AnomalyMoE: Towards a Language-free Generalist Model for Unified Visual Anomaly Detection

Zhaopeng Gu[1,2], **Bingke Zhu**[1,2*], **Guibo Zhu**[1,2*], **Yingying Chen**[1,2], **Wei Ge**[1,2], **Ming Tang**[1,2], **Jinqiao Wang**[1,2,3]

[1]Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]Objecteye Inc., Beijing, China
{guzhaopeng2023, gewei2023}@ia.ac.cn, {bingke.zhu, gbzhu, yingying.chen, tangm, jqwang}@nlpr.ia.ac.cn

Abstract

Anomaly detection is a critical task across numerous domains and modalities, yet existing methods are often highly specialized, limiting their generalizability. These specialized models, tailored for specific anomaly types like textural defects or logical errors, typically exhibit limited performance when deployed outside their designated contexts. To overcome this limitation, we propose AnomalyMoE, a novel and universal anomaly detection framework based on a Mixture-of-Experts (MoE) architecture. Our key insight is to decompose the complex anomaly detection problem into three distinct semantic hierarchies: local structural anomalies, component-level semantic anomalies, and global logical anomalies. AnomalyMoE correspondingly employs three dedicated expert networks at the patch, component, and global levels, and is specialized in reconstructing features and identifying deviations at its designated semantic level. This hierarchical design allows a single model to concurrently understand and detect a wide spectrum of anomalies. Furthermore, we introduce an Expert Information Repulsion (EIR) module to promote expert diversity and an Expert Selection Balancing (ESB) module to ensure the comprehensive utilization of all experts. Experiments on 8 challenging datasets spanning industrial imaging, 3D point clouds, medical imaging, video surveillance, and logical anomaly detection demonstrate that AnomalyMoE establishes new state-of-the-art performance, significantly outperforming specialized methods in their respective domains.

Introduction

Anomaly detection aims to identify anomalous samples that deviate from normal patterns. It is widely applied across various fields of production and daily life, including industrial defect detection (Gu et al. 2024b; Zhu et al. 2024), logical anomaly detection (Zhang et al. 2025; Liu et al. 2023), medical image diagnosis (Bao et al. 2024; Huang et al. 2024), and video surveillance (Yang et al. 2024), etc. Spanning multiple modalities such as images (Gu et al. 2025b), videos (Yang et al. 2024), and 3D point clouds (Ye et al. 2025), it represents a comprehensive, cross-domain, and cross-modal task with significant practical value, making it a prominent research topic in both academia and industry.

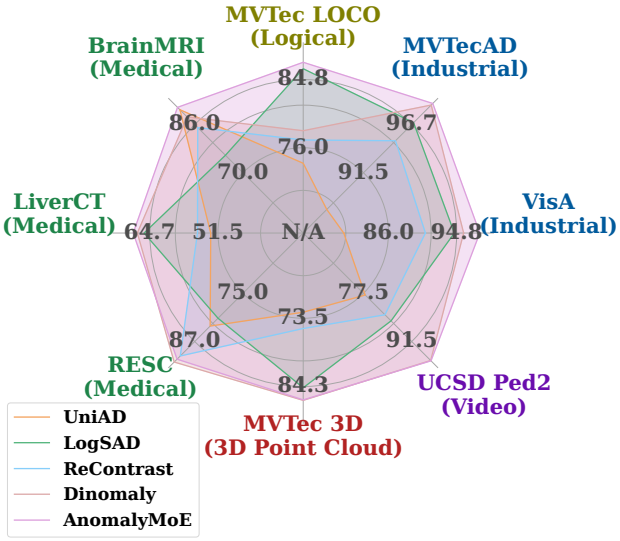

Figure 1: Performance comparison of AnomalyMoE with state-of-the-art anomaly detection methods.

As illustrated in Figure 2(a), existing methods are often highly specialized for particular domains and modalities (Zhu et al. 2024; Gu et al. 2024a). This specialization, while effective for specific tasks, inherently limits their generalizability. To address this, recent efforts have explored unified anomaly detection frameworks. However, these approaches still exhibit critical limitations. Some are confined to patch-level reconstruction, effectively capturing structural flaws but failing to comprehend higher-level logical anomalies. Conversely, others that excel at logical anomaly detection (Gu et al. 2025b; Zhang et al. 2025), heavily rely on component segmentation, rendering them incapable of identifying issues like missing components or abnormal arrangements, as shown in Figure 2(b). Furthermore, these advanced models increasingly depend on large vision-language or language models to interpret component semantics, introducing substantial computational overhead and a reliance on non-visual priors. This complexity poses a barrier to deployment and may not be optimal for purely visual tasks.

To address these issues, we propose AnomalyMoE, a unified and language-free framework that reconceptualizes the

*Corresponding authors.

