ATTENTION SHRINKAGE FOR VISION TRANSFORMERS VIA POSTERIOR WEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Transformers (ViTs) utilize multi-head self-attention to capture global context, yet their raw attention scores are inherently noisy and densely distributed, often causing attention to bleed into irrelevant background elements and thereby limiting model interpretability. Regularization of attention scores is understudied, while existing methods require architectural modification, use annotated segmentation, or aggressively hard-thresholding. To address these limitations, we propose adaptive posterior weighting methods – specifically Posterior Probability Shrinkage (PPS) and Wiener Shrinkage – which act as robust, data-driven regularizers. By penalizing attention values proportionally to their likelihood of belonging to a null distribution or by attenuating noisy spectral components, our approaches probabilistically refine attention maps. To overcome the computational burden of resampling, we introduce a joint bootstrap distribution that aggregates null statistics across the dataset. While architecture-agnostic, our proposed posterior weighting improves sparsity, shrinkage, and interpretability. *Without optimizing for localization or segmentation*, regularized attention maps enhance spatial coherence and smoothness. Evaluated on DUTS (a subset of the ImageNet) and LIDC-IDRI (thoracic CT scans) using three ViT architectures, our automated calibration and posterior weighting significantly improve the localization of target regions. Overall, the proposed regularization techniques for attention scores offer a superior balance by producing substantially smoother, less noisy attention distributions, without any hyperparameter tuning. These benefits are pronounced in architectures with inherently noisier attention mechanisms, demonstrating a statistically principled and robust path toward more interpretable visual representations.

1 INTRODUCTION

Vision Transformers (ViTs) have established themselves as a dominant approach in computer vision, exhibiting state-of-the-art representation capabilities across a wide array of image processing tasks (Dosovitskiy et al., 2021; Wu et al., 2020). Central to their success is the multi-head self-attention mechanism (Vaswani et al., 2017), which bypasses local inductive biases by computing attention scores that quantify the relevance between all pairs of input patches. Despite their proven utility in capturing global context, raw attention scores are notoriously noisy and densely distributed. In practice, these mechanisms often lack the necessary signal-to-noise ratio to reliably isolate the most salient features. Instead of focusing sharply on core semantic objects, attention maps frequently "bleed" into the background, highlighting spurious, correlated visual elements. This inherent noise complicates model interpretability and actively limits the efficacy of downstream applications.

To mitigate this noise, recent efforts have proposed various smoothing and sparsity-inducing regularization techniques for attention maps. However, the majority of these approaches require modification of an architecture (Qiang et al., 2025; Srivastava et al., 2014; Baig et al., 2025) or enforce heuristic constraints (Kobyzev et al., 2025; Campos et al., 2026). A recent work addressed this gap by proposing statistical frameworks (Shiraishi et al., 2024; Chung & Laletin, 2026). While the bootstrap method is effective in isolating statistically significant scores, it relies on a binary truncation that aggressively discards attention scores and yet leaves them beyond the threshold intact (Chung & Laletin, 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To probabilistically regularize the attention maps, we propose adaptive weighting methods based on the bootstrap procedure. Posterior probability shrinkage (PPS) smoothly penalizes attention values proportionally to their likelihood of belonging to the null distribution. Wiener shrinkage attenuates spectral components of the attention map according to their likelihood of constituting a noisy pattern, effectively regularizing attention scores not only by their intensity, but also by their spatial distribution. To overcome the computational burden of the bootstrap procedure, we introduce and demonstrate that the joint bootstrap distribution, aggregating the null statistics across samples in the dataset, can be highly efficient and scalable while maintaining accuracy.

We demonstrate using the natural images (DUTS, a subset of the ImageNet (Wang et al., 2017)) and medical images (LIDC-IDRI (Armato et al., 2011)) that the proposed methods operate effectively as robust, data-driven regularizers. Using a classical ViT (Dosovitskiy et al., 2020), DeiT (Touvron et al., 2021) and DINOv2 architecture (Oquab et al., 2024), we characterize the posterior weighting methods, in comparison to other baselines. The impact of regularization are quantified by using top-k attention mass and sparsity, whereas smoothness by average gradient magnitude (AGM).

Spatial coherence is organically improved through proposed regularization, *without optimizing the model for localization or segmentation tasks*. Note that a lack of ground truth makes evaluation of explainable AI methods challenging (Faber et al., 2021; Brocki & Chung, 2023; Rawal et al., 2025), which is why we quantify multiple properties of attention maps. The fact that the proposed methods do not pursue these secondary aims is critical for generalization of *attention regularization*, whereas the methods explicitly optimized for localization or segmentation would serve their distinct, specialized purposes in practice.

2 Related Works

To bridge the gap between the predictive power of ViTs and the necessity for interpretable feature attribution, there has been ex-ante and post-hoc approaches to modify attention.

Model Intervention Interpretable-Aware Vision Transformer (IA-ViT) employs a knowledge distillation loss to align the model's multi-head attention with an auxiliary single-head interpreter (Qiang et al., 2025). By introducing an explicit attention regularization term, IA-ViT forces the multi-head mechanism to align with a singular, unified interpretative focus, sharpening the attention distribution onto class-specific features (Qiang et al., 2025). By adapting Dropout (Srivastava et al., 2014) to attention mechanism, AttentionDrop is a stochastic regularizer designed to counteract the brittleness associated with overly sharp attention distributions (Baig et al., 2025). AttentionDrop formalizes distinct variants to manipulate the attention logits prior to the softmax operation: Hard Attention Masking stochastically zeroes out the top-attention logits per query to prevent overreliance on a fragile subset of dominant features; Blurred Attention Smoothing applies a dynamic one-dimensional Gaussian convolution over the attention logits to diffuse peaked distributions and increase spatial tolerance; and Consistency-Regularized AttentionDrop leverages a Kullback-Leibler (KL) divergence loss to enforce output stability across multiple masked perspectives (Baig et al., 2025). Compared to direct model interventions, our proposed method work with existing models in tact.

Denoising and Smoothing A parallel line of research focuses explicitly on the identification and removal of the "attention noise", where softmax self-attention assigns disproportionate weight to semantically uninformative tokens or background patches. Integral Transformer proposes a novel self-attention mechanism that denoises attention by integrating signals sampled from the logit distribution, mitigating noise while preserving the contributions of critical tokens (Kobyzev et al., 2025). Similarly, the Denoising Attention (DnA) framework models positive and negative interactions using two sets of queries (Campos et al., 2026). DnA projects these interactions onto two separate subspaces to segregate closely associated adversarial or noisy features from relevant ones.

Denoising has been a popular way to improve saliency maps in the context of convolutional neural networks (CNNs). Because vanilla saliency maps which back-propagate gradients from the output score down to the input features (Simonyan et al., 2013) can be noisy and unreliable under certain conditions, researchers have developed various extensions – such as modifying the back-propagation process or aggregating multiple maps (Sundararajan et al., 2017; Selvaraju et al., 2017; Smilkov et al., 2017; Shrikumar et al., 2016). Furthermore, gradient-based approaches have been adapted to

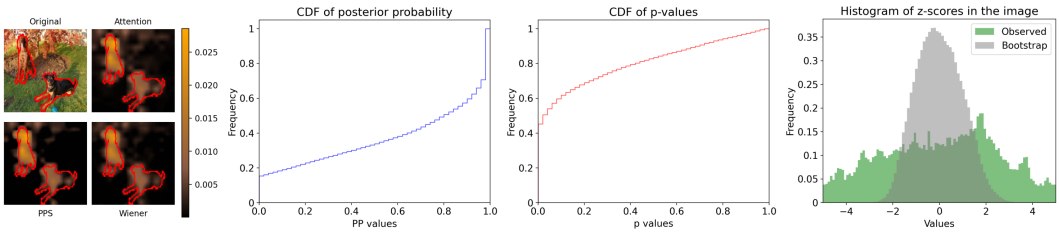

Figure 1: Example of proposed posterior shrinkage methods regularization for the attention map derived from the DINOv2 ViT for `ILSVRC2012_test_00059189` image from the DUTS dataset. The first 4 images include the original image and the attention map (top row) and the attention map after PPS and Wiener shrinkage (bottom row), with annotated segmentation masks in red. The second and third plots are the cumulative distribution function (CDF) of posterior probability (PP, blue) and p-values (red). The last histogram shows the z-score statistics for the observed attention scores (green) and bootstrap attention scores (grey).

enhance attention maps of transformers (Chefer et al., 2021; Brocki & Chung, 2019; Brocki et al., 2024).

Statistical Inference Recognizing the fundamental limitations of imposing arbitrary constraints on attention distributions, Shiraishi et al. (2024) proposes a hypothesis testing that the mean pixel intensity inside a segmentation mask differ from the mean pixel intensity outside that region in ViT. Selective p-value of (Shiraishi et al., 2024) requires and utilizes a segmentation mask in its hypothesis testing setup, resulting in significance on a sample level (i.e., a single p-values for an image). Our proposed methods neither test such a hypothesis, nor use a segmentation mask.

To address the necessity for a data-driven regularization of attention scores on a pixel level, Chung & Laletin (2026) introduces the bootstrap method for regularizing attention scores. This reconstructs the null distribution of attention scores via resampling without requiring a segmentation mask (Chung & Laletin, 2026). This utilized the null distributions from the bootstrap procedures to perform hard-thresholding, aggressively masking out attention scores that fell below a a statistical threshold.

Resampling Strategies Bootstrap techniques approximate the underlying data-generating mechanism through repeated resampling, estimating a statistic's sampling distribution without relying on parametric assumptions (Efron, 1979; Efron & Tibshirani, 1994). Recent progress in computational efficiency has enabled the scalable deployment of these algorithms on massive datasets, ultimately enhancing downstream predictive performance (Kleiner et al., 2014). With rapidly increasing size of unlabeled data, resampling strategies have been extended to unsupervised settings such as latent variable models (Chung & Storey, 2015) and clustering (Chung, 2020). Our proposed methods are founded upon bootstrapping the input data to estimate the distribution of attention scores.

3 METHODS

To obtain uncertainty measures, we follow the bootstrap procedure for attention score (Chung & Laletin, 2026). First, we compute the attention scores for an image $\mathbf{X}$ using model f. Then we create a sample of images $\mathbf{X}$ by either resampling *non-parametrically* from $\mathbf{X}$ or sampling the image pixels from a *parametric* distribution. In this work we use a *non-parametric bootstrap* method that resamples the pixels in the original image with repetition (examples in Figure S1).

Each bootstrap image is processed with the model f and the resulting attention scores constitute the bootstrap sample of attention $\mathbf{A}^*$. The estimated mean $\hat{\mu}$ and variance $\hat{\sigma}$ of $\mathbf{A}^*$ are used to calculate z-scores $z_i = \frac{a_i - \hat{\mu}}{\hat{\sigma}}$ and $z_i^* = \frac{a_i^* - \hat{\mu}}{\hat{\sigma}}$. Note that we use the same parameters $\hat{\mu}$ and $\hat{\sigma}$ to calculate both z and z^*. The z-scores introduced this way allow to evaluate the significance of attention scores in the original map directly or in terms of p-values.

In hard-thresholding, z-statistics and p-values are used to threshold attention scores (Chung & Laletin, 2026). When $z_{\text{th}} = 0$ is used, we refer it as *fixed-threshold* method. Similarly, using p-values, one nullifies the attention scores with p-values above a selected threshold. Sweeping across

different threshold values one can establish an optimal threshold for a desired level of regularization. In our study we perform the threshold-sensitivity analysis using AUROC and AUPRC metrics (see Appendix D.2). Lastly, by estimating a fraction of null scores π_0 (see more details in Section 3.1.1), we may identify a suitable threshold which we call π_0-thresholding.

3.1 Attention Regularization with Posterior Weighting

To regularize attention scores smoothly and probabilistically, we introduce two posterior weighting methods. Both of *posterior probability shrinkage* and *Wiener shrinkage* follow directly from the statistics obtained from the non-parametric bootstrap which is robust to a hyperparameter.

3.1.1 Posterior probability shrinkage

We are interested in estimating *posterior probabilities* of normalized attention scores. It can be interpreted in the Bayesian terms as probability that a feature receives non-zero attention:

$$\gamma_i = \Pr(z_i \neq 0 | p_i) = 1 - \Pr(z_i = 0 | p_i), \tag{1}$$

where i indicates a statistics related to a specific feature. When input features reshaped into 2-dimension, a matrix of posterior probability is denoted as $\mathbf{\Gamma}$. Our posterior probability shrinkage (PPS) weights the attention scores by their respective posterior probabilities (Efron et al., 2001; Storey, 2001):

$$\tilde{\mathbf{A}}^{\gamma} = \mathbf{A} \odot \mathbf{\Gamma}. \tag{2}$$

We can readily compute $\Pr(z_i = 0 \,|\, p_i)$ which is also called the local false discovery rate (LFDR) (Efron et al., 2001; Storey, 2001). Under the standard two-group model each z score is assumed to arise from the mixture $f(z) = \pi_0 f_0(z_i) + (1 - \pi_0) f_1(z_i)$, where $f_0(z_i)$ is the null density, $f_1(z_i)$ is the signal density and π_0 is the prior probability of the null. The LFDR (or l-value) is then defined as

$$l_i = \Pr(z_i = 0 | p_i) = \pi_0 \cdot \frac{f_0(z_i)}{f(z_i)}\,. \tag{3}$$

In practice, $f(z)$ is estimated non-parametrically (e.g. via Poisson regression on binned counts), while π_0 is estimated using Storey & Tibshirani (2003) method. In the most conservative case or due to a lack of sufficient features, assume $\pi_0 = 1$. Note that while we choose to implement PPS in an empirical Bayes framework, there are multiple ways to estimate posterior probability, including fully Bayesian approaches.

Unlike the hard-thresholding methods described above PPS suppresses unwanted attention, but may not remove it completely – hence, PPS typically achieves a softer and spatially-smoother regularization. If a stronger regularization is desired, this method can be applied in combination with π_0-thresholding, which we call PPS_{π_0}.

Algorithm 1 PPS regularization of attention scores

Require: Input sample $\mathbf{X}$, pretrained model f

$\mathbf{A} \leftarrow f(\mathbf{X})$ {observed attention map}
Generate bootstrap null samples $\{\mathbf{X}_1^*, \ldots, \mathbf{X}_B^*\}$ from $\mathbf{X}$ {resampling or parametric}
for $b = 1$ **to** B **do**
  $\mathbf{A}_b^* \leftarrow f(\mathbf{X}_b^*)$ {null (noise-only) attention map}
end for
$\hat{\mu}, \hat{\sigma} \leftarrow$ mean, std of $\mathbf{A}^*$ $\{\mathbf{A}^* = \{\mathbf{A}_1^*, \ldots, \mathbf{A}_B^*\}\}$
Compute standardized z-statistics $\mathbf{Z}, \mathbf{Z}^* \leftarrow \mathbf{A}, \mathbf{A}^*$
Compute p-values $\mathbf{p} \leftarrow \mathbf{Z}$ {see Eq. 7}
Filter out a_i with $z_i < 0$
Estimate the mixture density $f(z)$ {e.g. Poisson regression on binned counts}
Set the null density $f_0(z)$ and prior π_0 {e.g. $\pi_0 = 1$, or estimated}
$l_i \leftarrow \pi_0 \cdot f_0(z_i)/f(z_i)$ {LFDR, Eq. 3}
$\gamma_i \leftarrow 1 - l_i$ {posterior probability, Eq. 1}
$\tilde{\mathbf{A}}^{\gamma} \leftarrow \mathbf{A} \odot \mathbf{\Gamma}$ {shrink each attention score}

3.1.2 WIENER SHRINKAGE

The Wiener filter is grounded in the principle of minimizing the mean squared error (MSE) between the estimated and the true, uncorrupted signal, which makes it the optimal linear estimator under the assumption that signal and noise are wide-sense stationary with known (or estimable) second-order statistics. It has long served as a standard tool for image denoising (Pratt, 1972), with well-established applications across medical imaging (King et al., 1983; Giger et al., 1984; Benesty et al., 2010).

