

Multi-Modal Anomaly Detection: A Survey

Xudong Mou[✉], Zexin Wu[✉], Chuan Luo[✉], Shiru Chen[✉], Xudong Liu[✉], Chunming Hu[✉], Renyu Yang[✉]

Abstract—Multi-Modal Anomaly Detection (MMAD) detects rare abnormal events from heterogeneous data sources and is increasingly used in safety- and reliability-critical applications such as industrial inspection and cybersecurity. Yet the literature is fragmented across domains and modality combinations, and existing surveys usually group methods by architecture rather than by how abnormality is defined and separated in multi-modal settings. We survey MMAD from an assumption-driven perspective. We formalize the problem, identify five intrinsic characteristics underlying its core challenges, and organize prior work into two complementary paradigms. The first, *normality-assumption* methods, models regularity via representation learning, cross-modal alignment, and knowledge enhancement. The second, *abnormality-assumption* methods, sharpens decision boundaries through coarse-grained, structural, and semantic anomaly injection. We also investigate how foundation models are reshaping MMAD through scalable pretraining, flexible cross-modal transfer, and emerging reasoning capabilities. Finally, we compile representative benchmarks and evaluation protocols across domains and highlight open problems and future directions for robust, adaptive, and interpretable MMAD systems.

Index Terms—Multi-modal Anomaly Detection, Normality Assumption, Abnormality Assumption, Foundation Models, Assumption-driven Taxonomy.

I. INTRODUCTION

Anomaly Detection (AD) identifies observations that deviate from normal behavior [1], [2]. Because many AD applications are safety- and reliability-critical—such as industrial inspection, healthcare, surveillance, finance, and cybersecurity—the field has remained active for decades [3], [4]. Traditional methods typically use unimodal data and model normality in one observation space, but often fail when anomalies are subtle, partially observed, or modality-dependent.

With the rise of heterogeneous data sources, Multi-Modal Anomaly Detection (MMAD) has become a natural extension of classical AD. By combining evidence from modalities such as images, time series, text, audio, and sensor signals, MMAD can detect abnormal events that appear normal in

any single modality [5]. Multi-modal integration also improves robustness to noise, missing data, and partial failures, yielding a more accurate view of complex systems [6]. However, multi-modality also reshapes the problem by introducing heterogeneity and alignment issues, and by exacerbating challenges such as granularity mismatch, semantic ambiguity, and evolving data distributions.

Several surveys have examined MMAD and related areas. Existing reviews cover industrial visual anomaly detection with RGB, 3D, and multi-modal inputs [5], multi-modal datasets for audio-visual event understanding [7], and advances in multi-modal time-series analysis [8]. Other domain-focused surveys study multi-modal event detection and representation learning from collaboration and fusion perspectives [9], [10]. However, these works either target specific domains (e.g., RGB/3D industrial images, audio-visual benchmarks, time-series signals) or discuss general learning tools without offering a holistic view of MMAD.

Interest in applying Foundation Models (FMs) to MMAD is rapidly growing. Pretrained on large-scale, diverse corpora, FMs provide strong representation learning and cross-modal alignment that are hard to achieve with task-specific architectures [11], [12]. This has enabled promising zero-shot and few-shot anomaly detection, especially when labeled anomalies are scarce [13], [14]. FM-based methods also support flexible fusion and transfer, reducing handcrafted pipelines and easing adaptation to new modalities or anomaly types [15]–[17]. Nevertheless, existing FM-centric surveys largely describe generic AD pipelines (e.g., FM-as-encoder or FM-as-interpreter) without systematically linking multi-modal pretraining, cross-modal alignment, and FM-driven inference to MMAD’s specific challenges. Their taxonomies are thus weakly tied to the essence of anomaly detection, tending to emphasize architectural novelty over clear explanations of detection mechanisms.

In this survey, we take an assumption-driven view of MMAD. Instead of grouping methods by modality or model family, we organize them by their assumptions about *normality* and *abnormality*. We review MMAD approaches through two lenses: (i) normality-based methods that model regular patterns and define anomalies implicitly, and (ii) abnormality-based methods that explicitly construct contrasts to mark what lies beyond normality. Within each lens, we trace the development from classical unimodal techniques to modern multi-modal methods, showing how their assumptions interact with MMAD-specific challenges. Grounding the taxonomy in anomaly detection itself yields a unified view of how methods define, expose, and separate abnormal behavior in multi-modal

Manuscript received December 29, 2025; revised May 25, 2026; accepted 13 July 2026. This work is supported in part by National Key R&D Program of China (Grant No. 2024YFB4505901), Beijing Natural Science Foundation (No. L241050), National Natural Science Foundation of China (Grant No. 62402024), and the Fundamental Research Funds for the Central Universities. (Corresponding author: Renyu Yang)

Xudong Mou and Xudong Liu are with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China. (Email: mxd@buaa.edu.cn, liuxd@buaa.edu.cn).

Zexin Wu, Chuan Luo, Chunming Hu, and Renyu Yang are with the School of Software, Beihang University, Beijing 100191, China (Email: zexinwu@buaa.edu.cn, chuanluo@buaa.edu.cn, hucm@buaa.edu.cn, renyuyang@buaa.edu.cn).

Shiru Chen is with the Shandong Inspur Intelligent Production Technology Co., Ltd, Jinan, 250101, China (Email: chenshiru@inspur.com).

settings. This framework clarifies the strengths and limitations of existing approaches—including FM-based methods—and highlights open gaps and future research directions. Fig. 1 summarizes the roadmap of this survey.

The paper makes the following main contributions.

- To the best of our knowledge, we present the first systematic survey of MMAD from an assumption-driven perspective, providing a unified and interpretable view of the research landscape.
- We analyze the intrinsic nature of MMAD and distill a set of fundamental challenges that explain why many problems remain only partially resolved.
- We review a broad body of studies from leading conferences and journals through the complementary lenses of normality and abnormality assumptions, summarizing their objectives, capabilities, and limitations in addressing MMAD challenges.
- We identify emerging trends and formulate future research directions for MMAD through the lens of foundation-model-driven assumption fusion, highlighting reliable anomaly hypothesis generation, semantic cross-modal reasoning, and lifelong adaptation.

The remainder of this paper is organized as follows. Section II defines MMAD and analyzes its core challenges. Sections III and IV review assumption-driven methods from normality- and abnormality-centric perspectives, highlighting the impact of foundation models. Section V summarizes benchmarks and applications, Section VI discusses open problems, and Section VII concludes.

II. PRELIMINARIES

A. Definition of Multi-modal Anomaly Detection

Multi-modal anomaly detection seeks to integrate diverse and heterogeneous information to identify rare anomalous samples that deviate substantially from the normal distribution [1]. By exploiting cross-modal complementarity, it can reveal complex anomalous patterns that may be obscured in unimodal views due to limited data.

Let $\mathcal{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(K)}\}$ be a multi-modal sample with K heterogeneous modalities, where $\mathbf{x}^{(k)} \in \mathbb{R}^{d_k}$ is the observation from modality k (e.g., text, images, time series, audio). To model missing modalities, let $\mathcal{M} \subseteq \{1, \dots, K\}$ be the set of observed modalities, and define $\mathcal{X}_{\mathcal{M}} := \{\mathbf{x}^{(k)}\}_{k \in \mathcal{M}}$. MMAD aims to learn an anomaly score $\mathcal{S}$ via a scoring function $f(\cdot)$ that measures how far a sample deviates from the joint normal distribution $\mathcal{P}_{\text{normal}}$.

$$\mathcal{S} = f(\mathcal{X}, \mathcal{M}) := f(\mathcal{X}_{\mathcal{M}}) \in \mathbb{R}. \quad (1)$$

The decision rule for a test sample is:

$$\text{Label}(\mathcal{X}, \mathcal{M}) = \begin{cases} 1 \text{ (Anomaly)}, & \text{if } \mathcal{S} > \tau, \\ 0 \text{ (Normal)}, & \text{if } \mathcal{S} \leq \tau, \end{cases} \quad (2)$$

where τ is a predefined threshold typically determined on a validation set.

Compared with unimodal anomaly detection, MMAD must capture both within-modality regularities (in each $\mathbf{x}^{(k)}$) and

cross-modal dependencies (across $\mathcal{M}$), introducing challenges such as cross-modal inconsistency, modality imbalance, and missing modalities.

This formulation highlights three issues that deserve particular attention. (i) Since anomalies are defined relative to normality, multi-modal normality should be explicitly specified and aligned with the target task, capturing both within-modality regularities and cross-modal structural consistency; otherwise, detection results may be unreliable. (ii) Anomaly detection typically operates under extreme class imbalance, and in multi-modal scenarios, anomalies may arise from rare cross-modal interactions rather than from a single modality alone. (iii) Following [2], it is crucial to distinguish anomaly detection (AD), novelty detection (ND), outlier detection (OD), and out-of-distribution (OOD) detection, which differ in problem settings and assumptions and become more nuanced in multi-modal contexts due to interacting information sources. For example, AD detects deviations from established normal patterns, ND treats novel samples as potential new categories, OOD emphasizes robustness in multi-class settings, and OD identifies outliers within a given dataset.

These considerations underscore the inherent complexity of MMAD and motivate the detailed discussion of its key challenges in the next section.

B. Essential Problems in MMAD

The complexity of MMAD can be deconstructed into the inherent challenges of anomaly detection itself and the complications introduced by the integration of multiple data sources. We summarize these essential problems along two dimensions: the intrinsic nature of anomalies and the exacerbation introduced by multimodality.

1) *Intrinsic characteristics of anomaly detection*: At its core, anomaly detection addresses the pursuit of the “unknown” and the “rare”. The primary bottleneck is *anomaly scarcity*: real-world anomalies are extremely infrequent, making representative datasets of failure modes nearly impossible to collect. This data poverty leads to a chronic *knowledge deficit*, as the mechanisms, statistical distributions, and boundary conditions of anomalies remain poorly understood. *Anomaly diversity* also appears in unimodal settings, making modeling inherently challenging, for example, point- and pattern-wise anomalies in time series and pixel- and patch-level anomalies in images. *Concept drift* further challenges any deployment, as the notion of normality continuously evolves in non-stationary environments.

2) *Exacerbation introduced by multimodality*: Introducing multiple modalities brings *modal heterogeneity*, where data from disparate sensors exhibit inconsistent distributions, asynchronous sampling, and contradictory signals. Importantly, multimodality amplifies the inherent challenges of AD. Knowledge deficits are worsened when scarce abnormal signals must be reconciled across unaligned feature spaces. Anomaly diversity extends to the cross-modal level, besides anomalies that cross modalities, including spatial, relational, temporal, or semantic, an event may appear normal within a single stream but manifest as anomalous through inter-modal conflicts. Concept drift also becomes multi-dimensional

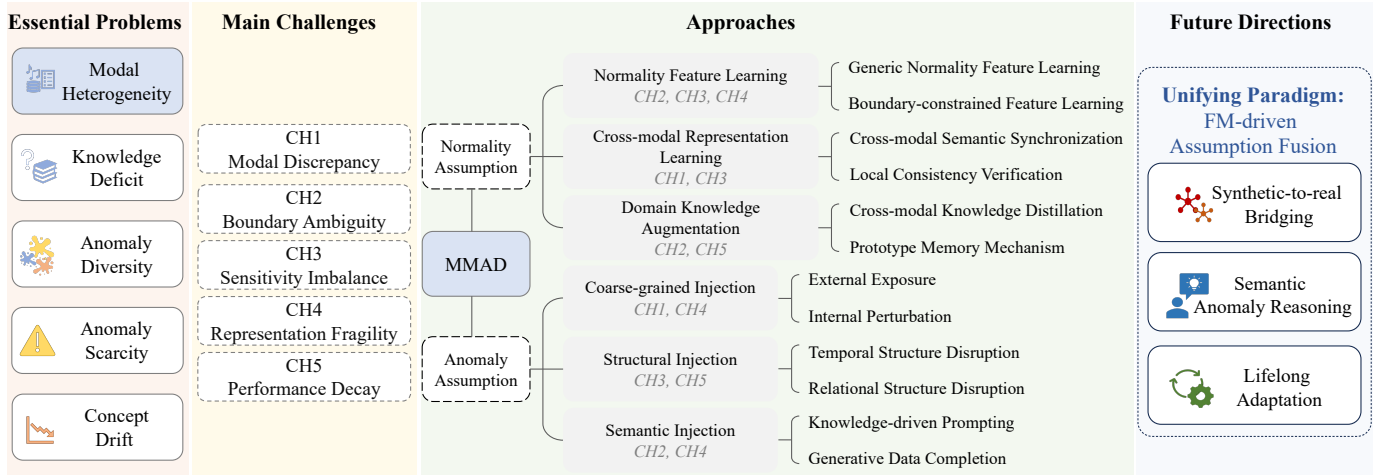


Fig. 1: Conceptual roadmap of multi-modal anomaly detection. The figure illustrates the progression from the intrinsic nature of anomaly detection to multi-modality-induced challenges, the assumption-driven modeling paradigms, and the emerging future opportunities in MMAD.

and harder to calibrate, as different modalities may evolve at varying rates or even in opposing directions.

In summary, the landscape of MMAD is defined by five key characteristics: modal heterogeneity, knowledge deficit, anomaly diversity, anomaly scarcity, and concept drift.

C. Main Challenges in MMAD

Despite recent progress in scaling and high-dimensional modeling, MMAD still faces several unresolved challenges.

1) *CH1: Modal discrepancy from heterogeneity*: Modalities such as text, images, time series, and audio exhibit substantial distributional heterogeneity in feature space [18]. Anomalies may manifest differently across modalities, and effective detection can require integrating multiple streams. For example, during robotic grasping, a micro-slip induces distinct friction changes in force/torque time-series signals but remains visually imperceptible. Without accurate semantic alignment, the model may exhibit visual bias, miss the slip, and even cause collisions during assembly. Achieving semantically aligned representations is therefore critical. The goal is for modalities to complement one another, rather than acting as “squabbling experts” that produce conflicting predictions.

2) *CH2: Boundary ambiguity from knowledge deficit*: The boundary between normality and anomaly is often blurred by limited expert guidance and weak or noisy annotations. A statistical outlier may be a rare yet legitimate operating state in specialized domains, complicating boundary optimization [19]. Conversely, data collected during the wear-out stage of the bathtub curve may naturally include failure patterns, contaminating the “normal” baseline. Although anomalies are frequently treated as unknowns, basic priors such as the 3σ rule provide indispensable constraints. Forcing a model to learn entirely from data without such domain priors is inefficient and often increases false-alarm rates. Leveraging limited domain knowledge to form robust and precise decision boundaries thus remains a key challenge.

3) *CH3: Sensitivity imbalance from anomaly diversity*: Anomalies vary widely in scale, ranging from point-level outliers to long-horizon patterns and large, patchy structural deviations. This requires sensitivity across multiple temporal and spatial granularities. For example, in intrusion detection, brute-force attacks can appear as high-frequency point anomalies, whereas low-and-slow exfiltration emerges over weeks as a subtle collective pattern. Emphasizing only global features can miss local irregularities, while over-indexing on local noise risks overfitting. The goal is to develop multi-scale mechanisms that capture the anomaly spectrum, from micro-structural glitches to macro-level semantic violations.