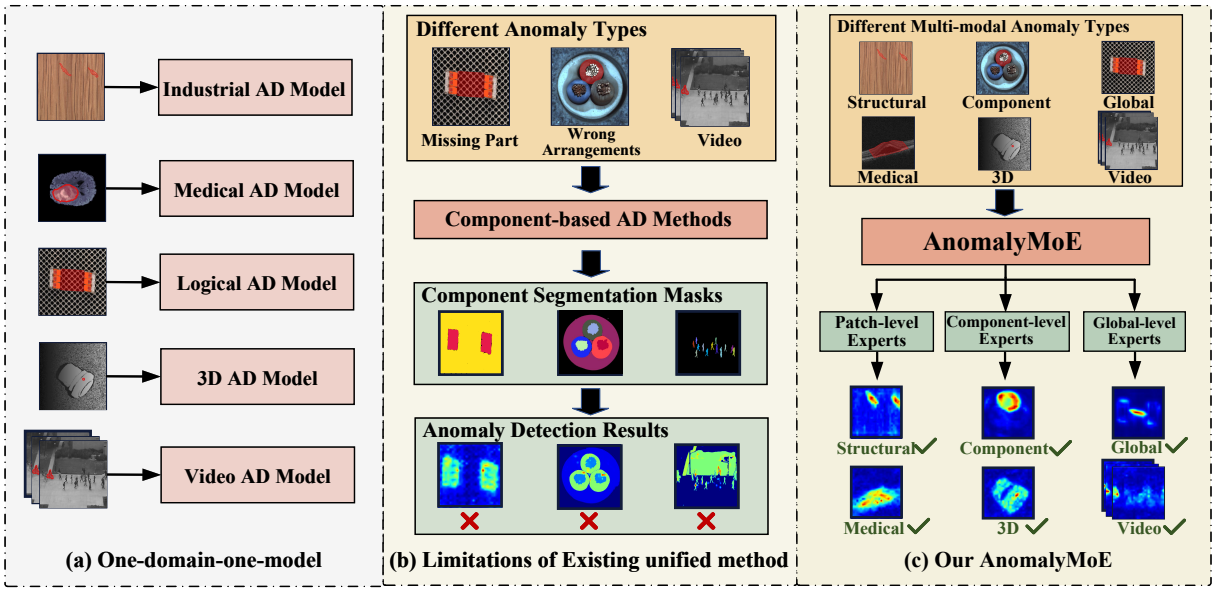

Figure 2: Illustration of the motivation for AnomalyMoE. (a) The conventional one-domain-one-model paradigm, where specialized models are not generalizable. (b) Existing unified methods that typically rely on component-level analysis, failing to detect global or compositional errors (e.g., missing components or incorrect assembly). (c) Our proposed AnomalyMoE, a single, universal model featuring a hierarchical Mixture of Experts. By dedicating experts to patch, component, and global semantic levels, it achieves comprehensive anomaly detection across diverse visual anomaly scenarios.

Mixture-of-Experts (MoE) architecture for anomaly detection. Its core idea is to decompose the task into a hierarchy of semantic sub-problems. As illustrated in Figure 2(c), this allows a single model with dedicated experts for local, component, and global semantics to comprehensively handle diverse anomaly types, directly addressing the shortcomings of prior work with superior performance and efficiency.

Specifically, we first categorize "anomalies" into three semantic levels: 1) Local structural anomalies, which manifest as fine-grained textural or pixel-level deviations (e.g., surface defects); 2) Component-level semantic anomalies, which appear as errors in an object part or region (e.g., incorrect components on a circuit board); and 3) Global logical anomalies, characterized by errors in global semantics (e.g., the incorrect assembly of correct components or abnormal behaviors in video surveillance). To detect these levels, AnomalyMoE employs three corresponding expert networks operating on a feature reconstruction paradigm. A Transformer-based (Vaswani et al. 2017) decoder expert targets fine-grained local details, while two autoencoder-based (Minhas and Zelek 2020) experts learn holistic representations for component and global semantics. Trained exclusively on normal samples, AnomalyMoE identifies anomalies via reconstruction errors. To ensure the experts are both diverse and fully utilized, we further introduce an Expert Information Repulsion (EIR) module to maximize their specialization and an Expert Selection Balancing (ESB) module to prevent model collapse. This integrated design not only enables targeted detection at each semantic level but also allows their features to mutually enhance one another, leading to robust and superior performance.

We conduct experiments on 8 datasets across diverse domains, including industrial imaging (Bergmann et al. 2021a), industrial 3D (Bergmann et al. 2021b), medical imaging (Bao et al. 2024), video surveillance (Wang and Miao 2010), and logical anomaly detection (Bergmann et al. 2022). The results demonstrate that AnomalyMoE achieves top performance in anomaly detection across multiple domains. Furthermore, the experts at different semantic levels successfully detect their corresponding types of anomalies, showcasing the comprehensive capability of AnomalyMoE.

Our contributions can be summarized as follows:

- We propose AnomalyMoE, a language-free, generalist framework for visual anomaly detection, significantly advancing the field of unified anomaly detection with its MoE architecture.
- AnomalyMoE's three-level experts interpret test samples from different semantic levels to accurately detect various types of anomalies. The Expert Information Repulsion and Expert Selection Balancing modules further enhance the diversity and balance among the experts.
- Extensive experiments on 8 datasets across domains such as industrial image, industrial 3D, medical image, video surveillance, and logical AD demonstrate that AnomalyMoE surpasses specialized methods in each respective field, achieving state-of-the-art performance.

Related Work

Anomaly Detection

Anomaly detection is a comprehensive task spanning multiple modalities and domains (Bergmann et al. 2021a,b; Wang

and Miao 2010). However, the nature of anomalies, ranging from industrial defects and logical errors to medical pathologies and abnormal behaviors, varies drastically in features and semantic levels. This diversity has led to a fragmented research landscape following a "one domain, one method" paradigm. For instance, industrial vision methods focus on patch-level feature analysis (Gu et al. 2024a, 2025a; Defard et al. 2021; Wang et al. 2025) for local structural defects, while logical anomaly detection relies on component-based segmentation (Liu et al. 2023; Hsieh and Lai 2024; Kim et al. 2024) to identify relational errors. Similarly, video surveillance uses sequence models like RNNs to capture temporal deviations (Naji et al. 2022), and other domains like medical imaging or 3D point clouds employ specialized techniques targeting semantic or geometric properties (Liang et al. 2025; Zhou et al. 2024). This specialization has created a landscape of incompatible specialized tools, hindering the development of universal anomaly detection.

