Statistical Test for Attention Maps in Vision Transformers

Tomohiro Shiraishi [1] **Daiki Miwa** [2] **Teruyuki Katsuoka** [1] **Vo Nguyen Le Duy** [3 4 5] **Kouichi Taji** [1] **Ichiro Takeuchi** [1 5]

Abstract

The Vision Transformer (ViT) demonstrates exceptional performance in various computer vision tasks. Attention is crucial for ViT to capture complex wide-ranging relationships among image patches, allowing the model to weigh the importance of image patches and aiding our understanding of the decision-making process. However, when utilizing the attention of ViT as evidence in high-stakes decision-making tasks such as medical diagnostics, a challenge arises due to the potential of attention mechanisms erroneously focusing on irrelevant regions. In this study, we propose a statistical test for ViT's attentions, enabling us to use the attentions as reliable quantitative evidence indicators for ViT's decision-making with a rigorously controlled error rate. Using the framework called selective inference, we quantify the statistical significance of attentions in the form of p-values, which enables the theoretically grounded quantification of the false positive detection probability of attentions. We demonstrate the validity and the effectiveness of the proposed method through numerical experiments and applications to brain image diagnoses.

1. Introduction

The Vision Transformer (ViT) (Dosovitskiy et al., 2020) demonstrates exceptional performance in various computer vision tasks by replacing traditional Convolutional Neural Networks (CNNs) with transformer-based architectures. ViT divides images into fixed-size patches and processes them using self-attention mechanisms, capturing wide-range dependencies. This enables the model to effectively learn spatial relationships and contextual information, surpassing the limitations of CNNs in handling global context (Wu et al., 2020; Henaff, 2020; Xiao et al., 2021; Touvron et al., 2021; Jia et al., 2021; Khan et al., 2022).

In ViT, attention plays a pivotal role in capturing complex visual relationships by allowing the model to weigh the importance of different image regions. The interpretability of attention mechanisms is crucial for understanding how the model makes decisions. ViT's attention mechanisms enable the identification of salient features and contribute to the model's ability to recognize patterns.

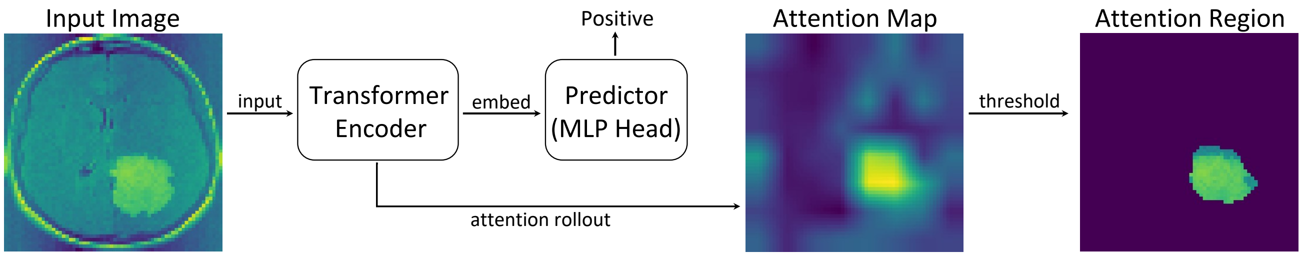
However, when utilizing ViT's attentions as evidence in high-stakes decision-making tasks such as medical diagnostics or autonomous driving (Dai et al., 2021; He et al., 2023; Prakash et al., 2021; Hu et al., 2022), a challenge arises due to the potential of attention mechanisms erroneously focusing on irrelevant regions. In this study, we propose a statistical test for ViT's attentions, enabling the quantification of the false positive detection probability of attentions in the form of p-values. This enables us to use the attentions as reliable quantitative evidence indicators for ViT's decision-making with a rigorously controlled error rate.

For example, in medical field, wrongly identifying the brain tumor regions leads to wrong treatments and patients may become apprehensive about seeking medical help, fearing misdiagnosis and unnecessary treatments. Similarly, in autonomous driving, wrongly identifying the road signs or obstacles can lead to inappropriate actions, resulting in traffic accidents. Therefore, in high-stakes decision-making tasks, it is crucial to control the probability of false positives, often referred to as the type I error rate in the statistical inference literature.

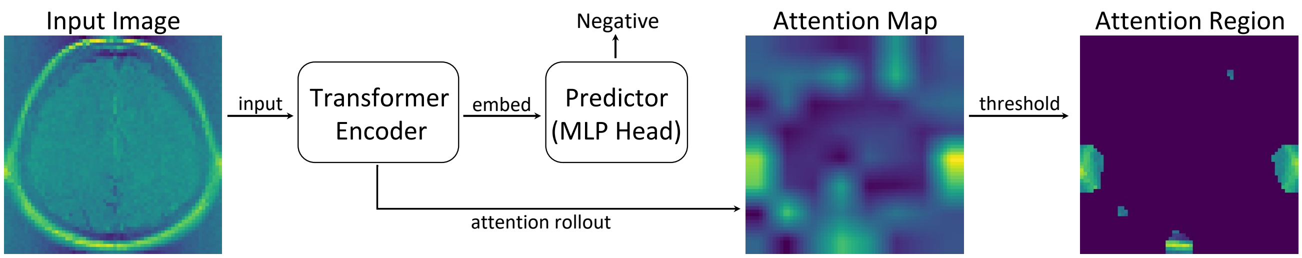
To our knowledge, there are no prior studies that investigate the statistical significance of ViT's attentions. The challenge in assessing the statistical significance of ViT's attentions stems from the inherent selection bias in ViT's attention mechanism. Testing image patches with high attention is biased, given that the ViT selects these patches by looking at the image itself. Consequently, it becomes imperative to develop an appropriate statistical test that can account for selection bias by properly considering the complex attention mechanism of ViT.

[1]Nagoya University, Aichi, Japan [2]Nagoya Institute of Technology, Aichi, Japan [3]University of Information Technology, Ho Chi Minh City, Vietnam [4]Vietnam National University, Ho Chi Minh City, Vietnam [5]RIKEN, Tokyo, Japan. Correspondence to: Ichiro Takeuchi <takeuchi.ichiro.n6@f.mail.nagoya-u.ac.jp>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(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 0.801 (true negative).

Figure 1. Schematic illustration of the problem setup and the proposed method on a brain image dataset. By inputting a brain image into the trained ViT classifier, the attention map is obtained, which indicates the area on which the ViT model 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 §4) cannot be used to properly control the false positive (type I error) rate. Instead, the selective p-value (introduced in §2) can be used to detect true positives while controlling the false positive rate at the specified level.

In this study, we address this challenge by employing selective inference (SI) (Lee et al., 2016; Taylor & Tibshirani, 2015). SI is a statistical inference framework that has gained recent interest for testing data-driven hypotheses. By considering the selection process itself as part of the statistical analysis, SI effectively addresses the selection bias issue in statistical testing when the hypotheses are selected in a data-driven manner.

Related Works. SI was initially developed for the statistical inference for feature selection in linear models (Fithian et al., 2015; Tibshirani et al., 2016; Loftus & Taylor, 2014; Suzumura et al., 2017; Le Duy & Takeuchi, 2021; Sugiyama et al., 2021) and later extended to other problem settings (Lee et al., 2015; Choi et al., 2017; Chen & Bien, 2020; Tanizaki et al., 2020; Duy et al., 2020; Gao et al., 2022). In the context of deep learning, SI was first introduced by Duy et al. (2022) and Miwa et al. (2023), where the authors proposed a computational algorithm for SI by exploiting the fact that a class of CNNs can be described as piecewise linear functions of the input image. Later, based on the algorithm proposed by Duy et al. (2022) and Miwa et al. (2023), SI was extended to anomaly detection with VAEs (Miwa et al., 2024) and to diffusion models (Katsuoka et al., 2024). However, this algorithm cannot be applied to the transformer-based architectures, as the self-attention mechanism are not piecewise linear. In this study, we introduce a new computational approach to develop SI for ViT's attentions.

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 brain image diagnosis task. The upper panel shows a brain image with a tumor region, in which we want the attentions 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 tumor regions, in which we want the attentions to be determined as statistically not 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 0.801, concluding that it is not statistically significant (true negative).