4) *CH4: Representation fragility from anomaly scarcity*: Real-world anomalies are naturally scarce, creating severe class imbalance and overwhelming models with normal samples. This scarcity forces detectors to rely heavily on the normality assumption, often degrading recall for rare or unseen patterns. For instance, in critical infrastructure monitoring, catastrophic failures may never occur in the observed history, yielding “black swan” events that models cannot extrapolate to. The central challenge is to complete the flip side of the distribution to sharpen the decision boundary and reliably identify rare threats.

5) *CH5: Performance decay from concept drift*: In non-stationary environments, the notion of normality evolves, gradually eroding model efficacy. Drift arises when environmental changes or system aging alter data statistics, making previously learned normal manifolds obsolete. For example, in industrial IoT, sensor readings for healthy machines can drift as components wear down or seasonal temperature shifts occur. Without adaptation mechanisms, models may flag these legitimate shifts as anomalies. Developing lifelong learning capabilities that separate true anomalies from a new normality is therefore essential for sustained deployment.

III. METHODS BASED ON NORMALITY ASSUMPTION

In MMAD, the extreme scarcity of anomalies often makes direct discriminative learning unreliable. Hence, many meth-

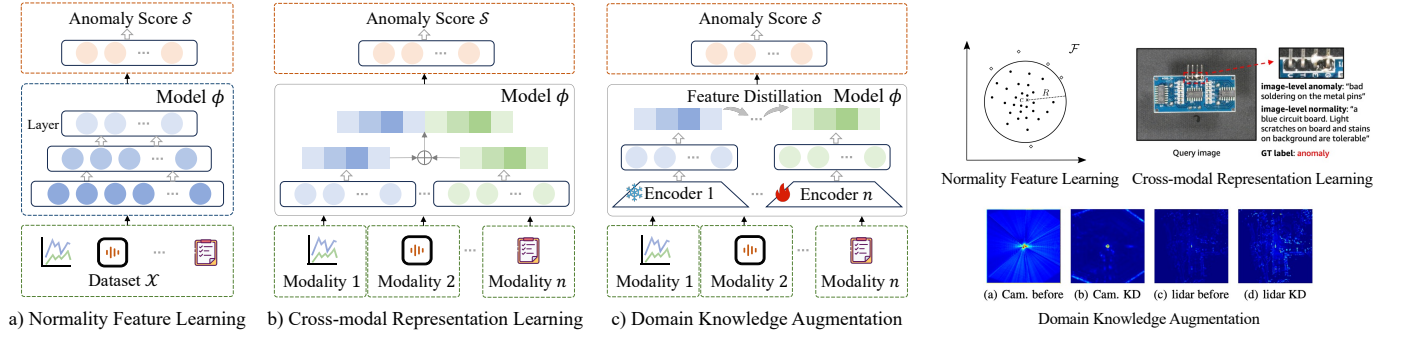


Fig. 2: Overview of normality-assumption methods across three groups: a) normal feature learning learns compact normal representations; b) cross-modal representation learning captures inter-modality relations to enrich the normal space; and c) domain knowledge enhancement leverages priors to refine the normal boundary. The right panel shows schematic or qualitative results adapted from Deep SVDD [20], WinCLIP [14], and VeXKD [21].

ods adopt the *normality assumption*. Under this assumption, normal multi-modal data exhibit consistent and compact association patterns, and anomalies are identified as deviations from these norms.

Normality-based methods typically learn a representation z that captures both modality-specific regularities and cross-modal consistency. The scoring function f is often a two-stage mapping with an encoder ϕ and a scoring head ς , i.e., $f = \varsigma \circ \phi$. Given a test sample with observed modality set $\mathcal{M}$, the anomaly score is computed as

$$\mathcal{S} = f(\mathcal{X}, \mathcal{M}) = \varsigma(\phi(\mathcal{X}_{\mathcal{M}})). \quad (3)$$

Here, $\varsigma(\cdot)$ can implement different criteria, such as reconstruction error, likelihood under a normal density model, or distance to a compact normal set. By modeling normal patterns during training, these methods establish a reference for evaluation and flag samples that deviate at test time. As illustrated in Fig. 2, we categorize normality-assumption methods into three groups: normal feature learning, cross-modal representation learning, and domain knowledge enhancement.

A. Normality Feature Learning

In unsupervised or weakly supervised MMAD, anomaly scarcity and variability make it natural to learn a representation of *normal* patterns from predominantly normal data, thereby characterizing normality and defining the decision boundary (CH2). This often mitigates sensitivity imbalance (CH3) and improves anomaly recall (CH4).

Normality feature learning aims to obtain a multimodal embedding z that captures modality-specific regularities and cross-modal consistency. We abstract $\phi(\cdot)$ as a holistic normality encoder that maps a (possibly incomplete) observation to an embedding:

$$z = \phi(\mathcal{X}_{\mathcal{M}}). \quad (4)$$

In this view, $\phi(\cdot)$ hides architectural choices and can be instantiated via reconstruction, prediction, or density modeling. A key practical issue is *uneven modality maturity* (CH1): strong foundation encoders for vision/text may coexist with shallow, task-specific encoders for bespoke sensors, biasing representation learning; we discuss remedies in Sec. III-B. The

embedding z thus serves as an implicit normality prior. We group normality feature learning into generic and boundary-constrained methods.

1) *Generic Normality Feature Learning*: This category focuses on learning features of normal patterns, effectively addressing the challenges arising from the diversity (CH3) and rarity (CH4) of anomalous samples. By accurately characterizing the distribution structure of normal data, the model can identify inputs located in low-probability density regions as anomalous, thereby achieving generalized detection of unknown anomalies.

Foundations in unimodal settings. Inspired by classical data compression and dimensionality reduction techniques, reconstruction-based anomaly detection methods aim to learn compact representations of normal data by enforcing the model to capture essential structural patterns while minimizing reconstruction errors. In time-series analysis, LSTM-ED [22] proposes an encoder-decoder scheme based on long-short-term memory networks that learns to reconstruct normal time-series behavior. TranAD [23] uses attention-based sequence encoders to swiftly perform inference with the knowledge of the broader temporal trends in the data. Generative adversarial networks [24] further advanced reconstruction-based detection. For images domain, f-AnoGAN [25] builds a generative model of healthy training data. Subsequent works extended this paradigm to general and industrial images. GANomaly [26] uses a conditional generative adversarial network that jointly learns the generation of high-dimensional image space and the inference of latent space. Another key approach is prediction-based anomaly detection, which uses patterns in previous normal data to predict current data instances to learn feature representations. This approach is popular in video anomaly detection, such as Liu et al. [27] and AnoPCN [28] leverage the difference between a predicted future frame and its ground truth to detect an abnormal event.

Extensions to multi-modal scenarios. Building upon unimodal research, researchers have extended normal feature learning to multi-modal contexts to leverage complementary information from different data sources. In this scenario, the research objective shifts from modeling a single distribution to capturing the joint distribution of multiple modalities. For

example, Feng et al. [29] train an autoregressive model to generate sequences of audio-visual features, using feature sets that capture the temporal synchronization between video frames and sound. Willibald et al. [30] propose a mixture-of-experts framework that integrates the complementary detection mechanisms with a visual-language model for environment monitoring and a Gaussian-mixture regression-based detector for tracking deviations in interaction forces and robot motions. MAVAD [31] combines visual and audio features extracted from video sequences by means of cross-attention.

Advantages. Such methods are relatively simple in structure and exhibit strong generality, making them applicable to various modalities and easily integrable with different neural network architectures for anomaly detection, which enables effective generalization to previously unseen abnormal patterns.

Limitations. When the training data are contaminated, the learned feature distributions may inevitably become biased, leading to an inaccurate characterization of normal patterns. Moreover, methods that primarily rely on modeling normal data distributions often fail to capture fine-grained semantic consistency across different modalities, which can further limit discriminative capability and degrade detection performance in multi-modal settings.

2) *Boundary-constrained Feature Learning:* The paradigm introduces explicit anomaly measures to construct a tight decision boundary that encloses normal samples in the feature space, thereby directly alleviating the issue of ambiguous decision boundaries (CH2).

Foundations in unimodal settings. Classical one-class methods (e.g., OC-SVM [32] and SVDD [33]) build explicit boundaries—a hyperplane or minimal-volume hypersphere—to separate normal data from anomalies in the input or kernel space. Deep extensions learn more expressive latent boundaries, exemplified by Deep SVDD [20] and its variants such as Deep SAD [34]. For time series, COCA [35] and AOC [36] adapt hypersphere-based objectives with temporal encoders, while RoCA [37] and FOCA [38] further improve robustness under contamination and foundation backbones.

Extensions to multi-modal scenarios. In multi-modal scenarios, boundary-constrained feature learning incorporates explicit decision boundaries—e.g., hyperspheres or ellipsoids—over fused representations to model normality and mitigate distributional discrepancies jointly. An early spectral framework [39] uses tensor or graph embeddings to enclose consistent multi-source patterns within a compact low-rank manifold. Crucially, optimizing such compactness in high-dimensional multimodal spaces risks *representation collapse*, where models may converge to trivial mappings (e.g., collapsing all inputs to a constant point) to trivially minimize the boundary loss. To avoid this, recent extensions explicitly integrate structural regularizations. For instance, MCDSVDD [40] trains a neural network to map different normal classes to separate hyperspheres in the latent space, maintaining spatial variance rather than collapsing to a single center. This method has been successfully applied to astronomical data with multiple object types. DEMS-SVDD [41] constructs SVDD boundaries in multi-modal subspaces, combining distribution-entropy and graph regularization to prevent dimensional col-

lapse and handle heterogeneous modality dimensionalities and distributions. Similarly, MSVDD [42] captures complementary structures across modalities by optimizing multiple Euclidean hyperspheres, and supports kernel extensions to model nonlinear boundaries while avoiding degenerate solutions.

Advantages. Boundary-constrained methods provide explicit, interpretable decision boundaries, directly mitigating ambiguity and offering distance-based anomaly scores. Focusing solely on normal compactness, they handle anomaly diversity and scarcity effectively, which avoids over-generalization by enforcing the building boundary.

Limitations. The compactness assumption can break down for heterogeneous or non-stationary data, yielding loose or unstable decision boundaries. Moreover, multimodal boundary objectives are still sensitive to encoder capacity gaps, which can cause *modality dominance* and *collapse* in weaker branches. Thus, compactness constraints alone often require explicit alignment and robust fusion (Sec. III-B).

B. Cross-modal Representation Learning

In MMAD, cross-modal heterogeneity can place modality-specific features in disparate spaces, worsening modal discrepancy (CH1) and causing uneven sensitivity across modalities (CH3). In many cases, CH3 is driven by granularity gaps, e.g., aligning dense visual patches with coarse global text or sparse sensor streams. Assuming that normal samples are cross-modally consistent while anomalies induce conflicts, cross-modal representation learning detects anomalies by modeling inter-modality relations and scoring deviations from the learned consistency.

These approaches typically encode each observed modality and then fuse or interact in the representation space:

$$z = \Theta\left(\{\phi_k(\mathbf{x}^{(k)})\}_{k \in \mathcal{M}}; \mathcal{M}\right), \quad (5)$$

where $\phi_k(\cdot)$ denotes the encoder for modality k and $\Theta(\cdot)$ is an interaction operator (e.g., aggregation, attention, message passing, or shared-space projection). The resulting z is scored by cross-modal disagreement or distance to a learned normal region. We broadly group existing methods into two families: cross-modal semantic synchronization and local consistency verification.

1) *Cross-modal Semantic Synchronization:* This paradigm aims to synchronize high-level global semantics by mapping features from heterogeneous modalities into a shared semantic space, thereby achieving high-level consistency across modalities (CH1).

Cross-modal fusion strategies. Fusion-based methods instantiate the interaction operator $\Theta(\cdot)$ by combining modalities in three canonical ways [43]. (1) *Early fusion* aggregates raw inputs before encoding, e.g., $z = \phi(\text{Concat}(\{\mathbf{x}^{(k)}\}_{k \in \mathcal{M}}))$. It preserves fine-grained correlations [44] but is most sensitive to modal discrepancy (CH1) (e.g., temporal asynchrony and heterogeneous dimensionality). (2) *Late fusion* first computes modality-specific decisions and then aggregates them at the score level, e.g., $S = \Theta_{\text{score}}(\{f_k(\phi_k(\mathbf{x}^{(k)}))\}_{k \in \mathcal{M}})$. This design can reduce modality dominance in multi-sensor settings [45], but it weakens early semantic interaction and may miss

anomalies expressed primarily as cross-modal inconsistency. (3) *Hybrid/Intermediate fusion* balances the above extremes by introducing interactions across intermediate layers, e.g., $z^{(l)} = \Theta^{(l)}(\{z_k^{(l-1)}\}_{k \in \mathcal{M}})$. Although more computationally demanding, it is often most effective for complex inter-modal dependencies in MMAD; representative methods include M3DM [18], adaptive unbalanced fusion designs [46], [47], and attention-guided 2D/3D decoders [48].

Cross-modal alignment. Cross-modal alignment reduces modality discrepancy by mapping heterogeneous representations into a shared semantic space via contrastive objectives or feature mapping, often building on CLIP-style vision–language pretraining [49]. By bridging representation spaces without explicit feature fusion, anomalies can be exposed as semantic misalignment or low agreement between modalities. Prompt-based adaptation further tailors the aligned space to downstream localization and detection, as in VadCLIP, AdaCLIP, and AA-CLIP [50]–[52]. To alleviate sensitivity imbalance from anomaly diversity (CH3), MMAD increasingly uses hierarchical alignment to address *granularity imbalance*: *global* clip–text matching captures macro semantic deviations, while *token-level* patch–word alignment reveals subtle localized defects. Beyond CLIP-style alignment, related designs include correspondence learning via cross-modal feature mapping [6], decoupled contrastive semantic alignment [53], and entropy-regularized alignment/optimization that sharpens normal structures [54], [55].

However, *uneven modality maturity* can induce *representation collapse*: with large capacity gaps (e.g., a foundation vision backbone vs. a shallow sensor encoder), contrastive training such as InfoNCE may push the weaker modality toward trivial alignment and suppress anomaly-relevant cues. Mitigations include asymmetric optimization (e.g., freezing the mature encoder and adapting the weaker branch via lightweight adapters) and auxiliary self-supervised reconstruction to preserve the weaker modality’s structure.

Advantages. This paradigm bridges modality heterogeneity by synchronizing high-level global semantics and effectively leveraging complementary information across modalities to build robust normal representations. As a result, it generally improves the model’s ability to distinguish diverse types of anomalies.

