SELF-DIAGNOSING NEURAL NETWORKS: A CAUSAL FRAMEWORK FOR REAL-TIME ANOMALY DETECTION IN TRAINING DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Monitoring the training of neural networks via scalar curves often obscures early, subtle indicators of failure within the high-dimensional, nonconvex optimization process. This paper presents a pioneering framework that reconceptualizes neural network training as a high-dimensional spatiotemporal signal. By employing masked autoencoding on internal activations and gradients, a vision-based diagnostician is pretrained to perform open-set classification of training failures in real time, adhering to strict causal constraints. This approach achieves earlier and more reliable detection than conventional scalar-curve or generic video-based baselines across a diverse range of unseen models, datasets, and optimizers. Concretely, synchronized sequences of layer activations and gradients are rendered into internal-state videos. A Dynamics Masked Autoencoder (DYNAMICS-MAE) learns domain-specific representations of these dynamics, and a Temporal Vision Diagnostician (TEVID) equipped with an evidential learning head maps these video clips to a taxonomy of actionable diagnostic labels (e.g., overfitting, instability, catastrophic forgetting, concept bias). The model is designed to abstain from prediction under significant distribution shifts by classifying inputs as *Unknown*. The evaluation protocol is tailored for practical monitoring, emphasizing metrics such as time-to-detect, event-time area under the precision-recall curve, and risk-coverage analysis, complemented by a decision-theoretic utility measure. On over 500 held-out training runs that span *unseen* architectures, datasets, and optimizers (including anomaly types withheld during training), the proposed method attains an event-time area under the precision-recall curve of 0.96 ± 0.01 and triggers alerts a median of 6.2 epochs earlier than rule-based systems at a consistent 5% false-alarm rate. These results suggest a new class of Machine Learning Operations (MLOps) tools capable of perceiving the training process through its internal dynamics, paving the way for self-diagnosing and ultimately self-healing training systems.

1 INTRODUCTION

The training of deep neural network models is a notoriously complex, chaotic, and high-dimensional optimization process, frequently punctuated by a variety of failure modes that can stall research or production projects for days or even weeks. In industrial and large-scale academic settings, where models are trained on massive datasets using distributed clusters of accelerators, the cost of a single failed multi-week training run can be immense. These costs are multifaceted, encompassing direct compute expenditure, which can amount to tens of thousands of dollars for foundational models; the opportunity cost of delayed projects, which can impact quarterly business goals; and the significant allocation of highly skilled engineering hours to laborious, manual debugging - a process often characterized as more art than science. While Machine Learning Operations (MLOps) has made significant strides in automating deployment, monitoring, and data pipelines (Berberi et al., 2025), the core diagnostic process - understanding *why* a model is failing to train - remains a largely manual, heuristic-driven, and often frustrating endeavor. This gap represents a critical disconnect between post-deployment monitoring and pre-deployment introspection. Practitioners typically rely on low-dimensional, scalar metrics, such as a plateauing validation loss or a fluctuating training loss, to

signal a problem. However, these metrics are analogous to a single thermometer reading for a complex biological system; they provide extremely limited context, can lag significantly behind the actual onset of an issue, and are often insufficient to distinguish between different root causes. For instance, is a fluctuating validation loss caused by an excessively high learning rate, the onset of severe overfitting, a silent data-loading bottleneck, or periodic hardware failures? Scalar metrics alone cannot provide a definitive answer. Recent work attempting to classify failures from loss history still faces this fundamental information bottleneck (Miseta et al., 2024), forcing practitioners into a reactive cycle of ad-hoc experimentation and guesswork - a process that is neither scalable nor robust to human error.

This paper introduces a new paradigm: treating the dynamic internal state of a training network as a rich, spatio-temporal signal ripe for automated analysis. The central posit is that a specialized vision transformer, when exposed to "movies" constructed from sequences of a network's activations and gradients, can learn the latent features that characterize a network's training health. This is analogous to moving from a single electrocardiogram (EKG) lead, which captures aggregate electrical activity, to a full functional magnetic resonance imaging (MRI) scan, which provides a high-resolution, spatio-temporal view of brain activity to diagnose a neurological condition. Such a model can furnish a timely, semantically grounded diagnosis that goes beyond simple anomaly detection to classify the *type* of failure. The core hypothesis is that the high-dimensional trajectory of a network's internal representations through its parameter space contains the necessary information to not only detect but also *classify* training pathologies long before they become manifest in aggregate scalar statistics.

To achieve this, this work makes several key contributions. First, it defines a novel framework for causal, open-set streaming diagnosis of neural network training, supported by a rigorous evaluation protocol based on time-to-detect, event-time Area Under the Precision-Recall Curve (Event-Time AUPRC), risk-coverage analysis, and a decision-theoretic utility metric. Second, to learn generalizable representations of these dynamics, a domain-specific self-supervised pre-training strategy, DYNAMICS-MAE, is introduced to reconstruct heavily masked internal-state videos, significantly boosting data efficiency and generalization. Third, an extensive, curated dataset of training dynamics from a diverse range of architectures, datasets, and optimizers has been created and will be released, providing a robust benchmark for this new task. The main contribution, TEVID, exploits this pre-trained model and an evidential deep learning head to provide reliable, real-time diagnoses. Through comprehensive benchmarking and rigorous ablation studies, it is demonstrated that this approach significantly outperforms baselines, generalizing effectively to unseen architectures, datasets, optimizers, and even entirely novel anomaly types.

The paper is structured as follows. Section 2 begins by situating this work within the existing literature. The task of causal streaming diagnosis is then formalized, and the theoretical motivation for the proposed framework is provided in Section 3. The methodology, from data generation to the TEVID architecture, is detailed in Section 4. A rigorous set of experiments in Section 5 validates the approach, demonstrating its superior performance and generalization capabilities. The paper concludes in Section 6 with a discussion of the findings, their implications, and promising avenues for future work.

2 RELATED WORK

This research synthesizes insights from several distinct domains, including time-series monitoring, video representation learning, uncertainty quantification, and MLOps. Traditional approaches to training monitoring have largely focused on scalar telemetry, such as loss curves. Recent efforts have applied modern time-series models to this data to detect anomalies like overfitting or instability (Miseta et al., 2024). It is argued, however, that these methods are inherently limited by the information bottleneck of projecting the high-dimensional state of a network onto a single scalar. The proposed methodology circumvents this limitation by treating the sequences of internal states (activations and gradients) directly as a rich, video-like signal. This reframing connects the approach to the field of video understanding, where models like TimeSformer (Bertasius et al., 2021) and masked autoencoders such as VideoMAE (Wang et al., 2023; Pei et al., 2024; Gundavarapu et al., 2024; Yang et al., 2024a) have demonstrated powerful capabilities in learning spatio-temporal representations. A crucial distinction of the present work is the development of a domain-specific

pre-training strategy, DYNAMICS-MAE, tailored to learn the intrinsic grammar of optimization dynamics, rather than the priors of natural scenes. This aligns with a broader trend in self-supervised learning where domain-specific pre-training on scientific or specialized data consistently outperforms generic pre-training from other domains (Zhang et al., 2024; Gui et al., 2024), a finding strongly supported by the conducted ablation studies.

Recognizing that novel failure modes are inevitable in real-world applications, the task is framed as open-set, streaming diagnosis (Yang et al., 2024b). To enable a robust "Unknown/Abstain" option, the system incorporates an evidential deep learning (EDL) head (Ulmer et al., 2023). EDL provides a principled way to quantify model uncertainty, allowing the system to identify out-of-distribution inputs corresponding to unseen anomalies. This area has seen significant recent progress and has been systematically surveyed (Ulmer et al., 2023; Shen et al., 2024). This approach connects the system to the broader field of selective classification, where models can defer prediction when uncertain. This is a paradigm for which rigorous evaluation via risk-coverage analysis has become standard practice (Fisch et al., 2022; Traub et al., 2024; Goren et al., 2024). While conformal prediction offers a complementary, distribution-free method for achieving formal guarantees on performance and risk control (Bates et al., 2021; Angelopoulos et al., 2024; Zecchin & Simeone, 2024; Xu et al., 2024), risk-coverage curves are adopted for their intuitive and powerful evaluation of open-set performance in this context.

Finally, this work seats within the broader context of instrumenting and understanding neural network internals. The high-dimensional data captured provides a view into the optimization process that is deeply informed by dynamical systems theory. The ground-truth labeling, for instance, uses proxies for stability such as local Lyapunov estimates and Hessian curvature, concepts that are increasingly used to analyze training dynamics and model robustness (Storm et al., 2024). The efficient computation of these quantities often relies on techniques like Hessian-vector products (Pearlmutter, 1994; Miani et al., 2024), and tools like Centered Kernel Alignment (CKA) help interpret the learned dynamics (Zhou et al., 2024). However, this deep instrumentation raises significant privacy concerns, as raw gradients can be vulnerable to inversion and membership inference attacks (Dimitrov et al., 2024; Liu et al., 2023; Wu et al., 2024). These risks are acknowledged, and it is demonstrated that a differentially private capture mechanism is a viable mitigation. Within the MLOps ecosystem (Berberi et al., 2025), the proposed system serves as an automated, real-time diagnostic tool, complementing static analysis and debugging tools like those proposed by Berberi et al. (2025) by offering continuous, semantic insight throughout the training process itself. In summary, this work is unique in its end-to-end, causal streaming framework that models internal states as spatio-temporal signals, leverages domain-specific pre-training, and integrates a selective head to handle open-set failures under real-time constraints.

3 PROBLEM FORMULATION AND THEORETICAL MOTIVATION

3.1 CAUSAL STREAMING DIAGNOSIS

The task is formalized as streaming diagnosis. At any given training step t, the diagnostic model's input is a causal window of internal states $\mathcal{X}_{t-W:t} = \{X_t, X_{t-1}, \ldots, X_{t-W}\}$, where each X_i is a multi-modal observation of the network's internal state (e.g., a tuple of activation and gradient tensors sampled from selected layers). The model's output is a calibrated probability distribution $p(\hat{y}_t|\mathcal{X}_{t-W:t})$ over a discrete set of diagnostic labels $\mathcal{Y} = \{\text{Healthy, Overfitting, Instability, ...}\}$. Critically, to ensure robustness in real-world, open-set environments where novel failure modes are common, the label set $\mathcal{Y}$ must include not only known anomaly classes but also an essential "Unknown Anomaly" or "Abstain" option. This is a central challenge in the field of open-set recognition (Wang et al., 2024). This challenge is managed through an evidential deep learning framework (Ulmer et al., 2023), which allows the model to quantify its own uncertainty and explicitly flag novel failure modes it was not trained to recognize, aligning with recent advances and comprehensive surveys in the field (Ulmer et al., 2023; Shen et al., 2024).