Contributions. Our contributions in this study are as follows. The first contribution is the introduction of a theoretically guaranteed framework for testing the statistical significance of ViT's Attention (§2). The second contribution involves the development of the SI method for ViT's attention, for which we introduce a new computational method for computing the p-values without selection bias (§3). The third contribution involves demonstrating the effectiveness of the proposed method through its applications to synthetic data simulations and brain image diagnosis (§4). For reproducibility, our implementation is available at https://github.com/shirara1016/statistical_test_for_vit_attention.

2. Statistical Test for ViT's Attentions

In this study, we aim to quantify the statistical significance of the attention regions identified by a trained ViT model. The details of the structure of the ViT model we used in our experiments are shown in Appendix A.1.

Notations. Let us consider an n-dimensional image as a random variable

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

where $\boldsymbol{\mu} \in \mathbb{R}^n$ is the pixel intensity vector 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 pixel intensities $\boldsymbol{\mu}$, 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 $\boldsymbol{X}$ as input and outputs attention scores $\mathcal{A}_i(\boldsymbol{X}) \in [0,1]$ for each pixel $i \in [n]$. The details of its computation are given in Appendix A.2. We define the attention region $\mathcal{M}_{\boldsymbol{X}}$ of an image $\boldsymbol{X}$ as the set of pixels with attention scores greater than a given threshold value $\tau \in (0,1)$, i.e.,

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

Statistical Inference. To quantify the statistical significance of the attention region $\mathcal{M}_{\boldsymbol{X}}$ of an image $\boldsymbol{X}$, we propose to consider the following hypothesis testing problem:

$$\begin{aligned} &\mathrm{H}_0\colon \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|} \sum_{i \in \mathcal{M}_{\boldsymbol{X}}} \mu_i = \frac{1}{|\mathcal{M}_{\boldsymbol{X}}^c|} \sum_{i \notin \mathcal{M}_{\boldsymbol{X}}} \mu_i \\ &\qquad\qquad \text{v.s.} \\ &\mathrm{H}_1\colon \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|} \sum_{i \in \mathcal{M}_{\boldsymbol{X}}} \mu_i \neq \frac{1}{|\mathcal{M}_{\boldsymbol{X}}^c|} \sum_{i \notin \mathcal{M}_{\boldsymbol{X}}} \mu_i, \end{aligned} \tag{2}$$

where H_0 is the null hypothesis that the mean pixel intensity inside and outside the attention region are equal, while H_1 is the alternative hypothesis that they are not equal. A reasonable choice of the test statistic for the statistical test in (2) is the difference in the average pixel values between inside and outside the attention region, i.e.,

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

where $\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}} = \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|} \mathbf{1}^n_{\mathcal{M}_{\boldsymbol{X}}} - \frac{1}{|\mathcal{M}^c_{\boldsymbol{X}}|} \mathbf{1}^n_{\mathcal{M}^c_{\boldsymbol{X}}}$ 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 vector whose elements are set to 1 if they belong to the set $\mathcal{C} \subset [n]$, and 0 otherwise. In this study, we consider the following standardized test statistic:

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

The p-value for the hypothesis testing problem 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.

Our main idea in this formulation is to quantify whether pixels selected as attention regions by ViT are statistically significantly different from the regions that were not selected. Although we consider the average difference in pixel values in the above formulation for the sake of simplicity, similar formulations are also possible for other image features obtained by applying appropriate image filters.

Conditional Distribution. To compute the p-value, we need to identify the sampling distribution of the test statistic $T(\boldsymbol{X})$. 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 ViT), the sampling distribution of the test statistic $T(\boldsymbol{X})$ is too complicated to characterize. Then, we consider the conditional sampling distribution of the test statistic $T(\boldsymbol{X})$ given the event $\{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{X}^{\mathrm{obs}}}\}$, i.e.,

$$T(\boldsymbol{X}) \mid \{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{X}^{\mathrm{obs}}}\}, \tag{3}$$

where $\boldsymbol{X}^{\mathrm{obs}}$ is the observed image. This conditioning means that we consider the rarity of the observation $\boldsymbol{X}^{\mathrm{obs}}$ only in the case where the same attention region $\mathcal{M}_{\boldsymbol{X}}$ as observed $\mathcal{M}_{\boldsymbol{X}^{\mathrm{obs}}}$ is obtained. The advantage of considering the conditional sampling distribution in (3) is that, by conditioning on the attention region $\mathcal{M}_{\boldsymbol{X}}$, the test statistic $T(\boldsymbol{X})$ is written as a linear function of $\boldsymbol{X}$, which allows us to characterize the sampling property of the test statistic $T(\boldsymbol{X})$.

Selective p-value. Statistical hypothesis testing based on the conditional sampling distribution has been studied within the framework of SI (also known as post-selection

[1]We discuss the robustness of the proposed method when the covariance matrix is unknown and the noise deviates from the Gaussian distribution in our experiments (§4).

inference). In this study, we also utilize the SI framework to perform statistical hypothesis testing in (2) based on the conditional sampling distribution in (3). For the tractable computation of the conditional sampling distribution in (3), we consider an additional condition on the sufficient statistic of the nuisance parameter $\mathcal{Q}_{\boldsymbol{X}}$, defined as

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

Remark 2.1. The nuisance component $\mathcal{Q}_{\boldsymbol{X}}$ corresponds to the component $\boldsymbol{z}$ in the seminal paper (Lee et al., 2016) (see Sec. 5, Eq. (5.2), and Theorem 5.2). We note that additionally conditioning on $\mathcal{Q}_{\boldsymbol{X}}$, which is required for technical purpose, is a standard approach in the SI literature and it is used in almost all the SI-related works that we cited.

The selective p-value is then computed as

$$p_{\text{selective}} = \mathbb{P}_{\text{H}_0}(|T(\boldsymbol{X})| > |T(\boldsymbol{X}^{\text{obs}})| \mid \boldsymbol{X} \in \mathcal{X}), \quad (4)$$

where $\mathcal{X} = \{\boldsymbol{X} \in \mathbb{R}^n \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{X}^{\text{obs}}}, \mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{X}^{\text{obs}}}\}$.

Theorem 2.2. *The selective p-value in (4) 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 2.2 is presented in Appendix B.1. This theorem guarantees that the selective p-value is uniformly distributed under the null hypothesis H_0 and then used to conduct the valid statistical inference for the attention region $\mathcal{M}_{\boldsymbol{X}}$.

3. Computing Selective p-values

In this section, we propose a novel computational procedure for the selective p-values in (4).

Characterization of the Conditional Data Space. To compute the selective p-values in (4), we need to characterize the conditional data space $\mathcal{X}$. According to the conditioning on the nuisance parameter $\mathcal{Q}_{\boldsymbol{X}}$, the conditional data space $\mathcal{X}$ is restricted to a one-dimensional line in $\mathbb{R}^n$.

Lemma 3.1. *The set $\mathcal{X}$ can be re-written, using a scalar parameter $z \in \mathbb{R}$, as*

$$\mathcal{X} = \{\boldsymbol{X}(z) \in \mathbb{R}^n \mid \boldsymbol{X}(z) = \boldsymbol{a} + \boldsymbol{b}z, \ z \in \mathcal{Z}\}$$

where vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ are defined as

$$\boldsymbol{a} = \mathcal{Q}_{\boldsymbol{X}^{\text{obs}}}, \ \boldsymbol{b} = \Sigma \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}^{\text{obs}}}} \Big/ \sqrt{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}^{\text{obs}}}}^\top \Sigma \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}^{\text{obs}}}}} \ ,$$

and the region $\mathcal{Z}$ is defined as

$$\mathcal{Z} = \{z \in \mathbb{R} \mid \mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{M}_{\boldsymbol{X}^{\text{obs}}}\}.$$

The proof of Lemma 3.1 is presented in Appendix B.2. This characterization of the conditional data space is first proposed by Liu et al. (2018) and used in many other SI studies. Let us consider a random variable $Z \in \mathbb{R}$ and its observation $z^{\text{obs}} \in \mathbb{R}$ such that they respectively satisfy $\boldsymbol{X} = \boldsymbol{a} + \boldsymbol{b}Z$ and $\boldsymbol{X}^{\text{obs}} = \boldsymbol{a} + \boldsymbol{b}z^{\text{obs}}$. The selective p-value in (4) is re-written as

$$p_{\text{selective}} = \mathbb{P}_{\text{H}_0}(|Z| > |z^{\text{obs}}| \mid Z \in \mathcal{Z}). \quad (5)$$

Because the unconditional variable $Z \sim \mathcal{N}(0, 1)$ under the null hypothesis H_0[2], the conditional random variable $Z \mid Z \in \mathcal{Z}$ follows the truncated standard Gaussian distribution. Once the truncated region $\mathcal{Z}$ is identified, the selective p-value in (5) can be easily computed. Thus, the remaining task is reduced to the characterization of $\mathcal{Z}$.

