Generation-Based Multi-Modal Anomaly Detection for Nuclear Fusion Target Polishing

Sherif Abdelkarim[1] **Antonios Alexos**[1] **Shashank Galla**[2] **Vikash Sunkara**[2]
Kshitij Bhardwaj[3] **Sean Hayes**[3] **Monika Biener**[3]
Satish Bukkapatnam[2] **Pierre Baldi**[1] **Suhas Bhandarkar**[3]
[1]University of California, Irvine
[2]Texas A&M University
[3]Lawrence Livermore National Lab

Abstract

Real-time detection of anomalous operating states is crucial for manufacturing systems, but prior works often separate visual modality and vibration signals, limiting robustness. We present the first multi-modal anomaly-detection study on a nuclear-fusion target polishing testbed, analyzing uncompressed video alongside vibration waveforms. Framing the task as a generation problem rather than classification, we introduce a reconstruction-based anomaly score that extends the autoencoder paradigm to autoregressive models. We compare two architectures: (i) a spatiotemporal Vision Transformer and (ii) Large Language Models adapted to time-frequency tokens. Both approaches are evaluated on our new Polishing-Fusion-200 benchmark (196 synchronized video-vibration episodes), with ablations on individual modalities and minimal fine-tuning across LLM families. This study introduces an end-to-end pipeline for video-plus-vibration anomaly detection and demonstrates a generation-based scoring strategy that avoids domain-specific heads.

1 Introduction

Inertial Confinement Fusion (ICF) is a key path to clean energy. Diamond (i.e., the target for lasers in ICF) fabrication is a complex process with sub-micrometer tolerances and polishing runs lasting 10–14 hours. Any tiny defects in targets during polishing can seed hydrodynamic instabilities that slash fusion yield. Large-scale fusion experiments represent immense capital investments, and failures or significant delays consequently result in substantial financial losses. Because polishing is stochastic—cracks/pits in targets degrade quality—real-time anomaly detection is essential. Ward et al. [2005]

Traditional uni-modal monitoring approaches face inherent limitations that motivate our multi-modal approach. Video cameras provide rich visual information about surface conditions but suffer from field-of-view constraints, while vibration sensors capture the underlying process dynamics without revealing specific visual defects. These complementary yet incomplete capabilities highlight the critical need for a unified detection framework that leverages both modalities simultaneously.

Despite extensive research in multi-modal anomaly detection across domains like video-audio and image-text analysis, joint video–vibration processing remains largely unexplored in precision manufacturing. We present the first multi-modal pipeline for ICF target polishing using LLMs, framing anomaly detection as generation rather than classification and extending reconstruction-error scoring from autoencoders to auto-regressive models over video frames and vibration signals .

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

Our approach compares spatiotemporal Vision Transformers against LLMs adapted to time-frequency tokens, evaluated on our benchmark: Polishing-Fusion-200, establishing the foundation for automated quality control in fusion energy target production. Our contributions include:

- We introduce the first multi-modal anomaly detection study for nuclear-fusion target polishing, fusing video and vibration data.
- We propose a reconstruction-style anomaly score for auto-regressive models, avoiding the need for domain-specific classification heads.
- We develop and compare two multi-modal architectures: a Vision Transformer, and adapted LLMs for generation-based scoring. LLMs outperform the Vision transformers.
- We demonstrate that our pipeline achieves robust detection of anomalous operating states, paving the way for future multi-modal safety research in nuclear fusion target polishing.

2 Related Work

Recent work explores multi-modal LLMs for time-series anomaly detection. Yang et al. [2025] use MLLMs to *refine* outputs of existing detectors—filtering candidate alarms rather than replacing the detector. They combine visual inspection of anomaly segments with textual context about the data-generating process to distinguish true anomalies from false positives, boosting simple baselines (e.g., k-NN) at lower LLM inference cost and reducing human verification effort.

Xu et al. [2025] introduce *VisualTimeAnomaly*, the first large-scale benchmark for MLLMs on time-series anomaly detection, by rendering numerical series as visuals and evaluating with LLMs. Their study spans univariate, multivariate, and irregular series, with point-, range-, and variate-wise anomalies. MLLMs more effectively detect range- and variate-wise anomalies than point-wise ones and remain robust with up to 25% missing data.

Gu et al. [2024] present *AnomalyGPT*, a few-shot industrial framework that leverages pretrained vision-language models (e.g., CLIP) via anomaly-specific prompts and multi-modal outputs. Across public and proprietary datasets, prompt-based CLIP inference outperforms traditional unsupervised image-encoder methods without extensive fine-tuning, substantially reducing labeling/training overhead while maintaining high accuracy. Despite these advances, none target joint *video* + *vibration* modalities in precision manufacturing.

