Token-Based Detection of Spurious Correlations in Vision Transformers

Anonymous authors
Paper under double-blind review

Abstract

Due to their powerful feature association capabilities, neural network-based computer vision models have the ability to detect and exploit unintended patterns within the data, potentially leading to correct predictions based on incorrect or unintended but statistically relevant signals. These clues may vary from simple color aberrations to small texts within the image. In situations where these unintended signals align with the predictive task, models can mistakenly link these features with the task and rely on them for making predictions. This phenomenon is referred to as spurious correlations, where patterns appear to be associated with the task but are actually coincidental. As a result, detection and mitigation of spurious correlations have become crucial tasks for building trustworthy, reliable, and generalizable machine learning models. In this work, we present a novel token-based method to detect spurious correlations in vision transformers, a type of neural network architecture that gained significant popularity in recent years. Using both supervised and self-supervised trained models, we present large-scale experiments on the ImageNet dataset demonstrating the ability of the proposed method to identify spurious correlations. We also find that, even if the same architecture is used, the training methodology has a significant impact on the model's reliance on spurious correlations. Furthermore, we show that certain classes in the ImageNet dataset contain spurious signals that are easily detected by the models and discuss the underlying reasons for those spurious signals. In light of our findings, we provide an exhaustive list of the aforementioned images and call for caution in their use in future research efforts. Lastly, we present a case study investigating spurious signals in invasive breast mass classification, grounding our work in a real-world scenario.

1 Introduction

Although the computer vision (CV) community has benefited greatly from the improved performance of neural networks (NNs) on many complex vision tasks (He et al., 2016; Szegedy et al., 2016), recent studies have shown that unintended shortcut learning and spurious correlations are more common than previously thought (Geirhos et al., 2020; Oakden-Rayner et al., 2020). In the context of CV, shortcut learning can be described by what are called core and non-core features (Singla & Feizi, 2021). Here, core features are associated with the desired, semantically meaningful attributes of the image under consideration, whereas non-core features exhibit some degree of correlation with the desired task, while being peripheral or incidental to the main task (Kirichenko et al., 2022). As a result, the prevalence of non-core features may lead to models relying on superficial cues rather than truly understanding the underlying concepts. Analogous concerns have been raised in NLP with the advent of large language models (LLMs): transformer-based models like BERT and GPT-3 have demonstrated extraordinary capabilities, but they too can latch onto dataset artifacts or spurious cues, which can propagate flaws across many downstream applications. Naturally, this phenomenon poses significant challenges to the reliability and generalization capability of NNs across vision and language tasks, highlighting the pressing need for more robust and interpretable learning algorithms in the field (Geirhos et al., 2020).

One of the primary reasons for the existence of non-core features is the tremendous feature association capability of NNs, which can discover and learn signals that are not intended (Geirhos et al., 2020). The

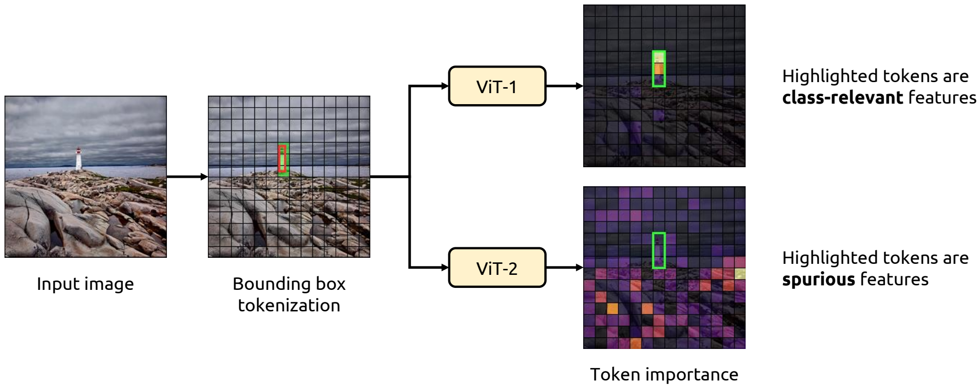


Figure 1: Visual illustration of the approach used in this work to detect spurious correlations. Given an input image, we identify important tokens using a token discarding method and analyze these tokens based on the location of the object of interest. In the first case (ViT-1), the important features lie within the object's bounding box, while in the second case (ViT-2), the important features highlight other elements, indicating the presence of spurious correlations.

introduction of transformer-based models such as vision transformer (ViT) further exacerbated the aforementioned issue due to the lack of inductive bias and the attention mechanism, which can discover and associate complex signals from even the most distant regions of an image (Vaswani et al., 2017).

Most of the work related to spurious correlations can be categorized into two main areas: (1) robust training methods designed to prevent spurious correlations during the training phase(Gong et al., 2019; Zhou et al., 2022), and (2) post-training detection of spurious correlations on a per-image basis (Lapuschkin et al., 2019; McCoy et al., 2019). The majority of research in the second category relies on interpretability methods, such as GradCAM (Selvaraju et al., 2017), as proxies for identifying spurious correlations. However, these interpretability methods are known to occasionally highlight misleading or irrelevant regions of the image, raising concerns about their reliability (Adebayo et al., 2018; Kindermans et al., 2019).

To address these issues, we propose a novel method for detecting the presence or absence of spurious correlations in a given image using a trained model. Our method relies on intrinsic properties of vision transformers and leverages the token discarding mechanism to identify spurious correlations through influential tokens that significantly impact the model's predictions. This token-based strategy relies on the fact that ViTs are capable of processing inputs with variable sequence lengths and builds on recent findings that many patch tokens in ViTs can be removed with minimal effect on model accuracy (Rao et al., 2021; Fayyaz et al., 2022; Bolya et al., 2022). Notably, eliminating such tokens not only speeds up inference but can also make model decisions more interpretable by revealing which regions the model deems important (Rao et al., 2021). Building on this insight, our approach repurposes token discarding as a diagnostic tool rather than an efficiency measure: by identifying which tokens a ViT model relies on most for a given image, we can tell whether the model's prediction is based on core features or on potentially spurious cues.

This strategy allows us to assess the trustworthiness of ViTs in terms of their ability to utilize core features and is widely applicable to any transformer-based CV model. Subsequently, we compare the positions of these influential tokens with the provided bounding box information to quantify the extent of spurious correlation using two novel metrics: the Average Token Spuriosity Index (A-TSI) and the Maximum Token Spuriosity Index (M-TSI). A visual overview of the proposed method is provided in Figure 1. Through large-scale experiments on both supervised and self-supervised ViTs on the widely-used ImageNet dataset (Russakovsky et al., 2015), we discover that model training significantly impacts the reliance of models on core and non-core features. Using the proposed approach, we also identify the problematic classes in the ImageNet dataset, i.e., those which contain a large number of images that are persistently classified based on non-core features by different models, thus highlighting issues with the ImageNet dataset.

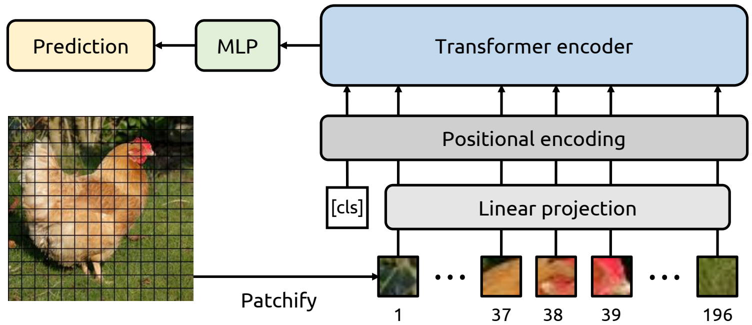


Figure 2: An overview of the ViT architecture and the tokenization of image patches.

2 Related work

Spurious correlations have been investigated in different contexts in the literature, such as invariant learning (Arjovsky et al., 2019), domain generalization (Wang et al., 2022), group robustness (Liu et al., 2021), shortcut learning (Geirhos et al., 2020; Du et al., 2023) and simplicity bias (Tiwari & Shenoy, 2023). Traditionally, most of the research focus has been on training methods which attempt to prevent spurious correlations from being learned, while our method focuses on the detection of spurious correlations after the training is complete, which is a comparatively unexplored area of research.