Reformulation of the Truncated Region. Based on the definition of the attention region in (1), the condition part of the set $\mathcal{Z}$ can be reformulated as

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

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

$$f_i(z) = \begin{cases} \tau - \mathcal{A}_i(\boldsymbol{a} + \boldsymbol{b}z) & (i \in \mathcal{M}_{\boldsymbol{X}^{\text{obs}}}) \\ \mathcal{A}_i(\boldsymbol{a} + \boldsymbol{b}z) - \tau & (i \notin \mathcal{M}_{\boldsymbol{X}^{\text{obs}}}) \end{cases}. \quad (6)$$

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

$$\mathcal{Z} = \bigcap_{i \in [n]} \{z \in \mathbb{R} \mid f_i(z) < 0\}. \quad (7)$$

Selective p-value Computation by Adaptive Grid Search. The problem of finding $\mathcal{Z}$ in (7) is reduced to the problem of enumerating all solutions to the nonlinear equations $f_i(z) = 0$ for each $i \in [n]$ in (6). The difficulty of this problem depends on the continuity, differentiability, and smoothness of the functions $f_i, \ i \in [n]$. Fortunately, since the function f_i is a part of the attention map computation in the ViT model, it is continuous, (sub)differentiable, and possesses a certain level of smoothness (except for pathological cases). Assuming the certain degree of smoothness of the function f_i, by adaptively generating grid points in the one-dimensional space $z \in \mathbb{R}$ and computing the values

[2] The random variable Z corresponds to the test statistic $T(\boldsymbol{X})$. Then, $Z \sim \mathcal{N}(0, 1)$ is obtained by the linearity of the test statistic $T(\boldsymbol{X})$ with respect to $\boldsymbol{X}$ and the fact that the test statistic $T(\boldsymbol{X})$ is already standardized in the definition.

of $f_i(z)$ at each grid point, it is possible to identify $\mathcal{Z}$ in (7) with sufficient accuracy. This further means that it is possible to compute the selective p-value in (5) with sufficient accuracy (as stated in Theorem 3.2 later).

The overall procedure for estimating the selective p-value by an adaptive grid search method is summarized in Algorithm 1. Here, S represents the grid search interval $[-S, S]$, $\varepsilon_{\min}$ and $\varepsilon_{\max}$ represent the minimum and maximum grid width, respectively. Note that, in line 9 of Algorithm 1, we added the interval $J(z^{\mathrm{obs}})$ that overlaps with the grid points for computational simplicity. The key of Algorithm 1 lies in how to determine the adaptive grid size $d(z_j)$.

Algorithm 1 Selective p-value Computation by Adaptive Grid Search

Require: $S, \varepsilon_{\min}, \varepsilon_{\max}, \{f_i\}_{i\in[n]}$ and $z^{\mathrm{obs}} := T(\boldsymbol{X}^{\mathrm{obs}})$

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: **end while**
7: $d^{\mathrm{obs}} \leftarrow \min(\varepsilon_{\max}, d(z^{\mathrm{obs}}))$
8: $J(z^{\mathrm{obs}}) \leftarrow [z^{\mathrm{obs}} - d^{\mathrm{obs}}, z^{\mathrm{obs}} + d^{\mathrm{obs}}]$
9: $\mathcal{Z}^{\mathrm{grid}} \leftarrow \cup_{j|z_j\in\mathcal{Z}}[z_j, z_{j+1}] \cup J(z^{\mathrm{obs}})$
10: $p_{\mathrm{grid}} \leftarrow \mathbb{P}_{\mathrm{H}_0}(|Z| > |z^{\mathrm{obs}}| \mid Z \in \mathcal{Z}_{\mathrm{grid}})$, where $Z \sim \mathcal{N}(0,1)$

Ensure: 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 3.2. *Assume that f_i is differentiable and Lipschitz continuous for all $i \in [n]$. Assume further that f_i has at most only a finite number of zeros, at any of which the value of f_i' 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

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

The proof of Theorem 3.2 is presented in Appendix B.3. The following lemma suggests why it is reasonable to define the grid width as $d(z_j)$ in Theorem 3.2.

Lemma 3.3. *For the grid width $d(z_j)$ defined in Theorem 3.2, 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 3.3 is presented in Appendix B.4. The main idea of the proof is that f_i has 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 shown in Figure 2). In Algorithm 1, we take the max operation in line 4 to avoid the case where the grid width is too small and then the grid point is stuck in $\mathcal{Z}$ or $\mathbb{R} \setminus \mathcal{Z}$.

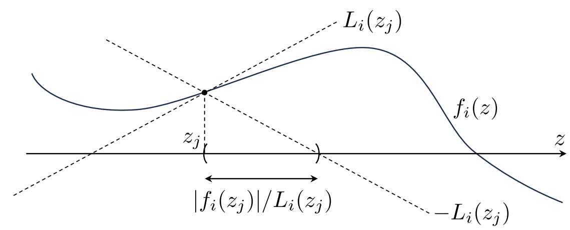


Figure 2. Schematic illustration of the relationship between grid width and Lipschitz constant.

Implementation Teqniques. For implementation, we need to define the grid width $d(z_j)$ in a computable form. Actually, computing $d(z_j)$ as defined in Theorem 3.2 presents a challenge since it requires the computation of the Lipschitz constant $L_i(z_j)$ of the attention score in the vicinity of z_j. In this study, we define $d(z_j)$ as in Theorem 3.2, by 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 z^{obs}. In the case where z_j is far from z^{obs} (i.e., $|z_j - z^{\mathrm{obs}}| > 0.1$), 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 i such that $f_i(z_j)$ and $f_i'(z_j)$ have the same sign, since f_i is assumed to be approximated by a linear function. This can be implemented by taking the min or max operation only for i such that $f_i(z_j)f_i'(z_j) < 0$. In contrast, in the case where z_j is close to z^{obs} (i.e., $|z_j - z^{\mathrm{obs}}| < 0.1$), 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, because it implies that the grid point z_j is close to the boundary of $\mathcal{Z}$. Therefore, it may not be reasonable to utilize the derivative of f_i in the same way as above, so we assume that $L_i(z_j) = 1$. This assumption is highly conservative, since the range of the attention score $\mathcal{A}_i$ is $[0, 1]$. The schematic illustration of these heuristics is shown in Figure 3.

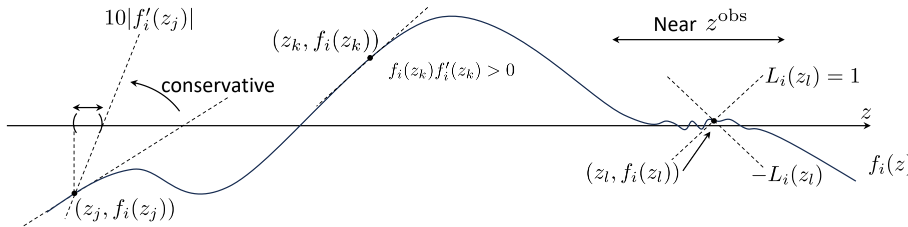


Figure 3. Schematic illustration of the introduced heuristics. The left and central part of the figure show the case where the grid point is far from z^{obs} and the right part of the figure shows the case where the grid point is close to z^{obs}. In the left part where z_j is far from z^{obs}, the function f_i is approximated by a linear function and the Lipschitz constant $L_i(z_j)$ is conservatively set to $10|f_i'(z_j)|$. In the central part where the function f_i is approximated by a linear function, the sign does not change as long as $f_i(z_k)f_i'(z_k)$ is positive. In the right part where z_l is close to z^{obs}, the function f_i may exhibit a flat shape or micro oscillations and tends to take values close to zero.

Derivative of the Attention Map. 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 $\mathcal{A}$, which is the output of the ViT model. Auto differentiation, which is implemented in many deep learning frameworks (e.g., TensorFlow and PyTorch), can be used to compute this derivative. It should be noted that we are to differentiate an n-dimensional attention map with respect to a scalar input z_j. When output dimension is large, reverse-mode auto differentiation (also called backpropagation) is generally inefficient. In these cases, forward-mode auto differentiation is a better option. However, it is not well supported in many frameworks and may require more implementation costs than using the back-mode auto differentiation. We modularized the operations specific to the ViT model for differentiating the attention map using forward-mode auto differentiation in TensorFlow. This allows us to differentiate the attention map for ViTs of any architecture without incurring additional implementation costs. For details, see our implementation code.

