##To Reviewer ByS8

Thank you so much for generously giving me your valuable time and advice. Apologies for the oversight, the code has been anonymously hosted at: github.com/yuesuoqingqiu/SensitiveHUE. Updated figures, experimental results can be found in the '**README.md**'.

The detailed replies will be provided as follows.

Deficiencies:

D1: The paper's structure is disorganized and challenging to read, with issues such as experimental results being prematurely introduced in the introduction and a lack of clear distinction between existing and proposed methodologies within the methodology section.

Answer1: Sorry for the lack of disorganization.

1. The confusion of the experimental results in the introduction: Apologies for the confusion. The purpose of this experimental results is to assess the sensitivity of the current reconstruction strategies towards normal patterns. In lines 103-125, we provided an overview of our experimental approach and results, demonstrating that complete disruption of either temporal or spatial relationships under existing reconstruction strategies had minimal impact on detection performance. It illustrates that the insensitivity of existing reconstruction methods to spatio-temporal dependency hampers their performance improvements.
2. The distinction of our method: As in lines 156-206, we briefly introduce our contribution and difference from existing methods. SensitiveHUE is a totally different probabilistic network, which belongs to the region of heteroscedastic uncertainty estimation (HUE). It utilizes probability loss to simultaneously achieve reconstruction and uncertainty estimation. Key differences from reconstruction methods include: (1) Loss function and framework: SensitiveHUE comprises two separate decoders for reconstruction and uncertainty estimation (Fig. 1). These two parts are jointly optimized through our MTS-NLL designed for anomaly detection task. This ensures the model can quantify the magnitude of noise, avoid the influence of noise in anomaly detection; (2) Reconstruction strategy: Statistical Feature Removal (SFR) is conducted in each input window. SensitiveHUE is aim to restore the original series before SFR, ensuring sensitivity to spatio-temporal dependencies.

D2: The document contains numerous grammatical and notational errors, further contributing to its overall difficulty to read.

Answer2: We have carefully corrected the errors in the text and figures. line 327, line 918, and the references pointed out in "redefine references" are due to the automatic formatting of the provided paper template, so they could not be changed.

D3: Critical concepts central to this research, such as "sensitivity" and "heteroscedastic uncertainty," are not defined in the abstract or introduction. This omission makes it difficult for readers, especially those not closely familiar with the field, to grasp what the study aims to address and resolve.

Answer3: Due to space constraints, we did not provide detailed explanations of sensitivity and heteroscedastic uncertainty in the abstract. In Lines 95-155, we focused on explaining how existing methods are insensitive to normal patterns through experiments and figures. The definition of sensitivity is specifically defined in Lines 282-338. We will add more background knowledge about Heteroscedastic Uncertainty Estimation (HUE) and the specific implementation of SensitiveHUE to related parts, along with clearer explanations of technical terms.

D4: The document frequently uses many specialized terminology without clear definitions, making it difficult to read and understand. It's essential not to neglect the effort to explain each term explicitly.

Answer4: We have clarified the explanations of technical terms, including the explanation of abbreviation, the definition of term, and so on

D5: The comparative analysis of methodological performance appears biased, with the proposed method's counterparts depicted as underperforming relative to their original benchmarks.

Answer5: We have checked all the results of comparison models and ensured consistency with the original results. For the unpublished results, we replicate the experiment using the provided code and fine-tune it according to the instructions. For the adjustments in Tab.3, please refer to '**README.md**'.

D6: The paper does not address existing critiques regarding the inability of certain datasets, like MSL and SMD, to provide a fair evaluation, as noted by [Wu & Keogh, 2021]. For instance, while the original paper of TranAD [Tuli+, 2022] cites F1 scores of 81.51 for SWaT, 49.51 for WADI, 94.94 for MSL, and 96.05 for SMD, your results in Table 3 show TranAD with significantly lower performance, suggesting an unfair comparison.