Training-based methods. Invariant learning methods such as Invariant Risk Minimization (Arjovsky et al., 2019) aim to learn features that remain predictive across multiple training environments. This is typically done by constraining the model to rely on consistent, environment-invariant signals rather than spurious correlations present in individual data distributions.

Group robust training focuses on improving model performance across different subgroups within a dataset, particularly those that are underrepresented or prone to high error rates. Optimization strategies include minimizing the loss of the worst-case subgroup loss (Sagawa et al., 2019) and iteratively identifying and correcting errors in hard-to-learn subgroups (Liu et al., 2021), reducing the reliance on shortcut features that perform well only for the majority group.

Domain Generalization techniques aim to improve model robustness by ensuring that learned features remain representative when shifting from the training dataset to unseen data (Wang et al., 2022). Some approaches achieve this by adversarial training (Ganin et al., 2016; Tzeng et al., 2017), encouraging the model to ignore domain-specific variations, while others promote feature alignment across different datasets (Jin et al., 2020; Lu et al., 2022).

Unlike the aforementioned training-time intervention methods, which actively modify the learning process of models to reduce reliance on spurious correlations, we tackle the problem of post-training identification of spurious correlations.

Attribution methods. Another line of research, similar to ours, identifies spurious correlations by investigating the attributions of models to determine whether they are focusing on semantically relevant regions (Ghosal & Li, 2024). Research efforts that follow this approach often analyze prediction changes when specific features are removed using polygon masks or occlusion (Plumb et al., 2021; Zeiler & Fergus, 2014). Region selection is typically guided by interpretability techniques such as GradCAM (Selvaraju et al., 2017), LIME (Ribeiro et al., 2016), or Integrated Gradients (Sundararajan et al., 2017). However, the reliability of these interpretability techniques is often under debate, since their effectiveness depends on assumptions about the underlying model behavior and feature importance (Adebayo et al., 2018; Kindermans et al., 2019). For instance, studies have shown that some attribution methods can produce visually plausible explanations even when applied to randomly initialized networks, raising concerns about their faithfulness to the decision-making process of models (Adebayo et al., 2018). While our approach shares similarities with these methods, it leverages the intrinsic properties of ViTs and employs a more principled strategy using token discarding.

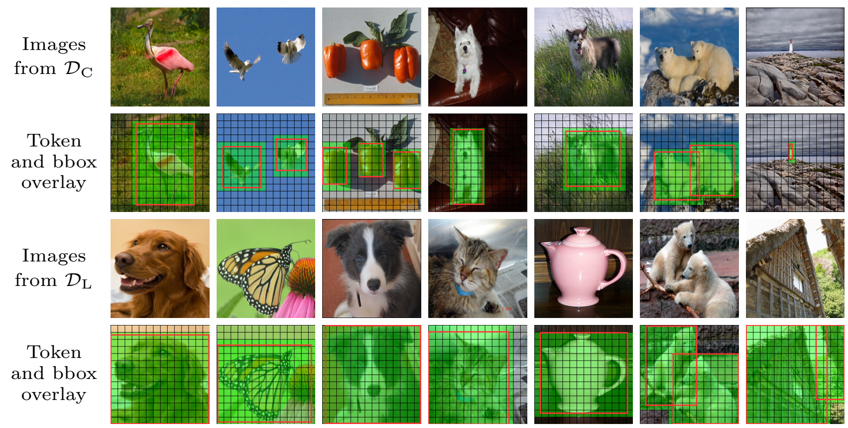


Figure 3: Example images from $\mathcal{D}_\mathrm{C}$ and $\mathcal{D}_\mathrm{L}$ are provided, along with their bounding boxes (highlighted in red) and corresponding tokens based on 16×16 patches (highlighted in green). As can be seen, the bounding boxes (bbox) for images from $\mathcal{D}_\mathrm{L}$ cover almost the entirety of the image.

Datasets. The majority of research investigating spurious correlations after training focuses on datasets with human-identifiable features, such as the CelebA dataset (Liu et al., 2015), where celebrities are categorized by characteristics like hair color and gender. As a result, many research efforts examine spurious correlations as a group-level phenomenon, using methods such as worst-case generalization of group performance (Sagawa et al., 2019; Ghosal & Li, 2024). Unfortunately, this approach makes the proposed methods largely inapplicable to datasets that lack features that are easily identifiable by humans.

3 Methodology

3.1 Models

We employ the most commonly used Vision Transformer architecture: ViT-B/16 (Dosovitskiy et al., 2020). The ViT-B/16 architecture tokenizes the input image into patches of size 16×16 (see Figure 2), resulting in 196 tokens for a standard ImageNet image of size 224×224.

In addition to the model trained in a supervised fashion, we also use two additional models based on the same architecture but trained in a self-supervised manner: Self-Distillation with No labels (DINO) (Caron et al., 2021) and Masked AutoEncoders (MAE) (He et al., 2022). DINO employs a novel self-supervised learning approach that relies on contrastive learning whereas MAE further advances this concept by learning to reconstruct missing parts of input images, effectively capturing intricate details and dependencies within the data. Further details on the employed self-supervised training methods can be found in their respective papers and in the surveys of (Khan et al., 2022; Ozbulak et al., 2023).

3.2 Dataset

Following past research efforts and to maximize the applicability of our findings, we utilize the ImageNet validation set for our experiments. This dataset contains approximately 50,000 images spread over 1,000 classes, all of which come with bounding box information highlighting the object of interest in the image. Instead of performing experiments at an aggregate level on all images in the ImageNet validation set, we separate them into three groups based on the properties detailed in the following.

- **Bounding box for the object of interest ($\mathcal{D}_\mathrm{L}$).** Recall that the images in the ImageNet dataset are tokenized into 196 tokens by ViTs. Surprisingly, we discovered that a large subset of images in this dataset have almost the entirety of the image highlighted with bounding boxes (over 160 tokens in bounding boxes). Some examples of images that exhibit this phenomenon are shown in Figure 3. For a faithful analysis, we filter out those $11,221$ images and do not use them in our experiments.

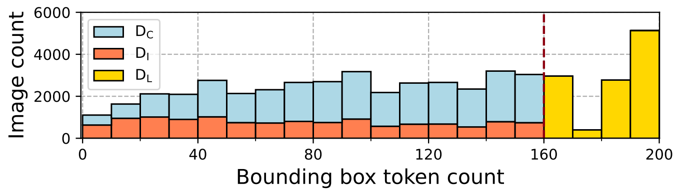


Figure 4: Stacked histograms showing the distribution of images in the ImageNet validation dataset based on the coverage of bounding boxes, illustrated in relation to the number of tokens.

This criterion ensures a targeted evaluation of the degree of spurious correlation by placing emphasis on images where the object of interest does not constitute a significant portion of the image.

- **Classification accuracy**. Inspired by the work of Ozbulak et al. (2021a), we divide the remaining images into two categories: those that are classified correctly by all selected models and those that are incorrectly classified by at least one model. This division ensures that our experiments capture differences between images that are easy to classify (i.e., those that are correctly classified by all models) and those that are comparatively harder to classify.
  - **Images that are correctly classified by all models ($\mathcal{D}_\mathbf{C}$)**. This filtering operation ensures that the selected set of images has a consensus among all models in their classification, allowing for a more controlled evaluation across models. Based on this criterion, we find that $26,317$ images are correctly classified by all three selected models.
  - **Images that are incorrectly classified by at least one model ($\mathcal{D}_\mathbf{I}$)**. To explore potential relationships between spurious correlations and classification accuracy, we conduct a separate analysis for images that are incorrectly classified by at least one model. By isolating these images, our idea is to gain insight into the specific challenges that arise in the classification process and spurious correlations. This category comprises $12,458$ images.