The pursuit of "a single model for all anomalies" is a key research frontier. UniAD (You et al. 2022) is an early pioneer, using a Transformer model with learnable queries to address the "identity mapping" problem for multi-class detection. ReContrast (Guo et al. 2023) and UniNet (Wei, Jiang, and Xu 2025) integrate contrastive learning and domain adaptation, while Dinomaly (Guo et al. 2025) proposes a simpler reconstruction network. However, these methods, being confined to patch-level reconstruction, primarily excel at industrial defects. To bridge this gap, UniVAD (Gu et al. 2025b) and LogSAD (Zhang et al. 2025) augment their frameworks with a component-level branch. Specifically, these methods leverage large vision-language models to interpret component semantics. Despite this advance, they face critical issues: a heavy reliance on segmentation accuracy, which fails on "unsegmentable" anomalies, and a lack of a global perspective to assess overall assembly. Moreover, their dependence on large language models introduces significant computational overhead. Our efficient, language-free approach overcomes these limitations while delivering superior performance.

In summary, existing unified models are either limited to local details or dependent on flawed, non-global analysis. They fail to provide a framework covering the three hierarchical levels of local structure, component semantics, and global logic. This research gap is the core motivation for our proposed AnomalyMoE.

Mixture-of-Experts

The Mixture-of-Experts (MoE) is a conditional computation architecture featuring a gating network and multiple expert networks (Shazeer et al. 2017). By sparsely activating experts based on the input, MoE enables massive model scaling with constant computational cost. GShard (Lepikhin et al. 2020) first combines MoE with the Transformer (Vaswani et al. 2017) architecture, enhancing multilingual machine translation by adding multiple parallel FFNs. Switch Transformer (Fedus, Zoph, and Shazeer 2022) integrates an MoE architecture into the T5 (Raffel et al. 2020) model, yielding a more powerful pretrained language model. V-MoE (Riquelme et al. 2021) brings this paradigm to the vision domain, leading to the training of the largest vision models to date. The first application in anomaly detection, Adapted-MoE (Lei et al. 2024), uses homogeneous experts to handle intra-class variations within datasets.

While existing MoE applications focus on model scaling, we reframe its purpose for effective problem decomposition. The value of MoE in our framework stems not from an increase in parameter count, but from its ability to dissect the complex anomaly detection task into distinct, semantically clear sub-problems. We assign specialized, heterogeneous experts to each sub-task and introduce Expert Information Repulsion and Expert Selection Balancing modules to govern their collaboration. This approach re-purposes the MoE architecture, transforming it from a tool for achieving greater model capacity into a sophisticated framework that performs hierarchical and multi-level semantic analysis.

Method

As illustrated in Figure 3, the core of AnomalyMoE is a hierarchical Mixture-of-Experts architecture. An input sample is first processed by a frozen, pre-trained visual encoder to extract patch embeddings and a global `[cls]` embedding. The `[cls]` embedding then directs a trainable router to orchestrate a parallel ensemble of three heterogeneous expert groups, each tailored to a distinct semantic hierarchy: local structure, component semantics, and global logic. The training process is optimized by two novel auxiliary modules, Expert Information Repulsion (EIR) and Expert Selection Balancing (ESB), which ensure functional diversity and balanced utilization. We will first detail this novel mechanism, followed by the specific implementations of the expert groups, and conclude with the final training objective.

Expert Routing and Collaboration Mechanism

To effectively manage expert networks and encourage them to learn diverse representations, we design a sophisticated expert routing and collaboration mechanism. At its core is a dynamic routing module, supplemented by the Expert Selection Balancing module and the Expert Information Repulsion module. These components work in synergy to address common challenges in MoE architectures, such as expert functional redundancy and training instability.

Top-K Sparse Expert Routing The Router module is a lightweight, trainable feed-forward network that dynamically allocates weights to experts based on the global `[cls]` embedding, denoted as $E_{[\text{cls}]}$. To ensure efficiency, we employ a Top-K sparse routing mechanism. The router first generates a vector of scores, or logits, for all experts by applying a learnable weight matrix $\mathbf{W}_{\text{gate}}$. We then activate only the K experts with the highest logit values (where K=3 in our implementation). The final gating weights are computed by applying a Softmax function exclusively to the logits of these selected experts. This process is formalized as:

$$\text{Indices}_{\text{topK}} = \text{TopK}(\mathbf{W}_{\text{gate}} \cdot E_{[\text{cls}]}, K), \tag{1}$$

$$G_j = \begin{cases} \frac{\exp(\text{logit}_j)}{\sum_{k \in \text{Indices}_{\text{topK}}} \exp(\text{logit}_k)}, & \text{if } j \in \text{Indices}_{\text{topK}}, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