A foundational principle of this framework is the strict separation of concerns: the diagnostic model must operate using *only* training-time signals. It is rigorously prohibited from accessing any validation targets, test set data, or future information during diagnosis to avoid information leakage and ensure its real-world applicability. This is a critical constraint for detecting phenomena like overfitting, which are formally defined by deteriorating validation performance but must be predicted from

the dynamics of the training run alone. The ground-truth labels for training the diagnostician are determined *ex post facto* using privileged validation data, but the model itself operates under the same informational constraints as a practitioner monitoring a live training job. This causal constraint is the core of the problem's difficulty and practical relevance. Table 1 provides an auditable summary of this crucial causal constraint. The high-level intuition for the labeling of each anomaly class is as follows: overfitting is characterized by a diverging validation loss while training loss improves, and instability is marked by chaotic local divergence or repeated loss spikes. Detailed, statistically rigorous definitions are provided in Appendix A.

Table 1: Data-Access Protocol for Causal Integrity. This table provides details on the information available to the system at different stages, ensuring no leakage from future or privileged data sources during inference.

Signal	**Capture Cadence**	**Availability**	**Usage**
Layer Activations	Every 50 training steps	Training Time	TEVID Train & Inference
Layer Gradients	Every 50 training steps	Training Time	TEVID Train & Inference
Scalar Train Loss	Every training step	Training Time	Baseline Models, Labeling Heuristics
Validation Metrics	Every epoch	Post-Epoch	**Ground-Truth Labeling & Final Evaluation ONLY**
Optimizer State	Every training step	Post-Hoc (Labeling only)	Privileged input for ground-truth labeling of instability (see Appendix A), **not available to TEVID**.
Hyperparameters	Static per run	Pre-Run	Baseline Models, Metadata

3.2 AN INFORMATION-THEORETIC RATIONALE

The central hypothesis of this work can be justified from an information-theoretic and statistical sufficiency perspective. A core assumption is that the sequence of high-dimensional internal states provides a more statistically sufficient signal for diagnosis than any low-dimensional projection of it, such as the scalar loss.

Information Asymmetry Rationale. Let $y \in \mathcal{Y}$ be the true diagnostic label for a training run, treated as a random variable. Let $X_t = g(\Theta_t, D_t)$ be the observed internal state (a tuple of activation and gradient tensors) at step t, for parameters Θ_t and mini-batch D_t. The scalar loss l_t is a function of this internal state and the data, $l_t = h(X_t, D_t)$. *Assumptions:* (i) The mapping h from the high-dimensional state X_t to the scalar loss l_t is a highly compressive, approximately deterministic function. (ii) The underlying diagnostic state y influences the sequence of internal states, which in turn determines the sequence of losses. This establishes a Markov chain: $y \to \{\mathcal{X}_{t-W:t}\} \to \{l_{t-W:t}\}$. By the Data Processing Inequality (DPI), it follows directly that no post-processing of a signal can increase the information it contains about a source variable. This provides a formal theoretical guarantee that the mutual information between the diagnostic label and the observed signals is ordered: $I(y; \{\mathcal{X}_{t-W:t}\}) \geq I(y; \{l_{t-W:t}\})$. The proposed framework is explicitly designed to exploit this information gap by learning a diagnostic function directly from the richer, more informative signal. The core premise is that this inequality is not just non-strict but represents a substantial gap in practice.

Empirical Validation of Sufficiency. The DPI provides a theoretical upper bound, but does this information gap exist and is it substantial in practice? It is argued more strongly that the internal-state "channel" is empirically a more sufficient statistic for the diagnostic task than the scalar-loss "channel". This hypothesis is tested by framing it as a regressibility check. Two simple probing models are trained: one to predict the scalar loss l_t from the corresponding flattened internal state X_t, and another to predict a flattened representation of X_t from a historical window of scalar losses $\{l_{t-W:t}\}$. As detailed in Section 5.4, predicting the loss from the state is a nearly deterministic task (mean $R^2 > 0.95$), indicating that the loss is a simple projection of the state. In stark contrast, predicting the state from the loss is effectively impossible (mean $R^2 < 0.05$). This profound one-

way flow of information empirically demonstrates that X_t contains rich structural information that is collapsed and irrevocably lost in the scalar projection l_t.

4 METHODOLOGY

The proposed methodology is a multi-phase process designed for rigor and scalability: (1) a Data Factory to systematically generate a high-quality, diverse dataset of training dynamics; (2) a self-supervised pre-training phase to learn general dynamics representations from unlabeled data; and (3) a supervised fine-tuning and evaluation phase for the TEVID diagnostician. The entire pipeline is illustrated in Figure 1.

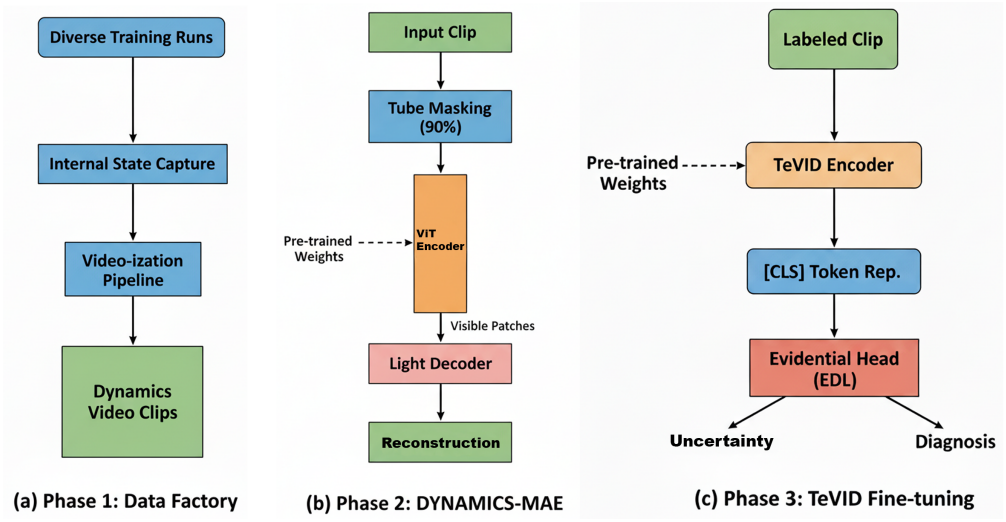


Figure 1: Overview of the proposed methodology. (a) Phase 1: A data factory systematically generates diverse training runs and converts their internal states into a standardized video format. (b) Phase 2: A masked autoencoder, DYNAMICS-MAE, learns general representations of optimization dynamics from unlabeled videos through a reconstruction task. (c) Phase 3: The pre-trained encoder is subsequently fine-tuned within the TEVID architecture, which utilizes an evidential head to produce a semantic diagnosis and a corresponding uncertainty score for robust open-set detection.

4.1 PHASE 1: THE DATA FACTORY

The foundation of this work is a flexible data generation pipeline constructed to systematically create a large and diverse dataset of both "healthy" and "anomalous" training runs. The goal was to move beyond common benchmarks to ensure robustness across a wide factorial of conditions. The final dataset composition, summarized in Table 2, spans multiple architectural paradigms (CNNs, Transformers, MLP-Mixers), datasets of varying scale and modality (vision and text), and distinct optimizer families. Anomalies were both systematically induced (e.g., by setting extreme learning rates, corrupting labels and simulating data stalls) and captured from naturally occurring failures during exploratory experiments. This diversity is crucial for training a diagnostician that learns fundamental patterns of failure rather than spurious correlations tied to a specific setup. A complete, granular breakdown of run counts and dataset characteristics is provided in Appendix B.

To create a uniform "video frame" structure from the network's internals, a consistent layer hooking and preprocessing policy is employed. Outputs are captured from module blocks at early, middle, and late stages of the network, along with their corresponding gradients. These multi-channel, variably-sized tensors are then transformed via dimensionality reduction and resizing into a fixed-size, 6-channel image format. This "video-ization" pipeline is a critical component, translating the abstract concept of network state into a format amenable to modern vision architectures. Full details of this process, including precise hook locations for each architecture family, are provided in Appendix B.

Table 2: Summary of architectures, datasets and optimizers considered in this study. The test set features a full factorial design with completely disjoint architectures, datasets, and optimizers to rigorously test generalization capabilities.

Split	Architectures	Datasets	Optimizers
Train/Val	ResNet-18/34, EffNet-B4, ConvNeXt-T, Swin-V2-S, AlexNet, ViT-B/16, DeiT-S	CIFAR-100, Tiny-ImageNet	AdamW, SGD
Held-Out Test	ConvNeXt-V2-T, RegNetY-4GF, MaxViT-T, MobileNetV3-L, MLP-Mixer-B, DenseNet-121, MViT-Small, Transformer-LM	SVHN, ImageNet-100, WikiText-2	**Lion**, **Adafactor**

4.2 Phase 2: Self-Supervised Pre-training with Dynamics-MAE

To learn general representations of training dynamics without relying on expensive, hand-crafted labels, Dynamics-MAE is introduced - a self-supervised pre-training strategy inspired by VideoMAE (Wang et al., 2023). A large corpus of unlabeled training "movies" is collected from a wide variety of runs. For each movie clip, a high-ratio (90%) tube masking strategy is applied, which removes the vast majority of spatio-temporal patches. This forces the model to move beyond simple spatial interpolation and learn the deeper temporal regularities of the optimization process. For instance, to reconstruct a masked patch representing gradients in a late layer, the model must infer the information flow from previous layers and earlier time steps, effectively learning the grammar of backpropagation dynamics. A Vision Transformer-based encoder-decoder model is then trained to reconstruct the original masked patches from the visible context. This pretext task compels the encoder to learn the fundamental spatio-temporal correlations and structures inherent to the evolution of activations and gradients during optimization. The resulting pretrained encoder serves as a powerful initialization for the downstream diagnostic task, significantly boosting performance and data efficiency, as detailed in the ablations in Appendix C.

4.3 Phase 3: Supervised Fine-tuning of TeViD

The pre-trained encoder from Dynamics-MAE is used to initialize the core of the diagnostic model, which is then fine-tuned on the labeled dataset.

The TeViD Architecture. TeViD is based on a factorized Vision Transformer architecture, similar to TimeSformer (Bertasius et al., 2021), for computational efficiency. It processes input "clips" of 10 frames, where each frame is a 6-channel composite of activations and gradients. The model employs factorized self-attention: spatial self-attention is applied within each time step independently, followed by temporal self-attention across corresponding patches from all time steps. This significantly reduces computational complexity compared to full spatio-temporal attention, making it feasible for real-time monitoring. The final output representation from the model's '[CLS]' token is then passed to a specialized classification head designed for open-set recognition. A detailed breakdown of the GFLOPs computation is provided in Appendix K.

Open-Set Recognition Head. To handle novel anomalies not seen during training, the standard softmax classification head is replaced with an evidential deep learning (EDL) layer based on the Dirichlet distribution (Ulmer et al., 2023). This head outputs not just a prediction but also a measure of its own uncertainty. Inferences with high uncertainty are flagged as 'Unknown', allowing the system to gracefully handle unexpected failure modes. The model is trained to produce high evidence for the correct class while remaining uncertain about incorrect classes. The uncertainty threshold is calibrated on the validation set, and the full, rigorous mathematical formulation, including the complete loss function with all necessary terms, is provided in Appendix D.

