additional Details of Latent Capsule

Storage. To avoid excessive memory usage, the memory size M of latent capsule $\mathcal{C}$ is limited, which is shared for all training steps. Besides, to further reduce memory usage, we apply spatial pooling on tokens. Concretely, $z_m \in \mathbb{R}^{d\times n}$ is reshaped to $z_m^{\text{reshape}} \in \mathbb{R}^{d\times\sqrt{n}\times\sqrt{n}}$, where d is the dimension and n is the number of tokens. Then, we adopt a pooling function:

$$z_m^{\text{pooled}} = \text{Flatten}(\text{AdaptiveAvgPool2d}_{\sqrt{n}/2\times\sqrt{n}/2}(z_m^{\text{reshape}})) \in \mathbb{R}^{d\times(\sqrt{n}/2)^2}. \tag{17}$$

In particular, the router weights Π also follow Eq. (17). After pooling operation, $\left\{z_m^{\text{pooled},l}, \Pi_m^{\text{pooled},l}\right\}$ are stored in $\mathcal{C}$ with half precision during l-th training step.

Replay. While storing, we also replay previous data $\left\{z_m^{\text{pooled},<l}, \Pi_m^{\text{pooled},<l}\right\}_{l\neq 1}$ from $\mathcal{C}$. During replay, we do not apply masking and modality dropout on. $z_m^{\text{pooled},<l}$ is directly processed by ϕ using Eq. (5) to obtain multimodal local contexts:

$$c_r^{\text{pooled},<l} = z_r^{\text{pooled},<l} + \phi(z_d^{\text{pooled},<l}), \quad c_d^{\text{pooled},<l} = z_d^{\text{pooled},<l} + \phi(z_r^{\text{pooled},<l}). \tag{18}$$

We obtain reconstruction features of each residual expert following Eq. (7):

$$\hat{z}_m^{\text{pooled},<l,(k)} = E_0(c_m^{\text{pooled},<l}) + E_k(c_m^{\text{pooled},<l}). \tag{19}$$

Then DEMO can mitigate forgetting through the replay loss as follows:

$$\mathcal{L}_{\text{replay}}^t = \sum\nolimits_m d(z_m^{\text{pooled},<l}, \bar{\Pi}_{mk^\star}^{\text{pooled},<l} \cdot \hat{z}_m^{\text{pooled},<l,(k^\star)}) + \mathcal{L}_{\text{route}}^t(\Pi_m^{\text{pooled},<l}, \Pi_m^t), \quad k^\star = \arg\max_k \bar{\Pi}_{mk}^{\text{pooled},<l}, \tag{20}$$

where $\bar{\Pi}_m^{\text{pooled},<l} = \text{Softmax}(\mathcal{R}(c_m^{\text{pooled},<l}))$.

Memory analysis. We need to store the latent representations and routing weights for each modality, with the total of 4. Each requires 2 bytes due to half precision. Thus, overall memory usage is $4\cdot(n')^2\cdot d\cdot M\cdot 2 = 8(n')^2 dM$. In DEMO, we set $M = 32$ and use DINOv2 ($d = 768$ and $n = 784$ after removing class token). Thus, overall memory usage is roughly 37 MB during CL throughout. We compare DEMO with other CL-based anomaly detection methods in terms of memory usage in Table 3.

B Theoretical Proofs

B.1 Proof of Proposition 1

Recall:

Proposition 1 (Non-entailment)**.** *Even if each expert E_k receives an identical number of tokens, it cannot be inferred that each modality m obtains non-zero effective normality reconstruction utility u_m.*

Proof. Consider two modalities and two experts E_1 and E_2. RGB and depth each have N tokens, and the capacity of each expert is N. Let the gains for all tokens be:

$$g_{ri1} = 1, \quad g_{ri2} = 1, \qquad g_{di1} = 1, \quad g_{di2} = 0. \tag{21}$$

Let the router assign all RGB tokens to expert E_1 and all depth tokens to expert E_2. Therefore, we ensure expert load balancing, *i.e.,* $\text{Load}(E_1) = \text{Load}(E_2) = N$. However, according to Eq. (10), we know that $u_r = 1$ and $u_d = 0$. In fact, if the router assigns all RGB tokens to expert E_2 and all depth tokens to expert E_1, the same capacity and load balancing will still be met, and $u_r = u_d = 1$. Therefore, Proposition 1 holds. □

Remark. The counterexample shows that an expert-load loss can prevent collapse in expert utilization without guaranteeing that tokens are dispatched to experts that benefit their modality. DEMO therefore augments capacity control with a modality-wise minimum-service mechanism.