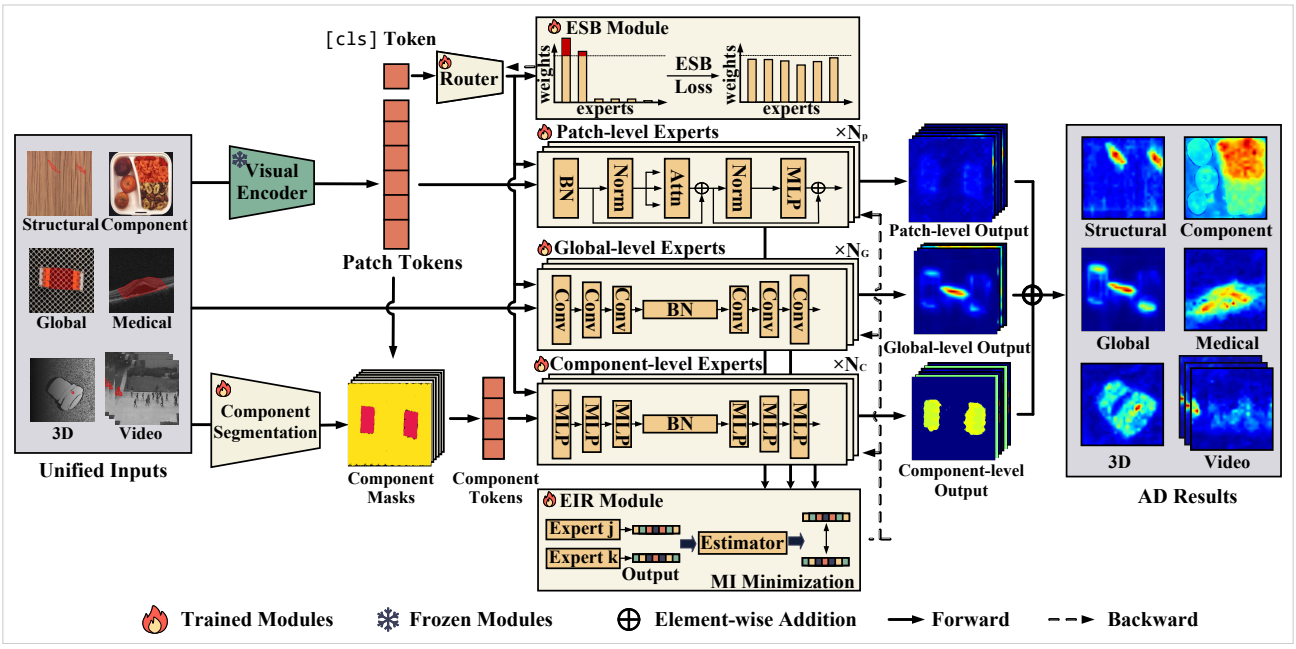

Figure 3: The overall architecture of AnomalyMoE. An input sample is processed by a frozen encoder to extract patch and [cls] embeddings. The [cls] token guides a router to activate a subset of three heterogeneous expert groups (Patch, Component, and Global level). Each expert produces an anomaly map, which are aggregated for the final result. The training is regularized by the Expert Selection Balancing (ESB) and Expert Information Repulsion (EIR) modules.

where logit_j is the score for the j-th expert, $\text{Indices}_{\text{topK}}$ contains the indices of the selected experts, and G_j is the final gating weight for the j-th expert. By activating only a small subset of experts per input, this approach significantly reduces both computational load and memory usage.

Expert Selection Balancing A common challenge in MoE training is router convergence to a state where only a few experts are consistently activated. To prevent this, our Expert Selection Balancing (ESB) module introduces a multi-part load balancing auxiliary loss, $\mathcal{L}_{\text{ESB}}$. For a training batch of size B, we define: 1) Expert Importance P_j: The average gating probability for the j-th expert, $P_j = \frac{1}{B}\sum_{b=1}^{B} G_{b,j}$. 2) Expert Count C_j: The number of samples assigned to the j-th expert. The total ESB loss is a weighted sum:

$$\mathcal{L}_{\text{ESB}} = \lambda_{\text{imp}}\mathcal{L}_{\text{importance}} + \lambda_{\text{load}}\mathcal{L}_{\text{load}} + \lambda_z\mathcal{L}_z. \quad (3)$$

Here, the importance loss, $\mathcal{L}_{\text{importance}} = N_{\text{exp}}\sum_j P_j^2$, encourages a uniform distribution of expert probabilities, where N_{exp} is the total number of experts. The load loss, $\mathcal{L}_{\text{load}} = \sum_j \max(C_j - \text{Capacity}, 0)^2$, enforces a hard capacity constraint by penalizing any expert whose count C_j exceeds a defined $Capacity$. Finally, the z-loss, $\mathcal{L}_z = \frac{1}{B}\sum_{b,j}(\text{logit}_{b,j})^2$, is a regularization term applied to the pre-softmax logits to enhance training stability. The terms λ_{imp}, λ_{load}, and λ_z are set to 1 in our experiments.

Expert Information Repulsion While ESB ensures expert utilization, the EIR module promotes their functional specialization by reducing redundancy. Inspired by information-theoretic methods (Zhang et al. 2024), EIR minimizes the Mutual Information (MI) between the output representations of different experts within the same group. Let Z_j and Z_k be random variables representing the output feature representations of the j-th and k-th experts, respectively. Our goal is to minimize $I(Z_j; Z_k)$.

Since direct MI computation is intractable, we employ the Contrastive Log-ratio Upper Bound (CLUB) (Cheng et al. 2020) to estimate and minimize this value. This is achieved using a variational network, $q_\theta(z_k|z_j)$, to approximate the conditional probability, where z_j and z_k are specific samples of the random variables Z_j and Z_k. The MI upper bound is then calculated via Monte Carlo estimation:

$$\begin{aligned} I_{\text{CLUB}}(Z_j; Z_k) \approx\ & \mathbb{E}_{p(z_j,z_k)}[\log q_\theta(z_k|z_j)] \\ & - \mathbb{E}_{p(z_j)p(z_k)}[\log q_\theta(z_k|z_j)]. \end{aligned} \quad (4)$$

The first term is the expected log-likelihood over matched sample pairs $(z_j^{(b)}, z_k^{(b)})$ from the batch, and the second is over non-matched pairs created by permutation. The total EIR loss, $\mathcal{L}_{\text{EIR}}$, is the sum of these CLUB estimates over all expert pairs within each group. By minimizing this loss, we push the representation spaces of experts apart, compelling them to learn distinct, complementary functions.

Hierarchical Expert Implementations

The effectiveness of our framework relies on three groups of heterogeneous experts, each designed to target a specific semantic level of anomalies.