Any image can be represented not only as an array of pixel intensities in the spatial domain, but equivalently as a superposition of elementary oscillatory patterns (spectral components) in the frequency domain – this decomposition if performed by the two-dimensional discrete Fourier transform (2D DFT)

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y)\, e^{-i2\pi(u\cdot x/M + v\cdot y/N)}\,, \tag{4}$$

which maps an image $f(x,y)$ of size $M \times N$ defined over (discreet) spatial coordinates (x,y) into a complex-valued function $F(u,v)$ defined over spatial-frequency coordinates (u,v). The basis functions are two-dimensional sinusoidal plane waves. Low-frequency components vary slowly across the image and capture coarse structure such as overall brightness and smooth intensity gradients, whereas high-frequency components oscillate rapidly and encode fine detail, sharp edges, and texture.

Conceptually, the filter attenuates each spectral component according to its local signal-to-noise ratio: frequencies dominated by the underlying signal are preserved, while those dominated by noise are suppressed. For the two-group model distribution (similar to the one discussed in Section 3.1.1 above Equation 3), this amounts to weighting each frequency component by the factor $S_f/(S_f+S_n)$, where S_f and S_n denote the power spectral (PSD) $S \propto |F(u,v)|^2$ of the signal and the noise (null), respectively. We propose to estimate the null PSD from the bootstrap attention maps $\mathbf{A}^*$ that undergo 2D DFT. Unlike the methods that we considered before, the spatial structure of the bootstrap attention maps is important. The noise PSD is constructed as a sum of the individual bootstrap PSD. In contrast with the PPS, we attenuate the attention scores in the *frequency space* (2D DFT representation $\mathbf{G}$) by the corresponding posterior probability weights $H \propto S_f/(S_f+S_n)$

$$\hat{\mathbf{F}} \leftarrow \mathbf{G} \odot \mathbf{H}\,. \tag{5}$$

The regularized attention map is further reconstructed from the spectral distribution using an inverse DFT. This adaptive, frequency-selective weighting enables the Wiener shrinkage to regularize attention maps while retaining salient structural detail.

Algorithm 2 Wiener shrinkage of attention maps

Require: Image sample $\mathbf{X}$, pretrained model f

$\mathbf{A} \leftarrow f(\mathbf{X})$ {attention map of the original image $\mathbf{X}$}
Generate bootstrap null samples $\{\mathbf{X}_1^*, \ldots, \mathbf{X}_B^*\}$ from $\mathbf{X}$
for $b = 1$ **to** B **do**
  $\mathbf{A}_b^* \leftarrow f(\mathbf{X}_b^*)$ {null (noise-only) attention map}
  $\mathbf{N}_b^* \leftarrow \mathcal{F}(\mathbf{A}_b^*)$ {2D DFT of null map}
end for
$\mathbf{S}_{nn} \leftarrow \frac{1}{B}\sum_{b=1}^{B} |\mathbf{N}_b^*|^2$ {estimate noise PSD}
$\mathbf{G} \leftarrow \mathcal{F}(\mathbf{A})$ {2D DFT of observed map}
$\mathbf{S}_{ff} \leftarrow \max\big(|\mathbf{G}|^2 - \mathbf{S}_{nn},\ 0\big)$ {estimate signal PSD}
$\mathbf{H} \leftarrow \dfrac{\mathbf{S}_{ff}}{\mathbf{S}_{ff} + \mathbf{S}_{nn} + \epsilon}$ {Wiener filter, $H \in [0,1]$}
$\hat{\mathbf{F}} \leftarrow \mathbf{G} \odot \mathbf{H}$ {apply filter elementwise}
$\tilde{\mathbf{A}}^w \leftarrow \mathrm{Re}\Big(\mathcal{F}^{-1}(\hat{\mathbf{F}})\Big)$ {inverse 2D DFT}

4 EXPERIMENTS AND MATERIALS

In this Section, we briefly describe the datasets, models, bootstrap procedure, and evaluation metrics used in our experiments. Further details and descriptions are available in Appendix A.

4.1 DATA

We perform our analysis on two different datasets of images together with the corresponding binary mask annotations indicating salient objects in these images.

Natural Images. The DUTS dataset is a large-scale salient object detection benchmark with an explicit training/test protocol (Wang et al., 2017), based on the ImageNet (Deng et al., 2009). The dataset comprises $10{,}553$ training images (DUTS-TR), collected from the ImageNet DET training and validation sets, and $5{,}019$ test images (DUTS-TE), collected from the ImageNet DET test set (Deng et al., 2009) and the SUN database (Xiao et al., 2010). Each image is paired with a pixel-level ground-truth saliency mask manually annotated by 50 subjects, and both subsets contain challenging scenarios for saliency detection. As our analysis is performed on held-out data, we use only the DUTS-TE subset. From it, we discard any image whose ground-truth mask contains fewer than 490 non-zero pixels, ensuring a sufficient spatial extent of the salient object for the analysis. We randomly sample from the remaining images yielding a final evaluation set of 350 images, each consisting of an RGB image and its spatially aligned binary saliency mask.

Medical Images. The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) (Armato et al., 2011) is a publicly available collection of thoracic computed tomography (CT) scans designed for lung nodule analysis. It comprises $1{,}018$ CT studies acquired across a set of institutions, scanners, and acquisition protocols – each study was independently reviewed by up to four radiologists providing subjective annotations and contour segmentations of the important regions. From the full CT scans, we derive a nodule-centric dataset (854 nodules from 523 distinct CT scans) and from each volumetric crop, we extract a single central axial slice. For rigorous evaluation, we use only the test partitions and we retain the malignant nodules from folds 1 through 4 and discard any instance whose mask, at the selected slice, contains 490 or fewer non-zero pixels, resulting in 322 images, which ensures a sufficient spatial extent for the analysis.

4.2 MODELS

We use three ViT models: supervised (Vanilla) ViT with a patch size of 16 (Dosovitskiy et al., 2021), self-supervised DINOv2 with registers with a patch size of 14 (Oquab et al., 2024; Darcet et al., 2024), and distilled DeiT with a patch size of 16 (Touvron et al., 2021). For the analysis on the LIDC-IDRI dataset we use the models that are fine-tuned for binary classification between malignant and benign nodules. From each fine-tuned model we extract the attention of the `[CLS]` token over the patch tokens in the final transformer layer, averaged across the 12 attention heads. For the supervised ViT this is the `[CLS]`-to-patch attention over the 14×14 grid. For DINOv2 we likewise take the `[CLS]` query row but discard the entries corresponding to the `[CLS]` and register tokens, retaining only the attention over the 256 patch tokens of the 16×16 grid. For DeiT we also take the `[CLS]` query row discarding the `[DIST]` token, yielding 14×14 attention patches.

4.3 BOOTSTRAP METHODS

To generate bootstrapped images, we non-parametrically resample an image with replacement which makes no distributional assumption about pixel values. The key advantage of our proposed methods is a lack of hyperparameter tuning. In the bootstrap, a single hyperparameter B denotes the number of bootstrap iterations. In general, after a reasonable amount of z-scores from the bootstrap samples, B only has an impact on the resolution of estimated probabilities. To characterize this, we conducted an experiment where B is varied, while p-values and posterior probabilities are computed. In our study, we use $B = 20$ as the size of the bootstrap image sample for each image, which balances the fidelity of the sample against computational cost, as we demonstrate in Appendix B.

Constructing a reliable bootstrap distribution for a single image typically requires generating several bootstrap images and passing each through the ViT, so that an marginal distribution of null attention scores are estimated accurately at a reasonably high resolution. To reduce the computational time

for bootstrap generation we propose to employ a **joint bootstrap distribution**, which is constructed from a large number of images in a comparable distribution and then take B samples out of it. Given that CT scans in LIDC-IDRI dataset are naturally similar, we demonstrate and validate this procedure for the analysis on LIDC-IDRI dataset and we prepare the joint bootstrap distribution from 3 bootstrap distributions for each CT scan in the dataset. Further details about comparison between the marginal and joint bootstrap processes are in Appendix C.

4.4 Evaluation Metrics

We evaluate attention maps in shrinkage, sparsity, contrast, smoothness, and other properties. While their metrics are briefly introduced, the details can be found in Appendix A. We refer to the region of the image annotated by the mask as the region of interest (ROI) and the rest of the image we denote as region outside of ROI or as background.

Top-k attention mass. Top-k attention mass indicates the minimum number of patches/pixels that are required to capture top k percent of the total attention mass. In this work we use Top-50 attention mass, which we denote as TP50 and measure it in the percentage of the total number of attention map pixels. We also report the **TP50 ratio**, which is the value of TP50 after regularization divided by its value prior to regularization.

Sparsity. We measure the fraction of non-zero attention scores after shrinkage, while noting that attention scores before regularization are almost always non-zero in practice. Additionally, we use the **non-zero pixel fraction**, which is the sum of non-zero pixels in and outside of ROI divided by the total image size.

Average gradient magnitude. To quantify how visually contrasting the attention maps are, we use the average gradient magnitude (AGM) contrast: $\text{AGM} = \frac{1}{m}\sum_{i=1}^{m}\sqrt{(\nabla_x \mathbf{A})_i^2 + (\nabla_y \mathbf{A})_i^2}$, where x and y denote the horizontal and vertical gradient directions respectively.

Energy ratio. To assess how well an attention map localizes the annotated ROI we employ the energy ratio (ER). This quantity measures the fraction of total attention mass a that falls within ROI defined as $\text{ER} = \sum_i (a_i \in \text{ROI}) \,/\, \sum_i a_i$.

AUROC and AUPRC. Area under the Receiver Operating Characteristic (ROC) curve (recall vs. false-positive rate) summarizes localization in a single threshold-free number. AUROC admits a direct probabilistic interpretation: it equals the probability that a randomly chosen in-ROI pixel receives a higher attention value than a randomly chosen out-of-ROI pixel.

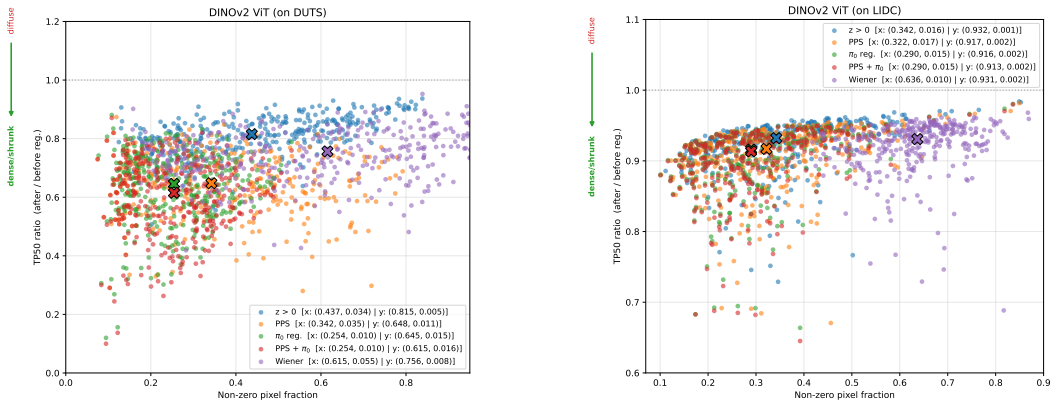

Figure 2: The ratio of TP50 metric after and before regularization (y-axis) vs. the fraction of non-zero attention scores (x-axis) for DINOv2 ViT on DUTS (left) and LIDC-IDRI (right) for different methods: z-thresholding (blue), PPS (orange), π_0-thresholding (green), PPS_{π_0} (red) and Wiener (purple). Each dot corresponds to the results for one image. The median and variance for each distribution dimension are reported next to each legend entry: [x: (median, var), . . .]. The intersection of the corresponding median values for each method are shown with the cross signs of the respective color.

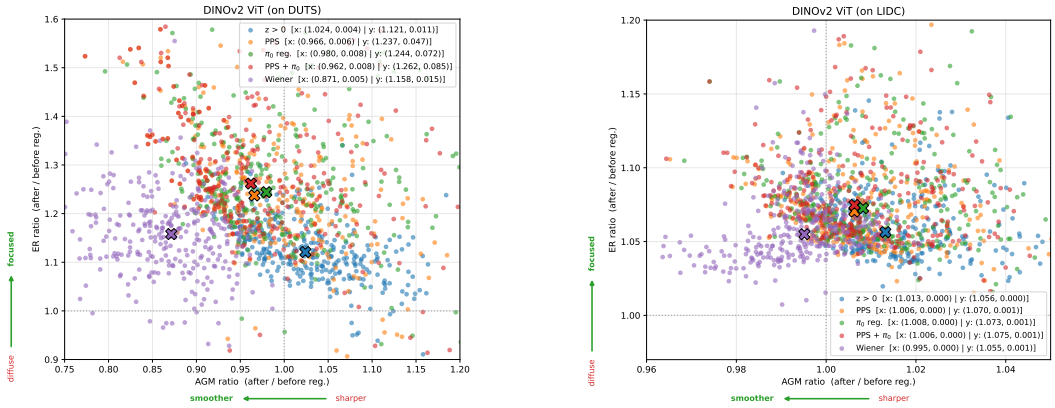

Figure 3: The ratios of two metric values after and before regularization: ER ratio (y-axis) vs. AGM contrast (x-axis) for DINOv2 ViT on DUTS (left) and LIDC-IDRI (right). The panel arrangement and graphical conventions follow those of Figure 2.

The Precision-Recall curve is constructed over the same range of thresholds as ROC, but plots recall against *precision* (fraction of true positives out of all predicted positives); its area, the average precision (AUPRC), measures how strongly the highest-attention pixels concentrate within the ROI. We report the AUPRC divided by the fraction of the ROI in the image π.

5 Results

We evaluate our posterior weighting methods on the attention maps obtained for every image in each dataset with three vision transformers. For each attention map we apply several regularization methods and, for each method, compute the metrics introduced in the previous section. We highlight the key features of our results using DINOv2, referring to the other models (see Appendices E and F) where their behaviour differs or provides additional insights. An example of the regularization produced by the posterior weighting methods for the attention map from DINOv2 on DUTS image, together with the underlying statistics used to derive it, is shown in Figure 1. Other examples of regularization for DINOv2 on both datasets can be found in Appendix D.1, in Appendix E.1 for Vanilla ViT and in Appendix F.1 for DeiT.

Our regularization methods consistently concentrate attention into fewer pixels and sparser maps, with the different methods forming a clear ordering in how aggressively they compress the attention mass, as can be seen in Figure 2 for DINOv2 and in Figure S25 and Figure S36 for Vanilla ViT and DeiT respectively, where we plot the TP50 ratio vs. the non-zero pixel fraction. The posterior weighting methods typically yield the levels of TP50, which sit in the middle between pre-regularization and π_0-thresholding results. PPS_{π_0} shrinkage suppresses the attention mass a bit further than π_0-thresholding itself and slightly increases the concentration of attention. The non-zero pixel fraction values suggest that the suppression of attention for posterior weighting methods happens mainly due to the intensity scaling instead of nullifying attention scores explicitly, especially for the Wiener shrinkage, which yield the lowest sparsity (although a detailed analysis of in-ROI and out-of-ROI non-zero pixels presented in Figure S16 shows that the Wiener shrinkage nullifies the least pixels in the important regions). This is consistent with the softer regularization and higher smoothness seen in Figure 3.

The concentration of attention demonstrated above by the decrease of TP50 results in the improvement of the attention localization, which can be seen in Figure 3, where we plot the ER ratio vs. the AGM contrast ratio for DINOv2 on both datasets (see also Figure S26 for Vanilla ViT and Figure S37 for DeiT). A point above unity on the vertical axis denotes better localization and a point below unity on the horizontal axis denotes a smoother, less contrasting map[1]. For DINOv2 and

[1]As a single metric can be optimized aggressively, it is important to consider multiple metrics when characterizing regularization. When comparing to hard-thresholding methods, we consider a smoother attention distribution as a feature of probabilistic soft-thresholding.

