Journal of Machine Learning Research 27 (2026) 1-43 Submitted 7/25; Revised 2/26; Published 3/26

Statistical Test for Attention in Transformers for Images and Time Series

Tomohiro Shiraishi SHIRAISHI.TOMOHIRO.NAGOYAML@GMAIL.COM
Nagoya University and RIKEN, Nagoya, Aichi 464-8601, Japan

Daiki Miwa MIWA.DAIKI.MLLAB.NIT@GMAIL.COM
Nagoya University, Nagoya, Aichi 464-8601, Japan

Teruyuki Katsuoka KATSUOKA.TERUYUKI.NAGOYAML@GMAIL.COM
Nagoya University, Nagoya, Aichi 464-8601, Japan

Vo Nguyen Le Duy DUY.MLLAB.NIT@GMAIL.COM
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam, and RIKEN

Shuichi Nishino NISHINO.SHUICHI.NAGOYAML@GMAIL.COM
Nagoya University and RIKEN, Nagoya, Aichi 464-8601, Japan

Kouichi Taji TAJI@NAGOYA-U.JP
Nagoya University, Nagoya, Aichi 464-8601, Japan

Ichiro Takeuchi TAKEUCHI.ICHIRO.N6@F.MAIL.NAGOYA-U.AC.JP
Nagoya University and RIKEN, Nagoya, Aichi 464-8601, Japan

Editor: Christian Shelton

Abstract

Transformer models have achieved exceptional performance in various domains, including computer vision and time-series analysis. Their core attention mechanism is widely used to interpret model decisions by assigning importance weights to input regions, such as image patches or time series intervals. However, the reliability of these interpretations remains a major concern. High-attention weights do not necessarily indicate genuinely significant features; they may instead be artifacts of the model's computation, undermining their reliabilities in high-stakes applications such as medical diagnostics. To address this, we propose a novel statistical framework designed to quantify the significance of high-attention regions in Transformer models. Our framework is built on selective inference (SI) to correct for the inherent selection bias that arises from testing regions chosen through the complex attention computation of the Transformer models. A key contribution of this work is a novel computational method that extends SI to the complex non-linearity of self-attention, enabling the computation of valid p-values for high-attention regions. These p-values serve as a reliable measure of significance, strengthening the interpretability of Transformer decisions. The validity and effectiveness of our approach are demonstrated through numerical experiments and applications to brain image diagnosis and electroencephalography (EEG) data analysis.

Keywords: statistical test, transformer, attention mechanism, selective inference, medical data analysis

1. Introduction

Transformer architectures have achieved remarkable success across a wide range of domains, including natural language processing, computer vision, and time-series modeling (Lin et al., 2022). While early research primarily focused on language models (Vaswani et al., 2017; Devlin et al., 2019), recent developments have shown that Transformers are also highly effective for modeling structured, continuous data such as images (Dosovitskiy et al., 2020; Liu et al., 2021; Khan et al., 2022) and time-series signals (Lim et al., 2021; Zhou et al., 2021; Wen et al., 2023). In this study, we focus on Transformers applied to these continuous data modalities—specifically in the vision and signal domains—with a particular emphasis on high-stakes fields such as medical applications, where interpretability and reliability are of paramount importance.

A central component of the Transformer is the attention mechanism, which dynamically weights the importance of input elements. In addition to improving performance, attention offers insight into the model's decision-making process (Xu et al., 2015; Choi et al., 2016). In vision and signal applications, attention maps—derived from attention weights—are commonly used to identify which spatial or temporal regions the model considers important (Dosovitskiy et al., 2020; Lim et al., 2021). For example, attention maps can highlight diagnostically relevant areas in medical images (Ferdous et al., 2023; Hong et al., 2024), or anomalous segments in physiological signals such as EEG or ECG (Xie et al., 2022; Hu et al., 2023). This interpretability is particularly valuable in healthcare, where understanding the model's rationale is critical for trust and clinical adoption (Tonekaboni et al., 2019). Consequently, attention visualization techniques are now widely used in medical image diagnosis and physiological signal analysis.

Despite their intuitive appeal, attention maps should be interpreted with caution. Several studies have shown that high attention weights do not necessarily correspond to causally important features, and that attention can be influenced by noise or spurious correlations (Jain and Wallace, 2019; Bibal et al., 2022). These observations raise concerns about the reliability of attention-based interpretations—particularly whether high-attention regions genuinely reflect meaningful signals or merely result from internal artifacts of the model. To answer this question, we consider a statistical framework for evaluating attention reliability, aiming to support more trustworthy interpretability of Transformer decisions. Specifically, we propose a method to assess the *statistical significance* of attention in the form of p-values that can be used to test whether high-attention regions differ significantly from the rest of the input.

Unfortunately, classical hypothesis testing methods are not directly applicable in this setting. At first glance, these regions often appear different from the rest but this is expected, as the Transformer model selects them based on the input itself. This introduces a selection bias: we are testing regions that have already been chosen precisely because they appear distinctive. As a result, simply observing differences does not guarantee that those regions are genuinely important—they may merely reflect noise or artifacts induced by the model. To address this issue, we propose a statistical test that explicitly accounts for this selection bias, enabling us to evaluate whether the observed differences are statistically significant or simply a byproduct of the attention mechanism.

To this end, we adopt the framework of selective inference (SI) (Lee et al., 2016; Taylor and Tibshirani, 2015), a statistical methodology for valid inference following data-driven hypothesis selection. SI is well suited to our setting, where high-attention regions are chosen based on the model's internal computations. By conditioning on the fact that attention values are model-derived, we enable valid statistical testing even in finite-sample regimes. A key technical contribution of this work is a new computational method for applying SI to Transformer models. Unlike prior SI methods that assume the selection event is defined by linear or quadratic constraints, Transformer self-attention involves complex nonlinear operations. To address this, we develop an approximation strategy that extends SI to this setting. Our approach broadens the applicability of SI to modern neural architectures and offers a principled way to evaluate the interpretability of Transformer attention mechanisms.

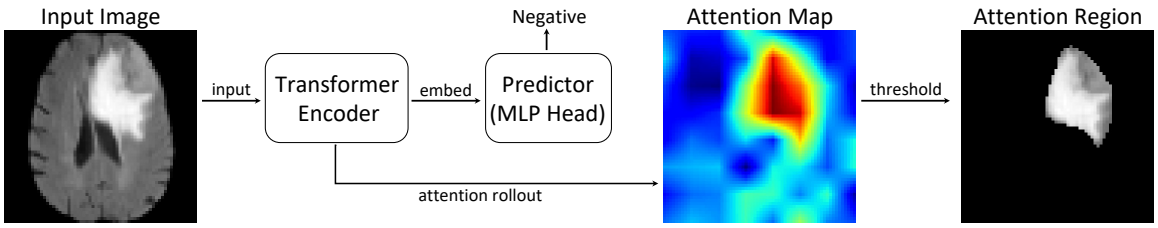
Related Works. It is widely believed that attention mechanisms provide interpretable explanations for model outputs (Xu et al., 2015; Choi et al., 2016; Martins and Astudillo, 2016; Xie et al., 2017; Mullenbach et al., 2018), based on the implicit assumption that tokens receiving higher attention weights play a more substantial role in the final prediction. However, Jain and Wallace (2019) raised concerns about this assumption, pointing out its lack of formal evaluation. They argued that for attention weights to serve as faithful explanations, they must correlate with established feature importance metrics and that counterfactual modifications of these weights should lead to different outcomes. Their experiments empirically demonstrated that neither of these conditions consistently holds, leading them to conclude that attention is not a faithful explanation for model predictions. On the other hand, Wiegreffe and Pinter (2019) argued that these criteria might be overly stringent, proposing alternative evaluation protocols and reasserting the utility of attention for explanation. This debate on the faithfulness of attention has spurred a significant body of subsequent research, including theoretical studies on factors that impair faithfulness (Bai et al., 2021; Brunner et al., 2020; Sun and Lu, 2020; Tutek and Šnajder, 2020) and extensive work on quantitative evaluation metrics (Liu et al., 2020; Ju et al., 2022; Madsen et al., 2022). Furthermore, various approaches have been proposed to enhance attention faithfulness by intervening in the training process or model architecture, such as improving hidden state representations (Chrysostomou and Aletras, 2021; Mohankumar et al., 2020), incorporating regularization terms into the objective function (Tutek and Šnajder, 2020; Moradi et al., 2021), and enforcing sparsity (Zhang et al., 2021; Meister et al., 2021). While all these studies primarily investigate the relationship between the model's final output and its attention mechanism, a definitive gold-standard evaluation metric has yet to be established. In this work, we introduce a new approach to quantifying the reliability of attention in terms of statistical significance, focusing specifically on whether high-attention components (e.g., patches in images, intervals in time series) genuinely contain meaningful and distinctive signals.

Selective Inference (SI), also known as post-selection inference, provides a framework for valid statistical inference on data-driven hypotheses. The central idea is to correct for selection bias by conditioning on the event that a particular hypothesis was selected based on the data. This approach allows for the calculation of p-values and confidence intervals that maintain the nominal significance level, thereby ensuring the reliability of statistical conclusions even when the same data is used for both hypothesis selection and testing. SI

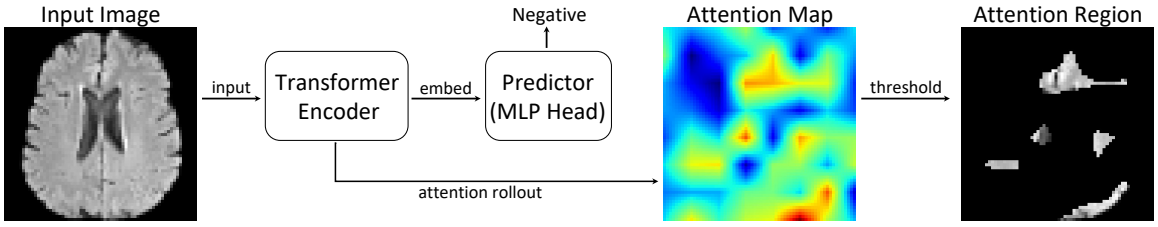
initially attracted substantial interest as a method for inference after feature selection in linear regression models. A seminal work by Lee et al. (2016) introduced an exact (non-asymptotic) inference procedure for the regression coefficients of features selected by the Lasso. Building on this foundation, the scope of SI has been extended to various other feature selection methods, such as marginal screening (Lee and Taylor, 2014), stepwise feature selection (Tibshirani et al., 2016), generalized linear models (Taylor and Tibshirani, 2018), among many others (Loftus and Taylor, 2015; Loftus, 2015; Yang et al., 2016; Charkhi and Claeskens, 2018; Hyun et al., 2018; Sugiyama et al., 2021; Rügamer and Greven, 2020; Zhao et al., 2022; Nguyen et al., 2025).

Concurrently, substantial theoretical and computational advancements have been made to improve its statistical power (Fithian et al., 2014; Tian and Taylor, 2018; Panigrahi et al., 2021, 2023; Terada and Shimodaira, 2017; Liu et al., 2018; Duy and Takeuchi, 2022; Hyun et al., 2021; Jewell et al., 2022; Carrington and Fearnhead, 2025). The application of SI has also been broadened beyond feature selection to encompass a diverse range of machine learning domains. These include clustering (Lee et al., 2015; Gao et al., 2022; Chen and Witten, 2023; Yun and Foygel Barber, 2023), principal component analysis (Choi et al., 2017; Perry et al., 2025), change-point detection (Hyun et al., 2018, 2021; Duy et al., 2020; Jewell et al., 2022), decision trees (Neufeld et al., 2022; Bakshi et al., 2024), deep neural networks (Duy et al., 2022; Miwa et al., 2023), among many others (Suzumura et al., 2017; Das et al., 2022; Yamada et al., 2018; Chen and Bien, 2020; Tsukurimichi et al., 2022; Rügamer et al., 2022; Duy and Takeuchi, 2025). In the context of deep learning, the work of Duy et al. (2022) and Miwa et al. (2023) pioneered the application of SI to neural networks. They observed that a class of models, such as Convolutional Neural Networks (CNNs), can be characterized as piecewise linear functions. This property allows the selection event in neural networks to be described by a set of linear constraints. Subsequent research has built upon their algorithm to extend SI to a variety of other deep learning models (Shiraishi et al., 2024b; Miwa et al., 2024; Katsuoka et al., 2024; Nishino et al., 2025). However, due to the non-linearity of the self-attention mechanism (e.g., the inner product of queries and keys, and a smooth activation function such as GELU), Transformer-based architectures cannot be represented as piecewise-linear functions, which makes the above specific existing algorithmic approach inapplicable. While similar challenges arise in adaptive regularized estimation approaches, for which selective inference methods have been proposed by Pirenne and Claeskens (2024), these methods are akin to fixed-width grid searches. Such approaches can be computationally prohibitive or limited in approximation accuracy when applied to the complex architecture of attention mechanisms. Therefore, a more efficient strategy is required. In this study, we introduce a new computational approach to develop SI for attention in Transformers.

Demonstration. Figure 1 illustrates the problem setup considered in this study, where we applied a naive statistical test, which does not consider selection bias, and our proposed statistical test to a brain image diagnosis task. The upper panel shows a brain image with a tumor region, where attention on the tumor region is expected to be declared as statistically significant (with a small p-value). Here, both the naive test and the proposed test conclude that the identified attention is statistically significant with p-values nearly 0. In contrast, the lower panel displays a brain image without a tumor region, where attention is expected



(a) Brain image with tumor. The naive p-value is 0.000 (true positive) and the selective p-value is 0.000 (true positive).



(b) Brain image without tumor. The naive p-value is 0.000 (false positive) and the selective p-value is 0.875 (true negative).

Figure 1: Schematic illustration of the problem setup and the proposed method on a brain image data set. By inputting a brain image into the trained Transformer classifier, the attention map is obtained, which indicates the area on which the attention mechanism focuses. Our objective is to provide the statistical significance of the attention map using the p-value. To achieve this objective, we consider testing the attention region, which consists of pixels with high attention levels by thresholding the attention map. The results suggest that the naive p-value (see Section 4) cannot be used to properly control the false positive (type I error) rate. Instead, the selective p-value (introduced in Section 2) can be used to detect true positives while controlling the false positive rate at the specified level.

to be determined as statistically non-significant (with a large p-value). In this case, the naive test falsely detects significance (false positive) with an almost zero p-value, while the proposed method yields a p-value of 0.875, concluding that it is not statistically significant (true negative).

Contributions. This paper significantly extends our preliminary work presented at International Conference on Machine Learning (ICML) 2024 (Shiraishi et al., 2024a). While the conference version was limited to a one-sample test for Vision Transformers, this work generalizes the framework to arbitrary continuous-valued inputs (including time series), introduces new hypothesis testing variants (e.g., two-sample tests), and provides a more applicable and optimized implementation as a Python package to facilitate practical use. Specifically, our main contributions are summarized as follows:

- We introduce a theoretically guaranteed framework to quantify the statistical significance of attention in Transformers, with a specific focus on image and time series data (see Section 2).

- We develop a selective inference method for attention in Transformers, introducing a novel computational approach to compute p-values without selection bias (see Section 3).
- We demonstrate the effectiveness of our proposed method through applications to synthetic data simulations, as well as real-world tasks such as brain image diagnosis using MRI data and electroencephalography (EEG) data analysis (see Section 4).

An implementation of our method for PyTorch-defined Transformer models is available as the Python package `si4attention` at `https://pypi.org/project/si4attention/`. For reproducibility, the experimental code is available at `https://github.com/shirara1016/statistical_test_for_attention_in_transformers`.

2. Statistical Test for Attention in Transformers

This study focuses on evaluating the statistical significance of attention in Transformers applied to image and time-series data. We consider the setting where a Transformer model has already been trained on a labeled data set—such as brain images for diagnosis or EEG signals for neurological assessment. Given a trained Transformer model, our goal is to assess the attention maps generated when a new test input is provided—for example, a brain scan or EEG recording from a new patient. Specifically, we test whether the regions highlighted by the attention truly contain meaningful and distinctive information. This is important because while attention values may appear focused, it remains unclear whether they correspond to informative patterns or are influenced by noise or model artifacts. Our statistical framework allows this question to be tested rigorously, providing trustworthy insights into the interpretability of Transformers applied to real-world applications. The details of the architecture of the Transformer model employed in our experiments are provided in Appendix A.1.

