

Article

Not peer-reviewed version

Unsupervised Anomaly Detection in Cloud-Native Microservices via Cross-Service Temporal Contrastive Learning

[Zizhao Zhang](#), Weicheng Liu, Jingzhou Tao, [Haotian Zhu](#), Sijia Li, Yujie Xiao *

Posted Date: 30 December 2025

doi: 10.20944/preprints202512.2703.v1

Keywords: cross-service temporal representation; contrastive learning; microservice anomaly detection; structural dependency



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Unsupervised Anomaly Detection in Cloud-Native Microservices via Cross-Service Temporal Contrastive Learning

Zizhao Zhang, Weicheng Liu, Jingzhou Tao, Haotian Zhu, Sijia Li and Yujie Xiao *

¹ University of Michigan, Ann Arbor, USA

² University of Texas at Dallas, Richardson, USA

³ Northeastern University, Seattle, USA

⁴ New York University, New York, USA

⁵ University of Michigan, Ann Arbor, USA

⁶ University of California, Berkeley, Berkeley, USA

* Correspondence: yujiexiao@berkeley.edu

Abstract

We address a key challenge in cloud native microservice anomaly detection. Existing methods struggle to characterize cross-service dependency propagation and complex temporal dynamics. We propose a temporal representation framework based on cross-service contrastive learning. The method first constructs multi-scale views from service call relations and historical runtime data. It embeds system-level collaborative behaviors into temporal window representations. This yields structure-enhanced input features. It then applies a dual view contrastive encoding strategy with weak augmentation across different time windows. This strategy pulls normal collaborative patterns closer and separates abnormal deviation patterns in the feature space. It improves sensitivity and robustness. The framework further performs unsupervised anomaly scoring through feature aggregation and service-level prototype construction. This avoids extra labeling costs and reduces dependence on manual thresholds. To evaluate the framework, we conduct systematic validation using metrics, logs, and distributed traces from a real open source microservice benchmark system. We also perform multi-dimensional hyperparameter sensitivity analyses across learning rate, batch size, time window length, and the temperature coefficient of the contrastive loss. These analyses reveal stability and adaptability under different dynamic conditions. In comparison with existing baseline models, the method achieves higher accuracy, stronger recall, and better overall performance. This confirms the advantage of cross-service contrastive strategies for capturing anomaly propagation patterns and structural characteristics. Overall, the proposed unified structure of temporal representations and contrastive learning mechanism enables effective unsupervised anomaly detection in collaborative microservice environments. The framework shows strong discriminative capacity and robustness under complex system operations. It provides a reproducible technical solution for improving intelligent operations in cloud native systems.

Keywords: Cross-service temporal representation; contrastive learning; microservice anomaly detection; structural dependency

I. INTRODUCTION

The widespread adoption of cloud native architecture has accelerated the shift of enterprise applications from monolithic systems to microservice ecosystems. Microservices improve delivery speed through fine-grained decomposition, independent deployment, and elastic scaling. They also introduce a more complex runtime landscape. Service counts grow rapidly. Call chains change dynamically. Versions roll frequently. Resource sharing and isolation coexist. Observability data expands at the same time. Metrics, logs, and traces show strong non-stationarity over time. They also

exhibit cross-service coupling and co-variation in the system space. Traditional operations approaches designed for single hosts or single applications do not adapt well to this highly distributed, dependency-intensive, and fast-evolving environment. As a result, microservice anomaly detection has become a core capability for ensuring cloud native reliability and business continuity[1].

The causes of microservice anomalies are diverse and often show system-level propagation. A seemingly local event, such as resource jitter, configuration drift, dependency degradation, or transient network congestion, can be amplified along the call chain into a cascading failure. This process can appear as cross-service temporal resonance and delayed transmission. At the same time, services differ greatly in roles and workload patterns. Gateway services, compute-intensive services, storage and messaging middleware, and business components with different threading models generate highly heterogeneous time series. They differ in metric distributions, periodicity, and burstiness. Detection methods that rely only on a single service view or static thresholds often fail to separate true anomalies from normal fluctuations. These fluctuations can be caused by traffic peaks and valleys, release windows, or cache warm-up. The result is both false alarms and missed detections. This weakens alert credibility and reduces the efficiency of incident response decisions[2].

At the data and labeling levels, anomaly detection in microservice settings faces practical constraints of limited labels, weak priors, and rapid change. Real failure samples are scarce and unevenly distributed. Labeling standards across teams are hard to unify. Anomaly categories also evolve with changes in business forms, infrastructure, and dependency ecosystems. The assumptions required by supervised methods are difficult to maintain in long-term production. Purely unsupervised methods are also vulnerable. They can be disrupted by environmental noise, periodic trends, and multi-source heterogeneity. Therefore, a key challenge is to learn robust representations under weakly labeled or unlabeled conditions. Another challenge is to adapt to continuously changing runtime environments.

