Review

Not peer-reviewed version

Agentic and LLM-Based Multimodal Anomaly Detection: Architectures, Challenges, and Prospects

Mohammed Ayalew Belay *, Amirshayan Haghipour, Adil Rasheed, Pierluigi Salvo Rossi

Posted Date: 25 February 2026

doi: 10.20944/preprints202602.1368.v1

Keywords: agentic anomaly detection; agents; large language models; multimodal; cross-modal fusion; agentic AI

 doi:10.20944/preprints202602.1368.v1

Review

Agentic and LLM-Based Multimodal Anomaly Detection: Architectures, Challenges, and Prospects

Mohammed Ayalew Belay [1,*], Amirshayan Haghipour [1], Adil Rasheed [2] and Pierluigi Salvo Rossi [1]

[1] Department of Electronic Systems, Norwegian University of Science and Technology, 7034 Trondheim, Norway
[2] Department of Engineering Cybernetics, Norwegian University of Science and Technology, 7034 Trondheim, Norway
* Correspondence: ayalew@simula.no or mohammed.a.belay@ntnu.no

Abstract

Anomaly detection is crucial for maintaining the safety, reliability, and optimal performance of complex systems across diverse domains such as industrial manufacturing, cybersecurity, and autonomous systems. Conventional methods typically handle single data modalities, limiting their effectiveness in the multimodal and dynamic real-world environments. The integration of multimodal data sources, including visual, audio, and sensor data, has emerged as a key advancement, improving detection robustness and accuracy. Simultaneously, the rise of agentic artificial intelligence (AI), characterized by autonomous, goal-oriented agents capable of reasoning and utilizing tools, presents significant opportunities for enhancing anomaly detection systems. This paper provides a comprehensive review of recent advancements at the intersection of agentic AI and multimodal anomaly detection. We propose a novel taxonomy categorizing existing methods by agent architecture, reasoning capabilities, tool integration, and modality scope. We survey foundation model-based detectors, cross-modal fusion techniques, and LLM-driven agents that facilitate dynamic and interpretable anomaly reasoning. Furthermore, we present recent benchmark datasets, critical challenges, mitigations, and future research directions.

Keywords: agentic anomaly detection; agents; large language models; multimodal; cross-modal fusion; agentic AI

1. Introduction

Anomaly detection is increasingly becoming crucial for ensuring the reliability, security, and optimal performance of complex systems across a wide range of domains. From industrial manufacturing [1,2] and financial services [3] to healthcare [4], cybersecurity [5–9], and autonomous systems [10], the early identification of abnormal patterns plays a vital role in preventing costly failures, supporting timely decision-making, and mitigating risks. At its core, anomaly detection involves identifying data patterns that deviate from expected or normal behavior [11,12]. These anomalies often signify critical events or faults: for instance, unusual network traffic may suggest a cyberattack, aberrant sensor readings could indicate impending equipment failure, and outlier financial transactions might reveal fraudulent activity. Because anomalies typically represent rare but significant occurrences, anomaly detection has been extensively studied in the past few years [13–15].

 doi:10.20944/preprints202602.1368.v1

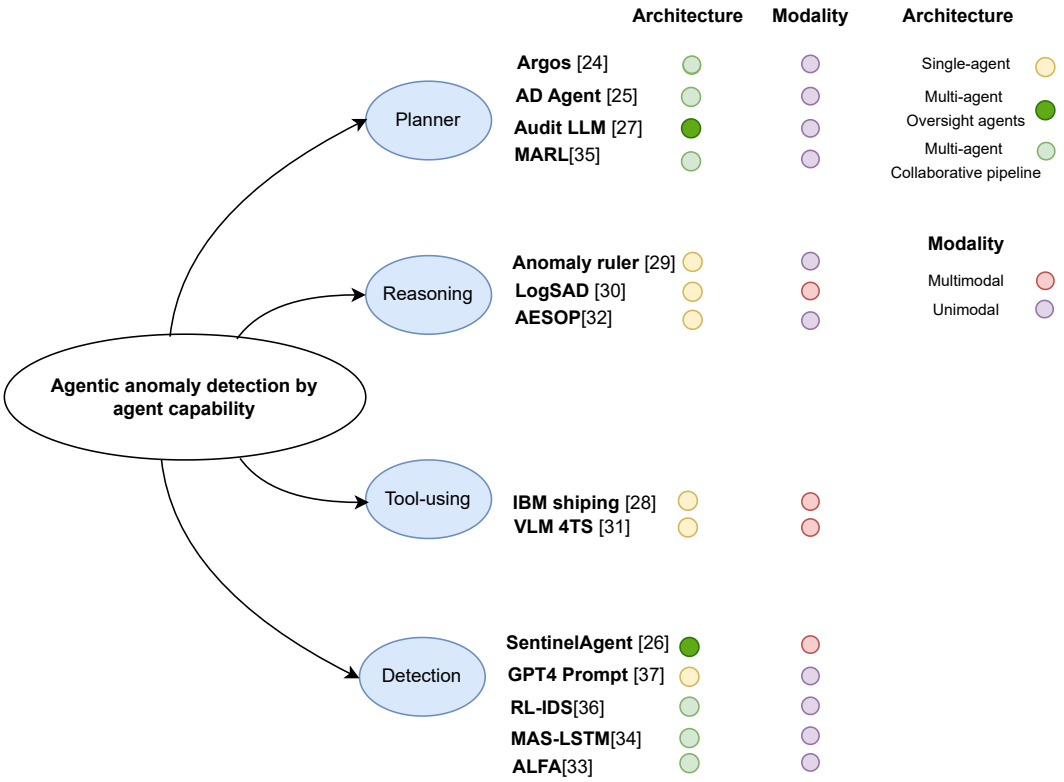

Figure 1. Agentic anomaly detection.

Conventional anomaly detectors have typically modeled a single data modality and a fixed normal profile (e.g., via statistical thresholds, one-class SVMs, nearest-neighbor, or density models). However, the explosion of data complexity and volume has challenged these methods. Deep learning has become a dominant paradigm in recent years. Pang *et al.* [16] highlight that deep anomaly detection (using autoencoders, GANs, normalizing flows, etc.) has emerged as a critical direction with unique problem complexities [17]. These methods can learn rich representations of normal data, but most still assume a single input modality or sensor [18,19]. As a result, purely single-modal approaches may miss anomalies that only manifest when multiple data sources are considered together. To address this, multimodal anomaly detection has gained prominence. Modern systems often have heterogeneous sensors (e.g., image, audio, thermal, etc.), and fusing their information can improve robustness [20,21]. Lin *et al.* [22] presents unsupervised industrial anomaly detection that integrates RGB images with 3D scans and yields much better performance than either alone. Similarly, Li *et al.* [23] introduce MulSen-AD, a multi-sensor anomaly dataset that unifies RGB cameras, laser scans, and infrared thermography. Other domains see analogous trends: in robotics, for example, anomaly detectors now combine proprioceptive signals (forces, motions) with visual-language models to capture both mechanical and environmental failures [24]. In surveillance, fusing video and audio cues has been explored to capture anomalies that are invisible in vision alone. In general, modern anomaly detection increasingly relies on multi-modal inputs to capture complex anomalies that single sensors might miss [23].

Another recent development is agentic AI, autonomous, goal-driven AI agents that can reason, plan, and use tools with minimal human intervention [25]. Agentic AI is designed to pursue complex goals with characteristics such as adaptability and decision-making under uncertainty [26]. Unlike conventional or purely generative AI, agentic systems are built to operate in uncontrolled environments, setting their own subgoals and using external tools or information as needed [27]. Many modern large language model (LLM) applications (e.g., AutoGPT, ReAct agents, and tool-augmented chatbots) fall under this paradigm. Agentic anomaly detection offers the ability of such systems to incorporate reasoning, context, and tool use into the detection process. For instance, LLM-based agents can

 doi:10.20944/preprints202602.1368.v1

utilize contextual knowledge, parse unstructured data (such as system logs or maintenance notes), and iteratively refine anomaly criteria. Recently, several agentic and multimodal anomaly detection methods have been proposed. Russell-Gilbert *et al.* [28] demonstrate framing time-series anomaly detection as a language task, using a pre-trained LLM to interpret sensor readings in context without retraining. Liu *et al.* [17] present a method that combines generative diffusion models with LLMs for human-interpretable anomaly explanations and contextual reasoning. In practice, agentic anomaly systems autonomously query multiple data sources (e.g., visual feeds, databases, knowledge graphs) and apply chain-of-thought reasoning to decide if an anomaly is significant.

Several recent works on anomaly detection provide extensive surveys on conventional and deep learning-based methods. Chandola *et al.* [11] established foundational taxonomies across multiple application domains. Chalapathy *et al.* [29] and Pang *et al.* [16] specifically addressed deep learning techniques, categorizing key neural architectures and methodological advances. Domain-specific reviews such as Cook *et al.* [30] target IoT-based systems, while Erhan *et al.* [31] focus on sensor-centric anomaly detection. Moreover, Choi *et al.* [32] and Garg *et al.* [33] provided practical guidelines, benchmarking methodologies, and comparative evaluations. However, despite this progress, no existing survey comprehensively focuses on the intersection of agentic and multimodal anomaly detection.

In this paper, we fill this gap by providing a comprehensive and up-to-date synthesis of recent advancements in anomaly detection that leverage both agentic AI systems (autonomous, reasoning-capable agents) and multimodal data integration frameworks. We present a unified taxonomy, compare key methods, identify prevailing trends, and outline open research challenges in this rapidly evolving field. Specifically, the contributions of this work are:

- We provide a structured survey of anomaly detection approaches that incorporate both agentic AI and multimodal data fusion.
- We introduce a novel taxonomy to classify existing methods based on agent architecture (single-agent vs. multi-agent), reasoning capability, tool integration, and modality scope.
- We review recent benchmark datasets and evaluation methods for multimodal anomaly detection.
- We present key challenges and summarize mitigation strategies and future directions in agentic and multimodal anomaly detection.

The remainder of the paper is organized as follows: Section 3 and Section 4 review multimodal anomaly detection methods and fusion techniques. Section 2 discusses agentic anomaly detection and classifies agent architectures and capabilities. Section 5 summarizes multimodal anomaly detection datasets and benchmarks. Section 6 highlights key challenges, mitigation strategies, and open research problems. Finally, Section 7 concludes the paper and outlines future research directions.

2. Agentic Anomaly Detection

Agentic anomaly detection (AAD) extends conventional anomaly detection paradigms by integrating autonomy, reasoning, and tool use into detection pipelines. These systems, often built upon large language models (LLMs) or reinforcement learning (RL) agents, enable dynamic decomposition of complex tasks, contextual interpretation of multimodal signals, and adaptive orchestration of detection tools and models. In conventional anomaly detection, a scoring function $s(x)$ is defined for input x and flags an anomaly if $s(x)$ exceeds a threshold τ (or equivalently if the estimated probability of x being generated by a normal data distribution falls below some value α). Unlike static detectors, agentic systems possess the ability to interact with external knowledge sources, communicate between agents, and iteratively refine their predictions based on observations and reasoning outcomes. Agentic systems transform the one-shot detection function into an interactive process that can incorporate external knowledge and multi-step reasoning to improve accuracy and explainability. We categorize agentic AD methods along three principal paradigms: agent architecture, agent capability, and modality scope. Table 1 summarizes representative agentic AD frameworks and their attributes.

 doi:10.20944/preprints202602.1368.v1

Table 1. Taxonomy of agentic anomaly detection methods, grouped by agent architecture, key capabilities, data modality, and evaluation datasets.