Based on the grouping detailed above, we can represent the ImageNet validation dataset as a combination of three disjoint subsets $\mathcal{D}_{\text{ImageNet}} = \mathcal{D}_L \cup \mathcal{D}_C \cup \mathcal{D}_I$ (see Figure 4). For experiments, we will use $\mathcal{D}_C$ and $\mathcal{D}_I$.

3.3 Identifying spurious correlations via token discarding

We propose a two-step procedure to identify spurious correlations in ViTs. Given an image and its corresponding bounding box we:

- **Step 1**: Discover influential tokens that contribute to the prediction made by the model.
- **Step 2**: Identify spurious correlations based on influential tokens and bounding box information.

3.3.1 Discovering influential tokens

Token discarding emerged as a method that is unique to transformer-based architectures including ViTs (Rao et al., 2021). This method can be utilized for a variety of purposes including creating robust models (Renggli et al., 2022), speeding up the training process (Bolya et al., 2022; Rao et al., 2021; Long et al., 2023), and interpretability (Haurum et al., 2023; Pan et al., 2021). In this work, we utilize token discarding in order to discover influential tokens that contribute to the prediction.

Given a correctly classified image $\mathtt{X}$, we denote the corresponding prediction confidence for the correct class of that image as $\hat{y} = g_\theta(\mathtt{X})$ obtained from a ViT with parameters θ. As described above, ViT-B/16 tokenizes an input image $\mathtt{X}$ of size 224×224 into 196 tokens which we will denote as $\mathtt{X} = [\boldsymbol{x}_i]_{i \in \{1,\dots,196\}}$. To identify the influential tokens, we systematically remove one token at a time from the image and observe how each removal affects the model's prediction compared to its initial prediction. As a result, we create a dictionary of confidence changes $\mathtt{Z} = \{z_k\}_{k \in \{1,\dots,196\}}$ where $z_k = |\hat{y} - \hat{y}^{(-k)}|$, with $\hat{y}^{(-k)} = g_\theta(\mathtt{X}^{(-k)})$ being the prediction

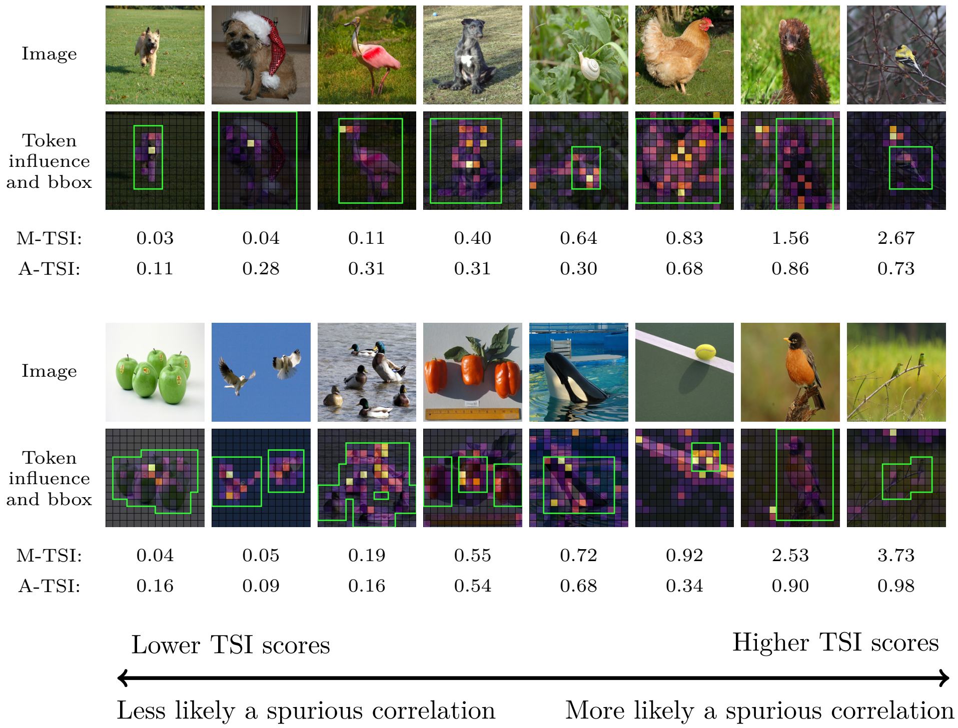


Figure 5: An example set of images is presented with their corresponding token influence maps generated using ViTs, as well as M-TSI and A-TSI scores. Green boundaries represent the bounding box information, highlighting the object of interest in the image. The images are sorted according to an increasing value of M-TSI to provide a clear qualitative view of spurious correlations identified by the proposed method.

confidence of the image with one fewer token (the kth one), described as $\mathtt{X}^{(-k)} = [\boldsymbol{x}_i]_{i\in\{1,\ldots,196\}\setminus\{k\}}$. Based on this dictionary, we can create a heatmap of token influence as shown in Figure 5. This approach can be viewed as the CV counterpart of the method proposed in Madsen et al. (2023), which assesses the importance of tokens in natural language processing tasks. The proposed method is also highly related to the topic of patch-based masking and missingness bias, which we discuss in detail in the Appendix.

3.3.2 Identifying spurious correlations

Given an image with bounding box information, we first identify the tokens that lie within the bounding box as shown in Figure 3. Based on this information, we create two sets of tokens IDs, one for those that lie in the bounding box ($\mathtt{B}_{\text{in}}$) and one for those that are outside it ($\mathtt{B}_{\text{out}}$). Afterwards, using the previously created token influence maps, we are able to quantify whether the identified influential tokens lie within the bounding box or not. We propose the following two metrics to identify spurious correlations:

- **Average Token Spuriosity Index (A-TSI)**. The first metric we propose is based on the average token influence and is described as follows:

$$\text{A-TSI}(\mathtt{Z}, \mathtt{B}_{\text{in}}, \mathtt{B}_{\text{out}}) = \frac{\frac{1}{n_{\text{out}}} \sum_{i\in\mathtt{B}_{\text{out}}} z_i}{\frac{1}{n_{\text{in}}} \sum_{i\in\mathtt{B}_{\text{in}}} z_i} \,. \tag{1}$$

- **Maximum Token Spuriosity Index (M-TSI)**. The second metric we propose is based on the maximum token influence and is described as follows:

$$\text{M-TSI}(\mathtt{Z}, \mathtt{B}_{\text{in}}, \mathtt{B}_{\text{out}}) = \frac{\max(\{z_i\}_{i\in\mathtt{B}_{\text{out}}})}{\max(\{z_i\}_{i\in\mathtt{B}_{\text{in}}})} \,. \tag{2}$$

Table 1: Using ViTs trained with supervised learning, DINO, and MAE, mean (standard deviation) TSI is calculated for $\mathcal{D}_C$ and $\mathcal{D}_I$ for all images within subsets as well as certain groups of images based on the coverage of the bounding box in terms of tokens.

$\mathcal{D}_C$	Tokens	Supervised	DINO	MAE
M-TSI	All	0.64 (0.76)	0.35 (0.64)	0.68 (0.69)
	1-40	0.88 (1.22)	0.64 (1.44)	0.86 (1.19)
	41-80	0.68 (0.73)	0.37 (0.43)	0.73 (0.63)
	81-120	0.62 (0.67)	0.30 (0.33)	0.67 (0.55)
	121-160	0.51 (0.53)	0.24 (0.26)	0.57 (0.47)
A-TSI	All	0.54 (0.30)	0.33 (0.19)	0.61 (0.27)
	1-40	0.37 (0.28)	0.26 (0.23)	0.41 (0.25)
	41-80	0.47 (0.24)	0.30 (0.16)	0.56 (0.23)
	81-120	0.57 (0.28)	0.34 (0.17)	0.65 (0.24)
	121-160	0.64 (0.32)	0.37 (0.18)	0.71 (0.27)

$\mathcal{D}_I$	Tokens	Supervised	DINO	MAE
M-TSI	All	0.94 (1.64)	0.83 (1.30)	0.84 (1.38)
	1-40	1.63 (2.47)	1.47 (1.82)	1.38 (1.96)
	41-80	0.85 (1.13)	0.76 (1.14)	0.81 (1.14)
	81-120	0.57 (0.70)	0.45 (0.52)	0.53 (0.56)
	121-160	0.39 (0.42)	0.32 (0.39)	0.40 (0.76)
A-TSI	All	0.51 (0.34)	0.44 (0.30)	0.49 (0.29)
	1-40	0.51 (0.40)	0.47 (0.34)	0.46 (0.33)
	41-80	0.50 (0.32)	0.45 (0.32)	0.50 (0.30)
	81-120	0.50 (0.31)	0.42 (0.28)	0.51 (0.26)
	121-160	0.52 (0.29)	0.41 (0.24)	0.52 (0.26)

While both A-TSI and M-TSI are useful in identifying spurious correlations, A-TSI measures average spurious correlation influence relative to core features, while M-TSI targets scenarios where a few influential tokens outside the bounding box signal significant correlations.

