DOI: https://doi.org/10.15276/hait.09.2026.04
UDC 378.147:004.6:331.5

Anomaly detection on time series data for autonomous robot operation

Maksym A. Shavarskyi[1)]
ORCID: https://orcid.org/0000-0002-1379-3244; maksym.a.shavarskyi@lpnu. Scopus Author ID: 58179182100
Yurii P. Kryvechuk[1)]
ORCID: https://orcid.org/0000-0002-2504-5833; Yurii.P.Kryvenchuk@lpnu.ua. Scopus Author ID: 57198358655
[1)] Lviv Polytechnic National University, 12, S. Bandera Str, Lviv, 79013, Ukraine

ABSTRACT

Autonomous robots deployed in safety-critical applications require real-time monitoring systems to detect both environmental and mechanical anomalies. While anomalies in visual sensor data (changing lighting, unexpected obstacles) are often obvious, mechanical failures – such as motor malfunctions or wheel slippage – may not be directly observable in images, making detection challenging. Additionally, real-world operational data is severely imbalanced: normal behavior is well-represented, but anomalous events are rare. This imbalance renders traditional supervised learning approaches ineffective. To address these challenges, this paper presents a staged multimodal autoencoder architecture – a neural network system that processes both visual information (RGB camera images) and motion sensor data simultaneously. Unlike conventional multimodal systems that train all components jointly and suffer from modality competition, proposed architecture employs a three-stage training curriculum that trains visual and motion encoders independently before joint optimization, preventing gradient imbalance and ensuring robust representations. The system performs anomaly detection through reconstruction error analysis: lower errors indicate normal operation patterns, while deviations signal potential anomalies. The method requires only normal operational data for training – no labeled anomaly samples are necessary. Experimental validation demonstrates that the architecture detects visual anomalies (color distortions, unexpected objects) and motion anomalies (sudden stops, jerks, velocity changes) in real-time. The proposed method is needed for safety-critical applications like autonomous robot navigation and warehouse automation, where detecting mechanical and environmental anomalies is essential for operational safety.

Keywords: Anomaly detection; staged training; robot operation; multimodal fusion; imbalanced data

***For citation*: Shavarskyi M. A., Kryvechuk Y. P. “Anomaly detection on time series data for autonomous robot operation”. *Herald of Advanced of Information Technology*. 2026; Vol.9 No.1: 46–59. DOI: https://doi.org/10.15276/hait.09.2026.04**

INTRODUCTION

Anomalies in robotic systems can manifest in two primary ways. First, environmental anomalies may appear in the camera feed – such as unexpected obstacles, changing conditions, or unusual scenes that deviate from trained scenarios. Second, mechanical anomalies may occur within the robot itself, such as wheel slippage, motor malfunctions, or sensor degradation, which may not always be directly visible in images, but significantly affects the robot's behavior and safety. The challenge of anomaly detection in robotics is compounded by the severe class imbalance inherent in real-world operational data. Normal behavior samples vastly outnumber anomalous cases, with imbalance ratios often exceeding 50:1. This characteristic makes traditional supervised learning approaches ineffective, as they tend to optimize for overall accuracy and consequently fail to detect rare but critical failure modes [1], [2]. Recent research in anomaly detection for robotic systems has examined various methodologies, primarily focusing on two approaches: reconstruction-based methods using autoencoders and one-class classification techniques. Reconstruction-based methods (such as variational and denoising autoencoders) have gained prominence because they can learn normal behavior patterns without requiring extensive labeled anomaly data – a significant practical advantage. However, most existing approaches suffer from critical limitations: single-modality focus and modal competition during training. Most systems process either visual data OR sensor data independently, missing opportunities for cross-modal validation. When motion anomalies occur that produce subtle visual artifacts, single-modality approaches may fail to detect them. Also, preliminary research shows that naive joint training of multimodal systems can overshadow lower-level modalities (e.g., motion data) by higher-dimensional modalities (e.g., images), leading to poor motion anomaly detection. To solve these challenges, this paper develops a staged multimodal autoencoder architecture that processes both visual information (RGB camera images) and motion sensor data in a un framework. The system employs a three-stage training curriculum that prevents modality competitionand gradient imbalance, enabling robust detection of anomalies in both environmental and mechanical domains. By combining complementary information from multiple modalities and implementing a curriculum learning strategy, a hypothesis was put forward that the system will achieve superior generalization to novel, unseen failure modes compared to existing single-modality or naively multimodal approaches. The scientific contribution lies in the joint processing of visual and sensor modalities through a unified reconstruction-based framework that operates in a semi-supervised manner, requiring only normal operational data for training. The practical relevance is demonstrated through real-time performance suitable for safety-critical applications, with the system capable of detecting both environmental obstacles and mechanical failures, such as wheel slippage, without requiring labeled anomaly samples during training.

In conclusion, it is necessary to note the relevance of this topic. Autonomous robots are rapidly transforming manufacturing, logistics, healthcare, and defense. The market is expanding significantly, projected to grow 14 % annually through 2030. This explosive growth is driven by labor shortages, rising operational costs, and improvements in sensor and computing technologies. However, widespread deployment of autonomous systems in safety-critical environments introduces significant risk: undetected failures can lead to product damage, facility downtime, environmental contamin-ation, or even personnel injury. Safety regulations are evolving to solve these risks. The ISO/IEC 27001 and ISO 10218 standards now mandate continuous monitoring and anomaly detection in robotic systems. Similarly, autonomous vehicle regulations (SAE Level 3-4) require real-time failure detection and intervention protocols. These regulatory pressures have made robust anomaly detection a non-negotiable requirement for commercial deployment.

In autonomous robotics, ensuring safe operation requires continuous monitoring of the robot's behavior and environment. Modern robotic systems process time series of images captured by onboard cameras during task execution. Under normal conditions, robots operate as expected, but occasionally, anomalies may occur that require immediate detection and intervention to prevent system failures or hazardous situations.

LITERATURE REVIEW AND PROBLEM STATEMENT

Recent research in anomaly detection for robotic systems has examined various approaches. Reconstruction-based methods using autoencoders have become prominent because they can learn normal behavior patterns without needing extensive labeled anomaly data [3], [4]. Chen et al. [5] proposed a sliding-window convolutional variational autoencoder for industrial robots that processes multivariate time series data, enabling real-time anomaly detection. Zhong et al. [6] created a one-dimensional convolutional autoencoder for vibration anomaly detection in industrial robots. Their method showed that convolutional networks can effectively extract spatiotemporal features from sensor data without needing extensive domain knowledge. Likewise, Li et al. [7] introduced a time-frequency convolutional autoencoder with IMU error calibration using Kalman filtering to address noise and joint errors in industrial robots. Multimodal fusion techniques have proven effective in industrial anomaly detection. Wang et al. [8] developed a hybrid fusion method that integrates point cloud and RGB features for industrial applications. Zhou et al. [9] introduced a multimodal fusion algorithm that utilizes time series and image data from agricultural wireless sensors, demonstrating that cross-modal attention mechanisms can enhance anomaly detection performance. Recent comprehensive surveys [10], [11] have categorized multimodal anomaly detection methods into early fusion, middle fusion, late fusion, and hybrid fusion strategies. These studies emphasize that proper feature fusion is critical for capturing complementary information from different modalities. The MFGAN framework [12] demonstrated that multimodal feature fusion mechanisms can significantly improve F1-scores compared to single-modality approaches. Variational autoencoders have been widely explored for anomaly detection tasks. Research by Bouman and colleagues [13] recently questioned the reliability of autoencoders for anomaly detection, showing that anomalies can sometimes be perfectly reconstructed. However, comparative studies [14], [15] have shown that vision transformer-based VAEs exhibit exemplary performance across various scenarios when properly configured. The problem of imbalanced datasets in anomaly detection has been addressed through various strategies. Semi-supervised learning approaches [16], [17] have proven effective by training models exclusively on normal samples and detecting deviations during inference. Liu et al. [18] proposed an autoencoder-based method for the detection of optical failure with imbalanced data, achieving 96.8% accuracy. Self-supervised representation learning enhanced by data augmentation using StyleGAN has also shown promise for manufacturing imbalanced data [19]. Despite these advances, several challenges remain unaddressed. Most existing methods focus either on visual data or sensor data independently, missing opportunities for cross-modal validation. Methods that do incorporate multimodal data often use simple concatenation strategies rather than learning complex cross-modal correlations. Furthermore, many approaches require labeled anomaly samples during training, which is impractical given the rarity and unpredictability of failure modes in real-world robotic systems. As a result of the literature review, only a few approaches are able to do simultaneous anomaly detection on different types of data. To achieve it was decided to use separate autoencoders for visual and sensor data. Differences will be in the training process, which will be divided into 3 stages for learning complex cross-modal correlations more effectively. Further comparisons will be made with model MFGAN from paper [9], because they are most relevant to this paper.