4. Numerical Experiments

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

Experimental Setup. We first trained the ViT classifier model on the synthetic dataset. We created a synthetic dataset by generating 1,000 negative images $\boldsymbol{X} = (X_1, \ldots, X_n) \sim \mathcal{N}(\mathbf{0}, I)$ and 1,000 positive images $\boldsymbol{X} = (X_1, \ldots, X_n) \sim \mathcal{N}(\boldsymbol{\mu}, I)$. The pixel intensity vector $\boldsymbol{\mu}$ was set to $\mu_i = \Delta,\ \forall i \in \mathcal{S}$ and $\mu_i = 0,\ \forall i \in [n] \setminus \mathcal{S}$, where Δ was uniformly sampled from $\mathcal{U}_{[1,4]}$ and $\mathcal{S}$ is the region to focus on whose location was randomly determined. After training process, we experimented with the trained ViT model on the test dataset. We input the test image to the trained ViT model and obtained the attention map, and then performed the statistical test for the obtained attention map. In all experiments, we set the threshold value $\tau = 0.6$, the grid search interval $[-S, S]$ with $S = 10 + |z^{\mathrm{obs}}|$, the minimum grid width $\varepsilon_{\min} = 10^{-4}$, the maximum grid width $\varepsilon_{\max} = 0.2$, and the significance level $\alpha = 0.05$. We considered two types of covariance matrices: $\Sigma = I_n \in \mathbb{R}^{n \times n}$ (independence) and $\Sigma = (0.5^{|i-j|})_{ij} \in \mathbb{R}^{n \times n}$ (correlation).

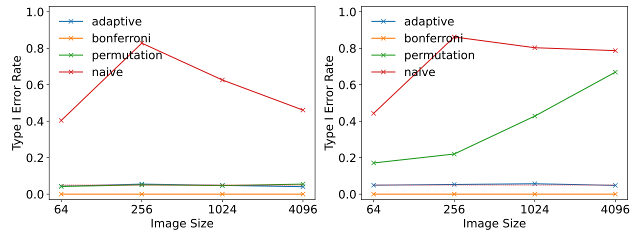
For the experiments to see the type I error rate, we considered two options: for image size in {64, 256, 1024, 4096} and for architecture in {small, base, large, huge} (the details of architectures are presented in Appendix C.2). If not specified, we used the image size of 256 and the architecture of base. In each setting, we generated 100 null test images $X = (X_1, \ldots, X_n) \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and ran 10 trials (i.e., 1,000 null images in total). Here, the first 2 trials were also used for comparing the computation time. Regarding our proposed method, we ran additional 90 trials to carefully check the validity in controlling the type I error rate at three significance levels $\alpha = 0.05, 0.01, 0.10$. To investigate the power, we set image size to 256 and architecture to base and generated 1,000 test images $\boldsymbol{X} = (X_1, \ldots, X_n) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. The pixel intensity vector $\boldsymbol{\mu}$ was set to $\mu_i = \Delta,\ \forall i \in \mathcal{S}$ and $\mu_i = 0,\ \forall i \in [n] \setminus \mathcal{S}$, where $\mathcal{S}$ is the region to focus on whose location was randomly determined. We set $\Delta \in \{1.0, 2.0, 3.0, 4.0\}$.

Results. The results of type I error rate are shown in Figures 4 and 5. The `adaptive` and `bonferroni` successfully controlled the type I error rate under the significance level in all settings, whereas the other two methods `naive` and `permutation` could not. Because the `naive` and `permutation` failed to control the type I error rate, we no longer considered their powers. The results of power comparison are shown in Figure 6 and we confirmed that the `adaptive` has much higher power than the `bonferroni` in all settings. The results of computation time are shown in Figures 7 and 8. In all settings, the `adaptive` outperforms the `fixed` and `combination` while utilizing the smallest minimum grid width. The results of additional trials for the type I error rate are shown in Figures 9 and 10. We confirmed that our proposed method can properly control the type I error rate at multiple significance levels in all settings for more trials.

Additionally, we confirmed the robustness of our proposed method in terms of type I error rate control for two cases: where the covariance matrix is estimated from the same data, and where the noise follows one of five non-Gaussian distribution families. More details can be found in Appendix C.3. The results are shown in Figures 11 and 12. Our method still maintains good performance in type I error rate control.

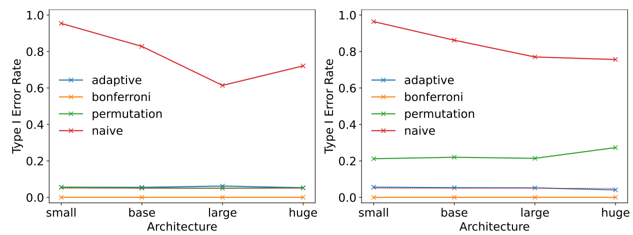
Discussion. Our experiments confirmed that the approximation approach works well with the heuristics considered in §3 for the attention map in the ViT model. We assess the reasonableness of the heuristics in §3 by presenting several examples of our target function f_i defined in (6) in Figure 13. The plots demonstrate that the function f_i is generally consistent with the heuristics, having a shape that can be approximated linearly when z is away from $z^{\rm obs}$ and tending to take values close to zero when z is close to $z^{\rm obs}$. The input image of the ViT model is written as $\boldsymbol{a}+\boldsymbol{b}z = \boldsymbol{X}^{\rm obs}+\boldsymbol{b}(z-z^{\rm obs})$, where $\boldsymbol{b}$ is the vector parallel to $\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}^{\rm obs}}}$ from the definition. For z away from $z^{\rm obs}$, the input $\boldsymbol{X}^{\rm obs} + \boldsymbol{b}(z - z^{\rm obs})$ results in an image where the pixel intensity in the attention region $\mathcal{M}_{\boldsymbol{X}^{\rm obs}}$ is highlighted. Then, each attention score $\mathcal{A}_i$ may exhibit a gradual trend. On the other hand, for z close to $z^{\rm obs}$, from the definition of f_i in (6) and continuity, it is expected that some f_i are close to zero. However, it is unclear whether these heuristics are always valid for any complex Transformer architectures. In some cases, it may be necessary to sufficiently reduce the grid size to account for highly nonlinear functions with increased computational costs.

Real Data Experiments. We examined the brain image dataset extracted from the dataset used in Buda et al. (2019), which included 939 and 941 images with and without tumors, respectively. We selected 100 images without tumors to estimate the variance and used 700 images each with and without tumors for training the ViT classifier model. The remaining images with and without tumors were used for testing, i.e., to demonstrate the advantages of the proposed method. The results of the `adaptive` and `naive` are shown in Figure 14. The naive p-values remain small even for images without tumors, which indicates that naive p-values cannot be used to quantify the reliability of the attention regions. In contrast, the adaptive p-values are large for images without tumors and small for images with tumors. This result indicates that the `adaptive` can detect true positive cases while avoiding false positive detections.



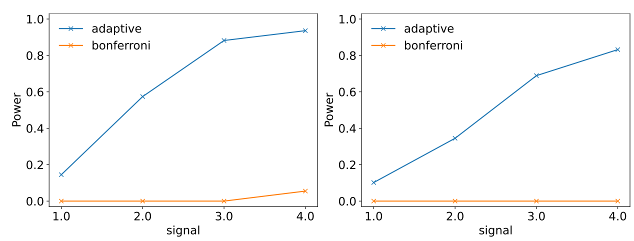
(a) Independence (b) Correlation

Figure 4. Type I Error Rate when changing the image size. Only our proposed method and the bonferroni correction are able to control the type I error rate in all settings.



(a) Independence (b) Correlation

Figure 5. Type I Error Rate when changing the architecture. Only our proposed method and the bonferroni correction are able to control the type I error rate in all settings.

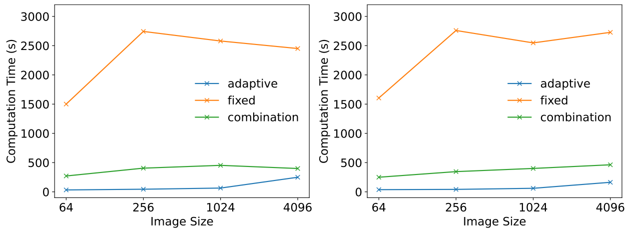


(a) Independence (b) Correlation

Figure 6. Power when changing the signal intensity. Our proposed method has much higher power than the bonferroni correction.