Interpreting TSI scores. Both A-TSI and M-TSI are straightforward to understand in terms of what their scores indicate.

- TSI $\in (0, 1)$ implies that the tokens within the bounding box are more influential for the prediction compared to those that are outside, meaning that the prediction is based on features that originate from regions of the input image related to the class information.
- TSI $= 1$ indicates that features outside the bounding box are as important as those that are inside for prediction.
- TSI > 1 the influence of tokens outside is greater than those inside, indicating the existence of spurious correlations. As TSI increases beyond 1, the intensity of spurious correlations increases.

In order to provide a straightforward understanding of the proposed method, we present Figure 5, which displays input images and their corresponding token influence maps overlaid with bounding box information. For each image, we calculate M-TSI and A-TSI, sorting the images from left to right according to increasing M-TSI scores. As measured by TSI, images on the left side display less spurious correlations since most of the important tokens are inside the bounding box, whereas images on the right side demonstrate higher levels of spurious correlations since the important tokens are outside the bounding box, indicating that the model is relying on irrelevant features or background information for its predictions.

4 Experimental results and insights on spurious correlations

Using all images in $\mathcal{D}_C$ and $\mathcal{D}_I$ and using three ViTs, we calculate A-TSI and M-TSI with token influence maps and present the mean and standard deviation in Table 1. Apart from providing these parameters for the group of all images, we also filter images according to the size of the bounding box in terms of token coverage and calculate the aforementioned metrics over four groups: 1-40, 41-80, 81-120, and 121-160 tokens. We also provide extensive histograms of TSI in the Appendix (Figure 11) for $\mathcal{D}_C$ and $\mathcal{D}_I$. Based on these results, we can make the following observations.

Correctly classified images contain fewer spurious features. A-TSI and M-TSI values for correctly classified images are generally lower across all token coverage groups, suggesting that models focus more on relevant features when making accurate predictions. Conversely, misclassified images tend to have higher TSI values, indicating the presence of more spurious or irrelevant features that may mislead the model. Notably, the M-TSI in the $[2, 2+]$ bin in Figure 11 highlights that a large number of images are incorrectly classified by at least one model, likely due to strong spurious correlations as identified by M-TSI. In contrast, this bin

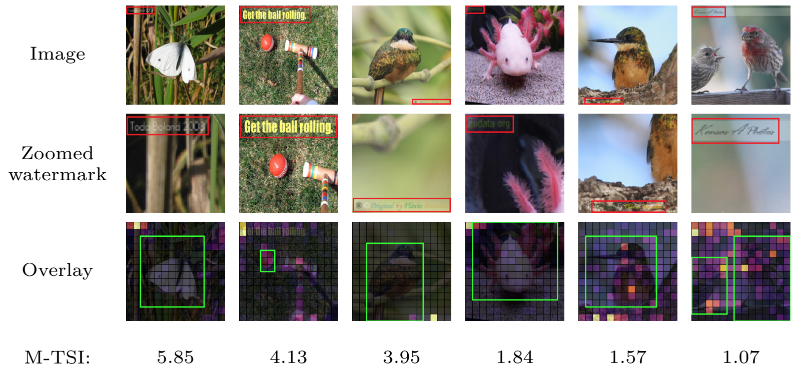


Figure 6: Example images obtained from the ImageNet validation dataset where tokens highlight watermarks over the image instead of the objects of interest, leading to high TSI scores.

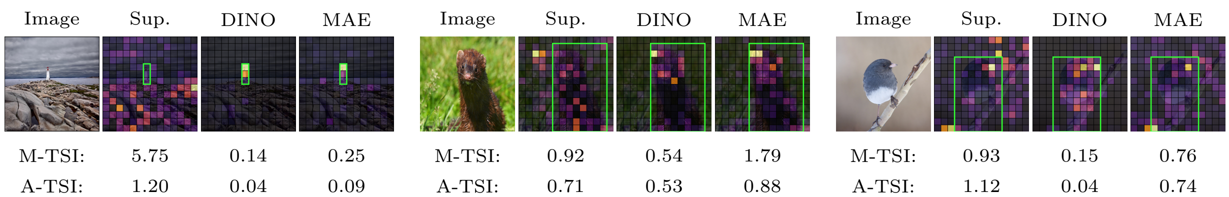


Figure 7: Example images in the ImageNet dataset, where the important tokens are correctly identified by DINO and MAE, resulting in low TSI scores, whereas the supervised model fails to do so, leading to high TSI scores.

contains proportionally fewer images in $\mathcal{D}_C$, despite the fact that $\mathcal{D}_C$ has a larger overall number of images compared to $\mathcal{D}_I$.

The training method influences spurious correlations. Although the three selected models share the same architecture, we observe differing results when analyzing TSI. Specifically, for images in $\mathcal{D}_C$, DINO exhibits fewer spurious correlations compared to the other two models, indicating that the training method influences the model's reliance on spurious correlations. Notably, while DINO has the lowest TSI among the three, MAE shows the highest, suggesting that self-supervised training does not necessarily reduce spurious correlations. In contrast, for images in $\mathcal{D}_I$, the TSI of the different models are less different, with DINO again having the lowest median TSI by a small margin. Consequently, DINO appears to be the more robust ViT. Supporting the claims made by Caron et al. (2021), we believe the robustness of DINO stems from its distillation-based training routine which mostly transfers robust and useful features between teacher and student networks.

M-TSI captures spurious correlations more effectively for small objects.. While A-TSI remains consistent across images with varying bounding box sizes, M-TSI differs significantly, particularly for images containing relatively small objects of interest. Specifically, images with bounding boxes covering up to 40 tokens show higher M-TSI values compared to other groups and their corresponding A-TSI. This suggests that when the object of interest is relatively small, the M-TSI score is more effective in detecting spurious correlations than the A-TSI score.

M-TSI identifies strong spurious signals such as watermarks. In Figure 6, we provide another set of qualitative examples where the tokens covering the watermark are identified as influential tokens, leading to high TSI scores. This discovery reveals yet another use-case for the proposed method where spurious correlations based on watermarks can be identified using M-TSI. This capability demonstrates the versatility and robustness of M-TSI in uncovering subtle, yet impactful, spurious correlations that could compromise the model's performance in real-world scenarios.

Table 2: For images in subset $\mathcal{D}_I$ (i.e., images misclassified into incorrect categories), M-TSI and A-TSI are computed and grouped into one of four bins based on prediction confidence. For each bin, the mean (std) TSI values are calculated.

$\mathcal{D}_I$	Model	0-25%	25-50%	50-75%	75-100%
M-TSI	Supervised	1.13 (1.44)	0.83 (1.22)	0.77 (1.27)	0.87 (1.91)
	DINO	0.98 (1.19)	0.72 (0.96)	0.64 (1.00)	0.79 (1.51)
	MAE	1.07 (1.67)	0.76 (1.00)	0.73 (1.29)	0.81 (1.39)
A-TSI	Supervised	0.54 (0.37)	0.48 (0.32)	0.47 (0.32)	0.50 (0.32)
	DINO	0.47 (0.32)	0.43 (0.29)	0.41 (0.28)	0.42 (0.28)
	MAE	0.50 (0.36)	0.47 (0.30)	0.46 (0.26)	0.53 (0.28)

Table 3: Classes with the highest average M-TSI scores in the ImageNet validation dataset calculated for $\mathcal{D}_C$ (i.e., images that are correctly classified) and $\mathcal{D}_I$ (i.e., images that are misclassified). Repeating classes are highlighted in bold.