Framework	Agent Type	Capabilities	Modality / Scope	Evaluation / Dataset
ARGOS [34]	Planner (Multi-agent LLM)	Workflow planning, tool use, external retrieval	Multimodal (TS + logs + web-text + meta)	KPI, Yahoo, internal Microsoft data
AD-Agent [35]	Multi-agent (LLM pipeline)	Instruction parsing, model selection, code generation	Tabular / Graph / Time-series	ADBench
SentinelAgent [36]	Oversight (LLM tool-user)	Graph modeling, oversight, cognitive inconsistency detection	Multimodal logs + plans + MAS interactions	Simulated email assistant; Magnetic-One
Audit-LLM [37]	Planner (Multi-agent)	Task decomposition, feedback, log auditing	Logs + metadata	Cybersecurity benchmarks
IBM Shipping [38]	Single-agent (LLM+tools)	Reasoning on multimodal sensor/knowledge graph (KG) data	Sensor + KG (maritime)	Real shipping operation logs
AnomalyRuler [39]	Single-agent (Reasoning)	Rule induction, chain-of-thought reasoning	Video	Few-shot video AD benchmarks
LogSAD [40]	Single-agent (Reasoning)	Compositional vision–language reasoning (GPT-4V+CLIP)	Mixed (image + text)	Industrial image and text AD tasks
VLM4TS [41]	Tool-using agent	Vision–language transformation, retrieval-augmented AD	Multimodal (time-series + text)	Time-series AD benchmarks
AESOP [42]	Single-agent (LLM)	Fast anomaly classification with fallback planning	Visual (robotics)	Quadrotor & vehicle simulations
ALFA [43]	VLM (LLM+vision)	Zero-shot visual anomaly detection via prompts	Visual (images)	MVTec, VisA anomaly datasets
MAS-LSTM [44]	Multi-agent (LSTM)	Local LSTM voting-based fusion	Time-series (IIoT)	Industrial IoT traffic
MARL [45]	Multi-agent (RL)	Decentralized RNN predictors, normality scoring	Observations (MARL)	StarCraft (multi-agent env.)
RL-IDS [46]	Multi-agent (RL)	Parallel DQN agents, cost-sensitive learning	Network traffic	CIC-IDS-2017 network dataset
GPT-4 Prompt [47]	Detection-only agent	Zero-shot anomaly classification via prompting	Time-series, Text	Prompt-based scoring benchmarks

 doi:10.20944/preprints202602.1368.v1

2.1. Architectures

Agentic anomaly detection frameworks can be broadly categorized into two architectural paradigms: *single-agent systems* and *multi-agent systems*. A single-agent system relies on one central intelligent agent (typically an LLM or a single RL policy) responsible for the end-to-end anomaly detection task. In contrast, a multi-agent system decomposes the detection pipeline into specialized components, each operated by a dedicated agent with clearly defined roles and responsibilities. Below, we discuss these architectures and their internal operation.

2.1.1. Single-Agent Systems

Single-agent AD frameworks operate with a single decision-making entity and often take on one of two roles: (i) *a tool-augmented orchestrator* or (ii) *a reasoner*. In the first setup, an LLM-based agent serves as a central controller that interfaces with external analytics tools, databases, or knowledge graphs to enrich its decision-making. The agent functions as an orchestrator that queries tools and synthesizes their results before making a final anomaly judgement. For example, Timms *et al.* [38] introduce a GPT-4-based maritime anomaly detection system that augments LLM capabilities with access to sensor logs, environmental databases, and a domain-specific knowledge graph. In the second role, the agent utilizes the inferential and few-shot reasoning abilities of LLMs to derive human-interpretable rules or patterns from normal data and then uses those rules to detect anomalies in new data. Yang *et al.* [39] propose *AnomalyRuler*, a two-stage zero-shot AD framework for video streams using GPT-4V. The agent first induces behavioral rules from a few-shot set of normal videos and then applies these rules to detect and explain anomalies in unseen data. Similarly, Zhang *et al.* [40] introduce *LogSAD*, which combines GPT-4V, CLIP, and SAM to form a rule-based vision-language system. By generating compositional rules aligned with industrial visual anomalies, LogSAD detects structural defects beyond the reach of conventional pixel-level models without model fine-tuning.

2.1.2. Multi-Agent Systems

Multi-agent systems adopt a distributed approach where multiple specialized agents cooperate to accomplish the anomaly detection task. The agents may operate in sequence or in parallel and can even oversee each other's performance. We distinguish two common patterns: collaborative pipelines, in which agents collectively execute different stages of a detection workflow, and oversight architectures, in which certain agents monitor or verify the actions of others to ensure reliability and coherence.

- *Collaborative Pipelines*: In a collaborative multi-agent pipeline, each agent is assigned a specific subtask (data preprocessing, feature extraction, anomaly scoring, explanation, etc.), and the output of one agent feeds into the input of the next. If we label the agents $A_1, A_2, \ldots, A_N$ as they appear in the pipeline, the overall detection function can be viewed as a composite of their operations:

$$y = A_N\big(A_{N-1}(\cdots A_2(A_1(x))\cdots)\big), \tag{1}$$

  where x is the initial input and y the final anomaly decision or score. Each A_i focuses on a delimited aspect of the task, and their coordination can be managed via an LLM-based planner. For example, Gu *et al.* [34] propose *ARGOS*, a multi-agent time-series AD framework that autonomously generates, validates, and refines detection rules using collaborative agents. Similarly, Yang et al.[35] introduce *AD-AGENT*, which employs a team of LLM agents to interactively build a complete anomaly detection pipeline from a high-level user instruction. Not all pipelines are strictly sequential; some agents might work in parallel on different data streams or features. For instance, Qin et al. [44] propose *MAS-LSTM*, where multiple LSTM-based detector agents each monitor a different subset of IIoT sensor streams, and anomalies are decided by voting or averaging their scores.
- *Oversight Agents:* As agentic systems become more complex, ensuring reliability and consistency is essential. Oversight architectures introduce dedicated agents that monitor and verify the outputs of task-oriented agents, effectively performing anomaly detection on the multi-agent system

 doi:10.20944/preprints202602.1368.v1

itself. These agents catch logical inconsistencies, hallucinations, or coordination failures that could compromise anomaly decisions. Formally, let $\{o_i\}$ be a collection of observations or outputs from various agents during an investigation; the oversight agent computes a consistency score or logical coherence measure $C(o_1, o_2, \ldots, o_k)$ over these. If C falls below a threshold (indicating incoherence or inconsistency), the oversight agent flags a meta-anomaly and can intervene (e.g., by resetting certain agents or requesting additional information). This adds a layer of fault tolerance and accountability to the agentic AD pipeline. For example, He et al. [36] propose *SentinelAgent*, which deploys an LLM oversight agent to supervise a team of collaborating agents. Similarly, in the *Audit-LLM* framework for security logs [37], a critic agent reviews the decisions made by a Detector agent and either approves them or asks for refinement, ensuring that high-stakes anomaly alerts

2.2. Agent Capability

The rapid development of autonomous agents has led to frameworks with varying depths of reasoning and autonomy. Plaat et al. [27] categorize agents by their ability to reason, act, and interact, emphasizing how agents decompose tasks, use external tools, and collaborate towards goals. Building on this idea, we classify agentic anomaly detection systems by their functional capabilities, resulting in four major categories: *detection-only agents*, *reasoning agents*, *tool-using agents*, and *planner agents*. Table 2 summarizes these categories, including key examples, strengths, and limitations.

Table 2. Agentic anomaly detection paradigms by agent capability, with representative examples.

Agent Type	Key Capability	Examples	Strengths	Limitations
Detection-Only Agents	Direct labeling via prompt or model output	SigLLM [48]; GPT-4V zero-shot [47]	Simple deployment, fast inference	Prone to LLM errors (hallucinations), no deep reasoning
Reasoning Agents	Chain-of-thought, rule induction from few-shot normals	AnomalyRuler [39]; LogSAD [40]	Explainable decisions; uses in-context learning	Sensitive to prompt design; needs clean normal data
Tool-Using Agents	External API or tool integration	ARGOS [34]; VLM4TS [41]; SentinelAgent [36]	Context-aware and domain-grounded	Tool dependency; higher latency and complexity
Planner Agents	Workflow decomposition, memory usage, multi-step planning	Audit-LLM [37]; ARGOS pipeline; LLM-based FM [3]	Tackles complex multi-stage tasks; dynamic adaptation	Complex architecture; more costly to develop/maintain

2.2.1. Detection-Only Agents

These agents operate similarly to traditional anomaly detectors, except an LLM (or other foundation model) is used to directly produce anomaly labels or scores from the raw input, without performing explicit multi-step reasoning or tool use. In other words, the agent's policy is essentially a single-step mapping $f : x \mapsto y$. For instance, one can prompt a large language model to output whether an input is anomalous or not in a zero-shot approach:

$$y = LLM(Prompt(x)), \tag{2}$$

where the prompt might be a template converting x (which could be a time series, log message, etc.) into a descriptive question for the LLM and y is the model prediction. Detection-only agents forego

 doi:10.20944/preprints202602.1368.v1

complex reasoning in favor of direct inference, making them straightforward to deploy and fast at runtime. Yang *et al.* [49] introduces AD-LLM, the first benchmark to evaluate the capabilities of LLMs for zero-shot anomaly detection, data augmentation, and model selection. Cao et al. [50] introduce TAD-Bench, a benchmark that uses state-of-the-art language model embeddings (from BERT, GPT, etc.) combined with classic anomaly detection algorithms (like isolation forests and autoencoders) for text anomalies. In the time-series domain, Alnegheimish et al. [48] propose SigLLM, a framework that serializes time-series data into textual descriptions and then uses an LLM (through prompting) to identify anomalies. Another line of work directly evaluates the zero-shot detection prowess of GPT-4 and similar models on various data. Dong et al. [47] tested GPT-4 on time-series anomaly detection by providing the model with sequences of values as input and asking it to highlight anomalies. Similarly, Derakhshan *et al.* [51] demonstrates that GPT-4o can effectively detect rework anomalies in business process event logs, achieving high accuracy across different anomaly distributions using zero-shot, one-shot, and few-shot prompting strategies. In general, detection-only agents benefit from the rich prior knowledge of foundation models and their ability to generalize from just a description of the task. They remain useful where low-latency or low-complexity anomaly flagging is required, especially for high-throughput tasks such as real-time sensor screening or log filtering. However, they are prone to the well-known pitfalls of LLMs, such as hallucination and limited transparency.

2.2.2. Reasoning Agents

Reasoning agents incorporate intermediate inference or rule induction into the anomaly detection process, enabling more interpretable and generalizable decisions. Instead of mapping input x directly to an output y, these agents construct a reasoning chain $c_{1:T}$, which guides their final decision:

$$c_{1:T} = \pi(x), \quad y = f(x, c_{1:T}), \tag{3}$$

where $c_{1:T}$ represents a sequence of intermediate steps (e.g., logical inferences, verbalized rules, or structured descriptions), π is a reasoning policy, and f is a decision function that uses both the input and reasoning trace. These steps act as latent variables, often human-readable, providing explainability. For instance, Yang et al. [39] propose *AnomalyRuler*, a zero-shot video anomaly detection framework using GPT-4V. The agent first observes normal video clips to induce temporal and behavioral rules, then applies these rules to new data to detect anomalies via chain-of-thought reasoning. Zhang *et al.* [40] combines GPT-4V and CLIP/SAM to perform logical and structural reasoning over industrial images without any training. Reasoning agents are particularly valuable in domains where explainability or concept-level anomaly interpretation is critical (e.g., surveillance, medical imaging). However, they heavily rely on well-constructed prompts and representative normal examples. Long reasoning chains may also introduce error propagation.