B.2 PROOF OF THEOREM 2

Before proving Theorem 2, we first provide the following preliminaries.

Debt-aware DPP Routing. To expose the exact DPP form, we define the bounded routing cost:

$$c_t(\Pi) = C_t - \sum_{m,i,k} \Pi_{mik} b^t_{mik} - \frac{\tau}{V} H(\Pi), \tag{22}$$

where C_t is a constant independent of Π and is chosen to make the cost non-negative without altering the optimal solution, and $V = \frac{1}{\beta}$. In Eq. (12), because the base hindsight scores are clipped, the feasible set is compact, and entropy is bounded on that set, there exist finite constants such that

$$c_{\min} \le c_t(\Pi) \le c_{\max}. \tag{23}$$

Then the ideal debt-aware routing solves

$$\Pi^*_t \in \arg\min_{\Pi \in \mathcal{P}_t} \left\{ V c_t(\Pi) - \sum_m Q_m(t) u^t_m(\Pi) \right\}. \tag{24}$$

where $\mathcal{P}_t$ is the feasible set in Eq. (12) and $V > 0$ controls the trade-off between base hindsight score and debt. Specifically, since $u^t_m(\Pi) = \sum_{i,k} \Pi_{mik} \bar{g}^t_{mik}$, maximizing Eq. (12) is exactly equivalent to minimizing the above formula (Eq. (24)). The post-hoc routing teacher is computed with Sinkhorn (Cuturi, 2013), while the deployed router imitates the teacher and use hard routing. We therefore explicitly assume a δ_t-approximate deployed decision: for every $\Pi \in \mathcal{P}_t$,

$$V c_t(\Pi^t) - \sum_m Q_m(t) u^t_m(\Pi^t) \le V c_t(\Pi) - \sum_m Q_m(t) u^t_m(\Pi) + \delta_t, \tag{25}$$

with $0 \le \delta_t \le \delta$. We assume that the deployed decision Π^t is feasible, *i.e.*, $\Pi^t \in \mathcal{P}_t$, and satisfies the δ_t-approximate DPP inequality, as the KL loss alone does not guarantee either condition. It accounts for finite Sinkhorn iterations, teacher-router approximation, and hard-routing discretization.

Remark. The boundedness assumption Eq. (23) and the general primal-dual structure of Eq. (24) follow the standard DPP method (Tassiulas & Ephremides, 1992; Neely, 2010). The virtual-queue Lyapunov DPP argument used below follows Neely (2010). These references supply the generic framework. We specialize the framework to modality service by defining virtual service as the normalized reconstruction utility under expert-capacity constraints.

Definition 3 (Lyapunov function and one-step drift)**.** *Define the quadratic Lyapunov function as:*

$$L(Q(t)) = \frac{1}{2} \sum_{m \in \mathcal{M}} Q_m(t)^2. \tag{26}$$

The corresponding one-step drift is denoted by:

$$\Delta_t = L(Q(t+1)) - L(Q(t)). \tag{27}$$

Lemma 4 (One-step drift upper bound)**.** *For any feasible routing Π^t, we have*

$$\Delta_t \le B + \sum_m Q_m(t) \big(a^t_m \gamma_m - u^t_m(\Pi^t) \big), \tag{28}$$

where

$$B = \frac{1}{2} \sum_m \max\{\gamma_m^2, (1-\gamma_m)^2\} \le \frac{|\mathcal{M}|}{2}. \tag{29}$$

Proof. Let $h_m(t) = a_m^t \gamma_m - u_m^t(\Pi^t)$. Since $a_m^t \in \{0,1\}$, $u_m^t \in [0,1]$, and $u_m^t = 0$ whenever $a_m^t = 0$, we have

$$h_m(t) \in [\gamma_m - 1, \gamma_m].$$

We use $[x]_+^2 \le x^2$ and the queue update (Eq. (11)),

$$Q_m(t+1)^2 = [Q_m(t) + h_m(t)]_+^2 \tag{30}$$
$$\le [Q_m(t) + h_m(t)]^2 \tag{31}$$
$$= Q_m(t)^2 + h_m(t)^2 + 2Q_m(t)h_m(t). \tag{32}$$