Patch-level Experts This group of N_p experts identifies fine-grained structural anomalies. Their objective is to re-

construct a target feature map $F_{\text{target}} \in \mathbb{R}^{N \times D}$, which fuses multi-level information from the encoder. A key challenge is the "identity mapping" problem, where a powerful network can learn to perfectly copy its input, rendering it unable to detect anomalies. To address this, we first corrupt F_{target} with sequential Gaussian noise $\mathcal{N}(\cdot)$ and dropout $\mathcal{D}(\cdot; p)$ with a dropout rate of p, which is set to 0.2 in our experiments:

$$F_{\text{rec\_in}} = \mathcal{D}(\mathcal{N}(F_{\text{target}}); p). \tag{5}$$

Second, within each Transformer-based expert, we replace standard Softmax Attention with Linear Attention (Han et al. 2023; Guo et al. 2025). Conventional Softmax Attention can learn highly localized mappings where a query token attends sharply to its corresponding key, facilitating trivial copying. In contrast, Linear Attention promotes a more diffuse attention distribution by replacing the exponential function with a linear operator. This property compels the model to integrate global context for reconstruction, serving as an effective regularizer while also reducing computational complexity from $O(N^2d)$ to $O(Nd^2)$. The training objective for each patch expert is to minimize the reconstruction loss $\mathcal{L}_p$, defined as the mean cosine distance between the target feature vectors and their reconstructions $F_{\text{rec\_out}}$:

$$\mathcal{L}_p = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{F_{\text{target}}(i) \cdot F_{\text{rec\_out}}(i)}{\|F_{\text{target}}(i)\|\|F_{\text{rec\_out}}(i)\|}\right)\right], \tag{6}$$

where i is the patch index. During inference, the pointwise calculation of this distance yields the anomaly score map S_p.

Component-level Experts This group of N_c experts identifies semantic anomalies, where a component's local texture may be normal but its identity is incorrect. To extract features for individual components, we first establish a "component knowledge base" for each object class by applying K-Means clustering to the patch embeddings of all its normal training samples. During inference, each patch of a test sample is assigned to the nearest cluster center, generating component masks. We then use these masks to perform Masked Average Pooling on the feature map F_{target}, yielding a single embedding vector $c_k \in \mathbb{R}^D$ for each component.

This expert group consists of N_c identically structured experts, each being a lightweight MLP-based Autoencoder. This architecture is well-suited for processing the non-spatial, vectorized representations of the component embeddings $\{c_k\}$. Through its compressed bottleneck layer, each autoencoder is compelled to learn a low-dimensional manifold of the semantic identity of normal components, abstracting away from visual texture. When a semantically anomalous component is input, its embedding c_k deviates from this learned normal manifold, leading to inaccurate reconstruction. The training loss $\mathcal{L}_c$ minimizes the mean cosine distance between the original embeddings and their reconstructions $\hat{c}_k$:

$$\mathcal{L}_c = \mathbb{E}\left[\frac{1}{K_c}\sum_{k=1}^{K_c}(1 - \cos(c_k, \hat{c}_k))\right], \tag{7}$$

where K_c is the number of components. The individual component scores $S_c^{(k)}$ are calculated using this same cosine distance during inference.

Global-level Experts This group of N_g experts identifies logical anomalies that require a holistic view, such as incorrect component arrangements. Each expert is a convolutional Autoencoder that takes the entire sample I as input and aims to reconstruct the target feature map F_{target}. The architecture comprises a convolutional encoder, E_g, which progressively downsamples the sample to capture holistic spatial information, and a corresponding decoder, D_g. The information bottleneck between E_g and D_g forces the model to learn a compact, low-dimensional manifold representing the normal global layout and inter-component relations of a scene. An sample with a global anomaly cannot be encoded onto this normal manifold, leading to a high reconstruction error, which is quantified by the squared Euclidean distance:

$$\mathcal{L}_g = \mathbb{E}_{I \sim \mathcal{D}_{\text{normal}}}\left[\|F_{\text{target}} - D_g(E_g(I))\|_2^2\right]. \tag{8}$$

During inference, the corresponding anomaly score map S_g is generated by computing this same pointwise squared Euclidean distance between the target and its reconstruction.

Final Training Objective

The final training objective of AnomalyMoE, $\mathcal{L}_{\text{total}}$, combines the weighted reconstruction loss from the activated experts with our two auxiliary regularization losses:

$$\mathcal{L}_{\text{total}} = \sum_{j \in \text{Indices}_{\text{topK}}} G_j \cdot \mathcal{L}_{\text{rec}}^{(j)} + \lambda_{\text{ESB}}\mathcal{L}_{\text{ESB}} + \lambda_{\text{EIR}}\mathcal{L}_{\text{EIR}}. \tag{9}$$

Here, $\mathcal{L}_{\text{rec}}^{(j)}$ is the reconstruction loss for the j-th activated expert, corresponding to one of the losses ($\mathcal{L}_p$, $\mathcal{L}_c$, or $\mathcal{L}_g$) defined in the previous subsections, depending on the expert's type. It is weighted by its gate score G_j. Crucially, the auxiliary losses $\mathcal{L}_{\text{ESB}}$ and $\mathcal{L}_{\text{EIR}}$ are computed using the full pre-softmax logit distribution from the router. This ensures that all experts, including inactive ones, receive gradients for load balancing and information repulsion. This integrated objective guides the framework toward a state of high efficiency, diversity, and functional specialization.

Experiments

Experimental Setup

Datasets To comprehensively evaluate the generalizability and performance of AnomalyMoE, we conduct experiments on a diverse suite of eight challenging datasets. These benchmarks span multiple domains, including industrial structural anomalies (MVTec AD (Bergmann et al. 2021a), VisA (Zou et al. 2022)), 3D point clouds (MVTec 3D-AD (Bergmann et al. 2021b)), logical anomalies (MVTec-LOCO (Bergmann et al. 2022)), medical imaging (BrainMRI (Baid et al. 2021), LiverCT (Bilic et al. 2023), RESC (Hu, Chen, and Yi 2019)), and video surveillance (UCSD Ped2 (Wang and Miao 2010)). This collection provides a robust and varied testbed for our generalist model, covering a wide spectrum of visual anomaly types.