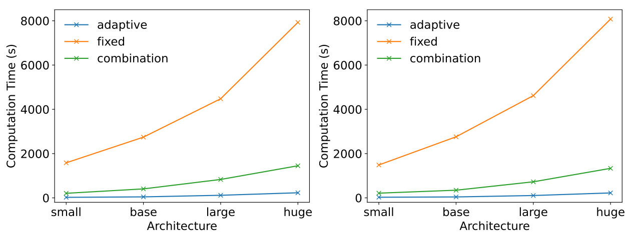
5. Conclusion

In this study, we introduced a novel framework for testing the statistical significance of ViT's Attention based on the concept of SI. We developed a new computational method for calculating the p-values, which are used as an indicator of statistical significance. One current limitation of the proposed method is its computational cost, which makes it difficult to apply to high-resolution images and very huge architectures. The introduction of reasonable and stronger heuristics is a possible future improvement that would make the proposed method more widely applicable. We believe that this study opens an important direction in ensuring the reliability of ViT's Attention.



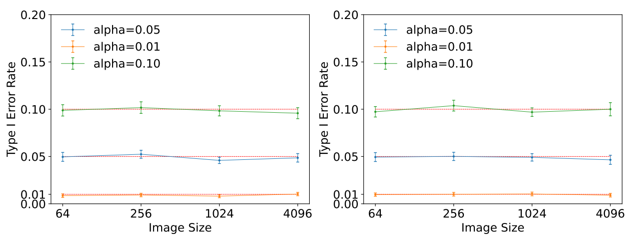
(a) Independence (b) Correlation

Figure 7. Computational Time when changing the image size. Our proposed method outperforms the other two grid search options while utilizing the smallest minimum grid width.



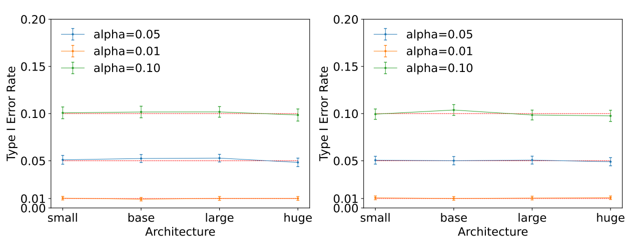
(a) Independence (b) Correlation

Figure 8. Computation Time when changing the architecture. Our proposed method outperforms the other two grid search options in all settings while utilizing the smallest minimum grid width.



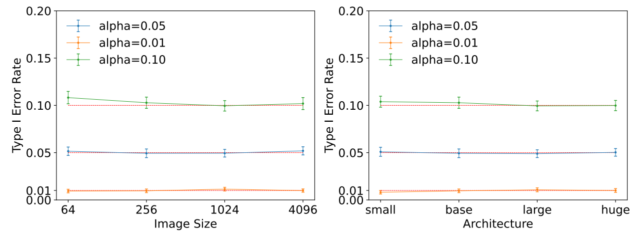
(a) Independence (b) Correlation

Figure 9. Type I Error Rate from 10,000 null images for three significance levels, when changing the image size. Error bars indicate the 95% confidence interval of the type I error rate. Our proposed method can properly control the type I error rate at multiple significance levels in all settings for more trials.



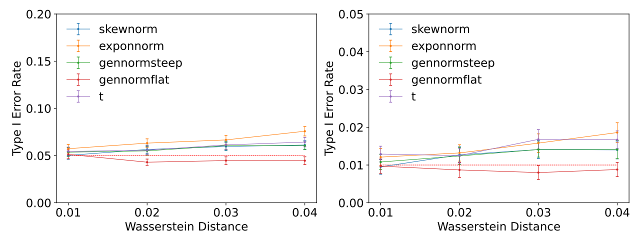
(a) Independence (b) Correlation

Figure 10. Type I Error Rate from 10,000 null images for three significance levels, when changing the architecture. Error bars indicate the 95% confidence interval of the type I error rate. Our proposed method can properly control the type I error rate at multiple significance levels in all settings for more trials.



(a) Image Size (b) Architecture

Figure 11. Robustness of Type I Error Rate Control. Our proposed method can robustly control the type I error rate even when the covariance matrix is estimated from the same data.



(a) Significance Level 0.05 (b) Significance Level 0.01

Figure 12. Robustness of Type I Error Rate Control. Our proposed method can robustly control the type I error rate, albeit slightly above the significance level, even when the noise follows non-Gaussian distributions.

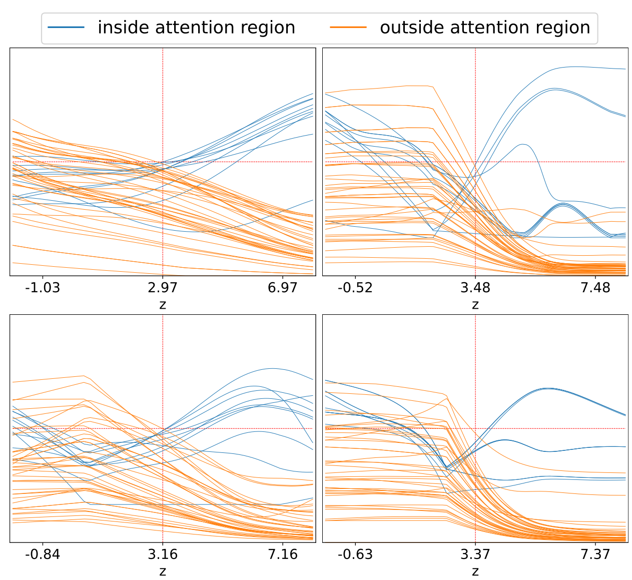
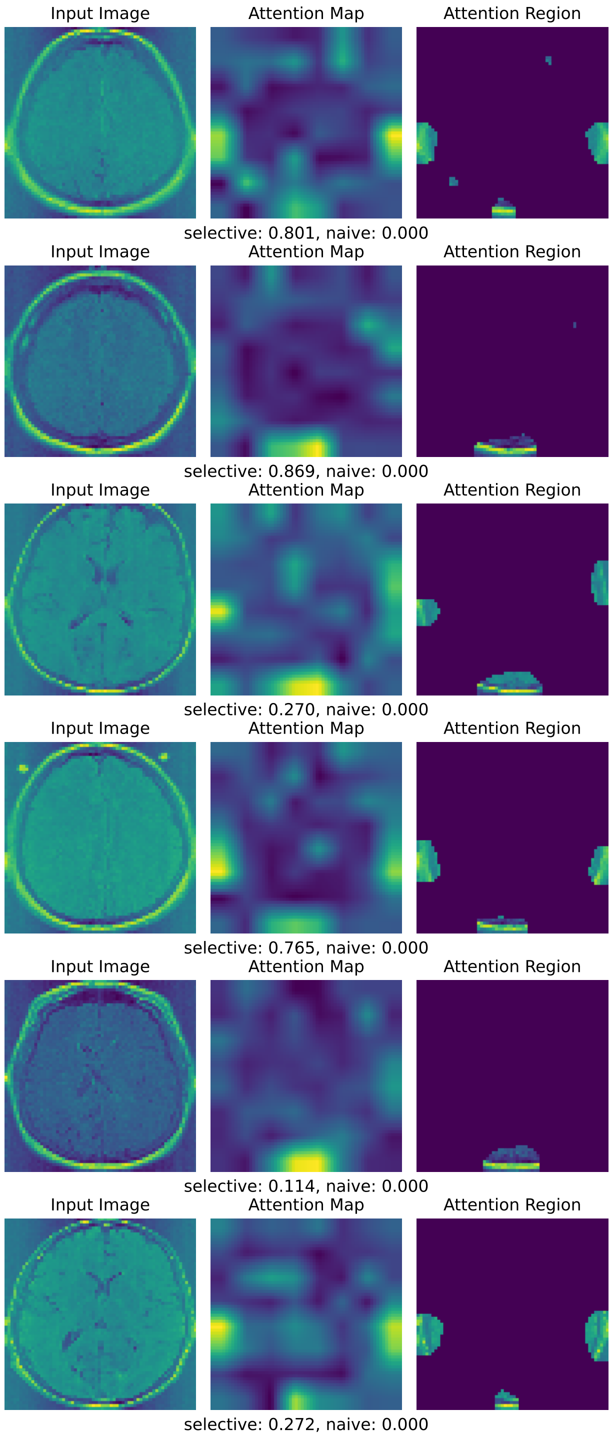
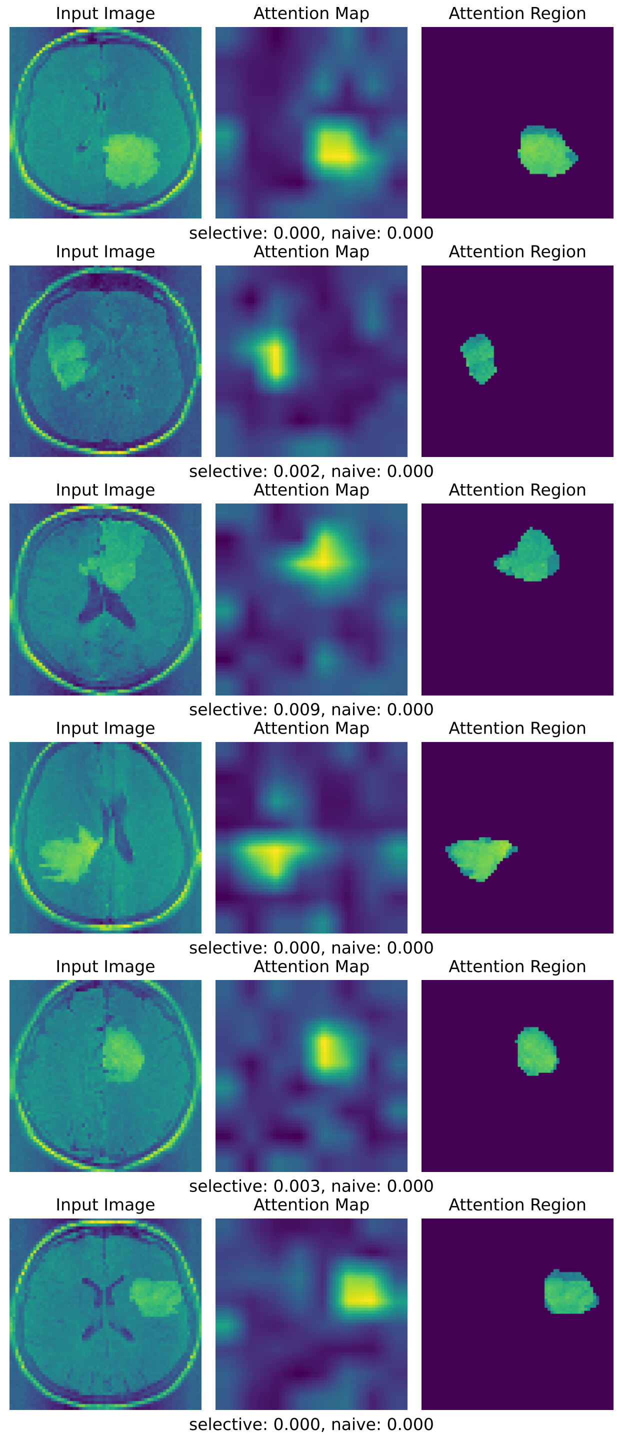