2.1 Assumptions and Notations

Let us denote the input image with n pixels or the input time series with n time points as $\boldsymbol{x} \in \mathbb{R}^n$. We assume that the input data $\boldsymbol{x}$ is a realization of the following random vector

$$\boldsymbol{X} = (X_1, \ldots, X_n)^\top = \boldsymbol{\mu} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ is the true underlying value vector, which corresponds to the pixel intensities for image data or signal intensities for time series data, and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the noise vector with covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$. We do not pose any assumption on the true value vector $\boldsymbol{\mu}$, while we assume that the noise vector $\boldsymbol{\epsilon}$ follows the Gaussian distribution with the covariance matrix Σ known or estimable from external independent data.[1]

We define the computation of the attention map as a mapping $\mathcal{A}\colon \mathbb{R}^n \ni \boldsymbol{x} \mapsto \mathcal{A}(\boldsymbol{x}) \in [0,1]^n$, which takes an image or a times-series $\boldsymbol{x}$ as input and outputs attention scores $\mathcal{A}_i(\boldsymbol{x}) \in [0,1]$ for each pixel or each time point $i \in [n]$. The details of its computation are provided in Appendix A.2. We then define the attention region $\mathcal{M}_{\boldsymbol{x}}$ of an input data $\boldsymbol{x}$ as

1. We discuss the robustness of our proposed method when the covariance matrix is unknown and the noise deviates from the Gaussian distribution in Appendix F.

the set of its constituent elements (pixels for image data or time points for time series data) that have attention scores greater than a pre-specified[2] threshold value $\tau \in (0, 1)$, i.e.,

$$\mathcal{M}_{\boldsymbol{x}} = \{i \in [n] \mid \mathcal{A}_i(\boldsymbol{x}) > \tau\}. \tag{1}$$

2.2 Formulation of Statistical Test

Hypotheses and Test Statistic. To quantify the statistical significance of the attention region $\mathcal{M}_{\boldsymbol{x}}$ of an input data $\boldsymbol{x}$, we propose the following statistical hypothesis test problem:

$$\mathrm{H}_0 \colon \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|} \sum_{i \in \mathcal{M}_{\boldsymbol{x}}} \mu_i = \frac{1}{|\mathcal{M}_{\boldsymbol{x}}^c|} \sum_{i \notin \mathcal{M}_{\boldsymbol{x}}} \mu_i \quad \text{vs.} \quad \mathrm{H}_1 \colon \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|} \sum_{i \in \mathcal{M}_{\boldsymbol{x}}} \mu_i \neq \frac{1}{|\mathcal{M}_{\boldsymbol{x}}^c|} \sum_{i \notin \mathcal{M}_{\boldsymbol{x}}} \mu_i, \tag{2}$$

where H_0 is the null hypothesis that the mean true value inside and outside the attention region are equal, while H_1 is the alternative hypothesis that they are not equal. To conduct the statistical test in (2), a natural choice for the test statistic is the difference in the mean values between inside and outside the attention region, i.e.,

$$\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^\top \boldsymbol{X} = \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|} \sum_{i \in \mathcal{M}_{\boldsymbol{x}}} X_i - \frac{1}{|\mathcal{M}_{\boldsymbol{x}}^c|} \sum_{i \notin \mathcal{M}_{\boldsymbol{x}}} X_i,$$

where $\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}} = \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|} \mathbf{1}^n_{\mathcal{M}_{\boldsymbol{x}}} - \frac{1}{|\mathcal{M}_{\boldsymbol{x}}^c|} \mathbf{1}^n_{\mathcal{M}_{\boldsymbol{x}}^c}$ is a vector that depends on the attention region $\mathcal{M}_{\boldsymbol{x}}$, and $\mathbf{1}^n_{\mathcal{C}} \in \mathbb{R}^n$ is an n-dimensional indicator vector whose elements are set to 1 if they belong to the set $\mathcal{C} \subset [n]$, and 0 otherwise. In this paper, for simplicity, we employ the following standardized test statistic without loss of generality:

$$T(\boldsymbol{X}) = \frac{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^\top \boldsymbol{X}}{\sqrt{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^\top \Sigma \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}}}. \tag{3}$$

The p-value for the statistical test in (2) can be used to quantify the statistical significance of the attention region $\mathcal{M}_{\boldsymbol{x}}$. Given a significance level $\alpha \in (0, 1)$, e.g., 0.05, we reject the null hypothesis H_0 if the p-value is less than α, indicating that the attention region $\mathcal{M}_{\boldsymbol{x}}$ is significantly different from the outside of the attention region. Otherwise, we fail to state that the attention region $\mathcal{M}_{\boldsymbol{x}}$ is statistically significant.

In the statistical test in (2), the statistical significance of the attention region is quantified by evaluating whether the elements (pixels for image data or time points for time series data) in the attention region selected by the Transformer model differ significantly from those that were not selected. While the above formulation focuses on the difference in mean true values for simplicity, similar approaches can be applied to other features derived using appropriate filters. Other formulations are also possible (e.g., significance can be evaluated using reference data), and these are discussed in detail in Appendix B. In cases where multiple features are tested simultaneously (e.g., testing both mean intensity and another feature), the family-wise error rate (FWER) can be controlled by applying standard multiple testing corrections, such as the Bonferroni method, after performing the valid hypothesis tests proposed in this work for each feature.

2. The choice of τ can be determined based on domain knowledge or specific application requirements. We provide an extended discussion on a data-driven approach for selecting τ in Appendix C.

Conditional Distribution. To compute the p-value, we need to identify the sampling distribution of the test statistic $T(\boldsymbol{X})$ defined in (3). However, as the vector $\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}$ depends on the attention region $\mathcal{M}_{\boldsymbol{x}}$ (i.e., depends on $\boldsymbol{x}$ through a complicated computation in the Transformer model), it is necessary to consider how the data influenced the attention region through the complicated process of the Transformer model to derive the sampling distribution of the test statistic $T(\boldsymbol{X})$. To overcome this issue, we consider the conditional sampling distribution of the test statistic $T(\boldsymbol{X})$ given the event $\{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}\}$, i.e.,

$$T(\boldsymbol{X}) \mid \{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}\}. \tag{4}$$

This conditioning means that we consider the rarity of the observation (input data) $\boldsymbol{x}$ only in the case where the same attention region $\mathcal{M}_{\boldsymbol{X}}$ as observed $\mathcal{M}_{\boldsymbol{x}}$ is obtained. The advantage of considering the conditional sampling distribution in (4) is that, by conditioning on the attention region, the vector $\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}$ can be treated as a constant vector, and the test statistic $T(\boldsymbol{X})$ becomes a linear function of $\boldsymbol{X}$, which allows us to characterize the sampling distribution of the test statistic $T(\boldsymbol{X})$.

Selective p-value. Statistical hypothesis testing based on conditional sampling distributions has been studied within the framework of SI, also known as post-selection inference. In this study, we utilize the SI framework to perform the statistical hypothesis test defined in (2), based on the conditional sampling distribution specified in (4). For the tractable computation of the conditional sampling distribution in (4), we introduce an additional condition on the sufficient statistic for the nuisance parameter $\mathcal{Q}_{\boldsymbol{X}}$, defined as

$$\mathcal{Q}_{\boldsymbol{X}} = \left(I_n - \frac{\Sigma \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}} \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^{\top}}{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^{\top} \Sigma \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}} \right) \boldsymbol{X}.$$

This additional conditioning on $\mathcal{Q}_{\boldsymbol{X}}$ is a standard practice in the SI literature required for computational tractability.[3] The selective p-value is then defined as

$$p_{\text{selective}} = \mathbb{P}_{\text{H}_0}(|T(\boldsymbol{X})| > |T(\boldsymbol{x})| \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}, \mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}}). \tag{5}$$

Theorem 1 *The selective p-value in* (5) *satisfies the following property of a valid p-value:*

$$\mathbb{P}_{\text{H}_0}\left(p_{\text{selective}} \leq \alpha\right) = \alpha, \ \forall \alpha \in (0, 1).$$

The proof of Theorem 1 is provided in Appendix G.1. This theorem guarantees that the selective p-value is uniformly distributed under the null hypothesis H_0, which is the defining property of a valid measure of statistical significance. This uniformity allows for valid statistical inference on the attention region $\mathcal{M}_{\boldsymbol{x}}$. Furthermore, this framework supports interval estimation of the parameter of interest (e.g., the difference in mean values) by inverting the proposed statistical test. We provide the detailed formulation and experimental evaluation of these confidence intervals in Appendix D.

3. Computation of Selective p-values

In this section, we propose a novel computational procedure for the selective p-values in (5).

3. The nuisance component $\mathcal{Q}_{\boldsymbol{X}}$ corresponds to the component $\boldsymbol{z}$ in the seminal paper by Lee et al. (2016) (see Sec. 5, Eq. (5.2), and Theorem 5.2) and is used in nearly all SI-related works cited in this paper.

3.1 Characterization of Conditional Distribution

To compute the selective p-values in (5), we need to characterize the following conditional distribution of the test statistic:

$$T(\boldsymbol{X}) \mid \{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}, \mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}}\}. \tag{6}$$

Due to conditioning on the sufficient statistic for the nuisance parameter $\mathcal{Q}_{\boldsymbol{X}}$, the event $\{\boldsymbol{X} \in \mathbb{R}^n \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}, \mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}}\}$ is restricted to a one-dimensional line within $\mathbb{R}^n$. Consequently, the conditional distribution of the test statistic $T(\boldsymbol{X})$ given this event is a truncated normal distribution defined on this one-dimensional event.

Theorem 2 *Under the null hypothesis* H_0 *defined in* (2)*, the conditional distribution of the test statistic defined in* (6) *is a truncated standard normal distribution, with the truncated region* $\mathcal{Z}$ *defined as*

$$\mathcal{Z} = \{z \in \mathbb{R} \mid \mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{M}_{\boldsymbol{x}}\},\ \boldsymbol{a} = \mathcal{Q}_{\boldsymbol{x}},\ \boldsymbol{b} = (\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^{\top}\Sigma\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}})^{-1/2}\Sigma\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}.$$

The proof of Theorem 2 is provided in Appendix G.2. Note that, from the definition of the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, the observed value of the test statistic $T(\boldsymbol{x})$ satisfies $\boldsymbol{x} = \boldsymbol{a} + \boldsymbol{b}T(\boldsymbol{x})$ and thus belongs to the truncated region $\mathcal{Z}$. According to Theorem 2, once the truncated region $\mathcal{Z}$ is identified, the selective p-value in (5) can be immediately computed based on the tail probability of the truncated standard normal distribution, i.e.,

$$p_{\text{selective}} = \mathbb{P}(|Z| > |T(\boldsymbol{x})| \mid Z \in \mathcal{Z}), \text{ where } Z \sim \mathcal{N}(0,1).$$

Thus, the remaining task is reduced to the characterization of $\mathcal{Z}$. Based on the definition of the attention region in (1), the condition part of the truncated region $\mathcal{Z}$ defined in Theorem 2 can be reformulated as

$$\begin{aligned}
\mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{M}_{\boldsymbol{x}} \Leftrightarrow & \{i \in [n] \mid \mathcal{A}_i(\boldsymbol{a}+\boldsymbol{b}z) > \tau\} = \mathcal{M}_{\boldsymbol{x}} \\
\Leftrightarrow & \begin{cases} \mathcal{A}_i(\boldsymbol{a}+\boldsymbol{b}z) > \tau,\ \forall i \in \mathcal{M}_{\boldsymbol{x}} \\ \mathcal{A}_i(\boldsymbol{a}+\boldsymbol{b}z) < \tau,\ \forall i \notin \mathcal{M}_{\boldsymbol{x}} \end{cases} \\
\Leftrightarrow & f_i(z) < 0,\ \forall i \in [n],
\end{aligned}$$

where $f_i \colon \mathbb{R} \to \mathbb{R},\ i \in [n]$ is defined as

$$f_i(z) = \begin{cases} \tau - \mathcal{A}_i(\boldsymbol{a}+\boldsymbol{b}z) & (i \in \mathcal{M}_{\boldsymbol{x}}) \\ \mathcal{A}_i(\boldsymbol{a}+\boldsymbol{b}z) - \tau & (i \notin \mathcal{M}_{\boldsymbol{x}}) \end{cases}. \tag{7}$$

Therefore, we can finally reformulate truncated region $\mathcal{Z}$ as

$$\mathcal{Z} = \bigcap_{i \in [n]} \{z \in \mathbb{R} \mid f_i(z) < 0\}. \tag{8}$$

3.2 Adaptive Grid Search for Selective p-value Computation

The problem of finding $\mathcal{Z}$ in (8) is reduced to the problem of enumerating all solutions to the nonlinear equations $f_i(z) = 0$ for each $i \in [n]$ in (7). The difficulty of this problem depends on the complexity and nonlinearity of these functions f_i, $i \in [n]$. Fortunately, since each function f_i is part of the attention map computation in Transformer models, it is continuous, (sub)differentiable, and possesses a certain level of smoothness (except in pathological cases). Assuming a certain degree of smoothness for these functions f_i, by adaptively generating grid points in the one-dimensional space $z \in \mathbb{R}$ and computing the values of $f_i(z)$ at each grid point, it is possible to identify $\mathcal{Z}$ in (8) with sufficient accuracy. This further means that the selective p-value in (5) can be computed with sufficient accuracy (as stated in Theorem 3 later).

The overall procedure for estimating the selective p-value by an adaptive grid search method is summarized in Algorithm 1. Here, S defines the grid search interval $[-S, S]$, while $\varepsilon_{\min}$ and $\varepsilon_{\max}$ denote the minimum and maximum grid widths, respectively. Note that, in line 9 of Algorithm 1, the interval J is included for computational simplicity. The key concept of Algorithm 1 lies in how to determine the adaptive grid width $d(z_j)$.

Algorithm 1: Selective p-value Computation by Adaptive Grid Search

Input: S, $\varepsilon_{\min}$, $\varepsilon_{\max}$, $\{f_i\}_{i\in[n]}$ and $T(\boldsymbol{x})$

1 $j \leftarrow 0, z_0 \leftarrow -S$
2 **while** $z_j < S$ **do**
3 compute the adaptive grid width $d(z_j)$
4 $z_{j+1} \leftarrow z_j + \min(\varepsilon_{\max}, \max(d(z_j), \varepsilon_{\min}))$
5 $j \leftarrow j + 1$
6 $d' \leftarrow \min(\varepsilon_{\max}, d(T(\boldsymbol{x})))$
7 $J \leftarrow [T(\boldsymbol{x}) - d', T(\boldsymbol{x}) + d']$
8 $\mathcal{Z}^{\text{grid}} \leftarrow \cup_{j|z_j\in\mathcal{Z}}[z_j, z_{j+1}] \cup J$
9 $p_{\text{grid}} \leftarrow \mathbb{P}(|Z| > |T(\boldsymbol{x})| \mid Z \in \mathcal{Z}_{\text{grid}})$, where $Z \sim \mathcal{N}(0, 1)$

Output: p_{grid}

The following theorem states that, by utilizing the Lipschitz constant of f_i, it is possible to appropriately determine the adaptive grid width $d(z_j)$ and compute the selective p-value with sufficient accuracy.

Theorem 3 *Assume that f_i is differentiable and Lipschitz continuous for all $i \in [n]$. Assume further that f_i has at most a finite number of zeros, and that for any zero z^* of f_i, its derivative $f_i'(z^*)$ is non-zero, for all $i \in [n]$. Define the grid width $d(z_j)$ as*

$$d(z_j) = \begin{cases} \min\limits_{i\in[n], f_i(z_j)<0} \dfrac{|f_i(z_j)|}{L_i(z_j)} & (z_j \in \mathcal{Z}), \\ \max\limits_{i\in[n], f_i(z_j)\geq 0} \dfrac{|f_i(z_j)|}{L_i(z_j)} & (z_j \notin \mathcal{Z}), \end{cases}$$

where $L_i(z_j)$ is the Lipschitz constant of f_i in the $\varepsilon_{\max}$-neighborhood of z_j. Then, we have

$$|p_{\text{selective}} - p_{\text{grid}}| = O(\varepsilon_{\min} + \exp(-S^2/2)), \text{ where } \varepsilon_{\min} \to 0, S \to \infty.$$

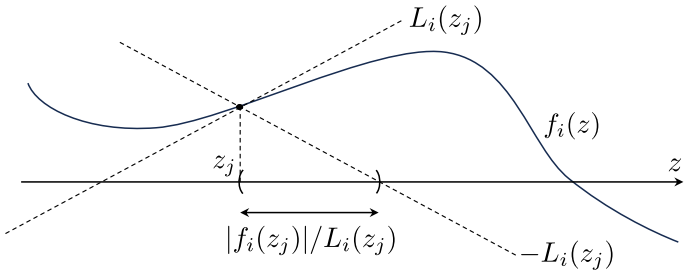


Figure 2: Schematic illustration of the relationship between grid width and Lipschitz constant. Let $L_i(z_j)$ be the Lipschitz constant of f_i in the $\varepsilon_{\max}$-neighborhood of z_j. Then, due to Lipschitz continuity, f_i maintains the same sign on the interval $[z_j, z_j + |f_i(z_j)|/L_i(z_j)]$. Note that this figure considers the case where $\varepsilon_{\max}$ is sufficiently large.

The proof of Theorem 3 is provided in Appendix G.3. The following lemma suggests why it is reasonable to define the grid width as $d(z_j)$ in Theorem 3.