DeiT (on LIDC-IDRI, right plot of Figure S37) the largest improvement in localization is achieved by the PPS and PPS$_{\pi_0}$ methods, while the Wiener shrinkage yields moderate localization improvement, but markedly smoother maps (for Vanilla ViT on LIDC-IDRI the Wiener shrinkage result is in the same ballpark in terms of localization as the other methods, as can be seen inthe right subplot of Figure S26). Notably, the combination of π_0-thresholding and PPS shrinkage not only produces a less contrastive map, but slightly improves localization. The effect on localization and smoothness for DINOv2 on LIDC-IDRI dataset is practically insignificant as the model produces very clean and focused attention maps from the outset (see examples in Figs. S8-S11). The results for Vanilla ViT and DeiT on the DUTS dataset (left subplots in Figure S26 and Figure S37) are quite scattered in terms of ER modification, yielding very moderate or even negative average localization improvements, which is the consequence of a poor quality of the attention maps produced by these models (as can be seen in Figs. S17-S24 for Vanilla ViT and Figs. S28-S35 for DeiT respectively these maps are rather diffuse with multiple attention artifacts that are more pronounced than the important salient features). The improvement of localization by our methods are corroborated by the results of the AUROC/AURPC analysis, which can be found in Appendix D.2.

In Appendix B, we show the properties of the bootstrapping and the joint bootstrap distribution for LIDC-IDRI which is demonstrated as an efficient and scalable solution (described in Section A.4). A marginal bootstrap distribution computed from a single sample approaches a joint bootstrap distribution as B increases. Figure S2 demonstrates how an individual bootstrap yields a coarse distribution, which are combined $B = 20$ times for a smooth marginal distribution for that sample. Repeating B bootstrap procedures for all samples may be bottlenecked by the compute resources. Thus, for CT scans of LIDC-IDRI, which share similar image characteristics, we propose and validate usage of the joint bootstrap distribution, which approximate marginal distributions when computing p-values and posterior probabilities (Figure 4 and Figure S3).

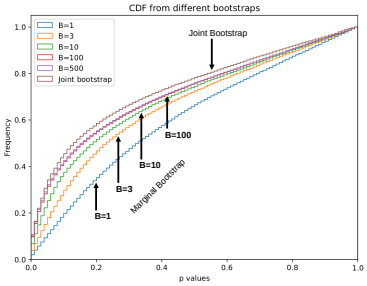

Figure 4: CDFs of p-values computed marginally with $B = 1$–500 and jointly (brown).

6 Conclusion

We introduce posterior weighting methods which attenuate attention values probabilistically, leading to smooth regularization. Namely, posterior probability shrinkage and Wiener shrinkage downweight attention in proportion to its posterior probability of belonging to the null distribution of attention magnitude or to the null spectral distribution, respectively. As these weights are derived directly from the model's own response to bootstrapped data, posterior weighting methods represent data-driven regularizers that require no manual hyperparameter tuning. Furthermore, we demonstrate that a joint bootstrap distribution is highly efficient and accurate with respect to marginally bootstrapping, paving the way for scalable implementation.

The proposed methods readily achieve great sparsity and interpretability without hyperparameter tuning in all ViT models across both domains. Our probabilistic framework provides a rigorous approach to regularization via the bootstrapped data, without relying on localization or segmentation. Yet, as a by-product, the proposed framework consistently improves the localization of attention maps, concentrating a larger share of the attention mass within ROI. Notably, the benefits of our regularization are most pronounced for the noisier attention maps produced by the original DINOv2 ViT on DUTS dataset, whereas the already clean maps produced by the fine-tuned DINOv2 ViT on LIDC-IDRI dataset leave comparatively little room for improvement, indicating that our methods are most valuable precisely where the underlying attention is least interpretable.

Future works include extending our approach to wider transformer architectures and studying how the smoothness-localization trade-off can be tuned to the needs of downstream applications. Overall, posterior weighting offers a robust, automatically calibrated, and principled route to attention regularization.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing (e.g., improving grammar, clarity, and readability), for retrieval and discovery of related work and to assist with writing some parts of our code. We have not used generative AI tools for generating research ideas, designing or conducting experiments or analyzing data. We have reviewed all AI-assisted work. All AI-drafted and AI-edited text was carefully read, revised, and checked by the authors to ensure it accurately reflects our methods, results, and claims. All related work suggested by AI tools was manually verified by the authors, who confirmed that each cited paper exists, read the relevant sources, and checked that it is accurately described and appropriately cited. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our results. Upon publication, we will make our main code framework publicly available. All datasets used in this work will also be made publicly available.

ACKNOWLEDGMENTS

We thank Jakub Binda for providing the LIDC-IDRI dataset preparation pipeline and the ViT models fine-tuned on this dataset for our study. This work was funded by the SONATA BIS [2023/50/E/ST6/00694] from Narodowe Centrum Nauki, with computational resources of Interdisciplinary Centre for Mathematical and Computational Modelling University of Warsaw [GDM-3540].

REFERENCES

Samuel G. Armato, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, Heber MacMahon, Edwin J. R. van Beek, David Yankelevitz, Andrea M. Biancardi, Pamela H. Bland, Matthew S. Brown, Roger M. Engelmann, Günter E. Laderach, David Max, R. C. Pais, Dean P.-Y. Qing, R. Y. Roberts, Anthony R. Smith, Adam Starkey, Puneet Batra, Paul Caligiuri, Ayman Farooqi, Gary W. Gladish, Christine M. Jude, Reginald F. Munden, Iva Petkovska, Leslie E. Quint, Lawrence H. Schwartz, Balamurugan Sundaram, and Gerald D. Dodd. The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical Physics*, 38(2):915–931, 2011. doi: 10.1118/1.3528204.

Mirza Samad Ahmed Baig, Syeda Anshrah Gillani, Abdul Akbar Khan, Shahid Munir Shah, and Muhammad Omer Khan. AttentionDrop: A novel regularization method for transformer models. *arXiv preprint arXiv:2504.12088*, 2025.

Jacob Benesty, Jingdong Chen, and Yiteng Huang. Study of the widely linear Wiener filter for noise reduction. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 205–208. IEEE, 2010. doi: 10.1109/ICASSP.2010.5496033. URL https://ieeexplore.ieee.org/document/5496033/.

Lennart Brocki and Neo Christopher Chung. Concept saliency maps to visualize relevant features in deep generative models. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pp. 1771–1778. IEEE, 2019.

Lennart Brocki and Neo Christopher Chung. Fidelity of interpretability methods and perturbation artifacts in neural networks. In Krystal Maughan, Rosanne Liu, and Thomas F. Burns (eds.), *Tiny Papers at International Conference on Learning Representations 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=nbqO93YTz-.

Lennart Brocki, Jakub Binda, and Neo Christopher Chung. Class-discriminative attention maps for vision transformers. *Transactions on Machine Learning Research*, 2024.

Ron Campos, Subhajit Maity, Xin Li, Srijan Das, and Aritra Dutta. DnA: Denoising attention for visual tasks. *arXiv preprint arXiv:2606.27372*, 2026.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021.

Neo Christopher Chung. Statistical significance of cluster membership for unsupervised evaluation of cell identities. *Bioinformatics*, 36(10):3107–3114, 2020.

Neo Christopher Chung and Maxim Laletin. Regularizing attention scores with bootstrapping. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026. URL https://arxiv.org/abs/2604.01339.

Neo Christopher Chung and John D. Storey. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, 31(4):545–554, feb 2015. doi: 10.1093/bioinformatics/btu674. URL https://doi.org/10.1093/bioinformatics/btu674.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

Lukas Faber, Amin K. Moghaddam, and Roger Wattenhofer. When comparing to ground truth is wrong: On evaluating gnn explanation methods. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 332–341, 2021.

Maryellen Lissak Giger, Kunio Doi, and Charles E. Metz. Investigation of basic imaging properties in digital radiography. 2. Noise Wiener spectrum. *Medical Physics*, 11(6):797–805, 1984. doi: 10.1118/1.595583. URL https://aapm.onlinelibrary.wiley.com/doi/10.1118/1.595583.

Michael A. King, Paul W. Doherty, Ronald B. Schwinger, and Bill C. Penney. A Wiener filter for nuclear medicine images. *Medical Physics*, 10(6):876–880, 1983. doi: 10.1118/1.595352. URL https://aapm.onlinelibrary.wiley.com/doi/10.1118/1.595352.

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(4):795–816, 2014.

Ivan Kobyzev, Abbas Ghaddar, Dingtao Hu, and Boxing Chen. Integral transformer: Denoising attention, not too much not too little. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2337–2354, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

William K. Pratt. Generalized Wiener filtering computation techniques. *IEEE Transactions on Computers*, C-21(7):636–641, 1972. doi: 10.1109/T-C.1972.223567. URL https://ieeexplore.ieee.org/document/1672160/.

Yao Qiang, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu. Interpretability-aware vision transformer. In *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025. URL https://arxiv.org/abs/2309.08035.

Kaivalya Rawal, Zihao Fu, Eoin Delaney, and Chris Russell. Evaluating model explanations without ground truth. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 3400–3411, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732219. URL https://doi.org/10.1145/3715275.3732219.

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.

Tomohiro Shiraishi, Daiki Miwa, Teruyuki Katsuoka, Vo Nguyen Le Duy, Kouichi Taji, and Ichiro Takeuchi. Statistical test for attention map in vision transformer. *arXiv*, 2024. doi: 10.48550/arxiv.2401.08169.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

John D Storey. *The false discovery rate: A Bayesian interpretation and the q-value*. Department of Statistics, Stanford University, 2001.

John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

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

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv e-prints*, Jun 2020.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

A Extended Methods and Evaluation Frameworks

A.1 p-values

p-values serve as a measure of the statistical significance of attention scores and can be computed from z-scores as the probability $P(z^* \geq z_i)$ that a certain z_i-score is lower than a random z^* value from the bootstrap distribution.

$$P(z^* \geq z_i) = \frac{\#\{z^*_{jb} \geq z_i;\ j = 1, \ldots, m,\ b = 1, \ldots, B\}}{mB}, \tag{6}$$

where m is the number of pixels and B is the number of bootstrap samples.

Thus, attention scores with p-values close to 0 can be interpreted as positive signal, while those with p-values close to 1 as the noise. We use these and similar statistics to decide which attention scores can be regularized. The corresponding p-values are obtained using a two-sided empirical formula.

$$p_i = 2 \times \min\left(P(z^* \geq z_i), P(z^* \leq z_i)\right). \tag{7}$$

A.2 Data

A.2.1 DUTS

We perform our experiments on DUTS Wang et al. (2017), a large-scale salient object detection benchmark with an explicit training/test protocol. The dataset comprises $10{,}553$ training images (DUTS-TR), collected from the ImageNet DET training and validation sets, and $5{,}019$ test images (DUTS-TE), collected from the ImageNet DET test set and the SUN database. Each image is paired with a pixel-level ground-truth saliency mask manually annotated by 50 subjects, and both subsets contain challenging scenarios for saliency detection.

Evaluation set. As our analysis is performed on held-out data, we use only the DUTS-TE subset. From it, we discard any image whose ground-truth mask contains fewer than 490 non-zero pixels, ensuring a sufficient spatial extent of the salient object for the analysis. We randomly sample from the remaining images yielding a final evaluation set of 350 images, each consisting of an RGB image and its spatially aligned binary saliency mask.

A.2.2 LIDC-IDRI

Additionally, we conduct our experiments on the dataset from the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) (Armato et al., 2011), which is a publicly available collection of thoracic computed tomography (CT) scans designed for lung nodule analysis. It comprises $1{,}018$ CT studies acquired across a set of institutions, scanners, and acquisition protocols. Each study was independently reviewed by up to four thoracic radiologists in a two-phase protocol, in which each radiologist first marked findings independently and then revised their marks in light of the others' anonymized assessments, without forced consensus. For nodules ≥ 3 mm, this yields slice-by-slice contour segmentations together with subjective ratings of nine characteristics, including likelihood of malignancy.

Nodule-centric preprocessing. From the raw data we derive a nodule-centric dataset in which each annotated nodule is represented by a $32 \times 32 \times 32$ volumetric crop centered on the nodule. The per-radiologist annotations are consolidated into a single reference representation: one binary segmentation mask and characteristic ratings averaged over the annotating radiologists. Each nodule receives a binary label derived from the radiologists' malignancy assessments (malignant vs. benign). This yields 854 nodule instances from 523 distinct CT scans (411 malignant, 443 benign), partitioned into five cross-validation folds with fixed train/test assignments. The folds are split at the nodule level, so nodules from one scan may appear across splits; we did not enforce scan-level grouping.

Slice extraction and normalization. From each volumetric crop we extract a single central axial slice. The slice is windowed to the $[-1000, 400]$ Hounsfield-unit range (air to soft tissue) and

min–max rescaled to a unit intensity range to standardize contrast across scanners. It is then replicated across three channels to match an ImageNet-pretrained backbone, resized to 224×224, and standardized with ImageNet channel statistics; the corresponding mask is carried through at the same slice position and resized with nearest-neighbor interpolation to preserve its binary values. No geometric augmentation is applied at evaluation time.

Evaluation set. As our study analyzes attention maps on held-out nodules, we use only the test partitions of the folds. From the test partitions of folds 1 through 4 we retain the malignant nodules and discard any instance whose mask, at the selected slice, contains 490 or fewer non-zero pixels, ensuring a sufficient spatial extent for the analysis. Pooling across the four folds yields a final set of 322 masked nodule images, each consisting of a 224×224 axial CT slice, a spatially aligned binary mask, and a class label.

A.3 Models and Attention

Our analysis operates entirely on attention maps: we extract and study the attention maps from each of the three ViT models. Our goal is *not* to improve classification accuracy through regularization. The classifier serves solely to adapt each model to the domain and to induce attention patterns; our study instead concentrates on the attention maps themselves – how regularization affects properties such as the overall attention mass distribution and the spatial localization of attention.

Architectures. We consider three ViT backbones:

- **Supervised ViT** (`google/vit-base-patch16-224`) (Dosovitskiy et al., 2021), pretrained in a supervised manner on ImageNet-21k. With a patch size of 16, it partitions a 224×224 input into a 14×14 grid of 196 patch tokens, prepended by a single classification (`[CLS]`) token. We shall refer to this ViT as **Vanilla ViT**.
- **DINOv2 with registers** (`facebook/dinov2-with-registers-base`) (Oquab et al., 2024; Darcet et al., 2024), a ViT pretrained with the DINOv2 self-supervised framework, which learns representations without labels through a self-distillation objective. We use the variant augmented with register tokens, which were introduced to absorb the high-norm artifact tokens that otherwise corrupt DINOv2 attention maps. With a patch size of 14, it produces a 16×16 grid of 256 patch tokens, alongside a `[CLS]` token and four register tokens.
- **Distilled DeiT** (`facebook/deit-base-distilled-patch16-224`) (Touvron et al., 2021), a ViT-Base trained in a supervised manner on ImageNet-1k only, using knowledge distillation from a convolutional teacher (RegNetY-16GF). To this end, it introduces an additional learnable distillation (`[DIST]`) token that interacts with the `[CLS]` and patch tokens through self-attention and is trained to reproduce the teacher's predictions, while the `[CLS]` token is supervised by the ground-truth labels. With a patch size of 16, it produces the same 14×14 grid of 196 patch tokens as the Vanilla ViT, preceded by a `[CLS]` token and a `[DIST]` token.

Fine-tuning. For the analysis on the LIDC-IDRI dataset (Section A.2.2) we use the models that are fine-tuned on the subset of this dataset for binary classification between malignant and benign nodules. Training uses the 683 images of fold 1 (328 malignant, 355 benign), with the remaining 171 images (83 malignant, 88 benign) used for evaluation. We minimize the cross-entropy loss with the AdamW optimizer at a learning rate of 1×10^{-4}. The supervised ViT was trained for 13 epochs, DINOv2 model for 14 epochs and DeiT for 12, reaching classification accuracies of 0.89, 0.90 and 0.88, respectively, on the fold-1 test set.