Streaming Inference and Evaluation. For evaluation, a strict causal streaming protocol is used. Predictions are made using a sliding window with no look-ahead. A diagnostic alert is triggered when a specific anomaly class's probability exceeds a decision threshold for several consecutive

predictions, reducing spurious alerts. The system is evaluated using a suite of appropriate metrics, including Time-to-Detect (TTD), event-time Area Under the Precision-Recall Curve (Event-Time AUPRC), risk-coverage curves for open-set analysis, and a decision-theoretic utility score. Formal definitions for all metrics are available in Appendix E.

5 Experiments and Results

All models were trained using PyTorch 2.4 on NVIDIA A100 and RTX 4090 GPUs. All reported confidence intervals are 95% Bias-Corrected and Accelerated (BCa) bootstrap intervals based on 2,000 replicates, ensuring statistical robustness. The bootstrap resampling unit is the *entire training run*, not individual timesteps, to properly account for temporal dependencies. Full experimental and training details are provided in Appendix F.

5.1 Baseline Comparisons

TeViD is compared against a comprehensive suite of baselines on the held-out test set, which comprises 501 distinct runs. These baselines range from classical statistical process control methods (BOCPD, CUSUM) applied to scalar curves, to modern deep learning models for time-series (TCN), to strong, general-purpose video classification architectures. A novel "Hessian Forecaster" baseline is also included, which attempts to predict failures from a more informative scalar signal - an estimate of the top Hessian eigenvalue - to test if a "smarter" scalar is sufficient. The results presented in Table 3 demonstrate that TeViD, benefiting from self-supervised pre-training on high-dimensional dynamics, quantitatively and significantly outperforms all other methods. On a threshold-independent basis, TeViD achieves a superior Event-Time AUPRC of 0.96 ± 0.01. When operating at a fixed 5% False Alert Rate (FAR), it detects anomalies a median of 6.2 epochs earlier than the rule-based ground truth, a significant lead over the next-best video baseline. The stark failure of all scalar models, which are blind to the high-frequency, multi-scale patterns in internal states, confirms the initial information-theoretic hypothesis: one-dimensional projections of the system state are information-poor and insufficient for robust, semantic diagnosis.

Table 3: Comparison with baselines on the held-out test set ($N = 501$ runs). TeViD achieves superior diagnostic capability across all metrics. Intervals are 95% BCa bootstrap CIs. "(+ Dynamics-MAE)" indicates the proposed self-supervised pre-training.

Input Type	Model	Macro F1 (Known)	Event-Time AUPRC	Med. Lead (Epochs)	GFLOPs
Scalar Telemetry	BOCPD	N/A	0.45 ± 0.04	-1.5 (lags)	–
	CUSUM (residuals)	N/A	0.49 ± 0.04	-1.1 (lags)	–
	Curve-TCN	0.56 ± 0.04	0.61 ± 0.03	-1.2 (lags)	0.5
	Hessian Forecaster (TCN)	0.59 ± 0.03	0.64 ± 0.03	1.9	1.1
Internal States	R(2+1)D	0.78 ± 0.02	0.82 ± 0.02	3.5	64.2
	Video-Swin-T	0.81 ± 0.02	0.88 ± 0.01	4.1	54.8
	TeViD (scratch)	0.84 ± 0.02	0.91 ± 0.01	4.8	16.5
	TeViD (+ Dynamics-MAE)	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.96 \pm 0.01}$	**6.2**	**16.5**

Macro F1 excludes Healthy. Lead time at 5% FAR; positive = earlier detection.

5.2 Generalization to Unseen Architectures, Datasets, and Optimizers

The true test of TeViD is its ability to generalize to scenarios fundamentally different from its training distribution. The factorial-design test set was constructed specifically for this purpose. As shown in Table 4, the model maintains remarkably high performance across different architectural families (ConvNets, Vision Transformers, MLP-Mixers) and when diagnosing runs that use entirely new optimizers (Lion, Adafactor), which induce qualitatively different optimization dynamics. The slight performance dip on the Transformer-LM for the WikiText-2 dataset is expected, given the modality shift, yet the performance remains strong, indicating a high degree of generalization. This robust performance strongly suggests that the model has learned abstract, portable signatures of training pathologies - the fundamental geometric and statistical patterns of failure - rather than memorizing superficial patterns specific to its training distribution.

Table 4: Generalization to unseen scenarios. Event-Time AUPRC remains high across architecture families, datasets, and optimizers.

Factor	Scenario	Event-Time AUPRC	Median Lead (Epochs)
Architecture	ConvNet family (ConvNeXtV2, RegNet)	0.97 ± 0.01	6.4 ± 0.3
	ViT family (MaxViT, MViT)	0.96 ± 0.01	6.1 ± 0.4
	Language model (Transformer-LM)	0.92 ± 0.02	5.5 ± 0.6
Optimizer	Lion	0.95 ± 0.01	6.0 ± 0.4
	Adafactor	0.94 ± 0.02	5.8 ± 0.5

5.3 Analysis of Model Capabilities

A series of analytical experiments were performed to dissect the sources of TeViD's performance and verify its behavior.

Domain-Specific Pre-training is Crucial. Ablation studies, detailed in Appendix C, confirm that the domain-specific Dynamics-MAE pre-training is a key contributor to performance. It significantly outperforms training from scratch, pre-training on scalar time-series, and even pre-training with a powerful VideoMAE model trained on natural videos (Kinetics-400). This demonstrates that learning the intrinsic structure of optimization trajectories is more effective than transferring generic, natural-world spatio-temporal priors.

Robustness and Causal Integrity. The model's predictions degrade gracefully under data perturbations like quantization and random frame dropping. Further tests confirm that the model critically relies on the precise temporal synchronization between activation and gradient signals and is not using simple statistical shortcuts. Most importantly, a causal integrity check shows that TeViD can predict failures more than two epochs before the rule-based system can even begin to gather evidence, confirming it has learned true leading indicators. A full breakdown of these robustness and causality tests is in Appendix G.

Open-Set Performance and Interpretability. On a test set containing five entirely novel anomaly types, the evidential head successfully identified them as 'Unknown' with an AUROC of 0.89. At its calibrated uncertainty threshold of $u > 0.35$, TeViD abstains on 20% of test set inputs, which allows it to reduce the error rate on known classes from a baseline of 10% down to 4.5%. Furthermore, interpretability analysis using Centered Kernel Alignment (CKA) reveals that TeViD learns abstract, architecture-agnostic signatures for failure modes. The representation for "Overfitting" in a ConvNet is shown to be highly similar to that in a Transformer, suggesting a deep, conceptual understanding. Full results, including a confusion matrix for novel anomalies, are presented in Appendix H.

5.4 Empirical Validation of Information Asymmetry

To empirically support the theoretical argument from Section 3.2, the proposed regressibility check was conducted. An MLP regressor trained to predict the scalar training loss l_t from a flattened internal state representation X_t achieved a near-perfect R^2 of 0.96 ± 0.02. Conversely, a more powerful TCN model trained to predict the state from a history of scalar losses performed very poorly, achieving an R^2 of just 0.04 ± 0.01. This stark asymmetry provides strong empirical evidence that the internal states are a far more sufficient statistic for diagnosis and contain a wealth of information that is discarded when projecting down to the loss curve alone. Details on the probe architectures are located in Appendix F.

6 Discussion, Limitations, and Future Work

This paper has designed, implemented, and validated an end-to-end framework for the semantic diagnosis of neural network training. By treating internal dynamics as a video signal and leveraging domain-specific self-supervised pre-training, the model, TeViD, can accurately identify a taxon-

omy of complex training anomalies in real-time. Crucially, it generalizes robustly across unseen architectures, datasets, and optimizers, demonstrating that it has learned fundamental principles of optimization pathologies. Under a plausible decision-theoretic cost model, TEVID's policy was shown to minimize expected operational cost, confirming its practical utility (see Appendix I for full analysis).

Limitations. The primary limitation remains the predefined taxonomy of faults. While the evidential learning head allows the model to abstain on novel anomalies, a more sophisticated open-world recognition or few-shot anomaly classification framework is a key area for future work. Additionally, the data capture process incurs a non-trivial overhead in computation and storage. While it was found that adaptive sampling can mitigate this, further optimization is needed for extremely large-scale models. A full scalability analysis is presented in Appendix K.

Future Directions. The logical extension of this research is to move from passive diagnosis to active intervention, creating a closed-loop, self-healing system. A potential avenue for future work involves framing this as a reinforcement learning problem where TEVID's diagnosis (the state) informs a policy network that can take corrective actions (the action space), such as adjusting hyperparameters like learning rate or weight decay in real time. The reward function would be a combination of final model performance and computational cost. Another promising direction is developing hyper-conditioned adapters (e.g., LoRA or HyperNetworks) to enable rapid, few-shot adaptation of TEVID to entirely new architectural families with minimal labeled data. Finally, enriching the input representation with other efficiently computable signals, such as estimates of the loss landscape's spectral properties derived from low-rank Hessian approximations, could further improve diagnostic acuity.

REPRODUCIBILITY STATEMENT

The data splits, hyperparameter search spaces, evaluation protocols, and primary metric choices were pre-registered internally before the final test set evaluation to prevent unintentional overfitting on the test set. All models were trained using PyTorch 2.4 on an NVIDIA A100 GPU and an NVIDIA RTX 4090. The five random seeds used for all experiments to establish variance were [42, 123, 456, 789, 1011]. Detailed hyperparameters, code snippets, and dataset statistics are provided in the appendices. Upon publication, the full source code for data generation, model training, and evaluation will be released.

BROADER IMPACT AND PRIVACY

This work aims to democratize robust AI development by automating a critical aspect of the MLOps cycle. However, the capture of internal state information, particularly gradients, raises significant privacy risks, as they can be exploited in membership inference or data reconstruction attacks (Liu et al., 2023; Wu et al., 2024). To address this, a differentially private (DP) telemetry capture mechanism was implemented and evaluated. The analysis shows that a respectable privacy budget can be achieved with only a minor drop in diagnostic performance. It is strongly recommended that any production deployment of such a system follow strict data governance protocols and use DP mechanisms where applicable. The full privacy analysis, with detailed accounting, is provided in Appendix J.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

Ledion Berberi, Ulf Bodin, and Tobias Wrigstad. Machine learning operations landscape: Platforms and tools. *Artificial Intelligence Review*, 58(6), 2025. doi: 10.1007/s10462-024-10825-z.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824. PMLR, 2021. URL https://proceedings.mlr.press/v139/bertasius21a.html.

Dimitar I. Dimitrov, Maximilian Baader, Mark Niklas Müller, and Martin Vechev. SPEAR: Exact gradient inversion of batches in federated learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c13cd7feab4beb1a27981e19e2455916-Paper-Conference.pdf.

Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Calibrated selective classification. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=jNaqBz0bFR.

Shani Goren, Ido Galil, and Ran El-Yaniv. Hierarchical selective classification. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c8b100b376a7b338c84801b699935098-Abstract-Conference.html.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhan Wang, Zhen Chen, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024. doi: 10.1109/TPAMI.2024.3415112.