Evaluation Metrics To ensure a fair and comprehensive comparison with prior work, we adopt standard evaluation protocols for anomaly detection. We evaluate performance using the Area Under the Receiver Operating Characteristic

Metric	Dataset	Task	UniAD	Recontrast	UniNet	Dinomaly	UniVAD	LogSAD	**AnomalyMoE**
Image-level (AUC)	MVTec-AD	Industrial	87.5	95.0	84.7	98.8	97.8	96.9	**99.5**
	VisA	Industrial	80.2	90.1	87.5	95.7	93.5	94.5	**98.1**
	MVTec LOCO	Logical	71.3	79.1	77.8	78.2	71.0	86.2	**87.5**
	BrainMRI	Medical	90.9	85.0	54.1	88.5	80.2	75.8	**92.1**
	LiverCT	Medical	52.9	55.5	47.5	67.1	70.0	65.8	**71.1**
	RESC	Medical	83.2	94.6	70.3	**92.2**	85.5	81.2	**92.2**
	MVTec 3D	3D	72.5	75.1	70.0	86.5	80.2	84.5	**87.2**
	Ped2	Video	72.5	83.7	61.7	**97.1**	94.3	85.5	**97.1**
Pixel-level (AUC)	MVTec-AD	Industrial	94.2	96.3	91.2	97.6	96.5	97.0	**97.7**
	VisA	Industrial	96.3	90.8	87.3	96.4	98.2	97.1	**99.0**
	MVTec LOCO	Logical	76.4	74.1	73.2	71.5	75.1	79.4	**82.2**
	BrainMRI	Medical	98.0	96.3	86.4	95.8	96.8	95.5	**97.3**
	LiverCT	Medical	96.3	96.6	95.0	97.2	96.3	96.6	**97.2**
	RESC	Medical	95.9	94.6	78.4	95.3	94.9	91.9	**96.0**
	MVTec 3D	3D	95.3	75.1	95.2	98.7	94.6	98.2	**98.9**
	Ped2	Video	95.4	97.1	96.3	98.3	97.4	98.3	**98.4**

Table 1: Quantitative comparison of AnomalyMoE with state-of-the-art methods across eight diverse datasets under the unified, single-model setting. We report Image-level and Pixel-level AUC (%). The best-performing method is highlighted in bold.

curve (AUC). We report both Image-level AUC to assess the model's detection capability and Pixel-level AUC to evaluate its localization accuracy.

Implementation Details We use a pre-trained DINOv2 model (ViT-B/14) (Oquab et al. 2023) as the frozen visual encoder. All input images are resized to 448×448 and then center-cropped to 392×392. For our MoE architecture, we set the number of experts in each group to $N_p = 6$, $N_c = 6$, and $N_g = 6$, with a Top-K routing of $K = 3$. The model is trained for 50,000 iterations using the StableAdamW optimizer with a learning rate of 5×10^{-4} and a weight decay of 10^{-4}. We use a batch size of 16 on 4 NVIDIA A6000 GPUs. The hyperparameters for the auxiliary losses are set to $\lambda_{\text{ESB}} = 0.01$ and $\lambda_{\text{EIR}} = 0.0001$.

Baselines We compare AnomalyMoE against a comprehensive set of state-of-the-art methods to demonstrate its superiority across different paradigms. Our baselines include top-performing generalist patch-level reconstruction-based models like UniAD (You et al. 2022), ReContrast (Guo et al. 2023), UniNet (Wei, Jiang, and Xu 2025), and Dinomaly (Guo et al. 2025), and advanced component-based methods that leverage language priors, specifically UniVAD (Gu et al. 2025b) and LogSAD (Zhang et al. 2025). For a fair comparison, we reproduce them using their officially released code under our unified evaluation protocol.

Main Quantitative Results

We present the main quantitative results in Table 1, establishing AnomalyMoE as the new state-of-the-art in generalist anomaly detection. The limitations of specialized approaches become clear in this unified setting. While reconstruction-based methods like Dinomaly excel on structural datasets, they falter on logical ones. On the other hand, component-based methods like LogSAD performs exceptionally well on its target MVTec-LOCO dataset, yet its performance significantly degrades on other domains. This demonstrates that while specialized methods often suffer a significant performance drop when generalized, AnomalyMoE's hierarchical MoE architecture delivers robust, top-tier performance across all anomaly types, validating its superior design.

Beyond its superior accuracy, AnomalyMoE offers a significant advantage in efficiency. Unlike methods such as UniVAD (Gu et al. 2025b) and LogSAD (Zhang et al. 2025) that depend on computationally expensive large language models, our framework is entirely language-free, resulting in an inference speed that is 11.3 times faster than UniVAD (58.5 ms vs. 658.5 ms) and 33 times faster than LogSAD (58.5 ms vs. 1932.6 ms).

Ablation Studies

To validate our key design choices, we conduct a series of comprehensive ablation studies, with results summarized in Table 2 and Table 3. The results first demonstrate the necessity of our hierarchical expert structure. As shown in Table 2, no single expert level is sufficient for robust, generalist performance. For instance, the Patch-only model excels on the structural anomalies of MVTec AD but fails on the logical tasks in MVTec LOCO. The full three-level configuration consistently outperforms all partial combinations, validating that the synergy of all three expert types is essential to cover the full spectrum of anomalies.

Furthermore, Table 3 confirms the importance of our auxiliary training modules. Removing the Expert Selection Balancing (ESB) module degrades performance by causing suboptimal router convergence and wasting model capacity. Similarly, removing the Expert Information Repulsion (EIR) module allows experts to learn redundant features, diminishing the benefits of specialization. Collectively, these ablations empirically confirm that each core component of AnomalyMoE, including its hierarchical structure, selection balancing, and information repulsion, is essential for its state-of-the-art, generalist performance.