Attention-map extraction. From each fine-tuned model we extract the attention of the `[CLS]` token over the patch tokens in the final transformer layer, averaged across the 12 attention heads. For the supervised ViT this is the `[CLS]`-to-patch attention over the 14×14 grid. For DINOv2 we likewise take the `[CLS]` query row but discard the entries corresponding to the `[CLS]` and register tokens, retaining only the attention over the 256 patch tokens of the 16×16 grid. For DeiT we also take the `[CLS]` query row discarding the `[DIST]` token, yielding 14×14 attention patches. In all three cases the resulting attention map is rescaled to the input resolution of 224×224 using bilinear interpolation, and all subsequent regularization and analysis are performed on these rescaled maps.

A.4 BOOTSTRAP METHOD

To generate bootstrapped images, we non-parametrically resample an image with replacement which makes no distributional assumption about pixel values. Particularly, the bootstrap is performed at the level of whole pixels rather than individual channel values, so the channel structure of each pixel is preserved. As a result, the bootstrap image approximately preserves the marginal distribution of pixel intensities of the original image while its spatial structure is destroyed, yielding a spatially unstructured, noise-like image that carries no coherent object or shape. More details about the bootstrap method, including the examples of bootstrap images and the corresponding attention maps can be found in Appendix B.

The bootstrap iteration. The key advantage of our proposed methods is a lack of hyperparameter tuning. In the bootstrap, a single hyperparameter B denotes the number of bootstrap iterations. In general, after a reasonable amount of z-scores from the bootstrap samples, B only has an impact on the resolution of estimated probabilities.

To characterize this, we conducted an experiment where B is increased from 1 to 100, while p-values are computed. For certain pixels, we plotted their p-values with varying B (we note that p-values stabilizes/converge quickly). In our experiments, we use $B = 20$ as the size of the bootstrap image sample for each image from the dataset, which balances the fidelity of the sample against computational cost.

The joint bootstrap distribution. Constructing a reliable bootstrap distribution for a single image typically requires generating several bootstrap images and passing each through the ViT, so that an marginal distribution of null attention scores are estimated accurately at a reasonably high resolution. Repeating this for every image in the dataset is expensive, as the number of ViT forward passes scales with both the number of images and the number of bootstrap samples per image.

We observe that this cost can be reduced drastically when the dataset is homogeneous in its low-level characteristics, as in the case of LIDC-IDRI dataset. In this regime, the per-pixel intensity distributions, from which the bootstrap resamples, are comparable across images. This motivates sharing marginal distributions for samples with the *joint bootstrap distribution* pooling across the dataset. We generate bootstrap images for every image in LIDC-IDRI dataset, pass all of them through each ViT once, and pool the resulting attention scores into a single distribution. We draw from this joint distribution a random sample of the same size that per-image bootstrapping would have produced.

This procedure offers two advantages. First, because it aggregates statistics for attention scores across the entire dataset, the joint distribution is smoother and more balanced than a marginal distribution estimated from a single image's bootstrap samples, which is based on far fewer draws (see Figure S2). Second, the ViT passes are performed only once and the pooled distribution is reused for all images, which accelerates attention regularization and makes the approach practical as a reusable component in other applications. We compare the individual sampling and joint bootstrap sampling in detail in Appendix C.

A.5 EVALUATION METRICS

In this Section we list all the evaluation metrics that we use in our analysis. We generally refer to the region of the image annotated by the mask as the region of interest (ROI) and the rest of the image we denote as region outside of ROI or as background.

Top-k attention mass. Top-k attention mass indicates the minimum number of patches/pixels that are required to capture top k percent of the total attention mass. Diffuse attention maps, where scores are spread across many irrelevant patches, obscure the semantic reasoning of ViTs and hinder interpretability. Top-k attention mass provides a principled measure of attention concentration: lower values indicate tight focus on a coherent set of regions, while higher values suggest diffuse, uninformative activation. This makes it a natural tool for evaluating whether regularization improves the spatial selectivity of attention maps.

In this work we use Top-50 attention mass, which we denote as TP50 and measure it in the percentage of the total number of attention map pixels.

Average gradient magnitude. To quantify how visually contrasting the attention maps are, we use the average gradient magnitude (AGM) contrast.

$$\text{AGM} = \frac{1}{m}\sum_{i=1}^{m}\sqrt{(\nabla_x \mathbf{A})_i^2 + (\nabla_y \mathbf{A})_i^2}\,, \tag{8}$$

where x and y denote the horizontal and vertical gradient directions respectively. Numerically the gradients are computed for each attention score using `numpy.gradient(A)` function. A low value of AGM indicates a diffuse map in which attention is distributed nearly uniformly, whereas a high value indicates an inhomogeneous map with sharply peaked features.

Sparsity. As all the regularization methods that we consider (except for Wiener shrinkage) explicitly set to 0 the attention scores attributed to the noise distribution, the number of remaining **non-zero attention pixels** indicates the strength of regularization. By measuring the number of non-zero pixels in different regions of the image one can deduce how the regularization preserves important features. In this work, we measure the fraction of non-zero pixels before and after shrinkage while noting that raw attention scores (before regularization) are almost always non-zero in practice. We also report the fraction of non-zero pixels in and outside of ROI.

Energy ratio. To assess how well an attention map localizes the annotated ROI we employ the energy ratio (ER). This quantity measures the fraction of total attention mass a that falls within ROI defined as

$$\text{ER} = \frac{\sum_i a_i \in \text{ROI}}{\sum_i a_i}\,. \tag{9}$$

It ranges from 0 to 1. Crucially, the metric penalises competing activations outside the ROI, as any attention mass leaked to irrelevant patches inflates the denominator without contributing to the numerator. A higher energy ratio directly implies greater localization of the ROI, as it indicates that the model's focus is concentrated on regions that a human expert would identify as meaningful, rather than dispersed across uninformative background patches.

AUROC. To further investigate how well our regularization preserve meaningful spatial structures, we frame localization as an ad hoc classification problem at a pixel level: each pixel's attention value serves as a score, and the mask provides the ground-truth label where pixels inside and outside the mask are positives and negatives, respectively. Sweeping a decision threshold across the full range of attention values and recording the true-positive rate (*recall* - fraction of true positives out of all ground-truth positives) at each threshold against the false-positive rate (fraction of false positives out of all ground-truth negatives) traces the Receiver Operating Characteristic (ROC) curve; the area under this curve (AUROC) summarizes localization in a single threshold-free number.

AUROC admits a direct probabilistic interpretation: it equals the probability that a randomly chosen in-ROI pixel receives a higher attention value than a randomly chosen out-of-ROI pixel. A value of 0.5 corresponds to no localization ability (attention is unrelated to the mask), while 1 indicates that every in-ROI pixel is ranked above every out-of-ROI pixel.

AUPRC. The Precision-Recall curve is constructed over the same range of thresholds as ROC, but plots recall against *precision* (fraction of true positives out of all predicted positives); its area, the average precision (AUPRC), measures how strongly the highest-attention pixels concentrate within the ROI. The essential difference from AUROC lies in the baseline. AUROC compares in-ROI against out-of-ROI pixels symmetrically, and its no-skill value is fixed at 0.5 regardless of how large the mask is. In its turn, precision does not include true negatives, so an uninformative map yields an AUPRC equal to the prevalence of the positive class (the fraction of pixels inside the ROI). Because the masks in our dataset vary considerably in size, raw AUPRC values are not comparable across images. Thus, we use AUPRC divided by the ROI fraction π, which expresses performance as follows: a value of 1 denotes no localization ability, values above 1 indicate attention concentrated within the mask beyond chance expectation, and the maximum attainable value is the reciprocal of the mask fraction $1/\pi$. This normalization makes AUPRC comparable across differently sized regions. Since precision doesn't include a large number of easily-classifiable background pixels normalized AUPRC is more sensitive than AUROC to whether the strongest attention actually falls inside the ROI. Thus, whenever we refer to AUPRC below we imply that it is divided by the ROI fraction π.

Beyond reporting AUROC and AUPRC as summary measures of localization, we use them to characterize how the threshold of p- and l-regularization methods affects performance: for each value of this threshold we produce a distinct regularized attention map and compute its AUROC and AUPRC, tracing how localization quality varies as the threshold is swept. It is important to distinguish this regularization threshold from the decision threshold that is internal to the construction of the ROC and precision–recall curves themselves. The latter is swept implicitly over the attention values of a single fixed map. The former is a hyperparameter of the hard-thresholding regularization method.

B BOOTSTRAP DISTRIBUTIONS

This appendix illustrates the bootstrap procedure underlying our regularization framework and examines the stability of the resulting null distributions. Figure S1 shows the construction of the bootstrap samples: starting from images of the LIDC-IDRI dataset (top row), we generate bootstrap images by pixel resampling with replacement (middle row) and pass them through the Vanilla ViT to obtain the corresponding bootstrap attention maps (bottom row). These maps constitute the null attention against which the observed attention is compared, providing the reference distribution from which the attention uncertainty and posterior null probabilities are computed.

Figure S2 characterizes the variability of this null distribution. The top row compares, for each example image, the z-score histograms obtained from individual bootstrap samples B_i (coloured) against the distribution obtained by combining all twenty samples (black contour), showing how the individual bootstrap draws scatter around the aggregated distribution. The bottom row compares the per-image combined bootstrap ($B = 20$, black contour) with a joint bootstrap sample pooled across images (grey histogram). The close agreement between the per-image and pooled distributions supports the use of a shared, universal bootstrap sample in place of recomputing a fresh bootstrap for every image.

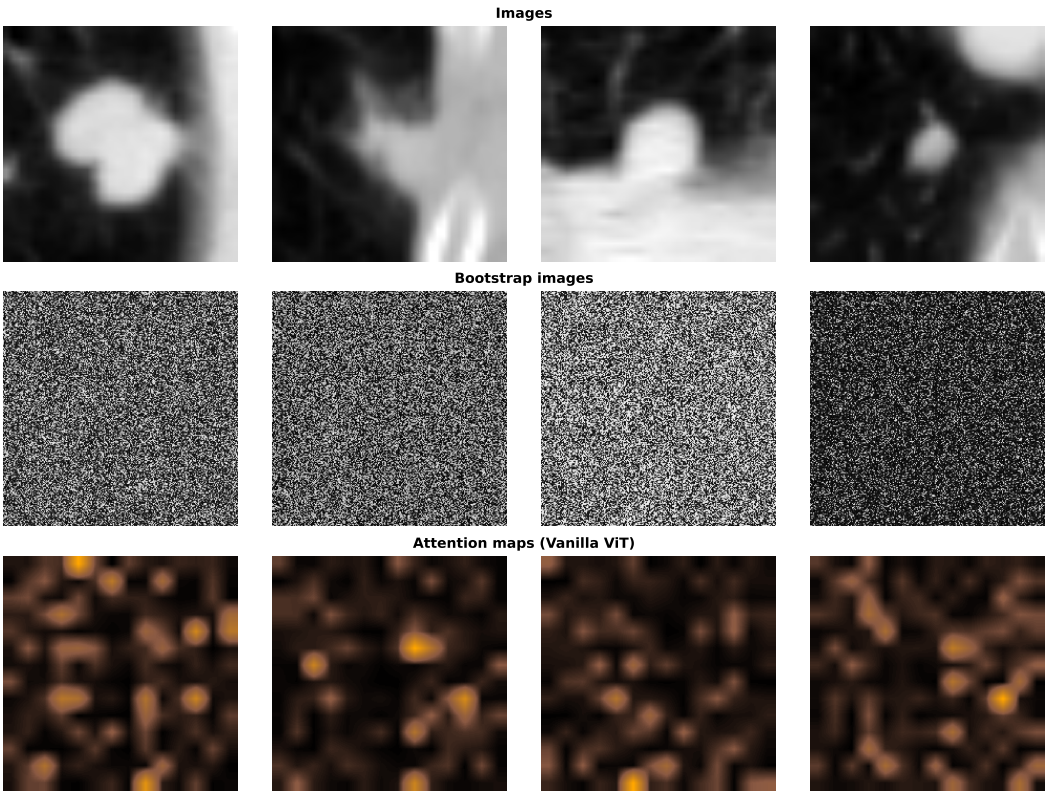

Figure S1: Examples of images from the LIDC-IDRI dataset (top row), bootstrap images generated from them using pixel resampling with replacement (middle row) and the corresponding bootstrap attention maps obtained with Vanilla ViT (bottom row).

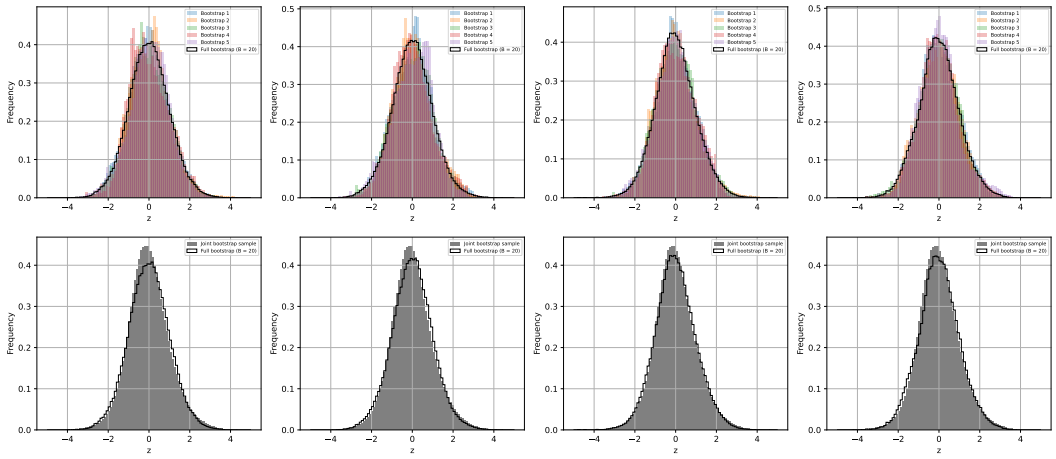

Figure S2: Top row: histograms of z-scores for different bootstrap samples obtained from the images in Figure S1 (in the same order) with the Vanilla ViT. Colored histograms correspond to different single bootstrap samples B_i, while the black contour shows the bootstrap distribution for the combined bootstrap ($\sum_i^{20} B_i$). Bottom row: comparison between the individual bootstrap for each image (black contour, $B = 20$) and the joint bootstrap sample (grey histogram).

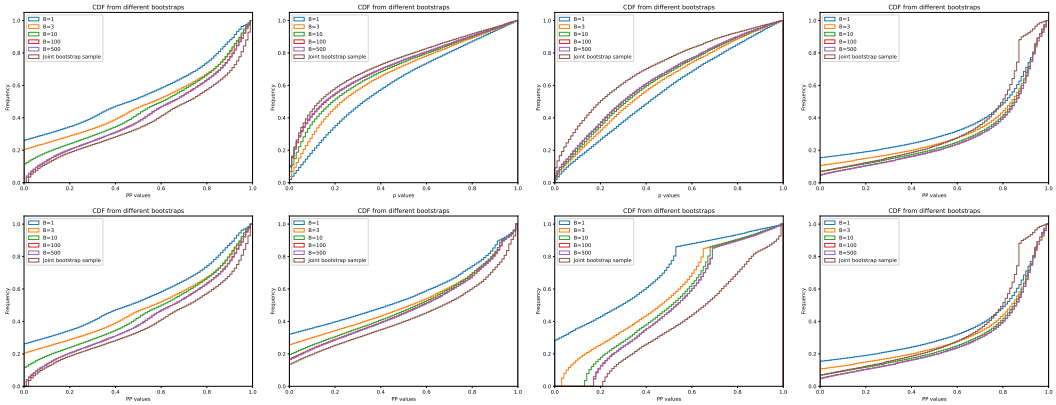

Figure S3: Cumulative distribution functions (CDF) of p-values (top row) and posterior probability (bottom row) for the images in Figure S1 (in the same order) for different bootstrap samples. Blue, orange, green, red and magenta curves correspond to different B numbers for each image: 1, 3, 10, 100 and 500 respectively. Brown curve corresponds to the CDF from the universal bootstrap sample.