2.2.3. Tool-Using Agents

Tool-using agents augment their own capabilities by calling external tools (knowledge graphs, search engines, anomaly scoring modules, or third-party APIs) during the anomaly detection process. These agents treat tool outputs as additional observations that inform their final decision. We can describe a tool-using agent's operation in two phases: (1) deciding which tool to use and what query to send, and (2) integrating the tool's result into anomaly inference. If q_i is a query formulated by the agent and $T_1, T_2, \ldots, T_k$ denotes external tool functions, the final decision could be represented as

$$y = A\big(x; T_1(q_1), T_2(q_2), \ldots, T_k(q_k)\big), \tag{4}$$

where x is the input data (or its derived features), A is the tool-using agent, and y is the output anomaly decision or score. The LLM agent A thus combines raw input with tool outputs $T_i(q_i)$ to produce a context-informed anomaly assessment. For example, Timms et al. [38] propose a GPT-4-based maritime anomaly detector that ingests real-time vessel sensor logs with queries from external

 doi:10.20944/preprints202602.1368.v1

maritime knowledge graphs and weather APIs. He *et al.* [36] propose SentinelAgent, a framework that uses a graph-based oversight model where agents query external APIs and maintain long-term tool state. He *et al.* [41] introduce VLM4TS, which reformulates time series into vision-language representations and then uses GPT-4V and retrieval tools to reason about anomalies. Gu *et al.* [34] propose ARGOS, which integrates LLM-generated rules with classical rule-based methods for time-series anomaly detection in cloud infrastructure. Tool-using agents blend the power of LLMs with domain-grounded computation. They are suitable for hybrid setups such as industrial pipelines with structured anomaly detection.

2.2.4. Planner Agents

Planner agents orchestrate multi-step workflows for anomaly detection, combining reasoning, action selection, and coordination across tools or sub-agents. These agents are capable of decomposing high-level goals into executable plans, managing task dependencies, and adapting dynamically to intermediate outcomes. Formally, a planner seeks an optimal action sequence Π^* from a space of possible action sequences $\mathcal{S}$ that maximizes some utility (e.g., detection accuracy or information gain):

$$\Pi^*(x) = \arg\max_{\Pi \in \mathcal{S}} U(\Pi; x), \tag{5}$$

where $\Pi = [a_1, a_2, \ldots, a_n]$ is a sequence of actions, such as invoking tools, querying sub-agents, or acquiring new data. In practice, planners often use heuristic or learned strategies rather than solving this optimization exactly, but the formulation emphasizes the agent's need to anticipate consequences and coordinate across steps. A representative example is *Audit-LLM* [37], a framework for insider threat detection in audit logs. An LLM-based planner (Coordinator) oversees three roles: a *Decomposer* breaks high-level questions into queries, an *Executor* runs them on the log data, and a *Critic* validates the results. The planner governs when to trigger deeper analysis or terminate the investigation, enabling adaptive depth. In the financial domain, Park et al. [3] design a multi-agent planner that coordinates tasks like searching external sources (e.g., news or market data), running anomaly checks, and synthesizing findings into a final report. *ARGOS* [34] also includes a planning component that iteratively refines anomaly detection rules by designing and validating experiments. This allows the system to not only detect but also improve its detection logic over time. Planner agents are best suited for real-world, multi-stage, and high-stakes AD scenarios where the detection task requires exploration, hypothesis testing, or intervention. However, their complexity presents challenges: debugging multi-step plans is difficult, coordination errors may arise, and latency increases with plan depth. Moreover, ensuring safe and reliable behavior in these agents remains an open problem.

2.3. Modality Integration

Modern anomaly detection (AD) systems are increasingly required to handle diverse data sources from heterogeneous sensor streams or data sources. Agentic AD methods can therefore be distinguished by their modality scope. A unimodal agent specializes in a single data modality, whereas a multimodal agent is designed to fuse and reason over multiple modalities.

2.3.1. Unimodal Agentic Detectors

Unimodal agents specialize in one modality, such as vision, time series, or logs. These agents are particularly effective in domains where anomalies are confined to a single data type. For example, Dong et al. [47] treat GPT-4 as a time-series agent by prompting it with numeric sequences to identify abnormal patterns. Zhu et al. [43] introduce ALFA, a vision-language model that performs zero-shot image anomaly detection by generating visual descriptions from image inputs and identifying deviations, effectively acting as a vision-centric unimodal agent. Audit-LLM [37], focused on event logs and metadata, operates entirely within the textual domain, making it a log/text unimodal agent. Unimodal agents provide simpler architectures, lower data alignment complexity, and efficient

 doi:10.20944/preprints202602.1368.v1

computation due to reduced input dimensionality. However, they are less effective when anomalies emerge from cross-modal interactions.

2.3.2. Multimodal Agentic Detectors

Multimodal agents process and reason over multiple data modalities, such as vision, time series, text, or structured metadata, to detect anomalies that are only apparent through their interaction. These agents must solve two core challenges: (i) fusing features from heterogeneous modalities and (ii) performing cross-modal reasoning to detect inconsistencies or correlations. These agents must align features across modalities and perform joint reasoning. A common solution is to extract features from each modality and combine them in a shared representation:

$$z^{(i)} = \phi_i(x^{(i)}), \quad z = [z^{(1)}, z^{(2)}, \ldots, z^{(M)}], \tag{6}$$

where ϕ_i is a modality-specific encoder and $[\cdot]$ denotes fusion (e.g., concatenation or attention). An anomaly score $s(z)$ or decision function is then applied, often via an LLM or neural classifier. For example, *SentinelAgent* [36] fuses textual inter-agent messages with structured environment state to detect coordination failures. *LogSAD* [40] jointly analyzes industrial images and machine logs to detect visual–textual inconsistencies. *VLM4TS* [41] reformulates time series as image–text pairs and analyzes them using GPT-4V, effectively treating time-series anomalies as vision-language problems. Similarly, the maritime anomaly detection system by Timms et al. [38] combines numeric sensor data with knowledge graph context to reason about situational normality. Multimodal agents are particularly useful in ambiguous or noisy settings, where cross-modal evidence reduces uncertainty. Mathematically, such agents implicitly learn the joint distribution $P(x^{(1)}, x^{(2)}, \ldots, x^{(M)})$, enabling them to detect anomalies in the co-occurrence structure. These agents also enhance explainability by pointing to complementary evidence. However, they come with added complexity: synchronization across modalities, higher model capacity, and potential complexity if one modality introduces noise.

3. Multimodal Anomaly Detection

Multimodal anomaly detection (MAD) methods detect anomalous patterns by jointly analyzing data from heterogeneous or multiple modalities, such as images, video, audio, text/logs, and sensor streams [22,52]. Traditional anomaly detection methods often operate within a single modality, but recent advances in deep learning and foundation models have enabled the fusion of heterogeneous sources, leveraging cross-modal context to improve robustness, interpretability, and generalization [53, 54]. The emergence of large multimodal models (LMMs), vision-language models (VLMs), and LLMs with reasoning capabilities has further accelerated this shift, enabling more powerful and generalizable anomaly detection pipelines. Recently, multimodal AD methods are utilized in industrial inspection, video surveillance, autonomous systems, and more. In industrial quality control, multimodal AD often fuses visual (RGB camera) and 3D information (depth or point clouds) to detect manufacturing defects. Depth data can reveal surface anomalies or misalignments invisible in color images, significantly improving detection under varying lighting or textures [55]. In multimodal video AD systems, video frames and synchronized audio are combined so that visual cues capture physical actions, while audio can signal alarms [56]. In robotics and autonomous systems, heterogeneous modalities are fused to detect anomalies in robot behavior or environment interactions [24]. We categorize recent advances in MAD into three primary paradigms: foundation models, cross-modal fusion models, and multimodal augmentation and synthesis. Table 3 outlines representative recent methods, their modalities, and key innovations:

 doi:10.20944/preprints202602.1368.v1

Table 3. Multimodal Anomaly Detection Methods and Their Innovations.

Method	Modalities	Key Idea/Innovation
BTF (Back to the Feature) [57]	RGB + 3D (depth)	First RGB–3D industrial AD: add 3D point-cloud features to a pre-trained 2D CNN representation; uses a memory bank of normal feature patches for anomaly scoring.
M3DM (Multimodal 3D AD via Hybrid Fusion) [58]	RGB + 3D	Uses frozen transformers (ViT and Point-MAE) to extract rich features and a memory bank for multimodal features. Hybrid fusion strategy with point-level feature alignment improves on BTF.
CFM (Crossmodal Feature Mapping) [55]	RGB + 3D	Learns mapping functions to translate 2D features into 3D and vice versa using only normal data. Anomalies detected via disagreement beyond threshold; eliminates memory bank.
CPIR (Cross-modal Prediction & Intra-modal Reconstruction) [54]	RGB + 3D	Enhances cross-modal mapping by adding autoencoder reconstruction and a shared latent bridge (LB3M) to ensure anomalies in one modality are caught while maintaining consistency.
3D-ADNAS (AD via neural architecture search) [59]	RGB + 3D	Uses Neural Architecture Search to find optimal multimodal fusion. Two-level search: intra-module (fusion ops and stage) and inter-module (connection of fusion modules).
WS-VAD (weakly supervised video anomaly detection) [56]	Video + Audio	Weakly supervised video anomaly detection using Cross-Modal Fusion Adapter (CFA) and Hyperbolic Graph Attention (HLGAtt). CFA gates noisy/dominant modalities; HLGAtt links segments via hyperbolic embeddings.
AnomalyGPT [60]	Image + Text	Uses large vision–language model (MiniGPT-4) for AD. Prompted fine-tuning maps simulated anomaly images to descriptive text. Image decoder adds fine-grained vision; learned prompts adapt LVLM.
LAD-Reasoner [61]	Image + Text	"Tiny" (3B) multimodal language model trained for logical anomaly detection with natural language explanations. Two-stage training: supervised fine-tuning (SFT) + GRPO reinforcement for reasoning. Based on Qwen-VL.

 doi:10.20944/preprints202602.1368.v1

3.1. Foundation Models

Foundation models such as CLIP, GPT-4V, and MiniGPT-4 are increasingly applied to MAD as unified encoders and interpretable reasoning agents. These models embed multimodal inputs into shared semantic spaces, enabling zero-shot or few-shot anomaly detection with natural language explanations [62]. Ren et al. [39] provide a taxonomy of foundation-model anomaly detectors as encoders, detectors, or interpreters, highlighting their strengths in anomaly representation, direct detection, or explanation generation. In practice, foundation models enable detecting anomalies across modalities with minimal supervision. However, pre-trained foundation models may lack domain-specific sensitivity or precision. Recent works demonstrate that specialized prompting and multimodal instruction tuning significantly improve both detection accuracy and the interpretability of reasoning outputs relative to generic foundation models. For instance, Anomaly-OV [63] integrates GPT-4o with a Look-Twice Feature Matching mechanism to achieve strong zero-shot performance and reliable anomaly descriptions on manufacturing benchmarks (e.g., using Anomaly-Instruct-125k and datasets like VisA-D&R), significantly outperforming generic LLMs. Similarly, LogSAD uses a match-of-thought architecture with GPT-4V, CLIP, and SAM to detect both structural and logical defects in industrial images without any training while also generating compositional rules and calibrated anomaly scores [40].