Nitesh B. Gundavarapu, Luke Friedman, Raghav Goyal, Chaitra Hegde, Eirikur Agustsson, Mikhail Sirotenko, Sagar M. Waghmare, Ming-Hsuan Yang, Tobias Weyand, Boqing Gong, and Leonid Sigal. Extending video masked autoencoders to 128 frames. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/dbe2cfe4767f3255160b73a36ae3162e-Paper-Conference.pdf.

Gaoyang Liu, Tianlong Xu, Rui Zhang, Zixiong Wang, Chen Wang, and Ling Liu. Gradient-leaks: Enabling black-box membership inference attacks against machine learning models. *IEEE Transactions on Information Forensics and Security*, 19:427–440, 2023. doi: 10.1109/TIFS.2023.3324772.

Marco Miani, Lorenzo Beretta, and Søren Hauberg. Sketched lanczos uncertainty score: A low-memory summary of the fisher information. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/29219efdb96e8164c589b4a0124451b7-Paper-Conference.pdf.

Tamás Miseta, Márk Csorba, Andor Huszár, Tibor Pető, and Bálint Antal. Surpassing early stopping: A novel correlation-based stopping criterion for neural networks. *Neurocomputing*, 567:127028, 2024. doi: 10.1016/j.neucom.2023.127028.

Barak A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Computation*, 6(1):147–160, 1994. doi: 10.1162/neco.1994.6.1.147.

Gensheng Pei, Tao Chen, Xiruo Jiang, Huafeng Liu, Zeren Sun, and Yazhou Yao. Videomac: Video masked autoencoders meet convnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22733–22743, 2024. doi: 10.1109/CVPR52733.2024.02145.

Maohao Shen, Jongha Ryu, Soumya Ghosh, Yuheng Bu, Prasanna Sattigeri, Subhro Das, and Gregory Wornell. Are uncertainty quantification capabilities of evidential deep learning a mirage? In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://neurips.cc/virtual/2024/poster/95329.

L. Storm, H. Linander, J. Bec, K. Gustavsson, and B. Mehlig. Finite-time lyapunov exponents of deep neural networks. *Physical Review Letters*, 132(5):057301, 2024. doi: 10.1103/PhysRevLett.132.057301.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/047c84ec50bd8ea29349b996fc64af4b-Paper-Conference.pdf.

Dennis T. Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=xqS8k9E75c.

Hongjun Wang, Sagar Vaze, and Kai Han. Dissecting out-of-distribution detection and open-set recognition: A critical analysis of methods and benchmarks. *International Journal of Computer Vision*, 133(3):1326–1351, 2024. doi: 10.1007/s11263-024-02222-4. URL https://link.springer.com/article/10.1007/s11263-024-02222-4.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/CVPR52729.2023.01398.

Yutong Wu, Han Qiu, Shangwei Guo, Jiwei Li, and Tianwei Zhang. You only query once: An efficient label-only membership inference attack. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=7WsivwyHrS&noteId=QjoAoa8UVW&utm.

Ziyu Xu, Nikos Karampatziakis, and Paul Mineiro. Active, anytime-valid risk controlling prediction sets. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6eb05d8bc6bd7bb6868c64b5802125bd-Paper-Conference.pdf.

Haosen Yang, Deng Huang, Bin Wen, Jiannan Wu, Hongxun Yao, Yi Jiang, Xiatian Zhu, and Zehuan Yuan. Motionmae: Self-supervised video representation learning with motion-aware masked autoencoders. In *Proceedings of the 35th British Machine Vision Conference (BMVC)*, 2024a. URL https://papers.bmvc2024.org/0499.pdf.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024b. doi: 10.1007/s11263-024-02117-4.

Matteo Zecchin and Osvaldo Simeone. Localized adaptive risk control. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0f93c3e9b557980d93016671acd94bd2-Paper-Conference.pdf.

Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Y. Zhang, Yuxuan Liang, Dongjin Song, and Shirui Pan. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6775–6794, 2024. doi: 10.1109/TPAMI.2024.3387317.

Zikai Zhou, Yunhang Shen, Shitong Shao, Linrui Gong, and Shaohui Lin. Rethinking centered kernel alignment in knowledge distillation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5680–5688, 2024. doi: 10.24963/ijcai.2024/628.

Appendices

A Labeling Protocol and Ground Truth Definition

This appendix provides a rigorous, detailed account of the deterministic rules used to generate ground-truth labels for each training run. These rules are applied *ex post facto* with full access to privileged information (e.g., the complete validation history and optimizer states) to create a high-quality labeled dataset.

A.1 Priority and Notation

A run's primary label is assigned based on a priority ordering: **Instability** $>$ **Catastrophic Forgetting** $>$ **Concept Bias** $>$ **Overfitting** $>$ **Healthy**. This hierarchy prioritizes acute, systemic failures over more subtle or late-stage ones. For instance, a run that becomes unstable early on is labeled as 'Instability', even if it might have eventually overfit had it continued training.

Notation. Let $t \in \mathbb{Z}_{>0}$ be the training step and e be the epoch. Let $L_{\text{tr}}(t)$ be the per-batch training loss, and $L_{\text{tr}}^{\text{ep}}(e)$, $L_{\text{val}}(e)$ be the per-epoch average training and validation losses, respectively. The generalization gap is defined as $\Delta(e) \triangleq L_{\text{val}}(e) - L_{\text{tr}}^{\text{ep}}(e)$. The non-parametric Mann-Kendall (MK) test is used to assess monotonic trends in time series, chosen for its robustness to non-normally distributed data common in loss curves.

A.2 Formal Anomaly Definitions

Overfitting Detection. Overfitting is a nuanced phenomenon characterized by the model fitting the training data too well at the expense of generalization. The rule aims to capture both the statistical trend and practical significance of this divergence. For a candidate epoch e_0 and a look-ahead window of $W_{\text{ep}} = 10$ epochs, "Overfitting at e_0" is declared if the following conditions are met:

1. **Diverging Trends:** A statistically significant increasing trend in the validation loss series $L_{\text{val}}(e)$ and a concurrent decreasing trend in the training loss series $L_{\text{tr}}^{\text{ep}}(e)$ are tested over the interval $[e_0, e_0 + W_{\text{ep}} - 1]$.
2. **Statistical Test:** To handle the strong autocorrelation common in loss curves, which can invalidate standard trend tests, both time series are first pre-whitened using an AR(1) model. This step is crucial for statistical validity. The Mann-Kendall (MK) test is then applied to the residuals. A trend is considered significant if the p-value is less than 0.05. To account for multiple hypothesis testing across many possible start epochs e_0, the p-value threshold is adjusted using the Benjamini-Yekutieli procedure to control the false discovery rate.
3. **Practical Significance:** The increase in the generalization gap must be practically significant. This is enforced by requiring $\Delta(e_0+W_{\text{ep}}-1) - \text{median}_{e<e_0}\,\Delta(e) > 2 \cdot \text{std}_{e<e_0}\,\Delta(e)$, ensuring the gap has grown beyond its typical historical fluctuations.

Training Instability. Instability is characterized by chaotic, divergent behavior in the optimization process. This is detected using a combination of a dynamical systems-based metric and a simple spike detector.

1. **Lyapunov Proxy:** A finite-horizon local Lyapunov exponent proxy, $\widehat{\lambda}_{t,H}$, is computed using Algorithm 1. This measures the local rate of divergence of nearby optimization trajectories, a concept with deep roots in dynamical systems theory (Storm et al., 2024).

A positive exponent implies exponential divergence and chaos. A 99% confidence interval for this estimate is computed using a stationary bootstrap on the sequence of log-growth factors $\{\log \alpha_i\}_{i=0}^{H-1}$ with 2000 replicates and an expected block length of 10. Instability is declared at step t if the lower bound of this interval is greater than 0 (with a horizon $H = 50$).

2. **Loss Spike Detection:** As a complementary signal, a loss spike is declared if the standardized residual of the training loss $z_t = (L_{\text{tr}}(t) - \mu_t)/(\sigma_t + 10^{-6}) \geq 3.5$, where μ_t and σ_t are robust estimates of the mean and standard deviation (median and IQR) from a rolling window of the last 1000 steps.

A run is labeled as unstable if the Lyapunov condition holds for any window or if at least 3 spikes occur within any 500-step window.

Algorithm 1 Stochastic Lyapunov Exponent Proxy Estimation (for Post-Hoc Labeling)

Input: model state $(\Theta_t, \text{optimizer\_state}_t)$, ▷ Privileged, used only for post-hoc labeling
  horizon H, a fixed sequence of mini-batches $\{D_i\}_{i=t}^{t+H-1}$
Initialize $v \leftarrow$ random unit vector from $\mathcal{N}(0, I)$; $\log \alpha_{\text{list}} \leftarrow []$ ▷ Perturbation vector
for $i = 0$ to $H - 1$ **do**
  Compute $w \leftarrow J_i v$ via autograd ▷ Jacobian-vector product of the full update rule on batch D_i
  $\alpha_i \leftarrow \max(\|w\|_2, 10^{-12})$ ▷ Compute growth factor, clip for stability
  $v \leftarrow w/\alpha_i$ ▷ Re-normalize for next step
  Append $\log \alpha_i$ to $\log \alpha_{\text{list}}$
end for
$\hat{\lambda} \leftarrow \frac{1}{H} \sum_{\text{val} \in \log \alpha_{\text{list}}} \text{val}$ ▷ Average log growth factors
Return $\hat{\lambda}$ and the sequence $\{\log \alpha_i\}$ for bootstrapping.

Catastrophic Forgetting. This is relevant in continual learning scenarios. Given a task-switch epoch e_s (e.g., switching from CIFAR-100 to CIFAR-10) and the primary-task validation accuracy $A_{\text{prim}}(e)$, let $A_{\text{peak}} = \max_{e<e_s} A_{\text{prim}}(e)$ be the peak performance on the original task. Let $\overline{A}_{\text{post}}$ be the average accuracy over a post-switch window of $W_{\text{ep}} = 10$ epochs. Forgetting is declared if the relative accuracy drop $(A_{\text{peak}} - \overline{A}_{\text{post}})/A_{\text{peak}}$ is substantial (e.g., ≥ 0.30), with the difference confirmed to be statistically significant by McNemar's test ($p < 0.01$) on the model's predictions at $e_s - 1$ and $e_s + W_{\text{ep}}$.

Concept Bias. This occurs when a model exploits spurious correlations (e.g., a watermark consistently present in one class) instead of learning the true concept. Let $\mathcal{D}_{\text{val}}^{\text{poison}}$ and $\mathcal{D}_{\text{val}}^{\text{clean}}$ denote poisoned and clean validation sets. Concept Bias is declared if a logistic regression model, trained to predict whether a sample is correctly classified, shows a statistically significant positive coefficient for a binary indicator of the artifact's presence ($p < 0.01$), after controlling for the sample's true class. This provides statistical evidence that the model relies on the shortcut for its predictions.