Expert Levels			Datasets						
Patch	Component	Global	MVTec-AD	VisA	MVTec LOCO	BrainMRI	RESC	MVTec 3D	Ped2
✓			(98.2, 96.0)	(97.0, 98.8)	(78.2, 75.9)	(91.6, 97.1)	(91.2, 95.9)	(85.7, 98.6)	(96.2, 98.5)
✓	✓		(98.4, 95.1)	(97.9, 98.9)	(84.7, 79.4)	(89.5, 96.4)	(88.7, 91.2)	(85.9, 85.8)	(96.9, 97.4)
✓		✓	(99.1, 96.9)	(97.7, 98.9)	(86.8, 81.1)	(89.2, 97.1)	(91.5, 95.5)	(86.3, 97.3)	(97.0, 98.3)
✓	✓	✓	**(99.5, 97.7)**	**(98.1, 99.0)**	**(87.5, 82.2)**	**(92.1, 97.3)**	**(92.2, 96.0)**	**(87.2, 98.9)**	**(97.1, 98.4)**

Table 2: Ablation study on the hierarchical expert structure of AnomalyMoE. We report the (Image-level AUC, Pixel-level AUC) in percent for different combinations of expert groups. The full three-level configuration achieves the best performance across all datasets, confirming the contribution of each expert.

Dataset	w/o ESB	w/o EIR	**AnomalyMoE**
MVTec-AD	(99.2, 97.2)	(99.0, 97.1)	**(99.5, 97.7)**
VisA	(97.3, 98.8)	(97.4, 98.9)	**(98.1, 99.0)**
MVTec LOCO	(84.2, 81.1)	(85.6, 92.0)	**(87.5, 82.2)**
BrainMRI	(91.2, 97.0)	(91.6, 97.2)	**(92.1, 97.3)**
LiverCT	(69.3, 97.1)	(69.4, 97.1)	**(71.1, 97.2)**
RESC	(91.2, 95.3)	(91.2, 95.7)	**(92.2, 96.0)**
MVTec 3D	(85.7, 98.0)	(85.9, 98.7)	**(87.2, 98.9)**
Ped2	(96.0, 97.9)	(96.6, 98.3)	**(97.1, 98.4)**

Table 3: Ablation study on the ESB and EIR auxiliary modules in AnomalyMoE. The table compares the performance of the full model against versions without either the ESB or EIR loss. Results are reported as (Image-level AUC, Pixel-level AUC) in percent.

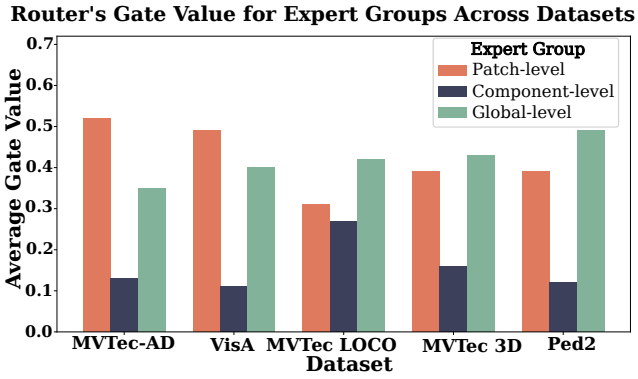

Figure 4: Analysis of the router's behavior. The histogram shows the average gate values assigned to each expert group across various datasets.

Qualitative Analysis and Visualization

To provide an intuitive understanding of our framework, we conduct a qualitative analysis of the router's behavior and the experts' specialization. As shown in Figure 4, the router learns to dynamically allocate resources, assigning higher gate values to Patch-level experts on structurally anomalous datasets like MVTec AD, while prioritizing Global-level experts on logical datasets like MVTec-LOCO. Figure 5 provides direct visual confirmation of this functional specialization. For a structural defect (e.g., "wood"), the Patch-level expert generates a precise activation, whereas for a logical missing part anomaly ("splicing_connectors"), the Global-level expert correctly identifies the compositional error. In all cases, the final aggregated result effectively integrates the most salient signals from the relevant experts. These visualizations clearly illustrate the complementary and synergistic nature of our hierarchical design.

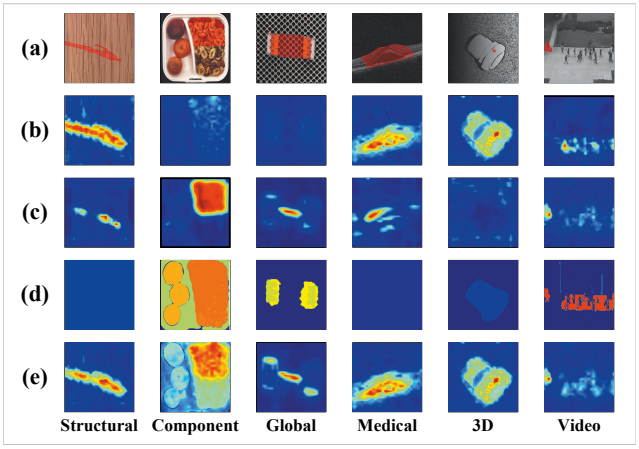

Figure 5: Visualization of the anomaly detection process. Rows from top to bottom: (a) Input, (b) Patch-level expert output, (c) Global-level expert output, (d) Component-level expert output, and (e) final AnomalyMoE result.

Conclusion

We introduce AnomalyMoE, a novel, language-free generalist framework designed to address the fragmented nature of visual anomaly detection. By decomposing anomalies into a three-level semantic hierarchy, our Mixture-of-Experts architecture leverages specialized patch, component, and global experts to achieve comprehensive detection capabilities. Governed by a sophisticated routing mechanism enhanced by our proposed Expert Selection Balancing and Expert Information Repulsion modules, AnomalyMoE learns a diverse and functionally disentangled set of representations. Extensive experiments on eight challenging datasets demonstrate that our approach not only establishes a new state-of-the-art but also consistently outperforms specialized methods in their respective domains. AnomalyMoE represents a significant step towards a truly universal, efficient, and scalable anomaly detection system, paving the way for more robust real-world applications.

Acknowledgments

This work was supported by National Key R&D Program of China under Grant No.2022ZD0160601, in part by National Natural Science Foundation of China (No. 62276260, 62076235, 62472423), and Beijing Natural Science Foundation (L252036).

References

Baid, U.; Ghodasara, S.; Mohan, S.; Bilello, M.; Calabrese, E.; Colak, E.; Farahani, K.; Kalpathy-Cramer, J.; Kitamura, F. C.; Pati, S.; et al. 2021. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*.