Contrastive learning offers a promising direction for this learning setting. It is well-suited to extracting structured representations from large-scale normal operation data. Temporal contrastive learning emphasizes the construction of positive and negative relationships. It aims to learn time series representations that are discriminative, transferable, and noise resistant. It can better capture local bursts, long-term trends, and multi-scale dependencies. Extending contrastive learning to the cross-service dimension may place system correlation at the center of representation learning. One path is to use call dependencies and synchronized load changes to capture collaborative patterns. Another path is to leverage inter-service differences and asynchronous disturbances to learn clearer decision boundaries. Such cross-service temporal representations may reveal propagation routes and early warning signals. They can support more accurate localization and earlier alerts for complex failures[3].

A microservice anomaly detection framework based on cross-service temporal contrastive learning has important theoretical and engineering value. Theoretically, it moves anomaly detection beyond assumptions of single variables, single services, and static distributions. It advances toward system modeling with multi-service linkage, temporal causal coupling, and dynamic evolution. This shift may help establish representation learning paradigms that better reflect real cloud native operating principles. From an engineering perspective, this approach can improve alert robustness and transferability with lower labeling costs. It can reduce reliance on manual rules and experience-based thresholds. It can also increase tolerance to release disturbances, workload fluctuations, and environmental noise. This provides more reliable anomaly signals for SRE and automated operations. As microservice scale continues to grow and platform-level intelligent operations become more critical, contrastive frameworks that model cross-service temporal relations may become a key path toward high availability, lower cost, and sustainable evolution.

I. RELATED WORK

Anomaly detection in cloud and microservice environments has been broadly surveyed, with recurring conclusions that practical systems must handle heterogeneous telemetry, non-stationary workloads, and dependency-driven propagation. Systematic reviews summarize common detection

paradigms and emphasize robustness and operational constraints in cloud settings [4], while microservice-focused surveys further highlight that dependency-aware modeling and propagation reasoning are central to effective detection and diagnosis in service ecosystems [5]. Practical overviews and case-driven studies reinforce these points by showing that anomalies can manifest as cross-component effects under dynamic deployments and diverse runtime behaviors, motivating learning-based approaches that go beyond static rules and thresholds [6-8].

To reduce labeling cost while capturing cross-service coupling, recent work increasingly adopts unsupervised representation learning with multi-source fusion and structural modeling. Multi-source anomaly detection frameworks demonstrate that jointly leveraging heterogeneous observability streams can improve sensitivity and reduce reliance on handcrafted rules [9], while interaction prediction and security-oriented monitoring suggest that learning cross-component behavioral relationships is useful for identifying cascading or coordinated deviations [10-11]. Graph-based reconstruction methods (e.g., nested graph diffusion) further show that encoding dependency structure during reconstruction can expose propagation-style anomalies more explicitly [12]. In parallel, contrastive objectives have been used to build separable embedding spaces for normal versus abnormal patterns without labels, including contrastive multi-modal representation clustering and dependency-aware contrastive modeling, which provide direct methodological grounding for cross-service temporal contrastive strategies [13-14]. Federated or distributed update settings have also been explored to support continual improvement under decentralized data collection, reinforcing the value of representation learning pipelines that can update under realistic deployment constraints without requiring dense anomaly labels [15].

Beyond microservice-specific designs, several modeling components are commonly reused to strengthen temporal sensitivity, generalization, and pipeline stability. Attention-based sequence encoders remain a practical backbone for capturing salient temporal segments in anomaly detection workflows [16], and attention-based long-range dependency modeling has also been used to improve robustness in complex time series more generally [17]. Transformer-style modeling of heterogeneous records provides transferable techniques for aligning mixed signals across time and feature granularities [18], while graph neural representations show that relational aggregation can recover latent system-level outcomes from interaction structure, which supports dependency-graph-based anomaly representations [19]. At the system design level, modular task decomposition and dynamic collaboration offer a general recipe for structuring complex learning pipelines into coordinated components [20]. Parameter-efficient adaptation (including LoRA-style multi-scale tuning and related fine-tuning strategies) provides lightweight mechanisms for adjusting models under evolving environments while limiting destabilizing parameter drift [21-22]. Finally, adaptive control perspectives based on reinforcement learning—ranging from deep Q-learning for scheduling to multi-agent orchestration—highlight policy-based mechanisms that react to changing system conditions and can complement anomaly detection loops in automated operations [23-24], while collaborative evolution studies and risk-aware multi-agent optimization provide additional principles for maintaining stable coordinated behavior under dynamic objectives [25-26]. Knowledge-augmented agent frameworks further motivate integrating structured reasoning components for more interpretable downstream analysis built on learned representations, without changing the core unsupervised detection objective [27].

I. METHOD

A. General Introduction

We represent a microservice system as a coupled entity of a call graph evolving and a sequence of multivariate metrics. Let the set of microservices be $V = \{v_1, \dots, v_N\}$, and the call relationships at time t be E_t . Then the system structure can be represented as:

$$G_t = (V, E_t) \quad (1)$$

Each service has a d -dimensional observation vector $X_{v,t}$ at time t , and service-level time segments are constructed using a sliding window of length w :

$$X_{v,t} = [x_{v,t-w+1}, \dots, x_{v,t}] \quad (2)$$

The core of the proposed framework is to leverage cross-service associations as a structural perspective and integrate multi-scale historical windows as a temporal perspective. The unified encoder is designed to learn both alignable and distinguishable service representations by fusing structural and temporal information, which enhances the robustness of anomaly discrimination under conditions of limited or absent annotation. To achieve this, the model draws on adaptive deep learning strategies proposed by Lyu et al. [28] for handling complex multi-objective interactions in microservice environments, ensuring that cross-service dependency signals are effectively embedded in the representation. The contrastive learning objective, inspired by the dependency modeling approach of Xing et al. [29], encourages the model to align normal collaborative behaviors and separate abnormal deviations in the learned feature space, even without strong supervision. Additionally, the temporal encoding module utilizes attention-based convolutional techniques to ensure that both local bursts and long-term patterns are adequately captured [30]. The overall architecture of the framework is illustrated in Figure 1.