Limitations. This paradigm relies on strong cross-modal consistency among normal samples. Although effective at bridging high-level heterogeneity (CH1), global synchronization can smooth out local details and reduce sensitivity to fine-grained anomalies (CH3), especially when violations occupy only a small fraction of the input. In practice, weak/noisy relations or missing/imbalanced modalities can cause alignment and fusion to break down, limiting generalization. Deployment is also challenging: dual-stream foundation backbones are compute- and memory-intensive, and transformer-based synchronization often scales as $\mathcal{O}(N^2)$ with sequence length. Dense fusion further enlarges the feature space and memory footprint, complicating real-time use on resource-constrained edge devices.

2) *Local Consistency Verification*: This paradigm establishes precise correspondences or dynamic constraints at the level of image patches, points, or regions. By shifting from

global semantic mapping to token-level verification, these methods directly address the granularity imbalance (CH3) discussed earlier. By emphasizing fine-grained verification, these methods become more sensitive to diverse and complex anomalies (CH3), enabling more accurate anomaly localization and improved overall detection performance.

Fine-grained correspondence matching. This design exploits geometric or structural priors in normal samples to enforce precise local alignments and reveals anomalies through matching failures or high correspondence errors. Such methods are commonly used in multi-modal scenarios with strong spatial correspondences, such as RGB–depth, image–point cloud, and 3D industrial inspection settings. Typical approaches include FiLo++ [56], a zero-/few-shot method that fuses fine-grained descriptions with deformable localization for patch-level cross-modal matching in industrial and medical anomaly detection. MPN [57], CPMF [58] and Bergmann et al. [59] enable robust point-to-point matching and local consistency verification in 3D point clouds. In RGB–depth scenarios, DADA [60] enables learning a general discrete latent space that jointly models RGB and 3D data for 3D surface anomaly detection. M3DM-NR [61] incorporates patch-level correspondence verification between noisy RGB and depth modalities to resist interference and capture fine-grained geometric discrepancies.

Attention-guided consistency constraint. Attention-guided consistency constraint methods dynamically enforce local consistency through cross-modal attention mechanisms or regularization. This allows adaptive discovery of semantically related regions without strict geometric priors. Such methods often leverage cross-attention to capture inter-modal dependencies and constrain attention maps or activations to remain consistent across modalities. Recent works such as Patel et al. [62], MOTCat [63], Gu et al. [64], and Su et al. [65] introduce cross-modal attention to align semantic features for guiding local consistency. PathoGraph [66] leverages structured clinical records, temporally evolving symptom graphs, and medical ontologies to build semantically interpretable latent spaces. ASMF [67] develops cross-modal consistency regularization with aesthetic similarity constraints.

Advantages. Local Consistency Verification improves MMAD by providing fine-grained sensitivity to subtle local deviations that are often missed by global semantic synchronization, thereby enhancing anomaly localization and overall detection accuracy. By aligning modalities at the finest possible granularity (e.g., patch-to-token), it prevents high-density information in one modality from being suppressed by lower-density signals in another, effectively mitigating the granularity mismatch in CH3.

Limitations. It relies heavily on reliable local correspondences or stable attention patterns in normal samples. When modalities are strongly misaligned, noisy, or partially missing, these assumptions may no longer hold, which can lead to reduced performance in real-world heterogeneous data. In addition, such methods often introduce high computational cost due to fine-grained matching or attention operations, which limits their scalability to high-resolution or large-scale multi-modal inputs.

C. Domain Knowledge Augmentation

To address the limitations of knowledge deficit (CH2) and concept drift (CH5), this approach incorporates structured knowledge priors into the normality assumption framework, thereby constructing more compact and robust representations of normal patterns. By introducing knowledge constraints, these methods guide models to capture the underlying semantic logic of normality in complex and heterogeneous multi-modal data, as follows:

$$z = \Psi(\{\phi_k(\mathbf{x}^{(k)})\}_{k \in \mathcal{M}} | \mathcal{K}) \quad (6)$$

where Θ denotes the approach exploiting knowledge, $\mathcal{K}$ denotes external knowledge or semantic priors. According to the manner in which domain knowledge is introduced and exploited, existing approaches can be broadly categorized into two types: cross-modal knowledge distillation and Prototype Memory Mechanism approaches.

1) *Cross-modal Knowledge Distillation*: This paradigm aims to transfer the discriminative and semantic knowledge contained in a more informative modality to the target modality or a lightweight student model, thereby injecting domain priors, enhancing the model's ability to characterize the discrimination boundary (CH2), and improving its robustness in concept drift (CH5).

Logit-level knowledge distillation. Logit-level knowledge distillation transfers softmax logits or soft probabilities from a teacher model trained with richer multi-modal inputs to a student model operating on a target or single modality, enabling the student to inherit inter-class relationships and refine decision boundaries. Early work, such as LCKD [68] and MMRD [69] introduce distillation paradigm by teacher logits knowledge to handle missing modalities during both training and inference. C²KD [70] employs bidirectional distillation with dynamic on-the-fly selection to generate customized soft labels, effectively bridging the modality gap. CorrKD [71] proposes a correlation-decoupled schema that models inter-sample, inter-category, and inter-response correlations in logits. RichKD [72] fuses logits from a dataset-specific unimodal teacher with cross-modal predictions from a pre-trained CLIP model, enriching supervisory signals with semantic diversity.

Feature-level knowledge distillation. Feature-level knowledge distillation transfers cross-modal structural and semantic information by aligning the intermediate feature representations of teacher and student models. Typical approaches include VeXKD [21], TRD [73], A2RD [74], and CMDIAD [75], which employ cross-modal reverse distillation to enhance inter-modality interactions, aiming to selectively transfer beneficial spatial features to single-modal students. FD-CMKD [76] disentangles modality-generic and specific information in the frequency domain to facilitate balanced knowledge transfer across heterogeneous modalities. MUKDF [77] incorporates modality-uncertainty-aware mechanisms with dual-branch extraction and counterfactual reasoning for robust feature distillation in multi-modal sentiment analysis under missing modalities.

Advantages. Cross-modal knowledge distillation injects domain priors by transferring discriminative and semantic in-

formation from informative modalities to target models, improving the characterization of normality decision boundaries. Logit-level methods are lightweight and easy to integrate, while feature-level variants provide deeper structural knowledge, often yielding state-of-the-art results.

Limitations. Distillation performance strongly depends on teacher quality and training-time modality availability, which may hinder robustness in highly dynamic multi-modal environments. The inherent modality heterogeneity can lead to information loss, misalignment, or diluted anomaly signals in fused teacher representations.

2) *Prototype Memory Mechanism*: By maintaining a prototype memory consisting only of normal patterns as a domain prior, this paradigm constrains the discrimination boundary (CH2), which assumes that the feature distribution of normal samples can be aggregated into a finite number of representative prototypes in the embedding space.

Memory bank with prototype clustering. This paradigm constructs a compact representation of normality by extracting feature embeddings from training samples and applying clustering algorithms to derive a finite set of representative prototypes. Typical methods include CIF [78] and CPIR [79], which employ memory to perform prototype clustering that models high-order inter-modal relationships, effectively capturing structural commonality in multi-modal anomaly detection. MTRMB [80] constructs a multi-modal task representation memory bank via key-prompt-guided cross-modal feature interaction and clustering. Information-theoretic extensions further refine prototype quality by minimizing structural entropy to capture intrinsic relational organization in normal data, as demonstrated in hypergraph-based recommendation pre-training [81] and hierarchical text classification optimization [82].

Dynamic memory mechanisms. Dynamic memory mechanisms adaptively refine or update stored prototypes and memory entries to accommodate intra-class variations, concept drift, and streaming data, enabling continuous refinement of normality representations without requiring full model retraining. Recent work, such as DPU [83], which introduces a plug-and-play framework for multi-modal out-of-distribution detection that dynamically adjusts class-specific prototypes based on sample variance and multi-modal prediction discrepancies. DNPR [84] proposes dynamic normal prototype refinement integrated with progressive masked geometric registration and vision-language models. FastRef [85] develops an efficient test-time prototype refinement framework for few-shot industrial anomaly detection, iteratively optimizing transport probabilities for anomaly suppression and transformation matrices for characteristic transfer from query images. EG-MPC [86] enables high-quality frame prediction and reconstruction under background interference by adaptively refining normality patterns across scales in multi-modal video streams.

Advantages. This paradigm effectively incorporates multi-modal domain knowledge through clustering or dynamic updates, improving robustness to rare anomalies and unbalanced detection.

Limitations. Despite their effectiveness, prototype-based memory mechanisms face a critical trade-off between memory

capacity and representational diversity. Insufficient capacity limits the coverage of normal variations, whereas excessive capacity increases the risk of memorizing anomalous samples or noise, thereby degrading the purity of normality modeling.

IV. METHODS BASED ON ABNORMALITY ASSUMPTION

The abnormality assumption complements the normality view by explicitly constructing the “flip side” of the data distribution. Rather than only reconstructing or enforcing self-consistency on normal data, it leverages domain knowledge to synthesize contrastive signals that expose deviations. This paradigm differs from standard self-supervised learning: while both use transformations (masking, perturbation, permutation), self-supervision learns *invariant* representations, whereas pseudo-anomaly generation creates *hard negatives* to make detectors *sensitive* to abnormality.

Given a normal observation $\mathcal{X}_M$, an injected sample is generated as $\tilde{\mathcal{X}}_M = T(\mathcal{X}_M)$. The detector enforces $f(\mathcal{X}_M) < f(\tilde{\mathcal{X}}_M)$ by minimizing

$$\mathcal{L}_{AA} = \mathbb{E}_{\mathcal{X}_M \sim \mathcal{D}_n} [\ell(f(\mathcal{X}_M), f(\tilde{\mathcal{X}}_M))]. \quad (7)$$

Here $\ell(\cdot, \cdot)$ can be a ranking loss or a binary classification loss treating injected samples as pseudo anomalies. Injected samples are not meant to match the true anomaly distribution; they probe outside normal support to tighten the boundary.

We review abnormality-assumption methods in three levels (Fig. 3): *coarse-grained injection* (physical intuition), *structural injection* (relational disruption), and *semantic injection* (knowledge-guided abnormality), progressing from heuristic to semantically grounded boundary learning. A shared bottleneck is validating the *rationality* of pseudo-anomalies, which becomes harder from coarse to semantic injection. Coarse/structural schemes typically enforce heuristic or physical constraints (e.g., CutPaste defects within the foreground), whereas semantic completion needs different criteria. Under extreme scarcity in CH4, metrics such as FID may overfit, motivating reference-free measures (e.g., Inception Score, IC-LPIPS [87]) with strict alignment checks; for time series/video, rationality must also respect physical constraints and temporal inertia.

A. Coarse-grained Injection

Relying solely on reconstruction of normal data can lead to a loose decision boundary, as discussed in CH4. Coarse-grained injection offers a direct way to “complete the flip side” of the data distribution by exposing the detector to low-level corruptions or externally sourced outliers. By contrasting clean samples with their corrupted counterparts, these methods promote a sharper separation between normal and abnormal regions. Crucially, coarse-grained injection largely bypasses data heterogeneity in CH1: simple operations such as masking, cutout, and noise injection are modality-agnostic and can be applied to images, time series, and text without semantic alignment. Conceptually, the transformation operator $T(\cdot)$ covers two complementary branches. The first is *internal perturbation*, which applies unstructured, data-level augmentations within each observed modality, namely $T(\mathcal{X}_M) =$

$\{\mathcal{A}(\mathbf{x}^{(k)})\}_{k \in \mathcal{M}}$, where $\mathcal{A}(\cdot)$ may instantiate additive Gaussian noise $\mathbf{x}^{(k)} + \epsilon$, Mixup, or adversarial perturbations. The second is *outlier exposure*, which samples proxy anomalies from an auxiliary external dataset, namely $\tilde{\mathcal{X}}_M^{\text{out}} \sim \mathcal{D}_{\text{out}}$. Together, these injections provide coarse contrasts that regularize the boundary and improve robustness to unseen outliers.

1) *External Exposure*: A direct way to address CH4 is to introduce proxy anomalies from external sources. This subclass follows an open-world intuition: any data lying outside the training distribution can be treated as anomalous.

Foundations in unimodal settings. In the CV domain, Outlier Exposure (OE) [90] shows that training detectors with an auxiliary outlier dataset improves generalization to unseen anomalies. Exposure to many “foreign” examples effectively compacts the decision boundary around normal data. Similarly, Deep SAD [34] incorporates a small number of external samples, indicating that even limited auxiliary supervision can substantially increase latent-space separation.

Building on OE, subsequent studies extended external exposure to other unimodal domains before moving to multi-modal scenarios. Latent Outlier Exposure (LOE) [91] revisits this paradigm under contaminated training data, using auxiliary outliers as latent-space regularizers to prevent representation collapse; it remains effective in other contaminated settings [37]. Domain-specific variants further adapt the idea to time series [92], robotic vision [93], and anomalous sound detection [94] by introducing exogenous temporal patterns, visually irrelevant scenes, or auxiliary audio as proxy anomalies. In graph-structured data, HGOE [95] combines cross-domain graph outliers with boundary-aware exposure losses in a model-agnostic manner, and has inspired multi-modal adaptations.

Extensions to multi-modal scenarios. As anomaly detection increasingly involves multi-modal data, external exposure has been extended to cross-modal settings. Recent works [96] show that simple multi-modal outliers, including cross-modal mismatches and random pairs, can serve as effective proxy anomalies for both detection and segmentation. These approaches leverage inter-modality misalignment as a strong negative signal, avoiding costly outlier collection and domain-specific annotation. To mitigate representation bias when the auxiliary exposure set is noisy or partially in-distribution, Taylor Outlier Exposure [97] proposes a Taylor-expanded loss to selectively downweight unreliable samples, addressing cross-modal contamination in MMAD. More broadly, the survey in [98] summarizes extensive OE-based methods in visual inspection, while emerging multi-modal benchmarks increasingly adopt exposure-style regularization to better align representations across inputs.

Advantages. External exposure provides a direct, scalable, and largely modality-agnostic remedy for anomaly scarcity, and it integrates easily with a wide range of detectors and inputs. In multi-modal settings, it supplies a unified negative signal that encourages cross-modal alignment and improves generalization to unseen anomalies without requiring task-specific annotations.