3.2. Cross-Modal Fusion Models

Cross-modal fusion models effectively combine heterogeneous modalities, such as vision, audio, time series, and text, during model training or inference to improve anomaly detection performance. Peng et al. [64] introduce AVadCLIP, which fuses audio and visual inputs via CLIP embeddings, adaptive audio-visual prompts, and uncertainty-driven distillation to enhance detection under noise and occlusion. Ghadiya et al. [56] propose a Cross-Modal Fusion Adapter (CFA) with hyperbolic graph attention to dynamically balance audio-visual streams, yielding state-of-the-art weakly supervised video AD performance. Barusco et al. [65] adapt vision-based localization methods to spectrograms, enabling fine-grained temporal-frequency anomaly detection in audio with improved interpretability. Lee et al. [66] present CLIPFUSION, which combines CLIP's global discriminative features with conditional diffusion-based reconstruction to jointly improve segmentation and classification on MVTec-AD and VisA datasets. Meanwhile, Wang et al. [67] introduce STADNet, a dual-stream 3D-convolutional model with spatio-temporal attention that fuses RGB and motion for improved surveillance anomaly detection on benchmarks like UCSD Ped2. Finally, Qu et al. [68] propose MFGAN, which integrates temperature, vibration, and acoustic sensor data via attention-based autoencoders to enhance anomaly detection in industrial manufacturing, achieving higher F1 scores than unimodal baselines. These cross-modal fusion methods consistently demonstrate that leveraging complementary modalities significantly enhances anomaly detection performance in both controlled and real-world environments.

3.3. Multimodal Augmentation

Multimodal generative frameworks augment anomaly detection pipelines by simulating rare or hard-to-label anomalies using domain knowledge and conditional generation. For instance, Key Knowledge Augmentation (KKA) [69] uses an LLM to synthesize plausible hard vs. easy anomalies based on domain knowledge, enriching training data and improving classifier boundaries. In audio, FS-TWFR-GMM [70] leverages metadata (e.g., machine types) with diffusion-based generation to simulate anomalies in zero- or few-shot scenarios. Such approaches blur the line between generative modeling and anomaly detection, with LLMs or multimodal models serving both as data generators and reasoning engines. The agentic and multimodal paradigm also extends to more specialized domains. In robotic processes, context-sensitive visual anomaly detection is enabled by combining procedural text with visual inputs through hierarchical VLM prompting [71]. In medical imaging, masked diffusion models trained on healthy scans can detect subtle pathologies without labels [72], while

 doi:10.20944/preprints202602.1368.v1

vision-language models promise even richer context-aware anomaly detection in future healthcare applications.

4. Multimodal Fusion Methods

In multimodal anomaly detection, fusion methods determine how data, features, or decisions from each modality are combined to produce an anomaly score or classification. Fusion strategies can be categorized based on when the modalities are combined (fusion level) and how they are combined (fusion operation).

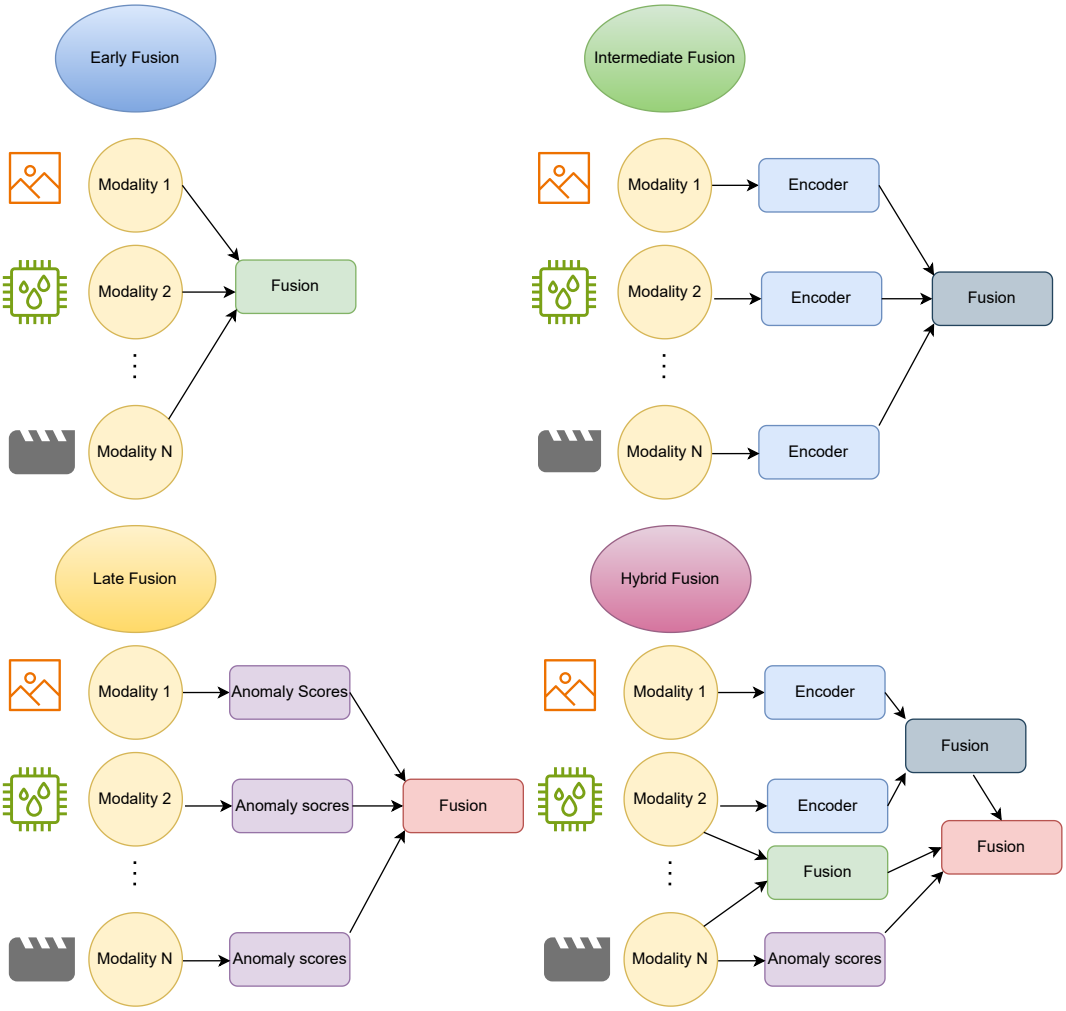

Figure 2. Multimodal Fusion Methods.

4.1. Fusion Stages

Fusion stages can be broadly categorized into early (data-level), intermediate (feature-level), and late (decision-level) fusion, depending on the stage at which modalities are combined within the processing pipeline. Recent multimodal anomaly detection systems often combine multiple fusion levels for improved robustness. For example, a system might perform intermediate fusion between closely related modalities while also applying a late fusion to incorporate a separate expert model decision. There are also hierarchical fusion architectures that integrate modalities at multiple stages (e.g., fusing some modalities early and fusing again at a higher feature level) to capture both low- and high-level cross-modal interactions. Additionally, multitask or multitier fusion frameworks are emerging approaches where different fusion stages are optimized for different objectives (for instance, earlier layers fuse modalities for a representation learning task, while later layers fuse decisions for a specific anomaly classification task) [73]. In general, early fusion maximizes direct cross-modal interaction, late fusion maximizes modularity, and intermediate fusion seeks a balance;

 doi:10.20944/preprints202602.1368.v1

hybrid approaches leverage the benefits of each. Table 4 summarizes common fusion methods along with their advantages and drawbacks.

Table 4. Fusion Methods in Multimodal Anomaly Detection.

Fusion Methods	Description	Pros	Cons
Early Fusion	Merge raw inputs or low-level features, then a single model processes them. *E.g.*, treat LiDAR depth as extra image channels.	Captures raw cross-modal correlations; simple implementation.	Modalities must be aligned; model may be overwhelmed by heterogeneous input.
Intermediate Fusion	Separate encoders, fuse at intermediate layer(s) via concat, add, attention, etc.	Balances modality specialization and interaction; learnable fusion can emphasize important features.	Need to choose *when* and *how* to fuse (hyperparameters); improper fusion point can hurt performance.
Late Fusion	Independent anomaly scores or decisions per modality, combined at end (e.g., weighted average or voting).	Each modality can be optimized/tuned separately; interpretable contributions; robust if one modality fails (others still contribute).	Loses benefit of joint feature learning; needs method to set weights or logic for combining decisions.
Hybrid / Multi-stage	Fuse at multiple points or use a mix of the above (including multi-modal transformers).	Very flexible, can capture both low-level and high-level interactions; often highest accuracy.	Increased complexity; requires sufficient data; harder to interpret and configure.

4.1.1. Early Fusion (Data-Level)

Early fusion concatenates raw or low-level features from all modalities into a joint representation, allowing a single model to learn cross-modal correlations directly. For example, one might concatenate image pixels and depth values channel-wise to form a joint input or merge sensor readings into one feature vector before anomaly modeling. Costanzino et al. [55] propose mapping depth and RGB features via a lightweight cross-modal feature mapper during inference, detecting anomalies by inconsistencies between observed and mapped features. Beyond simple concatenation, some approaches fuse RGB frames with optical flow and depth in a unified tensor before passing through a single network [74]. In log-based monitoring, early fusion has also been applied by integrating multiple system log streams into one high-dimensional feature vector for end-to-end anomaly detection [75]. Early fusion captures direct correlations between modalities from the start but can be susceptible if modalities are not well-aligned or if one modality has much higher dimensionality (it may dominate the representation) [56]. In practice, early fusion is simple but may force the model to learn a very complex mapping from heterogeneous raw data to anomalies.

4.1.2. Intermediate Fusion (Feature-Level)

In feature-level fusion, each modality is first processed by a dedicated encoder to extract higher-level features, and fusion occurs at one or more intermediate layers of the network. For instance, a framework might include a CNN to encode images and an RNN to encode time-series sensor data; their latent feature representations $h^{(1)}$ and $h^{(2)}$ are then fused (e.g., concatenated or combined via attention) at a certain layer to produce a joint representation for anomaly detection. This approach allows each modality to contribute more distilled, modality-specific features to the fusion, often

 doi:10.20944/preprints202602.1368.v1

making the combined representation more informative [76,77]. The fusion point can be a single layer or multiple layers (multi-stage fusion). A common pattern is a dual-stream (or multi-stream) network that processes modalities separately up to a point and then merges their feature streams. Choosing the right layer (or layers) at which to fuse is critical: fusing too early might mix noisy, incompatible raw signals, whereas fusing too late might fail to capture important cross-modal interactions that could be learned at intermediate levels [59]. Learnable fusion operations such as attention or gating can be inserted at the fusion layer to adaptively weight the contributions of each modality. Intermediate fusion thus aims to balance modality specialization with cross-modal learning.

4.1.3. Late Fusion (Decision-Level)

Late fusion involves training separate models for each modality and combines their anomaly scores or decisions via methods such as weighted sums, majority voting, dynamic gating, or mixture-of-experts schemes [78]. This approach treats each modality expert independently up to the decision stage, offering straightforward implementation and modular explainability. It is particularly useful when modalities are largely independent or when safety-critical applications demand that any single modality alone can trigger an alert. However, because it relies on heuristics or separate reconciliation models, late fusion does not learn an integrated cross-modal representation and may struggle when modalities produce conflicting signals. Adaptive late fusion methods, such as mixture-of-experts, address this by learning confidence weights for each expert. For instance, Willibald *et al.* [24] propose a mixture-of-experts framework combining a Gaussian-mixture regression detector on proprioceptive signals with a CLIP-based visual-language model, dynamically selecting the most reliable expert at inference. Lin *et al.*[22] explore a dynamic weighting of video and audio anomaly scores based on environmental conditions.

4.2. Fusion Operations/architectures

Fusion operations or architectures determine how modalities are combined. Formally, let $h^{(1)}$ and $h^{(2)}$ be feature representations from two modalities (these could be raw inputs in an early-fusion setting or intermediate features in a mid-fusion setting). A fusion operation is a function F that combines these representations into a joint feature z:

$$z = F(h^{(1)}, h^{(2)}) \tag{7}$$

There are many choices for F, ranging from simple arithmetic combinations to complex learned modules. Below, we describe several common fusion operations and modules, along with mathematical formulations and examples from recent anomaly detection literature.