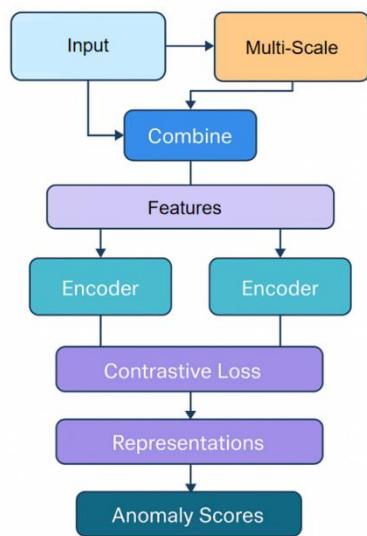


Figure 1. Overall model architecture diagram.

A. Cross-Service Time-Series View Construction

To explicitly characterize the cross-service propagation and coupling of anomalies, we introduce neighborhood aggregation for each service to form a structurally enhanced temporal input. Let A_t denote the adjacency matrix at time t , and we obtain the neighborhood context using simple normalization aggregation:

$$x_{v,t} = \sum_{u \in V} A_t(u, v) x_{v,t} \quad (3)$$

The local observations and neighborhood context are then concatenated into a structure-temporal joint token:

$$z_{v,t} = \text{Concat}(x_{v,t}, \tilde{x}_{v,t}) \quad (4)$$

And form a window-level input:

$$Z_{v,t} = [Z_{v,t-w+1}, \dots, z_{v,t}] \quad (5)$$

This construction method injects weakly structured information about "who affects whom and when the impact occurs" into the time model without relying on complex priors, enabling subsequent encoders to simultaneously perceive local fluctuations and linkages.

A. Cross-Service Time-Series Comparison Learning Objectives

We construct two lightweight augmented views, $h_{v,t}^a$ and $h_{v,t}^b$ for the same service window, such as time jitter, subsequence pruning, feature masking, or slight perturbations based on the call neighborhood, and feed them into a shared encoder f_θ to obtain the representations:

$$h_{v,t}^a = f_\theta(a(Z_{v,t})), \quad h_{v,t}^b = f_\theta(b(Z_{v,t})) \quad (6)$$

Cosine similarity is used as the alignment metric:

$$\text{sim}(p, q) = \frac{p^T q}{\|p\| \|q\|} \quad (7)$$

And use the contrast loss in the form of service-level InfoNCE:

$$L_{\text{svc}} = -\log \frac{\exp(\text{sim}(h_{v,t}^a, h_{v,t}^b) / \tau)}{\sum_{(u,s) \in B} \exp(\text{sim}(h_{v,t}^a, h_{v,t}^b) / \tau)} \quad (8)$$

To ensure system-level consistency, prototype alignment of clusters or call subgraphs can be further introduced. Let $c(v)$ be the cluster to which the service belongs, and p be the prototype. The overall goal can be written as:

$$L = L_{\text{svc}} + \lambda L_{\text{clu}} \quad (9)$$

This goal encourages the same service to maintain its representation unchanged under different enhancements, while also differentiating between different services or different time segments, thereby strengthening the stable structure of the "normal collaborative mode" at the representation level.

A. Anomaly Scoring and Online Inference

After training, we construct a normal prototype for each service using the representation of the historical reference interval:

$$p_v = \frac{1}{T_0} \sum_{t=1}^{T_0} h_{v,t} \quad (10)$$

During the online phase, a representation $h_{v,t}$ is calculated for the new window, and the degree of deviation from the normal prototype is used as the service-level anomaly score:

$$s_{v,t} = 1 - \text{sim}(h_{v,t}, p_v) \quad (11)$$

Considering the varying importance of different services to the business chain, the system-level anomaly strength can be obtained by aggregating the weights w_v :

$$S_t = \sum_{v \in V} w_v s_{v,t} \quad (12)$$

When S_t or critical service $s_{v,t}$ continues to rise, further root cause analysis and call chain backtracking can be triggered. Overall, this method takes cross-service temporal comparative learning as the main thread, unifying structural dependencies, temporal dynamics, and unsupervised representations into the same framework, providing a scalable modeling paradigm for early identification and robust alerting of complex anomalies in microservices.

I. EXPERIMENTAL RESULTS

A. Dataset

This study uses an open-source anomaly dataset based on the Train Ticket benchmark system as the experimental data source. Train Ticket is a representative microservice application for ticket purchasing. It contains about 41 microservices. These services cover gateways, orders, payments, users, notifications, and other functional components. The system supports containerized and Kubernetes-based deployment. It can effectively simulate cross-service call relationships and runtime fluctuations in real cloud native environments.