Limitations. External exposure remains coarse-grained, offering only weak, exclusion-based supervision that may miss

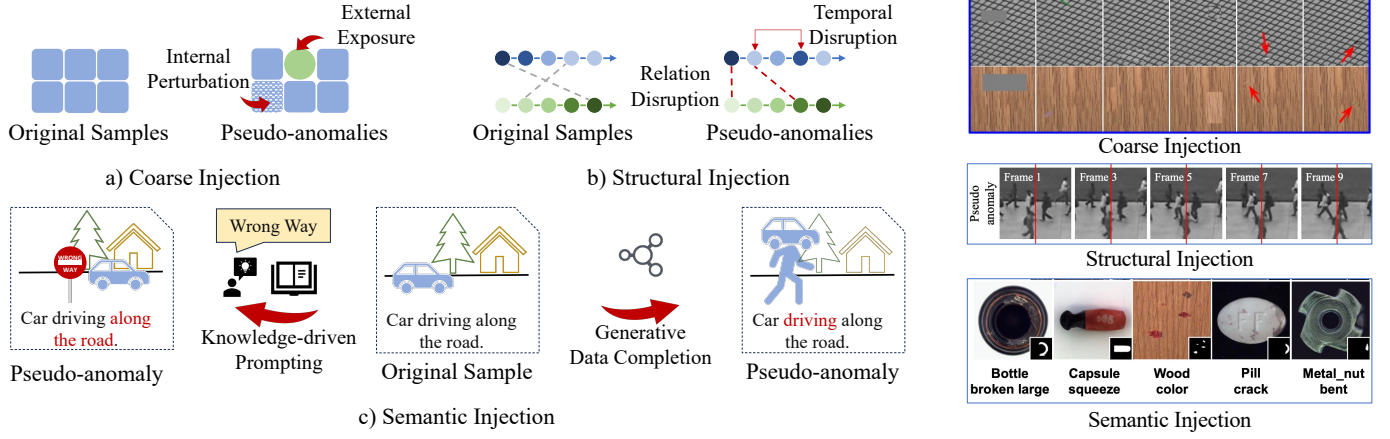


Fig. 3: Overview of abnormality-assumption methods that inject pseudo anomalies at three granularities: a) *coarse-grained injection* corrupts low-level statistics via within-modality perturbations or outlier exposure; b) *structural injection* breaks relational integrity by disturbing links or temporal order within/across modalities; and c) *semantic injection* synthesizes high-level logical inconsistencies using external knowledge or generative models. The right panel shows illustrative pseudo-anomaly visualizations adapted from CutPaste [88], STEAL Net [89], and AnomalyDiffusion [87].

complex deviations or subtle cross-modal mismatches. Its effectiveness is also sensitive to exposure quality: poorly aligned or contaminated auxiliary sets can distort the shared boundary.

2) *Internal Perturbation*: This subclass follows the physical intuition that “corrupted is anomalous.” Rather than relying on external data, internal perturbation synthesizes pseudo-anomalies by locally disrupting continuity or statistics in the observed data. It is particularly effective at bypassing data heterogeneity in CH1, since operations such as masking and shuffling are modality-agnostic primitives.

Foundations in unimodal settings. Internal perturbation originates from data augmentation, later repurposed to generate pseudo anomalies for unsupervised detection, most prominently in vision. Early masking-based methods (e.g., Hide-and-Seek [99], Cutout [100], GridMask [101], Random Erasing [102]) remove or corrupt local regions and can serve as effective pseudo anomalies. More realistic defects are created via patch manipulation (e.g., CutPaste [88]), while interpolative perturbations such as Mixup [103] broaden the boundary in input/feature space and have been validated beyond vision, including physiological and time-series settings [104], [105].

Extensions to multi-modal scenarios. Internal perturbation extends naturally to MMAD because it relies on simple, modality-agnostic operations and requires minimal domain knowledge. Although fewer works exist than in unimodal vision, recent studies have started adapting masking and perturbation strategies to heterogeneous inputs. For example, SeMANd [106] applies self-supervised masking to paired vector geometries and imagery, creating pseudo-anomalies by corrupting one modality while maintaining cross-modal consistency checks. Related designs use random masking or feature-level misalignment to exploit inter-modality inconsistencies as proxy signals. In industrial and security contexts, autoencoder-based fusion models follow the same intuition. Khan et al. [107] apply reconstruction perturbations across sensors for real-time vehicle-damage detection, while inte-

grated deep learning autoencoders [108] use coarse internal corruptions on diverse streams for cloud anomaly classification. Extensions of CutPaste-style operations to multi-modal fusion and adversarial internal augmentations [109] further show promise for industrial and cyber-physical systems.

Advantages. Internal perturbation offers a simple, computationally efficient, and highly modality-agnostic remedy for anomaly scarcity by generating diverse pseudo-anomalies from in-distribution data. These coarse heuristics capture local deviations effectively and avoid the domain-shift risks of external exposure; in MMAD, they can also encourage cross-modal consistency and robustness.

Limitations. Internal perturbation remains heuristic and coarse-grained, and may miss subtle semantic or structurally complex anomalies, particularly when fine-grained inter-modality dependencies matter. Performance can be sensitive to hyperparameters, and unrealistic artifacts may hurt generalization.

B. Structural Injection

Structural injection extends coarse-grained perturbations under the intuition that *disrupted structure is anomalous*. It synthesizes pseudo-anomalies by violating temporal or relational regularities of normal data, enabling multi-scale disruptions (from local subsequences to global relations) to address granularity imbalance (CH3) and improve robustness under drift (CH5). Concretely, structural injection perturbs the multi-modal observation $\mathcal{X}_{\mathcal{M}}$ via a structure-disrupting transformation $T(\cdot)$. Typical instantiations include: (i) *temporal disruption*, which permutes sequential content within each observed modality, i.e., $T(\mathcal{X}_{\mathcal{M}}) = \{\mathbf{x}_{\pi}^{(k)}\}_{k \in \mathcal{M}}$ where π denotes a permutation operator (e.g., shuffling or reversal); and (ii) *relational disruption*, which replaces local neighborhoods or links, i.e., $T(\mathcal{X}_{\mathcal{M}}) = \{\mathbf{M}^{(k)} \odot \mathbf{x}^{(k)} + (1 - \mathbf{M}^{(k)}) \odot \psi(\mathbf{x}^{(k)})\}_{k \in \mathcal{M}}$, where $\mathbf{M}^{(k)}$ is a relational mask and $\psi(\cdot)$ performs a local relational replacement (e.g., edge dropping or

patch swapping). These contrasts guide the model to place the decision boundary around regions of structural incoherence, rather than around marginal feature deviations.

1) *Temporal Structure Disruption*: This sub-subclass targets the temporal consistency inherent in sequential data, operating on the intuition that “disordered chronology is anomalous.” By explicitly violating the natural order, speed, or predictive continuity of normal sequences, these methods generate pseudo-anomalies that simulate contextual deviations, offering finer granularity than coarse masking while remaining computationally tractable.

Foundations in unimodal settings. The paradigm of temporal structure disruption originated in unimodal time series and video analysis, where sequential integrity serves as a strong supervisory signal. Early Transformer-based approaches like TranAD [23] introduced attention masking and subsequence perturbations to enforce robust temporal encoding. Anomaly-BERT [110] and RedLamp [111] employ data degradation or transformation that combine masking with explicit order disruption and future prediction failures, creating diverse pseudo-anomalies for self-supervised learning. CutAddPaste [112] further refines this by cutting patches informed by abnormal knowledge and pasting them into normal contexts, deliberately breaking local temporal continuity to mimic realistic deviations. In video anomaly detection, synthetic temporal perturbations [89], [113] leverage snippet shuffling and reversal as negative augmentations to amplify temporal inconsistencies, while recent reviews [114] consolidate these order-based disruptions as foundational techniques for sequential representation learning.

Extensions to multi-modal scenarios. In multi-modal contexts, temporal disruption becomes particularly powerful by exploiting cross-modal consistency as an additional structural cue. Deliberate misalignment between modalities including such as audio-video temporal offsets or sensor stream desynchronization, generates highly informative pseudo-anomalies that reflect real-world multi-modal incoherence. Fusion frameworks [115] introduce controlled audio-video temporal misalignments during training to strengthen joint representations in complex environments. Similarly, Multi-modal VAD [116] incorporates audio-vision-language streams, using modality-specific temporal shifts as negative signals for intelligent monitoring systems. Recurrent architectures [117] combine spatio-temporal audio-visual features, simulating anomalies via selective shuffling or delay injection across modalities. Applications in crowd analysis [118] further demonstrate the efficacy of temporal inconsistencies between visual trajectories and contextual cues in open-access multi-modal setups.

Advantages. Temporal structure disruption provides finer-grained supervision than coarse perturbations, effectively capturing contextual and sequential anomalies (CH3) while enhancing robustness to distribution shifts (CH5) through structure-aware regularization. In multi-modal settings, cross-modal misalignment offers a semantically rich negative signal, promoting synchronized representations and improved generalization to real inter-modal deviations.

Limitations. Despite its advantages, reliance on explicit temporal assumptions can limit applicability to non-sequential

data, and aggressive disruptions risk introducing artifacts that deviate from realistic anomalies. Sensitivity to disruption strength (e.g., shuffle intensity or offset magnitude) remains a challenge, particularly in multi-modal scenarios where subtle misalignments dominate but are harder to calibrate consistently.

2) *Relational Structure Disruption*: This sub-subclass emphasizes the relational topology inherent in interconnected data, operating on the intuition that “fractured connections are anomalous.” By deliberately perturbing the dependencies among entities, such as through edge deletions, relation shuffling, or subgraph masking, these methods generate pseudo-anomalies that simulate network-level inconsistencies, addressing granularity imbalance (CH3) by capturing anomalies at varying relational scales while countering concept drift (CH5) via topology-robust regularizations.

Foundations in unimodal settings. Relational structure disruption stems from graph anomaly detection, where self-supervision uses topological perturbations (edge/attribute edits, relation shuffling) as supervisory signals. Representative methods include DAGAD [119] and related robustness-oriented designs [120]; in video, spatio-temporal perturbations can be interpreted as breaking structured dependencies [113]. Later work strengthens this idea with contrastive and adversarial perturbations, e.g., multi-scale contrastive views [121], dual-bootstrapped training [122], adversarial/multiplex edge dropping [123], [124], and unified perturbation frameworks [125], [126].

Extensions to multi-modal scenarios. In multi-modal contexts, relational disruption extends to dependencies among modals, such as perturbing alignments between visual graphs and textual relations or fusing point cloud topologies with image features. This creates pseudo-anomalies that exploit relational voids across modalities, like mismatched object predicates or desynchronized sensor graphs. Zhang et al. [127] disrupts cross-modal relations via matching cost filtering in point cloud-image pairs. Generative approaches [128] incorporate relation perturbations in multi-modal node attributes for semi-supervised learning. These techniques demonstrate the adaptability of relational disruption in fusing heterogeneous structures, enhancing detection in scenarios like industrial inspection or autonomous systems.

Advantages. Relational structure disruption captures intricate inter-entity dependencies at multiple scales (CH3), yielding interpretable pseudo-anomalies that bolster robustness to evolving distributions (CH5). In multi-modal settings, it uncovers and makes use of cross-modal inconsistencies, fostering unified representations that generalize across fused relational spaces.

Limitations. These methods assume well-defined relational structures, constraining applicability to sparse or unstructured data, and perturbation designs (e.g., drop ratios) may bias toward specific anomaly types. In multi-modal applications, aligning disruptions across modalities introduces calibration challenges, potentially amplifying noise in high-dimensional relational contexts.

C. Semantic Injection

In many real-world tasks, anomalies are semantic deviations that violate domain knowledge or contextual plausibility (CH2), which low-level or purely structural perturbations often fail to capture. Enabled by generative and multi-modal foundation models, semantic injection creates knowledge-guided contrasts to sharpen anomaly definitions and partially alleviate scarcity (CH4). It can further incorporate external knowledge or semantic priors $\mathcal{K}$ to construct stronger, context-violating contrasts. Typical instantiations include (i) *knowledge-driven prompting*, $T(\mathcal{X}_M) = \mathcal{X}_M \oplus f_{\text{LLM}}(\mathcal{K}_{\text{prompt}})$, where f_{LLM} generates rule-violating prompts guided by $\mathcal{K}_{\text{prompt}}$; and (ii) *generative data completion*, $T(\mathcal{X}_M) = f_{\text{gen}}(\mathbf{M} \odot \mathcal{X}_M; \mathcal{K}_{\text{gen}})$, where f_{gen} hallucinates out-of-distribution content under generative priors $\mathcal{K}_{\text{gen}}$. These contrasts tighten the decision boundary by emphasizing incoherence over marginal feature deviations.

1) *Knowledge-driven Prompting*: This sub-subclass follows the intuition that “semantically implausible is anomalous.” It is increasingly enabled by foundation models, ranging from contrastive vision–language models (VLMs) to generative multi-modal large language models (MLLMs). By injecting semantic priors through designed or learned prompts, these methods synthesize knowledge-guided pseudo-anomalies that violate domain-specific logic or contextual plausibility, without explicitly modeling structural relations.

Foundations in unimodal settings. Knowledge-driven prompting emerged from zero-/few-shot anomaly detection, where VLMs offer a semantic space to define abnormality via text. WinCLIP [14] introduced class-specific negative prompts, while AnomalyCLIP [129] learns object-agnostic prompts from normal data to reduce manual engineering. Later works sharpen this semantic boundary via simpler or learned prompting and context cues, including SimCLIP [130], PromptAD [131], and VCP-CLIP [132]; larger-model variants (e.g., AnomalyGPT [133], AA-CLIP [52]) improve reasoning and robustness, with UniAD [134] further stabilizing prompts across domains.

Extensions to multi-modal scenarios. In multi-modal contexts, knowledge-driven prompting uses cross-modal semantic priors to expose logical inconsistencies, enabling the synthesis of semantically implausible pseudo-anomalies by reasoning over multiple data streams. For example, Negative Prompting [135] uses negative descriptions to simulate anomalous patterns in a multi-modal latent space, while prior work [136] prompts large VLMs to surface cross-modal conflicts (e.g., images contradicting their textual descriptions).

Recently, integrating MLLMs has shifted the focus from semantic alignment to high-level reasoning. VMAD [137] exploits the spatial–semantic capabilities of MLLMs for location-aware zero-shot detection, and Holmes-VAD [138] uses MLLMs to generate unbiased, explainable justifications for video anomalies, moving beyond numerical scores. As highlighted in recent work [139], MLLM reasoning chains offer a conceptual roadmap for advancing MMAD from “perception” to “understanding,” where cross-modal logical conflicts serve as explicit evidence of abnormality.

Advantages. Knowledge-driven prompting provides seman-

tically rich supervision (CH2) by directly incorporating domain knowledge through VLMs, reducing definition ambiguity and enabling precise anomaly localization. This paradigm is also more explainable than normality-based reconstruction: whereas reconstruction error is a “black-box” signal, semantic injection offers explicit contrastive evidence. By comparing inputs to predefined abnormal concepts, these methods yield human-readable justifications, crucial for high-stakes big data applications. It further mitigates class imbalance (CH4) by generating diverse pseudo-anomalies on demand, supporting zero-/few-shot scenarios and strong cross-domain generalization without task-specific training data.

Limitations. These methods depend heavily on pre-trained LLM quality, limiting performance in domains with poor semantic coverage or noisy prompts. Prompt engineering is sensitive to design choices, and large models incur high computational costs. In multi-modal settings, maintaining consistent semantic alignment across modalities is especially difficult for rare or emerging anomaly types.

2) *Generative Data Completion*: This sub-subclass employs generative architectures to synthesize semantically implausible continuations or completions from normal data, operating on the intuition that “semantically incoherent synthesis is anomalous.” By conditioning generation on normal samples while introducing controlled violations of logical or contextual consistency, these methods produce high-fidelity pseudo-anomalies that capture subtle, knowledge-driven irregularities.