4.2.1. Concatenation

The simplest fusion operation is to concatenate the two feature vectors (denoted $[h^{(1)}; h^{(2)}]$) into one larger vector, which can then be processed by subsequent layers (often a linear layer or MLP is applied to the concatenated vector to mix the features). For example, one can define

$$z = \phi([h^{(1)}; h^{(2)}]), \tag{8}$$

where $[h^{(1)}; h^{(2)}]$ denotes the concatenation of $h^{(1)}$ and $h^{(2)}$ and ϕ could be an identity function (direct concatenation) or a learnable transformation (a fully-connected layer). Several multimodal networks use simple concatenation followed by feed-forward layers at the fusion stage [59]. Concatenation preserves all information from both modalities, but it assumes the features are already aligned or comparable in scale, and the burden is on the subsequent layers to identify cross-modal relationships.

 doi:10.20944/preprints202602.1368.v1

4.2.2. Element-wise Addition or Weighted Sum

Another common fusion method is element-wise addition of the feature vectors. If $h^{(1)}$ and $h^{(2)}$ are of the same dimensionality, one can fuse by summing corresponding elements: $z = h^{(1)} + h^{(2)}$. More generally, a weighted sum allows the model to learn the contribution of each modality:

$$z = w_1 \odot h^{(1)} + w_2 \odot h^{(2)}, \tag{9}$$

where w_1 and w_2 are scalar weights or weight vectors (learned or set by prior knowledge) and $\odot$ denotes element-wise multiplication (scaling each feature). This operation effectively averages or emphasizes modalities in each feature dimension. Addition-based fusion is computationally cheap and treats the two feature sets symmetrically. For instance, some network architectures treat a secondary modality as an extra channel appended to the primary modality and use element-wise addition in intermediate layers to blend the information [55]. A variant of weighted sum fusion is to use a gating mechanism that dynamically adjusts weights based on the data. In a gating approach, a function (often a small neural network with a sigmoid activation) computes a gating factor α between 0 and 1 from the inputs:

$$\alpha = \sigma\big(W\,[\,h^{(1)}; h^{(2)}\,]\big), \tag{10}$$

where W is a learnable weight matrix and σ is the sigmoid function. Then the fused output is

$$z = \alpha \odot h^{(1)} + (1 - \alpha) \odot h^{(2)}. \tag{11}$$

Here α (which could be a single scalar or a vector of gating values for different feature dimensions) acts as an adaptive trade-off: if α is close to 1, modality 1 dominates, and if α is close to 0, modality 2 dominates. This kind of gating is used in some attention-fusion modules; for example, the Cross-Modal Feature Attention (CFA) module in an audio-visual fusion model learns a dynamic weight to balance audio and video features for each time frame [56].

4.2.3. Multiplicative or Bilinear Fusion

Multiplicative fusion involves combining features by multiplication interactions rather than addition. The simplest form is element-wise multiplication:

$$z = h^{(1)} \odot h^{(2)}, \tag{12}$$

which yields a fused feature where each dimension $z_i = h_i^{(1)} \cdot h_i^{(2)}$. This operation can highlight feature dimensions that strongly agree across modalities (if both $h_i^{(1)}$ and $h_i^{(2)}$ are high) or dampen those that disagree in sign. Element-wise multiplication requires the two feature vectors to be the same size and implicitly assumes a one-to-one correspondence between their dimensions. A more general form is bilinear fusion, where each feature in one modality is multiplied by each feature in the other modality (an outer product), potentially with learnable weights. A bilinear fusion is defined as:

$$z_{jk} = h_j^{(1)} \cdot h_k^{(2)}, \tag{13}$$

which produces a matrix capturing all pairwise interactions between elements of $h^{(1)}$ and $h^{(2)}$. This outer product can then be vectorized or passed through pooling layers to form the final fused representation. In practice, it is common to introduce a parameterized bilinear form to control the dimensionality, such as:

$$z = \big(h^{(1)}\big)^{\mathsf{T}}\, W\, h^{(2)}, \tag{14}$$

where W is a learned matrix or tensor. This yields each component of z as a weighted combination of pairwise products of $h^{(1)}$ and $h^{(2)}$ elements. Bilinear fusion can capture complex interactions between modalities and has been employed in some vision–language models and multimodal classifiers [79,80].

 doi:10.20944/preprints202602.1368.v1

However, it typically produces very high-dimensional representations (especially if we explicitly take an outer product) and can easily overfit without large amounts of data or proper regularization.

4.2.4. Attention-Based Fusion

Attention mechanisms adaptively fuse modalities by weighting relevant features from one modality using another. In a cross-attention setup, one modality's features (the query set) are used to softly select or weight the features of another modality (the key and value set). Formally, given two sets of features $H^{(1)} = \{h_i^{(1)}\}$ and $H^{(2)} = \{h_j^{(2)}\}$, an attention-based fusion can compute for each feature $h_i^{(1)}$ a weighted combination of the $H^{(2)}$ features:

$$\text{Attn}(h_i^{(1)}, H^{(2)}) = \sum_j \alpha_{ij} h_j^{(2)}, \tag{15}$$

where the attention weights α_{ij} are given by

$$\alpha_{ij} = \text{softmax}_j(h_i^{(1)} W_Q \, (h_j^{(2)} W_K)^T) \tag{16}$$

in a typical formulation (with W_Q, W_K projection matrices for queries and keys, respectively). The result is that each element of one modality can attend to (i.e., extract information from) the elements of another modality that have high content similarity or relevance. The fused representation could be, for instance, the original $H^{(1)}$ features augmented with these attention outputs, or vice versa. To reduce computational cost, attention bottlenecks introduce B learnable latent vectors $b_1, \ldots, b_B$ as intermediaries [76]. Each modality attends to these bottlenecks (at most $N \times B$ interactions), then reverse-attends back, enforcing a compact shared representation. This two-step bottleneck attention generalizes frameworks such as LB3M for RGB+3D fusion in industrial anomaly detection [55], yielding efficient and effective cross-modal integration. Cross-attention is ideal when modalities align semantically (e.g., synchronized video–audio or image–text). For time-series anomaly detection, MST-GAT uses intra- and inter-modal graph attention to learn sensor correlations and improve anomaly scoring [52]. In video surveillance, stacked cross-attention layers fuse RGB, optical flow, and audio, dynamically weighting cues for superior detection [74,81].

4.2.5. Feature Mapping

Rather than fusing features through arithmetic or attention, this approach uses one modality to reconstruct or predict the features of another. A mapping function M is trained to translate from one modality to another: $M : h^{(1)} \mapsto \hat{h}^{(2)}$, where $\hat{h}^{(2)}$ is the predicted counterpart of $h^{(2)}$. During inference, the reconstruction error $|\hat{h}^{(2)} - h^{(2)}|$ serves as an anomaly score, i.e., large deviations indicate cross-modal inconsistency and potential anomalies. This method exploits the idea that, under normal conditions, modalities agree and can predict each other well. For example, Costanzino *et al.* map depth to RGB features and vice versa, detecting anomalies when prediction and observation diverge [55]. Beyond direct mapping, more advanced methods integrate contrastive learning and memory mechanisms. Wang *et al.* propose M3DM for RGB + 3D anomaly detection using patch-wise contrastive learning to align 3D point cloud and image features in a shared embedding space [58]. In general, these methods detect novelty by evaluating how well one modality can explain the other. A failure to do so suggests a modality-specific anomaly that would be missed in unimodal analysis.

4.2.6. Modality and Temporal Alignment

Effective fusion often requires aligning modalities spatially, semantically, and temporally before combining them. In spatially or feature-wise alignment, inputs x_1, x_2 are transformed via $f_1(x_1)$ and $f_2(x_2)$ into a common space (e.g., projecting depth maps or point clouds onto the image plane so each pixel's depth and RGB correspond) [55]. Alternatively, contrastive or correlational objectives (e.g., contrastive loss) train feature extractors so that paired inputs yield embeddings $h^{(1)}$ and $h^{(2)}$ that

 doi:10.20944/preprints202602.1368.v1

coincide, while mismatches diverge. Temporally, modalities sampled at different rates or with offsets must be synchronized. A basic strategy resamples or buffers slower streams (e.g., holding the latest $h^{(2)}$ constant for multiple $h^{(1)}$ frames), creating aligned pairs $(h_t^{(1)}, h_{\lfloor t/r \rfloor}^{(2)})$. More advanced methods employ temporal attention or sequence alignment when events lag across sensors, letting each feature at a time t attend over a window of the other modality timestamps.

4.2.7. Graph-Based Fusion

Graph-based fusion leverages Graph Neural Networks (GNNs) to integrate multimodal information by modeling modalities or components as nodes in a graph, with edges capturing relationships—whether physical, functional, or learned. A GNN then propagates and aggregates information across this graph to produce fused node or graph-level embeddings. In multimodal anomaly detection, each modality (sensor) can be treated as a node, with edges encoding inter-modality dependencies. Through message passing, each node updates its state by blending its own features with those of its neighbors:

$$h_u^{(l+1)} = \sigma\Big(W\, h_u^{(l)} + \sum_{v \in \mathcal{N}(u)} \Theta\, h_v^{(l)}\Big), \tag{17}$$

where $h_u^{(l)}$ is the representation at layer l, $\mathcal{N}(u)$ are neighbors, W, Θ are learnable weights, and σ is an activation function. Ektefaie *et al.* apply this approach to sensor graphs, learning node embeddings that capture both individual and contextual information for anomaly detection [82]. Passos *et al.* extend this with CCA-GNN, which jointly optimizes GNN embeddings and Canonical Correlation Analysis to enforce cross-modal alignment, helping detect anomalies that disrupt typical inter-modality correlations [83]. Xia *et al.* propose VLDFNet, which builds a views-graph over 2D projections of 3D data and fuses image and point cloud features via a disentangled latent space [84]. The graph structure enables modeling both intra- and inter-modal relationships, enhancing robustness in anomaly detection. In general, graph-based fusion is well-suited for scenarios where relations between multiple modalities are key to understanding anomalies. It flexibly encodes multimodal interactions by treating modality as another dimension of connectivity in the graph.

4.3. Fusion Design Principles and Trade-offs

Designing effective multimodal fusion requires balancing integration and modularity: modalities should inform each other without allowing noise or irrelevant signals to degrade performance. Recent research suggests several key design principles:

- **Fuse at Multiple Levels:** Combining mid-level feature fusion with late-stage score fusion enhances robustness. Early fusion may overlook modality-specific noise, while multi-scale fusion captures both structural and semantic information [59].
- **Selective Feature Fusion:** Not all layers contribute equally to cross-modal alignment. Shallow or mid-level features often provide better spatial or temporal grounding across modalities than very early or late representations.
- **Learnable and Adaptive Fusion:** Models that use attention, gating, or mixture-of-experts mechanisms dynamically adjust modality contributions, outperforming static fusion strategies. This adaptiveness is especially valuable in heterogeneous or noisy environments [56].
- **Efficiency:** Fusing mid-level or compact embeddings is computationally efficient compared to raw input fusion. Techniques such as bottleneck layers and dimensionality reduction maintain informativeness while reducing overhead.
- **Domain-Specific Design:** Fusion should be tailored to the task and data characteristics. For instance, if one modality is unreliable or frequently missing, late fusion provides robustness. If anomalies depend on fine-grained correlations across streams (e.g., visual flashes coinciding with audio spikes), then mid-level feature fusion is essential. Domain knowledge can guide initial architecture choices, with further refinement via ablation or architecture search.

 doi:10.20944/preprints202602.1368.v1

In general, multimodal fusion is not a one-size-fits-all problem. The optimal strategy depends on the nature of the modalities, the anomaly types, and the computational constraints of the deployment scenario.