RESEARCH AIM AND OBJECTIVES

This research aims to develop a multimodal anomaly detection system for autonomous robots that can identify both environmental and mechanical anomalies in real-time using time series images and sensor data, using only normal robot behavior data. This model should react to visual anomalies if no sensor anomalies are detected, and vice versa. This is important because these anomalies do not necessarily occur simultaneously. The effectiveness of the developed model can be measured using the reconstruction loss metric.

To achieve this aim, the following objectives were established:

1) design a multimodal autoencoder architecture that fuses visual information from camera images with speed data;

2) implement a semi-supervised learning approach that trains exclusively on normal behavior data without requiring labeled anomaly samples;

3) develop preprocessing pipelines for both image and sensor data that enable robust feature extraction and temporal correlation learning;

4) evaluate the system's performance against MFGAN method and demonstrate better detection of diverse anomaly types.

MATERIALS AND METHODS

Dataset Description: The dataset consists of temporal sequences collected during robotic manipulation tasks (PushT environment) where a robot arm pushes a T-shaped object to target positions. Each data sample includes RGB images of resolution 96×96×3 pixels, 2D action coordinates representing end-effector positions, and speed measurements calculated as Euclidean distance between consecutive action positions (speed = 0 at episode start). The dataset contains 206 normal behavior episodes comprising 20,520 individual frames for training (Fig. 1).

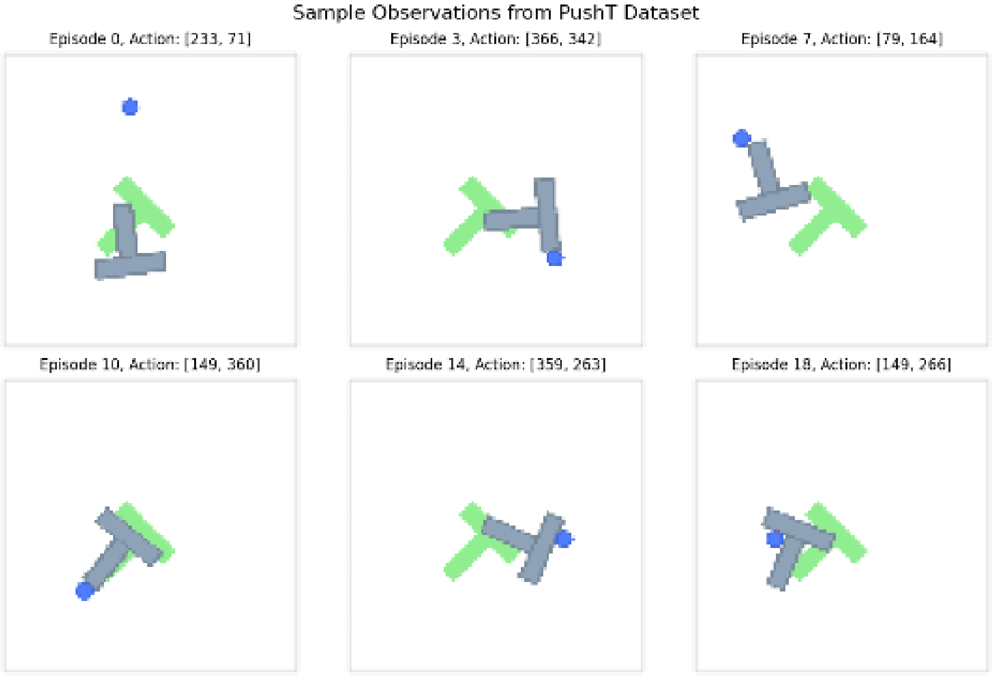

Sample Observations from PushT Dataset
Episode 0, Action: [233, 71]
Episode 3, Action: [366, 342]
Episode 7, Action: [79, 164]
Episode 10, Action: [149, 360]
Episode 14, Action: [359, 263]
Episode 18, Action: [149, 266]

Fig. 1. **Sample Observation from original PushT Dataset**

***Source*: compiled by the authors**

Only 30 anomaly episodes were distributed across visual anomalies, including robot appearance change, target object recoloring and random synthetic perturbations (Fig. 2).

Motion anomalies include sudden stops at high speeds, extended pauses during movement, and velocity jerks with speed spikes up to 20× normal maximum (Fig. 3).