Subtracting $Q_m(t)^2$ from both sides, summing over m, and dividing by 2, together with

$$h_m(t)^2 \le \max\{\gamma_m^2, (1-\gamma_m)^2\},$$

yields Eq. (28). Thus, Lemma 4 holds. □

Lemma 5 (Approximate DPP comparison inequality)**.** *If Eq. (25) holds, then for any comparison routing $\hat{\Pi}^t \in \mathcal{P}_t$,*

$$\Delta_t + Vc_t(\Pi^t) \le B + Vc_t(\hat{\Pi}^t) + \sum_m Q_m(t)\big(a_m^t\gamma_m - u_m^t(\hat{\Pi}^t)\big) + \delta_t. \tag{33}$$

Proof. Adding $Vc_t(\Pi^t)$ to both sides of Eq. (28) and then substituting Eq. (25) into the right-hand side, Lemma 5 holds. □

Assumption 6 (Stationary stochastic environment and feasibility)**.** *The exogenous batch state is stationary, independent of the current debt queue, and observed before routing. There exists a stationary randomized comparison policy $\pi^\star$ satisfying the following conditional feasibility property for every queue state:*

$$\mathbb{E}[u_m^t(\pi^\star) \mid Q(t)] \ge \gamma_m \mathbb{E}[a_m^t \mid Q(t)], \qquad \forall m, \tag{34}$$

and among all stationary policies satisfying Eq. (34), its expected cost is denoted by $c^\star$.

Assumption 7 (Slater margin)**.** *There exists another stationary randomized comparison policy π^{S} and constants $\epsilon > 0$ and $p_{\min} > 0$ such that, for every queue state,*

$$\mathbb{E}[u_m^t(\pi^{\mathrm{S}}) \mid Q(t)] \ge (\gamma_m + \epsilon)\mathbb{E}[a_m^t \mid Q(t)], \qquad \mathbb{E}[a_m^t \mid Q(t)] \ge p_{\min}, \quad \forall m, \tag{35}$$

and its expected cost is c^{S}. Equivalently, serviceable batches occur with non-vanishing frequency and the policy supplies a strict service margin on them.

Remark. Note that the Slater condition requires expert capacities and the utility structure to allow every modality to receive service above its minimum objective whenever serviceable demand occurs. If a modality is never serviceable, or the capacities are physically insufficient, then no algorithm can provide this starvation-free guarantee.

Theorem 8 (Cost optimality and queue stability)**.** *Under $u_m^t(\Pi) \in [0,1]$, Eq. (23), Eq. (25), and Assumptions 6 and 7,* DEMO *satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[c_t(\Pi^t)] \le c^\star + \frac{B+\delta}{V} + \frac{\mathbb{E}[L(Q(0))]}{VT}, \tag{36}$$

and

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_m \mathbb{E}[Q_m(t)] \le \frac{B + V(c^{\mathrm{S}} - c_{\min}) + \delta}{\epsilon p_{\min}} + \frac{\mathbb{E}[L(Q(0))]}{\epsilon p_{\min} T}. \tag{37}$$

Consequently, the cost gap to the best feasible stationary policy is $O(1/V)$, and the average queue length is $O(V)$.

Proof. Taking conditional expectations in Lemma 5 and choosing the comparison policy as $\pi^\star$, from Eq. (34) we have:

$$\mathbb{E}\left[\sum_m Q_m(t)\big(a_m^t\gamma_m - u_m^t(\pi^\star)\big) \mid Q(t)\right] \le 0.$$

Hence

$$\mathbb{E}[\Delta_t + Vc_t(\Pi^t)] \le B + Vc^\star + \delta. \tag{38}$$

Summing over $t = 0, \dots, T-1$, the drift terms telescope:

$$\mathbb{E}[L(Q(T))] - \mathbb{E}[L(Q(0))] + V\sum_{t=0}^{T-1}\mathbb{E}[c_t(\Pi^t)] \le T(B+\delta) + VTc^\star.$$

Using $L(Q(T)) \ge 0$ and dividing by VT yields Eq. (36).