C Comparison between joint and individual bootstrapping

Having shown in Appendix B that the per-image and pooled null distributions agree closely, we here quantify the effect of this simplification on the downstream metrics. Figure S4 shows, as violin plots, the relative difference between the metric values obtained using an individual bootstrap sample for each image (Y_I, $B = 20$) and those obtained using a single joint bootstrap sample of the same size (Y_U) from Vanilla ViT across the LIDC-IDRI dataset. The comparison spans three metrics (TP50, energy ratio, and AGM contrast) for three regularization methods (pre-regularization with $z > 0$, PPS, and Wiener shrinkage), with the mean and standard deviation of each distribution marked by the red point and error bar. Overally, the relative differences are concentrated near zero across all metrics and methods, confirming that replacing the per-image bootstrap with a single joint bootstrap sample leaves the regularization results essentially unchanged.

The dispersion of the relative differences in Figure S4 varies across methods and metrics, and reflects how sensitively each regularization method depends on the fine details of the bootstrap null. The

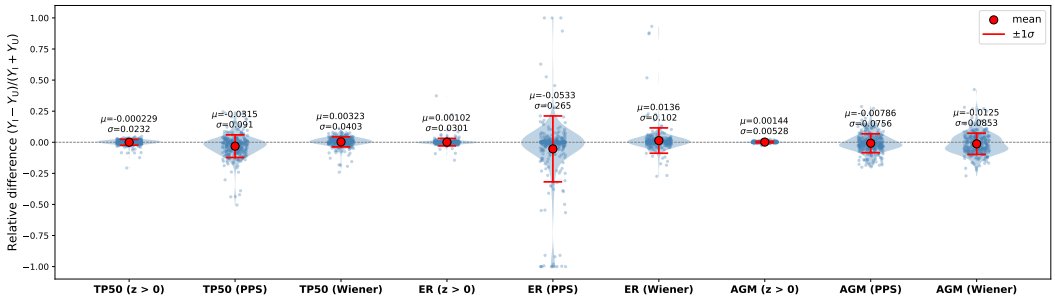

Figure S4: Violin plot for the relative difference between the values of various metrics obtained with using the individual bootstrap samples for each image from the dataset (Y_I, $B = 20$) vs. the values obtained with the joint bootstrap sample of the same size (Y_U) for Vanilla ViT and LIDC-IDRI dataset. We demonstrate TP50, energy ratio and AGM contrast for three regularization methods: pre-regularization ($z > 0$), PPS ans Wiener shrinkage. The mean value and the standard deviation are shown with the red circle and an error bar.

smallest dispersion is observed for pre-regularization ($z > 0$), which is expected: this method depends only on the *sign* of the z-score, which is the most robust quantity provided by the bootstrap. In contrast, PPS shows a large dispersion in the ER, because it attenuates each pixel by its posterior null probability, a continuous quantity computed from the full shape of the null distribution rather than merely its sign; small differences between the individual and joint bootstrap therefore propagate into the per-pixel weights and into a comparatively large spread. The relatively large dispersion of the TP50 for PPS, ER for Wiener shrinkage and that of the AGM contrast for both PPS and Wiener shrinkage have a similar origin.

D DINOv2 ViT

D.1 Examples of regularized images

D.1.1 DUTS

To complement the single example shown in the main text (Figure 1), we present additional representative cases of attention-map regularization for the DINOv2 ViT on DUTS. Together these examples illustrate the qualitative behavior of the different methods across images with varying attention quality and ROI characteristics. Note that no fine-tuning was performed on this dataset.

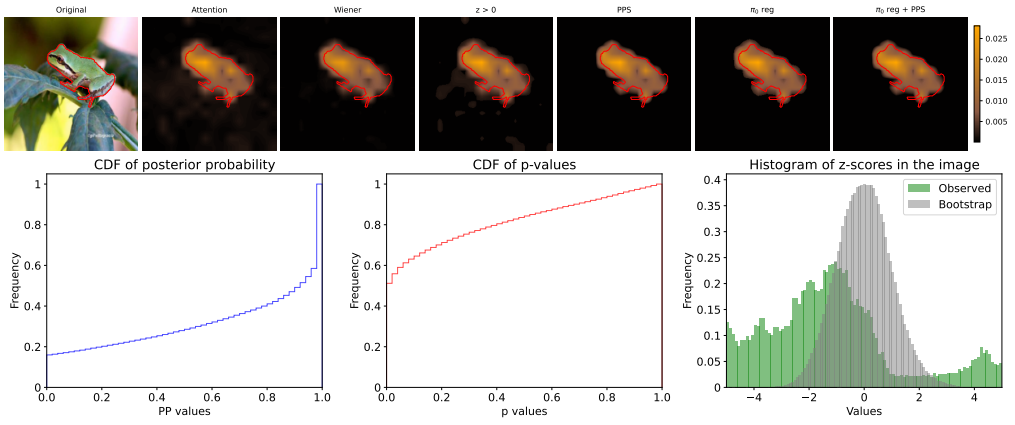

Figure S5: Example of regularization via different methods for the attention map derived from the DINOv2 ViT for image `ILSVRC2012_test_00003634`. The top row contains 7 images: original image, attention map, Wiener shrinkage, pre-regularization, PPS, π_0-thresholding and PPS$_{\pi_0}$. The red contour shows the annotated mask. The bottom row contains three statistics plots: CDF of posterior probabilities (PP, blue), CDF of p-values (red) and the histograms of the observed z-scores (green) and the bootstrap scores (grey).

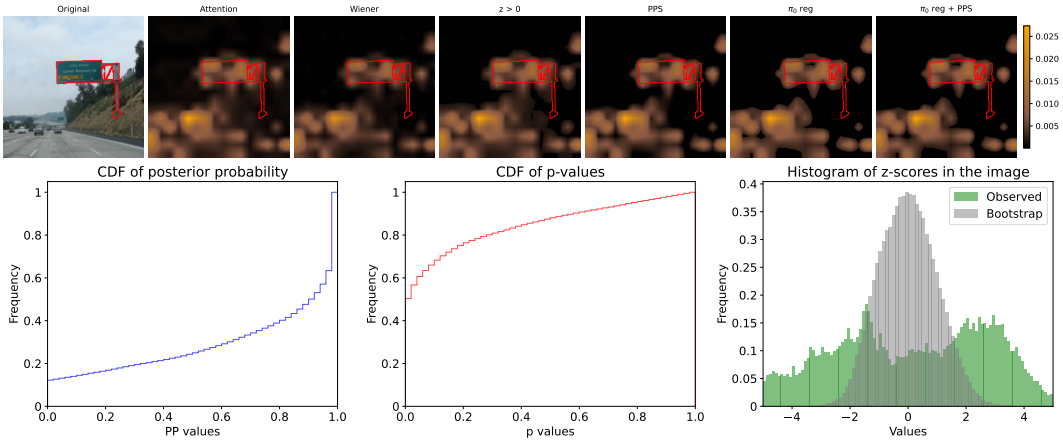

Figure S6: Example of regularization via different methods for the attention map derived from the DINOv2 ViT for image `sun_boeqjcuhmxrqmevn`. Same panel layout as in Figure S5.

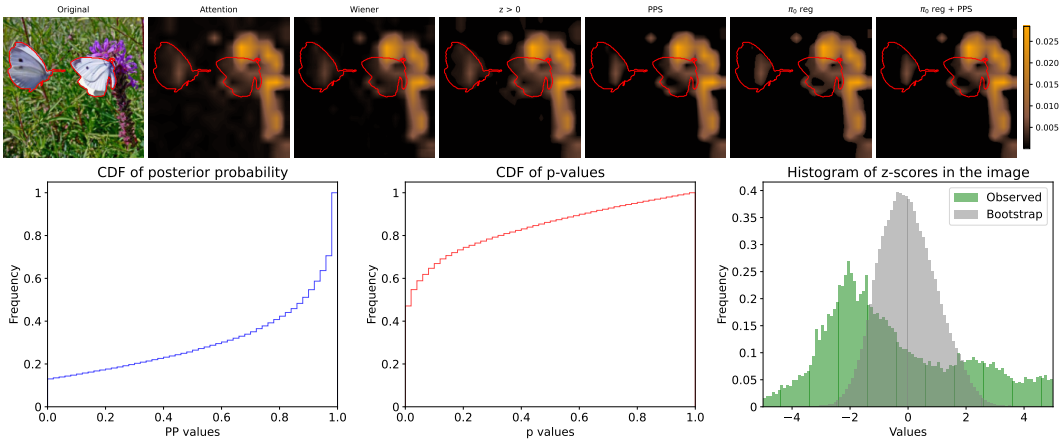

Figure S7: Example of regularization via different methods for the attention map derived from the DINOv2 ViT for image `ILSVRC2012_test_00087929`. Same panel layout as in Figure S5.

D.1.2 LIDC-IDRI

In Figs. S8-S11 we show the examples of regularization by different methods on the images drawn from distinct scans of the LIDC-IDRI dataset and processes by DINOv2 ViT. As we discuss in the main text, the raw DINOv2 attention maps for LIDC-IDRI are already substantially cleaner and better localized than those on DUTS, so the regularization methods have comparatively little noise to remove and their effect on the maps is correspondingly subtle – consistent with the marginal quantitative differences reported for DINOv2 in Figure 3.

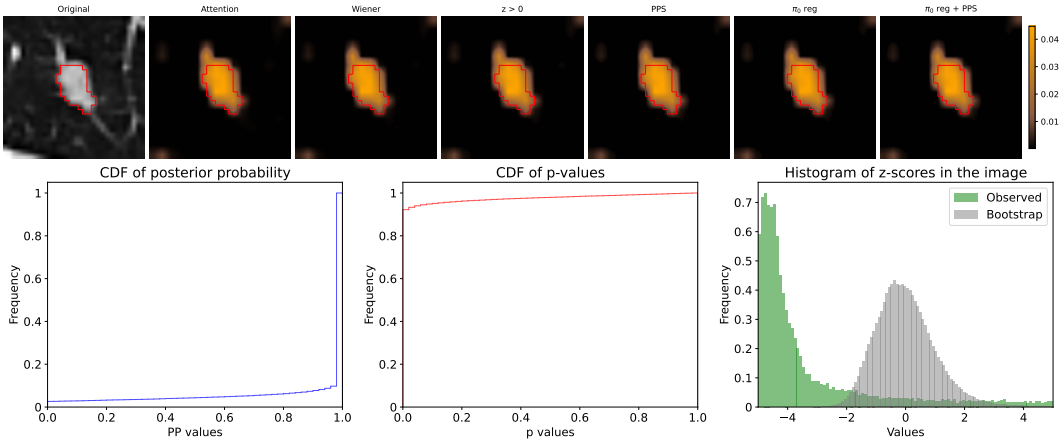

Figure S8: Example of regularization via different methods for the attention map derived from the DINOv2 ViT for image `LIDC-IDRI-0195_2`. Same panel layout as in Figure S5.

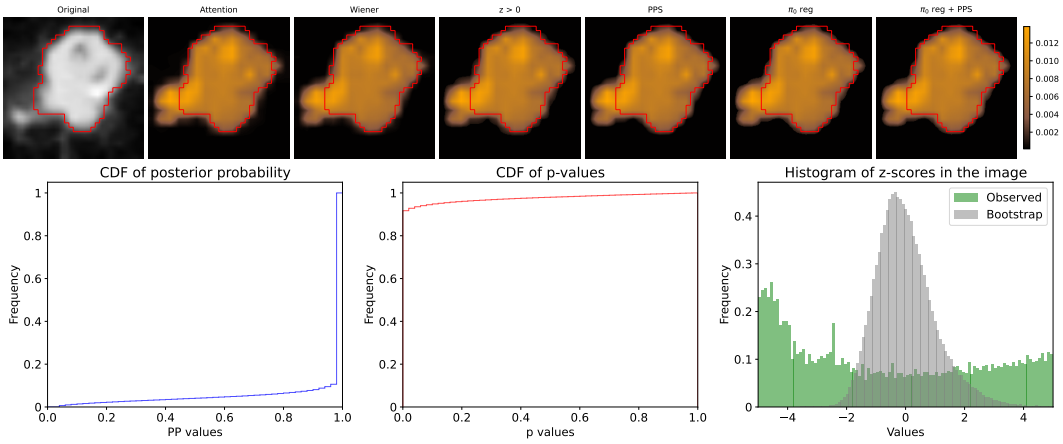

Figure S9: Example of regularization via different methods for the attention map derived from the DINOv2 ViT for image `LIDC-IDRI-0003_2`. Same panel layout as in Figure S9.

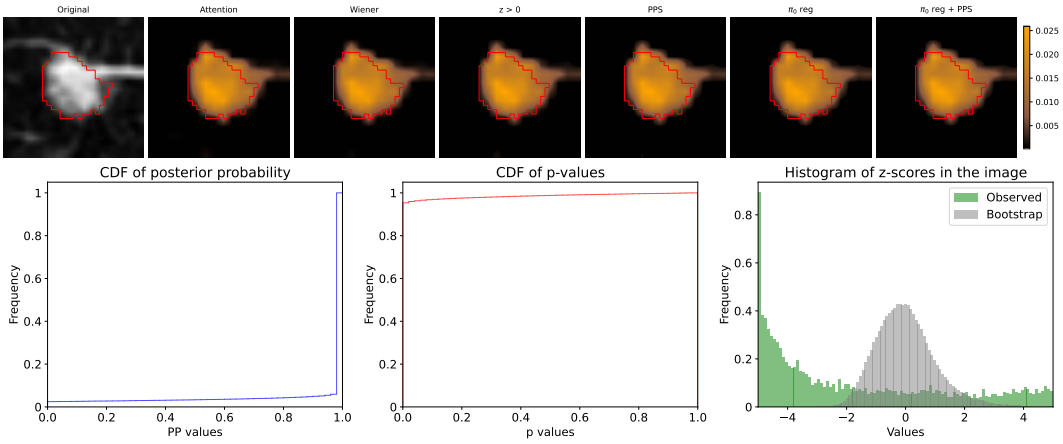

Figure S10: Example of regularization via different methods for the attention map derived from the DINOv2 ViT for image `LIDC-IDRI-0965_1`. Same panel layout as in Figure S5.

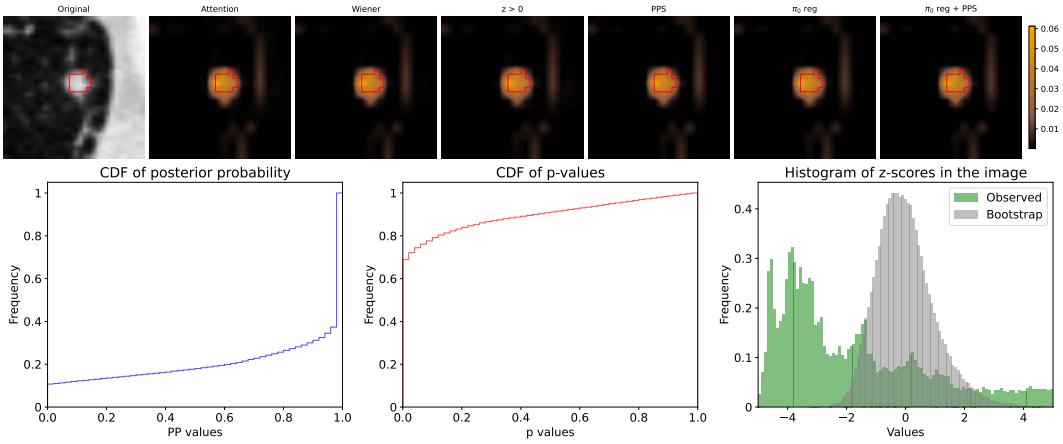

Figure S11: Example of regularization via different methods for the attention map derived from the DINOv2 ViT for image `LIDC-IDRI-0454_1`. Same panel layout as in Figure S5.