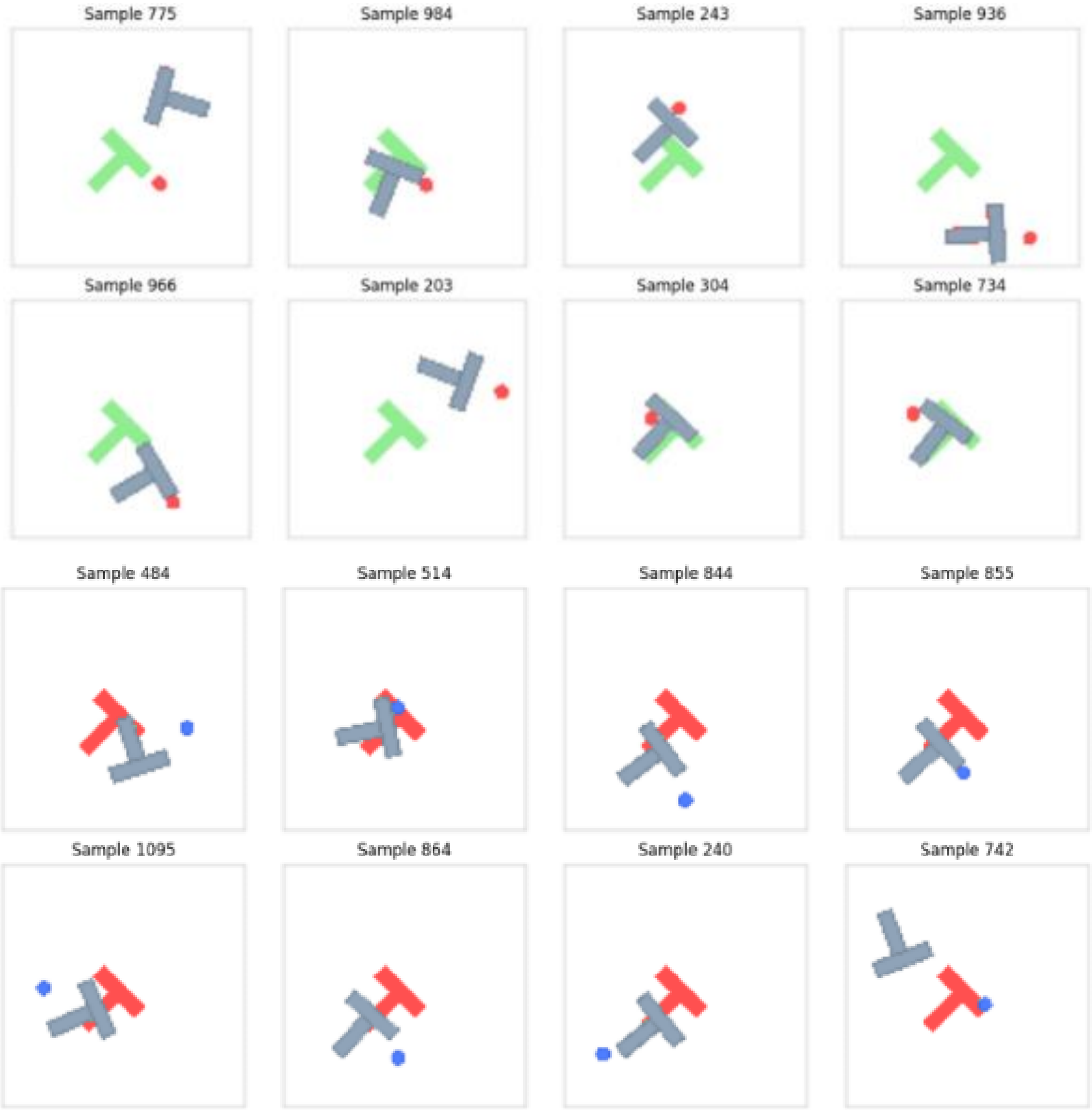


Fig. 2. **Sample Observation visual anomaly in PushT Dataset**

***Source*: compiled by the authors**

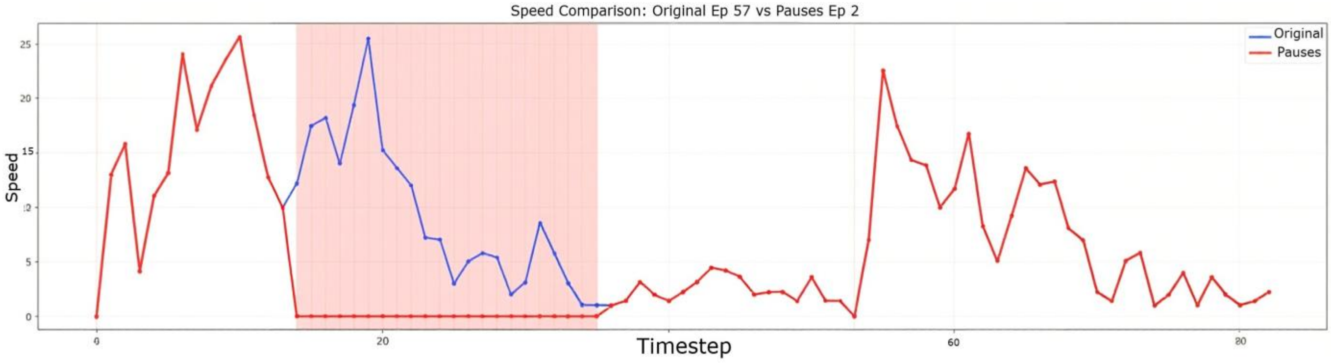


Fig. 3. **Sample Speed non-visual anomaly in PushT Dataset**

***Source*: compiled by the authors**

The total anomaly test dataset contains 42 normal episodes and 30 anomaly episodes: 10 episodes where the robot end effector (blue dot) was colored red and randomly placed fast speed spikes, velocity jerks, and stops. The dataset name is blue-to-red-anomaly-with-speed. 10 episodes where green “T” was colored red and randomly placed fast speed spikes, velocity jerks, and stops. The dataset name is green-to-red-anomaly-with-speed (Fig. 2). The last 10 episodes do not have any visual anomaly, but contain only high-speed parts. The dataset name is fast-speed-anomaly.

Preprocessing Pipeline: Image preprocessing involved extraction from parquet storage where images are stored as flattened uint8 arrays and reshaped to the native 96×96×3 resolution, normalization to [0,1] range through division by 255.0, and tensor conversion from HWC to CHW format for PyTorch compatibility. No resizing was required as images are captured at the target resolution. The current implementation uses no data augmentation to preserve anomaly characteristics, though the architecture supports optional augmentation via random horizontal flips, color jitter, and rotation for future extensions. For temporal multimodal training, sliding windows of 8 consecutive frames with a stride of 4 were used to capture short-term motion dynamics and temporal dependencies between visual states and velocity profiles. Speed data preprocessing included calculation from raw 2D action coordinates as Euclidean distance between consecutive positions with zero initialization at episode boundaries, Z-score normalization computed exclusively on the training set (mean: 10.14, std: 8.83) with 1e-8 epsilon added to prevent division by zero, and no outlier removal to preserve genuine high-speed events that may indicate anomalies. Speed sequences were temporally aligned with corresponding image frames through episode-indexed matching with perfect synchronization (both sampled from the same parquet records). Missing data handling was unnecessary as the dataset contains complete trajectories without gaps. Data splitting employed episode-level partitioning with 80% of episodes (165 episodes) allocated to training and 20 % (41 episodes) to validation using a fixed random seed (42) for reproducibility. This episode-based approach prevents temporal leakage where consecutive highly correlated frames from the same trajectory could appear in both the train and validation sets. Anomaly test sets were generated from held-out episodes not seen during training, ensuring true out-of-distribution evaluation.

The following metrics were used to evaluate the model: accuracy, precision, completeness, f1-score, ROC AUC. To calculate them, it is necessary to determine true positive (TP) and true negative (TN) cases. In this scenario, TP – a real anomaly in the episode occurred, and the model detected it, TN - the episode was normal and no anomaly was detected. The last thing to describe is which episode should be considered an anomaly. In this article, if 3 consecutive anomalies appear in a row, this episode is considered an anomaly. Recall was chosen as a crucial metric. The logic is next: it is better to detect all anomalies even if some of them could be false alarms (False Positive). Threshold, which was calculated and chosen during training for evaluation: 0.002 MSE thresholds for image and 0.03 MAE thresholds for speed.

MODEL ARCHITECTURE AND TRAINING PROCEDURE

Independent Encoders Multimodal Autoencoder Architecture (IEMA)

The architecture employs dual independent encoder branches with minimal fusion to prevent information bottlenecks. The image encoder processes 4 consecutive frames through a CNN with four convolutional blocks (32, 64, 128, 256 filters, kernel_size=3, stride=2) with batch normalization and ReLU activation. Each frame is encoded to 512 dimensions via Conv2D layers followed by flattering from 256×6×6 feature maps and FC projection, then mean-pooled across the sequence. The speed encoder processes 4 normalized speed values through three Conv1D layers (16, 32, 64 channels) with max pooling and FC compression to 16 dimensions. Fusion concatenates image (512-d) and speed (16-d) features into a 528-dimensional combined latent without compression bottleneck. Both decoders receive the full 528-d latent enabling cross-modal information flow. The image decoder uses FC expansion and four transposed convolution blocks (256→128→64→32→3 channels) with batch normalization and sigmoid activation to reconstruct 96×96×3 frames. The speed decoder applies FC projection and transposed 1D convolutions (64→32→16→1 channels) with ReLU for non-negative 4-value speed sequences. The architecture contains 10,422,868 trainable parameters (image encoder: 5.1M, image decoder: 5.3M, speed encoder: 9K, speed decoder: 42K).

Training uses a three-stage curriculum.

Stage 1 trains the image autoencoder (40 epochs, lr=1e-3), Stage 2 freezes the image branch and trains the speed branch with a weighted MAE loss emphasizing low speeds (60 epochs, lr=1e-3), Stage 3 fine-tunes all parameters jointly (30 epochs, lr=1e-4).

Training Procedure: The model employs a three-stage curriculum learning strategy to prevent modality competition and gradient imbalance between image and speed reconstruction tasks.

Stage 1 (Image Branch, 40 epochs): Trains only the image autoencoder using MSE loss while keeping speed branch parameters uninitialized. This establishes strong visual feature extraction before introducing speed modality, preventing the dominant image loss from overwhelming the smaller-magnitude speed loss during joint optimization.

Stage 2 (Speed Branch, 60 epochs): Freezes all image parameters and trains only the speed encoder/decoder with weighted MAE loss emphasizing low-speed samples. This allows the speed branch to converge independently without interference from simultaneously updating image parameters.

Stage 3 (Joint Fine-tuning, 30 epochs): Unfreezes all parameters and jointly optimizes both modalities with combined loss and reduced learning rate (1e-4) to achieve cross-modal coordination without disrupting pre-trained representations (Fig. 4).

Staged training prevents the high-magnitude image gradients (~1e-2) from dominating low-magnitude speed gradients (~1e-4) during simultaneous optimization, which would cause the speed decoder to produce near-zero outputs. Independent convergence ensures both modalities contribute meaningfully to the shared representation. Training was performed on an NVIDIA RTX 3090 using PyTorch, requiring approximately 5 minutes per stage.