Bao, J.; Sun, H.; Deng, H.; He, Y.; Zhang, Z.; and Li, X. 2024. Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4042–4053.

Bergmann, P.; Batzner, K.; Fauser, M.; Sattlegger, D.; and Steger, C. 2021a. The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4): 1038–1059.

Bergmann, P.; Batzner, K.; Fauser, M.; Sattlegger, D.; and Steger, C. 2022. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4): 947–969.

Bergmann, P.; Jin, X.; Sattlegger, D.; and Steger, C. 2021b. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. *arXiv preprint arXiv:2112.09045*.

Bilic, P.; Christ, P.; Li, H. B.; Vorontsov, E.; Ben-Cohen, A.; Kaissis, G.; Szeskin, A.; Jacobs, C.; Mamani, G. E. H.; Chartrand, G.; et al. 2023. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84: 102680.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, 1779–1788. PMLR.

Defard, T.; Setkov, A.; Loesch, A.; and Audigier, R. 2021. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, 475–489. Springer.

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

Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Li, H.; Tang, M.; and Wang, J. 2024a. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2041–2049.

Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Tang, M.; and Wang, J. 2024b. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 1932–1940.

Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Tang, M.; and Wang, J. 2025a. FiLo++: Zero-/Few-Shot Anomaly Detection by Fused Fine-Grained Descriptions and Deformable Localization. *arXiv preprint arXiv:2501.10067*.

Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Tang, M.; and Wang, J. 2025b. Univad: A training-free unified model for few-shot visual anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15194–15203.

Guo, J.; Lu, S.; Jia, L.; Zhang, W.; and Li, H. 2023. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. *Advances in Neural Information Processing Systems*, 36: 10721–10740.

Guo, J.; Lu, S.; Zhang, W.; Chen, F.; Li, H.; and Liao, H. 2025. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20405–20415.

Han, D.; Pan, X.; Han, Y.; Song, S.; and Huang, G. 2023. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5961–5971.

Hsieh, Y.-H.; and Lai, S.-H. 2024. Csad: Unsupervised component segmentation for logical anomaly detection. *arXiv preprint arXiv:2408.15628*.

Hu, J.; Chen, Y.; and Yi, Z. 2019. Automated segmentation of macular edema in OCT using deep neural networks. *Medical image analysis*, 55: 216–227.

Huang, C.; Jiang, A.; Feng, J.; Zhang, Y.; Wang, X.; and Wang, Y. 2024. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11375–11385.

Kim, S.; An, S.; Chikontwe, P.; Kang, M.; Adeli, E.; Pohl, K. M.; and Park, S. H. 2024. Few shot part segmentation reveals compositional logic for industrial anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 8591–8599.

Lei, T.; Chen, S.; Wang, B.; Jiang, Z.; and Zou, N. 2024. Adapted-moe: Mixture of experts with test-time adaption for anomaly detection. *arXiv preprint arXiv:2409.05611*.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Liang, H.; Xie, G.; Hou, C.; Wang, B.; Gao, C.; and Wang, J. 2025. Look inside for more: Internal spatial modality perception for 3D anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5146–5154.

Liu, T.; Li, B.; Du, X.; Jiang, B.; Jin, X.; Jin, L.; and Zhao, Z. 2023. Component-aware anomaly detection framework for adjustable and logical industrial visual inspection. *Advanced Engineering Informatics*, 58: 102161.

Minhas, M. S.; and Zelek, J. 2020. Semi-supervised anomaly detection using autoencoders. *arXiv preprint arXiv:2001.03674*.

Naji, Y.; Setkov, A.; Loesch, A.; Gouiffès, M.; and Audigier, R. 2022. Spatio-temporal predictive tasks for abnormal event detection in videos. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–8. IEEE.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, S.; and Miao, Z. 2010. Anomaly detection in crowd scene. In *IEEE 10th International Conference on Signal Processing Proceedings*, 1220–1223. IEEE.

Wang, X.; Wang, X.; Bai, H.; Lim, E. G.; and Xiao, J. 2025. CNC: Cross-modal Normality Constraint for Unsupervised Multi-class Anomaly Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7943–7951.

Wei, S.; Jiang, J.; and Xu, X. 2025. UniNet: A Contrastive Learning-guided Unified Framework with Feature Selection for Anomaly Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9994–10003.

Yang, Y.; Lee, K.; Dariush, B.; Cao, Y.; and Lo, S.-Y. 2024. Follow the rules: Reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision*, 304–322. Springer.

Ye, J.; Zhao, W.; Yang, X.; Cheng, G.; and Huang, K. 2025. Po3ad: Predicting point offsets toward better 3d point cloud anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1353–1362.

You, Z.; Cui, L.; Shen, Y.; Yang, K.; Lu, X.; Zheng, Y.; and Le, X. 2022. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35: 4571–4584.

Zhang, J.; Wang, G.; Jin, Y.; and Huang, D. 2025. Towards Training-free Anomaly Detection with Vision and Language Foundation Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15204–15213.

Zhang, Y.; Chen, Z.; Guo, L.; Xu, Y.; Hu, B.; Liu, Z.; Zhang, W.; and Chen, H. 2024. Multiple heads are better than one: Mixture of modality knowledge experts for entity representation learning. *arXiv preprint arXiv:2405.16869*.

Zhou, Q.; Yan, J.; He, S.; Meng, W.; and Chen, J. 2024. Pointad: Comprehending 3d anomalies from points and pixels for zero-shot 3d anomaly detection. *Advances in Neural Information Processing Systems*, 37: 84866–84896.

Zhu, B.; Gu, Z.; Zhu, G.; Chen, Y.; Tang, M.; and Wang, J. 2024. ADFormer: Generalizable few-shot anomaly detection with dual CNN-transformer architecture. *IEEE Transactions on Instrumentation and Measurement*.

Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; and Dabeer, O. 2022. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, 392–408. Springer.