Specifically, we adopt the publicly released dataset titled "Anomalies in Microservice Architecture (train-ticket) based on version configurations." The dataset uses Train Ticket as the running platform. It provides multiple anomaly scenarios triggered by changes in versions and dependency configurations. It also collects three key types of observability data in parallel. These include service logs, Jaeger distributed traces, and Prometheus metric data. This design forms a comprehensive monitoring view that covers structural dependencies, temporal evolution, and multi-source signals. The data are organized using a directory schema of "changed microservice_third party library_version_collection date." This structure facilitates mapping anomalies to specific service changes and their impact scope.

The dataset aligns directly with the theme of this paper on microservice anomaly detection based on cross-service temporal contrastive learning. On one hand, logs, metrics, and traces provide natural carriers for cross-service temporal alignment and multi-view augmentation. They support the construction of positive and negative samples at both service and system levels. On the other hand, performance degradation and behavioral anomalies caused by version or dependency changes show clear cross-service propagation and coupling. This property can help the model learn representation

boundaries between normal collaborative patterns and abnormal deviation patterns. With open accessibility, complete modalities, and well-defined anomaly causes, this dataset offers a stable and reproducible foundation. It supports the structure, temporal joint modeling, and cross-service alignment learning in our method.

A. Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results.

Model	ACC	Precision	Recall	F1-Score
SRdetector[31]	0.861	0.842	0.811	0.826
Uac-ad[32]	0.873	0.854	0.822	0.838
MADMM[33]	0.887	0.866	0.845	0.855
TraLogAnomaly[34]	0.901	0.884	0.862	0.873
TraceDAE[35]	0.912	0.895	0.876	0.885
Ours	0.931	0.918	0.903	0.910

In terms of overall performance, the proposed method achieves the best results across all four evaluation metrics. This indicates that cross-service temporal representations and the contrastive learning mechanism can capture collaborative behaviors and anomalous deviations more effectively in microservice systems. Compared with other methods, the advantage of ACC suggests a more stable global identification capability. The model maintains higher classification accuracy under complex dependency structures. This also shows that structured temporal representations play an important role in reducing both false positives and false negatives.

A closer look at Precision shows that the proposed method provides stronger positive discrimination than the other models. This means the model can distinguish normal and abnormal behaviors more precisely under disturbances such as runtime fluctuations and call chain changes. It does not produce excessive false alarms due to local errors or transient resource jitter. For real operational settings, higher Precision is especially valuable. It reduces ineffective alerts and improves the efficiency of operational resource allocation.

The method also remains superior on Recall. This suggests that the cross-service temporal alignment mechanism can describe anomaly propagation with higher sensitivity. It is more likely to capture deeper fault manifestations. Many traditional methods rely more on local service views. They therefore struggle to reflect the temporal transmission characteristics of call dependencies. In contrast, this cross-service contrastive framework can interpret anomalous signals under multi-service interactions in a more comprehensive way. This leads to higher recall.

Finally, the F1-Score further confirms stronger overall performance. It reflects an effective balance between accurate identification and comprehensive coverage. Compared with baseline models, this advantage suggests that cross-service contrastive learning improves representation quality. It also helps establish more robust decision boundaries for mapping anomalies at both service and system levels. Overall, these results indicate that the framework is suitable for complex and dynamic microservice environments. It has the potential to support more reliable anomaly detection and system-level early warning capabilities.

This paper also presents the impact of the learning rate on the experimental results, as shown in Table 2.

Table 2. The impact of the learning rate on experimental results.

Learning Rate	ACC	Precision	Recall	F1-Score
0.0006	0.914	0.892	0.875	0.883
0.0005	0.921	0.902	0.889	0.895
0.0004	0.928	0.910	0.897	0.904
0.0003	0.930	0.915	0.900	0.908

0.0002	0.931	0.917	0.902	0.909
0.0001	0.931	0.918	0.903	0.910

The learning rate sensitivity results show a clear trend. As the learning rate is gradually reduced from 0.0006, the model improves steadily on ACC, Precision, Recall, and F1-Score. This suggests that a higher learning rate makes parameter updates too aggressive. It introduces more structural noise during cross-service temporal feature extraction. The representation space becomes less stable. With a lower learning rate, the model can better exploit cross-service contrastive relations for consistent alignment. This leads to more robust system-level anomaly representations.

The Precision trend further indicates that an overly large learning rate can cause the model to overfit local service patterns. It then fails to capture truly system-correlated anomalous features. When the learning rate falls within 0.0003 to 0.0001, Precision rises noticeably. This implies that the model can achieve more accurate anomaly identification in cross-service collaborative structural representations. This is crucial for reducing false alerts. Microservice systems often experience large business fluctuations. A high false alarm rate can directly undermine the reliability of automated operations.

Recall increases as the learning rate decreases. This indicates that the temporal representations generated by contrastive learning generalize better under lower learning rates. The model becomes more capable of capturing anomaly propagation along call chains. Compared with other training settings, a lower learning rate allows the model to aggregate multi-scale cross-service variations more effectively. This improves coverage of implicit degradation and link-level propagation anomalies.