Now choose the comparison policy in Lemma 5 as π^{S}. From Eq. (35) we have:

$$\mathbb{E}[\Delta_t + Vc_t(\Pi^t) \mid Q(t)] \le B + Vc^{\mathrm{S}} + \delta - \epsilon\sum_m Q_m(t)\mathbb{E}[a_m^t \mid Q(t)] \le B + Vc^{\mathrm{S}} + \delta - \epsilon p_{\min}\sum_m Q_m(t).$$

Since $c_t(\Pi^t) \ge c_{\min}$, we obtain

$$\mathbb{E}[\Delta_t] \le B + V(c^{\mathrm{S}} - c_{\min}) + \delta - \epsilon p_{\min}\sum_m \mathbb{E}[Q_m(t)]. \tag{39}$$

Summing over t and using $L(Q(T)) \ge 0$, rearranging and dividing by $\epsilon p_{\min} T$ gives Eq. (37). □

Recall:

Theorem 2 (Starvation-free modality service)**.** *Assume a stationary stochastic routing environment, bounded normalized utility and routing cost, a feasible δ-approximate debt-aware DPP router, and a strictly feasible stationary policy. If serviceable batches occur with non-vanishing frequency, then the modality-debt queues are strongly stable and mean-rate stable, and every modality satisfies*

$$\liminf_{T\to\infty} \frac{\sum_{t=0}^{T-1}\mathbb{E}[u_m^t(\Pi^t)]}{\sum_{t=0}^{T-1}\mathbb{E}[a_m^t]} \ge \gamma_m > 0. \tag{40}$$

When every batch is serviceable ($a_m^t = 1$), this reduces to $\liminf_{T\to\infty} T^{-1}\sum_{t<T}\mathbb{E}[u_m^t] \ge \gamma_m$.

Proof. Since $[x]_+ \ge x$, from Eq. (11) we have:

$$Q_m(t+1) \ge Q_m(t) + a_m^t\gamma_m - u_m^t(\Pi^t).$$

Summing over $t = 0, \dots, T-1$ yields

$$\sum_{t=0}^{T-1} u_m^t(\Pi^t) \ge \gamma_m\sum_{t=0}^{T-1} a_m^t + Q_m(0) - Q_m(T). \tag{41}$$

We now prove $\mathbb{E}[Q_m(T)]/T \to 0$. Eq. (37) implies strong stability of the queue, *i.e.*, its time-average expectation has a finite upper bound. On the other hand, since $u_m^t, \gamma_m \in [0,1]$, the absolute change of the queue between two consecutive steps is at most 1. If there exist $\rho > 0$ and a subsequence T_j such that $\mathbb{E}[Q_m(T_j)] \ge \rho T_j$, then for every

$$t' \in [T_j - \lfloor \rho T_j/2 \rfloor,\ T_j],$$

we have

$$\mathbb{E}[Q_m(t')] \ge \mathbb{E}[Q_m(T_j)] - (T_j - t') \ge \rho T_j/2.$$

Therefore, the time-average of the queue up to T_j is at least a quantity that grows linearly with T_j, contradicting Eq. (37). Hence

$$\frac{\mathbb{E}[Q_m(T)]}{T} \to 0.$$

Let $A_m(T) = \sum_{t=0}^{T-1} \mathbb{E}[a_m^t]$. By the Slater assumption, $A_m(T) \geq p_{\min} T$. Taking expectations in Eq. (41) and dividing by $A_m(T)$, we have

$$\frac{\sum_{t=0}^{T-1} \mathbb{E}[u_m^t(\Pi^t)]}{A_m(T)} \geq \gamma_m + \frac{Q_m(0) - \mathbb{E}[Q_m(T)]}{A_m(T)}.$$

Because $A_m(T) \geq p_{\min} T$ and $\mathbb{E}[Q_m(T)]/T \to 0$, the final term vanishes. Taking the limit inferior yields Eq. (40). □

C Complete Experimental Setup

Datasets. Following recent works (Long et al., 2026), two MAD datasets were introduced: MVTec 3D (Bergmann et al., 2022) and Eyecandies (Bonfiglioli et al., 2022). MVTec 3D is the first MAD benchmark, which provides high-resolution RGB images together with precise 3D point cloud data. Eyecandies is a synthetically rendered dataset that simulates a candy production line, offering RGB, depth, and normal maps under multiple lighting conditions.