Answer6: For fairness, we employ MSL and SMD datasets for comparison, as most methods have been experimented on them. We will cite [Wu&Keogh, 2021] and point out the limitations of NASA dataset, such as the problem of excessive intensive annotation of outliers. The differences in results on TranAD are mainly due to different downsampling rates or

filtering of entities (note the F1 score in TranAD corresponds to $F1^*_{PA}$ in our paper).
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 $F1_{PA}$ scores and simplify them to $F1$, and only a few methods like GDN and NSPR provide the $F1$ score before adjustments. Note $F1_{PA}$ will overestimate the performance of a model [3], so we use both $F1$ and $F1_{PA}$ simultaneously to evaluate the model's performance. Obviously, we have achieved significant performance improvements, especially on the SWaT and WADI datasets.

Minor Corrections

C1: Punctuation related: Line 213, 240, 371, 389, 861, 909.
C2: Capitalization related: Line 281, 512, 850, 906, Figure 6.
C3: Italics related: Line 511, 693.
C4: abbreviation related: Line 210, 389, 671.
C5: Expression related: 208, 518, 665, 680, 689, 849.
C6: template related: 327, 918, Redefine reference

Answers: We are sorry for the minor errors in our presentation, all have been carefully corrected. For **C1**, we have added or removed some punctuations to ensure correctness, e.g., for Line 213, we have added a period to "in MTS" to make it "in MTS.". For **C2**, we have capitalized the corresponding characters, e.g., we have changed "Dependency-sensitive" to "Dependency sensitive.". For **C3**, we have removed the italics from “mean” and “F1” in Line 511 and 693 respectively. For **C4**, we have changed “SOTA” to “state-of-the-art”, “We'll” to “We will”, “Table 3” to “Tab. 3”. For **C5**, we have corrected the errors and added some extra prefixes, e.g., we have changed “republic” to “public” at Line 208 and have specified "anomaly scores" as "anomaly scores for SWaT datasets". For **C6**, they may due to the automatic formatting of paper template, so we did not change it.

Questions

Q1: Line 13: What is meant by "sensitivity"?

Answer1: Sensitivity refers to the model's sensitivity to normal patterns. In Lines 82-94, normal patterns include both spatiotemporal relationships and noise, while anomalies deviate from these patterns. For example, if a model is sufficiently sensitive, changes in spatiotemporal relationships should significantly impact its detection performance.

Q2: Line 20: What does "sensitivity modelling" refer to?

Answer2:, The "sensitive" denotes the model's ability to effectively learn normal patterns and maintain high sensitivity to them.

Q3: Line 22: What does SensitiveHUE stand for?

Answer3: SensitiveHUE emphasizes using heteroscedastic uncertainty estimation (HUE) in modeling, while the term 'sensitive' indicates that our model maintains a high sensitivity to normal patterns.

Q4: Line 25: Shouldn't there be an explanation of what NLL stands for?

Answer4: NLL refers to the negative log-likelihood and we have explained it in lines 418-422.

Q5: Line 47: What is the abbreviation CPS for?

Answer5: CPS stands for Cyber-Physical System.

Q6: Table 1: What does the bold text signify? It's advisable not to include experimental results in the introduction. The implementation methods for each strategy are inadequately explained. Mention of using the SWaT dataset should be included in the text. Besides the F1 score, the AUC PR should also be used as a metric.

Answer6: In time series anomaly detection, most studies use F1 score as the sole evaluation metric. To facilitate comparison with these studies, we use the F1 score, accompanied with precision and recall. Note that most studies only present the $\mathrm{F1}_{\mathrm{PA}}$ score after point ajustment, but for a rigorous evualtion, we use both $\mathrm{F1}$ and $\mathrm{F1}_{\mathrm{PA}}$ simultaneously to evaluate the model's performance. (Refer to Answer6 to C6).

For the implementation methods, sorry for the brevity in lines 104-120. We will add more details in main text. The experimental details are summarized as follows.

1. Details of reconstruction strategy: We adopted the original Transformer framework using existing reconstruction strategies individually (lines 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.
2. Details of disruption: Apology for the brevity in lines 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 ∞. 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.

Q7: Line 91: What is "sensitivity to normal patterns"? How is it different from “sensitivity to spatio-temporal dependencies”?

Answer7: In lines 82-94, normal patterns include both spatiotemporal relationships and noise, while anomalies deviate from these patterns. ‘Sensitivity to normal patterns’: The sensitivity of a model to normal patterns is demonstrated when it effectively captures spatiotemporal relationships in modeling and accurately identifies patterns of noise, distinguishing them from outliers.