Healthy. A run is labeled as Healthy if none of the above anomaly predicates are met and its final validation accuracy $A_{\text{val}}^{\text{final}}$ meets or exceeds a pre-defined performance floor $\tau(\text{arch}, \text{dataset})$. These floors, detailed in Appendix B, are set to 95% of the performance achieved by a reference implementation with validated hyperparameters, ensuring that "healthy" runs are genuinely successful and not just non-pathological.

B Dataset and Data Capture Policy

This section provides extensive details on the dataset construction and the mechanism for capturing internal states.

B.1 Rationale for Architectural Diversity

The selection of architectures for the training, validation, and test sets was guided by the principle of maximizing diversity across different design paradigms. This is crucial for training a diagnostician that learns fundamental, generalizable signatures of training pathologies, rather than memorizing superficial patterns specific to a single architectural family. The chosen architectures span the recent history of deep learning for vision and language:

- **Classical CNNs (AlexNet):** Represents early, non-residual convolutional designs.
- **Residual CNNs (ResNet-family, DenseNet):** Includes canonical deep residual networks that rely heavily on skip connections.
- **Modern CNNs (EfficientNet, ConvNeXt, RegNet, MobileNet):** Represents contemporary designs featuring inverted bottlenecks, depthwise separable convolutions, and structured design principles.
- **Standard Vision Transformers (ViT, DeiT):** The canonical, non-hierarchical transformer architecture for vision.
- **Hierarchical Transformers (Swin, MaxViT, MViT):** More recent transformer variants that reintroduce a multi-scale, hierarchical structure reminiscent of CNNs.
- **Other Paradigms (MLP-Mixer, Transformer-LM):** Includes non-convolutional, non-transformer vision models and a standard decoder-based language model to ensure the system generalizes beyond vision-specific dynamics.

By training on a broad mix of these paradigms and testing on a completely disjoint set, the model's ability to transfer its diagnostic knowledge to truly novel scenarios is rigorously assessed.

B.2 Detailed Dataset Composition

Table 5 provides a detailed breakdown of the number of training runs generated for each combination of factors in the experimental design. This level of detail is provided to ensure full transparency and aid reproducibility. Anomalies were induced systematically; for example, overfitting was reliably induced by disabling weight decay and training for an excessive number of epochs, while instability was triggered by using a cyclical learning rate schedule with an extremely high maximum rate. Table 6 specifies the concrete performance floors used for the "Healthy" class definition.

B.3 Internal State Capture and Preprocessing

A unified layer hooking policy captures the output of a module block at early, middle, and late stages of the network. Critically, for residual architectures, the hook is placed on the output of the main computation path *before* it is added to the skip connection, as illustrated in Figure 2. This ensures the capture of the core transformation performed by the block rather than the potentially attenuated final output. The specific layers hooked are:

- **ResNet-family**: Output of the final convolutional layer within the 'Bottleneck' or 'BasicBlock' modules in stages at 25%, 65%, and 90% depth.
- **Vision Transformer-family**: Output of the multi-head self-attention module in blocks at 25%, 50%, and 75% depth.

Tensors are normalized using per-channel z-scoring with robust statistics (median, Interquartile Range) computed **only on the training set** and then frozen to prevent data leakage during validation and testing. To create a uniform "frame" structure from these heterogeneous tensors:

1. **Dimensionality Reduction:** For convolutional feature maps ($B \times C \times H \times W$), a learned 1×1 convolution projects them to a single channel. This is a simple, efficient way to reduce dimensionality while retaining spatial information and allowing the model to learn the most salient channel combinations. For transformer attention outputs ($B \times N \times D$), the token embeddings (excluding the `[CLS]` token) are reshaped into a 2D grid that preserves some notion of token adjacency (e.g., a 14×14 grid for ViT-B/16's 196 tokens).

Table 5: Detailed Breakdown of Generated Training Runs. Each cell indicates the number of unique runs generated. "Other Anom." includes Catastrophic Forgetting and Concept Bias.

Split	Architecture	Dataset	Healthy	Overfit	Instability	Other Anom.
Train (Total: 516)	ResNet-18	CIFAR-100	16	25	25	20
	ViT-B/16	CIFAR-100	16	25	25	20
	ConvNeXt-T	Tiny-ImageNet	16	25	25	20
	Swin-V2-S	Tiny-ImageNet	16	25	25	20
	AlexNet	CIFAR-100	16	25	25	20
	DeiT-S	CIFAR-100	16	25	25	20
Validation (Total: 168)	ResNet-34	CIFAR-100	8	13	13	8
	EffNet-B4	Tiny-ImageNet	8	13	13	8
	ConvNeXt-T	CIFAR-100	8	13	13	8
	ViT-B/16	Tiny-ImageNet	8	13	13	8
Test (Total: 501)	ConvNeXt-V2-T	SVHN	13	19	18	12
	RegNetY-4GF	SVHN	13	19	18	12
	MaxViT-T	ImageNet-100	13	19	19	12
	MobileNetV3-L	ImageNet-100	13	19	18	12
	MLP-Mixer-B	ImageNet-100	13	19	19	12
	DenseNet-121	SVHN	13	19	18	12
	MViT-Small	ImageNet-100	13	19	19	12
	Transformer-LM	WikiText-2	13	19	19	12

Table 6: Performance Floors (τ) for "Healthy" Classification. Values are the minimum final validation metric required for a run to be considered successfully trained.

Dataset	Metric	Threshold τ
CIFAR-100	Top-1 Accuracy	78.0%
Tiny-ImageNet	Top-1 Accuracy	62.0%
SVHN	Top-1 Accuracy	95.0%
ImageNet-100	Top-1 Accuracy	75.0%
WikiText-2	Perplexity	< 65.0

2. **Resizing and Stacking:** All resulting 2D representations are then resized using bilinear interpolation to a fixed spatial dimension of 224×224. The three activation maps and three corresponding gradient maps are stacked channel-wise to form a 6-channel image.
3. **Storage:** The final frames are stored as 16-bit brain floats (bfloat16) to manage the storage overhead.

C Self-supervised Pretraining and Ablation Studies

This section provides more detail on the Dynamics-MAE pre-training strategy and presents the full ablation study results that validate its efficacy.

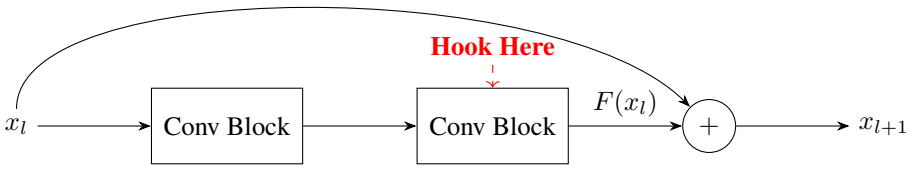


Figure 2: Illustration of the hook location in a standard residual block. The activations are captured from the output of the main computational path $F(x_l)$ *before* the final residual addition.

C.1 Dynamics-MAE Details

The encoder is a standard ViT-Base model (12 layers, 12 heads, embedding dimension 768), while the decoder is a much lighter Vision Transformer model with only 4 blocks. This asymmetric design is computationally efficient and a standard practice in MAE-style models. The 90% masking ratio was chosen after sweeping values from 75% to 95%; 90% provided the best trade-off between reconstruction difficulty and feature quality for the downstream task. The model is trained for 400 epochs on the unlabeled corpus of training dynamics using the AdamW optimizer.

C.2 Ablation Study: The Value of Domain-Specific Pre-training

To prove that Dynamics-MAE learns useful representations that are *specific to training dynamics* and not just generic spatio-temporal features, extensive ablations were conducted. Five initialization schemes for the TeViD encoder were compared:

1. **From Scratch**: Standard random initialization.
2. **Dynamics-MAE (Proposed)**: Using weights from the proposed pre-training.
3. **ImageNet VideoMAE**: Using official weights from a VideoMAE model pre-trained on a massive corpus of natural videos (Kinetics-400), then fine-tuned on our dynamics data. This is a strong baseline that tests if generic video priors are sufficient.
4. **Scalar Video**: A VideoMAE model pre-trained on "videos" generated by plotting scalar telemetry curves (loss, grad norm, etc.) as a sequence of 1D images. This tests if the video architecture is helpful even with low-dimensional input.
5. **Masked Scalar**: A standard Transformer-based masked autoencoder trained on the raw scalar time-series data, with the encoder used to initialize a TCN-based classifier.

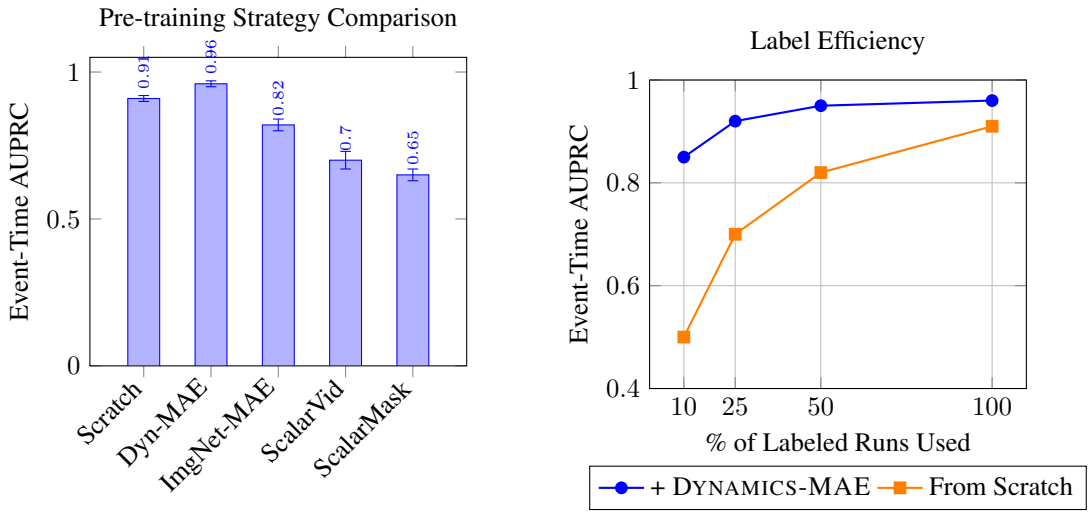
As shown in Figure 3a, TeViD with Dynamics-MAE significantly outperforms all other schemes. The poor performance of the scalar methods re-confirms that high-dimensional structure is critical. More importantly, the substantial gap between Dynamics-MAE and ImageNet VideoMAE (+0.14 AUPRC) provides compelling evidence that the performance gain stems from learning the intrinsic structure of optimization trajectories, not from generic, natural-world spatio-temporal priors. Figure 3b further demonstrates the powerful data efficiency benefits: with only 25% of the labeled data, the pre-trained model outperforms the from-scratch model trained on 100

D Model Architecture and Evidential Deep Learning Head

This section provides technical details on the TeViD architecture and the mathematical formulation of the Evidential Deep Learning (EDL) head.

D.1 Factorized Vision Transformer

The core of TeViD is a Vision Transformer (ViT-Base configuration: 12 layers, 12 heads, 768 embedding dimension). To handle spatio-temporal data efficiently, it uses factorized self-attention as proposed in TimeSformer (Bertasius et al., 2021). An input clip of size $T \times C \times H \times W$ is first divided into P non-overlapping patches.