Anomaly Detection Method. The proposed method employs reconstruction-based anomaly detection leveraging the autoencoder's inability to accurately reconstruct out-of-distribution samples. During training, the model learns to compress and reconstruct only normal operational patterns from the training distribution. When presented with anomalous inputs containing previously unseen visual appearances (color changes, object distortions) or abnormal motion patterns (sudden stops, jerks, pauses at high speeds), the model produces high reconstruction errors due to the mismatch between learned representations and anomalous features. A critical advantage of the three-stage training strategy is the emergence of cross-modal coupling during Stage 3 fine-tuning. When both branches are jointly optimized with small learning rate (1e-4), the image decoder learns to utilize speed information from the shared 528-dimensional latent representation, and conversely, the speed decoder incorporates visual context (Fig. 5).

This interdependence creates a beneficial anomaly amplification effect: when a speed anomaly occurs (e.g., sudden stop, jerk), the model not only fails to reconstruct the speed sequence accurately but also produces degraded image reconstructions, as the anomalous speed latent disrupts the expected visual-motion correlation learned during joint optimization. Similarly, visual anomalies (e.g., unexpected object colors) propagate reconstruction errors into the speed modality. This cross-modal error propagation enhances detection sensitivity compared to independent single-modality autoencoders, as anomalies manifest as elevated errors in both reconstruction channels even when the anomaly originates from a single modality.

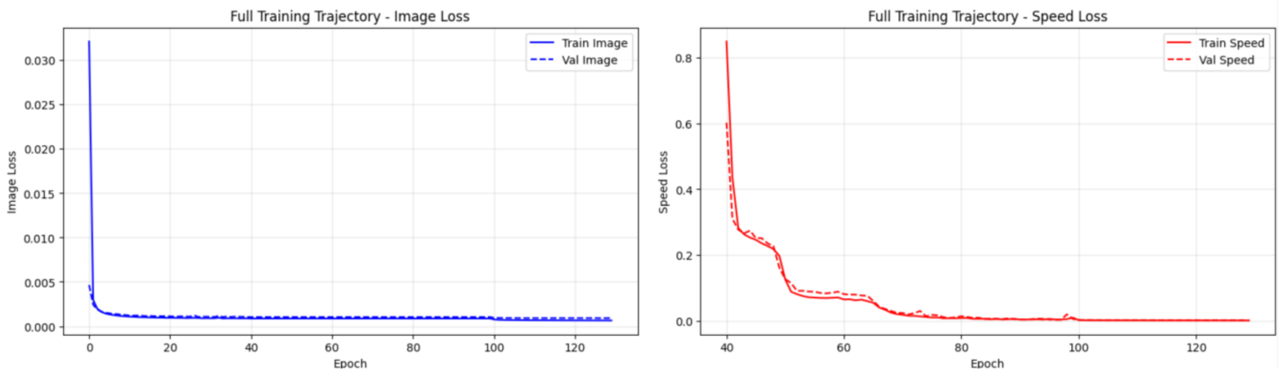


Fig. 4. **Image and Speed reconstruction loss**

***Source*: compiled by the authors**

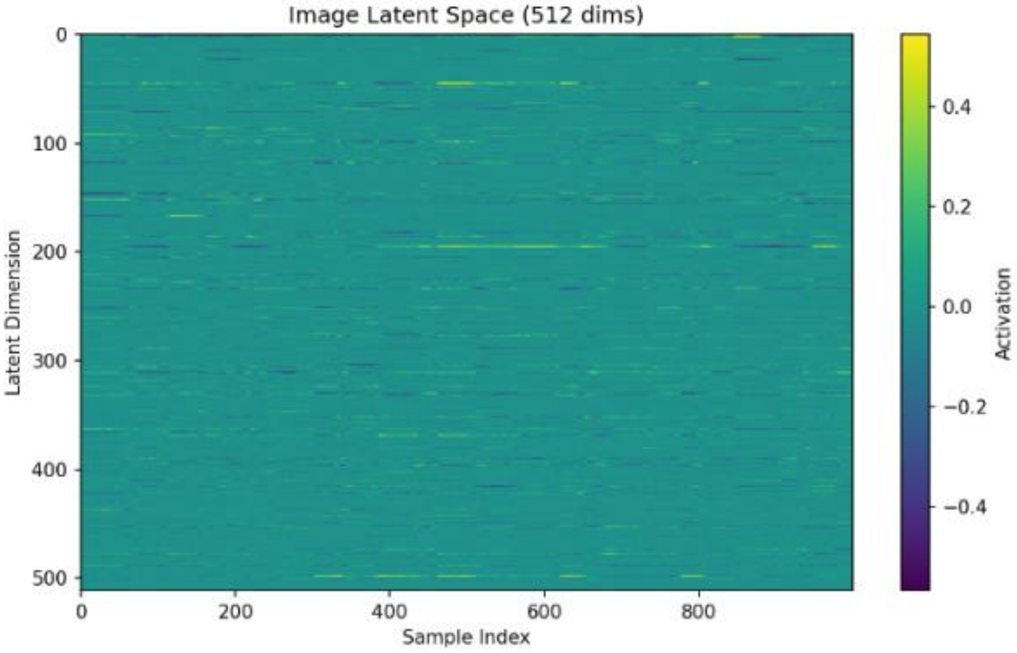


Fig. 5. **Image latent space activation**

***Source*: compiled by the authors**

RESEARCH RESULTS

A multimodal autoencoder architecture was trained to detect anomalies in both visual and motion modalities for robotic manipulation tasks. The model demonstrates the capability to identify visual anomalies (object appearance changes, color distortions) and motion anomalies (sudden stops, pauses, velocity jerks) independently, while also detecting coupled anomalies where both modalities are affected. Critically, the cross-modal coupling established during three-stage training enables the model to flag motion anomalies even when visual appearance remains normal, and conversely, to detect visual anomalies through their impact on predicted motion patterns.

Analysis of the learned 528-dimensional combined latent space (512-d image + 16-d speed) reveals well-structured representations with clear separation between normal operational modes (Fig. 5 and Fig. 6).

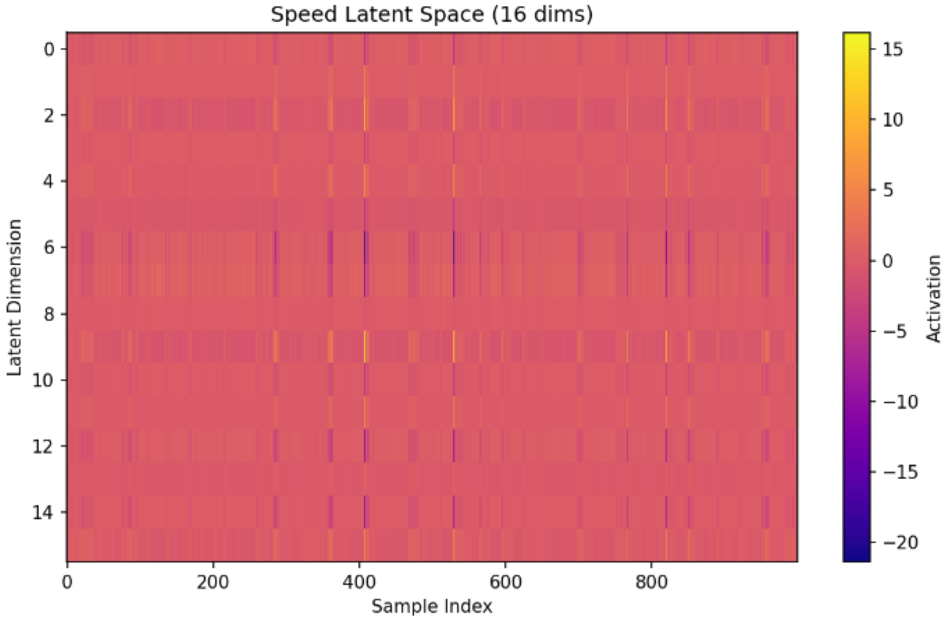


Fig. 6. **Speed latent space activation**

***Source*: compiled by the authors**

PCA visualization shows that the latent space captures primary variance directions corresponding to distinct manipulation phases (approach, contact, push), while correlation analysis confirms that the speed latent dimensions encode motion information complementary to visual features (Fig. 7).