Q8: Line 119: What does "points from the other channels" refer to?

Answer8: In lines 85-87, the relationship between different channels (such as sensors) is referred to as spatial relations. "Points from other channels" refers to data from other channels in multi-channel data. To make it clearer, we will use "data" instead of "points."

Q9: Line 126: What are "potential shortcuts"?

Answer9: This refers to the model's tendency to achieve good reconstruction results without relying on spatiotemporal relationships (i.e., by depending on only one of them or none at all).

Q10: Line 136: What is meant by "normal noise"? Does it not include anomalies?

Answer10: As described in lines 88-91, noise is inherent in the normal pattern of MTS, whereas abnormality is a complete deviation from the normal pattern.

Q11: Line 147: The sentence starting with "However" is too jargon-heavy for a first-time reader to understand.

Answer11: Sorry, we will try to express it in a more friendly and readable way.

Q12: Figure 1: Which method's results are shown? Is the reconstruction error squared or

absolute? Can lowering the threshold in this example detect anomalies?

Answer12: The results in Figure 1 are from our SensitiveHUE model using MSE as the loss function. The reconstruction error is the squared error. We chose the threshold that maximizes F1, although lowering the threshold can detect anomalies, it may also lead to more false positives.

Q13: Line 194: What is SensitiveHUE short for?

Answer13: refer to the Answer3 to Q3.

Q14: Line 333: In Table 2, the error changes slightly before and after the disruption, so doesn't it technically not satisfy Definition 3.2?

Answer14: Using the compression strategy as an example, the minimal deviation observed in the reconstruction results after disrupting temporal and spatial relations indicates their independence in both temporal and spatial perspectives (Def. 3.1). They may not meet the 'zero deviation' of being dependency sensitive (Def. 3.2). In practice, due to network architecture, it's hard to fully meet Def. 3.1. For instance, with attention mechanisms, the likelihood of the network completely ignores other time points (with a diagonal attention matrix) is very low. In general, in the existing network design, it is almost impossible to get a completely zero deviation after being disrupted. If the deviation is small and the disruption hardly affects the performance of anomaly detection, the strategy is considered to be approximately dependency insensitive. We will add 'approximately' in Line 314 and 321.

Q15: Theorem 3.3: Is it correct to understand that g represents the encoder, h the decoder, and f the entire autoencoder? Would a more intuitive explanation help the reader?

Answer15: $f: X \mapsto h(X^g)$ 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, the roles of encoder and decoder do not apply to them.

Q16: Line 357, Line 388: Change "satisfies Theorem A.1 well." to "is dependency sensitive." The conclusion should be understandable without needing to refer to the appendix.

Answer16: We have corrected these inaccuracies. We have replaced 'satisfies Theorem A.1 well' with 'is dependency sensitive'.

Q17: Line 379: Isn't it unfair to directly compare the values of deviation after disruption when

the Original MSE sizes are significantly different between the proposed method and the comparison group? Shouldn't the rate of change be compared?

Answer17: Thank you for your suggestion. Using the rate of change indeed provides a more intuitive understanding of the magnitude of changes. However, when the deviation itself is small, the magnitude may not reflect its impact on anomaly detection performance, as seen in the results in Table 1, where the performance are nearly unchanged. We provide the deviation primarily to illustrate that after disrupting the relations, the model's reconstruction results may hardly change, leading to the small variation in F1 scores. We have incorporated the rate of change into Table 2 as follows.

Strategy	Original MSE	Temporal disruption		Spatial disruption	
		Deviation	Change	Deviation	Change
Compression	0.0040	0.0004	10.0%	0.0091	227.5%
Mask	0.0364	0.8804	2418.7%	0.0768	211.0%
Ours	0.1012	21.18	**20928.9%**	0.6727	**664.7%**

Table 2: The deviation of restored results under different reconstruction strategies by disrupting spatial or temporal dependency in SWaT [21] dataset.

Q18: Line 381: Since statistical feature reconstruction treats each channel independently, isn't spatial dependency considered?

Answer18: The statistical feature removal strategy is a simple operation on input windows, aiming at removing statistical information from each individual sequence within the window. This forces the model to rely on spatio-temporal relationships to reconstruct the original values before the statistical features were removed. Specifically, in MTS, this removed information must be inferred simultaneously through both temporal and spatial dependencies.