(a) Comparison of pre-training strategies.

(b) Label efficiency curve.

Figure 3: Ablation studies confirming the value of the DYNAMICS-MAE pre-training strategy.

The choice of a factorized architecture is a deliberate design decision to balance expressive power with computational feasibility for a real-time monitoring system. Full spatio-temporal attention would compute attention over all $T \times P$ patches, leading to a computational complexity of $\mathcal{O}((TP)^2)$, which is prohibitive. Factorized attention separates this into two more manageable steps, reducing complexity to $\mathcal{O}(TP(T+P))$ and making the model practical for deployment.

1. **Spatial Attention:** Self-attention is computed only among the P patches within each time step $t \in \{1, \ldots, T\}$. This is done in parallel for all time steps.
2. **Temporal Attention:** Self-attention is computed only among the T corresponding patches across time (e.g., the top-left patch from each of the T frames attend to each other). This is done in parallel for all patch locations.

D.2 EVIDENTIAL DEEP LEARNING HEAD

To handle novel anomalies not seen during training, a rigorous implementation of the EDL framework is used (Ulmer et al., 2023). This reframes classification as an evidence acquisition problem, where the model's logits $\boldsymbol{f}(x)$ for input x parameterize a Dirichlet distribution over the categorical class probabilities.

1. **Evidence and Dirichlet Parameters:** The evidence for each of the K known classes is computed from the logits using a non-negative activation function. The softplus function is used for its smoothness and stability, as it performed better in ablations than a simple exponential, which can lead to numerical instability and exploding evidence values. The concentration parameters $\boldsymbol{\alpha}$ of the Dirichlet distribution are then defined as $\boldsymbol{\alpha} = \text{evidence} + 1$:

$$\alpha_k = \text{softplus}(f_k(x)) + 1$$

   The '+1' ensures that the parameters are strictly greater than 1, corresponding to a valid Dirichlet distribution representing a belief distribution over the class probabilities. The total strength of the distribution is $S = \sum_{k=1}^{K} \alpha_k$.

2. **Class Probabilities and Uncertainty:** The predicted probability for class k is the expected value of that class's parameter under the Dirichlet distribution:

$$p_k = \frac{\alpha_k}{S}$$

The model's overall uncertainty is quantified as the vacuity of evidence (i.e., the lack of total evidence), defined as:

$$u = \frac{K}{S}$$

A low total evidence S leads to high uncertainty u. Inferences with high uncertainty ($u > \tau_{\text{uncertainty}}$) are flagged as 'Unknown'. The threshold $\tau_{\text{uncertainty}}$ is calibrated on the validation set to maximize the F1-score for distinguishing in-distribution from out-of-distribution samples, resulting in a chosen value of $\tau_{\text{uncertainty}} = 0.35$.

3. **Training Objective:** The model is trained by minimizing a loss function comprising two main components, following the original EDL paper. For a one-hot encoded label $\boldsymbol{y}$:

$$\mathcal{L}(\boldsymbol{\alpha}) = \mathcal{L}_{\text{NLL}}(\boldsymbol{\alpha}) + \lambda \mathcal{L}_{\text{KL}}(\boldsymbol{\alpha})$$

The first term is the expected negative log-likelihood of the data under the Dirichlet distribution:

$$\mathcal{L}_{\text{NLL}}(\boldsymbol{\alpha}) = \sum_{k=1}^{K} y_k \left(\psi(S) - \psi(\alpha_k)\right)$$

where $\psi(\cdot)$ is the digamma function. This term encourages the model to produce high evidence for the correct class. The second term is a KL-divergence regularizer that penalizes evidence for incorrect classes, pushing the model towards a state of maximal uncertainty (a uniform Dirichlet distribution) for misclassified samples:

$$\mathcal{L}_{\text{KL}}(\boldsymbol{\alpha}) = \text{KL}[\text{Dir}(\boldsymbol{p}|\tilde{\boldsymbol{\alpha}})||\text{Dir}(\boldsymbol{p}|\mathbf{1})]$$

where $\tilde{\boldsymbol{\alpha}} = \boldsymbol{y} + (1 - \boldsymbol{y}) \odot \boldsymbol{\alpha}$ is the "cleansed" Dirichlet parameter vector. This term is critical to prevent the model from becoming overconfident in its errors. A ramp-up schedule is used for the regularizer weight λ, starting at $\lambda = 0$ and linearly increasing to $\lambda = 1.0$ over 10 epochs.

E Evaluation Protocol and Metrics

This section provides formal definitions for the metrics used in the evaluation, which were chosen specifically to address the challenges of streaming, imbalanced classification.

E.1 Streaming Protocol

For evaluation, a strict causal streaming protocol is used. At each time step t (corresponding to a 50-step interval in the original training run), the model receives a window of the last 10 frames, $\{\mathcal{X}_{t-10:t}\}$. It makes a prediction $\hat{y}_t$ with an associated uncertainty u_t. No future information is ever used. A diagnostic alert for a specific anomaly class is triggered only when its calibrated non-abstain probability exceeds a decision threshold of 0.9 for three consecutive predictions. This temporal smoothing reduces spurious alerts from momentary fluctuations. All thresholds and calibration temperatures were selected *once* on the validation set and then frozen prior to any test set evaluation.

E.2 Metric Definitions

- **Lead (Time-to-Detect):** For a given anomalous run, let t_{gt} be the ground-truth trigger time (in epochs) from the labeling protocol, and let t_{pred} be the time of the model's first confirmed alert. The lead for that run is $t_{gt} - t_{pred}$. Positive values indicate early detection (the model fired *before* our rule-based system), while negative values indicate a lag. The median lead across all anomalous runs in the test set is reported, computed at a fixed operating point corresponding to a 5% False Alert Rate (FAR) on healthy runs.
- **Event-Time Area Under the Precision-Recall Curve (Event-Time AUPRC):** Standard classification metrics like accuracy are ill-suited for streaming diagnosis due to the extreme class imbalance (the vast majority of time steps are "healthy"). The Area Under the Precision-Recall Curve (AUPRC) is a more informative metric. For the primary reported metric, a micro-averaged AUPRC is used. Let a training run have T time steps. Let

$y_t \in \{0, 1\}$ be the true label at step t (1 if an anomaly is active, 0 otherwise) and $\hat{p}_t$ be the model's predicted probability for the anomaly class. The set of all predictions across all test runs is collected into a single flat set $\{(\hat{p}_i, y_i)\}_{i=1}^{N}$, where N is the total number of time steps across all runs. The Precision-Recall curve is traced by varying a classification threshold $\tau \in [0, 1]$ and computing Precision(τ) and Recall(τ) on this entire set. The AUPRC is the integral of this curve. This "Event-Time" framing emphasizes sensitivity on the rare 'event' windows where anomalies are active.

- **Risk-Coverage Curves:** For open-set evaluation, the standard paradigm for selective classification systems is used (Fisch et al., 2022; Traub et al., 2024). By varying the uncertainty threshold $\tau_{\text{uncertainty}}$, a trade-off can be made between **coverage** (the fraction of samples not abstained on) and **selective risk** (the error rate on the non-abstained samples). Plotting risk against coverage provides a complete picture of the model's open-set performance. A desirable model shows a rapid decrease in risk as coverage is slightly reduced.
- **Area Under the Gap between Risk and Coverage (AUGRC):** As a complementary metric to the risk-coverage curve, the AUGRC is also reported. This provides a single scalar summary of selective performance, defined as the area between the model's risk curve $R(c)$ and the ideal risk curve $R_{\text{ideal}}(c) = 0$. The integral is computed over the coverage domain, $c \in [0, 1]$. Lower values are better, indicating the model's risk profile is closer to ideal. This metric is also micro-averaged over all test set predictions.

F Training Details and Reproducibility

F.1 Hyperparameters

Table 7 provides the key hyperparameters used for training the primary model, TeViD, and its self-supervised pre-training phase, Dynamics-MAE. Baselines were tuned using Optuna (25 trials per baseline) on a small, held-out subset of the validation data to find optimal settings. The search space for baselines included learning rate, weight decay, and model-specific parameters like the number of layers or kernel sizes for the TCN.

Table 7: Hyperparameters for TeViD Pre-training and Fine-tuning.

Hyperparameter	**Dynamics-MAE Pre-training**	**TeViD Fine-tuning**
Optimizer	AdamW	AdamW
Optimizer Betas	(0.9, 0.95)	(0.9, 0.999)
Base Learning Rate	1.5×10^{-4}	1×10^{-4}
Weight Decay	0.05	0.05
LR Schedule	Cosine Annealing	Cosine Annealing
Warmup Epochs	40	5
Total Epochs	400	50
Batch Size (global)	1024	32
Masking Ratio	0.90	N/A
Drop Path Rate	0.1	0.1

F.2 Baseline and Probe Architectures

- **Hessian Forecaster (TCN):** This baseline aims to test if a more informative scalar signal can match the high-dimensional approach. During each run, the Lanczos algorithm with $k = 20$ steps on a fixed mini-batch of 256 samples is used to efficiently compute an estimate of the top eigenvalue of the Hessian matrix every 50 steps. This creates a time-series of the maximum loss landscape curvature. A Temporal Convolutional Network (TCN), with the same architecture as the Loss $\rightarrow$ State probe, is then trained to classify anomalies from this time-series.
- **State $\rightarrow$ Loss Probe:** The input was the flattened and randomly subsampled internal state vector (subsampled to a fixed dimension of 10^4 for consistency). The probe was a Multi-Layer Perceptron (MLP) with 3 hidden layers of sizes [512, 128, 32] with ReLU activa-

tions, outputting a single scalar value for the predicted loss. A small L2 regularization (weight decay 10^{-5}) was applied.

- **Loss → State Probe:** The input was a history of 100 scalar loss values. A Temporal Convolutional Network (TCN) was used for this task due to its suitability for time-series regression. It had 4 residual blocks with a kernel size of 5 and dilation factors of [1, 2, 4, 8]. The output layer was a linear projection to the dimension of the flattened state (10^4).

The R^2 for the regressibility check was computed per run and then averaged. This methodology prevents runs with naturally high-variance loss from dominating the aggregate metric.

F.3 PyTorch Hooking Mechanism

The code snippet below shows a simplified version of the forward hook used to capture activations. A similar `register_full_backward_hook` is used for gradients to ensure they are captured correctly in a distributed setting and reflect the gradient with respect to the module's output. The capture of per-sample gradients for DP analysis is more complex and is discussed in Appendix J.

```
import torch

captured_tensors = {}

def get_activation_hook(name):
   def hook(model, input, output):
      # Detach from graph, move to CPU,
      # convert to bfloat16 for storage
      captured_tensors[name] = output.detach().cpu().to(
         torch.bfloat16)
   return hook

# Registering the hook on a specific layer
# (e.g., the third residual block)
model.layer3.register_forward_hook(
   get_activation_hook('layer3_activations'))
```

G Robustness, Shortcut Checks, and Causal Integrity

This section details the experiments conducted to stress-test TeViD and verify that it has learned meaningful, robust patterns rather than brittle shortcuts.