Latent magnitude distributions exhibit consistent statistics across validation samples, indicating the model does not produce degenerate embeddings. High values – strong, confident image features. Low values – weak, uncertain features (Fig. 8).

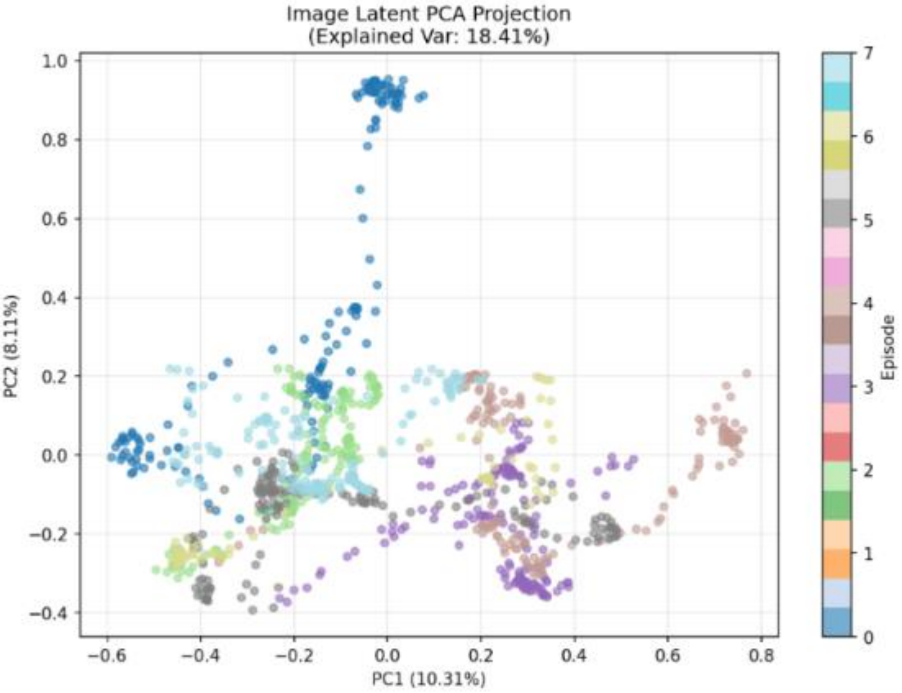


Fig. 7. **Image latent space using PCA projection**

***Source*: compiled by the authors**

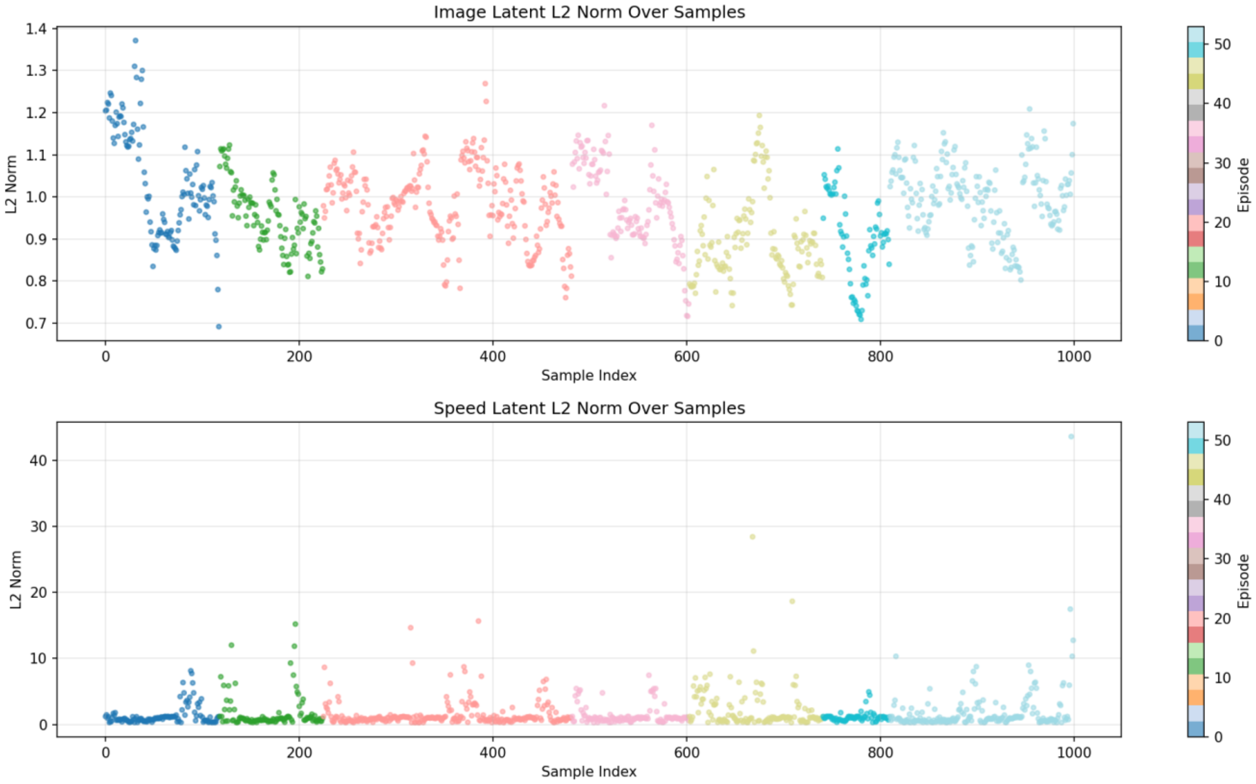
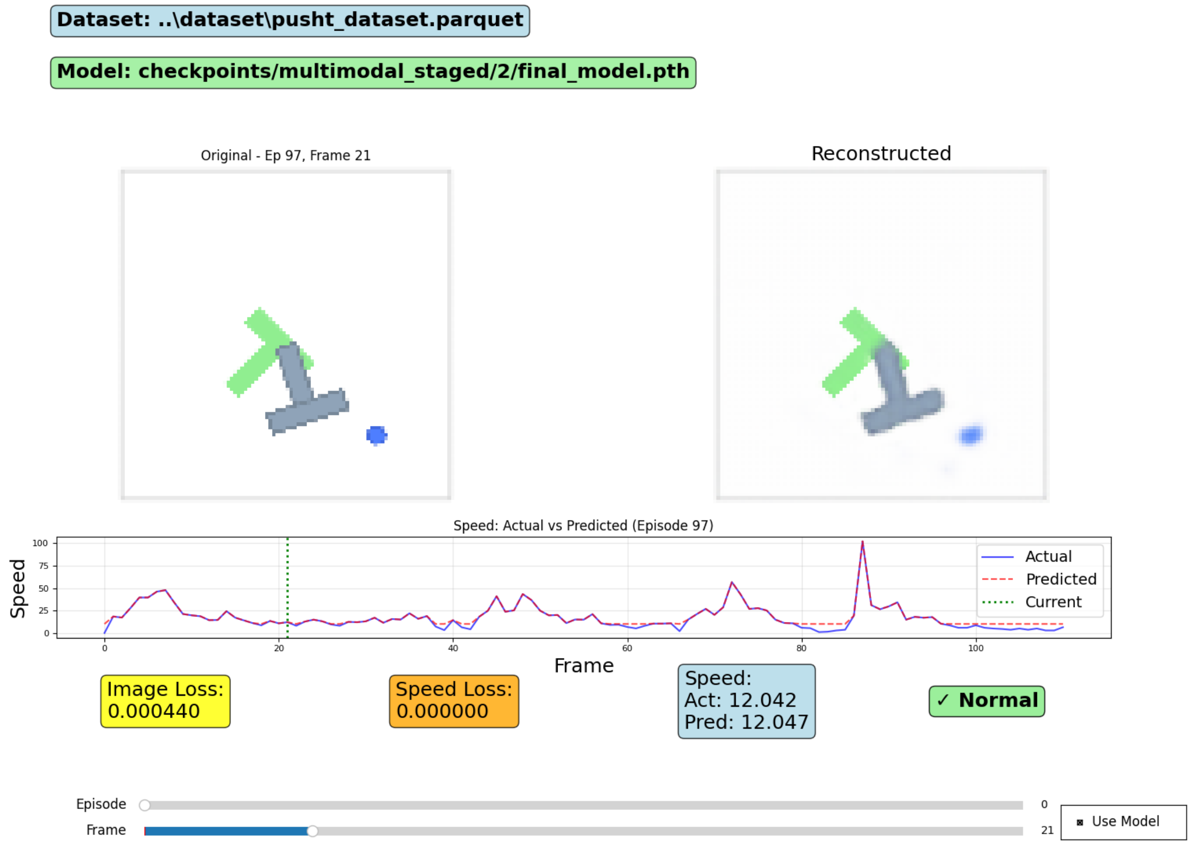


Fig. 8. **Image and Speed latent space L2 magnitudes**

***Source*: compiled by the authors**

The reconstruction examples illustrate how the model performs on normal data versus anomalies. Normal frames are reconstructed with low error, while anomalous frames – whether from color changes or speed variations – exhibit noticeably higher reconstruction loss. (Fig. 9, Fig. 10 and Fig.11).