Q19: Line 388: Is "$f: h(X) \rightarrow X$" truly the correct notation?

Answer19: Sorry for the mistake. We have changed $f: h(X) \mapsto X$ to $f: X \mapsto h(X^g)$.

Q20: Line 413: What are the definitions of the adorned μ and σ ?

Answer20: The adorned μ and σ represent the estimations of the mean and variance in the predictive Gaussian distribution (line 109).

Q21: Line 445: What is the definition of the stop gradient operation?

Answer21: The "stop gradient" operation refers to elements that do not participate in gradient backpropagation and are treated solely as coefficients (the operation of 'detach').

Q22: Line 453: If $\beta < \frac{1}{2}$, then aren't well-fit points overemphasized (assuming the assumption in Line 433)?

Answer22: For the β-NLL loss, its gradients with respect to $\hat{\mu}_{ts}$ and $\hat{\sigma}_{ts}$ are as follows:

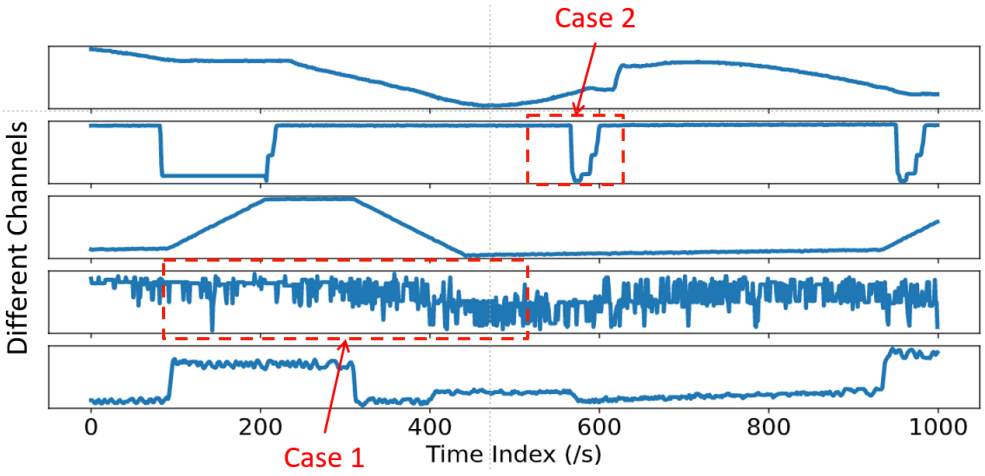
$$\nabla_{\hat{\mu}_{ts}} \mathcal{L}_{\beta-\mathrm{NLL}} = \frac{\hat{\mu}_{ts} - X_{ts}}{\hat{\sigma}_{ts}^{2-2\beta}},$$

$$\nabla_{\hat{\sigma}_{ts}^2} \mathcal{L}_{\beta-\mathrm{NLL}} = \frac{\hat{\sigma}_{ts}^2 - (\hat{\mu}_{ts} - X_{ts})^2}{2(\hat{\sigma}_{ts}^2)^{(2-\beta)}}.$$

Clearly, the gradient scales the error $\hat{\mu}_{ts} - X_{ts}$ by the factor $\hat{\sigma}_{ts}^{2-2\beta}$. When $0 < \beta < 1$, compared to NLL, as β increases, the excessive emphasis on easily fitting points gradually decreases. Specifically, the gradient of the mean converges to $\left(\frac{1}{\hat{\mu}_{ts} - X_{ts}}\right)^{1-2\beta}$ when the variance stabilizes at its stand point $\hat{\sigma}_{ts}^2 = (\hat{\mu}_{ts} - X_{ts})^2$. Therefore, theoretically, if the variance reaches the stable point, when $\beta > \frac{1}{2}$, the well-fit points will no longer be overemphasized.

Q23: Line 485: The explanations for case 1 and case 2 are too abstract. Without sufficient explanation using equations, figures, or concrete examples, it's challenging to convey what the issue is with the conventional methods.