Foundations in unimodal settings. Generative data completion began with GAN-based methods for synthesizing anomalies in unimodal industrial images. Early works such as Doping [140], Old is Gold [141], G2D [142], Defect-GAN [143], and DSG [144] use GANs to augment training with synthetic defects, improving reconstruction-based detectors by exposing them to generated outliers. Few-shot Defect Generation [145] extends this with defect-aware feature manipulation for limited data. Diffusion models then enabled higher-quality synthesis. Building on DDPM [146], AnoDDPM [147] employs simplex noise for robust anomaly scoring. Recent diffusion frameworks such as RealNet [148], AnomalyDiffusion [87], Multi-Class Diffusion [149], CAGEN [150], and DCP [151] provide controllable, few-shot anomaly generation with superior fidelity and diversity for unsupervised detection.

Extensions to multi-modal scenarios. Multi-modal extensions use generative models to synthesize cross-modal inconsistencies, such as implausible combinations of visual, depth, or sensor data. AnomalyXFusion [152] applies diffusion to fuse image-text-mask modalities and create semantically violating samples. MMCD [153] conditions generation on skeleton motion for video anomaly detection, while MAGE-ID [154] uses GAN–diffusion hybrids to synthesize minority-class threats in network–sensor data. Recent work further integrates LLMs to generate profile-conditioned yet semantically anomalous behaviors from historical user traces, enabling proactive graph evolution and adversarial detection [155]. These methods demonstrate that generative models can produce realistic multi-modal pseudo-anomalies, improving detection in industrial and surveillance settings.

TABLE I: Quantitative performance and theoretical taxonomy of representative MMAD methods across diverse domains.

Method	Modality	Paradigm	FM	Key Mechanism & Intuition	Representative Benchmark	Performance (Metrics)
<i>Image Anomaly Detection (Industrial, Medical & General Vision - Metrics: Image/Pixel-AUROC)</i>						
CutPaste [88]	I	AA (Coarse)	-	Heuristic patch self-augmentation	MVTec AD	95.2% (I-AUROC)
TaylorOE [97]	I	AA (Coarse)	-	Polynomial Taylor-expansion OE	CIFAR / ImageNet	99.3%+ (AUROC)
M3DM [18]	I+3D	NA (Alignment)	✓	Point-to-feature semantic alignment	MVTec 3D-AD	94.5% (I-AUROC)
WinCLIP [14]	I+L	AA (Semantic)	✓	Compositional prompt ensembling	MVTec AD / VisA (Zero-shot)	91.8% / 78.1% (AUROC)
PromptAD [131]	I+L	AA (Semantic)	✓	Knowledge-guided text descriptions	MVTec AD (Few-shot)	94.6% (I-AUROC) 95.9% (P-AUROC)
FastRef [85]	I	NA (Knowledge)	✓	Dynamic prototype memory distillation	MVTec AD	93.8% (I-AUROC) 95.7% (P-AUROC)
AdaCLIP [51]	I+L	NA (Alignment)	✓	Adaptive prompt-based fine-tuning	MVTec AD	89.2% (I-AUROC) 88.7% (P-AUROC)
AnomalyDiffusion [87]	I+L	AA (Semantic)	✓	Generative pseudo-anomaly synthesis	MVTec AD / VisA	99.1% (P-AUROC) 99.2% (I-AUROC)
MMRD [69]	I+3D	NA (Knowledge)	-	Teacher-student knowledge distillation	MVTec 3D-AD	95.0% (I-AUROC) 97.6% (PRO)
<i>Video Anomaly Detection (Surveillance & Events - Metrics: Frame-level AP / AUC)</i>						
VadCLIP [50]	V+L	NA (Alignment)	✓	Dual-branch visual-language alignment	XD-Violence	84.5% (AP)
DSANet [53]	V+L	NA (Alignment)	✓	Semantic-aware spatio-temporal modeling	UCF-Crime / XD-Violence	89.44% / 86.95% (AP)
STEAL Net [89]	V	AA (Structural)	-	Breaking temporal regularities	Pre2	98.4% (AUC)
<i>Time-Series, Sensors & IT Systems (Metrics: F1 / Precision)</i>						
TranAD [23]	T	NA (Feature)	-	Attention-based normal reconstruction	SMD	96.1% (PA F1) 99.7% (AUC)
FOCA [38]	T	NA (Feature)	✓	Boundary-constrained FM adaptation	WADI	74.9% (PA F1)
AnomalyBERT [110]	T	AA (Structural)	-	Sequence relational masking	SWaT	85.4% (PW F1) 92.5% (PA F1)
RedLamp [111]	T	AA (Structural)	-	Classification of perturbations	UCR	89.7% (VUS-AUC)
CutAddPaste [112]	T	AA (Structural)	-	Break decomposition structure	UCR	68.2% (RPA-F1)
Note: Modality abbreviations: I (Image), V (Video), L (Language/Text), A (Audio), T (Time-series), 3D (Point Cloud). Assumption abbreviations: NA = Normality Assumption, AA = Anomaly Assumption. Performance varies by metric (e.g., Pixel-AUROC for industrial defects vs. Frame-AP for video events).						

Advantages. Generative data completion produces high-fidelity, semantically coherent pseudo-anomalies (CH2), enabling precise modeling of complex irregularities via controllable synthesis. It mitigates anomaly rarity (CH4) by generating unlimited diverse samples, supporting few-shot and zero-shot generalization and surpassing heuristic methods in reconstruction quality and localization.

Limitations. These methods are computationally expensive, especially diffusion models with iterative sampling, limiting real-time use. GANs still face training instability and mode collapse, and multi-modal extensions require careful cross-modal conditioning to avoid artifacts. Generalization to highly diverse or unseen anomalies is limited by the generator’s training exposure. The computational cost of generative paradigms remains a central bottleneck: diffusion-based methods require iterative denoising with many forward passes, causing extreme inference latency that effectively rules out raw deployment in strict real-time anomaly detection and forces a trade-off between semantic anomaly diversity and detection efficiency.

Summary and FM perspective. Abnormality-based modeling progresses from low-level perturbations to structure-aware disruptions and, ultimately, knowledge-driven semantic reasoning, yielding increasingly tight and interpretable decision

boundaries through explicit contrastive evidence. Across this spectrum, Foundation models—including MLLMs—are best viewed as shared enablers: under the *Normality Assumption* (Sec. III) they serve as universal extractors/aligners that reduce heterogeneity and capacity gaps, whereas under the *Anomaly Assumption* (Sec. IV) they act as semantic generators/reasoners that support promptable pseudo-anomaly synthesis and, increasingly, natural-language rationales for violated cross-modal constraints. Table I summarizes representative methods from both paradigms and their FM integration; the next section then grounds this taxonomy with representative benchmarks and evaluation settings.

V. APPLICATIONS, BENCHMARKS, AND EVALUATION

MMAD mitigates the ambiguity of single-stream observations by integrating heterogeneous sources. Accordingly, this section organizes widely used benchmarks by their physical characteristics, spanning high-precision geometric inspection in industrial systems, life-critical sensing in medical applications, and semantically rich surveillance scenarios. A comprehensive taxonomy of these benchmarks—grouped by modality composition, spatiotemporal scale, and environmental dynamics—is provided in Table II.

TABLE II: Summary of multi-modal and unimodal anomaly detection datasets across diverse application domains.

Dataset	Year	Modality	Target / Highlight	Scale / Size	Anomaly Ratio	Annotation Quality	Dynamics
<i>Industrial & Manufacturing Inspection</i>							
MVTec AD [156]	2019	I	Structural defects (scratches, dents)	5,354 images	~23%	Pixel-level mask	Static
MVTec LOCO [158]	2022	I	Structural + Logical (rule violations)	3,644 images	~16%	Pixel-level mask	Static
MVTec 3D-AD [159]	2021	I+3D	Geometric shapes (holes, protrusions)	4,147 scans	~23%	Pixel & Voxel-level	Static
Real3D-AD [160]	2023	High-res 3D	Largest high-precision 360° dataset	1,250 3D models	~48%	Voxel-level mask	Static
VisA [157]	2022	I	Complex surfaces, multiple instances	10,821 images	~11%	Pixel-level mask	Static
Eyecandies [161]	2022	I+Depth	High-fidelity synthetic data avoiding bias	10,000 images	~25%	Pixel & Depth map	Static
BTAD [162]	2021	I	Real-world body and surface defects	2,830 images	~11%	Pixel-level mask	Static
MPDD [163]	2021	I	Painted metal parts manufacturing	1,346 images	~21%	Pixel-level mask	Static
SDD [164]	2020	I	Surface crack detection (weakly sup.)	400 images	~13%	Pixel-level mask	Static
DAGM [165]	2007	I	Historical reference for optical inspection	8,050 images	~12%	Image-level label	Static
HDFS [167]	2009	L+T	Distributed systems monitoring	24.4M logs	~0.1%	Sequence-level	Static
BGL [166]	2007	L+T	Supercomputer temporal alignment	4.7M logs	~7.3%	Log-level	Static
SWAT [168]	2016	T	Critical infrastructure physical attacks	11 days	~11.9%	Point-level	Static
WADI [169]	2017	T	Water treatment distribution attacks	16 days	~5.9%	Point-level	Concept Drift
SMD [170]	2019	T	Cloud server cluster health	5 weeks	~4.1%	Point-level	Static
<i>Medical Diagnostics & Healthcare</i>							
MMR [171]	2025	I+OCT	Retinal artery occlusion (multi-modal)	370 pairs	28%	Case & Pixel-level mask	Static
IDRiD [172]	2020	I	Diabetic retinopathy micro-lesions	516 images	Imbalanced	Pixel & Image-level	Static
HeadCT [173]	2018	I	Urgent care triage (hemorrhage)	21,586 scans	~12%	Image-level label	Static
BrainMRI [174]	2021	I	3D volume neuro-oncological imaging	8,160 mpMRI scans	100%	3D Voxel-level	Static
Br35H [175]	2025	I	Brain tumor presence (binary class.)	3,060 images	50%	Image-level label	Static
COVID-19 [176]	2020	I	Viral manifestations in lung imaging	3,487 images	~12%	Image-level label	Static
ISIC [177]	2018	I	Skin cancer / Melanoma standard	2,750 images	~19%	Pixel & Image-level & Label	Static
CVC-ClinicDB [178]	2015	I	Colon tumors and polyps on surfaces	612 frames	100%	Pixel-level mask	Static
Kvasir / Endo [179], [180]	2019/2021	I	Multi-class GI tract screening	8,000 images	~12%	Image-level & Pixel-level	Static
TN3k [181]	2021	I	Thyroid nodule segmentation priors	4,510 images	100%	Pixel-level mask	Static
MIT-BIH [182]	2001	T	Foundational bio-electrical series	48 records (116,000 beats)	~25%	Point-level	Static
ECG5000 [183]	2000	T	Large-scale heart signal screening	5,000 seqs	~42%	Sequence-level	Static
MIMIC-III [184]	2016	T+L	Critical care decompensation & risk	61,532 ICU stays	~11%	Patient-level	Concept Drift
<i>Video & Surveillance Analytics</i>							
UCF-Crime [185]	2018	V+Flow	13 types of criminal behavior	1,900 videos	~50%	Video & Temporal-level label	Static
XD-Violence [186]	2020	V+A	Largest violence dataset with audio	4,754 clips	~41%	Video & Frame-level label	Static
ShanghaiTech [187]	2017	V	Multi-scene campus surveillance	316,154 frames	~13%	Pixel & Frame-level label	Static
VAAR [188]	2025	V+A	Synchronized audio-visual cues	3,000 videos	~90%	Video-level label	Static
MSAD [114]	2024	V	Large-scale multi-scenario (14 envs)	1,010,480 frames	~31%	Frame & Video-level label	Static
UCF-Crime-DVS [189]	2025	V	First large-scale DVS benchmark	1,900 clips	~48%	Video & Temporal-level label	Static
Modality abbreviations: I (Image), V (Video), L (Language/Text), A (Audio), T (Time-series), 3D (Point Cloud), Depth (Depth Map), OCT (Optical Coherence Tomography). Annotation Quality: Defines the granularity of evaluation. Dynamics: Indicates whether the dataset assumes a stationary environment ('Static') or contains evolving conditions and non-stationary shifts ('Concept Drift').							

A. Application Domains and Benchmarks

Industrial and cyber-physical systems. Industrial anomaly detection is shifting from localized 2D surface defects to full-view 3D inspection and logical constraint verification, amplifying modal discrepancy (CH1) across heterogeneous 2D/3D representations and weakening the availability of explicit semantic priors (CH2). Representative inspection benchmarks span 2D/3D, depth, and logical settings [156]–[165]. Beyond manufacturing, system software monitoring and cyber-physical process control mix logs with continuous multivariate streams and emphasize long-horizon collective anomalies (CH3) and robustness under non-stationary drift (CH5) [166]–[170].

Medical diagnostics and healthcare. Medical MMAD is shaped by strict requirements on localization granularity and clinical reliability: anomalies are subtle, high-stakes, and often demand pixel/voxel-precise delineation (CH2–CH3). Multi-modal settings frequently combine 2D observations with 3D anatomy or longitudinal records, amplifying cross-modal alignment challenges (CH1) under severe data scarcity and subtype variability (CH4). Representative benchmarks span multi-modal ocular imaging, radiology, dermatology/endoscopy, and physiological monitoring [171]–[182],

[184], [190].

Video and surveillance analytics. VAD is commonly evaluated under weak supervision, where audio and language cues can reduce ambiguity in visually uncertain events (CH2) but introduce additional alignment and synchronization issues (CH1). Benchmarks span public-safety surveillance, audio-visual violence, multi-scenario settings, and event-based sensing, collectively stressing long-horizon modeling and granularity imbalance (CH3) [114], [185]–[189].

Cross-domain characteristics and benchmarking gaps. Table II suggests two recurring gaps. First, annotation *granularity* differs sharply: industrial/medical tasks often need pixel-/voxel-level masks, while surveillance typically provides only frame-level labels. Second, most benchmarks remain overly *stationary*; despite the importance of concept drift (CH5), few datasets reflect long-term non-stationarity as in WADI [169] or MIMIC-III [184].

B. Evaluation Metrics and Protocols

As MMAD spans diverse physical scenarios, a single unified evaluation metric is insufficient. While current evaluation *metrics* are largely inherited from established unimodal domains

to assess the final fused prediction, the emerging testing *protocols* are increasingly tailored to probe unique cross-modal dynamics. The choice of metric is deeply coupled with the underlying modality and the specific *granularity imbalance* (CH3) of the task. Furthermore, the rise of Foundation Models and generative paradigms has necessitated a shift from standard closed-set testing to more complex evaluation protocols.