Dataset: ..\dataset\pusht_dataset.parquet

Model: checkpoints/multimodal_staged/2/final_model.pth

Image Loss: 0.000440

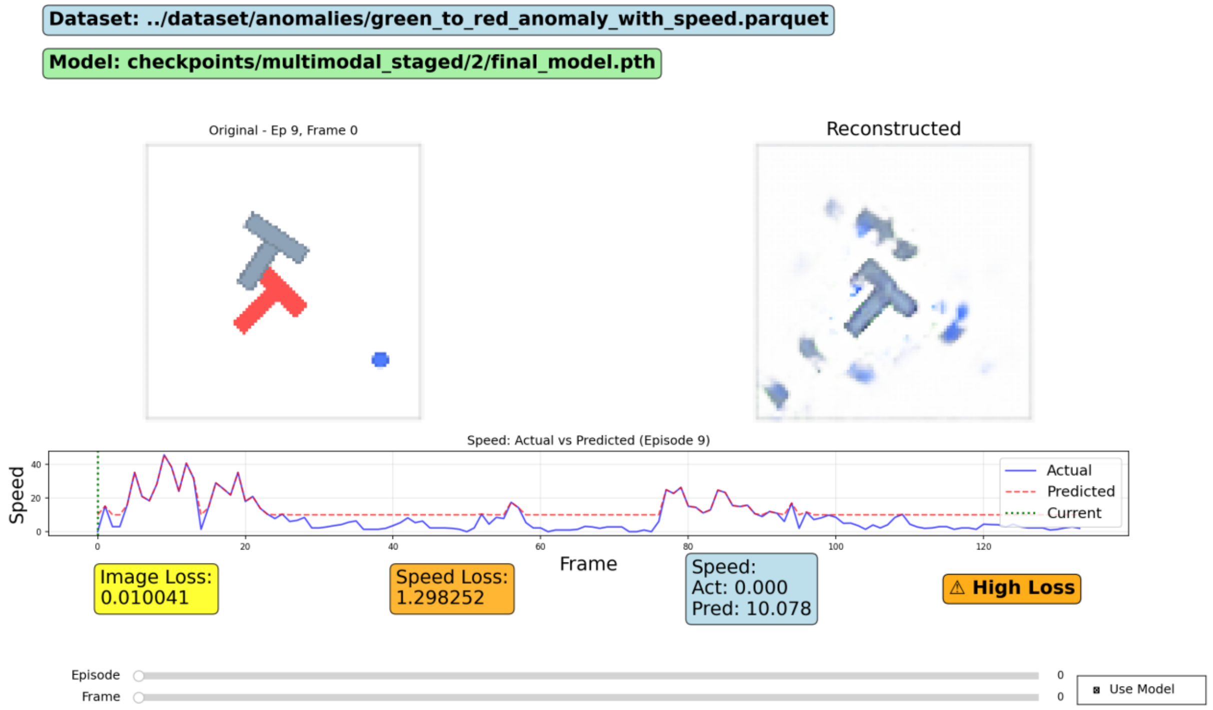
Speed Loss: 0.000000

Speed:
Act: 12.042
Pred: 12.047

✓ Normal

Fig. 9. **Image and Speed reconstruction example on the test dataset**

***Source*: compiled by the authors**



Dataset: ../dataset/anomalies/green_to_red_anomaly_with_speed.parquet

Model: checkpoints/multimodal_staged/2/final_model.pth

Image Loss: 0.010041

Speed Loss: 1.298252

Speed:
Act: 0.000
Pred: 10.078

⚠ High Loss

Fig. 10. **Image and Speed reconstruction example on the test dataset with wrong visual data**

***Source*: compiled by the authors**

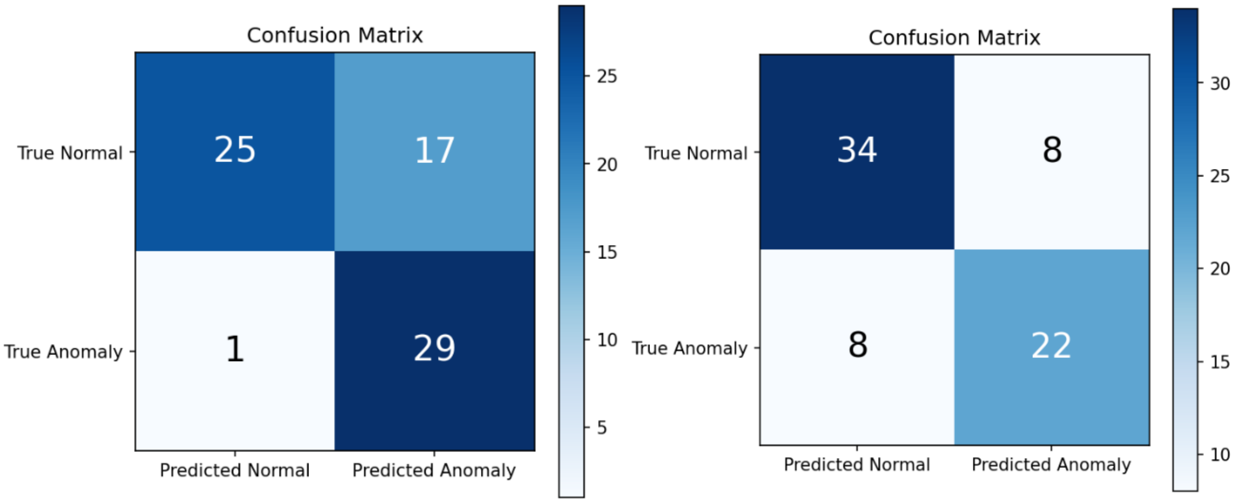


Fig. 11. **Confusion matrix comparison (left is IEMA (Ours), right is MFGAN)**

***Source*: compiled by the authors**

The multimodal anomaly detection model demonstrates strong performance, particularly in its ability to identify anomalies. The model achieved an overall accuracy of 75 % with a recall of 96.67 %, meaning it detected nearly all anomalous episodes in the dataset. The F1-score of 0.7632 reflects a reasonable balance between precision and recall, while the ROC AUC of 0.8603 indicates good discriminative ability. Breaking down detection by anomaly type, the model shows excellent performance across all categories: 100% detection rate for both green-to-red color anomalies and fast speed anomalies, with 90 % detection for blue-to-red anomalies (missing only 1 out of 10 episodes). The trade-off for this high sensitivity is a moderate false positive rate, with 17 normal episodes incorrectly flagged as anomalies (Fig. 12).

For comparison, MFGAN was trained on PoseT dataset. Since anomaly detection prioritizes catching all potential issues – where missing an anomaly could have serious consequences – recall serves as the primary comparison metric. Here is detection results by anomaly type for the proposed and compared model (Table 1 and Table 2).

IEMA (proposed) model outperforms the MFGAN in recall, detecting 29 out of 30 anomalies compared to only 22 out of 30. This advantage is particularly evident in the blue-to-red anomaly category (90 % vs 50 % detection) and fast speed anomalies (100 % vs 70 %). While the other model achieves slightly higher precision due to fewer false alarms, this comes at the cost of missing 8 anomalies – an unacceptable trade-off in safety-critical applications where undetected anomalies pose greater risks than false positives (Table 3).