A closely related body of work focuses specifically on anomaly detection for ICF target polishing. Galla et al. Galla et al. [2024b] pioneered a multi-sensor fusion approach for real-time monitoring, combining video and vibration data. They employed unsupervised learning to identify sensitive spectral bands in vibration signals and used a CNN with Explainable-AI (LIME) to detect subtle shell interaction events. Concurrently, other researchers have explored autoencoders with statistical process control for scalable anomaly detection in batch polishing runs Galla et al. [2024a], as well as statistical methods using Kernel Density Estimation for efficient post-polish screening of surface pit anomalies Canacoo et al. [2025]. While these studies establish the importance of the task, they rely on specialized, task-specific models. In contrast, our work explores the novel application of pre-trained Large Language Models, reframing anomaly detection as a generation task. This approach offers a flexible, multi-modal framework that avoids the need for custom-designed classification heads or spectral band identification, leveraging instead the general-purpose sequence modeling capabilities of LLMs.

3 Methodology

We study two architectures: a transformer-style autoencoder and large language models (LLMs), evaluating multiple LLM families. We frame the anomaly detection task as a generation problem rather than a classification problem for several key reasons. First, generation-based approaches, particularly those leveraging the next-token prediction objective of LLMs, offer inherent flexibility for multi-modal data without requiring complex, custom-designed fusion encoders. Second, this paradigm allows for effective few-shot or minimal fine-tuning of large pre-trained models, significantly reducing training overhead and data requirements. Third, it operates in a one-class setting, requiring only normal data for training, which is more practical and safer for real-world deployment where anomalous events are rare and costly to collect.

3.1 Transformer-Style Fusion Model

We use a **multimodal Vision Transformer (ViT) autoencoder** Vaswani et al. [2017], Dosovitskiy et al. [2020]: each modality is encoded independently, latents are optionally fused in a shared space, and two decoders reconstruct the inputs; reconstruction MSE is the anomaly score.

Video. Temporal order is injected via sinusoidal encodings. A spatial encoder (*FrameViT*) patchifies frames into 16×16 tokens with a `[CLS]` token and L_s layers; per-frame descriptors are aggregated by a temporal encoder (*SequenceEncoder*, depth L_t) over T frames to produce $\mathbf{h}^{(\mathrm{vid})}$.

Vibration. We operate on raw 900-sample windows: a linear projection followed by a temporal Transformer with sinusoidal positions; mean-pooling/normalization yields $\mathbf{h}^{(\mathrm{vib})}$.

Fusion/decoding. When both modalities are present, $\mathbf{h}^{(\mathrm{vid})}$ and $\mathbf{h}^{(\mathrm{vib})}$ are concatenated and passed through a two-layer ReLU MLP to obtain $\mathbf{h}^{(\mathrm{fusion})}$. Decoders are modality-specific (deconvolutional image decoder; MLP vibration decoder), trained with MSE in unimodal or multimodal settings.

3.2 LLMs

We leverage Transformer-based LLMs pre-trained with the next-token objective, $\max_\theta \sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t})$, where x_t is the token at position t and $x_{<t}$ its context. This training yields general sequence models with strong few-shot behavior. We repurpose LLMs for anomaly detection by framing the task as reconstruction-style scoring—analogous to autoencoders An and Cho [2015]—and detail the procedure in §3.3.

For data fusion, we present video frames as image inputs and append compact vibration-statistics text snippets in the same prompt, allowing the model to attend jointly across both modalities via its chat.

We note the asymmetry in vibration data representation between the Transformer and LLM approaches: the former uses raw waveforms while the latter uses statistical summaries. This design choice is intentional. The Vision Transformer is a dedicated architecture capable of processing raw, high-dimensional sequential data, and we aim to evaluate its performance under optimal conditions. Current LLMs, however, are predominantly trained on text and images; directly ingesting long, high-frequency time-series data remains a fundamental challenge. Providing statistical summaries is a pragmatic and effective way to leverage LLMs' reasoning capabilities for temporal data without requiring extensive architectural modification. This comparison thus evaluates each model class in its most native and effective operational regime, highlighting the current trade-offs between specialized and general-purpose models for this task.