Protocols. We introduced different MAD settings, including multiclass MAD (setting 1) and three CMAD settings: 6 – 4 with 1 step (setting 2), 6 – 1 with 4 steps (setting 3), and 1 – 1 with 9 steps (setting 4). In CL, "6 – 1 with 4 steps" means that the model is first trained on 6 base classes from the dataset, followed by incrementally introducing the remaining classes over 4 steps, namely one-step-one-class.

Evaluation metrics. We adopted two widely used metrics: image-level Area Under the Receiver Operation Characteristic (AUROC) and pixel-level Per-Region Overlap (PRO). To evaluate CL ability, the forgetting measure (FM) (Tang et al., 2024) was considered. We measured Average image-level AUROC and pixel-level PRO for all continual training steps:

$$\text{Average} = \begin{cases} \frac{1}{L} \sum_{l=1}^{L} \text{AUROC}_{L,l}^{\text{img}} \\ \frac{1}{L} \sum_{l=1}^{L} \text{PRO}_{L,l}^{\text{pix}} \end{cases}, \tag{42}$$

where $\text{AUROC}_{L,l}$ denotes that the performance of the model in l-th step after learning in L-th step. For FM metric, it is defined as follows:

$$\text{FM} = \begin{cases} \frac{1}{L-1} \sum_{l=1}^{L-1} \max_{L' \in \{1,\cdots,L-1\}} (\text{AUROC}_{L',l}^{\text{img}} - \text{AUROC}_{L,l}^{\text{img}}) \\ \frac{1}{L-1} \sum_{l=1}^{L-1} \max_{L' \in \{1,\cdots,L-1\}} (\text{PRO}_{L',l}^{\text{pix}} - \text{PRO}_{L,l}^{\text{pix}}) \end{cases}. \tag{43}$$

Implementation details. In DEMO, we employed ViT-B/14 pre-trained by DINOv2 (Oquab et al., 2023) as the feature extractor. The router and each residual expert consist of 2 simple linear layers with a layer normalization, while the anchor expert consists of 8 linear layers. The optimizer was based on StableAdamW (Wortsman et al., 2023) with weight decay of $1e$-4. The learning rate of our DEMO was set to $2e$-3 and reduced to $2e$-4 in CL tasks. We trained DEMO with a batch size of 12 for $10,000$ iterations. We initialized $Q(0)$ to zero. The main hyper-parameters, K_e, γ_m, and M, were set to 4, 0.8, and 32, respectively. The balancing hyper-parameters, λ_1, λ_2, τ, and β, were set to 0.5, 0.05, 0.2, and 2, respectively. We resized the input image to 448×448 and centre-cropped it to 392×392. To clearly understand the training process of DEMO, we provide the pseudo-code in Algorithm 1.

Baselines. We compared DEMO against several representative methods, including IUF (Tang et al., 2024), CDAD (Li et al., 2025b), IB-IUMAD (Long et al., 2026)), CFM (Costanzino et al., 2024), and AnomalyMoE (Gu et al., 2026). Both CFM and AnomalyMoE do not have the ability of CL, so we introduced EWC (Kirkpatrick et al., 2017) into them to mitigate forgetting. For fair comparison, their encoder or feature extractor was uniformly replaced with DINOv2 (Oquab et al., 2023) and we strictly followed their original configurations to reproduce the results under our protocols. Below, we briefly introduce these methods.

- CFM (Costanzino et al., 2024) is a representative MAD method, leveraging complementary information from RGB and 3D modalities to improve detection performance. By aligning and fusing features across modalities, it effectively captures both surface and geometric anomalies.