D.2 AUROC AND AUPRC

For general p- and l-regularization, we sweep across threshold values and report the resulting AUROC and AUPRC curves. Figure S12 reports the localization performance of the hard-thresholding methods on DINOv2 attention maps produced from the DUTS dataset, showing AUROC (left column) and AUPRC (right column) as a function of the regularization threshold for p-regularization (top row) and l-regularization (bottom row). The faint curves trace individual images and the thick black curve their median, while the horizontal dashed lines mark the median performance of the methods that either do not depend on thresholds (posterior weighting) or the threshold selection is automatized (combination of PPS with π_0) and the green dashed curve marks the median prior to any regularization.Note that the black curve position at the weakest threshold value on the right corresponds to the median value achieved via pre-regularization ($z > 0$). The role of these figures is twofold: to demonstrate how sensitive the hard-thresholding methods are to the value of the threshold and compare them alongside with the posterior weighting in terms of localization. Regularization based on LFDR threshold has a sharper dependence on the threshold, as the distribution of LFDR values is typically less skewed towards 0 as the distribution of p-values.

Figure S14 shows the distributions of AUROC and AUPRC ratios – the value after regularization divided by the value prior to the regularization (a ratio above unity indicates improved localization) – for the threshold-independent or fixed-threshold methods across the same pool of results that is presented in Figure S12. The per-method medians and variances reported in the legends quantify both the typical effect and its spread across images. Analyzing the median curves and values we see that on average hard-thresholding methods and PPS can hardly improve the localization measured by AUPRC and especially by AUROC. Whereas the ER rewards our regularization methods for concentrating attention mass within the ROI, the AUROC and AUPRC metrics tell a different story, because they depend on the ranking of pixel attention values rather than the magnitude of attention and are invariant to any order-preserving transformations. The hard-thresholding methods do not cause any reordering of pixels in the ranking, but only equalize low-rank pixels along the bottom of the ranking. When ROI pixels get nullified significantly it makes the precision decrease faster with the growth of false positive rate (for ROC) or recall (for PRC), hence leading to lower values of AUROC and AUPRC. PPS also acts largely along the existing attention ordering – it reweights pixels by their posterior probability, which typically yields smoother attention distributions w.r.t. hard-thresholding methods, and also leave more ROI pixels above zero, but does not substantially re-rank the attention. As a result, these methods improve the magnitude-based ER, but leave the ranking nearly unchanged, which decreases AUROC and AUPRC on the attention maps with large portions of highly-ranked background artifacts. Wiener shrinkage is the exception: by filtering against the spectrum of null attention inhomogeneities (that reflect not only the pixel intensity, but their spatial distribution), it can genuinely re-rank incoherent background artifacts below the region of interest, yielding the only consistent improvement in AUPRC. The improvement in AUROC, however, remains marginal even for the Wiener shrinkage, which is the consequence of the average

small size of the annotated masks. Under such strong foreground-background imbalance, the false-positive rate is diluted by the large pool of easily classified background, leaving AUROC inflated and largely unresponsive at baseline. In its turn the precision that does not involve true negatives is sensitive to false positive directly, so what improvement exists is concentrated in AUPRC rather than AUROC.

Figures S13 and S15 present the same results as Figures S12 and S14 respectively, but for the DINOv2 on LIDC-IDRI dataset. Quantitatively similar conclusions hold for these results, although the attention maps produced by the fine-tuned DINOv2 on this dataset are much cleaner than the ones in the experiments on DUTS, hence the p-values and LFDR values are concentrated very close to zero and the AUROC/AUPRC curves are almost flat and the improvement of these metric values for threshold-independent (or fixed-threshold) methods is very moderate.

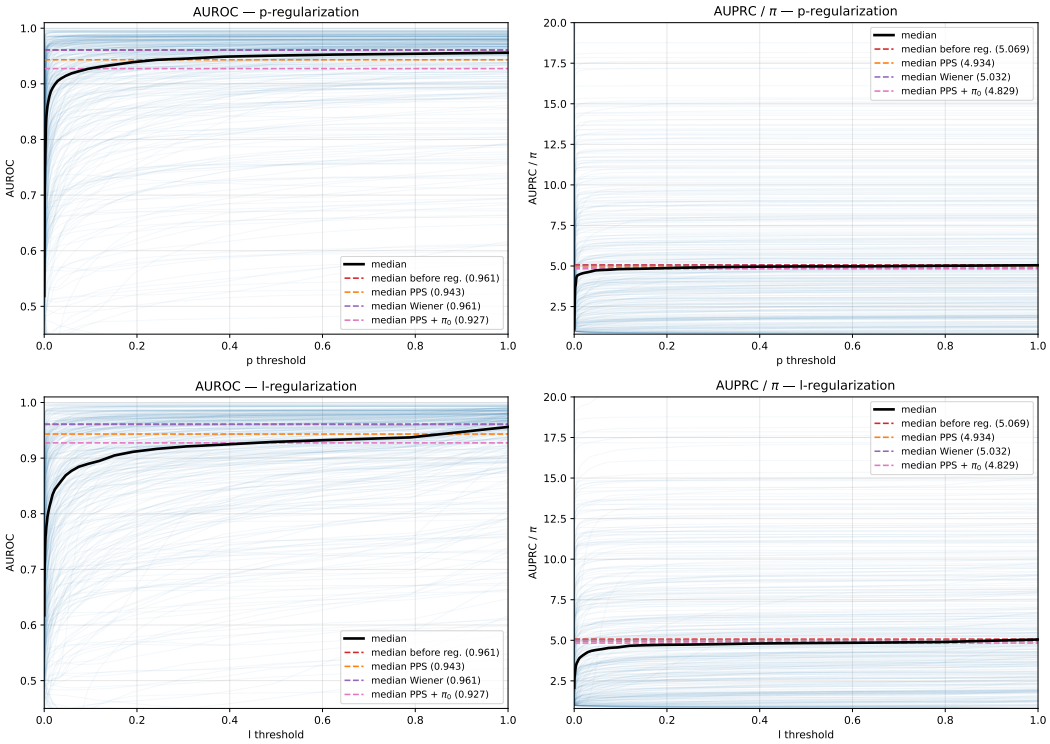

Figure S12: AUROC (*left column*) and AUPRC (*right column*) dependence on the regularization threshold for p-regularization (*top row*) and l-regularization (*bottom row*) of attention maps obtained with the DINOv2 ViT on DUTS. The faint blue curves correspond to the results for different images from the dataset, while the thick black line denotes a median curve. Colored dashed horizontal lines show the median results for 4 threshold-independent regularization methods: PPS (orange), Wiener shrinkage (purple) and PPS_{π_0} (pink). The red dashed curve show the median value prior to regularization. Note that the value of the median curve at the maximal threshold corresponds to the median value of z-thresholding pre-regularization ($z > 0$). The baseline value for AUROC measurements is 0.5, while for AUPRC it is 1. The limits of y-axis in all the plots are adopted to contain the 90% of the curves.

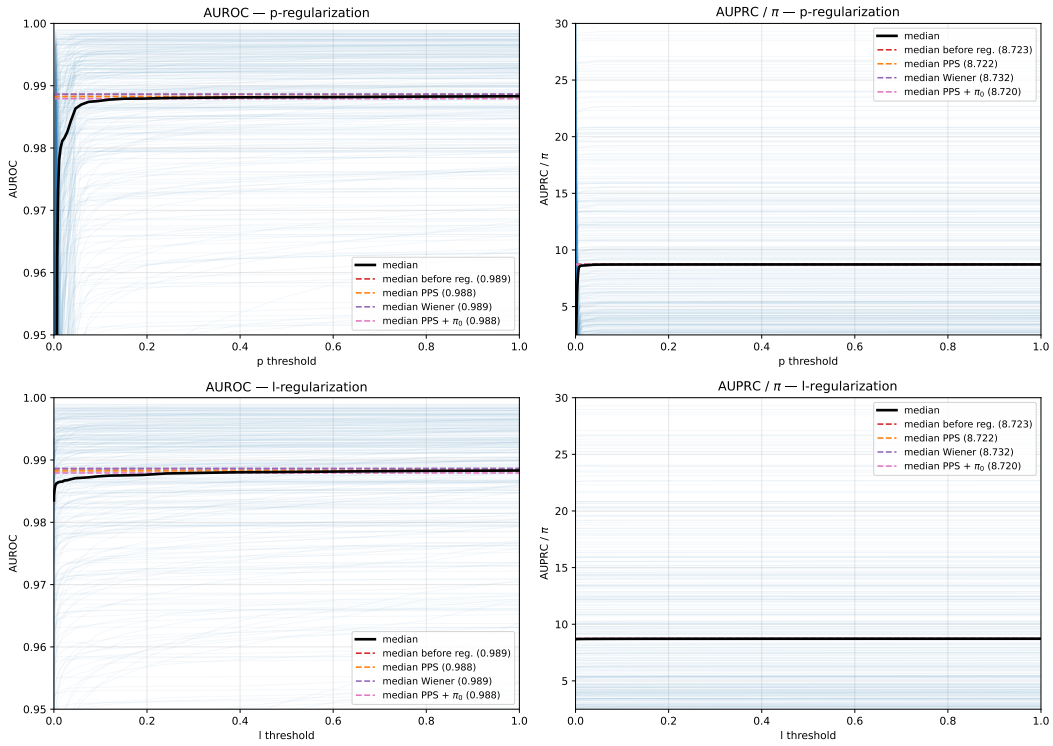

Figure S13: AUROC (*left column*) and AUPRC (*right column*) dependence on the regularization threshold for p-regularization (*top row*) and l-regularization (*bottom row*) of attention maps obtained with the DINOv2 ViT on LIDC-IDRI. The panel arrangement and graphical conventions follow those of Figure S12.

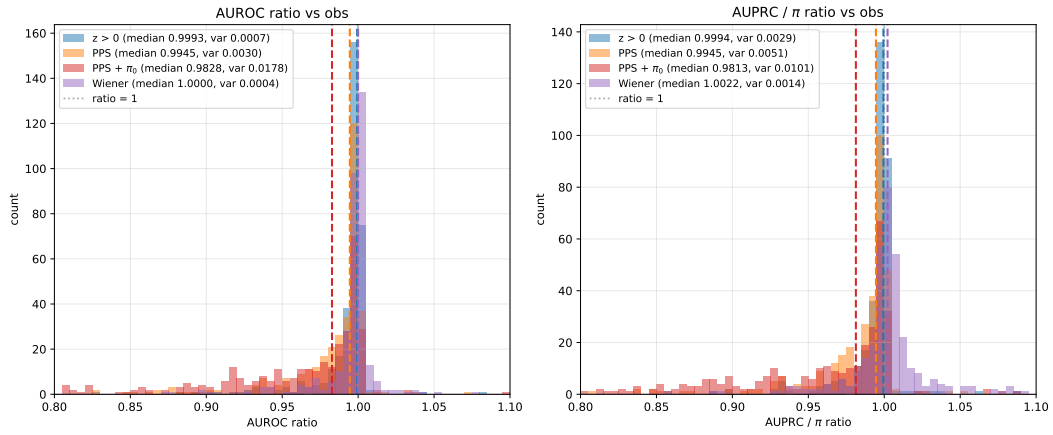

Figure S14: Histograms of AUROC (left) and AUPRC (right) ratios (after regularization vs. before regularization) for different threshold-independent or fixed-threshold methods presented in Figure S12 (DINOv2, DUTS). Dashed colored vertical lines represent median values for the distribution of values for each method (of the same color). The median values and variance are reported in the plot legend for each method.

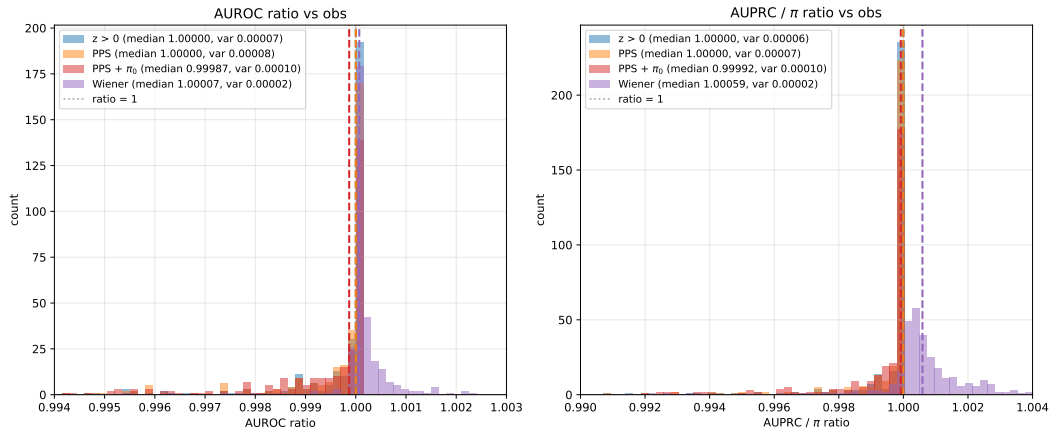

Figure S15: Histograms of AUROC (left) and AUPRC (right) ratios (after regularization vs. before regularization) for different threshold-independent or fixed-threshold methods presented in Figure S13 (DINOv2, LIDC-IDRI). The panel arrangement and graphical conventions follow those of Figure S14.

D.3 NON-ZERO PIXELS

To complement Figure 2 in Figure S16 we show the distributions of percentages of non-zero pixels in the attention maps produced by DINOv2 ViT from DUTS images (left plot) and from LIDC-IDRI images (right plot) in ROI (x-axis) and outside of ROI (y-axis). In these plots we do not show the π_0-thresholding regularization as it essentially coincide with the PPS$_{\pi_0}$ shrinkage results.

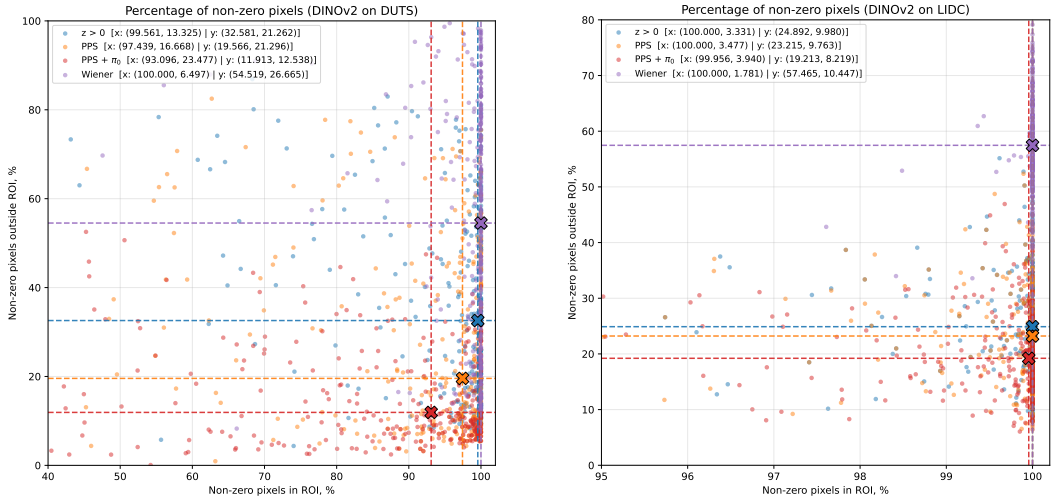

Figure S16: The percentage of non-zero pixels after regularization via different methods in ROI (x-axis) vs. outside of ROI (y-axis) for DINOv2 ViT on DUTS (left) and LIDC-IDRI (right). Each dot corresponds to the results for one image. The median and variance for each distribution dimension are reported next to each legend entry: [x: (median, var), . . .]. The intersection of the corresponding median values for each method are shown with the cross signs of the respective color.