Image subset	Supervised		DINO		MAE	
	Class	M-TSI	Class	M-TSI	Class	M-TSI
$\mathcal{D}_C$	**space bar**	4.99	**ping-pong ball**	4.83	**ping-pong ball**	3.18
	puck	3.84	**space bar**	3.69	**puck**	3.09
	ping-pong ball	3.35	**puck**	3.26	**space bar**	2.56
	geyser	1.90	**alp**	2.03	**rapeseed**	1.68
	rugby ball	1.87	balance beam	1.55	**basketball**	1.68
	laptop	1.78	**volleyball**	1.55	**volleyball**	1.68
	basketball	1.76	diaper	1.47	miniskirt	1.52
	shoji	1.76	**rapeseed**	1.45	apiary	1.51
	rapeseed	1.75	pickelhaube	1.37	**alp**	1.48
	lakeside	1.69	**rugby ball**	1.30	**geyser**	1.48
$\mathcal{D}_I$	shoe shop	7.06	**cockatoo**	7.17	**cockatoo**	7.73
	worm fence	7.00	**soccer ball**	5.35	parking meter	5.07
	airship	6.77	conch	4.75	radiator	4.46
	Granny Smith	5.80	traffic light	4.03	matchstick	3.97
	flagpole	5.33	baseball	3.89	cockroach	3.88
	bow tie	4.90	**worm fence**	3.65	brass	3.67
	soccer ball	4.85	pickelhaube	3.30	jean	3.26
	platypus	3.84	tennis ball	3.23	**platypus**	3.21
	slug	3.60	crash helmet	3.03	cello	3.06
	zebra	3.56	accordion	2.99	grille	3.06

Training routine affects TSI on identical images. We present Figure 7, which shows the TSI calculated using three ViT models for identical images. In the provided examples, the supervised model considers spurious features that surround the object for the prediction, while DINO and MAE successfully make use of the features originating from the class-related regions of the image. This shows that the features taken into account during prediction may vary by the type of learning, resulting in different degrees of spuriosity for different models on the same dataset. We further explore model differences in the next section.

Misclassifications with low confidence have higher TSI scores. Investigating TSI for misclassified images, we discover that misclassifications with low confidence (i.e., confidence between 0% and 25%) have substantially higher M-TSI and A-TSI scores, indicating that when the model is uncertain about its prediction, it is more likely to rely on non-core features rather than the primary object of interest (see Table 2). This trend is particularly pronounced for M-TSI, indicating that, in such cases, the most influential tokens for the prediction often reside outside the bounding box of the object, highlighting the model's tendency to use background artifacts or incidental cues when making uncertain decisions.

Class-based investigation with TSI scores. In order to investigate whether some classes are more prone to spurious correlations than others, we calculate the TSI for images within each class separately. In Table 3, we provide the 10 classes with the highest average M-TSI scores for each model along with their respective values for images in $\mathcal{D}_C$ and in $\mathcal{D}_I$. Notably, a large portion of the classes repeatedly appears across all three models, suggesting the presence of potentially systematic problems in the ImageNet dataset specific to these classes. This repeated appearance raises concerns about biases or artifacts that may be embedded within the dataset, leading models to rely on irrelevant features when making predictions. We also notice that classes

(a) Label inconsistency (b) Secondary objects (c) Small area of interest (d) Class similarities

Figure 8: Several examples from the ImageNet dataset and their annotations, highlighting (a) label inconsistencies, (b) the presence of secondary objects, (c) small areas of interest, and (d) class similarities, all of which contribute to higher TSI scores. In (a), two images originate from the same class (rapeseed), but in the first image, the annotation covers only a few individual plants, whereas in the second image, it encompasses the entire field. In (b), the presence of humans, along with other strong correlating signals such as uniforms or helmets, influences the focus of models. In (c), the objects of interest are extremely small, making it difficult for models to detect them. In (d), four images come from two distinct classes: typewriter keyboard and spacebar, despite the images and their backgrounds being nearly identical.

in $\mathcal{D}_I$ have substantially higher M-TSI scores compared to classes in $\mathcal{D}_C$, meaning that misclassifications often rely on unintended cues, rather than the features of the object of interest.

To uncover the underlying reasons for the presence of spurious correlations, we investigate images with high TSI scores, particularly M-TSI scores. After analyzing several hundred images, along with their token influence maps and TSI scores, we identify the following key factors contributing to spurious correlations:

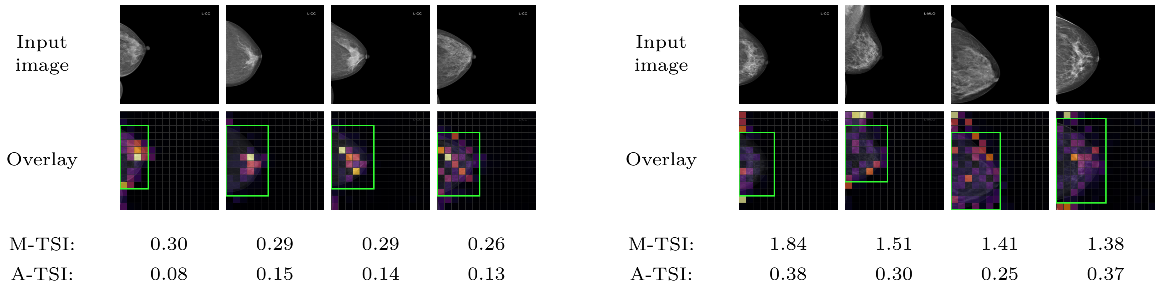
- Label inconsistency: In these cases, the bounding box either fails to highlight the correct object of interest, partially covers it, or includes unrelated parts of the image. Such inconsistencies can mislead models into associating predictions with irrelevant features.

- Secondary objects: Some images contain secondary objects that strongly correlate with the primary object of interest, leading models to focus on these unintended features. For example, several ImageNet classes related to sports frequently include humans in the images. In such cases, we find that models tend to focus on the person rather than the actual object of interest.

- Small area of interest: Some target objects in the dataset are extremely small, often occupying only a single token in the model's representation. An example of this are ping pong, tennis, or rugby balls, which appear in images but are too small for the model to reliably detect. Instead of focusing on the actual object, the model often relies on contextual cues from the rest of the image to infer the class, leading to high TSI.

- Class similarities: Certain classes exhibit significant visual similarities, making it difficult for the model to distinguish between them. For instance, the classes typewriter, computer keyboard, and typewriter keyboard are different classes in ImageNet but contain highly similar objects and backgrounds. This overlap can cause models to rely on features that are not unique to a single class, reducing their ability to generalize correctly.

For each of the aforementioned cases, we provide several examples and their descriptions in Figure 8. Note that our work is not the only research effort that identified these problems in the ImageNet dataset (Ozbulak et al., 2021b; Peychev et al., 2023; Beyer et al., 2020; Moayeri et al., 2022). However, our method can consistently discover such cases based on TSI scores.



(a) Examples without spurious correlations

(b) Examples with spurious correlations

Figure 9: (top) Images from the VinDR-Mammo dataset and (bottom) corresponding token influence maps obtained from fine-tuned models. The left side presents examples with lower M-TSI and A-TSI scores, indicating the absence of spurious correlations. The right side shows examples with higher M-TSI and A-TSI scores, where spurious correlations are evident in regions outside the breast area, such as chest fat tissue.

5 Case study on breast cancer detection

Based on our previous findings, we tackle a real-world scenario involving spurious correlations: invasive breast mass classification on MRI images. For this case study, we use the VinDr-Mammo dataset, which is designed to advance computer-aided diagnosis in full-field digital mammography (Nguyen et al., 2023). We fine-tune three pretrained ViTs on this dataset, making sure that the models achieve a performance that is close to the state-of-the-art on this dataset. Based on the trained models, we calculate TSI scores based on the breast area and investigate images with their corresponding TSI scores and present several cases in Figure 9.