The F1-Score trend provides a comprehensive view. Model performance becomes more stable as the learning rate decreases. The best balance is achieved at 0.0001 and 0.0002. This suggests that the collaborative representations formed by cross-service temporal contrastive learning benefit from small step updates. The optimization converges more reliably. The model can more fully characterize internal normal patterns and abnormal deviation patterns. This enhances the stability and credibility of anomaly detection decisions.

The impact of different batch sizes on cross-service contrastive learning is mainly reflected in two aspects: gradient estimation stability and representation space structure. Smaller batches can introduce more randomness, making cross-service time-series segments exhibit richer perturbation patterns during training, which is beneficial for contrastive learning to uncover more detailed differential features. However, this may also lead to excessively large gradient variance, affecting the stability of optimization convergence. Larger batches tend to provide smoother gradient estimates, enabling cross-service collaborative patterns to form a clearer aggregation structure in the representation space, but there is a risk of decreased sensitivity to local anomaly signals. Therefore, it is necessary to systematically examine the sensitivity of cross-service contrastive learning under different batch sizes, focusing on recall, an indicator that directly reflects anomaly coverage capability, to characterize the balance between "stable training" and "anomaly capture capability," and to provide an operational adjustment basis for hyperparameter configuration in actual deployment. The experimental results are shown in Figure 2.

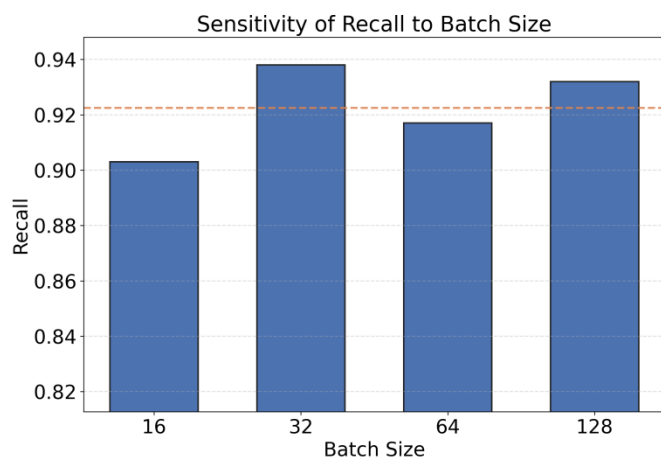


Figure 2. Sensitivity experiment of different batch sizes to cross-service contrastive learning recall.

The overall trend shows that different batch sizes have a clear impact on recall. This suggests a coupled relationship between the stability of cross-service contrastive learning, the optimization path, and the batch scale. With a larger batch, gradient estimates become smoother. Cross-service collaborative patterns in the representation space are easier to extract. This provides more consistent statistical cues for identifying anomaly propagation paths. Therefore, the model needs a relatively large batch to fully reveal the contribution of cross-service dependency structures to anomaly detection.

When the batch size is adjusted from 16 to 32, recall increases significantly. This indicates that small batches introduce high variance gradients. Such variance can hinder the convergence of the contrastive objective. Stable association patterns among key services then become difficult to form. In contrast, a batch size of 32 offers a more stable update rhythm for service level contrastive relations. It helps the model capture propagation dynamics along call chains. This improves recall.

It is worth noting that recall drops when the batch size increases to 64. This shows that an excessively large batch does not necessarily yield better cross-service anomaly representations. Overly smooth gradient updates may reduce sensitivity to fine-grained anomaly patterns. Some local propagation anomalies may become less visible. This implies that cross-service anomaly detection does not rely only on scaling the batch size. It requires a balance between extracting collaborative structures and preserving sensitivity to details.

When the batch size is further expanded to 128, recall rises again. This suggests that with higher sampling coverage, the overall dependency structure of the system is characterized more completely. The propagation paths of cross-service anomalies are more likely to be captured by aligned representations. This pattern also implies a possible mapping between batch size and dependency complexity. Thus, selecting an appropriate batch scale is an important lever for optimizing the performance of cross-service temporal contrastive learning.

The length of the time window directly determines the range of historical context that the model can observe across service time sequences, thus affecting the discriminability of anomaly patterns and the optimization difficulty of the contrastive learning objective. When the window is too short, the model can only see local fluctuations, making it difficult to perceive delay propagation and multi-scale superposition effects along the service chain; when the window is too long, it may introduce a large amount of background information unrelated to the current anomaly, causing noise interference when the contrastive learning constructs "normal collaborative patterns" and "abnormal deviation patterns" in the representation space. Therefore, it is necessary to systematically examine the sensitivity of a single model on multiple metrics around different time window lengths, and to analyze the nonlinear relationship between window length and cross-service contrastive learning performance through visual line analysis, providing a more operational basis for selecting the granularity of time sequence modeling in microservice anomaly detection. The experimental results are shown in Figure 3.