Lemma 4 *For the grid width $d(z_j)$ defined in Theorem 3, we have*

$$z_j \in \mathcal{Z} \Rightarrow [z_j, z_j + \min(\varepsilon_{\max}, d(z_j))] \subset \mathcal{Z},$$
$$z_j \notin \mathcal{Z} \Rightarrow [z_j, z_j + \min(\varepsilon_{\max}, d(z_j))] \subset \mathbb{R} \setminus \mathcal{Z}.$$

The proof of Lemma 4 is provided in Appendix G.4. The central idea of the proof is that f_i maintains the same sign on the interval $[z_j, z_j + \min(\varepsilon_{\max}, |f_i(z_j)|/L_i(z_j))]$ from Lipschitz continuity (as illustrated in Figure 2). In Algorithm 1, the max operation (in line 4) is employed to prevent the grid width from becoming so small that the search process gets stuck within either $\mathcal{Z}$ or $\mathbb{R} \setminus \mathcal{Z}$.

3.3 Implementation of Adaptive Grid Search

Heuristics for Practical Implementation. To compute the grid width $d(z_j)$ in Theorem 3, it is necessary to know the Lipschitz constant $L_i(z_j)$ of the attention score $\mathcal{A}_i$ (from which f_i is derived) in the vicinity of z_j. In this paper, for practical computation, $d(z_j)$ is determined by conservatively estimating the Lipschitz constant $L_i(z_j)$ using some heuristics. Specifically, we introduce two types of heuristics based on the relative positions of the current grid point z_j and the observed test statistic $T(\boldsymbol{x})$. In the case where z_j is far from $T(\boldsymbol{x})$ (i.e., $|z_j - T(\boldsymbol{x})| > 0.1$), we assume that f_i can be approximated by a linear function in the $\varepsilon_{\max}$-neighborhood of z_j. Then, we conservatively set $L_i(z_j) = 10|f_i'(z_j)|$. Here, we can also assume that the sign of f_i does not change on the interval $[z_j, z_j + \varepsilon_{\max}]$ for an index i where $f_i(z_j)$ and $f_i'(z_j)$ have the same sign, because f_i is assumed to be approximated by a linear function. This can be implemented by performing the min or max operation (from the definition of $d(z_j)$) only for indices i such that $f_i(z_j)f_i'(z_j) < 0$. In contrast, in the case where z_j is close to $T(\boldsymbol{x})$ (i.e., $|z_j - T(\boldsymbol{x})| < 0.1$), f_i may exhibit a flat shape or micro oscillations and tends to take values close to zero. Note that careful consideration is required when any f_i is close to zero, as this implies that the grid point

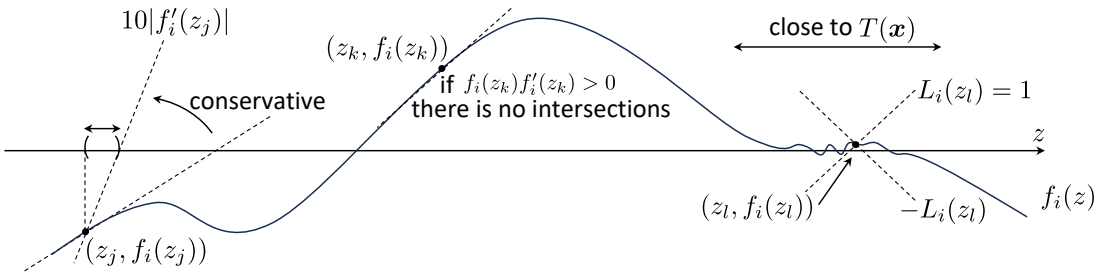


Figure 3: Schematic illustration of the heuristics we have employed in this paper. The left and central parts of the figure illustrate the case where a grid point is far from $T(\boldsymbol{x})$, while the right part illustrates the case where a grid point is close to $T(\boldsymbol{x})$. In the left part (representing z_j far from $T(\boldsymbol{x})$), the function f_i is approximated by a linear function, and its Lipschitz constant $L_i(z_j)$ is conservatively set to $10|f_i'(z_j)|$. In the central part (representing z_k far from $T(\boldsymbol{x})$ and f_i approximated linearly), the sign of f_i does not change as long as $f_i(z_k)f_i'(z_k)$ is positive. In the right part (representing z_l close to $T(\boldsymbol{x})$), the function f_i may exhibit a flat shape or micro-oscillations and tends to take values close to zero.

z_j is near the boundary of $\mathcal{Z}$. Therefore, it may not be reasonable to utilize the derivative of f_i in the same manner as above, so we set $L_i(z_j) = 1$ by assumption. This assumption is highly conservative, given that the range of the attention score $\mathcal{A}_i$ is $[0, 1]$. A schematic illustration of these heuristics is presented in Figure 3.

Derivative of Attention Maps. We considered utilizing the derivative of each f_i to compute the grid width $d(z_j)$. This necessitates computing the derivative of the attention map $\mathcal{A}$ (from which all f_i are derived), an output of the Transformer model. Automatic differentiation (AD), a technique implemented in many deep learning frameworks (e.g., TensorFlow and PyTorch), can be used to compute this derivative. It should be noted that the task involves differentiating an n-dimensional attention map with respect to a scalar input z_j. For differentiating a function with a scalar input and a high-dimensional vector output, reverse-mode AD (also known as backpropagation) is generally inefficient. Therefore, forward-mode AD (which computes Jacobian-vector products, or JVPs) is a more suitable option for this scenario, and we employ the forward-mode AD capabilities provided by PyTorch in our implementation. For details, see our implementation code.

Python Library. We implemented `si4attention`, a Python library based on PyTorch, to compute selective p-values for attention regions in Transformer models using our proposed Adaptive Grid Search. This library provides a primary function, `test_attention`. Its only mandatory arguments are the data to be tested and the Transformer model; it returns an attention map and its corresponding selective p-value. The input data is accepted as a `torch.Tensor`. The requirements for the Transformer model are minimal: it must be an instance of an `nn.Module` subclass and must possess a method to extract raw attention weights from each layer. These requirements align with standard practices in attention analysis for Transformers. For more details, please refer to our library repository.

4. Numerical Experiments

4.1 Methods for Comparison

We compared our proposed method (`adaptive`) with naive test (`naive`), permutation test (`permutation`), and Bonferroni correction (`bonferroni`), in terms of type I error rate and statistical power. Then, we compared our proposed method with other grid search options (`fixed` and `combination`) in terms of computation time. See Appendix E.1 for more details.

4.2 Synthetic Data Experiments

Setup. In our experiments, we employed an identical setup for both image and time series data. First, we trained Transformer classifier models on synthetic data sets. For each data type (image and time series), a synthetic data set with 1,000 negative samples (drawn from $\mathcal{N}(\mathbf{0}, I)$) and 1,000 positive samples (drawn from $\mathcal{N}(\boldsymbol{\mu}, I)$) was generated. The true value vector $\boldsymbol{\mu}$ was set to $\mu_i = \Delta$ for $i \in \mathcal{S}$ and $\mu_i = 0$ for $i \in [n] \setminus \mathcal{S}$, where Δ was uniformly sampled from $[1, 4]$ and $\mathcal{S}$ was a randomly located region of interest.

After training, we conducted experiments using the trained Transformer models on test data sets. A test sample (either an image or a time series) was input into the respective trained model to obtain an attention map, which was then used to perform a statistical test. In all experiments, we set the threshold value $\tau = 0.6$, the grid search interval $[-S, S]$ with $S = 10 + |T(\boldsymbol{x})|$, the minimum grid width $\varepsilon_{\min} = 10^{-4}$, the maximum grid width $\varepsilon_{\max} = 10^{-1}$, and the significance level $\alpha = 0.05$. We considered two types of covariance matrices $\Sigma \in \mathbb{R}^{n \times n}$ for generating the test data. The first represented independence: $\Sigma = I_n \in \mathbb{R}^{n \times n}$. The second represented correlation, structured as follows: for image data, $\Sigma = \Sigma' \otimes \Sigma' \in \mathbb{R}^{n \times n}$, where $\Sigma' = (0.5^{|i-j|})_{ij} \in \mathbb{R}^{\sqrt{n} \times \sqrt{n}}$; for time series data, $\Sigma = (0.5^{|i-j|})_{ij} \in \mathbb{R}^{n \times n}$.

To evaluate the type I error rate, we varied two primary factors: data size (i.e., the number of pixels n for images or the number of time points n for time series) and model architecture configuration. For data sizes, we varied the image size $n \in \{64, 256, 1024, 4096\}$ for image data and the time series length $n \in \{128, 256, 512, 1024\}$ for time series data. For model architecture configurations, we varied the architecture size in {small, base, large, huge} (details are in Appendix E.2). Unless otherwise specified, the default data size was 256 and the default architecture was base. For each experimental configuration, we generated 10,000 null test samples drawn from $\mathcal{N}(\mathbf{0}, \Sigma)$. The first 1,000 of these null samples also served for comparing computation times. To investigate the power, we fixed the data size to 256 and the architecture to base, then generated 1,000 test samples drawn from $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. For these power experiments, the true value vector $\boldsymbol{\mu}$ was set to $\mu_i = \Delta$ for $i \in \mathcal{S}$ and $\mu_i = 0$ for $i \in [n] \setminus \mathcal{S}$, where $\mathcal{S}$ is a randomly located region of interest and $\Delta \in \{1.0, 2.0, 3.0, 4.0\}$.

Results. The results for type I error rate are presented in Figures 4 and 5. The `adaptive` and `bonferroni` methods successfully controlled the type I error rate below the significance level across all configurations and for both image and time series data, whereas the other two methods, `naive` and `permutation`, could not. Because the `naive` and `permutation` methods failed to control the type I error rate, we did not evaluate their statistical power. The results for statistical power are presented in Figure 6. We confirmed that the `adaptive` method exhibited much higher power than the `bonferroni` method across all configurations and for both image and time series data. The results for computation time are presented in Figures 7

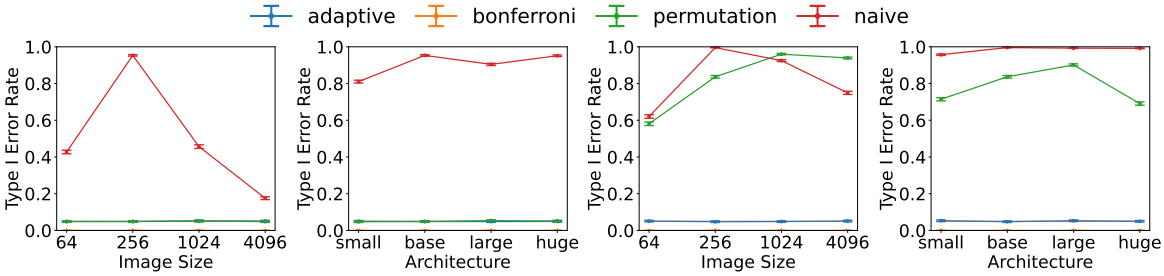
and 8. Across all configurations and for both image and time series data, the `adaptive` method outperformed the other compared grid strategies, `fixed` and `combination`, despite utilizing the smallest minimum grid width.

Discussion. Our experiments confirmed that the approximation approach, using the heuristics considered in Section 3, works well for attention maps in Transformer models. We further assess the reasonableness of these heuristics by presenting several examples of our target function f_i, defined in (7), in Figure 9. The plots demonstrate that the behavior of f_i is generally consistent with our heuristics: its shape can be approximated linearly when z is far from $T(\boldsymbol{x})$, and it tends to take values close to zero when z is close to $T(\boldsymbol{x})$. This observed behavior can be understood by considering the input to the Transformer model, which can be expressed as $\boldsymbol{a} + \boldsymbol{b}z = \boldsymbol{x} + \boldsymbol{b}(z - T(\boldsymbol{x}))$, where $\boldsymbol{b}$ is, by definition, parallel to $\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}$ by definition. When z is far from $T(\boldsymbol{x})$, the input $\boldsymbol{x} + \boldsymbol{b}(z - T(\boldsymbol{x}))$ results in data (an image or time series) where elements within the attention region $\mathcal{M}_{\boldsymbol{x}}$ (i.e., pixels or time points) are distinctively highlighted. Consequently, each attention score $\mathcal{A}_i$ (and thus f_i) often exhibits a gradual, near-linear trend. Conversely, when z is close to $T(\boldsymbol{x})$, the definition and continuity of f_i lead to the expectation that some f_i values will be close to zero. However, it remains an open question whether these heuristics are universally valid across all, especially more complex, Transformer architectures. In some instances, accurately modeling highly nonlinear functions might necessitate a significant reduction in grid size, which would, in turn, incur increased computational costs.

4.3 Real-World Data Experiments

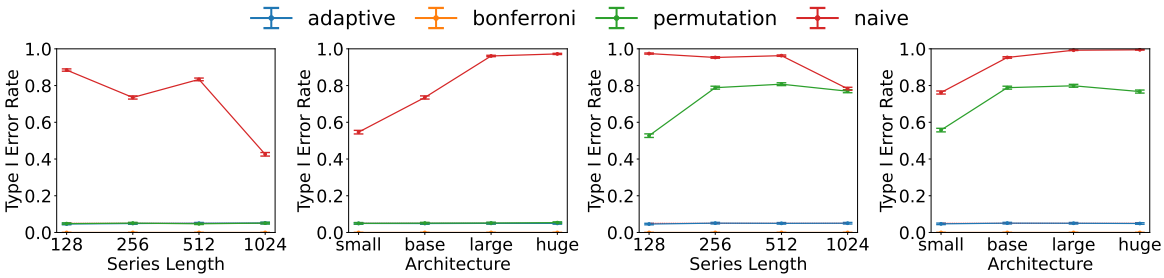
Setup. We conducted experiments on real-world data sets to demonstrate the effectiveness of our proposed method. For the image data, we utilized the T2-FLAIR MRI brain scans from the Brain Tumor Segmentation (BraTS) 2023 data set Karargyris et al. (2023); LaBella et al. (2023). We preprocessed this data set to obtain 402 positive images (with tumors) and 532 negative images (without tumors). From these images, 300 positive and 300 negative images were used to train the Transformer classifier model. The remaining images were used for testing, i.e., to conduct the statistical test. For the time series data, we utilized electroencephalography (EEG) signals recorded during the RSVP task (reaction to visual stimuli presentation) from the data set provided by Won et al. (2022). We preprocessed this data set to obtain 550 positive time series (identified by a stimulus-induced potential shift) and 7,700 negative time series (lacking such a shift). From these time series, 400 positive and 400 negative time series were used to train the Transformer classifier model. The remaining time series were used for testing, i.e., to conduct the statistical test. Further details on the data sets and preprocessing steps are provided in Appendix E.3.

Results. The results for the `adaptive` and `naive` methods are presented in Figure 10 for the MRI data set and Figure 11 for the EEG data set. The naive p-values remained small even for negative samples, indicating that they cannot be reliably used to quantify the statistical significance of the attention regions. In contrast, the selective p-values (computed from our `adaptive` method) were appropriately large for negative samples and small for positive samples. These results indicate that the `adaptive` method can effectively detect true positive cases while avoiding false positive detections.



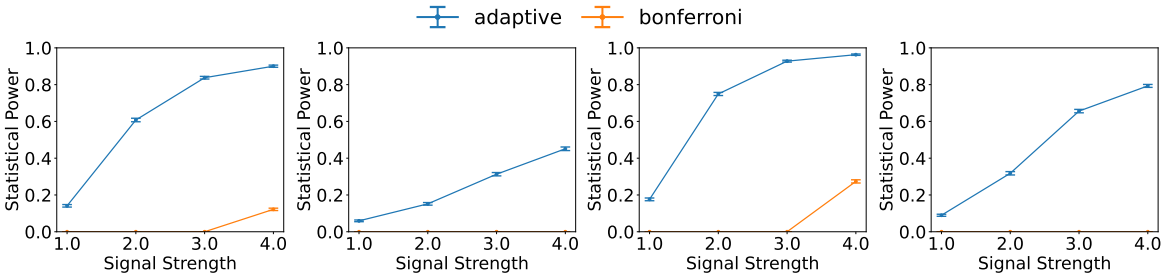
(a) Independence (b) Correlation

Figure 4: Type I Error Rate for synthetic image data set. Only our proposed method and Bonferroni correction are able to control the type I error rate across all configurations.



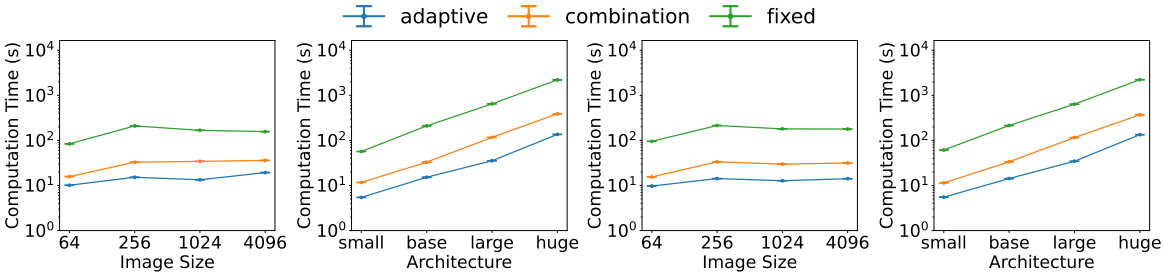
(a) Independence (b) Correlation

Figure 5: Type I Error Rate for synthetic time series data set. Only our proposed method and Bonferroni correction are able to control the type I error rate across all configurations.