1) *Muliti-granularity evaluation metrics.*: To rigorously assess detection performance at different resolutions, domains use metrics aligned with their annotation granularity:

- **Spatial localization (image/3D):** In industrial inspection and medical imaging, evaluation should address both coarse detection and precise localization. *Image-level AUROC* is the standard metric for sample-level decisions. For defect isolation, *Pixel-level AUROC* and Area Under the Per-Region Overlap curve (*AUPRO*) [156] are widely used; *AUPRO* is especially suitable for small, localized anomalies because it emphasizes region-wise overlap. In medical applications, overlap-based metrics such as the *Dice score* and *Intersection over Union (IoU)* are commonly reported to quantify structural consistency.
- **Spatio-temporal detection (video/audio):** For visual sequences and audio streams, anomalies stem from the *temporal evolution* of patterns rather than a single frame. Although *Frame-level AUROC* is widely used, mainstream benchmarks such as ShanghaiTech [27] increasingly report *Average Precision (AP)*, which is more informative under severe class imbalance. *Detection Latency* is also reported to measure how quickly a model responds after an anomalous event begins.
- **Sequential dynamics (time-series/sensors):** In industrial sensing and IT monitoring, evaluation targets unimodal or multi-modal temporal dependencies, but metrics remain fragmented due to continuous temporal drift and ambiguous boundaries. Traditional *Point-Wise (PW)* metrics underestimate performance when predictions are slightly time-shifted, while the widely used *Point-Adjusted (PA)* F1-score [170] often overestimates it by rewarding partially detected segments. To address this, several alternatives have been proposed: *Revised Point-Adjusted (RPA)* [22] treats each anomalous segment as a single event, *Affiliation metrics* [191] reward proximity to ground-truth intervals, and *Volume Under the Surface (VUS)* [192] integrates accuracy over varying window sizes to reduce window-induced bias.

2) *Emerging evaluation protocols.*: Beyond static metrics, robust assessment of MMAD systems increasingly relies on protocol-level testing:

- **Zero-shot and open-set assessment:** With vision-language models, protocols increasingly evaluate *zero-shot generalization* [14], [131], requiring models to detect unseen anomaly categories from text prompts and directly probing generalization under knowledge deficits (CH2).
- **Fairness in anomaly injection:** Following recent work [87], [88], pseudo-anomaly synthesis protocols

must prevent injected anomalies from leaking semantic labels or other shortcuts into training. Quantifying the synthetic-to-real distribution gap is thus a key fairness criterion.

- **Modality-missing robustness:** Since real deployments rarely provide complete data streams, robustness protocols now measure performance degradation when a modality is missing at inference, testing the resilience of cross-modal alignment mechanisms [18], [50] under severe modal discrepancy (CH1).

C. Computational Complexity and Scalability Analysis

Foundation models and generative approaches have greatly advanced MMAD, but large-scale deployment is often limited by compute and memory. In Big Data settings—with high-resolution inputs, long time spans, and continuous multi-modal streams—scalability becomes a primary concern, not a secondary one after accuracy.

From a system perspective, two dominant sources of computational overhead are commonly observed:

- **FM-based cross-modal alignment:** Many recent MMAD methods use high-capacity Transformers to align heterogeneous modalities [49], [51], [131]. Standard self-attention can incur up to $\mathcal{O}(N^2)$ computation and memory in sequence length N , so jointly processing high-resolution visual tokens with synchronized text or 3D representations can cause high GPU memory use and latency, limiting throughput in large-scale deployments.
- **Generative and diffusion-based paradigms:** Generative methods, especially diffusion-based anomaly synthesis [87], require iterative denoising over T steps, yielding overall complexity approximately proportional to $\mathcal{O}(T \cdot C_{\text{step}})$, where T denotes the denoising steps and C_{step} is the cost of one network evaluation. Although they produce diverse pseudo-anomalies, these multi-step procedures can incur substantial overhead, particularly for online generation or adaptation.

Beyond these backbone costs, cross-modal interaction adds burden because aligning heterogeneous representations often needs repeated fusion or joint encoding, with overhead especially high in streaming and high-throughput settings. Computational limits are not just system concerns but can directly reduce detection quality: efficiency demands often force lower spatial resolution or temporal sampling, which, though scalable, can reduce sensitivity to fine-grained or brief anomalies and worsen the granularity imbalance discussed in CH3.

To better support real-world deployment, scalable MMAD systems should explicitly navigate the accuracy–efficiency trade-off frontier. Potential directions include *model compression* (e.g., knowledge distillation [69]) and *parameter-efficient adaptation* [51], [131], which aim to reduce overhead while preserving cross-modal alignment capability.

VI. OPEN PROBLEMS AND FUTURE DIRECTIONS

MMAD is increasingly shaped by multi-modal foundation models and hybrid pipelines that couple normality modeling

with anomaly-oriented perturbations. Yet core challenges persist, including cross-modal misalignment (CH1), ambiguous boundaries (CH2), imbalanced sensitivity (CH3), a synthetic-to-real gap (CH4), and limited adaptability under drift (CH5). We condense future directions into one unifying paradigm and three core research questions (RQs).

Unifying paradigm: Foundation-model-driven assumption fusion. Rather than separating normality modeling and abnormality injection, future MMAD frameworks should leverage foundation models as building blocks to couple both: use their priors to align normal representations, and generate targeted semantic edits/prompts as hard negatives to tighten the boundary. In parallel, rigorous theory for such pipelines (e.g., information-theoretic bounds for fusion and generalization under drift) remains limited and is an important open frontier.

RQ1: How to generate reliable anomaly hypotheses? Pseudo anomalies can tighten boundaries, but their realism is rarely measurable; with scarce real anomalies, metrics such as FID may overfit. Reference-free protocols (auxiliary discriminators, alignment checks, and physics/temporal constraints) are needed to validate pseudo anomalies across modalities.

RQ2: How to perform semantic reasoning under cross-modal conflicts? To mitigate modality dominance and conflicting judgments, MMAD needs uncertainty-aware fusion such as evidential fusion or dynamic blending, plus semantic grounding. Foundation models, particularly MLLMs, can map anomaly evidence to human-readable rationales by linking continuous representations to violated cross-modal constraints.

RQ3: How to continually adapt under evolving environments? Real-world MMAD systems must continually adapt to evolving distributions while maintaining stable decision boundaries under constrained resources. Promising directions include PEFT adapters with replay/prototype retention, linear-time sequence models as alternatives to $\mathcal{O}(N^2)$ attention, and token pruning/distillation for real-time operation.

VII. CONCLUSIONS

This survey provided a structured, assumption-driven review of multi-modal anomaly detection, organizing prior work around two complementary paradigms: normality modeling and anomaly injection. By focusing on underlying assumptions rather than specific architectures or modalities, we offered a unified lens that clarifies both the strengths and limitations of existing approaches. We highlighted that MMAD complexity stems from the interplay between intrinsic anomaly detection challenges and multimodality-specific difficulties, particularly heterogeneity. Within this framework, we examined how current methods tackle these issues across multiple levels, from representation learning and structural perturbations to semantic injection and knowledge-guided modeling. Overall, MMAD is shifting from isolated detection pipelines toward holistic perception systems that combine representation learning, anomaly simulation, semantic reasoning, and deployment-aware adaptation. Foundation models are emerging as a key catalyst, reducing modal heterogeneity and enabling knowledge-rich anomaly definitions. By synthesizing recent progress and articulating actionable open questions, we aim for this survey to serve as

both a conceptual reference and a roadmap toward adaptive, interpretable, and reliable real-world MMAD systems.

REFERENCES

- [1] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM computing surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.
- [2] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5635–5662, 2024.
- [3] R. Wang, X. Mou, T. Wo, M. Zhang, Y. Liu, T. Wang, P. Liu, J. Yan, and X. Liu, "Acbot: an iiot platform for industrial robots," *Frontiers of Computer Science*, vol. 19, no. 4, p. 194203, 2025.
- [4] H. De Meulemeester, F. De Smet, J. van Dorst, E. Derroitte, and B. De Moor, "Explainable unsupervised anomaly detection for health-care insurance data," *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, p. 14, 2025.
- [5] Y. Lin, Y. Chang, X. Tong, J. Yu, A. Liotta, G. Huang, W. Song, D. Zeng, Z. Wu, Y. Wang *et al.*, "A survey on rgb, 3d, and multimodal approaches for unsupervised industrial image anomaly detection," *Information Fusion*, p. 103139, 2025.
- [6] A. Costanzino, P. Z. Ramirez, G. Lisanti, and L. Di Stefano, "Multi-modal industrial anomaly detection by crossmodal feature mapping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 234–17 243.
- [7] P. Kumari, A. K. Bedi, and M. Saini, "Multimedia datasets for anomaly detection: a review," *Multimedia Tools and Applications*, vol. 83, no. 19, pp. 56 785–56 835, 2024.
- [8] Y. Jiang, K. Ning, Z. Pan, X. Shen, J. Ni, W. Yu, A. Schneider, H. Chen, Y. Nevmyvaka, and D. Song, "Multi-modal time series analysis: A tutorial and survey," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025, pp. 6043–6053.
- [9] K. Xiao, Z. Qian, and B. Qin, "A survey of data representation for multi-modality event detection and evolution," *Applied Sciences*, vol. 12, no. 4, p. 2204, 2022.
- [10] Y. Wang, "Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1s, pp. 1–25, 2021.
- [11] J. Ren, T. Tang, H. Jia, Z. Xu, H. Fayek, X. Li, S. Ma, X. Xu, and F. Xia, "Foundation models for anomaly detection: Vision and challenges," *arXiv preprint arXiv:2502.06911*, 2025.
- [12] M. Ben Ammar, A. Mendoza, N. Belkhir, G. Franchi, and A. Manzanera, "Foundation models and transformers for anomaly detection: A survey," *Available at SSRN 5251498*.
- [13] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection," in *The Twelfth International Conference on Learning Representations*, 2024.
- [14] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "Winclip: Zero-/few-shot anomaly classification and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 606–19 616.
- [15] Y. Zhao, Y. Pang, L. Zhang, H. Liu, J. Zuo, H. Lu, and X. Zhao, "Unimad: Unified multi-modal and multi-class anomaly detection via moe-driven feature decomposition," *arXiv preprint arXiv:2509.25934*, 2025.
- [16] J. Zhang, G. Wang, Y. Jin, and D. Huang, "Towards training-free anomaly detection with vision and language foundation models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 15 204–15 213.
- [17] Y. Jiang, W. Luo, H. Zhang, Q. Chen, H. Yao, W. Shen, and Y. Cao, "Anomagic: Crossmodal prompt-driven zero-shot anomaly generation," *arXiv preprint arXiv:2511.10020*, 2025.
- [18] Y. Wang, J. Peng, J. Zhang, R. Yi, Y. Wang, and C. Wang, "Multimodal industrial anomaly detection via hybrid fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8032–8041.
- [19] H. Peng, X. Huang, S. Sun, R. Zhang, and X. Wang, "Adaptive and robust dbSCAN with multi-agent reinforcement learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [20] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.

- [21] Y. Ji, Y. Chen, L. Yang, R. Ding, M. Yang, and X. Zheng, "Vexkd: The versatile integration of cross-modal fusion and knowledge distillation for 3d perception," *Advances in Neural Information Processing Systems*, vol. 37, pp. 125 608–125 634, 2024.
- [22] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 387–395.
- [23] S. Tuli, G. Casale, and N. R. Jennings, "Tranad: Deep transformer networks for anomaly detection in multivariate time series data," *arXiv preprint arXiv:2201.07284*, 2022.
- [24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [25] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical image analysis*, vol. 54, pp. 30–44, 2019.
- [26] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian conference on computer vision*. Springer, 2018, pp. 622–637.
- [27] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—a new baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6536–6545.
- [28] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "Anopcn: Video anomaly detection via deep predictive coding network," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1805–1813.
- [29] C. Feng, Z. Chen, and A. Owens, "Self-supervised video forensics by audio-visual anomaly detection," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 491–10 503.
- [30] C. Willibald, D. Sliwowski, and D. Lee, "Multimodal anomaly detection with a mixture-of-experts," *arXiv preprint arXiv:2506.19077*, 2025.
- [31] B. Leporowski, A. Bakhtiarnia, N. Bonnici, A. Muscat, L. Zarella, Y. Wang, and A. Iosifidis, "Mavad: Audio-visual dataset and method for anomaly detection in traffic videos," in *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024, pp. 1106–1112.
- [32] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," *Advances in neural information processing systems*, vol. 12, 1999.
- [33] D. M. Tax and R. P. Duin, "Support vector domain description," *Pattern recognition letters*, vol. 20, no. 11–13, pp. 1191–1199, 1999.
- [34] L. Ruff, R. A. Vandermeulen, N. Göritz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft, "Deep semi-supervised anomaly detection," *arXiv preprint arXiv:1906.02694*, 2019.
- [35] R. Wang, C. Liu, X. Mou, K. Gao, X. Guo, P. Liu, T. Wo, and X. Liu, "Deep contrastive one-class time series anomaly detection," in *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 2023, pp. 694–702.
- [36] X. Mou, R. Wang, T. Wang, J. Sun, B. Li, T. Wo, and X. Liu, "Deep autoencoding one-class time series anomaly detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [37] X. Mou, R. Wang, B. Li, T. Wo, J. Sun, H. Wang, and X. Liu, "Roca: Robust contrastive one-class time series anomaly detection with contaminated data," *arXiv preprint arXiv:2503.18385*, 2025.
- [38] M. Ma, T. Wang, R. Wang, X. Mou, T. Wo, and X. Liu, "Foca: Foundation-model-based one-class anomaly detection for time series," 2025.
- [39] J. Gao, W. Fan, D. Turaga, S. Parthasarathy, and J. Han, "A spectral framework for detecting inconsistency across multi-source object relationships," in *2011 IEEE 11th International Conference on Data Mining*. IEEE, 2011, pp. 1050–1055.
- [40] M. Pérez-Carrasco, G. Cabrera-Vives, L. Hernández-García, F. Forster, P. Sánchez-Sáez, A. M. Arancibia, N. Astorga, F. Bauer, A. Bayo, M. Cádiz-Leyton *et al.*, "Multi-class deep svdd: Anomaly detection approach in astronomy with distinct inlier categories," *arXiv preprint arXiv:2308.05011*, 2023.
- [41] C. Wang, X. Ning, P. Qian, W. Hu, J. Yao, E.-Y.-K. Ng, K.-W. Lai, and S. Wang, "Distribution entropy regularized multimodal subspace support vector data description for anomaly detection," *Pattern Recognition*, p. 112478, 2025.
- [42] V. Blanco, I. Espejo, R. Páez, and A. M. Rodríguez-Chía, "A mathematical optimization approach to multisphere support vector data description," *arXiv preprint arXiv:2507.11106*, 2025.
- [43] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM computing surveys*, vol. 56, no. 9, pp. 1–36, 2024.
- [44] C. Tao, X. Cao, and J. Du, "G2sf: Geometry-guided score fusion for multimodal industrial anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 20 551–20 560.
- [45] W. Li, B. Zheng, X. Xu, J. Gan, F. Lu, X. Li, N. Ni, Z. Tian, X. Huang, S. Gao *et al.*, "Multi-sensor object anomaly detection: Unifying appearance, geometry, and internal properties," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9984–9993.
- [46] A. Ghadiya, P. Kar, V. Chudasama, and P. Wasnik, "Cross-modal fusion and attention mechanism for weakly supervised video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1965–1974.
- [47] H. Cheng, J. Luo, and X. Zhang, "Multimodal industrial anomaly detection via uni-modal and cross-modal fusion," *IEEE Transactions on Industrial Informatics*, 2025.
- [48] U. Ali, A. Zia, A. Rehman, U. Ramzan, Z. Hassan, T. Sattar, J. Wang, and W. Xiang, "2d_3d feature fusion via cross-modal latent synthesis and attention guided restoration for industrial anomaly detection," *arXiv preprint arXiv:2510.21793*, 2025.
- [49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.
- [50] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang, "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 6074–6082.
- [51] Y. Cao, J. Zhang, L. Frittoli, Y. Cheng, W. Shen, and G. Boracchi, "Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 55–72.
- [52] W. Ma, X. Zhang, Q. Yao, F. Tang, C. Wu, Y. Li, R. Yan, Z. Jiang, and S. K. Zhou, "Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4744–4754.
- [53] W. Yin, H. Zhang, X. Wang, Y. Lu, Y. Zhang, B. Gong, J. Zuo, L. Yu, C. Gao, and N. Sang, "Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 14, 2026, pp. 12 027–12 035.
- [54] G. Zeng, H. Peng, A. Li, J. Wu, C. Liu, and P. S. Yu, "Scalable semi-supervised clustering via structural entropy with different constraints," *IEEE Transactions on Knowledge and Data Engineering*, 2024.
- [55] X. Zeng, H. Peng, D. Su, and A. Li, "Hierarchical decision making based on structural information principles," *Journal of Machine Learning Research*, vol. 26, no. 182, pp. 1–55, 2025.
- [56] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "Filo++: Zero-few-shot anomaly detection by fused fine-grained descriptions and deformable localization," *arXiv preprint arXiv:2501.10067*, 2025.
- [57] T. Zhao, L. Li, T. Tian, J. Ma, and J. Tian, "Patch-guided point matching for point cloud registration with low overlap," *Pattern Recognition*, vol. 144, p. 109876, 2023.
- [58] Y. Cao, X. Xu, and W. Shen, "Complementary pseudo multimodal feature for point cloud anomaly detection," *Pattern Recognition*, vol. 156, p. 110761, 2024.
- [59] P. Bergmann and D. Sattlegger, "Anomaly detection in 3d point clouds using deep geometric descriptors," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2613–2623.
- [60] V. Zavrtanik, M. Kristan, and D. Skočaj, "Cheating depth: Enhancing 3d surface anomaly detection via depth simulation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 2164–2172.
- [61] C. Wang, H. Zhu, J. Peng, Y. Wang, R. Yi, Y. Wu, L. Ma, and J. Zhang, "M3dm-nr: Rgb-3d noisy-resistant industrial anomaly detection via multimodal denoising," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [62] A. Patel, P.-D. Tudosi, W. H. L. Pinaya, G. Cook, V. Goh, S. Ourselin, and M. J. Cardoso, "Cross attention transformers for multi-modal unsupervised whole-body pet anomaly detection," in *MICCAI Workshop on Deep Generative Models*. Springer, 2022, pp. 14–23.