Table 1. **Detection by anomaly type for our model**

Source	TP	FN	Detection rate
Blue-to-red-anomaly-with speed	9	1	90 %
Green-to-red-anomaly-with-speed	10	0	100 %
Fast-speed-anomaly	10	0	100 %

***Source*: compiled by the authors**

Table 2. **Detection by anomaly type for proposed model**

Source	TP	FN	Detection rate
Blue-to-red-anomaly-with speed	5	5	50 %
Green-to-red-anomaly-with-speed	10	0	100 %
Fast-speed-anomaly	7	3	70 %

***Source*: compiled by the authors**

Table 3. **Comparison with MFGAN**

Metric	IEMA (proposed)	MFGAN
Recall	96.67 %	73.33 %
Precision	63.04 %	73.33 %
Accuracy	75.00 %	77.78 %
F1-score	76.32 %	73.33 %

***Source*: compiled by the authors**

DISCUSSION OF RESULTS

The experimental results demonstrate that the proposed staged multimodal autoencoder architecture is able to address the critical challenge of anomaly detection in severely imbalanced robotic datasets through a novel reconstruction-based semi-supervised learning approach. The achieved image reconstruction MSE of 0.002 and speed reconstruction MAE of 0.03 on validation data indicate that the model has learned representations of normal operational patterns. This low reconstruction error on normal samples is critical for establishing a reliable baseline against which anomalous inputs will produce elevated errors, thereby enabling effective threshold-based detection without requiring labeled anomaly samples during training. The three-stage training strategy prevents modality competition by training image and speed encoders independently before joint optimization. This addresses gradient imbalance issues that plague standard multimodal approaches, where differences in gradient magnitude (~1e-2 for images vs. ~1e-4 for speed) would cause one modality to dominate. The result is a cross-modal amplification effect: anomalies in one modality propagate errors to the other, enhancing overall detection sensitivity compared to single-modality approaches. Proposed model outperforms the baseline in recall (96.67 % vs. 73.33 %), detecting 29 of 30 anomalies. While the MFGAN achieves higher precision, it misses 8 anomalies – an unacceptable trade-off for safety-critical applications where undetected failures pose greater risks than false positives. The small training-validation gap (<0.0005 for images) indicates strong generalization capability. Key limitations include validation on synthetic data only, with unknown transferability to physical robots. The test set contains only six known anomaly categories, and the 40% false positive rate may cause unnecessary shutdowns in real deployment. These issues should be addressed through real-world validation and threshold optimization in future work. The staged training approach solves a genuine technical problem—modality competition not explicitly addressed in existing multimodal anomaly detection methods. This contribution advances the field and extends beyond robot manipulation to autonomous vehicles, industrial machinery, and medical devices where visual and non-visual sensor fusion is critical.

LIMITATIONS AND FUTURE WORK

The current study presents several important limitations that must be acknowledged. First, all validation was conducted on synthetic data from the PushT simulation environment. While simulation provides controlled experimental conditions and comprehensive ground truth annotations, the transferability to physical robot platforms remains unverified. Real-world complications – including sensor noise, mechanical wear patterns, lighting variations, occlusions, and unmodeled dynamics – may differ substantially from simulation, potentially degrading detection performance. Second, the false positive rate warrants careful consideration. The model flags 17 normal episodes as anomalous out of 42 total flagged instances, representing a 40% false positive rate. In real operational scenarios, false alarms could cause unnecessary robot shutdowns, reducing efficiency and user confidence in the system. This trade-off – high recall for safety versus acceptable false positive rate for practicality – remains a critical challenge for deployment in unstructured environments. Finally, the current approach assumes that normal operational data is available and representative of actual operation. The model's ability to detect novel anomalies depends on learning a sufficiently rich representation of normal operation; rare but valid operational scenarios not present in training data may be incorrectly flagged as anomalies. What about future works, first of all needs to reduce False Positives. Dynamic threshold adaptation based on recent operation history and confidence calibration techniques (temperature scaling, uncertainty quantification) should be implemented to distinguish high-confidence anomalies from ambiguous near-normal states. Modality-specific error weighting – emphasizing motion cues for slip detection and visual cues for collision detection – could leverage the multimodal architecture's strengths to improve specificity. Then, real-world validation. Physical robot experiments on actual manipulation tasks are essential to assess transferability and identify domain-specific failures missing in simulation. Collecting data from multiple robot instances would evaluate robustness across hardware variations and wear patterns. Next is extended evaluation. The model should be tested on expanded anomaly types (mechanical damage, sensor degradation, electrical faults) and across different manipulation tasks (reaching, grasping, and assembly) to assess true generalization and out-of-distribution detection capability. Finally, model improvements. Online adaptation mechanisms would allow the system to learn new normal operational patterns without full retraining, critical for long-term deployment with gradual mechanical wear. Attention-based mechanisms could provide interpretability by highlighting which image regions and motion features trigger anomaly decisions—essential for safety-critical applications where operators must understand system behavior. These developments would advance the staged multimodal approach toward practical deployment in real-world autonomous robot safety systems.

CONCLUSIONS

This research solves the challenge of anomaly detection in severely imbalanced robotic datasets through a novel approach that achieves the following:

Objective 1. Multimodal Autoencoder Architecture: A staged multimodal autoencoder architecture was designed and trained to fuse visual information from camera images with speed data. The architecture achieved an image reconstruction MSE of 0.002 and a speed reconstruction MAE of 0.03, demonstrating the ability to learn compact representations of normal operational patterns. The three-stage training approach addresses modality competition by training image and speed encoders independently in Stage 1 and 2 before joint optimization in Stage 3. Cross-modal error propagation was demonstrated where anomalies in one modality produce elevated reconstruction errors in both channels. This objective was achieved.

Objective 2. Semi-supervised Learning Approach: A semi-supervised learning approach was implemented that trains exclusively on normal robot behavior data without requiring labeled anomaly samples. The model achieved 96.67 % recall on the test set, detecting 29 of 30 anomalies using only normal operation examples during training. The reconstruction-based paradigm enables anomaly detection through deviation from learned normal patterns rather than memorization of specific failure modes.

Objective 3. Preprocessing Pipelines: Preprocessing pipelines were developed for both visual and sensor data. Image preprocessing included extraction at 96×96×3 pixel resolution without data augmentation. Sensor preprocessing computed Euclidean distance between consecutive action positions with Z-score normalization applied exclusively on the training set (mean: 10.14, std: 8.83) with 1e-8 epsilon added to prevent division by zero. These pipelines enabled temporal correlation learning through 8-frame sliding windows with a 4-frame stride.

Objective 4. Evaluate the system's performance and compare against MFGAN: IEMA (proposed) model was compared with MFGAN [9]. The recall metric was chosen as a crucial metric. Results show that IEMA model achieves better recall than MFGAN on PoseT dataset. IEMA model outperforms the MFGAN in recall, detecting 29 out of 30 anomalies compared to only 22 out of 30 (~96% vs ~73%).

REFERENCES

1. Chen, T., Liu, X., Xia, B., Wang, W. & Lai, Y. “Unsupervised anomaly detection of industrial robots using sliding-window convolutional variational autoencoder”. *IEEE Access.* 2020; 8: 47072–47081, https://www.scopus.com/pages/publications/85082007681.
DOI: https://doi.org/10.1109/ACCESS.2020.2977892.

2. Alfeo, A. L., Cimino, M. G.C.A., Manco, G., Ritacco, E. & Vaglini, G. “Using an autoencoder in the design of an anomaly detector for smart manufacturing”. *Pattern Recognition Letters.* 2020; 136: 272–278, https://www.scopus.com/pages/publications/85086889361. DOI: https://doi.org/10.1016/j.patrec.2020.06.008.

3. Neloy, A. A. & Turgeon, M. “A comprehensive study of auto-encoders for anomaly detection: Efficiency and trade-offs”. *Results in Engineering.* 2024; 23: 102526, https://www.scopus.com/pages/publications/105027927578. DOI: https://doi.org/10.1016/j.mlwa.2024.100572.