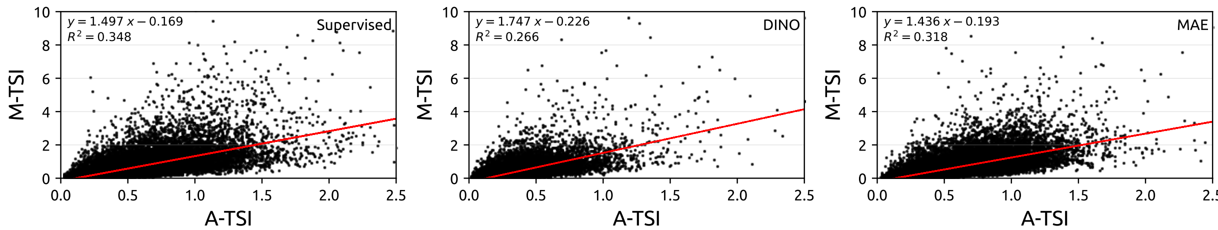
As illustrated by the images on the left side of Figure 9, where M-TSI and A-TSI scores are lower, the models concentrate on the pertinent breast tissue areas. In contrast, for images on the right side, which have higher M-TSI and A-TSI scores, the models assign importance to non-relevant regions outside the breast, such as chest fat tissue. These examples demonstrate the potential of our proposed method in effectively identifying spurious correlations in real-world scenarios, ensuring that model predictions are based on correct features (i.e., clinically meaningful features) rather than on irrelevant artifacts.

6 Selecting A-TSI vs M-TSI

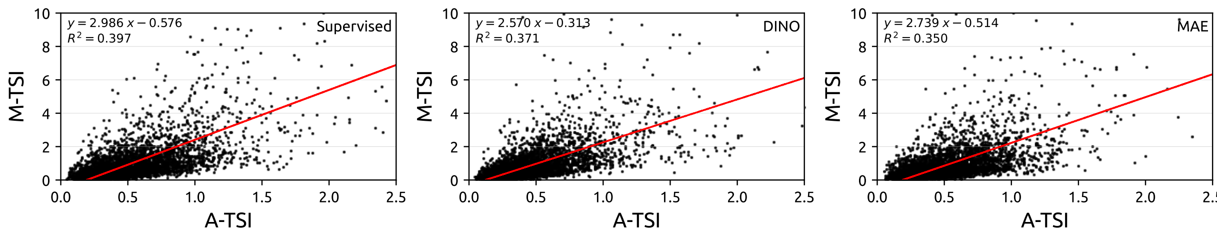
Throughout this paper, we presented results and examples containing both M-TSI and A-TSI. However, a fundamental question arises: given a model trained on a dataset, which one of the two is more appropriate to identify spurious correlations?

Based on previous observations, we investigate whether a large A-TSI is indicative of a large M-TSI, and vice versa. To explore this relationship, we provide Figure 10, which illustrates the correlation between A-TSI and M-TSI across all three models on ImageNet, using both D_C and D_I separately. The corresponding R^2 values fall within the range of low to moderate positive correlations, which means that although there is some relationship between A-TSI and M-TSI, it is not particularly strong.

Given the differences between A-TSI and M-TSI, it is important to determine which measure is more useful for identifying spurious correlations. Across all experiments that use token influence maps, we found that M-TSI provides a more consistent and reliable signal for detecting spurious correlations, particularly in cases where the relevant areas of the object of interest cover large portions of the image. In such cases, A-TSI tends to smooth out due to the large number of tokens covering the area, whereas M-TSI, which is calculated using only the maximum token influence, more reliably identifies spurious correlations. Therefore, based on our findings, M-TSI emerges as the more appropriate metric for evaluating the extent to which a model relies on spurious features under ideal conditions, making it a preferable choice for assessing model trustworthiness in real-world applications.



(a) $\mathcal{D}_C$



(b) $\mathcal{D}_I$

Figure 10: Scatterplots depicting the relationship between M-TSI and A-TSI for all images in (top) D_C and (bottom) D_I. For each graph, the R^2 value is provided in the top left, and the model information is displayed in the top right.

7 Conclusions and future perspectives

In this work, we proposed an easily usable technique as well as two new metrics to identify spurious correlations in ViTs. Through large-scale experiments, we demonstrated that the chosen training routine may have a drastic impact on the model's reliance on spurious correlations. Furthermore, we identified a number of classes in ImageNet that exhibit very strong spurious signals that may confuse models, and we extensively discussed the underlying reasons for their presence.

Two major limitations of the proposed method are (1) its reliance on annotations and (2) the computational demand required to obtain the most reliable results with token discarding. While our approach can be adapted to work with other forms of annotation that highlight the object of interest, such as segmentation masks, individual tokens, or approximate regions, it still requires manual intervention. Moreover, the computational cost of generating token influence maps highlights the need for a more efficient alternative in resource-constrained settings. We conduct ablation studies using GradCAM as well as attention maps and present preliminary results for strong directions of future research in the Appendix for interested readers.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Sriram Balasubramanian and Soheil Feizi. Towards improved input masking for convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1855–1865, 2023.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

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.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 67(1):110–120, 2023.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *European conference on computer vision*, pp. 396–414. Springer, 2022.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Soumya Suvra Ghosal and Yixuan Li. Are vision transformers robust to spurious correlations? *International Journal of Computer Vision*, 132(3):689–709, 2024.

Zhiqiang Gong, Ping Zhong, and Weidong Hu. Diversity in machine learning. *Ieee Access*, 7:64323–64350, 2019.

Joakim Bruslund Haurum, Sergio Escalera, Graham W Taylor, and Thomas B Moeslund. Which tokens to use? investigating token reduction in vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 773–783, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Saachi Jain, Hadi Salman, Eric Wong, Pengchuan Zhang, Vibhav Vineet, Sai Vemprala, and Aleksander Madry. Missingness bias in model debugging. *arXiv preprint arXiv:2204.08945*, 2022.

Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Feature alignment and restoration for domain generalization and adaptation. *arXiv preprint arXiv:2006.12009*, 2020.

Adnan Khan, Sarah AlBarri, and Muhammad Arslan Manzoor. Contrastive self-supervised learning: a survey on different architectures. In *2022 2nd International Conference on Artificial Intelligence (ICAI)*, pp. 1–6. IEEE, 2022.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pp. 267–280, 2019.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Sifan Long, Zhen Zhao, Jimin Pi, Shengsheng Wang, and Jingdong Wang. Beyond attentive tokens: Incorporating token importance and diversity for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10334–10343, 2023.

Wang Lu, Jindong Wang, Haoliang Li, Yiqiang Chen, and Xing Xie. Domain-invariant feature exploration for domain generalization. *arXiv preprint arXiv:2207.12020*, 2022.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Faithfulness measurable masked language models. *arXiv preprint arXiv:2310.07819*, 2023.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. *Advances in Neural Information Processing Systems*, 35:10068–10077, 2022.

Hieu T Nguyen, Ha Q Nguyen, Hieu H Pham, Khanh Lam, Linh T Le, Minh Dao, and Van Vu. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277, 2023.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

Utku Ozbulak, Esla Timothy Anzaku, Wesley De Neve, and Arnout Van Messem. Selection of source images heavily influences the effectiveness of adversarial attacks. *arXiv preprint arXiv:2106.07141*, 2021a.

Utku Ozbulak, Maura Pintor, Arnout Van Messem, and Wesley De Neve. Evaluating adversarial attacks on imagenet: A reality check on misclassification classes. *arXiv preprint arXiv:2111.11056*, 2021b.

Utku Ozbulak, Hyun Jung Lee, Beril Boga, Esla Timothy Anzaku, Ho-min Park, Arnout Van Messem, Wesley De Neve, and Joris Vankerschaver. Know your self-supervised learning: A survey on image-based generative and discriminative training. *Transactions on Machine Learning Research*, 2023.

Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogerio Feris, and Aude Oliva. $ia-red^2$: Interpretability-aware redundancy reduction for vision transformers. *Advances in Neural Information Processing Systems*, 34:24898–24911, 2021.

Momchil Peychev, Mark Müller, Marc Fischer, and Martin Vechev. Automated classification of model errors on imagenet. *Advances in Neural Information Processing Systems*, 36:36826–36885, 2023.

Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *arXiv preprint arXiv:2106.02112*, 2021.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Cedric Renggli, André Susano Pinto, Neil Houlsby, Basil Mustafa, Joan Puigcerver, and Carlos Riquelme. Learning to merge tokens in vision transformers. *arXiv preprint arXiv:2202.12015*, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? *arXiv preprint arXiv:2110.04301*, 2021.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Rishabh Tiwari and Pradeep Shenoy. Overcoming simplicity bias in deep networks using a feature sieve. In *International Conference on Machine Learning*, pp. 34330–34343. PMLR, 2023.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

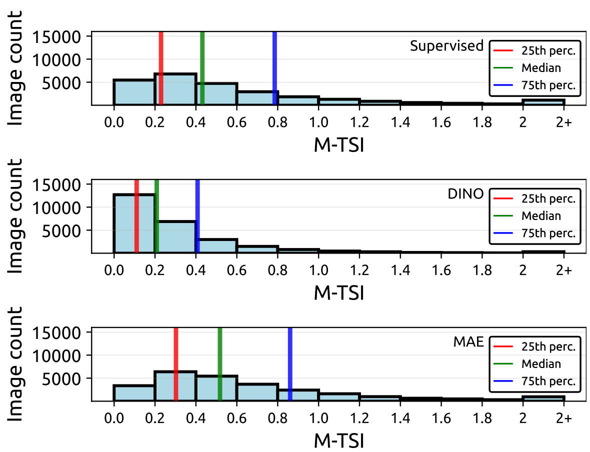
Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

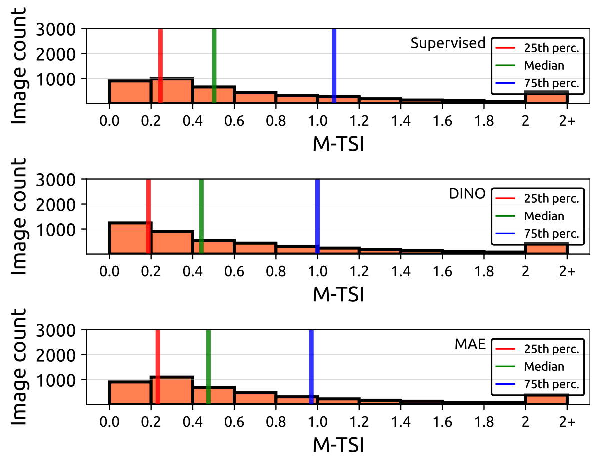
Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

Appendix

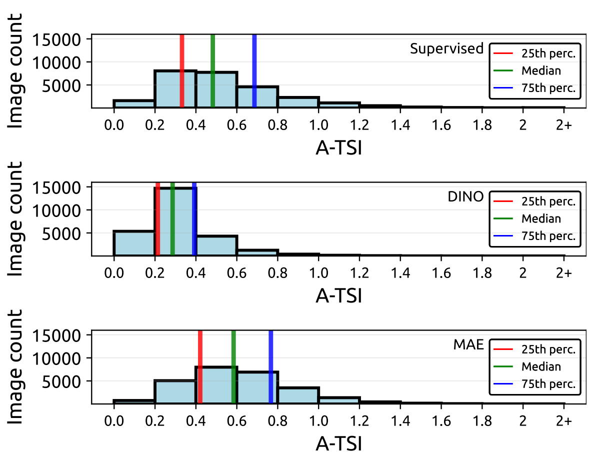
A Distributions of A-TSI and M-TSI on ImageNet



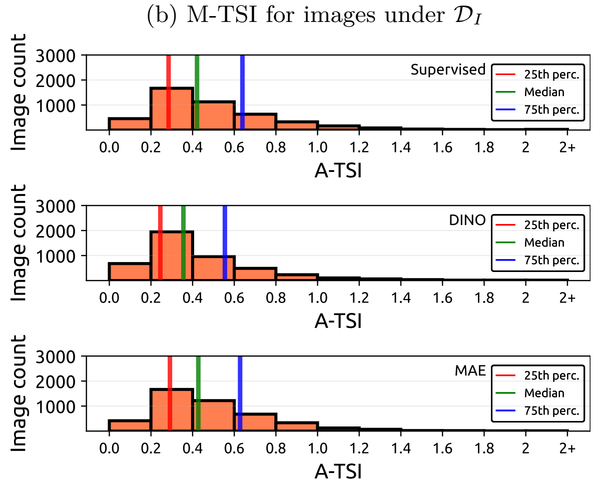
(a) M-TSI for images under $\mathcal{D}_C$



(b) M-TSI for images under $\mathcal{D}_I$



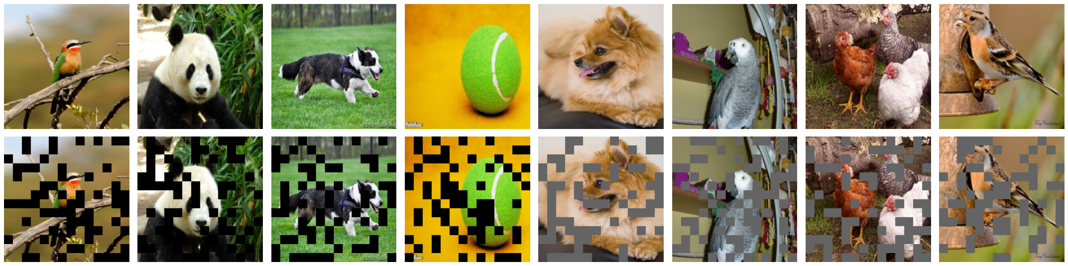
(c) A-TSI for images under $\mathcal{D}_C$



(d) A-TSI for images under $\mathcal{D}_I$

Figure 11: Histograms of M-TSI and A-TSI for ViTs trained using supervised learning, DINO, and MAE on $\mathcal{D}_C$ and $\mathcal{D}_I$. Images with a TSI value greater than 2 are grouped into the $[2, 2+]$ bin for clarity.

B On image masking, token discarding, and missingness bias



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 12: (Top) Input images and (bottom) their masked versions. While both the original and masked images are correctly classified by ViTs when masked tokens are discarded, the masked images are misclassified by ResNet-50, a prominent CNN architecture into plausible but incorrect categories due to missingness bias. The initial and masked predictions for the images are: (a) bee-eater → boa, (b) panda → soccer ball, (c) corgi → doormat, (d) tennis ball → mousetrap, (e) Pomeranian → African gray, (f) African gray → quill, (g) hen → spotlight, and (h) brambling → crossword puzzle.

Missingness bias in convolutional neural networks (CNNs) arises when features are removed or masked in a way that introduces unintended distortions in model predictions (Balasubramanian & Feizi, 2023; Jain et al., 2022). Since CNNs rely on convolutions that operate over a spatially contiguous image, they cannot naturally ignore missing regions (LeCun et al., 1998). Instead, missing pixels must be replaced with approximations, such as blacking them out, adding noise, or using a blurred region. These approximations inadvertently introduce biases because the model learns to associate the masking pattern itself with certain predictions, rather than relying on the remaining unmasked features. This effect not only affects model predictions but also interpretability techniques such as LIME, where missingness is used to infer feature importance. In contrast, ViTs offer a more natural implementation of missingness through token discarding. Since ViTs process images as a set of non-overlapping tokens rather than a spatially continuous grid, individual tokens corresponding to specific image regions can be entirely without causing missingness bias. As a result, ViTs mitigate missingness bias and enable more reliable model debugging, making them particularly advantageous for feature attribution and interpretability studies.