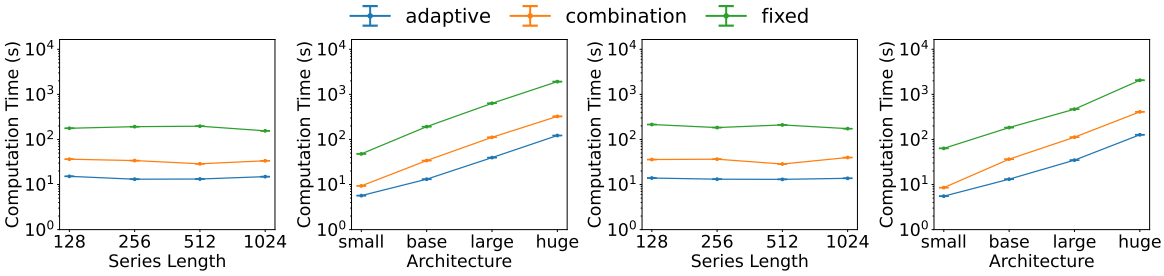
(a) Image (Independence and Correlation) (b) Time Series (Independence and Correlation)

Figure 6: Statistical Power for synthetic data sets. Our proposed method has much higher power than Bonferroni correction across all configurations.



(a) Independence (b) Correlation

Figure 7: Computation Time for synthetic image data set. Our proposed method outperforms the other two grid strategies across all configurations.



(a) Independence (b) Correlation

Figure 8: Computation Time for synthetic time series data set. Our proposed method outperforms the other two grid strategies across all configurations.

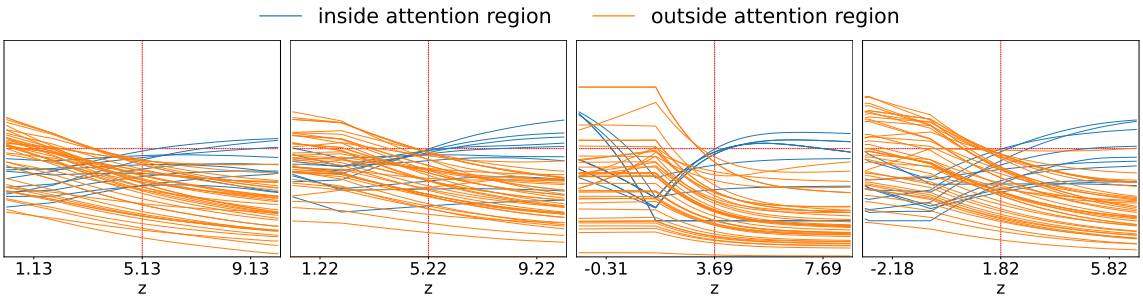
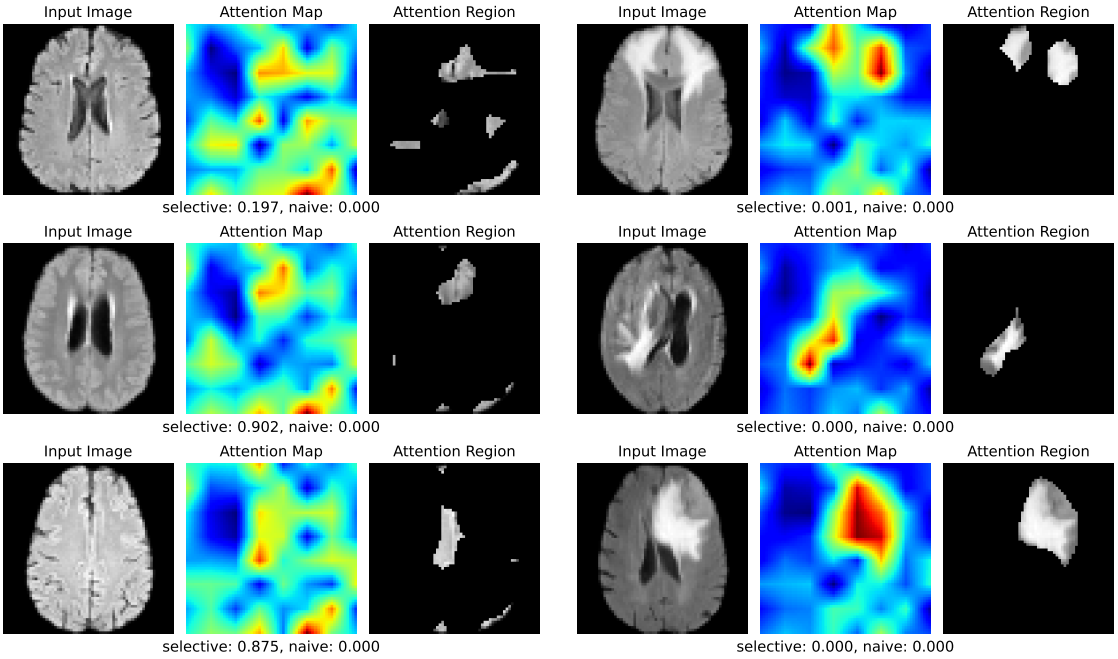


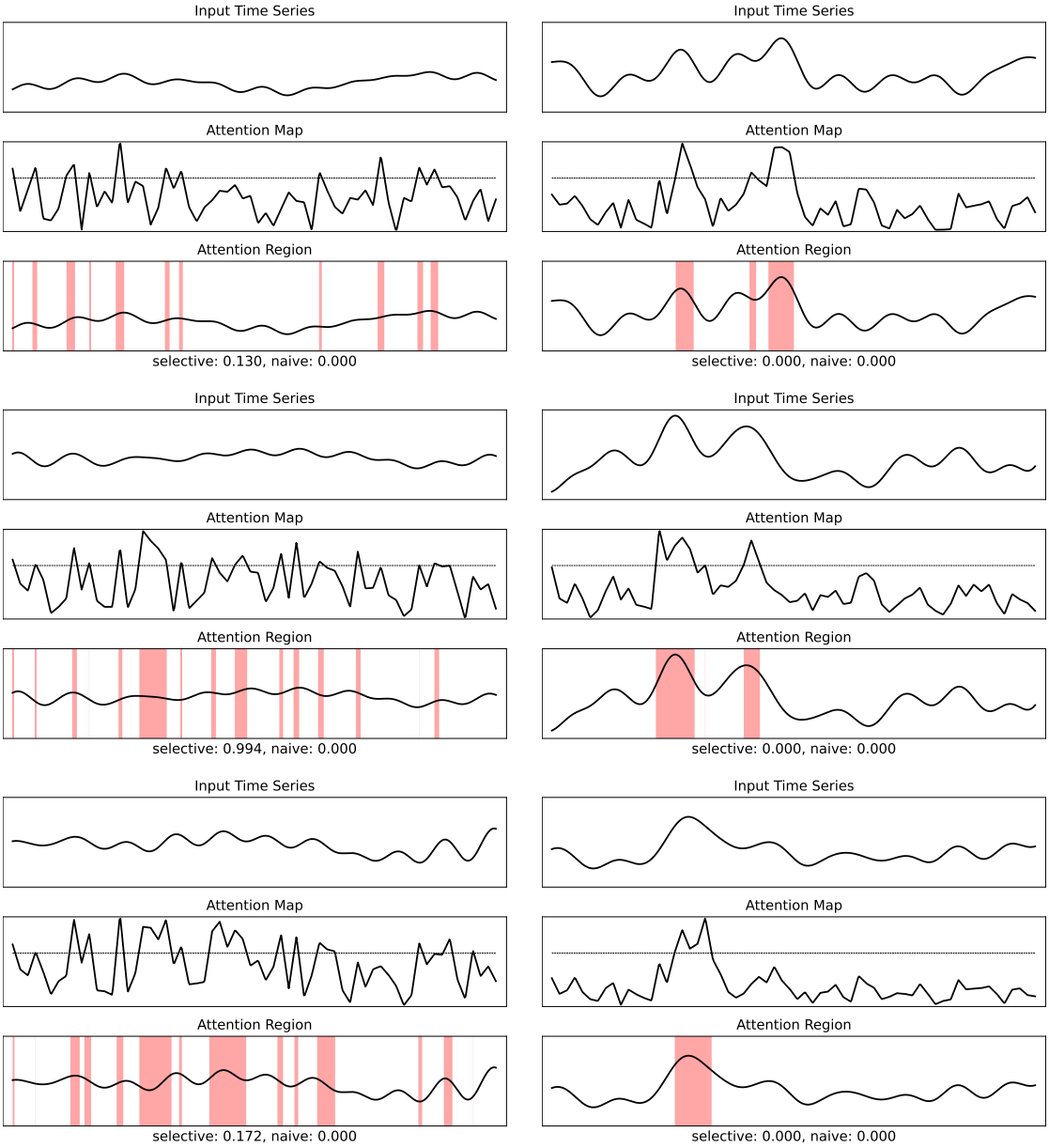
Figure 9: Demonstration of our target function f_i. For this demonstration, we generated a null image $\boldsymbol{x}$ from $\mathcal{N}(\mathbf{0}, I)$ with image size 256. We then input this image into a trained Transformer model with a base architecture to obtain the attention map. The vertical red line indicates the observed value of the test statistic, $T(\boldsymbol{x})$, and the horizontal red line represents zero. The blue plots display f_i values for 10 randomly selected indices i from the attention region $\mathcal{M}_{\boldsymbol{x}}$, while the orange plots display f_i values for 40 randomly selected indices i from its complement, $\mathcal{M}_{\boldsymbol{x}}^c$. Note that the region where all f_i values fall below zero (i.e., below the horizontal red line) corresponds to the truncated region $\mathcal{Z}$.



(a) Negative Images (without tumor)

(b) Positive Images (with tumor)

Figure 10: Demonstration on the MRI data set. Our proposed method conclude that attentions are statistically significant for positive images with tumors, while avoiding falsely detection of significance for negative images without tumors.



(a) Negative Time Series (without potential shift) (b) Positive Time Series (with potential shift)

Figure 11: Demonstration on the EEG data set. Our proposed method conclude that attentions are statistically significant for positive time series with potential shift, while avoiding falsely detection of significance for negative time series without potential shift.

5. Conclusion

In this study, we introduced a novel framework for quantifying the statistical significance of attention in Transformer, based on the concept of selective inference. We developed an innovative computational method for computing p-values, which serve as indicators of this statistical significance. One current limitation of our proposed method is its computational cost, which can make application to high-resolution data or very large architectures challenging. Introducing more efficient and robust heuristics is a key direction for future improvement, potentially broadening the applicability of our proposed method. We believe that this study opens an important avenue for ensuring the reliability of attention mechanisms in Transformer models.

Acknowledgments

This work was partially supported by JST CREST (JPMJCR21D3, JPMJCR22N2), JST Moonshot R&D (JPMJMS2033-05), RIKEN Center for Advanced Intelligence Project, and RIKEN Junior Research Associate Program.

Appendix A. Details of Transformer Models

A.1 Structure of Transformer Models

The overall structure of the Transformer model employed in our experiments is illustrated in Figure 12. Within the Multi-Layer Perceptron (MLP) components, we use two fully-connected layers and set the hidden dimension to four times the embedding dimension (#emb_dim). The Multi-Head Self-Attention mechanism utilizes a specified number of attention heads (#heads). Regarding patch embedding, the approach differs by data type: for image data, we set the patch size to $\min(2, \lfloor\sqrt{n}/8\rfloor)$ and the padding size equal to this patch size; while for time series data, we set the patch size to $\lfloor n/32\rfloor$ and the padding size to half of this patch size. For our default model configuration, we set the number of layers (#layers) to 8, the embedding dimension (#emb_dim) to 64, and the number of attention heads (#heads) to 4.

A.2 Computation of Attention Maps

Let N denote the number of patches (#patched), L the number of layers (#layers), and H the number of attention heads (#heads). We also define D as the dimension per head, calculated as the embedding dimension (#emb_dim) divided by the number of attention heads (#heads). In our experiments, we employed the Attention Rollout method, proposed by Abnar and Zuidema (2020), to compute the attention map. We now describe the formulation of attention weights and their aggregation, which ultimately produce the attention map $\mathcal{A}(\boldsymbol{x}) \in [0,1]^n$ from an input $\boldsymbol{x} \in \mathbb{R}^n$.

Formulation of Attention Weights. Within the l-th layer (for $l \in [L]$) and its h-th attention head (for $h \in [H]$), let $Q_{l,h} \in \mathbb{R}^{(N+1)\times D}$ and $K_{l,h} \in \mathbb{R}^{(N+1)\times D}$ denote the query and key matrices, respectively. The dimension $(N+1)$ accounts for N input patches plus

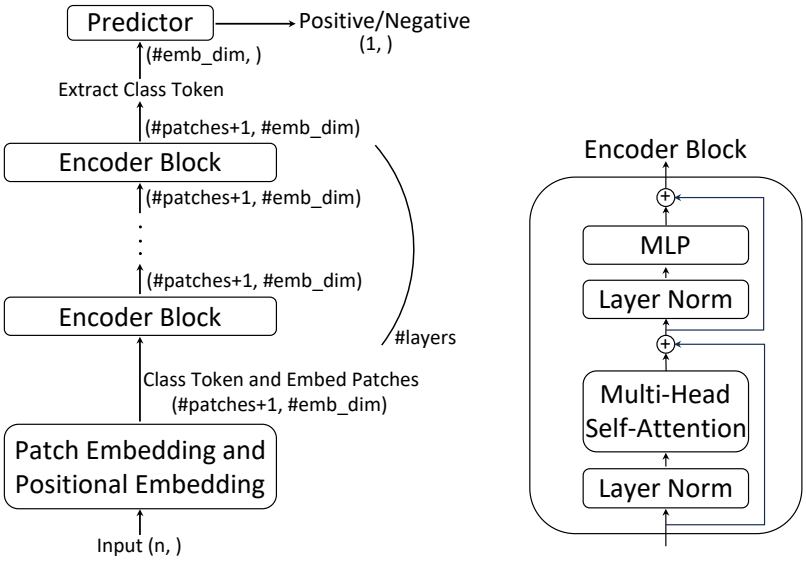


Figure 12: Structure of Transformer models.

an additional class token. The attention weights $A_{l,h} \in \mathbb{R}^{(N+1)\times(N+1)}$ are then computed as

$$A_{l,h} = \text{softmax}\left(\frac{Q_{l,h}K_{l,h}^{\top}}{\sqrt{D}}\right),\ (l,h) \in [L] \times [H],$$

where the softmax operation is applied row-wise to the input matrix. Note that the rows of $A_{l,h}$ correspond to queries and the columns correspond to keys.

Aggregating Attention Weights. First, we compute layer-wise attention weights $\hat{A}_l \in \mathbb{R}^{(N+1)\times(N+1)}$ by averaging the head-specific attention weights $A_{l,h}$ across all H attention heads for each layer $l \in [L]$:

$$\hat{A}_l = \frac{1}{H}\sum_{h\in[H]} A_{l,h},\ l \in [L].$$

Then, to aggregate these layer-wise attentions into a single weights matrix $\bar{A} \in \mathbb{R}^{(N+1)\times(N+1)}$, we form terms by adding the identity matrix $I \in \mathbb{R}^{(N+1)\times(N+1)}$ to each $\hat{A}_l$, and then compute the matrix product of these resulting terms across all layers:

$$\bar{A} = \prod_{l\in[L]} (\hat{A}_l + I),$$

where the addition of the identity matrix accounts for the skip connections within the Encoder Blocks, as depicted in Figure 12. Finally, we extract an N-dimensional attention vector $A \in \mathbb{R}^N$ from $\bar{A}$ by selecting the attentions from the class token to the N patches:

$$A = \bar{A}_{1,2:N+1},$$

which represents the aggregated attention from the class token (query) to each of the N patches (keys).

Post-Processing. We upscale the N-dimensional vector A to an n-dimensional vector A', whose size matches that of the original input data $\boldsymbol{x}$, using bilinear interpolation. Then, we obtain the final attention map $\mathcal{A}(\boldsymbol{x}) \in [0,1]^n$ by applying min-max normalization to A'.

Appendix B. Other Statistical Tests for Attentions in Transformers

In §2, to quantify the statistical significance of the attention region, we defined the null and alternative hypotheses in (2), thereby evaluating the attention region based on its entire exterior. However, other null and alternative hypotheses can also be considered for this purpose. For instance, we describe two such options below: one evaluating the attention region against its neighboring regions, and another evaluating it against reference data.