Figure 13. Demonstration of the target function f_i. We set image size to 256 and architecture to base, and the image was generated from the standard normal distribution $\mathcal{N}(\mathbf{0}, I)$. The vertical red line indicates the observed test statistic z^{obs} and the horizontal red line indicates zero. The blue plots display f_i values for 10 randomly selected i from $\mathcal{M}_{\boldsymbol{X}^{\mathrm{obs}}}$, while the orange plots display f_i values for 40 randomly selected i from $\mathcal{M}^{\mathrm{c}}_{\boldsymbol{X}^{\mathrm{obs}}}$. We note that the region on which all f_i values lie below zero (horizontal red line) is the truncated region $\mathcal{Z}$.



(a) Brain images without tumor



(b) Brain images with tumor

Figure 14. Demonstration on brain image dataset. Our proposed method conclude that attentions are statistically significant for images with tumors while avoiding falsely detection of significance for images without tumors.

Acknowledgements

This work was partially supported by MEXT KAKENHI (20H00601), JST CREST (JPMJCR21D3, JPMJCR22N2), JST Moonshot R&D (JPMJMS2033-05), JST AIP Acceleration Research (JPMJCR21U2), NEDO (JPNP18002, JPNP20006) and RIKEN Center for Advanced Intelligence Project.

Impact Statement

This work, which focuses on statistical tests for Vision Transformer's attentions, aims to enhance the reliability of AI and has the potential to broadly influence the machine learning community. On the other hand, it does not present significant ethical concerns or foreseeable societal consequences because this work is theoretical and, as of now, has no direct applications that might impact society or ethical considerations.

References

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

Buda, M., Saha, A., and Mazurowski, M. A. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology and Medicine*, 109:218–225, 2019. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2019.05.002. URL https://www.sciencedirect.com/science/article/pii/S0010482519301520.

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

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

Dai, Y., Gao, Y., and Liu, F. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8):1384, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Duy, V. N. L., Toda, H., Sugiyama, R., and Takeuchi, I. Computing valid p-value for optimal changepoint by selective inference using dynamic programming. In *Advances in Neural Information Processing Systems*, 2020.

Duy, V. N. L., Iwazaki, S., and Takeuchi, I. Quantifying statistical significance of neural network-based image segmentation by selective inference. *Advances in Neural Information Processing Systems*, 35:31627–31639, 2022.

Fithian, W., Taylor, J., Tibshirani, R., and Tibshirani, R. Selective sequential model selection. *arXiv preprint arXiv:1512.02565*, 2015.

Gao, L. L., Bien, J., and Witten, D. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, pp. 1–11, 2022.

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., and Shen, D. Transformers in medical image analysis. *Intelligent Medicine*, 3(1):59–78, 2023.

Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pp. 4182–4192. PMLR, 2020.

Hu, Z., Zhang, Y., Xing, Y., Zhao, Y., Cao, D., and Lv, C. Toward human-centered automated driving: A novel spatiotemporal vision transformer-enabled head tracker. *IEEE Vehicular Technology Magazine*, 17(4): 57–64, 2022.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Katsuoka, T., Shiraishi, T., Miwa, D., Duy, V. N. L., and Takeuchi, I. Statistical test for generated hypotheses by diffusion models. *arXiv preprint arXiv:2402.11789*, 2024.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Le Duy, V. N. and Takeuchi, I. Parametric programming approach for more powerful and general lasso selective inference. In *International conference on artificial intelligence and statistics*, pp. 901–909. PMLR, 2021.

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

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

Liu, K., Markovic, J., and Tibshirani, R. More powerful post-selection inference, with application to the lasso. *arXiv preprint arXiv:1801.09037*, 2018.

Loftus, J. R. and Taylor, J. E. A significance test for forward stepwise model selection. *arXiv preprint arXiv:1405.3920*, 2014.

Miwa, D., Le, D. V. N., and Takeuchi, I. Valid p-value for deep learning-driven salient region. In *Proceedings of the 11th International Conference on Learning Representation*, 2023.

Miwa, D., Shiraishi, T., Duy, V. N. L., Katsuoka, T., and Takeuchi, I. Statistical test for anomaly detections by variational auto-encoders. *arXiv preprint arXiv:2402.03724*, 2024.

Prakash, A., Chitta, K., and Geiger, A. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7077–7087, 2021.

Sugiyama, K., Le Duy, V. N., and Takeuchi, I. More powerful and general selective inference for stepwise feature selection using homotopy method. In *International Conference on Machine Learning*, pp. 9891–9901. PMLR, 2021.

Suzumura, S., Nakagawa, K., Umezu, Y., Tsuda, K., and Takeuchi, I. Selective inference for sparse high-order interaction models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3338–3347. JMLR. org, 2017.

Tanizaki, K., Hashimoto, N., Inatsu, Y., Hontani, H., and Takeuchi, I. Computing valid p-values for image segmentation by selective inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9553–9562, 2020.

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

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

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., and Vajda, P. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.

Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., and Girshick, R. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021.