Algorithm 1 Pseudo-code for DEMO training

Require: RGB and depth features z_m ($m \in \{r, d\}$) extracted from pre-trained ViT
Require: Flags: update_debt, feature_mask, modality_dropout
Set flags as False during replay.
$z_{r\to d} \leftarrow \phi(z_r)$
$z_{d\to r} \leftarrow \phi(z_d)$
if modality_dropout is True **then**
    Randomly zero $z_{d\to r}$ and $z_{r\to d}$.
end if
if feature_mask is True **then**
    $\tilde{z}_m \leftarrow \mathbf{M}_{m,mask}(z_m)$
end if
$c_r \leftarrow \tilde{z}_r + z_{d\to r}$
$c_d \leftarrow \tilde{z}_d + z_{r\to d}$
$\hat{z}_m^{(0)} \leftarrow E_0(c_m)$
for each modality m **do**
    $(\texttt{Logits}_m, \Pi_m) \leftarrow \mathcal{R}(c_m)$
    for each expert k **do**
        $\hat{z}_m^{(k)} \leftarrow \hat{z}_m^{(0)} + E_k(c_m)$
        Compute cosine error and normality reconstruction gain g_{mik}.
    end for
    Compute attainable gain G_m, demand mask a_m, and normalized gain $\bar{g}_m$.
end for
for each modality m **do**
    $u_m \leftarrow \left(\sum_{i,k} \Pi_{mik}\bar{g}_{mik}\right)$
end for
if update_debt is True **then**
    Update modality debt Q_m using u_m and a_m.
end if
$\Pi^* \leftarrow \text{Sinkhorn}(\texttt{Logits}_{mik}, g_{mik}, Q_m, \bar{g}_{mik})$
Optimize DEMO via Eq. 13.

- AnomalyMoE (Gu et al., 2026) is a MoE-based multiclass MAD method, which reconceptualizes the complex AD task by decomposing it into a three-level semantic hierarchy. In particular, each hierarchy is managed by different MoE models, which are respectively responsible for local structural anomalies, component-level semantic anomalies, and global logical anomalies.
- IUF (Tang et al., 2024), one of the early representative continual anomaly detection (CAD) methods, employs an object-aware self-attention to partition the semantic space of different objects and uses a semantic compression loss to reserve feature capacity for new objects. Besides, it further adopts a novel parameter update strategy to protect old knowledge.
- CDAD (Li et al., 2025b) is a recent top-performaing CAD method. It adapts pre-trained diffusion models to CL tasks by projecting the learning gradient of new tasks onto the orthogonal direction of that space and efficiently computing the important feature space of old tasks via iterative Singular Value Decomposition.
- IB-IUMAD (Long et al., 2026) is the first CMAD method, extending IUF to a multimodal framework, which mitigates forgetting by reducing spurious and redundant features from an information bottleneck perspective. It also introduces a Mamba decoder (Pei et al., 2025) to decouple features.

D SUPPLEMENTARY EXPERIMENTAL RESULTS

D.1 ADDITIONAL ANALYSES

Computational efficiency analysis of DEMO and AnomalyMoE. Here we compare DEMO with the recent MoE-based MAD method, AnomalyMoE (Gu et al., 2026). AnomalyMoE focuses on dissecting the complex anomaly detection task into multiple clear sub-problems but does not consider the increase in parameter count. As presented in Table 5, AnomalyMoE suffers from an excessive number of trainable parameters (over 180 M) and slow inference speed (18.14 fps), as it uses different networks (e.g., convolution and multilayer perception) as experts in MoE. In contrast, DEMO requires fewer parameters and achieves faster speed, gaining over a $3\times$ shrinkage in parameters and a $2.5\times$ increase in frame rate compared with AnomalyMoE.

Table 5: Computational efficiency analysis of DEMO and AnomalyMoE.

Method	Trainable params. (M)	Frame rate (fps)
AnomalyMoE	> 180	18.14
DEMO	**61.5**	**46.99**

How does modality competition affect localization. We analyse the impact of modality competition on final multimodal anomaly localization. The results obtained on the Eyecandies dataset are provided in Figure 7. When no intervention is applied to modality competition, one modality is often suppressed during training, which renders its information unavailable, resulting in incomplete normality modelling on that modality. This leads to the failure of final multimodal anomaly localization, particularly when anomalies resemble normal features (see the 1st row of Figure 7).

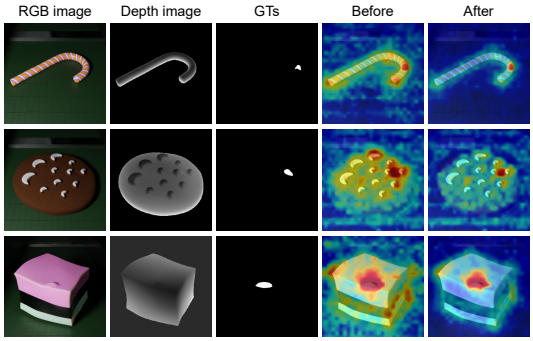


Figure 7: The impact of modality competition on multimodal anomaly localization. Before: without intervention; After: with intervention.