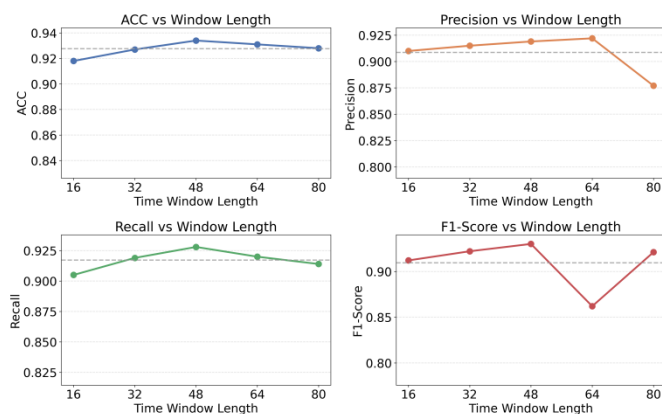


Figure 3. The effect of time window length variation on experimental results.

These results indicate that the length of the time window has a significant impact on representation stability for cross-service anomaly detection. On ACC, performance increases as the window grows from 16 to 48. It then slightly decreases when the window is further extended to 64 and 80. This pattern suggests two risks. A short window cannot capture the temporal propagation of cross-service dependencies. A long window may introduce noisy information and weaken structured linkage signals. Therefore, the time window should balance the coverage of temporal information and the level of noise. This balance is necessary to realize the advantage of contrastive learning for cross-service behavior modeling.

The changes in Precision and Recall show that the middle range performs best. This indicates that a moderate window helps the model distinguish normal collaborative patterns from abnormal deviation paths. When the window is too small, the main issue is insufficient information. The model cannot reveal cascading propagation among microservices. When the window is too large, it includes a longer history. It may also introduce background changes that are unrelated to the current anomaly. This can reduce discriminative capacity. This observation reflects the core value of cross-service contrastive learning. It does not depend only on local samples. It relies on multi-scale real collaborative behaviors. It also needs to avoid interference from irrelevant context.

Finally, the F1 trend shows that a window length of 48 achieves a more balanced overall performance. This highlights an important point for cross-service anomaly detection. Temporal modeling should not pursue longer ranges without restraint. In systems with complex call structures, a moderate window is more conducive to learning stable service correlation features. This suggests that time scale selection is closely related to the model's ability to capture anomaly propagation. Optimizing the time window from a system perspective is an important support for improving robustness and generalization in cross-service contrastive learning frameworks.

The temperature coefficient in the contrastive loss directly controls the "sharpness" of the similarity between positive and negative samples in softmax normalization, thus affecting the strength with which contrastive learning brings positive samples closer and pushes negative samples further apart in the representation space. In cross-service anomaly detection scenarios, if the temperature coefficient is too small, the model may overemphasize a few extreme negative samples, resulting in an overly tight representation space and difficulty in maintaining structural diversity between different services; if the temperature coefficient is too large, the differences between positive and negative samples are over-smoothed, weakening the ability to distinguish anomaly propagation patterns. Therefore, it is necessary to systematically examine the sensitivity of a single model on multiple metrics around different temperature coefficients to characterize the balance between "contrast strength" and cross-service representation quality. The experimental results are shown in Figure 4.

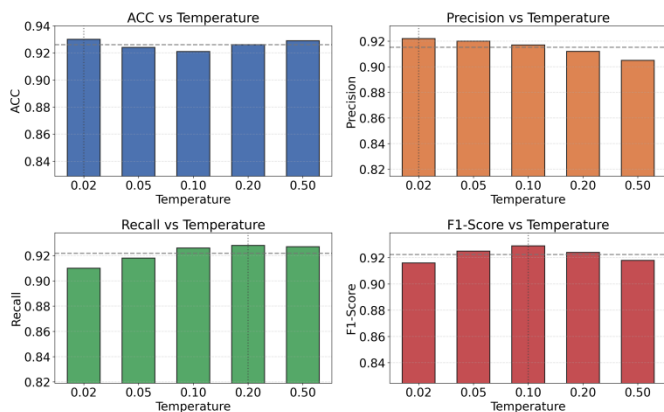


Figure 4. The effect of changes in the temperature coefficient of loss on the experimental results.

Across different temperature settings, the results indicate that neither a lower temperature nor a higher temperature is universally better. The key is to balance contrastive force and structural separability. When the temperature is too low, the model overemphasizes differences among samples. The cross-service temporal representation space becomes overly compact. Local service fluctuations can be amplified into anomalous features. This harms overall representation stability. In contrast, an excessively high temperature weakens sample discrimination. The model then fails to capture structural characteristics of cross-service anomaly propagation. This reduces anomaly recognition capability.

Among the metrics, the trends of F1 and Recall suggest that a middle temperature range provides a better balance. This implies that a moderate temperature helps construct a more structural service representation space. Such a space can capture key propagation paths in cross-service correlation modeling. It also preserves robust sensitivity to anomalies along service call chains. In other words, the temperature coefficient influences more than the loss scale in this framework. It implicitly shapes the geometry of the cross-service representation space.

Precision and ACC remain relatively stable, but they still decline under extreme temperatures. This indicates that temperature changes affect how the model delineates boundaries between anomalous patterns and normal collaborative behaviors. Cross-service anomalies often involve multi-scale propagation and link-level resonance. Temperature adjustment, therefore, controls the strength of aggregation and separation in the contrastive space. An appropriate temperature promotes reasonable alignment of cross-service temporal information. It helps the model maintain high anomaly sensitivity at the structural modeling level. It also avoids incorrect decisions caused by excessive alignment or excessive dispersion.