- [63] Y. Xu and H. Chen, "Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 21 241–21 251.
- [64] J. Gu, Y. Wang, J. Chen, M. Zhang, Z. Wang, and J. Chen, "Multi-modal contrastive causal consistency fusion for anomaly detection in additive manufacturing," *Additive Manufacturing*, p. 104816, 2025.
- [65] Y. Su, Y. Tan, S. An, M. Xing, and Z. Feng, "Semantic-driven dual consistency learning for weakly supervised video anomaly detection," *Pattern Recognition*, vol. 157, p. 110898, 2025.
- [66] J. Wu, "Anomaly detection in medical via multimodal foundation models," *Frontiers in Bioengineering and Biotechnology*, vol. 13, p. 1644697, 2025.
- [67] W. Fan and Z. Shi, "Cross-modal consistency with aesthetic similarity for multimodal false information detection," *Computers, Materials & Continua*, vol. 79, no. 2, 2024.
- [68] H. Wang, C. Ma, J. Zhang, Y. Zhang, J. Avery, L. Hull, and G. Carneiro, "Learnable cross-modal knowledge distillation for multi-modal learning with missing modality," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 216–226.
- [69] Z. Gu, J. Zhang, L. Liu, X. Chen, J. Peng, Z. Gan, G. Jiang, A. Shu, Y. Wang, and L. Ma, "Rethinking reverse distillation for multi-modal anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 8445–8453.
- [70] F. Huo, W. Xu, J. Guo, H. Wang, and S. Guo, "C2kd: Bridging the modality gap for cross-modal knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 006–16 015.
- [71] M. Li, D. Yang, X. Zhao, S. Wang, Y. Wang, K. Yang, M. Sun, D. Kou, Z. Qian, and L. Zhang, "Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 458–12 468.
- [72] A. M. Mansourian, A. M. Babaei, and S. Kasaei, "Enriching knowledge distillation with cross-modal teacher fusion," *arXiv preprint arXiv:2511.09286*, 2025.
- [73] X. Liu, J. Wang, B. Leng, and S. Zhang, "Multimodal industrial anomaly detection by crossmodal reverse distillation," *arXiv preprint arXiv:2412.08949*, 2024.
- [74] Z. Chen, B. Chen, J. Yang, Y. Chen, and Y. Zhang, "Enhancing multimodal anomaly detection via asymmetric dual-branch reverse distillation: Z. chen, b. chen, j. yang, y. chen, y. zhang," *The Visual Computer*, pp. 1–17, 2025.
- [75] W. Sui, D. Lichau, J. Lefèvre, and H. Phelippeau, "Incomplete multimodal industrial anomaly detection via cross-modal distillation," *Information Fusion*, p. 103572, 2025.
- [76] J. Liu, Y. Zhang, T. Huang, W. Xu, and R. Yang, "Distilling cross-modal knowledge via feature disentanglement," *arXiv preprint arXiv:2511.19887*, 2025.
- [77] N. Wang and Q. Wang, "Modality-uncertainty-aware knowledge distillation framework for multimodal sentiment analysis," *Complex & Intelligent Systems*, vol. 12, no. 1, p. 14, 2026.
- [78] Y. Lin, H. Yan, X. Tong, Y. Chang, H. Wang, Z. Zhou, S. Gao, Y. Wang, and W. Zhang, "Commonality in few: Few-shot multimodal anomaly detection via hypergraph-enhanced memory," *arXiv preprint arXiv:2511.05966*, 2025.
- [79] W. Shangguan, H. Wu, Y. Niu, H. Yin, J. Yu, B. Chen, and B. Huang, "Cpir: Multimodal industrial anomaly detection via latent bridged cross-modal prediction and intra-modal reconstruction," *Advanced Engineering Informatics*, vol. 65, p. 103240, 2025.
- [80] Y. Zhou, J. Zhao, D. Zeng, Z. Zuo, W. Liu, and Z. Wu, "Multimodal task representation memory bank vs. catastrophic forgetting in anomaly detection," *arXiv preprint arXiv:2502.06194*, 2025.
- [81] J. Zhang, H. Peng, M. Yang, and P. S. Yu, "Enhanced pre-training for recommendation via hypergraph structural entropy," *ACM Transactions on Information Systems*, 2025.
- [82] Q. Liu, H. Peng, X. Huang, Z. Hao, Q. Sun, Z. Yu, and P. S. Yu, "Hierarchical text classification optimization via structural entropy and singular smoothing," *IEEE Transactions on Knowledge and Data Engineering*, 2025.
- [83] S. Li, H. Gong, H. Dong, T. Yang, Z. Tu, and Y. Zhao, "Dpu: Dynamic prototype updating for multimodal out-of-distribution detection," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 10 193–10 202.
- [84] S. Li, L. Zhi, W. Wang, L. Zheng, and Y. Lu, "Dnpr: Zero-shot industrial anomaly detection via dynamic normal prototype refinement," *Available at SSRN 5786608*.
- [85] L. Tian, Y. Li, Y. Dai, W. Chen, X. Liu, and B. Chen, "Fastref: Fast prototype refinement for few-shot industrial anomaly detection," *arXiv preprint arXiv:2506.21398*, 2025.
- [86] H. Hu, W. Du, P. Liao, B. Wang, and S. Fan, "Noise-resistant video anomaly detection via rgb error-guided multiscale predictive coding and dynamic memory," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 109–19 119.
- [87] T. Hu, J. Zhang, R. Yi, Y. Du, X. Chen, L. Liu, Y. Wang, and C. Wang, "Anomalydiffusion: Few-shot anomaly image generation with diffusion model," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 8, 2024, pp. 8526–8534.
- [88] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9664–9674.
- [89] M. Astrid, M. Z. Zaheer, and S.-I. Lee, "Synthetic temporal anomaly guided end-to-end video anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 207–214.
- [90] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *International Conference on Learning Representations*, 2018.
- [91] C. Qiu, A. Li, M. Kloft, M. Rudolph, and S. Mandt, "Latent outlier exposure for anomaly detection with contaminated data," in *International conference on machine learning*. PMLR, 2022, pp. 18 153–18 167.
- [92] J. Feng, Z. Huang, J. Guo, and W. Qiu, "Unsupervised anomaly detection for time series with outlier exposure," in *Proceedings of the 33rd International Conference on Scientific and Statistical Database Management*, 2021, pp. 1–12.
- [93] D. Mantegazza, A. Giusti, L. M. Gambardella, and J. Guzzi, "An outlier exposure approach to improve visual anomaly detection performance for mobile robots," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 354–11 361, 2022.
- [94] Y. Zhang, H. Suo, Y. Wan, and M. Li, "Outlier-aware inlier modeling and multi-scale scoring for anomalous sound detection via multitask learning," *arXiv preprint arXiv:2309.07500*, 2023.
- [95] H. Junwei, Q. Xu, Y. Jiang, Z. Wang, Y. Sun, and Q. Huang, "Hgoe: Hybrid external and internal graph outlier exposure for graph out-of-distribution detection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1544–1553.
- [96] M. Liu, H. Dong, J. Kelly, O. Fink, and M. Trapp, "Extremely simple multimodal outlier synthesis for out-of-distribution detection and segmentation," *arXiv preprint arXiv:2505.16985*, 2025.
- [97] K. Fukuda and H. Aizawa, "Taylor outlier exposure," *arXiv preprint arXiv:2412.07219*, 2024.
- [98] Y. Shi, R. Xu, and Z. Qi, "Outlier exposure in visual industrial anomaly detection-survey," in *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2024, pp. 17–24.
- [99] K. Kumar Singh and Y. Jae Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3524–3533.
- [100] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.
- [101] P. Chen, S. Liu, H. Zhao, X. Wang, and J. Jia, "Gridmask data augmentation," *arXiv preprint arXiv:2001.04086*, 2020.
- [102] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.
- [103] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.
- [104] P. Guo, H. Yang, and A. Sano, "Empirical study of mix-based data augmentation methods in physiological time series data," in *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2023, pp. 206–213.
- [105] C. U. Carmona, F.-X. Aubet, V. Flunkert, and J. Gasthaus, "Neural contextual anomaly detection for time series," *IJCAI*, 2022.
- [106] D. Reshetova, S. Ganguli, C. K. Iyer, and V. Pandey, "Semand: Self-supervised anomaly detection in multimodal geospatial datasets," in *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, 2023, pp. 1–4.
- [107] S. Khan, M. Yüksel, and F. Kirchner, "Robust anomaly detection through multi-modal autoencoder fusion for small vehicle damage detection," *Machine Learning with Applications*, p. 100794, 2025.