Analysis on modality competition across steps. Here we additionally present the situation of modality competition during CL. The results yielded in setting 2 of MVTec 3D dataset are illustrated in Figure 8. As expected, before intervention, the RGB modality consistently suppresses the depth modality, leaving a large gap between the availability of modality information. However, this issue is effectively mitigated after intervention, as shown in Figure 8(b).

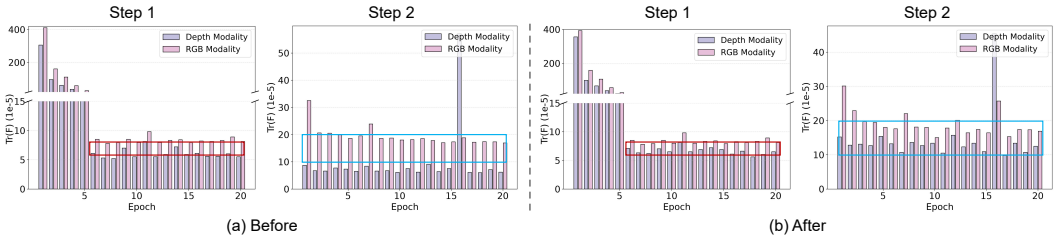


Figure 8: Analysis on the availability of modality information in the setting 2 of MVTec 3D dataset.

Analysis on modality service across steps. To further evaluate the modality service during CL, Figure 9 shows the results in setting 3 of MVTec 3D dataset. As recorded, despite in long training steps, the average expectation of modality services for DEMO is still larger than minimal service objective, which guarantees the feasibility of long-term multimodal learning, leading to promising CMAD performance.

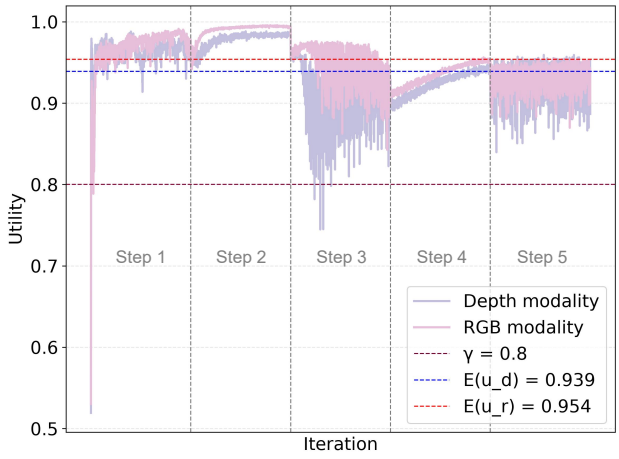


Figure 9: Modality service analysis in the setting 3 of MVTec 3D dataset.

Balancing hyper-parameter analysis. We analyse four balancing hyper-parameters, λ_1, λ_2, τ, and β, in setting 2 and 3 of MVec 3D dataset, as illustrated in Figure 10. As recorded, large values (e.g., 1) for the first three hyper-parameters may lead to significant performance degradation, with a decline of roughly 2%. Different small weights yield comparably promising results, such as 0.05 and 0.5 for λ_1, and 0.02 and 0.2 for τ. After comprehensive comparison, we set λ_1 and τ to 0.5 and 0.2, respectively. For λ_2, it is clear that the best results can achieved when we set it to 0.05. β controls the base hindsight score and debt. Conversely, larger values for it can result in better performance. From the Figure 10, the most reasonable value is 2.