E VANILLA VIT

E.1 EXAMPLES OF REGULARIZED IMAGES

Analogously to Appendix D.1 here we present the examples of regularization of the same images from DUTS dataset (Figs. S17-S20) and LIDC-IDRI dataset (Figs S21-S24) by the Vanilla ViT.

E.1.1 DUTS

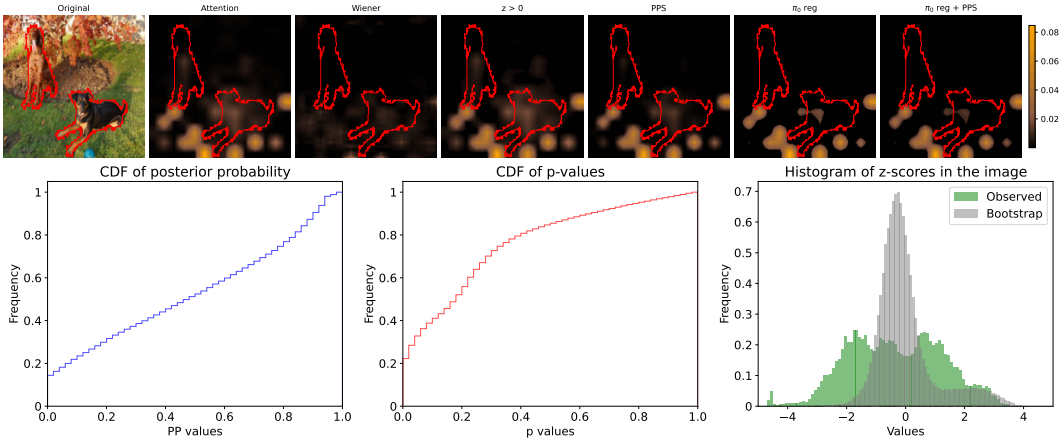

Figure S17: Example of regularization via different methods for the attention map derived from the Vanilla ViT for image `ILSVRC2012_test_00059189` (as in Fig. 1). Same panel layout as in Figure S5.

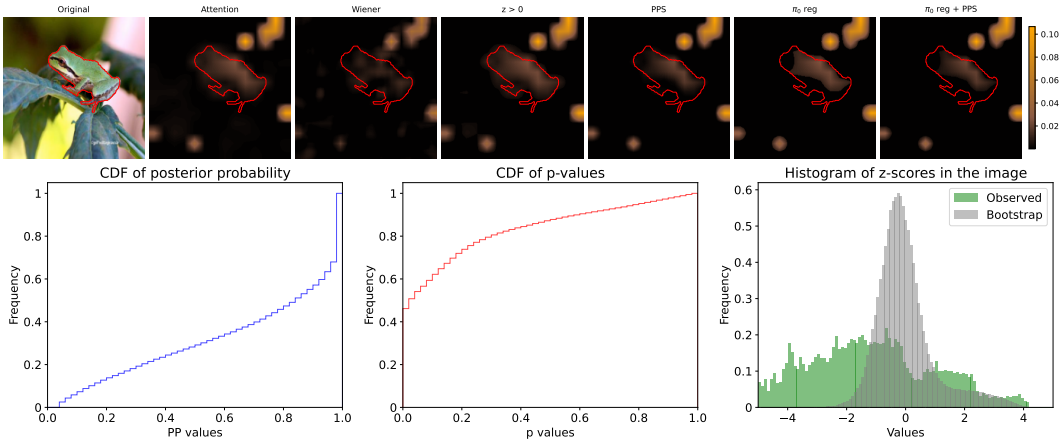

Figure S18: Example of regularization via different methods for the attention map derived from the Vanilla ViT for image `ILSVRC2012_test_00003634`. Same panel layout as in Figure S5.

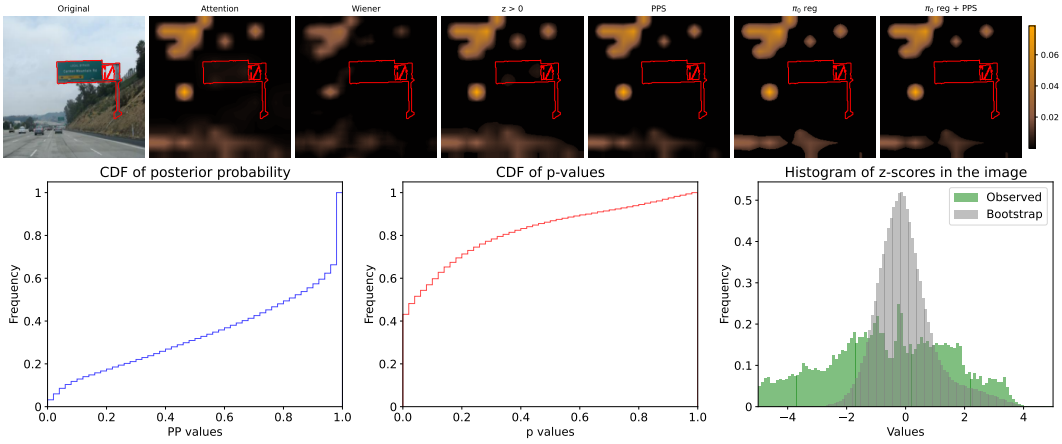

Figure S19: Example of regularization via different methods for the attention map derived from the Vanilla ViT for image `sun_boeqjcuhmxrqmevn`. Same panel layout as in Figure S5.

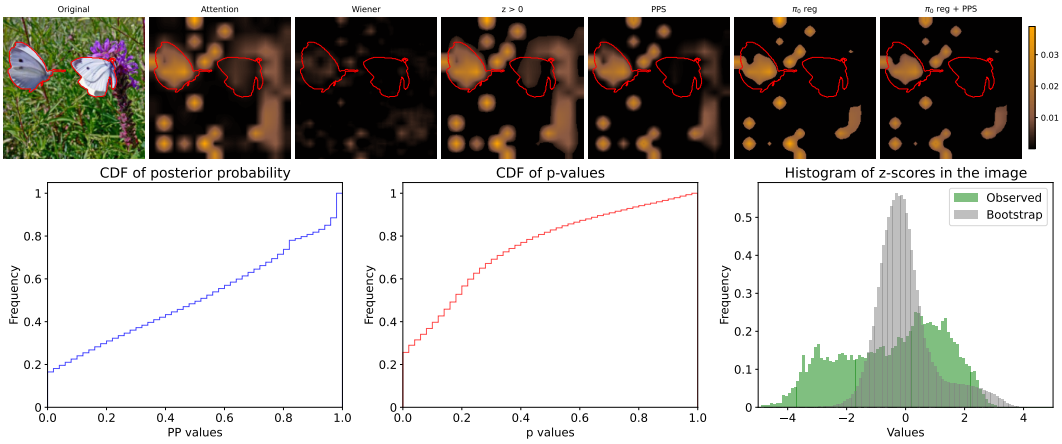

Figure S20: Example of regularization via different methods for the attention map derived from the Vanilla ViT for image `ILSVRC2012_test_00087929`. Same panel layout as in Figure S5.

E.1.2 LIDC-IDRI

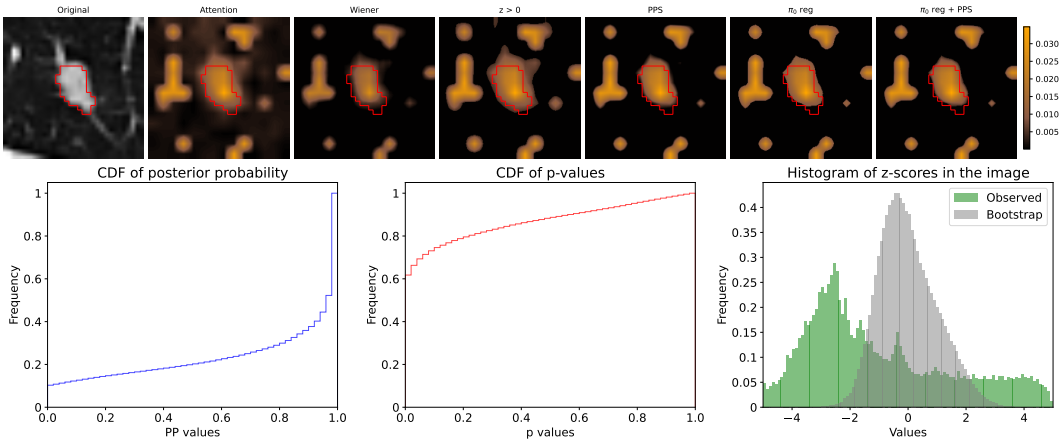

Figure S21: Example of regularization via different methods for the attention map derived from the Vanilla ViT for image `LIDC-IDRI-0195_2`. Same panel layout as in Figure S18.

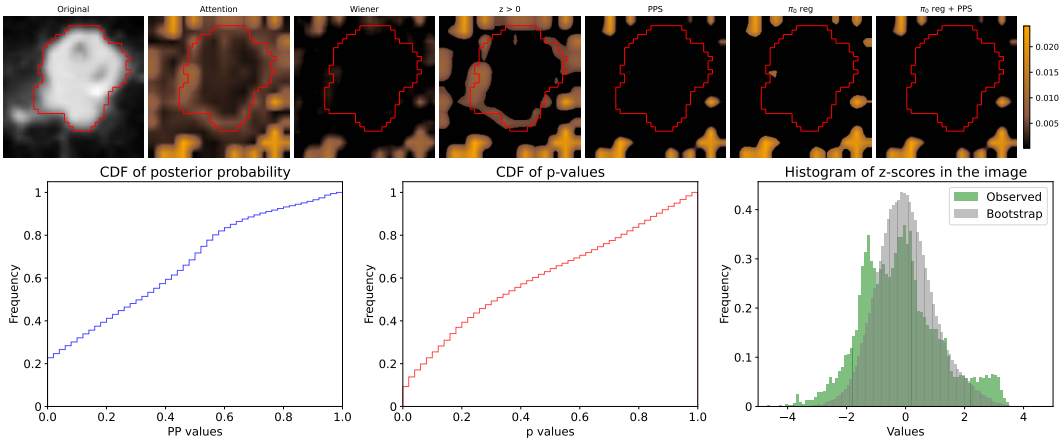

Figure S22: Example of regularization via different methods for the attention map derived from the Vanilla ViT for image `LIDC-IDRI-0003_2`. Same panel layout as in Figure S22.

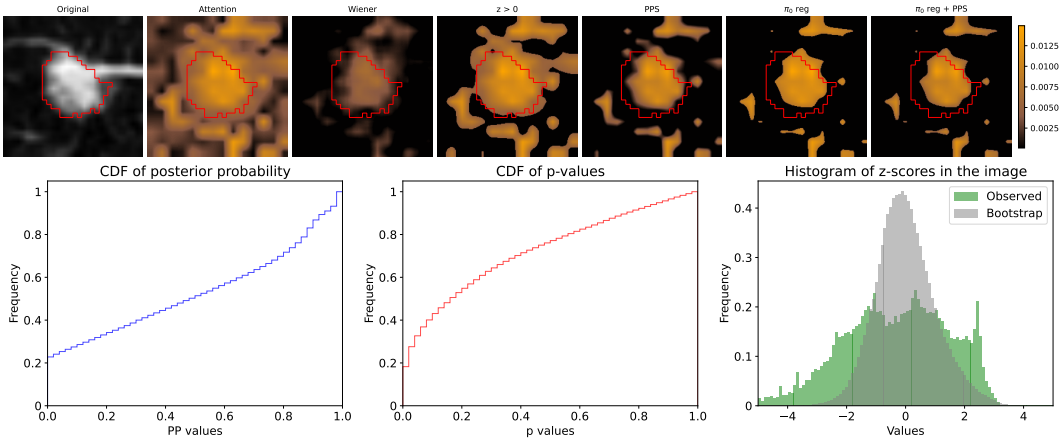

Figure S23: Example of regularization via different methods for the attention map derived from the Vanilla ViT for image `LIDC-IDRI-0965_1`. Same panel layout as in Figure S18.

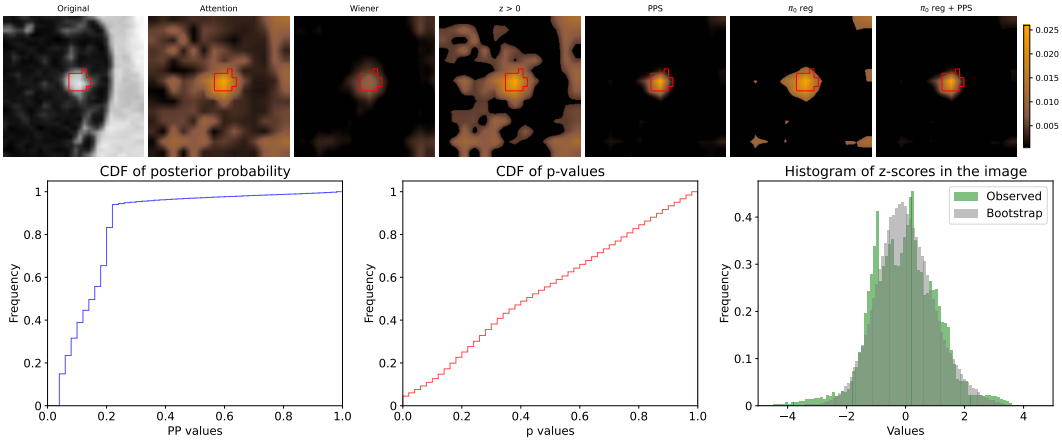

Figure S24: Example of regularization via different methods for the attention map derived from the Vanilla ViT for image `LIDC-IDRI-0454_1`. Same panel layout as in Figure S18.

E.2 OTHER METRIC DISTRIBUTIONS

In this section we demonstrate the metric evaluation results for various regularization methods computed by Vanilla ViT on DUTS (left column) and LIDC-IDRI (right column). Figure S25 shows the TP50 improvement ratio vs. the total non-zero pixel fraction (analogously to Figure 2), Figure S26 shows the ER ratio improvement vs. the AGM contrast improvement (analogously to Figure 3) and Figure S27 shows the percentages of non-zero pixels inside and outside of ROI (analogously to Figure S16).

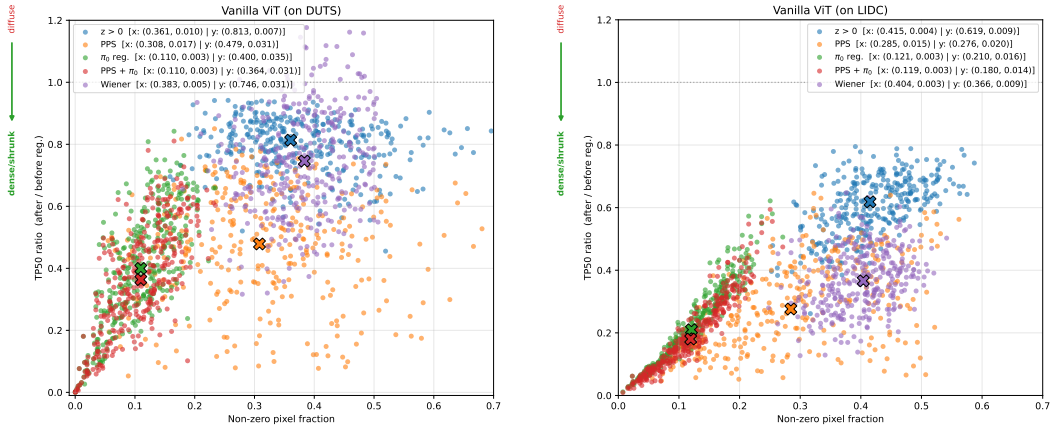

Figure S25: The ratio of TP50 metric after and before regularization (y-axis) vs. the fraction of non-zero attention scores (x-axis) for Vanilla ViT on DUTS (left) and LIDC-IDRI (right) for different methods: z-thresholding (blue), PPS (orange), π_0-thresholding (green), PPS_{π_0} (red) and Wiener (purple). Each dot corresponds to the results for one image. The median and variance for each distribution dimension are reported next to each legend entry: [x: (median, var), . . .]. The intersection of the corresponding median values for each method are shown with the cross signs of the respective color.