B.1 Difference from Neighborhood Region

For a given attention region $\mathcal{M}_{\boldsymbol{x}}$, we define its neighborhood region, $\mathcal{M}_{\boldsymbol{x}}^{\mathcal{N}}$, as the set of indices located within a certain distance from $\mathcal{M}_{\boldsymbol{x}}$ but not in $\mathcal{M}_{\boldsymbol{x}}$ itself:

$$\mathcal{M}_{\boldsymbol{x}}^{\mathcal{N}} = \{i \in [n] \setminus \mathcal{M}_{\boldsymbol{x}} \mid \exists j \in \mathcal{M}_{\boldsymbol{x}}, \rho(i,j) \leq R\},$$

where $\rho(i,j)$ is a distance metric between indices i and j, and R is a threshold distance. In this paper, we employed the Chebyshev distance as the distance for $\rho(i,j)$ and set R to $\max(1, \lfloor\sqrt{n}/16\rfloor)$ for the image data and $\lfloor n/16\rfloor$ for the time series data. Based on this neighborhood region, to quantify the statistical significance of the attention region, we can define the null and alternative hypotheses as:

$$\mathrm{H}_0 \colon \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{x}}}\mu_i = \frac{1}{|\mathcal{M}_{\boldsymbol{x}}^{\mathcal{N}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{x}}^{\mathcal{N}}}\mu_i, \quad \text{vs.} \quad \mathrm{H}_1 \colon \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{x}}}\mu_i \neq \frac{1}{|\mathcal{M}_{\boldsymbol{x}}^{\mathcal{N}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{x}}^{\mathcal{N}}}\mu_i, \tag{9}$$

To conduct the statistical test in (9), a natural choice for the test statistic is the difference in the mean values between the attention region and its neighborhood region, i.e.,

$$\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^{\top}\boldsymbol{X} = \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{x}}} X_i - \frac{1}{|\mathcal{M}_{\boldsymbol{x}}^{\mathcal{N}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{x}^{\mathcal{N}}}} X_i,$$

where $\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}} = \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|}\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{x}}} - \frac{1}{|\mathcal{M}_{\boldsymbol{x}}^{\mathcal{N}}|}\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{x}}^{\mathcal{N}}}$ is a vector that depends on the attention region $\mathcal{M}_{\boldsymbol{x}}$. For this statistical test, the theoretical framework discussed in Section 2 and Section 3 can be applied in exactly the same manner, thereby allowing the statistical test to be performed validly.

B.2 Difference from Reference Data

To quantify the statistical significance of the attention region, we can also utilize independent null reference data $\boldsymbol{x}^{\mathrm{ref}} \in \mathbb{R}^n$. Let us assume that this reference data $\boldsymbol{x}^{\mathrm{ref}}$ is a realization of the random vector

$$\boldsymbol{X}^{\mathrm{ref}} = (X_1^{\mathrm{ref}}, \ldots, X_n^{\mathrm{ref}})^{\top} = \boldsymbol{\mu}^{\mathrm{ref}} + \boldsymbol{\epsilon}^{\mathrm{ref}},\ \boldsymbol{\epsilon}^{\mathrm{ref}} \sim \mathcal{N}(0, \Sigma^{\mathrm{ref}}),$$

where $\boldsymbol{\mu}^{\mathrm{ref}}$ is the true underlying value vector and $\boldsymbol{\epsilon}^{\mathrm{ref}}$ is a Gaussian noise vector with a known or estimable covariance matrix Σ^{ref}. Based on this, we define the null and alternative hypotheses as:

$$\mathrm{H}_0 \colon \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|} \sum_{i \in \mathcal{M}_{\boldsymbol{x}}} \mu_i = \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|} \sum_{i \in \mathcal{M}_{\boldsymbol{x}}} \mu_i^{\mathrm{ref}}, \quad \text{vs.} \quad \mathrm{H}_1 \colon \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|} \sum_{i \in \mathcal{M}_{\boldsymbol{x}}} \mu_i \neq \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|} \sum_{i \in \mathcal{M}_{\boldsymbol{x}}} \mu_i^{\mathrm{ref}}, \quad (10)$$

To conduct the statistical test in (10), a natural choice for the test statistic is the difference in the mean values between the attention region and the corresponding region in the reference data, i.e.,

$$\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^{\top} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}^{\mathrm{ref}} \end{pmatrix} = \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|} \sum_{i \in \mathcal{M}_{\boldsymbol{x}}} X_i - \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|} \sum_{i \in \mathcal{M}_{\boldsymbol{x}}} X_i^{\mathrm{ref}},$$

where $\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}} = \frac{1}{|\mathcal{M}_{\boldsymbol{x}}|}((\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{x}}})^{\top}, -(\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{x}}})^{\top})^{\top} \in \mathbb{R}^{2n}$ is a vector that depends on the attention region $\mathcal{M}_{\boldsymbol{x}}$. For this statistical test, by considering the augmented random vector

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}^{\mathrm{ref}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}^{\mathrm{ref}} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon}^{\mathrm{ref}} \end{pmatrix}, \ \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon}^{\mathrm{ref}} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \Sigma & O \\ O & \Sigma^{\mathrm{ref}} \end{pmatrix}\right),$$

the framework discussed in Section 2 and Section 3 can again be applied in exactly the same manner, thereby allowing the statistical test to be performed validly.

B.3 Experimental Results

To evaluate the statistical tests defined in (9) and (10), we conducted experiments for type I error rate and statistical power, following the methodology described in Section 4. The results for the type I error rate are presented in Figures 13, 14, 16, and 17. The results for statistical power are presented in Figures 15 and 18. These figures show results similar to those in Section 4, indicating that our adaptive method also outperforms the comparator methods for these statistical tests.

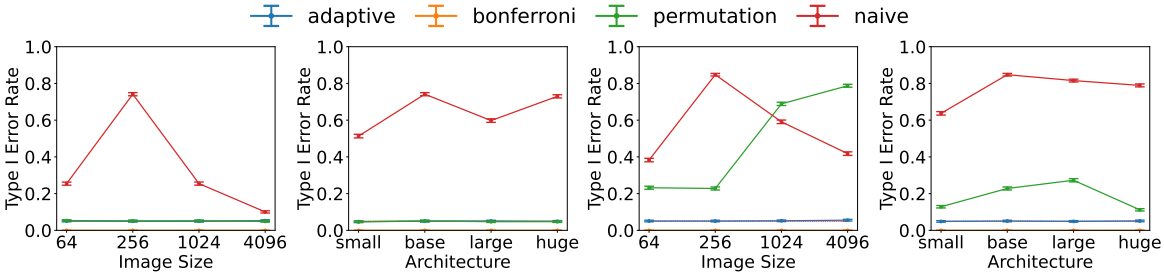
Appendix C. Data-Driven Approach for Selecting Threshold Value

In Section 2, we defined the attention region $\mathcal{M}_{\boldsymbol{x}}$ by thresholding the attention map $\mathcal{A}(\boldsymbol{x})$ with a pre-specified threshold τ. To extend our method to practical scenarios where an appropriate threshold is unknown a priori, we present a data-driven approach for selecting τ. A practical strategy is to select τ based on the empirical distribution of the attention scores $\{\mathcal{A}_i(\boldsymbol{x}) \mid i \in [n]\}$. For instance, we can set τ to the k-th largest attention score $\mathcal{A}_{(k)}(\boldsymbol{x})$ for a pre-specified integer $k \in [n]$. Crucially, the theoretical framework discussed in Section 2 and Section 3 remains valid under this selection strategy, necessitating only a modification to the reformulation of the truncated region in (8).

C.1 Modified Reformulation of Truncated Region

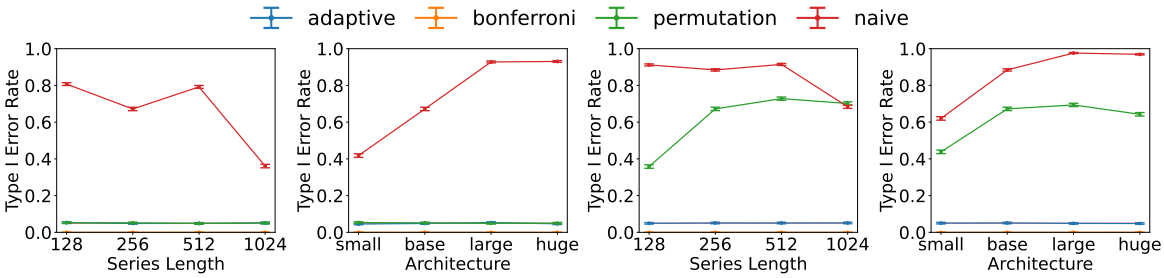
We consider the scenario where the threshold τ is set to the k-th largest attention score $\mathcal{A}_{(k)}(\boldsymbol{x})$. In this case, the attention region $\mathcal{M}_{\boldsymbol{x}}$ is defined as

$$\mathcal{M}_{\boldsymbol{x}} = \{i \in [n] \mid \mathcal{A}_i(\boldsymbol{x}) > \mathcal{A}_{(k)}(\boldsymbol{x})\},$$



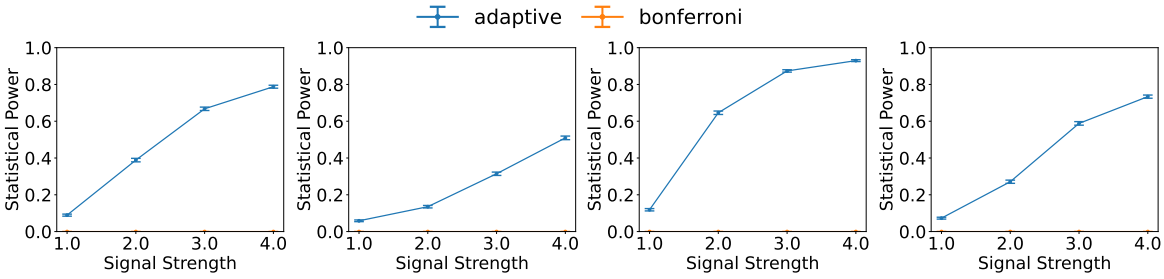
(a) Independence (b) Correlation

Figure 13: Type I Error Rate of statistical test in (9) for synthetic image data set.



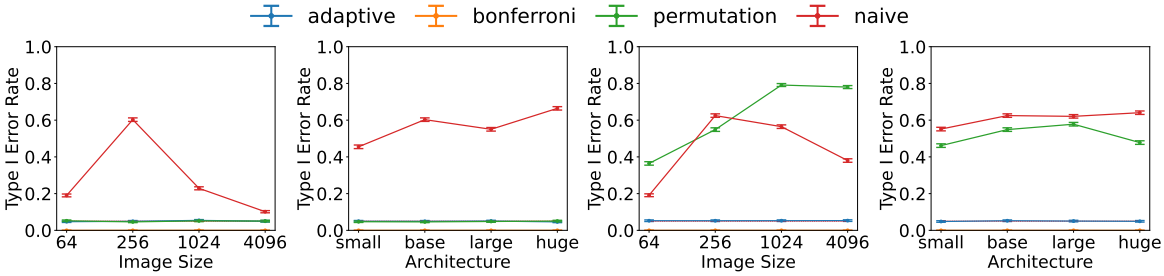
(a) Independence (b) Correlation

Figure 14: Type I Error Rate of statistical test in (9) for synthetic time series data set.



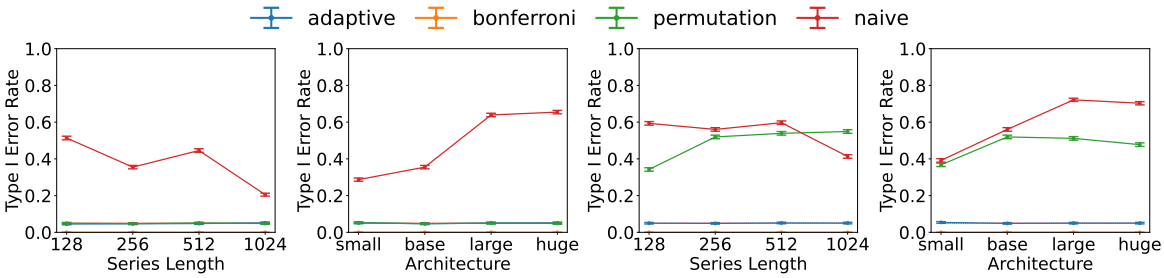
(a) Image (Independence and Correlation) (b) Time Series (Independence and Correlation)

Figure 15: Statistical Power of statistical test in (9) for synthetic data sets.



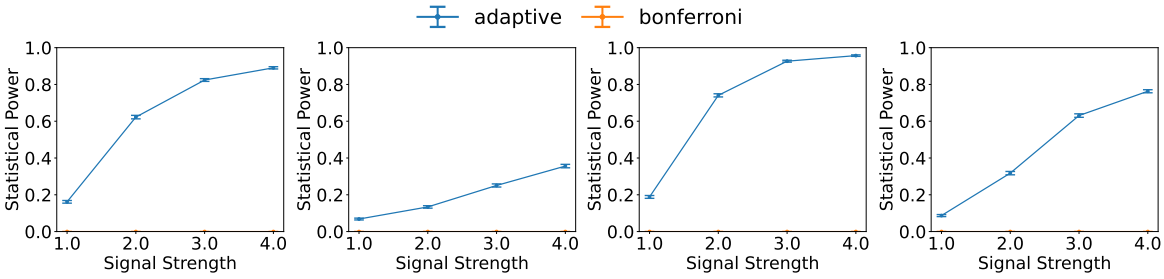
(a) Independence (b) Correlation

Figure 16: Type I Error Rate of statistical test in (10) for synthetic image data set.



(a) Independence (b) Correlation

Figure 17: Type I Error Rate of statistical test in (10) for synthetic time series data set.



(a) Image (Independence and Correlation) (b) Time Series (Independence and Correlation)

Figure 18: Statistical Power of statistical test in (10) for synthetic data sets.

which implies that the top $k-1$ elements with the highest attention scores are included in $\mathcal{M}_{\boldsymbol{x}}$. Using the fact that $\mathcal{A}_{(k)}(\boldsymbol{x}) = \max_{j\in[n]\setminus\mathcal{M}_{\boldsymbol{x}}} \mathcal{A}_j(\boldsymbol{x})$, the condition part of the truncated region $\mathcal{Z}$ defined in Theorem 2 can be reformulated as

$$\begin{aligned}
\mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{M}_{\boldsymbol{x}} \Leftrightarrow & \{i \in [n] \mid \mathcal{A}_i(\boldsymbol{a}+\boldsymbol{b}z) > \mathcal{A}_{(k)}(\boldsymbol{a}+\boldsymbol{b}z)\} = \mathcal{M}_{\boldsymbol{x}} \\
\Leftrightarrow & \begin{cases} \mathcal{A}_i(\boldsymbol{a}+\boldsymbol{b}z) > \max_{j\in[n]\setminus\mathcal{M}_{\boldsymbol{x}}} \mathcal{A}_j(\boldsymbol{a}+\boldsymbol{b}z), & \forall i \in \mathcal{M}_{\boldsymbol{x}} \\ \mathcal{A}_i(\boldsymbol{a}+\boldsymbol{b}z) \le \max_{j\in[n]\setminus\mathcal{M}_{\boldsymbol{x}}} \mathcal{A}_j(\boldsymbol{a}+\boldsymbol{b}z), & \forall i \in [n]\setminus\mathcal{M}_{\boldsymbol{x}} \end{cases} \\
\Leftrightarrow & \mathcal{A}_i(\boldsymbol{a}+\boldsymbol{b}z) > \mathcal{A}_j(\boldsymbol{a}+\boldsymbol{b}z), \ \forall i \in \mathcal{M}_{\boldsymbol{x}}, \forall j \in [n]\setminus\mathcal{M}_{\boldsymbol{x}} \\
\Leftrightarrow & f_{i,j}(z) < 0, \ \forall i \in \mathcal{M}_{\boldsymbol{x}}, \forall j \in [n]\setminus\mathcal{M}_{\boldsymbol{x}},
\end{aligned}$$

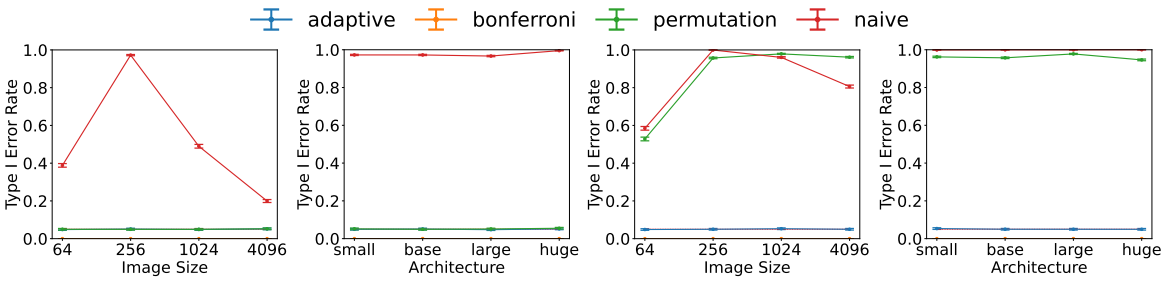
where $f_{i,j}(z) = \mathcal{A}_j(\boldsymbol{a}+\boldsymbol{b}z) - \mathcal{A}_i(\boldsymbol{a}+\boldsymbol{b}z)$. Therefore, we can finally reformulate truncated region $\mathcal{Z}$ as

$$\mathcal{Z} = \bigcap_{i\in\mathcal{M}_{\boldsymbol{x}}} \bigcap_{j\in[n]\setminus\mathcal{M}_{\boldsymbol{x}}} \{z \in \mathbb{R} \mid f_{i,j}(z) < 0\}.$$

Based on this reformulation, the computation of the selective p-value can be performed analogously to the procedure described in Section 3. Note that the total number of inequalities defining the truncated region $\mathcal{Z}$ is $|\mathcal{M}_{\boldsymbol{x}}| \cdot |[n]\setminus\mathcal{M}_{\boldsymbol{x}}|$, reaching a maximum of $\lfloor n^2/4 \rfloor$. While this number exceeds that of the fixed threshold case (where the number of inequalities is n), these inequalities are derived from the same n-dimensional attention map $\mathcal{A}(\boldsymbol{a}+\boldsymbol{b}z)$, ensuring that the overall computational complexity remains feasible.