5. Multimodal Datasets and Benchmarks

Multimodal anomaly detection relies on rich datasets that align multiple sensor modalities such as vision, audio, LiDAR, and text, yet truly multimodal benchmarks remain relatively scarce. Recent progress, however, has led to important datasets featuring synchronized multimodal recordings with detailed ground truth annotations. These datasets typically combine modalities like RGB and LiDAR for driving scenarios, audio and video streams, RGB images with depth information, and multisensor industrial data. Existing benchmarks often lack standardized evaluation metrics and underrepresent certain modality combinations, such as vision-text fusion in industrial anomaly detection or sensor-metadata fusion in manufacturing. The MMAD dataset [85] marks progress toward multimodal language model evaluations, although it primarily emphasizes vision. Table 5 presents multimodal datasets and benchmarks with details including modality, size, anomaly types, and label information.

Table 5. Multimodal Anomaly Detection Datasets.

Dataset	Modalities	Size/Scale	Anomaly Types	Ground Truth
AnoVox [86]	RGB + LiDAR	City-scale driving (multisensor)	Spatial and temporal road anomalies	Voxel-level segmentation
MAVAD [87]	Video + Audio	764 videos (11 classes)	Traffic anomalies (e.g. u-turns, obstructions)	Clip-level labels
MVTec LOCO [88]	RGB images	3,644 images (5 categories)	Structural & logical anomalies	Pixel-level masks
MVTec 3D-AD [89]	RGB + Depth (3D)	4,000+ high resolution scans (10 categories)	Surface & depth irregularities	2D masks + precise depth
MMAD [85]	RGB + Text prompts	8,366 images with 39,672 QA pairs	Caption-based AD behaviors	QA accuracy & response correctness
AURSAD [90]	Multisensor time-series	2,045 samples	Robot screwdriving anomalies	Sample-level labels
DoTA [91]	Video	4,677 dash-cam clips	Traffic accidents/anomalies	Temporal, spatial, categorical

6. Challenges, Mitigations and Future Works

Although agentic and multimodal anomaly detection systems provide robust anomaly detection, their deployment faces challenges including heterogeneous data fusion, data scarcity, and high inference costs [92]. LLM-based agentic pipelines compound these issues with added complexity surrounding agent coordination, trust, and security [36]. This section outlines key challenges, mitigations, and future research directions.

6.1. Data Scarcity

Anomalies are inherently rare, leading to severe data imbalance and scarcity of labeled examples for training. This scarcity impedes the learning of robust detectors and often biases models toward the abundant normal patterns. A promising solution to such a problem is synthetic anomaly generation using generative models. Early approaches employed GANs to perturb or reconstruct normal data, thereby creating pseudo-anomalies that enrich the training distribution [93]. More recently, diffusion models have shown the ability to produce high-fidelity synthetic anomalies [94,95]. Despite progress,

 doi:10.20944/preprints202602.1368.v1

ensuring that synthetic anomalies truly reflect real-world outliers remains a challenge. Future research should refine generative processes using more advanced conditional generation (e.g., text- or class-conditioned prompts to specify the anomaly type) and hybrid GAN–diffusion pipelines.

6.2. Multimodal Representation and Alignment

Multimodal anomaly detection requires integrating multiple data modalities into a unified representation, typically leveraging contrastive learning and attention mechanisms for aligning heterogeneous features [96]. Despite recent advances, modality imbalance and naive feature fusion remain major challenges, often suppressing modality-specific anomalies or limiting subtle cross-modal interactions [96]. Methods such as AVadCLIP and adaptive cross-modal fusion adapters dynamically balance and align modalities by adaptively weighting features based on context [56,64]. Future research must further develop adaptive fusion methods and robust alignment techniques capable of effectively handling missing or noisy modalities.

6.3. Agentic Reasoning and LLM Limitations

LLM-based agents enhance interpretability and reasoning but are prone to hallucination and domain mismatches [97]. Misinterpretations in specialized domains highlight the need for domain-adapted LLMs and structured knowledge integration. Future work should enhance LLM factuality and reasoning alignment and develop benchmarks for evaluating agent reasoning fidelity.

6.4. Real-Time Inference and Scalability

For several practical applications (e.g., security monitoring, autonomous driving), anomaly detection must operate under strict latency and throughput constraints. Complex agent frameworks can be computationally expensive, limiting their real-time use. Recent work has highlighted that detectors often involve very large networks, which are not suitable for cost-effective real-time applications [98]. A promising future direction is model compression and efficiency enhancement. Techniques such as knowledge distillation have demonstrated the ability to transfer anomaly detection capabilities from a heavy teacher model to a lightweight student, achieving similar accuracy with dramatically fewer parameters. Future work should explore dynamic model cascades, online model adaptation for data streams, and hardware-friendly implementations (e.g., via model quantization or edge computing) to ensure scalability.

6.5. Interpretability

Interpretability remains a major challenge in anomaly detection, particularly in high-stakes scenarios. Current approaches focus on generating natural-language rationalizations or employing explainable rule-based reasoning to enhance transparency [97]. Additionally, agent traceability, where each decision or action is logged and auditable, is increasingly emphasized to validate anomaly detection processes. Future work should explore developing intuitive interfaces for anomaly explanation, effective communication of uncertainty, and methods ensuring the reliability and trustworthiness of explanations provided to end-users [97].

6.6. Evaluation and Benchmarking

Evaluation and benchmarking in multimodal and agentic anomaly detection face significant limitations due to the lack of standardized benchmarks and consistent metrics, resulting in difficulties comparing methods across studies [97]. Recent efforts such as the MMAD [85] and AnoVox [99] datasets partially address these gaps but remain domain-specific. To progress, the field requires standardized anomaly taxonomies, evaluation protocols, and benchmarks that can handle complex multimodal and dynamic scenarios.

 doi:10.20944/preprints202602.1368.v1

6.7. Theoretical Foundations

The theoretical underpinning of agentic and multimodal anomaly detection remains insufficient, often relying on empirical heuristics. While progress has been made using graph-based anomaly detection for modeling agent interactions [36], there is a pressing need for rigorous theoretical frameworks that can characterize detector behavior under diverse conditions, establish formal performance bounds, and analyze the complex dynamics of multi-agent systems.

Table 6. Key challenges in multimodal/agentic anomaly detection and prospective research directions.

Challenge	Current Mitigation	Future Directions
Data scarcity	Synthetic-anomaly generation via GANs / diffusion, data augmentation	Advanced conditional generation (text / prompts), hybrid GAN–diffusion approaches, few-shot simulation
Modality alignment	Cross-modal embeddings (CLIP, contrastive losses), feature-fusion layers	Unified multimodal representations (transformers), dynamic fusion strategies, multimodal foundation models
LLM domain mismatch	Prompt engineering, few-shot normal exemplars, rule-based prompting	Domain-adapted LLMs, hybrid neuro-symbolic systems, anomaly-aware fine-tuning
Real-time & scalability	Knowledge distillation (LLM to lightweight student), fast/slow pipelines	Model compression, efficient LLM architectures, adaptive inference scheduling
Benchmarking & evaluation	Emerging datasets (MMAD, AnoVox)	Comprehensive multimodal AD benchmarks & metrics, standardized anomaly taxonomies
Theoretical foundations	Ad-hoc frameworks (e.g., graph models for multi-agent systems)	Formal analysis of LLM-agent behavior, robustness theory, multi-agent anomaly theory

7. Conclusions

Anomaly detection is evolving rapidly to meet the demands of increasingly complex, heterogeneous, and dynamic real-world environments. As intelligent systems become more autonomous and multimodal by design, the integration of agentic reasoning with cross-modal perception is not only natural but necessary. This review has presented a comprehensive overview of recent advances at the intersection of multimodal and agentic anomaly detection, two emerging paradigms that enable more robust, interpretable, and context-aware detection systems. We reviewed foundational work in anomaly detection and highlighted the shift from unimodal, static methods to multimodal fusion frameworks and reasoning-capable agents. We introduced a taxonomy that classifies agentic anomaly detection systems by architecture, capability, and modality scope, and we examined recent benchmarks, tools, and datasets that are driving this research frontier. We also identified core challenges and discussed current mitigation strategies and promising research directions. Future research should prioritize adaptive fusion strategies, theory-grounded agent design, and secure, explainable deployments. We hope this work serves as a foundation for researchers and practitioners seeking to develop next-generation anomaly detection systems that are both perceptually rich and autonomously intelligent.

Author Contributions: M.A.B. conducted the literature review, prepared the first draft of the manuscript, and edited the revised versions of the manuscript; A.H reviewed the draft, edited the manuscript, suggested part

 doi:10.20944/preprints202602.1368.v1

of data analysis, and finalized the modifications; A.R. and P.S.R. provided the guidance for the manuscript preparation, reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Research Council of Norway under the project DIGITAL TWIN within the PETROMAKS2 framework (project nr. 318899).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on the source mentioned in the text.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yin, S.; Ding, S.X.; Xie, X.; Luo, H. A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics* **2014**, *61*, 6418–6428. https://doi.org/10.1109/TIE.2014.2301773.
2. Nizam, H.; Zafar, S.; Lv, Z.; Wang, F.; Hu, X. Real-Time Deep Anomaly Detection Framework for Multivariate Time-Series Data in Industrial IoT. *IEEE Sensors Journal* **2022**, *22*, 22836–22849. https://doi.org/10.1109/JSEN.2022.3211874.
3. Park, T. Enhancing Anomaly Detection in Financial Markets with an LLM-based Multi-Agent Framework. *arXiv preprint arXiv:2403.19735* **2024**.
4. Ukil, A.; Bandyoapdhyay, S.; Puri, C.; Pal, A. IoT healthcare analytics: The importance of anomaly detection. In Proceedings of the International Conference on Advanced Information Networking and Applications (AINA), 2016, pp. 994–997. https://doi.org/10.1109/AINA.2016.158.
5. García-Teodoro, P.; Díaz-Verdejo, J.; Maciá-Fernández, G.; Vázquez, E. Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security* **2009**, *28*, 18–28. https://doi.org/10.1016/j.cose.2008.08.003.
6. Bhuyan, M.H.; Bhattacharyya, D.K.; Kalita, J.K. Network anomaly detection: Methods, systems and tools. *IEEE Communications Surveys & Tutorials* **2014**, *16*, 303–336. https://doi.org/10.1109/SURV.2013.052213.00046.
7. Belay, M.A.; Rasheed, A.; Salvo Rossi, P. Digital Twin-Based Federated Transfer Learning for Anomaly Detection in Industrial IoT. In Proceedings of the 2025 IEEE Symposium on Computational Intelligence on Engineering/Cyber Physical Systems (CIES), 2025. https://doi.org/10.1109/CIES64955.2025.11007631.
8. Belay, M.A.; Rasheed, A.; Salvo Rossi, P. Digital Twin Knowledge Distillation for Federated Semi-Supervised Industrial IoT DDoS Detection. In Proceedings of the 2025 IEEE Symposium on Computational Intelligence in Security, Defence and Biometrics Companion (CISDB Companion), 2025, pp. 1–5. https://doi.org/10.1109/CISDBCOMPANION65092.2025.11010678.
9. Belay, M.A.; Rasheed, A.; Rossi, P.S. Digital Twin-Driven Communication-Efficient Federated Anomaly Detection for Industrial IoT. *arXiv preprint arXiv:2601.01701* **2026**.
10. Van Wyk, F.; Wang, Y.; Khojandi, A.; Masoud, N. Real-time sensor anomaly detection and identification in automated vehicles. *IEEE Transactions on Intelligent Transportation Systems* **2020**, *21*, 1264–1276. https://doi.org/10.1109/TITS.2019.2906038.
11. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys* **2009**, *41*. https://doi.org/10.1145/1541880.1541882.
12. Belay, M.A.; Blakseth, S.S.; Rasheed, A.; Salvo Rossi, P. Unsupervised Anomaly Detection for IoT-Based Multivariate Time Series: Existing Solutions, Performance Analysis and Future Directions. *Sensors* **2023**, *23*, 2844. https://doi.org/10.3390/s23052844.
13. Belay, M.A.; Rasheed, A.; Salvo Rossi, P. Sparse Non-Linear Vector Autoregressive Networks for Multivariate Time Series Anomaly Detection. *IEEE Signal Processing Letters* **2025**, *32*, 331–335. https://doi.org/10.1109/LSP.2024.3520019.
14. Belay, M.A.; Bernardino, L.F.; Rasheed, A.; Montañés, R.M.; Salvo Rossi, P. Unsupervised Leak Detection for Heat Recovery Steam Generators in Combined-Cycle Gas and Steam Turbine Power Plants. *IEEE Sensors Journal* **2025**. Under review.