Answer23: Case 1 denotes noises with low signal-to-noise ratios, while Case 2 refers to areas that are difficult to reconstruct aside from noise. We provide a demonstration from SWaT data for this purpose, as follows:



Q24: Line 548: Can you explain using an equation?

Answer24: For clarity, we revised the phrase 'This ensures that for the data from each channel, $\mathcal{L}_{\mathrm{MTS-NLL}}$ has roughly ⋯ during different periods' to 'This ensures that in one batch, for the data from each channel, the gradients of means scales the error $\hat{\mu}_{ts} - X_{ts}$ by an identical factor $\hat{\sigma}_{s,\mathrm{mean}}^{2\alpha}$ in different time intervals'.

Q25: Line 572: What does "the model performance is monitored" mean? Is it referring to the anomaly score?

Answer25: Yes, the model's performance in anomaly detection is assessed using the original NLL as the anomaly score, while the MTS-NLL loss is only utilized in training for a better balance between different regions.

Q26: Line 578: It's confusing to denote the number of channels with S as well. Experiments: Ensure that the experiments section is somewhat readable even for readers who haven't thoroughly read the Methodology section.

Answer26: Thanks for your suggestions. We will change it to a more readable way.

Q27: Line 659: Mentions of precision and recall are missing. Does Table 3 include their maximum values when changing the threshold?

Answer27: We have added the Precision and Recall to Line 660. Yes, we search for the threshold corresponding to the best F1, details are shown in Line 1239-1245.

Q28: Figure 3: Isn't this evaluation possible with AUC PR without the need for such visualization?

Answer28: Although using AUC PR can reflect the issue illustrated in Figure 3, it may not be as intuitive. Here, using the distribution of anomaly scores can demonstrate the distinguishability between normal and anomalous instances in a more clear way, as seen in similar visualizations in other papers [1].

[1] Chih-Yu Lai, Fan-Keng Sun, Zhengqi Gao, Jeffrey Lang, and Duane S. Boning. 2023. Nominality Score Conditioned Time Series Anomaly Detection by Point/Sequential Reconstruction. In Proceedings of the 37th International Conference on Neural Information Processing Systems.

Q29: Line 799: What does "clear shortcut in this way" mean?

Answer29: The "clear shortcut" here refers to the identical shortcut, where the network tends to copy the input, and no meaningful representations can be learned (as discussed in [2]). This strategy is insensitive to dependencies.

[2] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Le Xinyi. 2022. A Unified Model for Multi-class Anomaly Detection. In Proceedings of the 36th International Conference on Neural Information Processing Systems.

Q30: Figure 4: Why are the results for MSL and SMD omitted?

Answer30: Due to space limitations, we only present results for two datasets to illustrate the impact of parameters on model performance. We will add the results for other datasets in the Appendix.

Q31: Line 855: What does "leading to constant mapping" mean?

Answer31: When the window size is too large, the statistical features $\{\mathbb{E}_s(X), \mathrm{Var}_s(X)\}$ of long-term series may remain unchanged. For each channel, the constant mapping denotes $X_{:s}^{g} \cdot \sqrt{\mathrm{Var}_s(X) + \epsilon_0} + \mathbb{E}_s(X)$, allowing for perfect reconstruction of all inputs.

Q32: Figure 5: Does "the estimated uncertainty" refer to $\hat{\sigma}$? How was the threshold for the bottom panels determined?

Answer32: Yes, the estimated uncertainty is indeed $\hat{\sigma}$. We select the optimal threshold to achieve the highest possible F1 score, as detailed in Line 1239~1245.

Q33: Figure 6: What variables in Section 3 correspond to "the learned correlations"? Why is it not a symmetric matrix if it represents correlations?

Answer33: Each variable denotes one channel, and we select the first 34 channels in SWaT dataset for visualization. The approach to get the spatial correlations is shown in Line 1373-1392, note it is not based on the correlation coefficient (e.g. Pearson), but intervention that is widely used in model interpretability and causal discovery. Indeed, the correlation matrix is not symmetric in MTS, as noted in GDN [3].

[3] Ailin Deng and Bryan Hooi. 2021. Graph neural network-based anomaly detection in multivariate time series. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. 4027–4035.

Thank you once again for graciously offering us your precious time. We will incorporate your valuable feedback to further improve our work.