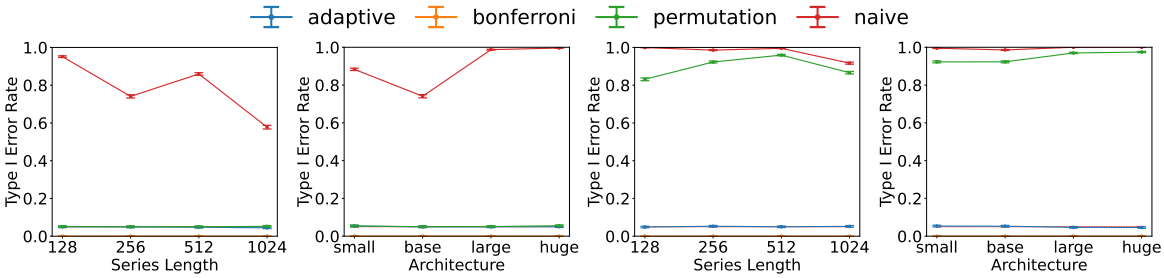
C.2 Experimental Results

To evaluate the data-driven threshold selection strategy, we set the threshold τ to the k-th largest attention score, with $k = \lceil 0.1n \rceil$. Using this thresholding approach, we conducted experiments for type I error rate and statistical power, following the methodology described in Section 4. The results for the type I error rate are presented in Figures 19 and 20. The results for statistical power are presented in Figure 21. These figures show results similar to those in Section 4, indicating that our adaptive method also outperforms the comparator methods even under this data-driven threshold selection strategy.



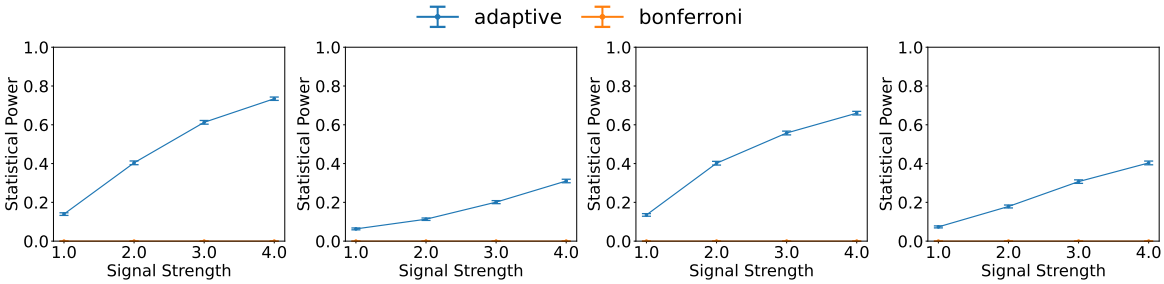
(a) Independence (b) Correlation

Figure 19: Type I Error Rate for synthetic image data set with data-driven threshold.



(a) Independence (b) Correlation

Figure 20: Type I Error Rate for synthetic time series data set with data-driven threshold.



(a) Image (Independence and Correlation) (b) Time Series (Independence and Correlation)

Figure 21: Statistical Power for synthetic data sets with data-driven threshold.

Appendix D. Confidence Intervals

In Section 2, we mainly focused on statistical hypothesis testing for evaluating the difference in mean values between the attention region and its complement. In addition to hypothesis testing, our framework can be utilized to construct confidence intervals for the parameter of interest. We define the scalar parameter δ as the standardized difference in mean values between the attention region $\mathcal{M}_{\boldsymbol{x}}$ and its complement:

$$\delta = \frac{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^{\top}\boldsymbol{\mu}}{\sqrt{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^{\top}\Sigma\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}}} = \frac{1}{\sqrt{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^{\top}\Sigma\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}}}\left(\frac{1}{|\mathcal{M}_{\boldsymbol{x}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{x}}}\mu_i - \frac{1}{|\mathcal{M}_{\boldsymbol{x}}^c|}\sum_{i\notin\mathcal{M}_{\boldsymbol{x}}}\mu_i\right).$$

Note that the test statistic $T(\boldsymbol{X})$ acts as an estimator for this parameter δ.

D.1 Construction of Confidence Intervals

The construction of confidence intervals for δ can be achieved by inverting the hypothesis test described in Section 3. From Theorem 2 and its proof, under the condition $\{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}, \mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}}\}$, the test statistic $T(\boldsymbol{X})$ follows a truncated normal distribution with mean δ, variance 1, and truncated region $\mathcal{Z}$. We denote the cumulative distribution function of this truncated normal distribution as

$$F_{\delta}(z) = \mathbb{P}(Z \leq z \mid Z \in \mathcal{Z}), \text{ where } Z \sim \mathcal{N}(\delta, 1).$$

Since $F_\delta(T(\boldsymbol{x}))$ is a pivot quantity that follows a uniform distribution, we can construct a $(1-\alpha)$ confidence interval $[\delta_{\mathrm{L}}, \delta_{\mathrm{U}}]$ for any significance level $\alpha \in (0,1)$ by finding the values of δ that satisfy

$$F_{\delta_{\mathrm{L}}}(T(\boldsymbol{x})) = 1 - \frac{\alpha}{2},\ F_{\delta_{\mathrm{U}}}(T(\boldsymbol{x})) = \frac{\alpha}{2}.$$

Since $F_\delta(z)$ is strictly decreasing with respect to δ, these endpoints δ_{L} and δ_{U} are uniquely determined and can be computed efficiently using numerical root-finding methods (e.g., the bisection method).

D.2 Experimental Results

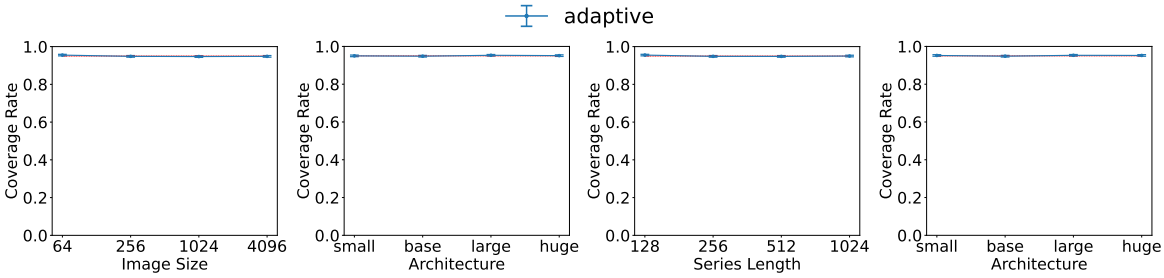
To evaluate the validity of the proposed confidence intervals, we conducted experiments following the experimental setup for type I error rate evaluation described in Section 4, with the covariance matrix set to $\Sigma = I_n$. We computed 95% confidence intervals ($\alpha = 0.05$) for the parameter δ and evaluated the coverage rate, defined as the proportion of trials where the constructed interval contains the true parameter δ. The results are presented in Figure 22. These results demonstrate that our adaptive method achieves coverage rates close to the nominal level of 95% across all configurations, confirming the validity of the proposed confidence intervals.

Appendix E. Experimental Details

E.1 Methods for Comparison

We compared our proposed method (`adaptive`) with the following methods:

- `naive`: The method uses a classical z-test without conditioning, i.e., the naive p-value is computed as $p_{\mathrm{naive}} = \mathbb{P}(|Z| > |T(\boldsymbol{x})|)$, where $Z \sim \mathcal{N}(0,1)$.
- `permutation`: This method uses a permutation test. The procedure is as follows: first, we compute the observed test statistic $T(\boldsymbol{x})$ by inputting the observed data $\boldsymbol{x} \in \mathbb{R}^n$ into the Transformer model. Then, we generate B permuted data (we set $B = 1,000$), denoted $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(B)} \in \mathbb{R}^n$, by permuting the elements of $\boldsymbol{x}$. For each



(a) Image (b) Time Series

Figure 22: Coverage Rate for synthetic data sets. Our proposed method achieves coverage rates close to the nominal level across all configurations.

permuted data $\boldsymbol{x}^{(b)}$, we compute its test statistic $T^{(i)}$ by inputting it into the same Transformer model. The permutation p-value is then computed as $p_{\text{permutation}} = \frac{1}{B}\sum_{b\in[B]} \mathbf{1}\{|T^{(b)}| > |T(\boldsymbol{x})|\}$, where $\mathbf{1}\{\cdot\}$ is the indicator function.

- `bonferroni`: This is a method to control the type I error rate by using the Bonferroni correction, a simple yet widely used method for multiple testing correction. The number of all possible attention regions is 2^n, then the Bonferroni corrected p-value is computed as $p_{\text{bonferroni}} = \min(1, 2^n \cdot p_{\text{naive}})$.

- `fixed`: This is a grid search method with a fixed grid width $\varepsilon = 10^{-3}$.

- `combination`: This is a grid search method with a fixed grid width $\varepsilon = 10^{-4}$ for grid points z_j satisfying $|z_j - T(\boldsymbol{x})| < 0.1$ and a with $\varepsilon = 10^{-2}$ for all other grid points.

Note that, in implementing the grid search methods (`adaptive`, `fixed`, and `combination`), a binary search is performed to find the boundary of the truncated region $\mathcal{Z}$ between adjacent grid points where one belongs to $\mathcal{Z}$ and the other does not.

E.2 Architectures to Compare

The architectures to compare in our experiments are presented in Table 1.

E.3 Details on Real-World Data Sets and Preprocessing

MRI Data Set. We utilized the T2-weighted Fluid-Attenuated Inversion Recovery (T2-FLAIR) MRI brain scans from the Brain Tumor Segmentation (BraTS) 2023 data set (Karargyris et al., 2023; LaBella et al., 2023). This data set comprises 934 3D T2-FLAIR scans, each with dimensions of $240 \times 240 \times 155$ voxels, presented without prior skull stripping.

As a preprocessing step, we first extracted a single 2D axial slice (240×240 pixels) from the 95th slice position of each 3D scan. Subsequently, we cropped each slice to remove the background outside the skull and then resized the cropped image to 64×64 pixels. Finally, based on the provided tumor annotations, these processed 2D images were categorized into 532 negative images (without tumors) and 402 positive images (with tumors). For standardization, we first estimated the mean and standard deviation of pixel intensities for each 2D image. These statistics were calculated using only pixel values from the brain region, excluding the background outside the skull and the tumor region identified in the ground truth. Each image was then standardized by subtracting its estimated mean and dividing by its estimated standard deviation.

Architecture	#layers	#hidden_dim	#heads	#parameters
small	4	32	2	53.2K
base	8	64	4	405K
large	12	128	8	2.39M
huge	16	256	16	12.7M

Table 1: Architectures to compare.

EEG Data Set. We utilized the electroencephalography (EEG) signals recorded during the Rapid Serial Visual Presentation (RSVP) task (focusing on the period following visual stimuli presentation) from the data set provided by Won et al. (2022). This data set comprises EEG recordings from 55 participants. For each participant, there are 40 positive samples (target stimuli) and 560 negative samples (non-target stimuli). Each raw sample consists of a 32-channel, 614 time-point signal, acquired from 32 electrodes at a sampling rate of 512 Hz. The epoch for these recordings spanned from −200 ms to 1000 ms relative to stimulus onset. In this paper, our analysis specifically targeted the post-stimulus period (0 ms to 1000 ms) to investigate the P300 response, a well-established neural marker of target detection in RSVP paradigms.

As a preprocessing step, we first applied a band-pass filter to the raw samples, following the methodology described by Won et al. (2022). Subsequently, to transform the multi-channel samples into univariate time series, we averaged the signals from three central electrodes: Fz, Cz, and Pz. Finally, we averaged sets of four samples from the same participant that shared the same label, resulting in 550 positive samples and 7,700 negative time series. For standardization, we first estimated the mean and standard deviation of signal values from a dedicated set of 100 negative samples that were not utilized in either model training or statistical testing. Each time series was then standardized by subtracting this estimated mean and dividing by this estimated standard deviation.

Appendix F. Robustness of Type I Error Rate Control

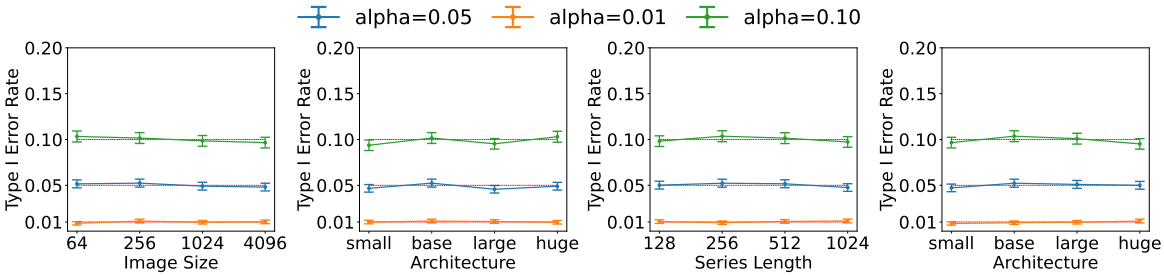
We confirmed the robustness of our proposed method in controlling the type I error rate by evaluating its performance in two scenarios: when the covariance matrix is estimated from the test data itself, and when the noise follows one of five non-Gaussian distribution families.

F.1 Estimated Covariance Matrix.

For the case where the covariance matrix is estimated from the test data, we conducted experiments to evaluate type I error rate. These experiments followed the same methodology as in Section 4 with the covariance matrix $\Sigma = I_n$. For each configuration, we generated 10,000 null test samples and performed a statistical test using an estimated covariance matrix $\hat{\sigma}^2 I_n$, where $\hat{\sigma}^2$ was the sample variance computed from each respective test sample. For these experiments, we considered three significance levels $\alpha \in \{0.05, 0.01, 0.10\}$. The results, presented in Figure 23, show that our proposed method robustly controlled the type I error rate across all configurations, even when the covariance matrix was estimated from the test data.

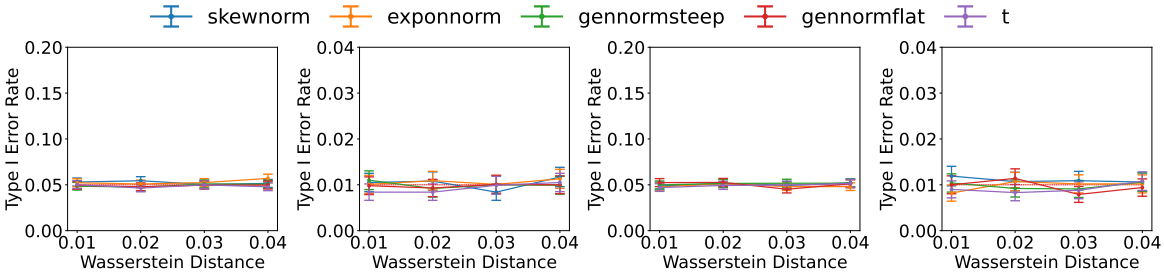
F.2 Non-Gaussian Noise.

For the case where the noise follows one of five non-Gaussian distribution families, we conducted experiments to evaluate type I error rate. We fixed the data size to 256 and the architecture to base, and considered the following five distribution families for the noise: skew normal distribution family (`skewnorm`), exponentially modified normal distribution family (`exponnorm`), generalized normal distribution family with a shape parameter $\beta < 2$ that is



(a) Image (b) Time Series

Figure 23: Type I Error Rate for synthetic data sets with estimated covariance matrix. Our proposed method robustly controlled the type I error rate across all configurations.



(a) Image (Significance Level 0.05 and 0.01) (b) Time Series (Significance Level 0.05 and 0.01)

Figure 24: Type I Error Rate for synthetic data sets with non-Gaussian noise. Our proposed method robustly controlled the type I error rate across all configurations.

steeper than the normal distribution (`gennormsteep`), generalized normal distribution family with a shape parameter $\beta > 2$ that is flatter than the normal distribution (`gennormflat`), and Student's t distribution family (`t`). For each family, we first identified specific distributions such that the 1-Wasserstein distance from $\mathcal{N}(0,1)$ is l, for $l \in \{0.01, 0.02, 0.03, 0.04\}$. We then generated 10,000 null test samples from each distribution and performed a statistical test. For these experiments, we considered two significance levels $\alpha = \{0.05, 0.01\}$. The results, presented in Figure 24, show that our proposed method robustly controlled the type I error rate across all configurations, even when the noise followed one of five non-Gaussian distribution families.