3.3 LLM Anomaly Metric

We propose an LLM-based *one-class* detector that—unlike prior binary classification with post-hoc explainability Gu et al. [2024]—is trained only on normal data. After each epoch, we evaluate on a small hold-out split (containing normals and anomalies) by forcing two predictions per sample: a "normal" loss L_{normal} and an "anomalous" loss $L_{\mathrm{anomalous}}$. We define the margin $\mathrm{margin} = L_{\mathrm{anomalous}} - L_{\mathrm{normal}}$, typically large and positive for normals and near zero/negative for anomalies. A decision threshold τ is calibrated as the p-th percentile (typically 95th) of margins computed *on hold-out normals only*; at test time, a sample is flagged anomalous if its margin $\leq \tau$.

This reframes anomaly detection as a generation task that aligns with the LLM's next-token objective, avoiding task-specific classification heads and labels while yielding a simple, model-agnostic score. The margin acts as a surrogate log-likelihood ratio between "normal" and "anomalous" hypotheses, and percentile calibration sets a target false-positive rate on normal data—robust to class imbalance and not tuned on anomalies. This protocol mirrors our transformer autoencoder setting (reconstruction error vs. margin) and supports fair, unified evaluation across modalities and backbones. The full procedure appears in algorithm 1.

Algorithm 1 LLM-Based Anomaly Detection

```
Require: Normal training set $\mathcal{D}_{\mathrm{train}}$, hold-out set $\mathcal{D}_{\mathrm{hold}}$, percentile $p$
Train LLM parameters $\theta$ on $\mathcal{D}_{\mathrm{train}}$
for all $x_i \in \mathcal{D}_{\mathrm{hold}}$ do
    $L_{\mathrm{n}}(x_i) = -\sum_{t=1}^{T_i} \log p_\theta(x_{i,t} \mid x_{i,<t}, \text{“normal”}), \qquad L_{\mathrm{a}}(x_i) = -\sum_{t=1}^{T_i} \log p_\theta(x_{i,t} \mid x_{i,<t}, \text{“anomalous”})$
    $\mathrm{margin}_i \leftarrow L_{\mathrm{a}}(x_i) - L_{\mathrm{n}}(x_i)$
end for
Set threshold $\tau$ to the $p$-th percentile of $\{\mathrm{margin}_i\}$

function DETECT($x$)
    $\mathrm{margin}_i \leftarrow L_{\mathrm{a}}(x_i) - L_{\mathrm{n}}(x_i)$
    if $\mathrm{margin}_x \leq \tau$ then
        return Anomalous
    else
        return Normal
    end if
end function
```

4 Experiments and Results

4.1 Data Collection and Pre-processing

A heterogeneous dataset, called Polishing-Fusion-200, serves as the experimental test bed. The dataset comprises 10 fps video and synchronized vibration telemetry (10 kHz) acquired during spherical shell polishing with a Bühler Echomet 300 system. Synchronization between modalities was achieved via a process-based physical trigger. Six experimental runs, each lasting between 2–5 minutes, were conducted, within which 19 prominent anomalous events (with shell–shell interaction) were identified. The dataset consists of 177 normal videos (with no shell-shell interactions) and 19 anomalous videos. Both modalities are fully synchronized (video frames and vibration segments aligned 1:1).

4.2 Experimental Design

We explore several different configurations:

1. **Video only.** The model receives the sequence of resized frames that visualize the polishing chamber. This is used by both Transformer and LLM approaches.
2. **Raw vibrations (Transformer-only).** The transformer autoencoder operates directly on raw time-series windows (e.g., 900-sample segments). *LLMs do not use raw vibrations.*
3. **Vibration statistics only.** Each one-second segment is converted to a short text fragment that summarizes descriptive statistics of its vibration waveform (mean, standard deviation, root mean square, dominant frequency and energy). This is used by LLMs only.
4. **Video plus raw vibration/statistics.** Transformer uses video and raw vibration data. In case of LLM, the prompt concatenates the visual placeholders with the vibration statistics text snippets so that the LLM can reason jointly over the two information streams.

All variants of each model keep exactly the same architecture and LoRA parameters and optimizer settings with only the input modality changes. Models are trained with the one-class schedule described in Section 3.3 using the normal training samples. Detection thresholds are chosen on the hold-out normal set. Test performance is evaluated based on the margin scores, with performance metrics including accuracy at the optimal threshold.