4. Torabi, H., Mirtaheri, S. L. & Greco, S. “Practical autoencoder-based anomaly detection by using vector reconstruction error’. *Cybersecurity.* 2023; 6 (1), https://www.scopus.com/pages/publications/85145601143. DOI: https://doi.org/10.1186/s42400-022-00134-9.

5. Zhong, Z., et al. “Industrial robot vibration anomaly detection based on sliding window one-dimensional convolution autoencoder”. *Shock and Vibratio.* 2022; 2022 (1): 1179192, https://www.scopus.com/pages/publications/85132345987. DOI: https://doi.org/10.1155/2022/1179192.

6. Zhou, Z., et al. "Multimodal fusion anomaly detection model for agricultural wireless sensors". *Engineering Reports*. 2024; 6 (10): e13021, https://www.scopus.com/pages/publications/85205894350. DOI: https://doi.org/10.1002/eng2.13021.

7. Lin, Y., et al. "A Survey on RGB, 3D, and Multimodal Approaches for Unsupervised Industrial Image Anomaly Detection". *Information Fusion*. 2025; 115: 103139, https://www.scopus.com/pages/publications/105002034587. DOI: https://doi.org/10.1016/j.inffus.2025.103139.

8. Liu, J., et al. "Deep industrial image anomaly detection: A survey". *Machine Intelligence Research*. 2024; 21 (1): 104–135, https://www.scopus.com/pages/publications/85182489543. DOI: https://doi.org/10.1007/s11633-023-1459-z.

9. Qu, X., et al. "MFGAN: Multimodal fusion for industrial anomaly detection using attention-based autoencoder and generative adversarial network". *Sensors*. 2024; 24 (2): 637, https://www.scopus.com/pages/publications/85183318391. DOI: https://doi.org/10.3390/s24020637.

10. Jian, C. & Ao, Y. "Imbalanced fault diagnosis based on semi-supervised ensemble learning". *Journal of Intelligent Manufacturing*. 2023; 34 (7): 3143–3158, https://www.scopus.com/pages/ publications/85136958787. DOI: https://doi.org/10.1007/s10845-022-01985-2.

11. Liu, S., et al. "Semi-supervised anomaly detection with imbalanced data for failure detection in optical networks". *Optical Fiber Communication Conference (OFC)*. 2021. – Available from: https://opg.optica.org/abstract.cfm?URI=OFC-2021-Th1A.24.

12. Kim, Y., et al. "Self-supervised representation learning anomaly detection methodology based on boosting algorithms enhanced by data augmentation using StyleGAN for manufacturing imbalanced data". *Computers in Industry*. 2023; 153 (332): 104024, https://www.scopus.com/pages/publications/8511210330. DOI: https://doi.org/10.1016/j.compind.2023.104024.

13. Olivato, M., et al. "A Comparative analysis on the use of autoencoders for robot security anomaly detection". In *Third IEEE International Conference on Robotic Computing*. Naples, Italy. 2019. p. 984–989, https://www.scopus.com/pages/publications/85081158048. DOI: 10.1109/IROS40897.2019.8968105.

14. Pang, G., et al. "Deep learning for anomaly detection: A Review". *ACM Computing Surveys*. 2021; 54 (2): 1–38, https://www.scopus.com/pages/publications/85102487714. DOI: https://doi.org/10.1145/3439950.

15. Chandola, V., Banerjee, A. & Kumar, V. "Anomaly detection: A survey". *ACM Computing Surveys*. 2009; 41 (3) 1–58, https://www.scopus.com/pages/publications/84859722266. DOI: https://doi.org/10.1145/1541880.1541882.

16. Liu, J., Huang, Y., Wu, D., et al. "Multi-channel multi-scale convolution attention variational autoencoder (MCA-VAE): An interpretable anomaly detection algorithm based on variational autoencoder". *Sensors*. 2024; 24 (16): 5316, https://www.scopus.com/pages/publications/85202452472. DOI: https://doi.org/10.3390/s24165316.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship, or other, which could influence the research and its results presented in this article

Received 04.12.2025
Received after revision 11.02.2026
Accepted 18.02.2026

DOI: https://doi.org/10.15276/hait.09.2026.04
УДК 378.147:004.6:331.5

Виявлення аномалій на часових рядах даних для автономної роботи робота

Шаварський Максим Андрійович[1)]
ORCID: https://orcid.org/0000-0002-1379-3244; maksym.a.shavarskyi@lpnu. Scopus Author ID: 58179182100
Кривенчук Юрій Павлович[1)]
ORCID: https://orcid.org/ 0000-0002-2504-5833; Yurii.P.Kryvenchuk@lpnu.ua. Scopus Author ID: 57198358655
[1)] Національний університет «Львівська політехніка», вул. С. Бандери, 12. Львів, 79013, Україна

АНОТАЦІЯ

Автономні роботи, що використовуються в критично важливих для безпеки застосуваннях, потребують систем моніторингу в режимі реального часу для виявлення як екологічних, так і механічних аномалій. Хоча аномалії в даних візуальних датчиків (зміна освітлення, неочікувані перешкоди) часто очевидні, механічні несправності, такі як несправності двигуна або пробуксовка коліс, можуть бути невидимими безпосередньо на зображеннях, що ускладнює їх виявлення. Крім того, реальні робочі дані сильно незбалансовані: нормальна поведінка добре представлена, але аномальні події трапляються рідко. Цей дисбаланс робить традиційні підходи до навчання з учителем неефективними. Щоб вирішити ці проблеми, у цій статті представлена поетапна архітектура мультимодального автоенкодера – нейронної мережі, яка одночасно обробляє як візуальну інформацію (зображення RGB-камери) та сенсорні дані. На відміну від традиційних мультимодальних систем, які навчають усі компоненти спільно та страждають від конкуренції модальностей, запропонована архітектура використовує триетапну навчальну програму, яка навчає візуальні та рухові кодери незалежно перед спільною оптимізацією, запобігаючи градієнтному дисбалансу та забезпечуючи надійні представлення. Система виконує виявлення аномалій за допомогою аналізу помилок реконструкції: менші помилки вказують на нормальні режими роботи, тоді як відхилення сигналізують про потенційні аномалії. Метод вимагає лише нормальних робочих даних для навчання – не потрібні позначені зразки аномалій. Експериментальна перевірка демонструє, що архітектура виявляє візуальні аномалії (спотворення кольорів, неочікувані об'єкти) та аномалії руху (раптові зупинки, ривки, зміни швидкості) у режимі реального часу. Запропонований метод виявляється необхідним для критично важливих для безпеки застосувань, таких як автономна навігація роботів та автоматизація складів, де виявлення механічних та екологічних аномалій є важливим для безпеки експлуатації.

Ключові слова: виявлення аномалій; поетапне навчання; оперування робота; мультимодальне злиття; незбалансовані дані

ABOUT THE AUTHORS

Maksym A. Shavarskyi - Postgraduate student, Department of Artificial Intelligence Systems. Lviv Polytechnic National University, 12, S. Bandera Str. Lviv, 79013, Ukraine
ORCID: https://orcid.org/0000-0002-1379-3244; maksym.a.shavarskyi@lpnu. Scopus Author ID: 58179182100
***Research field*:** Robotics; machine learning; imitation learning; data science

Шаварський Максим Андрійович - аспірант кафедри Системи штучного інтелекту. Національний університет «Львівська політехніка», вул. С. Бандери, 12. Львів, 79013, Україна

Yurii P. Kryvechuk - PhD in Engineering Sciences, Associate Professor, Department of Artificial Intelligence Systems, Deputy Director for Academic Affairs at Institute of Computer Science and Information Technologies. Lviv Polytechnic National University, 12, S. Bandera Str. Lviv, 79013, Ukraine
ORCID: https://orcid.org/0000-0002-2504-5833; Yurii.P.Kryvenchuk@lpnu.ua. Scopus Author ID: 57198358655
***Research field*:** Data science; machine learning

Кривенчук Юрій Павлович - кандидат технічних наук, доцент кафедри Системи штучного інтелекту, заступник директора з навчальної роботи Інституту комп'ютерних наук та інформаційних технологій. Національний університет «Львівська політехніка», вул. С. Бандери, 12. Львів, 79013, Україна