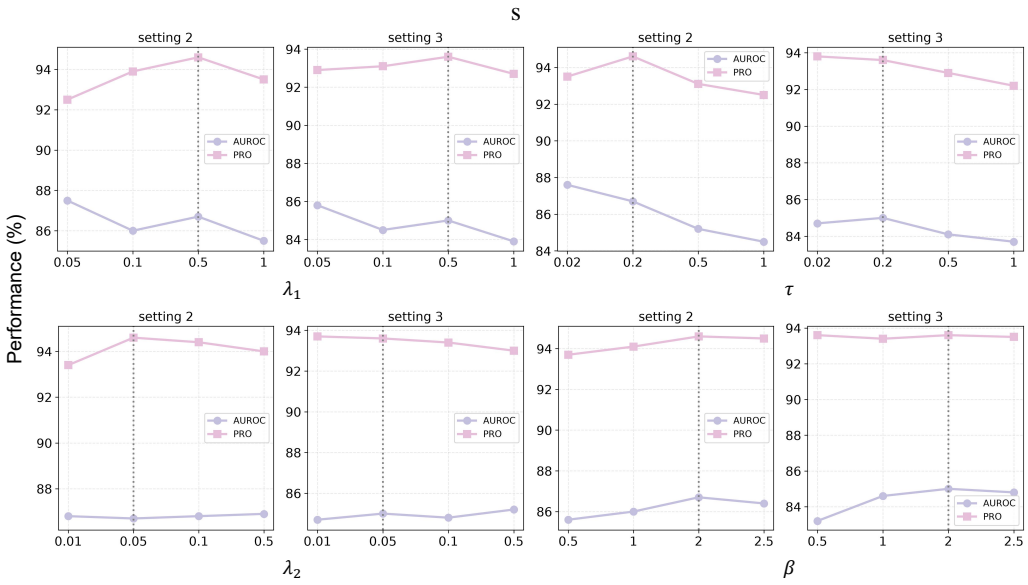


Figure 10: Sensitivity analysis of λ_1, λ_2, τ, and β in setting 2 and 3 of MVTec 3D dataset.

D.2 VISUALIZATIONS

In this section, we provide additional multimodal anomaly localization results in setting 1 and 2 on Eyecandies dataset, as illustrated in Figure 11 and 12.

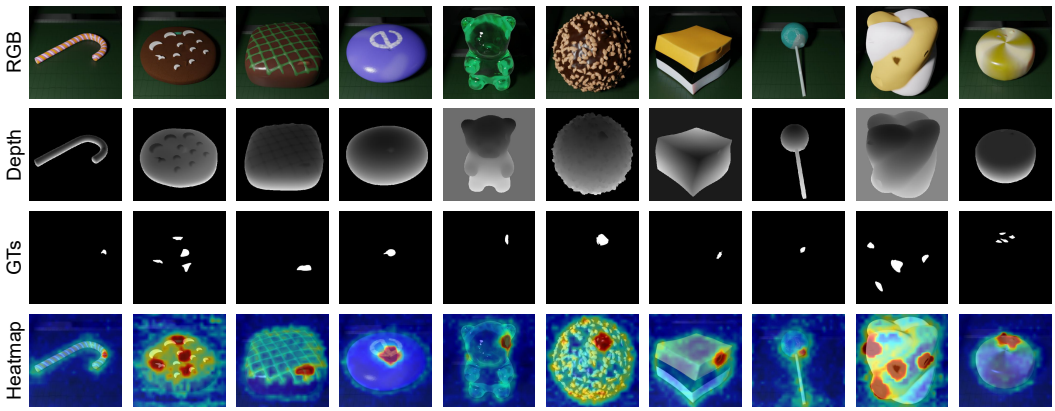


Figure 11: Visualization results in setting 1 of Eyecandies dataset.

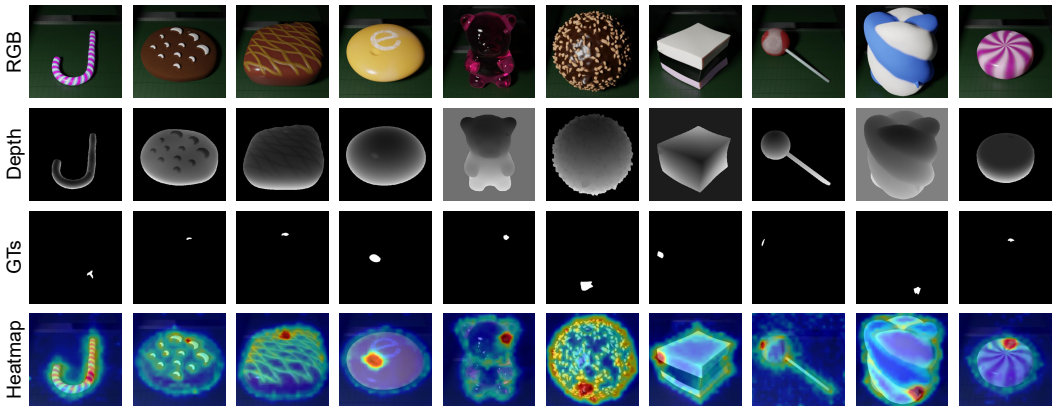


Figure 12: Visualization results in setting 2 of Eyecandies dataset.