##Response to Reviewer 1Wzx

Thanks for your valuable suggestions. The detailed replies will be provided as follows. The updated figures and experimental results can be found in the '**README.md**' file.

Cons:

C1: Inadequate disclosure of experimental settings. The descriptions of temporal and spatial disruption experiments (Lines 114-120 and Lines 308-310) are too simple to be confusing.

Answer1: Our detailed experimental details in Tab.1 and Tab.2 are divided into two aspects as follows:

1. Details of disruption: Apology for the brevity in Line 114-120. SWaT is divided into multiple windows of length 24, where the training minimizes the Mean Squared Error (MSE) between reconstruction results and inputs. During testing, temporal disruption is to hide data from other time steps when reconstructing data for a specific time point, achieved by setting the attention scores for other time points to $-\infty$. Spatial disruption entail that for each given channel, data from other channels are randomly shuffled using 'torch.randperm'. The disruption have no additional parameters. We will add details in main text.
2. Details of reconstruction strategy: We adopted the original Transformer framework using existing reconstruction strategies individually (Line 104-113). For the compression strategy, we set the feature dimension to a value (32 for SWaT) smaller than the number of input channels. For the mask strategy, we randomly set the value of all channels to 0 at some time points in a certain proportion (20% for SWaT), with the reconstruction targets being the original values of the masked portion. We will add it to the Appendix.

C2: Inconsistent and disordered academic terminology. See "target / restored results / reconstruction results / reconstruction error / restored error".

Answer2: We will consistently use 'reconstruction'. The term "target" will primarily be used in mapping functions.

C3: Definitions that lack mathematical rigor. "small" (Lines 328) in Definition 3.2 lacks mathematical meaning and is an invalid definition. "small enough"(Lines 342) in Theorem 3.3 lacks mathematical meaning and is an invalid definition.

Answer3: Sorry for any confusion caused. Thanks for your advice, we have changed 'small enough' to '$\rightarrow \mathbf{0}$'.

C4: Incorrect claims. Reconstruction with a mask requires function f to have the ability to recover the masked part based on the context that is not masked. Such f may not conform to definition 3.1, so f is dependency sensitive according to definition 3.2. See "compression and mask based reconstruction are both dependency insensitive."(Lines 332)

Answer4: We'll clarify 'dependency sensitive' in Definition 3.2 and emphasize that it refers to maintaining the sensitivity to both temporal and spatial dependencies. According to Definition 3.1, the masking strategy is not time-independent, but channel-independent. Thus, according to Definition 3.2, mask based reconstruction is not dependency-sensitive due to the insensitivity to spatial dependency. Besides, we apologize for the confusion in lines 328-330, which has been corrected to 'the reconstruction function f is neither time-independent nor channel-independent'.

C5: Incorrect mathematical expression. In Theorem 3.3.(Lines 339-348), the mathematical expression of condition 2 is inconsistent with condition 1, and the superscript is confused.

Answer5: Sorry for the mistakes, we have unified the superscript in Theorem 3.3.

C6: Undeclared symbol. Lines 347 claim that f is a composite mapping of the reconstruction mapping g and h, which is both undeclared and confusing.

Answer6: $f: X \mapsto h(X^g)$ denotes that X undergoes an invertible mapping $g(X)$ followed by function $h(X^g)$. The two steps are described separately in order to find $g(X)$ that satisfies Theorem 3.3.

C7: Loose and ambiguous language. See "operation / strategy / mapping"(Lines 349-357) for the same thing g(X).

Answer7: In the theorem, we still refer to it as mapping, indicating that $g(X)$ is a reversible function. For other parts, we will uniformly term it as 'strategy', denoting the approach satisfying the invertible mapping g in Theorem 3.3.

C8: Redescribe and name existing methods. The essence of the so-called Statistical Feature Removal (Proposition 3.4) is equivalent to a Z-Score operation on each univariate time series data, which is a common pre-processing method. (See "Its core includes a statistical feature removal strategy." Lines 24-25 in Abstract.)

Answer8: The differences with Z-score: We have different objectives. Z-score is a preprocessing method for the whole sequence, which is irrelevant to the model. While we process each input window, the goal is to remove the statistical information within the window, so that the model must restore the statistical features in the original sequence through spatio-temporal dependencies. In Line 381-387, ReVIN [2] also mentions that the practice in lines 359-369 serves to remove statistics, so we called it Statistical Feature Removal (SFR). In our work, SFR is just a simple reconstruction strategy we proposed in order to satisfy theorem 3.3.

[2] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. 2022. Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift. In Proceedings of the 10th International Conference on Learning Representations

C9: Wrong mathematical expression. In Lines 405, μ(X) is a statistic of Gaussian distribution, while the definition of μ in Lines 407 means that the mean of the Gaussian distribution has a dimension of T×S, which is not consistent with the multivariate Gaussian distribution statistic (vector). The same thing goes with σ.