A. Details of the Vision Transformers

A.1. Structure of the Vision Transformers

The overall structure of the ViT model is shown in Figure 15. In MLP, we use two fully-connected layers and set the hidden dimension to four times the #emb_dim. In Multi-Head Self-Attention, we use #heads self-attention mechanisms. Regarding the patch embedding, we set the patch size to $\min(2, \sqrt{n}/8)$ (i.e., for $\sqrt{n} = 16$ case, the patch size is 2 and then #patches is $(16/2)^2 = 64$). As the base model, we set the #layers to 8, the #emb_dim to 64, and the #heads to 4, respectively.

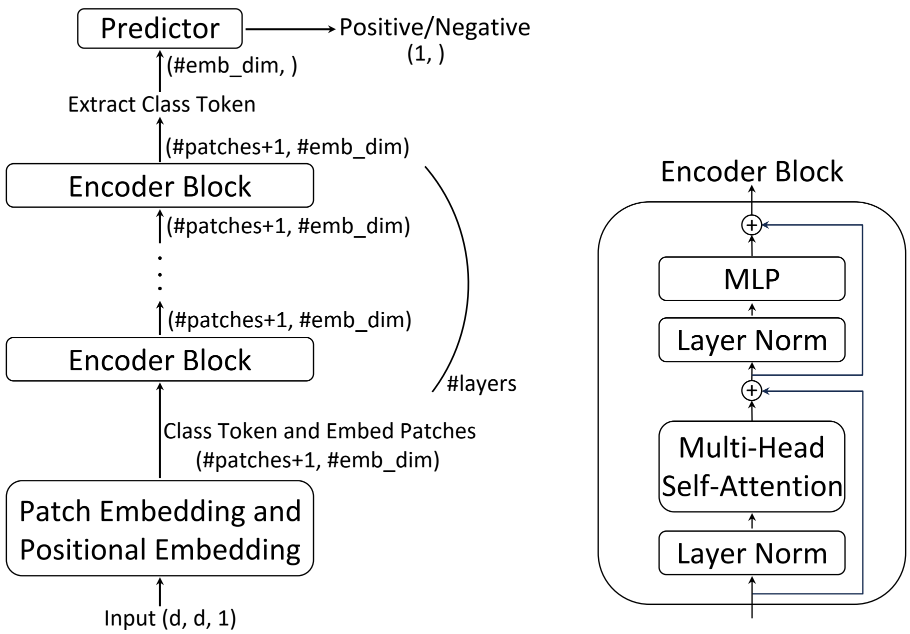


Figure 15. Structure of the Vision Transformer model.

A.2. Computation of the Attention Maps

Let we denote the #patches as N, #layers as L, the #heads as H, and the #emb_dim/#heads as D. We describe the computation of the attention map $\mathcal{A}(\boldsymbol{X}) \in [0, 1]^n$ for input image $\boldsymbol{X} \in \mathbb{R}^n$ from the ViT model based on (Abnar & Zuidema, 2020)

Obtain the Attention Weights. We reshape the input image $\boldsymbol{X} \in \mathbb{R}^n$ to $\boldsymbol{X}' \in \mathbb{R}^{d \times d}$ where $d = \sqrt{n}$, and then input it to the ViT model. In process of the ViT model, the input image $\boldsymbol{X}'$ is passing through the self-attention mechanism $H \times L$ times. In the h-th self-attention mechanism of the l-th layer, let we denote the query and key as $Q_{l,h} \in \mathbb{R}^{(N+1) \times D}$ and $K_{l,h} \in \mathbb{R}^{(N+1) \times D}$, respectively. Then, the attention weights $A_{l,h} \in \mathbb{R}^{(N+1) \times (N+1)}$ are 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 softmax operation is applied to each row of the matrix. Note that the row of $A_{l,h}$ corresponds to the queries and the column of $A_{l,h}$ corresponds to the keys. The attention map is computed by aggregating the all attention weights $\{A_{l,h}\}_{(l,h) \in [L] \times [H]}$.

Aggregate the Attention Weights. We compute the layer-wise attention weights $\hat{A}_l \in \mathbb{R}^{(N+1) \times (N+1)}$ by averaging the attention weights $A_{l,h}$ in heads direction as

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

Then, to aggregate the all attention weights to $\bar{A}$, we take the matrix product of each $\hat{A}_l$, adding the identity matrix $I \in \mathbb{R}^{(N+1)\times(N+1)}$, as

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

where matrix I represents the skip connection in Encoder Block as in Figure 15. Finally, we extract the N-dimensional vector $A \in \mathbb{R}^N$ from $\bar{A}$ as

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

which corresponds to the keys of each patch for the query of the class token as an aggregated form.

Post-Processing. We reshape the N-dimensional vector A to square matrix and upscale it to A' whose size is the same as the input image $\boldsymbol{X}'$ by using bilinear interpolation. Then, we obtain the attention map $\mathcal{A}(\boldsymbol{X}) \in [0,1]^n$ by normalizing A' with min-max normalization and flattening it to n-dimensional vector.

B. Proofs

B.1. Proof of Theorem 2.2

We proof this theorem based on the Lemma 3.1. Then, under the null hypothesis H_0, we have

$$T(\boldsymbol{X}) \mid \{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{X}^{\mathrm{obs}}}, \mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{X}^{\mathrm{obs}}}\} \sim \mathrm{TN}(0,1;\mathcal{Z}),$$

where $\mathrm{TN}(0,1;\mathcal{Z})$ is the truncated standard Gaussian distribution on $\mathcal{Z}$ and truncated region $\mathcal{Z}$ is defined in Lemma 3.1. Therefore, by probability integral transform, under the null hypothesis we have

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

which leads to

$$\mathbb{P}_{\mathrm{H}_0}(p_{\mathrm{selective}} \le \alpha \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{X}^{\mathrm{obs}}}, \mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}_{\boldsymbol{X}^{\mathrm{obs}}}) = \alpha,\ \forall \alpha \in (0,1).$$

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

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

Regarding the marginalization of the attention regions, we obtain the results as follows:

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

B.2. Proof of Lemma 3.1

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

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

where $z = T(\boldsymbol{X}) \in \mathbb{R}$. Then, we obtain the results as follows:

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

B.3. Proof of Theorem 3.2

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

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

Here, $R^1 \subset \mathcal{Z}^{\text{grid}}$ and $R^2 \subset \mathbb{R} \setminus \mathcal{Z}^{\text{grid}}$ are obvious from the definition of them, and from the Lemma 3.3, we have $R^1 \subset \mathcal{Z}$ and $R^2 \subset \mathbb{R} \setminus \mathcal{Z}$. Then, we have following subset relationships

$$R^1 \subset \mathcal{Z},\ \mathcal{Z}^{\text{grid}} \subset R^1 \cup R^3 \cup R^4 \tag{8}$$

Let us denote the probability density function of the standard Gaussian distribution as ϕ and the cumulative distribution function of that as Φ, and introduce the integrate function $\mathcal{I}$ as

$$\mathcal{I} : \mathcal{B}(\mathbb{R}) \ni R \mapsto \int_R \phi(z) dz \in [0, 1],$$

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 $R_{\text{in}}, R_{\text{out}} \in \mathcal{B}(\mathbb{R})$ as

$$R_{\text{in}} = R \cap [-|z^{\text{obs}}|, |z^{\text{obs}}|],\ R_{\text{out}} = R \setminus [-|z^{\text{obs}}|, |z^{\text{obs}}|].$$

Then, we have $p_{\text{selective}}$ and p_{grid} as

$$p_{\text{selective}} = \mathbb{P}_{\text{H}_0}(|Z| > |z^{\text{obs}}| \mid Z \in \mathcal{Z}) = \frac{\mathcal{I}(\mathcal{Z}_{\text{out}})}{\mathcal{I}(\mathcal{Z}_{\text{in}})}, \tag{9}$$

$$p_{\text{grid}} = \mathbb{P}_{\text{H}_0}(|Z| > |z^{\text{obs}}| \mid Z \in \mathcal{Z}^{\text{grid}}) = \frac{\mathcal{I}(\mathcal{Z}^{\text{grid}}_{\text{out}})}{\mathcal{I}(\mathcal{Z}^{\text{grid}}_{\text{in}})}, \tag{10}$$

respectively. Therefore, by considering the subset relationships in (8), 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

$$\begin{aligned} R^3 &\subset \bigcup_{j|I(z_j)=0} [z_j, z_{j+1}] = \bigcup_{j|I(z_j)=0} [z_j, z_j + \varepsilon_{\min}] \\ &\subset \bigcup_{j|L\varepsilon_{\min} > \min_{i \in [n]} |f_i(z_j)|} [z_j, z_j + \varepsilon_{\min}]. \end{aligned} \tag{11}$$