I. CONCLUSIONS

This paper addresses core challenges of anomaly detection in microservice systems. It proposes a temporal contrastive learning framework that targets cross-service dependencies. The framework jointly models multi-scale temporal dynamics and microservice call structures. It aims to achieve more robust anomaly pattern identification. By placing service-level local patterns and system-level collaborative behaviors in the same feature space, the learned structure temporal representations can describe how anomalies propagate along call chains. The model thus becomes more sensitive to abnormal signals in complex business workflows. This study goes beyond traditional single-service detection. It highlights the importance of cross-service dependency structures for anomaly detection.

At the representation learning level, the method introduces a contrastive mechanism. It uses soft constraints to enhance the aggregation of normal operating patterns. It also strengthens the separability of abnormal deviation patterns in the feature space. This design reduces reliance on large amounts of labeled data. The method can still function effectively under weak labels or imbalanced data. This matches practical requirements in real cloud native deployments. With flexible view augmentation, the framework shows good robustness to dynamic microservice changes, workload

fluctuations, and runtime disturbances. It offers new design insights for operations monitoring in large-scale containerized systems.

From an engineering perspective, this work supports the development of intelligent microservice operations. It can contribute to automatic alerting, root cause localization, performance degradation detection, and capacity planning. The proposed cross-service contrastive learning idea is not limited to business call chain structures. It can also be extended to the fusion analysis of multi-source heterogeneous runtime signals. This extension may provide more reliable observability-driven intelligent decisions for highly reliable cloud native platforms. As business scale grows and operational complexity increases, such a structured temporal collaborative representation modeling may become a key technical foundation for future intelligent monitoring systems.

There remains considerable room for further research. The interaction mechanism between cross-service graph structures and temporal information can be refined. Future work can explore higher-level representation strategies, such as hybrid attention, graph-level semantic aggregation, and cross-granularity alignment. The framework can also be extended beyond anomaly detection to predictive maintenance, system stability planning, and adaptive resource scheduling. This could further advance automated operations capabilities. With the development of online learning and incremental learning paradigms, the framework can be combined with continual learning mechanisms. This may help the model maintain agile adaptability under unstable environments. It can promote cloud native intelligent operations toward greater autonomy, higher efficiency, and stronger elasticity.

References

1. Hao, Junfeng, et al. "Multi-task federated learning-based system anomaly detection and multi-classification for microservices architecture." *Future Generation Computer Systems* 159 (2024): 77-90.
2. Dodda, Suresh, et al. "Enhancing microservice reliability in cloud environments using machine learning for anomaly detection." *2024 International Conference on Computing, Sciences and Communications (ICCS)*. IEEE, 2024.
3. Xing, S., Wang, Y., & Liu, W. (2025). Multi-Dimensional Anomaly Detection and Fault Localization in Microservice Architectures: A Dual-Channel Deep Learning Approach with Causal Inference for Intelligent Sensing. *Sensors*, 25(11), 3396.
4. T. Hagemann and K. Katsarou, "A Systematic Review on Anomaly Detection for Cloud Computing Environments," *Proceedings of the 2020 3rd Artificial Intelligence and Cloud Computing Conference*, pp. 83–96, 2020.
5. J. Soldani and A. Brogi, "Anomaly Detection and Failure Root Cause Analysis in (Micro) Service-Based Cloud Applications: A Survey," *ACM Computing Surveys*, vol. 55, no. 3, pp. 1–39, 2022.
6. C. Nwachukwu, K. Durodola-Tunde and C. Akwiwu-Uzoma, "AI-Driven Anomaly Detection in Cloud Computing Environments," *International Journal of Science and Research Archive*, vol. 13, no. 2, pp. 692–710, 2024.
7. Poddaturi, Santhosh. "AI for Microservice Monitoring & Anomaly Detection." *International Journal of Emerging Trends in Computer Science and Information Technology* (2025): 192-211.
8. Floroiu, Maximilian Stefan, et al. "Anomaly detection and root cause analysis of microservices energy consumption." *2024 IEEE International Conference on Web Services (ICWS)*. IEEE, 2024.
9. Li, Zhengxin, Junfeng Zhao, and Jia Kang. "Multi-source anomaly detection for microservice systems." *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings*. 2024.
10. Aktaş, Kemal, and H. Hakan Kilinc. "Interaction prediction and anomaly detection in a microservices-based telecommunication platform." *Proceedings of the 2024 International Conference on Software and Systems Processes*. 2024.
11. Ramamoorthi, Vijay. "Anomaly Detection and Automated Mitigation for Microservices Security with AI." *Applied Research in Artificial Intelligence and Cloud Computing* 7.6 (2024): 211-222.
12. Fan, Mengwei, et al. "Multi-modal anomaly detection for microservice system through nested graph diffusion reconstruction." *Applied Intelligence* 55.11 (2025): 784.