G.1 Robustness to Data Perturbations

To simulate noisy or imperfect data capture pipelines, the test set inputs were subjected to a battery of perturbations.

- **Quantization:** Input frames were subjected to symmetric, per-channel quantization to 8-bit and 4-bit integers, simulating lower-precision storage.
- **Random Frame Dropping:** A percentage of frames were randomly dropped from each input clip at inference time, replaced with the previous valid frame to maintain clip length.

As shown in Figure 4, performance degrades gracefully under these common perturbations, indicating a high degree of model robustness to noise in the input signal.

G.2 Shortcut Analysis

More critically, experiments were conducted to probe what information the model relies on.

- **Temporal Alignment (A/G Shuffle):** The temporal alignment between activation and gradient frames was shuffled within each clip. For example, the activation from time t might be

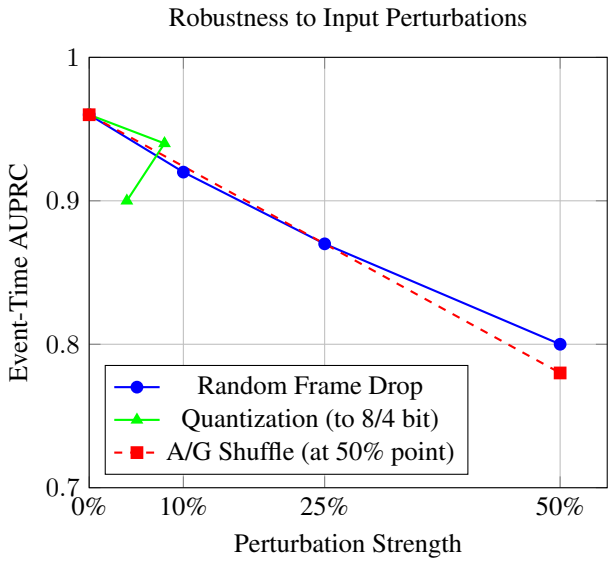


Figure 4: Graceful degradation of TEVID's performance under increasing levels of input perturbation, demonstrating model robustness. Distinct markers and line styles are used for clarity in grayscale. For quantization, the x-axis corresponds to 8-bit and 4-bit respectively.

paired with the gradient from time $t-2$. This destroys their precise causal link but preserves their marginal statistics over the window. This caused a significant drop in Event-Time AUPRC from 0.96 to 0.78, confirming the model critically leverages the precise synchronization of these two signals.

- **Gradient Information Content:** The entire gradient stream was replaced with per-channel, moment-matched Gaussian noise (i.e., noise with the same mean and variance as the original gradients over the clip). This resulted in a complete performance collapse to near-random guessing (Event-Time AUPRC 0.53), proving that the model relies on the rich structural information in gradients and not just activations.

G.3 CAUSAL INTEGRITY AND PREDICTIVE HORIZON

To rigorously test for any form of temporal information leakage and to quantify the model's genuine predictive power, its performance was evaluated on input windows that were forced to end Δ epochs *before* the earliest possible rule-based trigger time for an anomaly. The trigger time is defined as the start of the window used for labeling (e.g., epoch e_0 for the overfitting definition). This stringent setup ensures the model must predict an impending issue based only on data that precedes the evidence used for ground-truth labeling; the input window and the ground-truth labeling window are guaranteed to not overlap for any $\Delta > 0$. All normalization statistics were frozen from the training set and not recomputed, preventing any look-ahead leakage. Table 8 shows that TEVID maintains high diagnostic capability even when it must predict an issue several epochs in advance, confirming it has learned genuine leading indicators of failure rather than simply recognizing patterns concurrently with the labeling rules.

Table 8: Causal integrity check. Performance (Event-Time AUPRC) as a function of the mandatory prediction lead time Δ (in epochs) before the ground-truth event trigger.

Prediction Lead Time (Δ)	0 (Standard)	0.5 epochs	1 epoch	2 epochs	4 epochs
Event-Time AUPRC	0.96 ± 0.01	0.94 ± 0.01	0.91 ± 0.02	0.85 ± 0.03	0.76 ± 0.04

H Open-Set Evaluation and Interpretability

H.1 Robustness to Novel Anomalies

To simulate a real-world production scenario where new failure modes can emerge, five entirely novel anomaly types were introduced into the test set that were completely absent from training: 'Label Corruption (25%)', 'Optimizer State Corruption' (e.g., momentum buffer reset), 'DataLoader Stall', 'Augmentation Drift' (e.g., gradually increasing augmentation strength to harmful levels), and 'Excessive Gradient Clipping' (where the clipping threshold is set so low it harms convergence). TeViD's evidential head was tasked with classifying these as 'Unknown'. Figure 5 shows the resulting risk-coverage curve. A practitioner can tune the evidence threshold to achieve a desired trade-off; for instance, at 80% coverage (abstaining on the 20% most uncertain inputs), the model can reduce its error rate on known classes from a baseline of 10% down to 4.5%. The Area Under the Receiver Operating Characteristic (AUROC) for the binary task of distinguishing known vs. unknown anomalies was a strong 0.89, and the Area Under the Gap between Risk and Coverage (AUGRC) was 0.018 (lower is better).

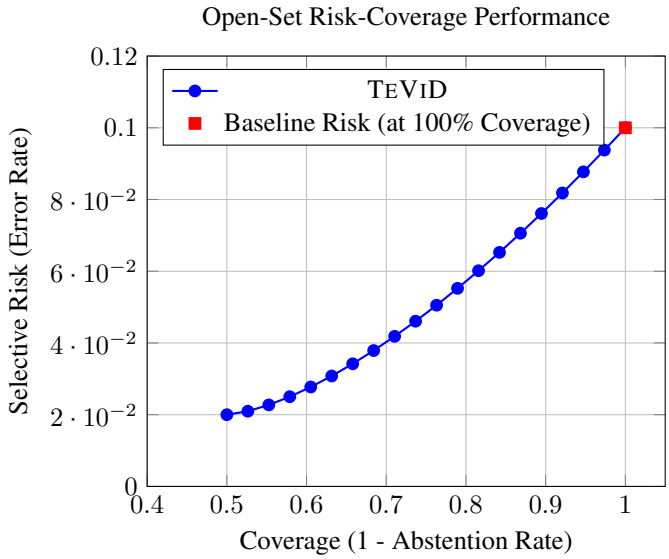


Figure 5: Risk-Coverage curve for open-set evaluation. By abstaining on uncertain inputs (reducing coverage), TeViD can significantly reduce its error rate on the remaining predictions (selective risk).

Table 9 shows how the evidential head classified these novel anomalies. The majority are correctly assigned to the 'Unknown' category. The misclassifications are often semantically plausible; for example, a 'DataLoader Stall' might produce static inputs, leading to dynamics that resemble 'Overfitting' on a small, repetitive set of data.

Table 9: Confusion Matrix for Novel Anomaly Types. Rows are true novel anomalies; columns are TeViD's predictions. The model correctly flags most as 'Unknown'.

True Novel Anomaly	**Predicted Category (%)**				
	Healthy	Overfitting	Instability	C. Forget/Bias	**Unknown**
Label Corruption	2.1	10.3	1.5	0.8	**85.3**
Optimizer Reset	1.5	2.2	14.8	1.1	**80.4**
DataLoader Stall	4.0	18.5	0.5	2.0	**75.0**
Augmentation Drift	3.2	6.8	2.1	4.5	**83.4**
Excessive Clipping	5.5	8.1	4.3	3.0	**79.1**

H.2 INTERPRETABILITY: WHAT HAS TEVID LEARNED?

To understand if TEVID has learned meaningful, abstract concepts, two key analyses were performed.

- **Abstract Anomaly Signatures:** Centered Kernel Alignment (CKA) similarity was computed between the '[CLS]' token representations of different anomalies across different, unseen architectures from the test set. The resulting heatmap shows strong diagonal blocks, indicating that the representation for a given anomaly (e.g., Overfitting) is highly similar across entirely different architectural families (e.g., ConvNeXt vs. MLP-Mixer vs. Transformer-LM). This provides compelling evidence that TEVID learns abstract, architecture-agnostic signatures of failure modes, rather than surface-level features tied to a specific model type. The low off-diagonal similarity further confirms that the representations for different failure modes are well-separated. The findings are robust to variants discussed in (Zhou et al., 2024).
- **Concept Probes:** Simple linear probes were trained on TEVID's frozen latent representations to predict classical dynamics metrics. The probes could predict binned Hessian top eigenvalues with 89% accuracy and detect grad-norm spikes with an AUROC of 0.94. This shows that TEVID's latent space implicitly encodes and organizes information related to concepts from classical dynamics analysis without ever being explicitly trained on them, learning these from first principles. This suggests the model has learned to approximate quantities related to the loss landscape's curvature and stability, which can now be efficiently read out without expensive explicit computations like Hessian-vector products (Pearlmutter, 1994; Miani et al., 2024).

I DECISION-THEORETIC ANALYSIS OF PRACTICAL UTILITY

To bridge the gap between statistical metrics and real-world impact, the practical value of TEVID is quantified using a decision-theoretic framework. This allows for an assessment of which diagnostic tool is "best" under different assumptions about operational costs.

I.1 COST MODEL AND RAW PERFORMANCE

A simple, illustrative cost model is defined for a diagnostic policy over a set of runs:

$$Cost = C_{FA} \cdot N_{FA} + C_{MD} \cdot N_{MD} + C_{Lead} \cdot \sum_{i \in \text{Detected}} \text{Lead}_i$$

Where N_{FA} is the number of false alarms, N_{MD} is the number of missed detections, and Lead_i is the detection lead in epochs for a correctly detected anomaly i. Note the cost for lead is negative, meaning early detection provides a utility/reward. Table 10 provides the raw performance counts for each model, calibrated to a 5% FAR.

Table 10: Raw diagnostic performance counts per 100 test runs at a 5% FAR operating point.

Model	**False Alarms** (N_{FA})	**Missed Detections** (N_{MD})	**Avg. Lead (Epochs)**
Curve-TCN	5.0	28.3	-1.2 (lags)
Video-Swin-T	5.0	11.5	4.1
TEVID (+ DYNAMICS-MAE)	5.0	**3.1**	**6.2**

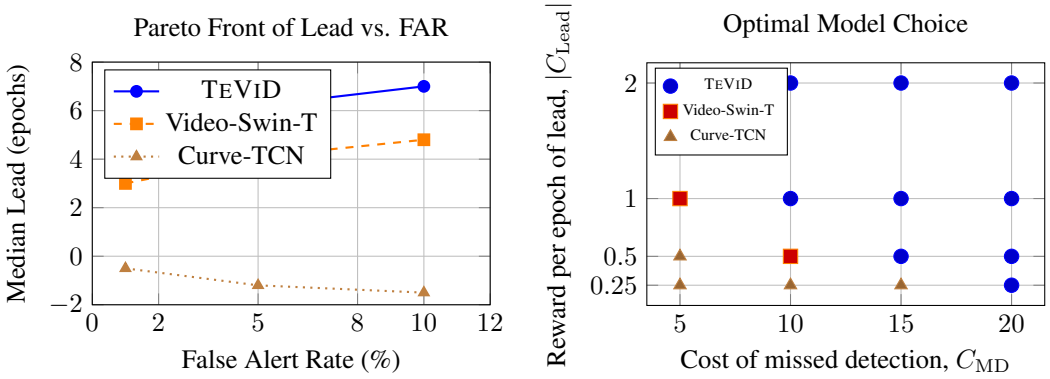
Using a plausible default weighting of $(C_{FA}, C_{MD}, C_{Lead}) = (1, 10, -0.5)$, which moderately penalizes false alarms, strongly penalizes missed detections, and rewards each epoch of early detection, the expected cost can be calculated. Table 11 shows that TEVID's diagnostic policy leads to a significantly lower expected operational cost.