- [108] A. Syed and M. I. Ahmad, "Advanced data collection techniques in cloud security: A multi-modal deep learning autoencoder approach," *arXiv preprint arXiv:2511.21795*, 2025.
- [109] S. Benabderrahmane and T. Rahwan, "Adversarial augmentation and active sampling for robust cyber anomaly detection," *arXiv preprint arXiv:2509.04999*, 2025.
- [110] Y. Jeong, E. Yang, J. H. Ryu, I. Park, and M. Kang, "Anomalybert: Self-supervised transformer for time series anomaly detection using data degradation scheme," *arXiv preprint arXiv:2305.04468*, 2023.
- [111] K. Obata, Y. Matsubara, and Y. Sakurai, "Robust and explainable detector of time series anomaly via augmenting multiclass pseudo-anomalies," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025, pp. 2198–2209.
- [112] R. Wang, X. Mou, R. Yang, K. Gao, P. Liu, C. Liu, T. Wo, and X. Liu, "Cutadpaste: Time series anomaly detection by exploiting abnormal knowledge," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 3176–3187.
- [113] A. K. Rai, T. Krishna, F. Hu, A. Drimbarean, K. McGuinness, A. F. Smeaton, and N. E. O'Connor, "Video anomaly detection via spatio-temporal pseudo-anomaly generation: A unified approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3887–3899.
- [114] L. Zhu, L. Wang, A. Raj, T. Gedeon, and C. Chen, "Advancing video anomaly detection: A concise review and a new dataset," *Advances in Neural Information Processing Systems*, vol. 37, pp. 89 943–89 977, 2024.
- [115] Y. Wang, Y. Zhao, Y. Huo, and Y. Lu, "Multimodal anomaly detection in complex environments using video and audio fusion," *Scientific Reports*, vol. 15, no. 1, p. 16291, 2025.
- [116] D. Wang, Q. Wang, Q. Hu, and K. Wu, "Multimodal vad: Visual anomaly detection in intelligent monitoring system via audio-vision-language," *IEEE Transactions on Instrumentation and Measurement*, 2025.
- [117] S. Tariq, A.-U. Rehman, M. Abubakar, W. Iqbal, H. S. Alsagri, Y. A. Alduraywish, and H. A. A. Alhakkani, "A recurrent neural network for multimodal anomaly detection by using spatio-temporal audio-visual data," *Computers, Materials & Continua*, vol. 81, no. 2, 2024.
- [118] A. Dionis-Ros, J. Vila-Francés, R. Magdalena-Benedito, F. Mateo, and A. J. Serrano-López, "Multimodal video analysis for crowd anomaly detection using open access tourism cameras," *Applied Sciences*, vol. 14, no. 23, p. 11075, 2024.
- [119] F. Liu, X. Ma, J. Wu, J. Yang, S. Xue, A. Beheshti, C. Zhou, H. Peng, Q. Z. Sheng, and C. C. Aggarwal, "Dagad: Data augmentation for graph anomaly detection," in *2022 IEEE international conference on data mining (ICDM)*. IEEE, 2022, pp. 259–268.
- [120] S. Zhou, X. Huang, N. Liu, H. Zhou, F.-L. Chung, and L.-K. Huang, "Improving generalizability of graph anomaly detection models via data augmentation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12 721–12 735, 2023.
- [121] J. Duan, S. Wang, P. Zhang, E. Zhu, J. Hu, H. Jin, Y. Liu, and Z. Dong, "Graph anomaly detection via multi-scale contrastive learning networks with augmented view," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 6, 2023, pp. 7459–7467.
- [122] H. Tang, X. Liang, J. Wang, and S. Zhang, "Dualgad: Dual-bootstrapped self-supervised learning for graph anomaly detection," *Information Sciences*, vol. 668, p. 120520, 2024.
- [123] Z. Chen, Z. Wu, Y. Sadikaj, C. Plant, H.-N. Dai, S. Wang, Y.-M. Cheung, and W. Guo, "Adedgedrop: Adversarial edge dropping for robust graph neural networks," *IEEE Transactions on Knowledge and Data Engineering*, 2025.
- [124] X. Li, J. Qi, Z. Zhao, G. Zheng, L. Cao, J. Dong, and Y. Yu, "Umgad: Unsupervised multiplex graph anomaly detection," in *2025 IEEE 41st International Conference on Data Engineering (ICDE)*. IEEE, 2025, pp. 3724–3737.
- [125] Y. Xu, Z. Peng, B. Shi, X. Hua, B. Dong, S. Wang, and C. Chen, "Revisiting graph contrastive learning on anomaly detection: A structural imbalance perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 12, 2025, pp. 12 972–12 980.
- [126] C. Song, X. Lin, H. Shen, Y. Shang, and Y. Cao, "Uniform: Towards unified framework for anomaly detection on graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 12, 2025, pp. 12 559–12 567.
- [127] Z. Zhang, M. Cai, G. Wu, J. Zhang, L. Liu, D. Tao, T. Chai, and X. Zhu, "Unified unsupervised anomaly detection via matching cost filtering," *arXiv preprint arXiv:2510.03363*, 2025.
- [128] H. Qiao, Q. Wen, X. Li, E.-P. Lim, and G. Pang, "Generative semi-supervised graph anomaly detection," *Advances in neural information processing systems*, vol. 37, pp. 4660–4688, 2024.
- [129] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection," *arXiv preprint arXiv:2310.18961*, 2023.
- [130] C. Deng, H. Xu, X. Chen, H. Xu, X. Tu, X. Ding, and Y. Huang, "Simclip: Refining image-text alignment with simple prompts for zero-/few-shot anomaly detection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1761–1770.
- [131] X. Li, Z. Zhang, X. Tan, C. Chen, Y. Qu, Y. Xie, and L. Ma, "Promptad: Learning prompts with only normal samples for few-shot anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 838–16 848.
- [132] Z. Qu, X. Tao, M. Prasad, F. Shen, Z. Zhang, X. Gong, and G. Ding, "Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 301–317.
- [133] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "Anomalygpt: Detecting industrial anomalies using large vision-language models," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 3, 2024, pp. 1932–1940.
- [134] X. Xiang, S. Luo, Y. Du, L. Zhang, and X. Zhen, "Uniad: Unified cross-modal prompt regularization for zero-shot anomaly detection across domains," *Neurocomputing*, p. 132372, 2025.
- [135] J. Nie, Y. Zhang, Z. Fang, T. Liu, B. Han, and X. Tian, "Out-of-distribution detection with negative prompts," in *The twelfth international conference on learning representations*, 2024.
- [136] J. Xu, S.-Y. Lo, B. Safaei, V. M. Patel, and I. Dwivedi, "Towards zero-shot anomaly detection and reasoning with multimodal large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 370–20 382.
- [137] H. Deng, H. Luo, W. Zhai, Y. Guo, Y. Cao, and Y. Kang, "Vmad: Visual-enhanced multimodal large language model for zero-shot anomaly detection," *IEEE Transactions on Automation Science and Engineering*, 2025.
- [138] H. Zhang, X. Xu, X. Wang, J. Zuo, C. Han, X. Huang, C. Gao, Y. Wang, and N. Sang, "Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm," *arXiv preprint arXiv:2406.12235*, 2024.
- [139] Y. Dang, K. Huang, J. Huo, Y. Yan, S. Huang, D. Liu, M. Gao, J. Zhang, C. Qian, K. Wang *et al.*, "Explainable and interpretable multimodal large language models: A comprehensive survey," *arXiv preprint arXiv:2412.02104*, 2024.
- [140] S. K. Lim, Y. Loo, N.-T. Tran, N.-M. Cheung, G. Roig, and Y. Elovici, "Doping: Generative data augmentation for unsupervised anomaly detection with gan," in *2018 IEEE international conference on data mining (ICDM)*. IEEE, 2018, pp. 1122–1127.
- [141] M. Z. Zaheer, J.-h. Lee, M. Astrid, and S.-I. Lee, "Old is gold: Redefining the adversarially learned one-class classifier training paradigm," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 183–14 193.
- [142] M. Pourreza, B. Mohammadi, M. Khaki, S. Bouindour, H. Snoussi, and M. Sabokrou, "G2d: Generate to detect anomaly," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2003–2012.
- [143] G. Zhang, K. Cui, T.-Y. Hung, and S. Lu, "Defect-gan: High-fidelity defect synthesis for automated defect inspection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2524–2534.
- [144] S. Niu, B. Li, X. Wang, and H. Lin, "Defect image sample generation with gan for improving defect recognition," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 3, pp. 1611–1622, 2020.
- [145] Y. Duan, Y. Hong, L. Niu, and L. Zhang, "Few-shot defect image generation via defect-aware feature manipulation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 1, 2023, pp. 571–578.
- [146] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [147] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, "Anoddpmm: Anomaly detection with denoising diffusion probabilistic models using simplex noise," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 650–656.
- [148] X. Zhang, M. Xu, and X. Zhou, "Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection," in *Proceedings*

- of the *IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 699–16 708.
- [149] H. He, J. Zhang, H. Chen, X. Chen, Z. Li, X. Chen, Y. Wang, C. Wang, and L. Xie, “A diffusion-based framework for multi-class anomaly detection,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 8, 2024, pp. 8472–8480.
- [150] B. Jiang, Y. Xie, J. Li, N. Li, Y. Jiang, and S.-T. Xia, “Cagen: Controllable anomaly generator using diffusion model,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 3110–3114.
- [151] S. Dai, Y. Wu, X. Li, and X. Xue, “Generating and reweighting dense contrastive patterns for unsupervised anomaly detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 1454–1462.
- [152] J. Hu, Y. Huang, Y. Lu, G. Xie, G. Jiang, Y. Zheng, and Z. Lu, “Anomalyxfusion: Multi-modal anomaly synthesis with diffusion,” *arXiv preprint arXiv:2404.19444*, 2024.
- [153] A. Flaborea, L. Collorone, G. M. D. Di Melendugno, S. D’Arrigo, B. Prenkaj, and F. Galasso, “Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 10 318–10 329.
- [154] M. A. Loodaricheh, M. H. Manshaei, and A. Raja, “Mage-id: A multimodal generative framework for intrusion detection systems,” *arXiv preprint arXiv:2512.03375*, 2025.
- [155] X. Zeng, H. Peng, and A. Li, “Proactive bot detection based on structural information principles,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [156] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, “Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9592–9600.
- [157] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, “Spot-the-difference self-supervised pre-training for anomaly detection and segmentation,” in *European conference on computer vision*. Springer, 2022, pp. 392–408.
- [158] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, “Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization,” *International Journal of Computer Vision*, vol. 130, no. 4, pp. 947–969, 2022.
- [159] P. Bergmann, X. Jin, D. Sattlegger, and C. Steger, “The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization,” *arXiv preprint arXiv:2112.09045*, 2021.
- [160] J. Liu, G. Xie, R. Chen, X. Li, J. Wang, Y. Liu, C. Wang, and F. Zheng, “Real3d-ad: A dataset of point cloud anomaly detection,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 30 402–30 415, 2023.
- [161] L. Bonfiglioli, M. Toschi, D. Silvestri, N. Fioraio, and D. De Gregorio, “The eyecandies dataset for unsupervised multimodal anomaly detection and localization,” in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 3586–3602.
- [162] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, “Vt-adl: A vision transformer network for image anomaly detection and localization,” in *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*. IEEE, 2021, pp. 01–06.
- [163] S. Jezek, M. Jonak, R. Burget, P. Dvorak, and M. Skotak, “Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions,” in *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*. IEEE, 2021, pp. 66–71.
- [164] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, “Segmentation-based deep-learning approach for surface-defect detection,” *Journal of Intelligent Manufacturing*, vol. 31, no. 3, pp. 759–776, 2020.
- [165] M. Wieler and T. Hahn, “Weakly supervised learning for industrial optical inspection,” in *DAGM symposium in*, vol. 6, 2007, p. 11.
- [166] A. Oliner and J. Stearley, “What supercomputers say: A study of five system logs,” in *37th annual IEEE/IFIP international conference on dependable systems and networks (DSN’07)*. IEEE, 2007, pp. 575–584.
- [167] W. Xu, L. Huang, A. Fox, D. Patterson, and M. I. Jordan, “Detecting large-scale system problems by mining console logs,” in *Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles*, 2009, pp. 117–132.
- [168] J. Goh, S. Adepu, K. N. Nunejo, and A. Mathur, “A dataset to support research in the design of secure water treatment systems,” in *International conference on critical information infrastructures security*. Springer, 2016, pp. 88–99.
- [169] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, “Wadi: a water distribution testbed for research in the design of secure cyber physical systems,” in *Proceedings of the 3rd international workshop on cyber-physical systems for smart water networks*, 2017, pp. 25–28.
- [170] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, “Robust anomaly detection for multivariate time series through stochastic recurrent neural network,” in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2828–2837.
- [171] J. Li, T. Chen, X. Wang, Y. Zhong, and X. Xiao, “Adapting the segment anything model for multi-modal retinal anomaly detection and localization,” *Information Fusion*, vol. 113, p. 102631, 2025.
- [172] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, J. Son, W. Bac, L. Liu, J. Wang, X. Liu, L. Gao *et al.*, “Idrid: Diabetic retinopathy-segmentation and grading challenge,” *Medical image analysis*, vol. 59, p. 101561, 2020.
- [173] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier, “Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study,” *The Lancet*, vol. 392, no. 10162, pp. 2388–2396, 2018.
- [174] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati *et al.*, “The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification,” *arXiv preprint arXiv:2107.02314*, 2021.
- [175] A. Hamada, “Br35h :: Brain tumor detection 2020,” 2025. [Online]. Available: <https://dx.doi.org/10.21227/tbkk-q937>
- [176] M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al Emadi *et al.*, “Can ai help in screening viral and covid-19 pneumonia?” *Ieee Access*, vol. 8, pp. 132 665–132 676, 2020.
- [177] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic),” in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.
- [178] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, “Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.
- [179] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen, “Kvasir-seg: A segmented polyp dataset,” in *International conference on multimedia modeling*. Springer, 2019, pp. 451–462.
- [180] S. A. Hicks, D. Jha, V. Thambawita, P. Halvorsen, H. L. Hammer, and M. A. Riegler, “The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy,” in *International Conference on Pattern Recognition*. Springer, 2021, pp. 263–274.
- [181] H. Gong, G. Chen, R. Wang, X. Xie, M. Mao, Y. Yu, F. Chen, and G. Li, “Multi-task learning for thyroid nodule segmentation with thyroid region prior,” in *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*. IEEE, 2021, pp. 257–261.
- [182] G. B. Moody and R. G. Mark, “The impact of the mit-bih arrhythmia database,” *IEEE engineering in medicine and biology magazine*, vol. 20, no. 3, pp. 45–50, 2001.
- [183] Y. Chen, Y. Hao, T. Rakthanmanon, J. Zakaria, B. Hu, and E. Keogh, “A general framework for never-ending learning from time series streams,” *Data mining and knowledge discovery*, vol. 29, no. 6, pp. 1622–1664, 2015.
- [184] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, “Mimic-iii, a freely accessible critical care database,” *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
- [185] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.
- [186] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, “Not only look, but also listen: Learning multimodal violence detection under weak supervision,” in *European conference on computer vision*. Springer, 2020, pp. 322–339.
- [187] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked rnn framework,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 341–349.

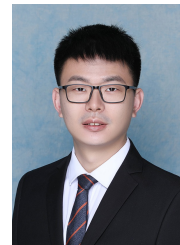
- [188] A. Ali, Z. A. Khan, A. Hussain, M. Munsif, A. Hussain, and S. W. Baik, "Avar-net: A lightweight audio-visual anomaly recognition framework with a benchmark dataset," *arXiv preprint arXiv:2510.13630*, 2025.
- [189] Y. Qian, S. Ye, C. Wang, X. Cai, J. Qian, and J. Wu, "Ucf-crime-dvs: A novel event-based dataset for video anomaly detection with spiking neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 6577–6585.
- [190] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [191] A. Huet, J. M. Navarro, and D. Rossi, "Local evaluation of time series anomaly detection algorithms," in *ACM SIGKDD*, 2022, pp. 635–645.
- [192] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. J. Elmore, and M. J. Franklin, "Volume under the surface: A new accuracy evaluation measure for time-series anomaly detection." *Proc. VLDB Endow.*, vol. 15, no. 11, pp. 2774–2787, 2022.



Xudong Mou received her M.S. degree in the School of Computer Science and Engineering, Beihang University, in 2021. She is working towards a Ph.D. at the School of Computer Science and Engineering, Beihang University, China. Her research interest is time series anomaly detection.



Zexin Wu received his B.S. degree in Electrical Engineering and Automation from China University of Geosciences, Beijing, in 2020. He is working towards an M.S. at the School of Software, Beihang University, China. His research interest is time series anomaly detection.



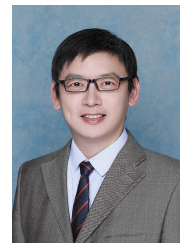
Chuan Luo received the Ph.D. degree in computer science from Peking University, Beijing, China, in 2016. He is currently an associate professor with School of Software, Beihang University, Beijing, China. His current research interests include constraint solving and heuristic search.



Shiru Chen received a PhD degree from Harbin Engineering University. He is currently a R&D Director of Shandong Inspur Intelligent Production Technology Co., Ltd. and has over 20 years of experience in the fields of communication, IoT, AI, distributed data processing, and industrial software systems. He has published 25+ academic papers, holds 60 invention patents, participated in 5 national standards writing, and authored 2 books.



Xudong Liu is a professor at the School of Computer Science and Engineering, Beihang University, Beijing, China. He has led several China 863 key projects and government projects. His research interests include software middleware technology, software development methods and tools, large-scale information technology projects, and the application of research and teaching.



Chunming Hu received the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2006. Currently, he is a Professor with the School of Software, Beihang University, Beijing, China. His research interests include distributed systems, system virtualization, data management, and processing systems.



Renyu Yang is an associate professor with the School of Software, Beihang University, China. Prior to this, he was with the University of Leeds UK, Alibaba Group China and Edgetic Ltd. UK, building large-scale computing/AI infrastructures. He is a recipient of Alan Turing Post-Doctoral Enrichment Award, 2022. His research interests include parallel and distributed computing, large-scale AI systems and software dependability. He is a member of IEEE.