Continuously, we evaluate the length of R^3 by show that the set in (11) 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

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

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

$$\min_{i \in [n]} \min_{k \in [K_i]} |f_i(q_{ik} + s)| \geq \frac{h}{2}|s|. \tag{12}$$

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 (12), 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

$$\begin{aligned} \mathcal{I}(R^3) &\leq \sum_{i \in [n]} \sum_{k \in [K_i]} \phi(0) \left(\frac{4L}{h} + 2 \right) \varepsilon_{\min} \\ &\leq nK\phi(0) \left(\frac{4L}{h} + 2 \right) \varepsilon_{\min} = M_1 \varepsilon_{\min}, \end{aligned} \tag{13}$$

where $M_1 = nK\phi(0)\,(4L/h+2)$ 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

$$\begin{aligned}\mathcal{I}(R^1_{\text{in}}) &\geq \mathcal{I}(J(z^{\text{obs}})_{\text{in}})\\ &= \mathcal{I}([z^{\text{obs}} - d^{\text{obs}}, z^{\text{obs}} + d^{\text{obs}}] \cap [-|z^{\text{obs}}|, |z^{\text{obs}}|])\\ &\geq \phi(z^{\text{obs}})d^{\text{obs}} = M_2,\end{aligned} \tag{14}$$

where $M_2 = \phi(z^{\text{obs}})d^{\text{obs}}$ is a positive constant independent of $\varepsilon_{\min}$ and S. Finally, regarding the $\mathcal{I}(R^4)$, we have

$$\mathcal{I}(R^4) = 2\Phi(-S) \tag{15}$$

Finally, we evaluate the error bound. From (8), (9) and (10), we have

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

Therefore, by using (13), (14) and (15), we have the following error bound

$$\begin{aligned}&|p_{\text{selective}} - p_{\text{grid}}|\\ \leq&\frac{\mathcal{I}((R^1 \cup R^3 \cup R^4)_{\text{out}})}{\mathcal{I}(R^1_{\text{in}})} - \frac{\mathcal{I}(R^1_{\text{out}})}{\mathcal{I}((R^1 \cup R^3 \cup R^4)_{\text{in}})}\\ =&\frac{\mathcal{I}((R^1 \cup R^3 \cup R^4)_{\text{out}})\mathcal{I}((R^1 \cup R^3 \cup R^4)_{\text{in}}) - \mathcal{I}(R^1_{\text{out}})\mathcal{I}(R^1_{\text{in}})}{\mathcal{I}(R^1_{\text{in}})\mathcal{I}((R^1 \cup R^3 \cup R^4)_{\text{in}})}\\ =&\frac{\mathcal{I}((R^3 \cup R^4)_{\text{out}})\mathcal{I}((R^3 \cup R^4)_{\text{in}}) + \mathcal{I}(R^1_{\text{out}})\mathcal{I}(R^3 \cup R^4)}{\mathcal{I}(R^1_{\text{in}})\mathcal{I}((R^1 \cup R^3 \cup R^4)_{\text{in}})}\\ \leq&\frac{\mathcal{I}(R^3 \cup R^4)^2 + \mathcal{I}(R^3 \cup R^4)}{\mathcal{I}(R^1_{\text{in}})^2}\\ \leq&\frac{2}{M_2^2}\mathcal{I}(R^3 \cup R^4) \leq \frac{2}{M_2^2}(M_1\varepsilon_{\min} + 2\Phi(-S)).\end{aligned} \tag{16}$$

Here, based on the three equations $\lim_{x\to\infty} \phi(x)/x\phi(x) = 0$, $\Phi'(-x) = (1-\Phi(x))' = -\phi(x)$ and $\phi'(x) = -x\phi(x)$, we have the $\lim_{x\to\infty} \Phi(-x)/\phi(x) = 0$ 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 $\lim_{x\to\infty} \Phi(-x)/\phi(x) = 0$). By combining this with (16), we have

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

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.

B.4. Proof of Lemma 3.3

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

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

C. Experimental Details

C.1. Methods for Comparison

We compared our proposed method with the following methods:

- `naive`: This method uses a classical z-test without conditioning, i.e., we compute the naive p-value as

$$
p_{\text{naive}} = \mathbb{P}_{\mathrm{H}_0}\left(|Z| > |z^{\text{obs}}|\right).
$$

- `permutation`: This method uses a permutation test. The procedure is as follows:
  - Compute the observed test statistic z^{obs} by inputting the observed image $\boldsymbol{X}^{\text{obs}}$ to the ViT model.
  - For $i = 1, \ldots, B$, compute the test statistic $z^{(i)}$ by inputting the permuted image $\boldsymbol{X}^{(i)}$ to the ViT model. Here, B is the number of permutations which is set to 1,000 in our experiments.
  - Compute the permutation p-value as

$$
p_{\text{permutation}} = \frac{1}{B} \sum_{b \in [B]} \mathbf{1}\{|z^{(b)}| > |z^{\text{obs}}|\},
$$

  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. The number of all possible attention regions is 2^n, then we compute the bonferroni p-value 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 the grid point z_j which satisfies $|z_j - z^{\text{obs}}| < 0.1$ and $\varepsilon = 10^{-2}$ for the remaining grid points.

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

C.2. Architectures to Compare

The architectures to compare are shown in Table 1. The details of the structure of the ViT model we used are shown in Appendix A.1.

C.3. Details of Robustness Experiment

Estimated Covariance Matrix. We considered the two options: for image size in {64, 256, 1024, 4096} and for architecture in {small, base, large, huge} as same as the type I error rate experiments in §4. In each setting, we generated 100 null images $X = (X_1, \ldots, X_n) \sim \mathcal{N}(\mathbf{0}, I)$ and estimated the covariance matrix as $\hat{\sigma}^2 I$ where $\hat{\sigma}^2$ is the sample variance of the same data. We ran 100 trials (i.e., 10,000 images in total) for three significance levels $\alpha = 0.05, 0.01, 0.10$.

Table 1. Architectures to compare

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

Non-Gaussian Noise. We set the image size to 256 and the architecture to base. As non-Gaussian distributions, we considered the following five distribution families:

- `skewnorm`: Skew normal distribution family.
- `exponnorm`: Exponentially modified normal distribution family.
- `gennormsteep`: Generalized normal distribution family (limit the shape parameter β to be steeper than the normal distribution, i.e., $\beta < 2$).
- `gennormflat`: Generalized normal distribution family (limit the shape parameter β to be flatter than the normal distribution, i.e., $\beta > 2$).
- `t`: Student's t distribution family.

We note that all of these distribution families include the Gaussian distribution and are standardized in the experiment. To conduct the experiment, we first obtained a distribution such that the 1-Wasserstein distance from $\mathcal{N}(0, 1)$ is d in each distribution family, for $d \in \{0.01, 0.02, 0.03, 0.04\}$. We then generated 100 images following each distribution and ran 100 trials (i.e., 10,000 images in total) for two significance levels $\alpha = 0.05, 0.01$.

We demonstrate the probability density functions for distributions from each distribution family such that the 1-Wasserstein distance from $\mathcal{N}(0, 1)$ is 0.04 in Figure 16

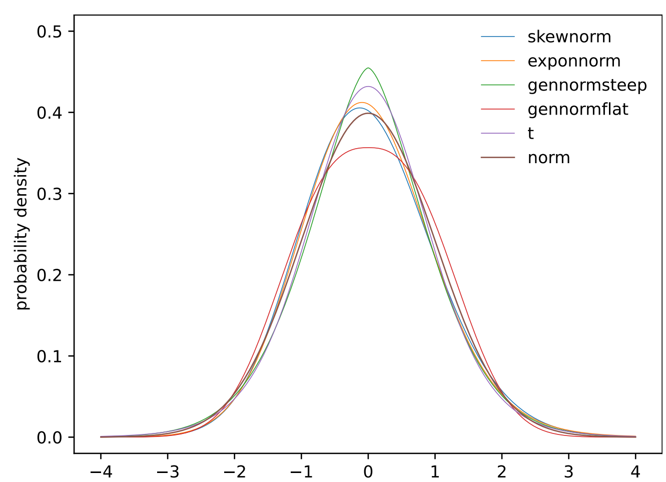


Figure 16. Demonstration of non-Gaussian distributions