13. Wang, Peipeng, et al. "Unsupervised microservice system anomaly detection via contrastive multi-modal representation clustering." *Information Processing & Management* 62.3 (2025): 104013.
14. Y. Xing, Y. Deng, H. Liu, M. Wang, Y. Zi and X. Sun, "Contrastive Learning-Based Dependency Modeling for Anomaly Detection in Cloud Services," arXiv preprint, arXiv:2510.13368, 2025.
15. Raeiszadeh, Mahsa, et al. "Asynchronous Real-Time Federated Learning for Anomaly Detection in Microservice Cloud Applications." *IEEE Transactions on Machine Learning in Communications and Networking* (2025).
16. H. Wang, C. Nie and C. Chiang, "Attention-Driven Deep Learning Framework for Intelligent Anomaly Detection in ETL Processes," 2025.
17. Q. Xu, W. Xu, X. Su, K. Ma, W. Sun and Y. Qin, "Enhancing Systemic Risk Forecasting with Deep Attention Models in Financial Time Series," *Proceedings of the 2025 2nd International Conference on Digital Economy, Blockchain and Artificial Intelligence*, pp. 340–344, 2025.
18. W. C. Chang, L. Dai and T. Xu, "Machine Learning Approaches to Clinical Risk Prediction: Multi-Scale Temporal Alignment in Electronic Health Records," arXiv preprint, arXiv:2511.21561, 2025.
19. R. Liu, R. Zhang and S. Wang, "Graph Neural Networks for User Satisfaction Classification in Human-Computer Interaction," arXiv preprint, arXiv:2511.04166, 2025.
20. S. Pan and D. Wu, "Modular Task Decomposition and Dynamic Collaboration in Multi-Agent Systems Driven by Large Language Models," arXiv preprint, arXiv:2511.01149, 2025.
21. Y. Huang, Y. Luan, J. Guo, X. Song and Y. Liu, "Parameter-Efficient Fine-Tuning with Differential Privacy for Robust Instruction Adaptation in Large Language Models," arXiv preprint, arXiv:2512.06711, 2025.
22. H. Zhang, L. Zhu, C. Peng, J. Zheng, J. Lin and R. Bao, "Intelligent Recommendation Systems Using Multi-Scale LoRA Fine-Tuning and Large Language Models," 2025.
23. K. Gao, Y. Hu, C. Nie and W. Li, "Deep Q-Learning-Based Intelligent Scheduling for ETL Optimization in Heterogeneous Data Environments," arXiv preprint, arXiv:2512.13060, 2025.
24. G. Yao, H. Liu and L. Dai, "Multi-Agent Reinforcement Learning for Adaptive Resource Orchestration in Cloud-Native Clusters," *Proceedings of the 2nd International Conference on Intelligent Computing and Data Analysis*, pp. 680–687, 2025.
25. Y. Li, S. Han, S. Wang, M. Wang and R. Meng, "Collaborative Evolution of Intelligent Agents in Large-Scale Microservice Systems," arXiv preprint, arXiv:2508.20508, 2025.
26. R. Ying, J. Lyu, J. Li, C. Nie and C. Chiang, "Dynamic Portfolio Optimization with Data-Aware Multi-Agent Reinforcement Learning and Adaptive Risk Control," 2025.
27. Q. Zhang, Y. Wang, C. Hua, Y. Huang and N. Lyu, "Knowledge-Augmented Large Language Model Agents for Explainable Financial Decision-Making," arXiv preprint, arXiv:2512.09440, 2025.
28. N. Lyu, Y. Wang, Z. Cheng, Q. Zhang and F. Chen, "Multi-Objective Adaptive Rate Limiting in Microservices Using Deep Reinforcement Learning," arXiv preprint, arXiv:2511.03279, 2025.
29. Y. Xing, Y. Deng, H. Liu, M. Wang, Y. Zi and X. Sun, "Contrastive Learning-Based Dependency Modeling for Anomaly Detection in Cloud Services," arXiv preprint, arXiv:2510.13368, 2025.
30. N. Lyu, F. Chen, C. Zhang, C. Shao and J. Jiang, "Deep Temporal Convolutional Neural Networks with Attention Mechanisms for Resource Contention Classification in Cloud Computing," 2025.
31. Ge, Haixin, et al. "SRdetector: Sequence Reconstruction Method for Microservice Anomaly Detection." *Electronics* 14.1 (2024): 65.
32. Liu, Hongyi, et al. "Uac-ad: Unsupervised adversarial contrastive learning for anomaly detection on multi-modal data in microservice systems." *IEEE Transactions on Services Computing* 17.6 (2024): 3887-3900.
33. Wang, Peipeng, et al. "MADMM: microservice system anomaly detection via multi-modal data and multi-feature extraction." *Neural Computing and Applications* 36.25 (2024): 15739-15757.
34. Wei, X., Sun, C. A., Yang, P., Zhang, X. Y., & Towey, D. (2025). TraLogAnomaly: A microservice system anomaly detection approach based on hybrid event sequences. *Science of Computer Programming*, 245, 103303.
35. Li, Junjun, et al. "TraceDAE: Trace-Based Anomaly Detection in Micro-Service Systems via Dual Autoencoder." *IEEE Transactions on Network and Service Management* (2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.