Appendix G. Proofs

G.1 Proof of Theorem 1

We proof this theorem based on the Theorem 2. Then, by probability integral transformation, under the null hypothesis, we have

$$p_{\text{selective}} \mid \{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}, \mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}}\} \sim \text{Unif}(0,1),$$

which leads to

$$\mathbb{P}_{\text{H}_0}(p_{\text{selective}} \leq \alpha \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}, \mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}}) = \alpha, \ \forall \alpha \in (0,1).$$

For any $\alpha \in (0,1)$, we firstly marginalize over all the values of the sufficient statistic for the nuisance parameter and then over all possible attention regions. Regarding the marginalization of the sufficient statistic for the nuisance parameter, we have

$$\begin{aligned}
&\mathbb{P}_{\text{H}_0}(p_{\text{selective}} \leq \alpha \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}) \\
=&\int_{\mathbb{R}^n} \mathbb{P}_{\text{H}_0}(p_{\text{selective}} \leq \alpha \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}, \mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}}) \mathbb{P}_{\text{H}_0}(\mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}} \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}) d\mathcal{Q}_{\boldsymbol{x}} \\
=&\alpha \int_{\mathbb{R}^n} \mathbb{P}_{\text{H}_0}(\mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}} \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}) d\mathcal{Q}_{\boldsymbol{x}} = \alpha.
\end{aligned}$$

Regarding the marginalization of the attention region, we have

$$\begin{aligned}
&\mathbb{P}_{\text{H}_0}(p_{\text{selective}} \leq \alpha) \\
=&\sum_{\mathcal{M}_{\boldsymbol{x}} \in 2^{[n]} \setminus \{\emptyset, [n]\}} \mathbb{P}_{\text{H}_0}(p_{\text{selective}} \leq \alpha \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}) \mathbb{P}_{\text{H}_0}(\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}) \\
=&\alpha \sum_{\mathcal{M}_{\boldsymbol{x}} \in 2^{[n]} \setminus \{\emptyset, [n]\}} \mathbb{P}_{\text{H}_0}(\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}) = \alpha.
\end{aligned}$$

■

G.2 Proof of Theorem 2

According to the conditioning on $\mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}}$, we have

$$\mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}} \Leftrightarrow \left(I_n - \frac{\Sigma \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}} \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^\top}{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}^\top \Sigma \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{x}}}} \right) \boldsymbol{X} = \mathcal{Q}_{\boldsymbol{x}} \Leftrightarrow \boldsymbol{X} = \boldsymbol{a} + \boldsymbol{b} z,$$

where $z = T(\boldsymbol{X}) \in \mathbb{R}$. Then, we have

$$\begin{aligned}
\mathcal{X} &= \{\boldsymbol{X} \in \mathbb{R}^n \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}, \mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}}\} \\
&= \{\boldsymbol{X} \in \mathbb{R}^n \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}, \boldsymbol{X} = \boldsymbol{a} + \boldsymbol{b} z, z \in \mathbb{R}\} \\
&= \{\boldsymbol{a} + \boldsymbol{b} z \in \mathbb{R}^n \mid \mathcal{M}_{\boldsymbol{a} + \boldsymbol{b} z} = \mathcal{M}_{\boldsymbol{x}}, z \in \mathbb{R}\} \\
&= \{\boldsymbol{a} + \boldsymbol{b} z \in \mathbb{R}^n \mid z \in \mathcal{Z}\}.
\end{aligned}$$

Therefore, we have that the following conditional distribution of the test statistic is the truncated standard normal distribution with the truncated region $\mathcal{Z}$:

$$T(\boldsymbol{X}) \mid \{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{x}}, \mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{x}}\}.$$

■

G.3 Proof of Theorem 3

We note that the $\varepsilon_{\max}$ is not necessarily to evaluate the error bound because it is introduced for implementation convenience. Let us define the indicator function $I(z_j)$ as

$$I(z_j) = \begin{cases} 1 & (\varepsilon_{\min} \leq d(z_j)) \\ 0 & (\varepsilon_{\min} > d(z_j)) \end{cases}$$

First, we divide $\mathbb{R}$ into the four unions of intervals such that any two of them have no intersection with length as

$$R^1 = \bigcup_{j|I(z_j)=1, z_j \in \mathcal{Z}} [z_j, z_{j+1}] \cup J,\ R^2 = \bigcup_{j|I(z_j)=1, z_j \notin \mathcal{Z}} [z_j, z_{j+1}],$$
$$R^3 = \bigcup_{j|I(z_j)=0} [z_j, z_{j+1}] \setminus J,\ R^4 = (-\infty, -S] \cup [S, \infty).$$

Here, $R^1 \subset \mathcal{Z}^{\mathrm{grid}}$ and $R^2 \subset \mathbb{R} \setminus \mathcal{Z}^{\mathrm{grid}}$ are obvious from the definition of them, and from the Lemma 4, we have $R^1 \subset \mathcal{Z}$ and $R^2 \subset \mathbb{R} \setminus \mathcal{Z}$. Then, we have following subset relationships:

$$R^1 \subset \mathcal{Z},\ \mathcal{Z}^{\mathrm{grid}} \subset R^1 \cup R^3 \cup R^4 \tag{11}$$

Let us denote the probability density function of the standard normal distribution as ϕ and the cumulative distribution function of that as Φ, and introduce the integrate function $\mathcal{I}\colon \mathcal{B}(\mathbb{R}) \ni R \mapsto \int_R \phi(z)dz \in [0,1]$, where $\mathcal{B}(\mathbb{R})$ is the Borel set of $\mathbb{R}$. Additionally, for any $R \in \mathcal{B}(\mathbb{R})$, we define the two sets $R_{\mathrm{in}}, R_{\mathrm{out}} \in \mathcal{B}(\mathbb{R})$ as

$$R_{\mathrm{in}} = R \cap [-|T(\boldsymbol{x})|, |T(\boldsymbol{x})|],\ R_{\mathrm{out}} = R \setminus [-|T(\boldsymbol{x})|, |T(\boldsymbol{x})|].$$

Then, using the standard normal distribution $Z \sim \mathcal{N}(0,1)$, we have $p_{\mathrm{selective}}$ and p_{grid} as

$$p_{\mathrm{selective}} = \mathbb{P}_{\mathrm{H}_0}(|Z| > |T(\boldsymbol{x})| \mid Z \in \mathcal{Z}) = \frac{\mathcal{I}(\mathcal{Z}_{\mathrm{out}})}{\mathcal{I}(\mathcal{Z})}, \tag{12}$$

$$p_{\mathrm{grid}} = \mathbb{P}_{\mathrm{H}_0}(|Z| > |T(\boldsymbol{x})| \mid Z \in \mathcal{Z}^{\mathrm{grid}}) = \frac{\mathcal{I}(\mathcal{Z}^{\mathrm{grid}}_{\mathrm{out}})}{\mathcal{I}(\mathcal{Z}^{\mathrm{grid}})}, \tag{13}$$

respectively. Therefore, by considering the subset relationships in (11), our goal of evaluating the error is casted into the evaluating the $\mathcal{I}(R^1)$, $\mathcal{I}(R^3)$, and $\mathcal{I}(R^4)$. To do so, we start to evaluate the length of R^3.

We denote the Lipschitz constant of f_i as $L_i > 0$ and the number of zeros of f_i as $K_i \in \mathbb{N}$. We define the $L > 0$ and $K \in \mathbb{N}$ as $L = \max_{i \in [n]} L_i$ and $K = \max_{i \in [n]} K_i$, respectively. Then, for any $z_j \in \mathcal{Z}$, we have

$$d(z_j) \geq \min_{i \in [n]} \frac{|f_i(z_j)|}{L_i(z_j)} \geq \min_{i \in [n]} \frac{|f_i(z_j)|}{L}$$

Furthermore, regarding the condition of R^3, we have $\varepsilon_{\min} > \min_{i \in [n]} |f_i(z_j)|/L$ from $I(z_j) = 0 \Leftrightarrow \varepsilon_{\min} > d(z_j)$. Therefore, we have the following subset relationship:

$$R^3 \subset \bigcup_{j|I(z_j)=0} [z_j, z_{j+1}] = \bigcup_{j|I(z_j)=0} [z_j, z_j + \varepsilon_{\min}] \subset \bigcup_{j|L\varepsilon_{\min} > \min_{i \in [n]} |f_i(z_j)|} [z_j, z_j + \varepsilon_{\min}]. \tag{14}$$

Continuously, we evaluate the length of R^3 by show that the set in (14) is restricted to the neighborhood of the zeros of f_i. For $i \in [n]$, we denote the k-th zeros of f_i as $q_{ik} (k \in [K_i])$, and the minimum value of $|f_i'|$ at the zeros of f_i as $h_i > 0$ (i.e., $h_i = \min_{k\in[K_i]} |f_i'(q_{ik})|$). Let us denote the $h > 0$ as $h = \min_{i\in[n]} h_i$. Here, by using these zeros, we define the set $D(r)$ for any $r > 0$, which is the union of the r-neighborhood of the zeros,

$$D(r) = \bigcup_{i\in[n]} \bigcup_{k\in[K_i]} [q_{ik} - r, q_{ik} + r].$$

Then, for any $i \in [n]$ and $k \in [K_i]$, from the definition of derivative function, there exists $\delta_{ik} > 0$ such that, for any s satisfying $0 < |s| < \delta_{ik}$,

$$\left| \frac{f_i(q_{ik} + s) - f_i(q_{ik})}{s} - f_i'(q_{ik}) \right| < \frac{h}{2}$$

holds. Therefore, from the triangle inequality and the definition of h, we have

$$\frac{h}{2} > \left| f'(q_{ik}) - \frac{f_i(q_{ik} + s)}{s} \right| \geq \left| f'(q_{ik}) \right| - \left| \frac{f_i(q_{ik} + s)}{s} \right| \geq h - \left| \frac{f_i(q_{ik} + s)}{s} \right|.$$

To summarize, we have $|f_i(q_{ik}+s)| \geq h|s|/2$ including the case of $s = 0$. Thus, let us denote the $\delta > 0$ as $\delta = \min_{i\in[n]} \min_{k\in[K_i]} \delta_{ik}$, then, for any s satisfying $|s| < \delta$, we have

$$\min_{i\in[n]} \min_{k\in[K_i]} |f_i(q_{ik} + s)| \geq \frac{h}{2}|s|. \tag{15}$$

Next, we consider the set $[-S, S] \setminus D(\delta)$, which is assumed to have its boundary points added. Then, this set is a compact set, and thus the minimum value of $\min_{i\in[n]} |f_i|$ in this set is attained and we denote it as $l > 0$ (because $l = 0$ violates the assumption of zeros of f_i and the definition of $D(\delta)$).

As follows, we consider the asymptotic case of $\varepsilon_{\min} \to 0$ and then only consider the case of $\varepsilon_{\min} < \min(h\delta/2L, l/L)$. In this case, we have $0 < 2L\varepsilon_{\min}/h < \delta$, thus, from (15), the infimum of $\min_{i\in[n]} |f_i|$ in $D(\delta)/D(2L\varepsilon_{\min}/h)$ is greater than or equal to $h(2L\varepsilon_{\min}/h)/2 = L\varepsilon_{\min}$. By combining this with the definition of l, for any $z \in [-S, S] \setminus D(2L\varepsilon_{\min}/h)$, we have

$$\min_{i\in[n]} |f_i(z)| \geq \min(L\varepsilon_{\min}, l) = L\varepsilon_{\min},$$

where we used the assumption of $\varepsilon_{\min} < l/L$. Therefore, we have

$$\begin{aligned} R^3 &\subset \bigcup_{j|L\varepsilon_{\min} > \min_{i\in[n]} |f_i(z_j)|} [z_j, z_j + \varepsilon_{\min}] \\ &\subset \bigcup_{j|z_j \in D(2L\varepsilon_{\min}/h)} [z_j, z_j + \varepsilon_{\min}] \subset D\left(\left(\frac{2L}{h} + 1 \right) \varepsilon_{\min} \right). \end{aligned}$$

Based on these results, we return to the evaluation of the $\mathcal{I}(R^1)$, $\mathcal{I}(R^3)$, and $\mathcal{I}(R^4)$. Regarding the $\mathcal{I}(R^3)$, we have

$$\begin{aligned}\mathcal{I}(R^3) &\leq \mathcal{I}\left(D\left(\left(\frac{2L}{h}+1\right)\varepsilon_{\min}\right)\right)\\ &\leq \sum_{i\in[n]}\sum_{k\in[K_i]}\mathcal{I}\left(\left[q_{ik}-\left(\left(\frac{2L}{h}+1\right)\varepsilon_{\min}\right), q_{ik}+\left(\left(\frac{2L}{h}+1\right)\varepsilon_{\min}\right)\right]\right)\\ &= \sum_{i\in[n]}\sum_{k\in[K_i]}\left\{\Phi\left(q_{ik}+\left(\frac{2L}{h}+1\right)\varepsilon_{\min}\right)-\Phi\left(q_{ik}-\left(\frac{2L}{h}+1\right)\varepsilon_{\min}\right)\right\}.\end{aligned}$$

By using the mean value theorem and the fact that ϕ has the maximum value at 0, then we have

$$\mathcal{I}(R^3) \leq \sum_{i\in[n]}\sum_{k\in[K_i]}\phi(0)\left(\frac{4L}{h}+2\right)\varepsilon_{\min} \leq nK\phi(0)\left(\frac{4L}{h}+2\right)\varepsilon_{\min} = M_1\varepsilon_{\min}, \tag{16}$$

where $M_1 = nK\phi(0)\left(4L/h+2\right)$ is a positive constant independent of $\varepsilon_{\min}$ and S. Next, regarding the $\mathcal{I}(R^1)$, from the mean value theorem, the symmetry of ϕ and the decreasing property of ϕ on $[0,\infty)$, we have

$$\mathcal{I}(R^1) \geq \mathcal{I}(J) \geq \mathcal{I}([T(\boldsymbol{x})-d', T(\boldsymbol{x})+d'] \cap [-|T(\boldsymbol{x})|, |T(\boldsymbol{x})|]) \geq \phi(T(\boldsymbol{x}))d' = M_2, \tag{17}$$

where $M_2 = \phi(T(\boldsymbol{x}))d'$ is a positive constant independent of $\varepsilon_{\min}$ and S. Finally, regarding the $\mathcal{I}(R^4)$, we have

$$\mathcal{I}(R^4) = 2\Phi(-S) \tag{18}$$

Finally, we evaluate the error bound. From (11), (12) and (13), we have

$$\frac{\mathcal{I}(R^1_{\text{out}})}{\mathcal{I}(R^1\cup R^3\cup R^4)} \leq p_{\text{selective}},\ p_{\text{grid}} \leq \frac{\mathcal{I}((R^1\cup R^3\cup R^4)_{\text{out}})}{\mathcal{I}(R^1)}.$$

Therefore, by using (16), (17) and (18), we have the following error bound:

$$\begin{aligned}|p_{\text{selective}} - p_{\text{grid}}| \leq& \frac{\mathcal{I}((R^1\cup R^3\cup R^4)_{\text{out}})}{\mathcal{I}(R^1)} - \frac{\mathcal{I}(R^1_{\text{out}})}{\mathcal{I}(R^1\cup R^3\cup R^4)}\\ =& \frac{\mathcal{I}((R^1\cup R^3\cup R^4)_{\text{out}})\mathcal{I}(R^1\cup R^3\cup R^4) - \mathcal{I}(R^1_{\text{out}})\mathcal{I}(R^1)}{\mathcal{I}(R^1)\mathcal{I}(R^1\cup R^3\cup R^4)}\\ =& \frac{\mathcal{I}(R^1_{\text{out}})\mathcal{I}(R^3\cup R^4) + \mathcal{I}((R^3\cup R^4)_{\text{out}})\mathcal{I}(R^1\cup R^3\cup R^4)}{\mathcal{I}(R^1)\mathcal{I}(R^1\cup R^3\cup R^4)}\\ \leq& \frac{\mathcal{I}(R^3\cup R^4) + \mathcal{I}((R^3\cup R^4)_{\text{out}})}{\mathcal{I}(R^1)^2}\\ \leq& \frac{2}{M_2^2}\mathcal{I}(R^3\cup R^4) \leq \frac{2}{M_2^2}(M_1\varepsilon_{\min} + 2\Phi(-S)).\end{aligned} \tag{19}$$

Here, based on the three equations $\lim_{z\to\infty}\phi(z)/z\phi(z) = 0$, $\Phi'(-z) = (1-\Phi(z))' = -\phi(z)$ and $\phi'(z) = -z\phi(z)$, we have the $\lim_{z\to\infty}\Phi(-z)/\phi(z) = 0$ from the l'Hôpital's rule. We

consider the asymptotic case of $S \to \infty$ and then only consider the case of S sufficiently large such that $\Phi(-S) \leq \exp(-S^2/2)$ holds (we can take such S because $\lim_{z\to\infty} \Phi(-z)/\phi(z) = 0$). By combining this with (19), we have

$$|p_{\text{selective}} - p_{\text{grid}}| \leq \frac{2}{M_2^2}(M_1\varepsilon_{\min} + 2\exp(-S^2/2)) \leq \frac{2M_1+4}{M_2^2}(\varepsilon_{\min} + \exp(-S^2/2)),$$

where the coefficient $(2M_1+4)/M_2^2$ is positive constant independent of $\varepsilon_{\min}$ and S. Thus, we have successfully showed that the error is bounded by $O(\varepsilon_{\min} + \exp(-S^2/2))$ in asymptotic case of $\varepsilon_{\min} \to 0$ and $S \to \infty$, which was what we wanted. ■

G.4 Proof of Lemma 4

In case of $z_j \in \mathcal{Z}$, we have

$$\begin{aligned}
[z_j, z_j + \min(\varepsilon_{\max}, d(z_j))] &= [z_j, z_j + \varepsilon_{\max}] \cap \left[z_j, z_j + \min_{i\in[n], f_i(z_j)<0} \frac{|f_i(z_j)|}{L_i(z_j)}\right] \\
&= \bigcap_{i\in[n], f_i(z_j)<0} [z_j, z_j + \varepsilon_{\max}] \cap \left[z_j, z_j + \frac{|f_i(z_j)|}{L_i(z_j)}\right] \\
&= \bigcap_{i\in[n], f_i(z_j)<0} \left[z_j, z_j + \min\left(\varepsilon_{\max}, \frac{|f_i(z_j)|}{L_i(z_j)}\right)\right] \subset \mathcal{Z}.
\end{aligned}$$

Similarly, in case of $z_j \notin \mathcal{Z}$, we have

$$\begin{aligned}
[z_j, z_j + \min(\varepsilon_{\max}, d(z_j))] &= [z_j, z_j + \varepsilon_{\max}] \cap \left[z_j, z_j + \max_{i\in[n], f_i(z_j)\geq 0} \frac{|f_i(z_j)|}{L_i(z_j)}\right] \\
&= \bigcup_{i\in[n], f_i(z_j)\geq 0} [z_j, z_j + \varepsilon_{\max}] \cap \left[z_j, z_j + \frac{|f_i(z_j)|}{L_i(z_j)}\right] \\
&= \bigcup_{i\in[n], f_i(z_j)\geq 0} \left[z_j, z_j + \min\left(\varepsilon_{\max}, \frac{|f_i(z_j)|}{L_i(z_j)}\right)\right] \subset \mathbb{R} \setminus \mathcal{Z}.
\end{aligned}$$ ■

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020.

Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. Why attentions may not be interpretable? In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 25–34, 2021.

Soham Bakshi, Yiling Huang, Snigdha Panigrahi, and Walter Dempsey. Inference with randomized regression trees. *arXiv preprint arXiv:2412.20535*, 2024.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, 2022.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *8th International Conference on Learning Representations (ICLR 2020)(virtual)*. International Conference on Learning Representations, 2020.

Rachel Carrington and Paul Fearnhead. Improving power by conditioning on less in post-selection inference for changepoints. *Statistics and Computing*, 35(1):1–23, 2025.

Ali Charkhi and Gerda Claeskens. Asymptotic post-selection inference for the akaike information criterion. *Biometrika*, 105(3):645–664, 2018.

Shuxiao Chen and Jacob Bien. Valid inference corrected for outlier removal. *Journal of Computational and Graphical Statistics*, 29(2):323–334, 2020.

Yiqun T Chen and Daniela M Witten. Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152):1–41, 2023.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

Yunjin Choi, Jonathan Taylor, and Robert Tibshirani. Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *The Annals of Statistics*, 45 (6):2590–2617, 2017.

George Chrysostomou and Nikolaos Aletras. Improving the faithfulness of attention-based explanations with task-specific information for text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 477–488, 2021.

Diptesh Das, Vo Nguyen Le Duy, Hiroyuki Hanada, Koji Tsuda, and Ichiro Takeuchi. Fast and more powerful selective inference for sparse high-order interaction model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9999–10007, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

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

Vo Nguyen Le Duy and Ichiro Takeuchi. More powerful conditional selective inference for generalized lasso by parametric programming. *The Journal of Machine Learning Research*, 23(1):13544–13580, 2022.

Vo Nguyen Le Duy and Ichiro Takeuchi. Statistical inference for the dynamic time warping distance, with application to abnormal time-series detection: Vnl duy, i. takeuchi. *Annals of the Institute of Statistical Mathematics*, pages 1–29, 2025.

Vo Nguyen Le Duy, Hiroki Toda, Ryota Sugiyama, and Ichiro Takeuchi. Computing valid p-value for optimal changepoint by selective inference using dynamic programming. In *Advances in Neural Information Processing Systems*, 2020.

Vo Nguyen Le Duy, Shogo Iwazaki, and Ichiro Takeuchi. Quantifying statistical significance of neural network-based image segmentation by selective inference. *Advances in Neural Information Processing Systems*, 35:31627–31639, 2022.

Gazi Jannatul Ferdous, Khaleda Akhter Sathi, Md Azad Hossain, Mohammed Moshiul Hoque, and M Ali Akber Dewan. Lcdeit: A linear complexity data-efficient image transformer for mri brain tumor classification. *IEEE Access*, 11:20337–20350, 2023.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Lucy L Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, pages 1–11, 2022.

Shuang Hong, Jin Wu, Lei Zhu, and Weijie Chen. Brain tumor classification in vit-b/16 based on relative position encoding and residual mlp. *Plos one*, 19(7):e0298102, 2024.

Shuaicong Hu, Jian Liu, Rui Yang, Ya'Nan Wang, Aiguo Wang, Kuanzheng Li, Wenxin Liu, and Cuiwei Yang. Exploring the applicability of transfer learning and feature engineering in epilepsy prediction using hybrid transformer model. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1321–1332, 2023.

Sangwon Hyun, Max G'sell, and Ryan J Tibshirani. Exact post-selection inference for the generalized lasso path. *Electronic Journal of Statistics*, 12(1):1053–1097, 2018.

Sangwon Hyun, Kevin Z Lin, Max G'Sell, and Ryan J Tibshirani. Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*, 77(3):1037–1049, 2021.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, 2019.

Sean Jewell, Paul Fearnhead, and Daniela Witten. Testing for a change in mean after changepoint detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1082–1104, 2022.

Yiming Ju, Yuanzhe Zhang, Zhao Yang, Zhongtao Jiang, Kang Liu, and Jun Zhao. Logic traps in evaluating attribution scores. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5911–5922, 2022.

Alexandros Karargyris, Renato Umeton, Micah J. Sheller, Alejandro Aristizabal, Johnu George, Anna Wuest, Sarthak Pati, Hasan Kassem, Maximilian Zenk, Ujjwal Baid, Prakash Narayana Moorthy, Alexander Chowdhury, Junyi Guo, Sahil Nalawade, Jacob Rosenthal, David Kanter, Maria Xenochristou, Daniel J. Beutel, Verena Chung, Timothy Bergquist, James Eddy, Abubakar Abid, Lewis Tunstall, Omar Sanseviero, Dimitrios Dimitriadis, Yiming Qian, Xinxing Xu, Yong Liu, Rick Siow Mong Goh, Srini Bala, Victor Bittorf, Sreekar Reddy Puchala, Biagio Ricciuti, Soujanya Samineni, Eshna Sengupta, Akshay Chaudhari, Cody Coleman, Bala Desinghu, Gregory Diamos, Debo Dutta, Diane Feddema, Grigori Fursin, Xinyuan Huang, Satyananda Kashyap, Nicholas Lane, Indranil Mallick, Pietro Mascagni, Virendra Mehta, Cassiano Ferro Moraes, Vivek Natarajan, Nikola Nikolov, Nicolas Padoy, Gennady Pekhimenko, Vijay Janapa Reddi, G. Anthony Reina, Pablo Ribalta, Abhishek Singh, Jayaraman J. Thiagarajan, Jacob Albrecht, Thomas Wolf, Geralyn Miller, Huazhu Fu, Prashant Shah, Daguang Xu, Poonam Yadav, David Talby, Mark M. Awad, Jeremy P. Howard, Michael Rosenthal, Luigi Marchionni, Massimo Loda, Jason M. Johnson, Spyridon Bakas, Peter Mattson, FeTS Consortium, BraTS-2020 Consortium, and AI4SafeChole Consortium. Federated benchmarking of medical artificial intelligence with medperf. *Nature Machine Intelligence*, 5(7):799–810, July 2023. doi: 10.1038/s42256-023-00652-2. URL `https://doi.org/10.1038/s42256-023-00652-2`.

Teruyuki Katsuoka, Tomohiro Shiraishi, Daiki Miwa, Vo Nguyen Le Duy, and Ichiro Takeuchi. Statistical test on diffusion model-based anomaly detection by selective inference. *arXiv preprint arXiv:2402.11789*, 2024.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Dominic LaBella, Maruf Adewole, Michelle Alonso-Basanta, Talissa Altes, Syed Muhammad Anwar, Ujjwal Baid, Timothy Bergquist, Radhika Bhalerao, Sully Chen, Verena Chung, Gian-Marco Conte, Farouk Dako, James Eddy, Ivan Ezhov, Devon Godfrey, Fathi Hilal, Ariana Familiar, Keyvan Farahani, Juan Eugenio Iglesias, Zhifan Jiang, Elaine Johanson, Anahita Fathi Kazerooni, Collin Kent, John Kirkpatrick, Florian Kofler, Koen Van Leemput, Hongwei Bran Li, Xinyang Liu, Aria Mahtabfar, Shan McBurney-Lin, Ryan McLean, Zeke Meier, Ahmed W Moawad, John Mongan, Pierre Nedelec, Maxence Pajot, Marie Piraud, Arif Rashid, Zachary Reitman, Russell Takeshi Shinohara, Yury Velichko, Chunhao Wang, Pranav Warman, Walter Wiggins, Mariam Aboian, Jake Albrecht, Udunna Anazodo, Spyridon Bakas, Adam Flanders, Anastasia Janas, Goldey Khanna, Marius George Linguraru, Bjoern Menze, Ayman Nada, Andreas M Rauschecker, Jeff Rudie, Nourel Hoda Tahon, Javier Villanueva-Meyer, Benedikt Wiestler, and Evan Calabrese. The asnr-miccai brain tumor segmentation (brats) challenge 2023: Intracranial meningioma, 2023. URL `https://arxiv.org/abs/2305.07642`.

Jason D Lee and Jonathan E Taylor. Exact post model selection inference for marginal screening. *Advances in neural information processing systems*, 27, 2014.

Jason D Lee, Yuekai Sun, and Jonathan E Taylor. Evaluating the statistical significance of biclusters. *Advances in neural information processing systems*, 28, 2015.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI open*, 3:111–132, 2022.

Keli Liu, Jelena Markovic, and Robert Tibshirani. More powerful post-selection inference, with application to the lasso. *arXiv preprint arXiv:1801.09037*, 2018.

Ninghao Liu, Yunsong Meng, Xia Hu, Tie Wang, and Bo Long. Are interpretations fairly evaluated? a definition driven pipeline for post-hoc interpretability. *arXiv preprint arXiv:2009.07494*, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Joshua R Loftus. Selective inference after cross-validation. *arXiv preprint arXiv:1511.08866*, 2015.

Joshua R Loftus and Jonathan E Taylor. Selective inference in regression models with groups of variables. *arXiv preprint arXiv:1511.01478*, 2015.

Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. Evaluating the faithfulness of importance measures in nlp by recursively masking allegedly important tokens and retraining. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1731–1751, 2022.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pages 1614–1623. PMLR, 2016.

Clara Meister, Stefan Lazov, Isabelle Augenstein, and Ryan Cotterell. Is sparse attention more interpretable? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 122–129, 2021.

Daiki Miwa, Duy Vo Nguyen Le, and Ichiro Takeuchi. Valid p-value for deep learning-driven salient region. In *Proceedings of the 11th International Conference on Learning Representation*, 2023.

Daiki Miwa, Tomohiro Shiraishi, Vo Nguyen Le Duy, Teruyuki Katsuoka, and Ichiro Takeuchi. Statistical test for anomaly detections by variational auto-encoders. *arXiv preprint arXiv:2402.03724*, 2024.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. Towards transparent and explainable attention models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4206–4216, 2020.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2791–2802, 2021.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, 2018.

Anna C Neufeld, Lucy L Gao, and Daniela M Witten. Tree-values: selective inference for regression trees. *Journal of Machine Learning Research*, 23(305):1–43, 2022.

Thang Loi Nguyen, Loc Duong, and Vo Nguyen Le Duy. Statistical inference for feature selection after optimal transport-based domain adaptation. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2025.

Shuichi Nishino, Tomohiro Shiraishi, Teruyuki Katsuoka, and Ichiro Takeuchi. Statistical test for saliency maps of graph neural networks via selective inference. *Transactions on Machine Learning Research*, 2025.

Snigdha Panigrahi, Jonathan Taylor, and Asaf Weinstein. Integrative methods for post-selection inference under convex constraints. *The Annals of Statistics*, 49(5):2803–2824, 2021.

Snigdha Panigrahi, Peter W MacDonald, and Daniel Kessler. Approximate post-selective inference for regression with the group lasso. *Journal of machine learning research*, 24 (79):1–49, 2023.

Ronan Perry, Snigdha Panigrahi, Jacob Bien, and Daniela Witten. Inference on the proportion of variance explained in principal component analysis. *Journal of the American Statistical Association*, pages 1–11, 2025.

Sarah Pirenne and Gerda Claeskens. Parametric programming-based approximate selective inference for adaptive lasso, adaptive elastic net and group lasso. *Journal of Statistical Computation and Simulation*, 94(11):2412–2435, 2024.

David Rügamer and Sonja Greven. Inference for l 2-boosting. *Statistics and computing*, 30 (2):279–289, 2020.

David Rügamer, Philipp FM Baumann, and Sonja Greven. Selective inference for additive and linear mixed models. *Computational Statistics & Data Analysis*, 167:107350, 2022.

Tomohiro Shiraishi, Daiki Miwa, Teruyuki Katsuoka, Vo Nguyen Le Duy, Kouichi Taji, and Ichiro Takeuchi. Statistical test for attention maps in vision transformers. In *International Conference on Machine Learning*, pages 45079–45096. PMLR, 2024a.

Tomohiro Shiraishi, Daiki Miwa, Vo Nguyen Le Duy, and Ichiro Takeuchi. Selective inference for change point detection by recurrent neural network. *Neural Computation*, 37(1):160–192, 2024b.

Kazuya Sugiyama, Vo Nguyen Le Duy, and Ichiro Takeuchi. More powerful and general selective inference for stepwise feature selection using homotopy method. In *International Conference on Machine Learning*, pages 9891–9901. PMLR, 2021.

Xiaobing Sun and Wei Lu. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428, 2020.

Shinya Suzumura, Kazuya Nakagawa, Yuta Umezu, Koji Tsuda, and Ichiro Takeuchi. Selective inference for sparse high-order interaction models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3338–3347. JMLR. org, 2017.

Jonathan Taylor and Robert Tibshirani. Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61, 2018.

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

Yoshikazu Terada and Hidetoshi Shimodaira. Selective inference for the problem of regions via multiscale bootstrap. *arXiv preprint arXiv:1711.00949*, 2017.

Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018.

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR, 2019.

Toshiaki Tsukurimichi, Yu Inatsu, Vo Nguyen Le Duy, and Ichiro Takeuchi. Conditional selective inference for robust regression and outlier detection using piecewise-linear homotopy continuation. *Annals of the Institute of Statistical Mathematics*, 74(6):1197–1228, 2022.

Martin Tutek and Jan Šnajder. Staying true to your word:(how) can attention become explanation? In *Association for Computational Linguistics*, pages 131–142, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: a survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6778–6786, 2023.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 11–20. Association for Computational Linguistics, 2019.

Kyungho Won, Moonyoung Kwon, Minkyu Ahn, and Sung Chan Jun. Eeg dataset for rsvp and p300 speller brain-computer interfaces. *Scientific Data*, 9(1):388, 2022.

Jin Xie, Jie Zhang, Jiayao Sun, Zheng Ma, Liuni Qin, Guanglin Li, Huihui Zhou, and Yang Zhan. A transformer-based approach combining deep learning network and spatial-temporal information for raw eeg classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2126–2136, 2022.

Qizhe Xie, Xuezhe Ma, Zihang Dai, and Eduard Hovy. An interpretable knowledge transfer model for knowledge base completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 950–962, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

Makoto Yamada, Yuta Umezu, Kenji Fukumizu, and Ichiro Takeuchi. Post selection inference with kernels. In *International conference on artificial intelligence and statistics*, pages 152–160. PMLR, 2018.

Fan Yang, Rina Foygel Barber, Prateek Jain, and John Lafferty. Selective inference for group-sparse linear models. In *Advances in Neural Information Processing Systems*, pages 2469–2477, 2016.

Young-Joo Yun and Rina Foygel Barber. Selective inference for clustering with unknown variance. *Electronic Journal of Statistics*, 17(2):1923–1946, 2023.

Biao Zhang, Ivan Titov, and Rico Sennrich. Sparse attention with linear units. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6507–6520, 2021.

Qingyuan Zhao, Dylan S Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):382–413, 2022.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35(12), pages 11106–11115, 2021.