 doi:10.20944/preprints202602.1368.v1

15. Belay, M.A.; Rasheed, A.; Salvo Rossi, P. Autoregressive Density Estimation Transformers for Multivariate Time Series Anomaly Detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025. https://doi.org/10.1109/ICASSP49660.2025.10888728.
16. Pang, G.; Shen, C.; Cao, L.; van den Hengel, A. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys* **2021**, *54*. https://doi.org/10.1145/3439950.
17. Liu, J.; Ma, Z.; Wang, Z.; Zou, C.; Ren, J.; Wang, Z.; Song, L.; Hu, B.; Liu, Y.; Leung, V.C.M. A Survey on Diffusion Models for Anomaly Detection. *arXiv preprint arXiv:2501.11430* **2025**.
18. Belay, M.A.; Rasheed, A.; Salvo Rossi, P. MTAD: Multiobjective Transformer Network for Unsupervised Multisensor Anomaly Detection. *IEEE Sensors Journal* **2024**, *24*, 20254–20265. https://doi.org/10.1109/JSEN.2024.3396690.
19. Belay, M.A.; Rasheed, A.; Salvo Rossi, P. Self-Supervised Modular Architecture for Multi-Sensor Anomaly Detection and Localization. In Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI), 2024, pp. 1278–1283. https://doi.org/10.1109/CAI59869.2024.00226.
20. Haghipour, A.; Tabella, G.; Stang, J.; Rossi, P.S. Sensor Validation in Carbon Capture and Storage Infrastructures. *IEEE Sensors Letters* **2025**.
21. Belay, M.A.; Rasheed, A.; Salvo Rossi, P. Multivariate Time Series Anomaly Detection via Low-Rank and Sparse Decomposition. *IEEE Sensors Journal* **2024**, *24*, 34942–34952. https://doi.org/10.1109/JSEN.2024.3452318.
22. Lin, Y.; Chang, Y.; Tong, X.; Yu, J.; Liotta, A.; Huang, G.; Song, W.; Zeng, D.; Wu, Z.; Wang, Y.; et al. A Survey on RGB, 3D, and Multimodal Approaches for Unsupervised Industrial Image Anomaly Detection. *arXiv preprint arXiv:2410.21982* **2025**.
23. Li, W.; Zheng, B.; Xu, X.; Gan, J.; Lu, F.; Li, X.; Ni, N.; Tian, Z.; Huang, X.; Gao, S.; et al. Multi-Sensor Object Anomaly Detection: Unifying Appearance, Geometry, and Internal Properties. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 9984–9993.
24. Willibald, C.; Sliwowski, D.; Lee, D. Multimodal Anomaly Detection with a Mixture-of-Experts. *arXiv preprint arXiv:2506.19077* **2025**.
25. Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. The Rise and Potential of Large Language Model Based Agents: A Survey. *arXiv preprint arXiv:2309.07864* **2023**.
26. Acharya, D.B.; Kuppan, K.; Divya, B. Agentic AI: Autonomous Intelligence for Complex Goals - A Comprehensive Survey. *IEEE Access* **2025**. https://doi.org/10.1109/ACCESS.2025.3532853.
27. Plaat, A.; Van Duijn, M.; Van Stein, N.; Preuss, M.; Van Der, P.; Kees, P.; Batenburg, J. Agentic Large Language Models, a survey. *arXiv preprint arXiv:2503.23037* **2025**.
28. Russell-Gilbert, A.; Sommers, A.; Thompson, A.; Cummins, L.; Mittal, S.; Rahimi, S.; Seale, M.; Jaboure, J.; Arnold, T.; Church, J. AAD-LLM: Adaptive Anomaly Detection Using Large Language Models. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), 2024, pp. 4194–4203. https://doi.org/10.1109/BIGDATA62323.2024.10825679.
29. Chalapathy, R.; Chawla, S. Deep Learning for Anomaly Detection: A Survey. *arXiv preprint arXiv:1901.03407* **2019**. https://doi.org/10.48550/arxiv.1901.03407.
30. Cook, A.A.; Misirli, G.; Fan, Z. Anomaly Detection for IoT Time-Series Data: A Survey. *IEEE Internet of Things Journal* **2020**, *7*, 6481–6494. https://doi.org/10.1109/JIOT.2019.2958185.
31. Erhan, L.; Ndubuaku, M.; Di Mauro, M.; Song, W.; Chen, M.; Fortino, G.; Bagdasar, O.; Liotta, A. Smart anomaly detection in sensor systems: A multi-perspective review. *Information Fusion* **2021**, *67*, 64–79. https://doi.org/10.1016/j.inffus.2020.10.001.
32. Choi, K.; Yi, J.; Park, C.; Yoon, S. Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines. *IEEE Access* **2021**, *9*, 120043–120065. https://doi.org/10.1109/ACCESS.2021.3107975.
33. Garg, A.; Zhang, W.; Samaran, J.; Savitha, R.; Foo, C.S. An Evaluation of Anomaly Detection and Diagnosis in Multivariate Time Series. *IEEE Transactions on Neural Networks and Learning Systems* **2022**, *33*, 2508–2517. https://doi.org/10.1109/TNNLS.2021.3105827.
34. Gu, Y.; Xiong, Y.; Mace, J.; Jiang, Y.; Hu, Y.; Kasikci, B.; Cheng, P. Argos: Agentic Time-Series Anomaly Detection with Autonomous Rule Generation via Large Language Models. *arXiv preprint arXiv:2501.14170* **2025**.
35. Yang, T.; Liu, J.; Siu, W.; Wang, J.; Qian, Z.; Song, C.; Cheng, C.; Hu, X.; Zhao, Y. AD-AGENT: A Multi-agent Framework for End-to-end Anomaly Detection. *arXiv preprint arXiv:2505.12594* **2025**.

 doi:10.20944/preprints202602.1368.v1

36. He, X.; Wu, D.; Zhai, Y.; Sun, K. SentinelAgent: Graph-based Anomaly Detection in Multi-Agent Systems. *arXiv preprint arXiv:2505.24201* **2025**.
37. Song, C.; Ma, L.; Zheng, J.; Liao, J.; Kuang, H.; Yang, L. Audit-LLM: Multi-Agent Collaboration for Log-based Insider Threat Detection. *arXiv preprint arXiv:2408.08902* **2024**.
38. Timms, A.; Langbridge, A.; O'Donncha, F. Agentic Anomaly Detection for Shipping. In Proceedings of the Proceedings of the NeurIPS 2024 Workshop on Open-World Agents, 2024.
39. Ren, J.; Tang, T.; Jia, H.; Xu, Z.; Fayek, H.; Li, X.; Ma, S.; Xu, X.; Xia, F. Foundation Models for Anomaly Detection: Vision and Challenges. *arXiv preprint arXiv:2502.06911* **2025**.
40. Zhang, J.; Wang, G.; Jin, Y.; Huang, D. Towards Training-free Anomaly Detection with Vision and Language Foundation Models. *arXiv preprint arXiv:2503.18325* **2025**.
41. He, Z.; Alnegheimish, S.; Reimherr, M. Harnessing Vision-Language Models for Time Series Anomaly Detection. *arXiv preprint* **2025**.
42. Sinha, R.; Elhafsi, A.; Agia, C.; Foutter, M.; Schmerling, E.; Pavone, M. Real-Time Anomaly Detection and Reactive Planning with Large Language Models. *arXiv preprint* **2024**.
43. Zhu, J.; Cai, S.; Deng, F.; Ooi, B.C.; Wu, J.; Ooi, C.; Wu, J. Do LLMs Understand Visual Anomalies? Uncovering LLM's Capabilities in Zero-shot Anomaly Detection. In Proceedings of the Proceedings of the ACM Multimedia Conference, 2024, p. 10. https://doi.org/10.1145/3664647.3681190.
44. Qin, Z.; Luo, Q.; Nong, X.; Chen, X.; Zhang, H.; Wong, C.U.I. MAS-LSTM: A Multi-Agent LSTM-Based Approach for Scalable Anomaly Detection in IIoT Networks. *Processes* **2025**, *13*, 753. https://doi.org/10.3390/PR13030753.
45. Kazari, K.; Shereen, E.; Dán, G. Decentralized Anomaly Detection in Cooperative Multi-Agent Reinforcement Learning. In Proceedings of the Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2023, Vol. 1, pp. 162–170. https://doi.org/10.24963/IJCAI.2023/19.
46. Tellache, A.; Mokhtari, A.; Korba, A.A.; Ghamri-Doudane, Y. Multi-agent Reinforcement Learning-based Network Intrusion Detection System. *arXiv preprint* **2024**.
47. Dong, M.; Huang, H.; Cao, L. Can LLMs Serve As Time Series Anomaly Detectors? *arXiv preprint* **2024**.
48. Alnegheimish, S.; Nguyen, L.; Berti-Equille, L.; Veeramachaneni, K. Large language models can be zero-shot anomaly detectors for time series? *arXiv preprint* **2024**.
49. Yang, T.; Nian, Y.; Li, S.; Xu, R.; Li, Y.; Li, J.; Xiao, Z.; Hu, X.; Rossi, R.; Ding, K.; et al. AD-LLM: Benchmarking Large Language Models for Anomaly Detection. *arXiv preprint* **2025**.
50. Cao, Y.; Yang, S.; Li, C.; Xiang, H.; Qi, L.; Liu, B.; Li, R.; Liu, M. TAD-Bench: A Comprehensive Benchmark for Embedding-Based Text Anomaly Detection. *arXiv preprint* **2025**.
51. Derakhshan, M.; Ceravolo, P.; Mohammadi, F. Leveraging GPT-4o Efficiency for Detecting Rework Anomaly in Business Processes. *arXiv preprint* **2025**.
52. Ding, C.; Sun, S.; Zhao, J. MST-GAT: A multimodal spatial–temporal graph attention network for time series anomaly detection. *Information Fusion* **2023**, *89*, 527–536. https://doi.org/10.1016/j.inffus.2022.08.011.
53. Han, X.; Chen, S.; Fu, Z.; Feng, Z.; Fan, L.; An, D.; Wang, C.; Guo, L.; Meng, W.; Zhang, X.; et al. Multimodal Fusion and Vision-Language Models: A Survey for Robot Vision. *arXiv preprint* **2025**.
54. Shangguan, W.; Wu, H.; Niu, Y.; Yin, H.; Yu, J.; Chen, B.; Huang, B. CPIR: Multimodal Industrial Anomaly Detection via Latent Bridged Cross-modal Prediction and Intra-modal Reconstruction. *Advanced Engineering Informatics* **2025**, *65*, 103240. https://doi.org/10.1016/j.aei.2025.103240.
55. Costanzino, A.; Ramirez, P.Z.; Lisanti, G.; Di Stefano, L. Multimodal Industrial Anomaly Detection by Crossmodal Feature Mapping. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. https://doi.org/10.1109/CVPR52733.2024.01631.
56. Ghadiya, A.; Kar, P.; Chudasama, V.; Wasnik, P. Cross-Modal Fusion and Attention Mechanism for Weakly Supervised Video Anomaly Detection. *arXiv preprint* **2024**.
57. Horwitz, E.; Hoshen, Y. Back to the Feature: Classical 3D Features are (Almost) All You Need for 3D Anomaly Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023, pp. 2968–2977. https://doi.org/10.1109/CVPRW59228.2023.00298.
58. Wang, Y.; Peng, J.; Zhang, J.; Yi, R.; Wang, Y.; Wang, C. Multimodal Industrial Anomaly Detection via Hybrid Fusion. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8032–8041. https://doi.org/10.1109/CVPR52729.2023.00776.
59. Long, K.; Xie, G.; Ma, L.; Liu, J.; Lu, Z. Revisiting Multimodal Fusion for 3D Anomaly Detection from an Architectural Perspective. *arXiv preprint* **2024**.