4.3 Results

Table 1 shows that multi-modal fusion ("Both") consistently outperforms uni-modal inputs across all models (e.g., TRANSFORMER: 92.5% vs. 85.0% for video; QWEN: 95.0% vs. 70.0% for video). This

Table 1: Results on the Polishing-Fusion-200 dataset

Model	Modality	Normal	Anom.	Overall	Precision	Recall	Accuracy
Transformer	Video	20/21	14/19	34/40	93.3%	73.7%	85.0%
	Raw Vibration	20/21	8/19	28/40	88.9%	42.1%	70.0%
	Both	19/21	18/19	37/40	90.0%	**94.7%**	92.5%
Llama	Video	17/21	16/19	33/40	76.2%	84.2%	82.5%
	Vibration Stats	8/21	16/19	24/40	55.2%	84.2%	60.0%
	Both	19/21	16/19	35/40	88.9%	84.2%	87.5%
Qwen	Video	13/21	15/19	28/40	65.2%	78.9%	70.0%
	Vibration Stats	13/21	12/19	25/40	60.0%	63.2%	62.5%
	Both	21/21	17/19	38/40	**100.0%**	89.5%	**95.0%**

confirms the complementary value of video and vibration data. TRANSFORMER+BOTH achieves high recall (18/19 anomalies), while QWEN+BOTH achieves perfect specificity on normals (21/21) and the highest overall accuracy (95.0%), minimizing false positives. Video-only models generally outperform vibration-only models, indicating that our current vibration encoding is too lossy. Fusion optimizes precision-recall trade-offs, favoring more balanced detection. Results are based on a small test set ($n = 40$); future work should evaluate on larger datasets.

4.4 Discussion

Our results demonstrate that multi-modal fusion outperforms single modality, confirming the complementary value of video and vibration data. The strong performance of adapted LLMs shows the next-token prediction objective can effectively function as a reconstruction-based anomaly score. The performance gap between modalities suggests our hand-crafted vibration summaries are suboptimal, indicating future work should develop direct tokenization strategies for raw time-series data as well as explore ensemble methods combining the sensitivity of Vision Transformers to raw signal anomalies with the robust temporal reasoning of LLMs, potentially yielding even stronger performance.

5 Conclusion

We presented a novel LLM-based framework for multi-modal anomaly detection in nuclear fusion target polishing, establishing the Polishing-Fusion-200 benchmark. Crucially, our adapted LLM approach outperformed a specialized transformer autoencoder (95.0% vs. 92.5% accuracy), demonstrating that generative language models can surpass dedicated architectures in complex sensing tasks. This superior performance highlights the potential of leveraging pre-trained LLMs for safety-critical monitoring in energy applications, providing a new paradigm for anomaly detection beyond traditional reconstruction-based methods.

References

Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18, 2015.

Sampson Canacoo, Shashank Galla, Yuhao Zhong, Saikiran Chary Nalband, Sean Michael Hayes, Monika Biener, Suhas Bhandarkar, and Satish T.S. Bukkapatnam. Efficient screening of rare large pit anomalies on polished surfaces using a minimalist sampling scheme. *Journal of Manufacturing Processes*, 147:80–87, 2025. ISSN 1526-6125. doi: https://doi.org/10.1016/j.jmapro.2025.03.115. URL `https://www.sciencedirect.com/science/article/pii/S1526612525003792`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Shashank Galla, Akash Tiwari, Kshitij Bhardwaj, Sean Michael Hayes, Satish Bukkapatnam, and Suhas Bhandarkar. Scalable anomaly detection in batch polishing processes for inertial confinement fusion shells. In *ICML 2024 AI for Science Workshop*, 2024a. URL `https://openreview.net/forum?id=j3gVxXYWDI`.

Shashank Galla, Akash Tiwari, Saikiran Chary Nalband, Sean Michael Hayes, Suhas Bhandarkar, and Satish Bukkapatnam. Detecting anomalous motions in ultraprecision shell-polishing process combining unsupervised spectral-band identification and explainable-ai. *Journal of Manufacturing Systems*, 75:278–287, 2024b. ISSN 0278-6125. doi: https://doi.org/10.1016/j.jmsy.2024.04.004. URL `https://www.sciencedirect.com/science/article/pii/S0278612524000700`.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 1932–1940, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

D.J. Ward, I. Cook, Y. Lechon, and R. Saez. The economic viability of fusion power. *Fusion Engineering and Design*, 75-79:1221–1227, 2005. ISSN 0920-3796. doi: https://doi.org/10.1016/j.fusengdes.2005.06.160. URL `https://www.sciencedirect.com/science/article/pii/S092037960500308X`. Proceedings of the 23rd Symposium of Fusion Technology.

Xiongxiao Xu, Haoran Wang, Yueqing Liang, Philip S Yu, Yue Zhao, and Kai Shu. Can multimodal llms perform time series anomaly detection? *arXiv preprint arXiv:2502.17812*, 2025.

Alan Yang, Yulin Chen, Sean Lee, and Venus Montes. Refining time series anomaly detectors using large language models. *arXiv preprint arXiv:2503.21833*, 2025.