Answer9: We mainly want to convey: each element $\hat{X}_{ts}$ is assumed to adhere to a Gaussian distribution with mean $\mathrm{\mu_{ts}}$ and variance $\mathrm{\sigma_{ts}^2}$. To avoid ambiguity, we have removed the following content: "it is generally assumed that $\mathrm{P}(\hat{X}|X) = \mathcal{N}(\mu(X), \sigma^2(X))$, where $\mathrm{\mu, \sigma}$: $\mathbb{R}^{\mathrm{T\times S}} \longmapsto \mathbb{R}^{T\times S}$, $\mathcal{N}$ denotes the Gaussian distribution." (Lines 404-407).

C10: Inappropriate illustration. The illustration in Figure 2 on the so-called Statistical Feature Removal is not intuitive.

Answer10: For the part in Figure 2 that depicts the statistical features removal (SFR), the data from several channels are transformed from different distributions on the left to Gaussian distribution with mean and variance equal to 0 and 1, respectively. The gray ellipses represent the same distribution cluster. We will further refine this part in subsequent revisions.

C11: Experimental comparisons lack recent algorithms. [P:94.03 R:82.96 F1:88.15] on WADI and [P:94.81 R:91.93 F1:93.35] on SMD is reported in [1], which is higher than proposed method.
[1] Zhang, Zhenwei, et al. "Unravel Anomalies: an End-to-End Seasonal-Trend Decomposition Approach for Time Series Anomaly Detection." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.

Answer11: We have carefully checked the results in [1], which presented the point-adjusted F1 score that corresponds to $\mathrm{F1_{PA}^*}$ in our paper. Our results are obviously better, achieving [P:90.88 R:93.66 F1_PA:92.25] on WADI and [P:96.40 R:96.26 F1_PA:96.33] on SMD (detailed results are shown in **README.md**). We will add this method for comparison. In Line 662, point adjustment (PA) can be explained as follows: if at least one timestamp in a consecutive anomalous sequence is detected, then all timestamps in that sequence are considered correctly detected. As far as we know, most studies still focus on $\mathrm{F1_{PA}}$ scores and simplify them to $\mathrm{F1}$, and only a few methods like GDN and NSPR provide the $\mathrm{F1}$ score before adjustments. Note $\mathrm{F1_{PA}}$ will overestimate the performance of a model [3], so we use both $\mathrm{F1}$ and $\mathrm{F1_{PA}}$ simultaneously to evaluate the model's performance. Obviously, we have achieved significant performance improvements, especially on the SWaT and WADI datasets.

[3] Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. 2022. Towards a rigorous evaluation of time-series anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. 7194–7201.

C12: There is occlusion in the illustrations. See Fig4 & Fig6.

Answer12: We are sorry for that and we have corrected these figures, as shown in '**README.md**'.

Questions

Q1: Why is the parameter setting of the comparison algorithm not disclosed, and how to ensure that the parameters of the comparison algorithm on the data set are the best results after tuning?

Answer1: We use published results for comparison algorithms whenever available. Otherwise, we replicate the experiment using the provided code and fine-tune it according to the reported instructions.

Q2: Why not show the visualizations of Figure 1 case for each algorithm?

Answer2: The purpose of Fig. 1 is to illustrate inherent issues with reconstruction errors, hence the use of a generic framework. Detailed model comparisons will be included in the appendix.

Q3: The higher the sensitivity to normal modalities, the lower the robustness to normal noise, and vice versa. How does SensitiveHUE trade off false positives and false negatives?

Answer3: Both spatio-temporal relationships and noise are normal patterns (lines 83-94). The sensitivity of the model to normal patterns implies the sensitivity to spatio-temporal dependency and noise, so 'normal modalities' and 'noise' are not two conflicting concepts. SensitiveHUE employs statistical feature removal to enhance sensitivity to spatio-temporal dependency and utilizes uncertainty estimation to quantify noise levels, effectively balancing model performance.

Q4: Is the time window a key variable affecting performance?

Answer4: Analysis of window size's impact on model performance is provided in Figure 4 and Lines 850–855. For the MTS anomaly detection methods, the appropriate window size is important for effectively learning spatio-temporal dependencies within each input window.

Q5: How to ensure that 30 epochs satisfy sufficient convergence conditions for all datasets?

Answer5: Early stop strategy (Lines 1266–1267) is used to determine model convergence, typically achieved within 30 epochs (see Fig. 10).

Q6: Why does WADI require 10% for the validation set and 25% for the other datasets?

Answer6: 10% is a common validation set percentage for the WADI dataset, as the size of this dataset is relatively large. It is also noted in literature [4].

[4] Zhihan Li, Youjian Zhao, Jiaqi Han, Ya Su, Rui Jiao, Xidao Wen, and Dan Pei. Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. KDD, 2021.

Q7: SWaT has a single variable of identity, and it does not have space-time dependence. Do these single variables affect performance?

Answer7: The term "spatio-temporal dependency" refers to the correlations within a single entity. In lines 85-87, "temporal dependency" denotes correlations between different time intervals, while "spatial dependency" represents relationships across different channels. For SWaT, the spatial dependency are relationships between data in 51 channels.

Thank you once again for generously giving us your valuable time. We will make further enhancements to the presentation based on your insightful feedback.