 doi:10.20944/preprints202602.1368.v1

60. Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Tang, M.; Wang, J. AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 1932–1940. https://doi.org/10.1609/aaai.v38i3.27963.
61. Li, W.; Chu, G.; Chen, J.; Xie, G.S.; Shan, C.; Zhao, F. LAD-Reasoner: Tiny Multimodal Models are Good Reasoners for Logical Anomaly Detection. *arXiv preprint arXiv:2504.12749* **2025**.
62. Xu, X.; Cao, Y.; Chen, Y.; Shen, W.; Huang, X. Customizing Visual-Language Foundation Models for Multi-modal Anomaly Detection and Reasoning. In Proceedings of the Proceedings of the IEEE 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2025.
63. Xu, J.; Lo, S.Y.; Safaei, B.; Patel, V.M.; Dwivedi, I. Towards Zero-Shot Anomaly Detection and Reasoning with Multimodal Large Language Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
64. Wu, P.; Su, W.; Pang, G.; Sun, Y.; Yan, Q.; Wang, P.; Zhang, Y. AVadCLIP: Audio-Visual Collaboration for Robust Video Anomaly Detection. *arXiv preprint arXiv:2504.04495* **2025**.
65. Barusco, M.; Borsatti, F.; Pezze, D.D.; Paissan, F.; Farella, E.; Susto, G.A. From Vision to Sound: Advancing Audio Anomaly Detection with Vision-Based Algorithms. *arXiv preprint arXiv:2502.18328* **2025**.
66. Lee, B.; Won, J.; Lee, S.; Shin, J. CLIP Meets Diffusion: A Synergistic Approach to Anomaly Detection. *arXiv preprint arXiv:2506.11772* **2025**.
67. Wang, Y.; Zhao, Y.; Huo, Y.; Lu, Y. Multimodal anomaly detection in complex environments using video and audio fusion. *Scientific Reports* **2025**, *15*, 1–22. https://doi.org/10.1038/s41598-025-01146-4.
68. Qu, X.; Liu, Z.; Wu, C.Q.; Hou, A.; Yin, X.; Chen, Z. MFGAN: Multimodal Fusion for Industrial Anomaly Detection Using Attention-Based Autoencoder and Generative Adversarial Network. *Sensors* **2024**, *24*, 637. https://doi.org/10.3390/s24020637.
69. Chen, D.; Hu, Z.; Fan, P.; Zhuang, Y.; Li, Y.; Liu, Q.; Jiang, X.; Xu, M. KKA: Improving Vision Anomaly Detection through Anomaly-related Knowledge from Large Language Models. *arXiv preprint arXiv:2502.14880* **2025**.
70. Zhang, H.; Zhu, Q.; Guan, J.; Liu, H.; Xiao, F.; Tian, J.; Mei, X.; Liu, X.; Wang, W. First-Shot Unsupervised Anomalous Sound Detection With Unknown Anomalies Estimated by Metadata-Assisted Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2023**, *32*, 1188–1201. https://doi.org/10.1109/TASLP.2023.3330335.
71. Lin, S.; Wang, C.; Ding, X.; Wang, Y.; Du, B.; Song, L.; Wang, C.; Liu, H. A VLM-based Method for Visual Anomaly Detection in Robotic Scientific Laboratories. *arXiv preprint arXiv:2506.05405* **2025**.
72. Iqbal, H.; Khalid, U.; Chen, C.; Hua, J. Unsupervised Anomaly Detection in Medical Images Using Masked Diffusion Model. In Proceedings of the Machine Learning in Medical Imaging. MLMI 2023 (Lecture Notes in Computer Science). Springer, 2023, Vol. 14348, pp. 372–381. https://doi.org/10.1007/978-3-031-45673-2_37.
73. Liang, M.; Yang, B.; Chen, Y.; Hu, R.; Urtasun, R. Multi-Task Multi-Sensor Fusion for 3D Object Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7337–7345. https://doi.org/10.1109/CVPR.2019.00752.
74. Kaneko, Y.; Miah, A.S.M.; Hassan, N.; Lee, H.S.; Jang, S.W.; Shin, J. Multimodal Attention-Enhanced Feature Fusion-based Weekly Supervised Anomaly Violence Detection. *IEEE Open Journal of the Computer Society* **2024**, *5*, 1–12. https://doi.org/10.1109/OJCS.2024.3517154.
75. Zang, R.; Guo, H.; Yang, J.; Liu, J.; Li, Z.; Zheng, T.; Shi, X.; Zheng, L.; Zhang, B. MLAD: A Unified Model for Multi-system Log Anomaly Detection. *arXiv preprint arXiv:2401.07655* **2024**.
76. Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; Sun, C. Attention Bottlenecks for Multimodal Fusion. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021, Vol. 34, pp. 14200–14213.
77. Gan, C.; Fu, X.; Feng, Q.; Zhu, Q.; Cao, Y.; Zhu, Y. A multimodal fusion network with attention mechanisms for visual–textual sentiment analysis. *Expert Systems with Applications* **2024**, *242*, 122731. https://doi.org/10.1016/j.eswa.2023.122731.
78. Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; González, F.A. Gated Multimodal Units for Information Fusion. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Workshop Track Proceedings, 2017.
79. Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear CNN Models for Fine-grained Visual Recognition. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1449–1457.

 doi:10.20944/preprints202602.1368.v1

80. Fukui, A.; Park, D.H.; Yang, D.; Rohrbach, A.; Darrell, T.; Rohrbach, M. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In Proceedings of the Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 457–468. https://doi.org/10.18653/v1/d16-1044.
81. Jeong, S.; Moloco, J.P.; Imani, M. Uncertainty-Weighted Image-Event Multimodal Fusion for Video Anomaly Detection. *arXiv preprint* **2025**.
82. Ektefaie, Y.; Dasoulas, G.; Noori, A.; Farhat, M.; Zitnik, M. Multimodal learning with graphs. *Nature Machine Intelligence* **2023**, *5*, 340–350. https://doi.org/10.1038/s42256-023-00624-6.
83. Passos, L.A.; Papa, J.P.; Del Ser, J.; Hussain, A.; Adeel, A. Multimodal audio-visual information fusion using canonical-correlated Graph Neural Network for energy-efficient speech enhancement. *Information Fusion* **2023**, *90*, 1–11. https://doi.org/10.1016/j.inffus.2022.09.006.
84. Xia, C.; Liu, C.; Zhou, Y.; Li, K.C. VLDFNet: Views-Graph and Latent Feature Disentangled Fusion Network for Multimodal Industrial Anomaly Detection. *IEEE Transactions on Instrumentation and Measurement* **2025**. https://doi.org/10.1109/TIM.2025.3553232.
85. Jiang, X.; Li, J.; Deng, H.; Liu, Y.; Gao, B.B.; Zhou, Y.; Li, J.; Wang, C.; Zheng, F. MMAD: A Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection. *arXiv preprint* **2025**.
86. Bogdoll, D.; Hamdard, I.; Rößler, L.N.; Geisler, F.; Bayram, M.; Wang, F.; Imhof, J.; de Campos, M.; Tabarov, A.; Yang, Y.; et al. AnoVox: A Benchmark for Multimodal Anomaly Detection in Autonomous Driving. *arXiv preprint arXiv:2402.13846* **2024**. Duplicate entry of BogdollAnoVox:Driving.
87. Leporowski, B.; Bakhtiarnia, A.; Bonnici, N.; Muscat, A.; Zanella, L.; Wang, Y.; Iosifidis, A. MAVAD: Audio-Visual Dataset and Method for Anomaly Detection in Traffic Videos. In Proceedings of the Proceedings of the IEEE International Conference on Image Processing (ICIP), 2024, pp. 1106–1112. https://doi.org/10.1109/ICIP51287.2024.10647874.
88. Bergmann, P.; Batzner, K.; Fauser, M.; Sattlegger, D.; Steger, C. Beyond Dents and Scratches: Logical Constraints in Unsupervised Anomaly Detection and Localization. *International Journal of Computer Vision* **2022**, *130*, 947–969. https://doi.org/10.1007/s11263-022-01578-9.
89. Bergmann, P.; Jin, X.; Sattlegger, D.; Steger, C. The MVTec 3D-AD Dataset for Unsupervised 3D Anomaly Detection and Localization. In Proceedings of the Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), 2021, Vol. 5, pp. 202–213. https://doi.org/10.5220/0010865000003124.
90. Leporowski, B.; Tola, D.; Hansen, C.; Iosifidis, A. AURSAD: Universal Robot Screwdriving Anomaly Detection Dataset. *arXiv preprint* **2021**.
91. Yao, Y.; Wang, X.; Xu, M.; Pu, Z.; Wang, Y.; Atkins, E.; Crandall, D.J. DoTA: Unsupervised Detection of Traffic Anomaly in Driving Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45*, 444–459. https://doi.org/10.1109/TPAMI.2022.3150763.
92. Zhao, T.; Zhang, L.; Ma, Y.; Cheng, L. A Survey on Safe Multi-Modal Learning System. *arXiv preprint* **2024**.
93. Li, Z.; Yan, Y.; Wang, X.; Ge, Y.; Meng, L. A survey of deep learning for industrial visual anomaly detection. *Artificial Intelligence Review* **2025**, *58*, 1–82. https://doi.org/10.1007/s10462-025-11287-7.
94. Zhang, X.; Xu, M.; Zhou, X. RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. https://doi.org/10.1109/CVPR52733.2024.01580.
95. Bi, Y.; Huang, L.; Clarenbach, R.; Ghotbi, R.; Karlas, A.; Navab, N.; Jiang, Z. Synomaly Noise and Multi-Stage Diffusion: A Novel Approach for Unsupervised Anomaly Detection in Ultrasound Imaging. *arXiv preprint* **2024**.
96. Ma, X.; Chen, H.; Deng, Y. Improving Multimodal Learning Balance and Sufficiency through Data Remixing. *arXiv preprint* **2025**.
97. Xu, R.; Ding, K. Large Language Models for Anomaly and Out-of-Distribution Detection: A Survey. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL, 2025, pp. 5992–6012. https://doi.org/10.18653/v1/2025.findings-naacl.333.
98. Rakhmonov, A.A.U.; Subramanian, B.; Olimov, B.; Kim, J. Extensive Knowledge Distillation Model: An End-to-End Effective Anomaly Detection Model for Real-Time Industrial Applications. *IEEE Access* **2023**, *11*, 69750–69761. https://doi.org/10.1109/ACCESS.2023.3293108.
99. Bogdoll, D.; Hamdard, I.; Rößler, L.N.; Geisler, F.; Bayram, M.; Wang, F.; Imhof, J.; de Campos, M.; Tabarov, A.; Yang, Y.; et al. AnoVox: A Benchmark for Multimodal Anomaly Detection in Autonomous Driving. *arXiv preprint arXiv:2402.13846* **2024**.

 doi:10.20944/preprints202602.1368.v1