In Figure 12, we provide several example cases where images have their content partially masked, demonstrating the impact of missingness bias on CNNs. While the original images are correctly classified by ResNet-50, their masked counterparts are misclassified into plausible but incorrect categories. This occurs because the missing regions alter key visual features, leading the model to rely on incomplete or misleading cues. The figure illustrates how different objects, such as animals and everyday items, shift in classification due to the absence of crucial details. These examples highlight the susceptibility of CNNs to missing information (i.e., masking) and the importance of understanding how they handle occlusions and data sparsity compared to token discarding in ViTs.

C Ablation studies

Table 4: Mean (standard deviation) prediction confidence changes for the correct class after masking various number of tokens according to GradCAM and Token Influence.

Number of Tokens	Supervised		DINO		MAE	
	GradCAM	Token Influence	GradCAM	Token Influence	GradCAM	Token Influence
1	0.02 (0.04)	0.06 (0.08)	0.02 (0.05)	0.05 (0.08)	0.01 (0.02)	0.03 (0.05)
3	0.04 (0.06)	0.09 (0.12)	0.04 (0.07)	0.09 (0.13)	0.02 (0.04)	0.05 (0.08)
5	0.06 (0.08)	0.11 (0.14)	0.05 (0.09)	0.11 (0.16)	0.03 (0.04)	0.05 (0.09)
10	0.08 (0.10)	0.14 (0.16)	0.06 (0.11)	0.15 (0.20)	0.04 (0.05)	0.07 (0.10)
20	0.11 (0.13)	0.20 (0.19)	0.09 (0.14)	0.21 (0.24)	0.04 (0.06)	0.08 (0.12)

C.1 GradCAM to identify token influence

We compare our proposed method for quantifying token influence against GradCAM, a widely used technique for visualizing and interpreting neural network predictions. Specifically, we generate both GradCAM heatmaps and token influence maps for each correctly classified image ($\mathcal{D}_C$). Next, we systematically mask the tokens identified as most important according to each method. For GradCAM, we mask the tokens corresponding to the highest activation values in the GradCAM heatmaps (as illustrated in Figure 12). Similarly, for the token influence maps, we mask the tokens with the highest importance scores. We repeat this masking process for varying numbers of tokens, using $n \in 1, 3, 5, 10, 20$. For each value of n, we record the resulting change in model's prediction confidence for the correct class after the masking is applied. Finally, we compute the average change in prediction confidence across all images and report average prediction changes as a measure of how sensitive the model's output is to the tokens identified by each interpretability method in Table 4.

As can be seen in Table 4, masking tokens based on our proposed token influence method consistently leads to larger decreases in prediction confidence for the correct class compared to GradCAM, across all models and token counts.

For all three models — Supervised, DINO, and MAE — the drop in confidence grows as more tokens are masked, indicating that both methods correctly identify influential tokens. However, the proposed token influence approach results in noticeably higher prediction shifts, especially when masking larger number of tokens. This trend highlights that our method more effectively captures tokens critical to the model's decision-making process compared to GradCAM, thereby demonstrating a more precise identification of influential tokens for predicting the correct class.

C.2 GradCAM and attention maps as alternatives to token discarding

As discussed in Section 3.3.1, generating token influence maps involves discarding each token individually and measuring the resulting change in the model's prediction for the correct class. This process requires multiple forward passes and can become infeasible in resource-constrained settings. To mitigate this, we investigate whether attention maps can serve as a more efficient substitute, as they can be obtained in a single forward pass without repeated token masking.

We generate attention maps for all images under $\mathcal{D}_C$ and $\mathcal{D}_I$ and for all three models. Using these attention maps, we compute M-TSI and A-TSI scores in the same manner as with token influence maps. We then assess the degree to which TSI scores using attention maps correlate with TSI scores calculated using token influence maps by calculating Pearson's correlation coefficient. These results are presented in Table 5. Furthermore, we also use the previously generated GradCAM heatmaps to perform the same type of analysis, and present those results in Table 5 as well.

Our results reveal varying levels of correlation across models. For the supervised model, we observe a weak positive correlation, while the MAE model exhibits a moderate positive correlation. In contrast, DINO demonstrates a strong positive correlation between attention-based and token influence-based TSI

Table 5: For both M-TSI and A-TSI, we compute the correlation between TSI scores derived from different interpretation methods across all images in $\mathcal{D}_C$ and $\mathcal{D}_I$. The first table shows the correlation between TSI scores obtained from Token Influence Maps and those from attention maps. The second table shows the correlation between Token Influence Maps and GradCAM patch attributions.

Subset	Metric	Attention Maps			GradCAM		
		Supervised	**DINO**	**MAE**	**Supervised**	**DINO**	**MAE**
$\mathcal{D}_C$	M-TSI	0.308	0.613	0.288	0.213	0.331	0.201
	A-TSI	0.228	0.806	0.329	0.084	0.023	0.007
$\mathcal{D}_I$	M-TSI	0.235	0.591	0.395	0.101	0.285	0.266
	A-TSI	0.168	0.825	0.584	-0.006	0.035	-0.001

scores. These findings suggest that attention maps may serve as a suitable proxy for token influence maps in discriminative self-supervised models such as DINO, but are less reliable for supervised and generative self-supervised models such as MAE. Overall, attention maps offer efficiency gains but do not consistently approximate token influence across models, and should not be relied upon in isolation for fine-grained spurious correlation detection. Moreover, although we explored the use of attention maps as a more efficient alternative to token influence maps, our results show that this substitution is model-dependent and does not consistently approximate token influence across different architectures.

In contrast, using GradCAM heatmaps yields weak to negligible correlation with token influence-based TSI scores across all models. This indicates that GradCAM highlights different regions than those identified as influential by token masking. As a consequence, we discover that GradCAM is not suitable as a proxy for token influence in the context of spurious correlation analysis.

C.3 Attention maps as alternatives to manual annotation

To explore scenarios where annotations are not available or practical, we investigate whether attention maps can serve as an alternative for identifying objects of interest and quantifying spurious correlations since attention maps have been shown to highlight objects of interest in prior research efforts, particularly for models trained in a self-supervised fashion. As such, instead of using annotations to select the object of interest and to identify spurious correlations based on tokens that lie inside the bounding box ($\mathtt{B}_{\text{in}}$) and outside ($\mathtt{B}_{\text{out}}$), we leverage attention maps generated by the model. We progressively select the top 5, 10, 20, 40, and 80 tokens from the attention maps and use them as $\mathtt{B}_{\text{in}}$. By doing so, we compute M-TSI and A-TSI using the token influence maps without relying on bounding boxes, thereby alleviating the need for explicit annotations.

Table 6: For all images in $\mathcal{D}_C$ and $\mathcal{D}_I$, the correlation is calculated between TSI scores obtained using ground-truth bounding boxes and TSI scores obtained using high-attention regions as a substitute (for bounding boxes).

Subset	**Metric**	**Supervised**	**DINO**	**MAE**
$\mathcal{D}_C$	M-TSI	0.487	0.200	0.429
	A-TSI	0.681	0.791	0.835
$\mathcal{D}_I$	M-TSI	0.568	0.482	0.438
	A-TSI	0.703	0.668	0.731

We then examine the correlation between the best-case scenario of TSI scores calculated using attention-based $\mathtt{B}_{\text{in}}$ selections and those computed with using annotated bounding boxes. In Table 6, we report Pearson's correlation coefficients computed across all images under $\mathcal{D}_C$ and $\mathcal{D}_I$.

We observe that M-TSI, which is highly sensitive to the single-largest attention score, exhibits poor correlation with TSI scores computed using token influence maps. In contrast, A-TSI, which aggregates contributions across all tokens in `Bin` and `Bout`, demonstrates moderate positive correlation ($r > 0.6$) and strong positive correlation ($r > 0.7$). This finding indicates that A-TSI scores computed from attention maps as a method of token selection instead of manual annotation have the potential to be a proxy for identifying spurious correlations and assessing model focus, particularly in scenarios where annotations are unavailable. That said, relying solely on attention maps for token selection remains insufficient to fully capture token influence (especially for M-TSI), and care should be taken when interpreting such results, especially in models where attention distributions do not reliably reflect model reasoning.