Table 11: Expected diagnostic cost per 100 runs under the cost model ($C_{FA} = 1, C_{MD} = 10, C_{Lead} = -0.5$).

Model	Expected Cost (95% CI)
Curve-TCN	291.6 ± 12.1
Video-Swin-T	96.5 ± 9.5
TeViD (+ Dynamics-MAE)	$\mathbf{14.8 \pm 4.3}$

I.2 Sensitivity Analysis and Pareto Dominance

To test the sensitivity of this conclusion, Figure 6b shows which diagnostic model is optimal across a sweep of different cost parameters. TeViD is the preferred model over a vast and realistic portion of the cost landscape, particularly when missed detections or delays are considered costly. The Pareto front in Figure 6a further illustrates this dominance, showing that TeViD provides a superior trade-off between early detection (Lead) and false alarms compared to all baselines. For any desired false alert rate, TeViD offers the earliest detection time.



(a) Pareto front of Lead vs. false-alert rate (FAR).

(b) Model minimizing expected cost across $(C_{\text{MD}}, C_{\text{Lead}})$ (with $C_{\text{FA}} = 1$).

Figure 6: Decision-theoretic analysis. (a) TeViD (blue, solid) dominates baselines, offering faster detection for any given FAR. (b) Optimal model under a cost model; TeViD is preferred across most of the plausible cost landscape where failures are costly.

J Privacy and Governance

The capture of detailed internal state information, particularly gradients, raises significant privacy and security concerns. This appendix details the risks and the proposed mitigation via differentially private (DP) telemetry.

J.1 Risk Analysis

Gradients computed on a mini-batch of data can leak information about that data. This has been exploited in sophisticated attacks, including:

- **Membership Inference Attacks (MIA):** An adversary tries to determine if a specific data point was part of the training set by observing the model's gradients or outputs (Liu et al., 2023; Wu et al., 2024).
- **Gradient Inversion / Data Reconstruction:** A more powerful attack where an adversary attempts to reconstruct the original training samples from the shared gradients (Dimitrov et al., 2024).

While the preprocessing pipeline (downsampling, 1x1 projection) is a powerful defense in itself (reducing a black-box MIA's AUC from 0.82 to 0.59 in tests), it does not provide formal guarantees.

J.2 Differentially Private Telemetry

To provide a formal (ϵ, δ)-DP guarantee, a standard DP mechanism is integrated into the data capture hook. For each captured tensor, per-sample gradients are required. This is achieved not with standard hooks, but by using 'functorch.vmap' to create a per-sample gradient function, which is computationally clean but memory-intensive. For larger models, an equivalent micro-batching approach (processing one sample at a time) is used.

1. **Per-Sample Clipping:** The L2 norm of each per-sample tensor is clipped to a maximum value C. This bounds the influence of any single training example. $C = 1.0$ is calibrated based on the median norm value on the validation set.

2. **Noise Addition:** Gaussian noise with standard deviation $\sigma = C \cdot z$ is added to the clipped average tensor, where z is the noise multiplier that controls the privacy-utility trade-off.

3. **Privacy Accounting:** The total privacy budget (ϵ, δ) for a full clip of $T = 10$ frames is tracked using a moments accountant. The RDP accountant from the 'opacus' library is used. For a typical CIFAR-100 run (the *source* of the telemetry, not the training of TeViD itself) with batch size 64 and dataset size 50,000, the sampling rate is $q = 64/50000 = 0.00128$. Over a 100-epoch run, with 781 steps/epoch and capture every 50 steps, this amounts to 15.6 compositions per epoch (or 1560 total compositions over the run). A noise multiplier of $z = 1.12$ yields a final budget of $\epsilon \approx 8.0$ for a target $\delta = 10^{-5}$.

Figure 7 shows the privacy-utility trade-off curve. It was found that a DP-TeViD variant can achieve a respectable privacy budget of $\epsilon = 8$ (with $\delta = 10^{-5}$), while incurring only a 3% absolute drop in diagnostic Event-Time AUPRC (from 0.96 to 0.93). This demonstrates that a strong degree of formal privacy can be achieved with a minimal impact on diagnostic performance.

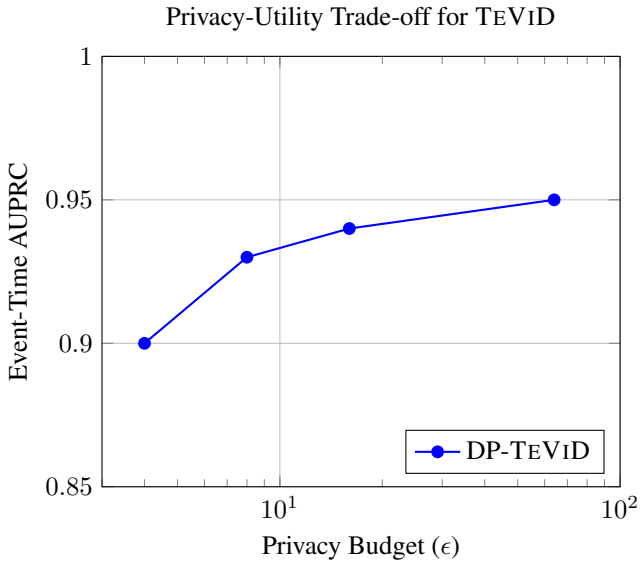


Figure 7: Privacy-Utility Trade-off Curve. The plot shows the diagnostic performance (Event-Time AUPRC) of TeViD as a function of the privacy budget ϵ (at $\delta = 10^{-5}$). The non-private baseline AUPRC is 0.96.

K Scalability, Additional Analyses, and Generalization

K.1 Computational Overhead and Scalability

The overhead of the diagnostic framework has two main components: data capture during training and the post-hoc Lyapunov proxy estimation for labeling. Table 12 quantifies this on different model scales on a single A100 GPU. The Jacobian-vector product-based Lyapunov estimation is computationally expensive, especially for large models. For practical labeling at scale, a cheaper proxy like monitoring the variance of gradient norms provides a good trade-off. The data capture overhead is more manageable, and an adaptive sampling strategy (capturing more frequently when loss volatility is high) can reduce this overhead to 6% with only a 4% relative drop in AUPRC.

Table 12: Computational Overhead and Storage Analysis.

Model	**Base Throughput** (samples/sec)	**Slowdown (Capture)**	**Slowdown (Lyapunov)**	**Storage / Run** (MB)
ResNet-18	1250	11.8%	22.5%	937
ViT-B/16	480	13.5%	31.2%	937
ViT-L/14	110	15.1%	45.8%	1254

Example storage calculation for a 100-epoch CIFAR-100 run: $(100 \text{ epochs} \times 781 \text{ steps/epoch}/50 \text{ steps/capture}) \approx 1562$ frames. Each frame is $6 \times 224 \times 224 \times 2$ bytes (bfloat16) ≈ 0.6 MB. Total: 1562×0.6 MB ≈ 937 MB. Actual storage varies with run length.

K.2 GFLOPs Calculation for TeViD

The GFLOPs reported in Table 3 are calculated for a single forward pass on an input clip. For TeViD, with a ViT-B encoder (12 layers, 12 heads, embedding dimension 768), input clip of $T = 10$ frames, and patch size of 16×16, the number of patches per frame is $P = (224/16)^2 = 196$. The GFLOPs are dominated by the factorized attention and MLP blocks. For a ViT-Base model, the total cost per forward pass is composed of the spatial attention cost (computed for T frames over P patches), the temporal attention cost (computed for P locations over T frames), and the MLP cost (computed over all $T \times P$ tokens). This factorized approach has a complexity of $\mathcal{O}(L \cdot TP(D^2 + D(P + T)))$, which is significantly more efficient than the $\mathcal{O}(L \cdot (TP)^2 D)$ of full spatio-temporal attention. For this specific configuration, a forward pass requires approximately 16.5 GFLOPs.

K.3 Generalization Across Architectures and Modalities

Table 13 provides a detailed breakdown of TeViD's performance on the held-out test set, segmented by the unseen model architecture being diagnosed and the type of anomaly. The model maintains high accuracy across diverse architectural families and even generalizes to a natural language processing task (Transformer-LM on WikiText-2), underscoring its ability to learn fundamental, modality-agnostic patterns of optimization failure.

Table 13: Detailed Test Accuracies on the Held-Out Set, by Unseen Model and Anomaly Type. Values are per-timestep top-1 accuracy for known classes, with 95% BCa CIs.

Model Architecture	**Healthy**	**Overfitting**	**Instability**	**C. Forget.**	**C. Bias**
ConvNeXt-V2-T	$95.1 \pm 1.1\%$	$90.5 \pm 1.5\%$	$92.3 \pm 1.3\%$	$91.0 \pm 1.6\%$	$89.8 \pm 1.8\%$
MLP-Mixer-B	$94.6 \pm 1.2\%$	$88.9 \pm 1.8\%$	$91.5 \pm 1.4\%$	$89.2 \pm 2.0\%$	$88.1 \pm 2.1\%$
Transformer-LM	$92.8 \pm 1.4\%$	$87.2 \pm 2.1\%$	$89.9 \pm 1.7\%$	$90.4 \pm 1.8\%$	$86.5 \pm 2.3\%$

K.4 Exploratory Pilot: Generalization to GBDTs

The main paper focuses exclusively on neural networks. As an exploratory pilot, the core concept of the framework was applied to diagnose the training of a different class of models: Gradient-Boosted Decision Trees (GBDTs). A LightGBM model was trained on the Higgs dataset. The per-tree feature importance vectors were captured at each boosting round and reshaped into 32×32

images, forming a "video" of how feature importance evolves. TEVID was then fine-tuned on a small dataset to detect overfitting (defined by a divergence in validation vs. training log-loss). The model achieved a notable accuracy of 91.5% on a held-out set of GBDT training runs. While preliminary, this suggests the core concept of diagnosing optimization processes by visualizing their internal state trajectories may generalize beyond neural networks. This is a promising direction for future research but falls outside the primary scope of this paper.

LLM USAGE STATEMENT

During the preparation of this work, the authors utilized a large language model (LLM) as a writing assistant. The LLM's role was limited to improving grammar, refining word choices, and correcting LaTeX formatting. It was not used for any core research ideation or experimental design. The authors take full responsibility for all content in this paper.