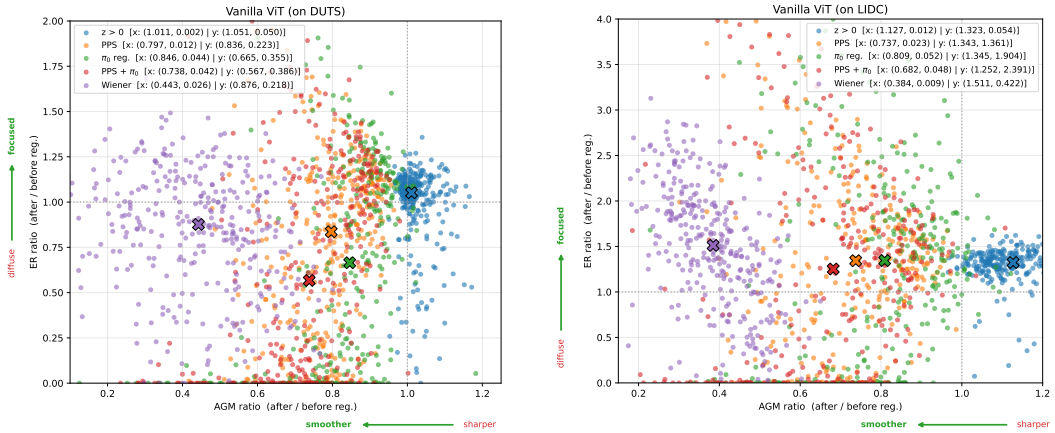

Figure S26: The ratios of two metric values after and before regularization: ER ratio (y-axis) vs. AGM contrast (x-axis) for Vanilla ViT on DUTS (left) and LIDC-IDRI (right) for different methods: z-thresholding (blue), PPS (orange), π_0-thresholding (green), PPS_{π_0} (red) and Wiener (purple). The panel arrangement and graphical conventions follow those of Figure S25.

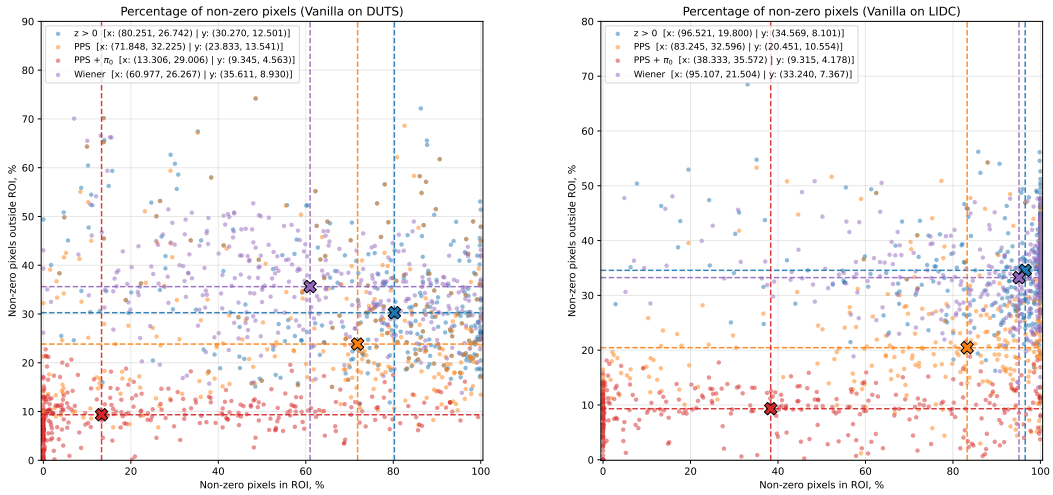

Figure S27: The percentage of non-zero pixels after regularization via different methods in ROI (x-axis) vs. outside of ROI (y-axis) for Vanilla ViT on DUTS (left) and LIDC-IDRI (right). The panel arrangement and graphical conventions follow those of Figure S16.

F DeiT ViT

F.1 Examples of regularized images

Analogously to Appendix D.1 here we present the examples of regularization of the same images from DUTS dataset (Figs. S28-S31) and LIDC-IDRI dataset (Figs S32-S35) by the DeiT ViT.

F.1.1 DUTS

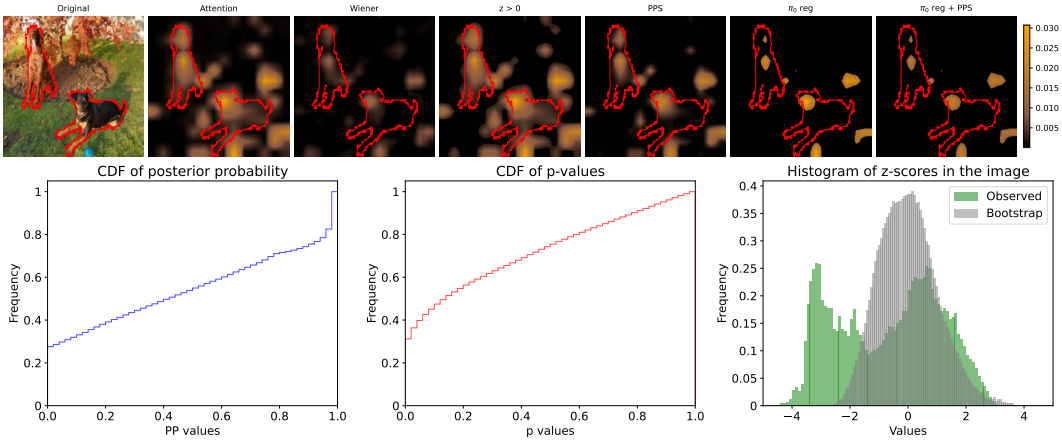

Figure S28: Example of regularization via different methods for the attention map derived from the DeiT ViT for image `ILSVRC2012_test_00059189` (as in Figure 1). Same panel layout as in Figure S5.

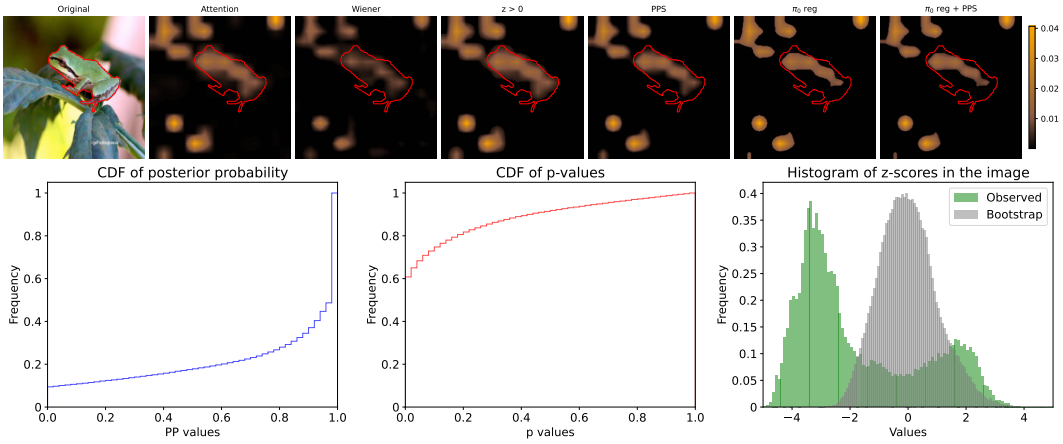

Figure S29: Example of regularization via different methods for the attention map derived from the DeiT ViT for image `ILSVRC2012_test_00003634`. Same panel layout as in Figure S5.

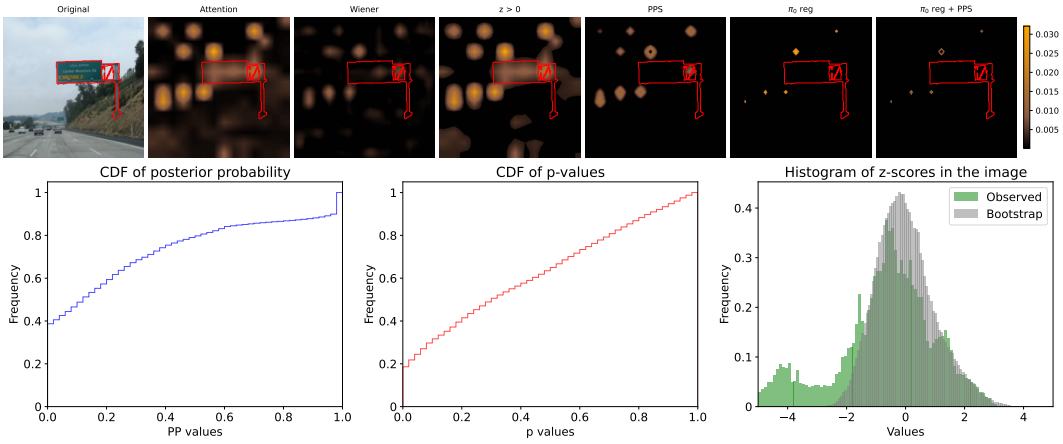

Figure S30: Example of regularization via different methods for the attention map derived from the DeiT ViT for image `sun_boeqjcuhmxrqmevn`. Same panel layout as in Figure S5.

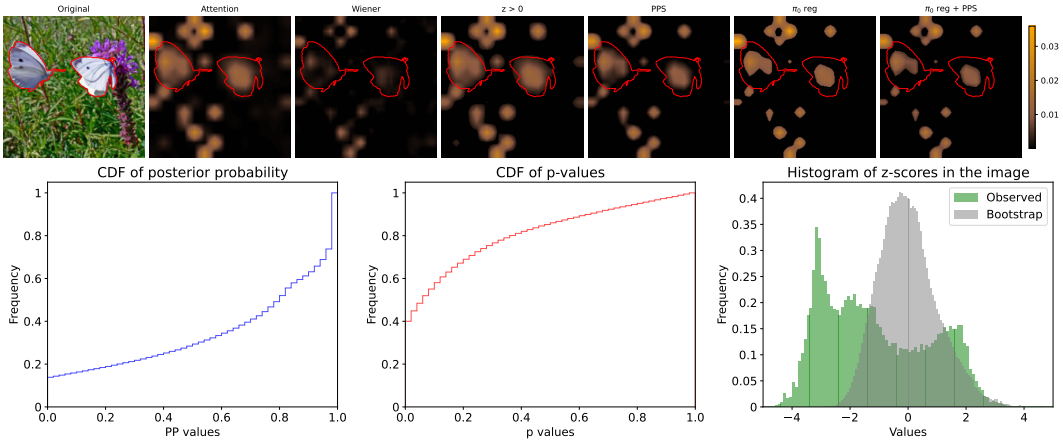

Figure S31: Example of regularization via different methods for the attention map derived from the DeiT ViT for image `ILSVRC2012_test_00087929`. Same panel layout as in Figure S5.

F.1.2 LIDC-IDRI

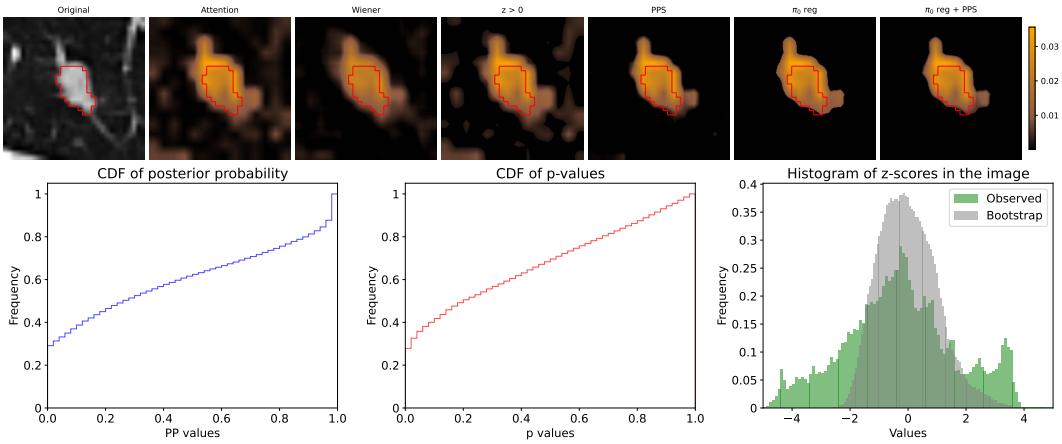

Figure S32: Example of regularization via different methods for the attention map derived from the DeiT ViT for image `LIDC-IDRI-0195_2`. Same panel layout as in Figure S29.

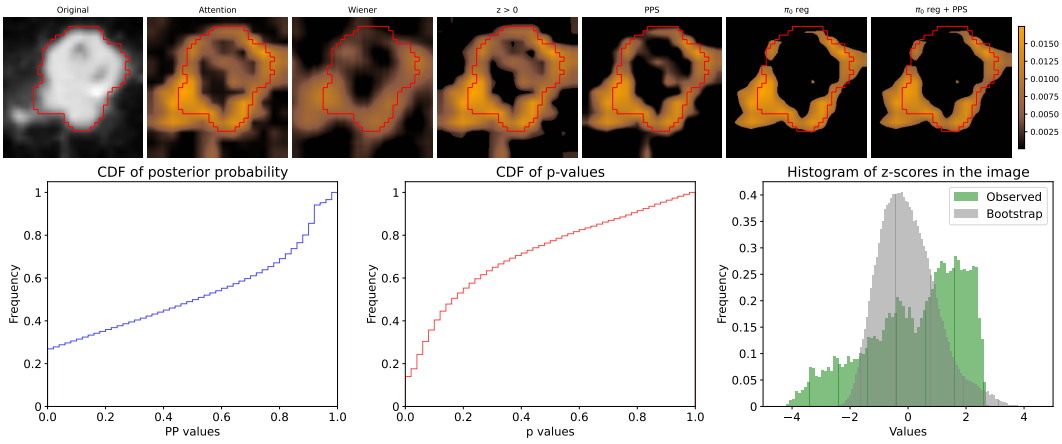

Figure S33: Example of regularization via different methods for the attention map derived from the DeiT ViT for image `LIDC-IDRI-0003_2`. Same panel layout as in Figure S33.

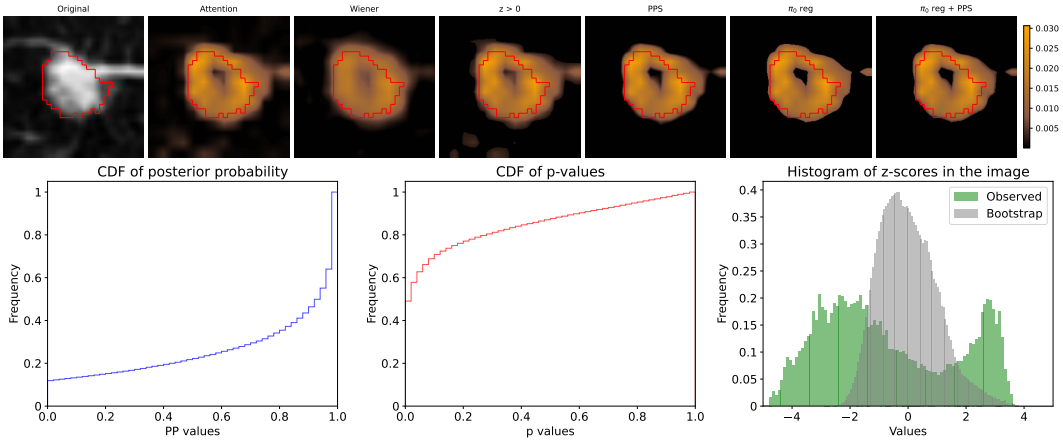

Figure S34: Example of regularization via different methods for the attention map derived from the DeiT ViT for image `LIDC-IDRI-0965_1`. Same panel layout as in Figure S29.

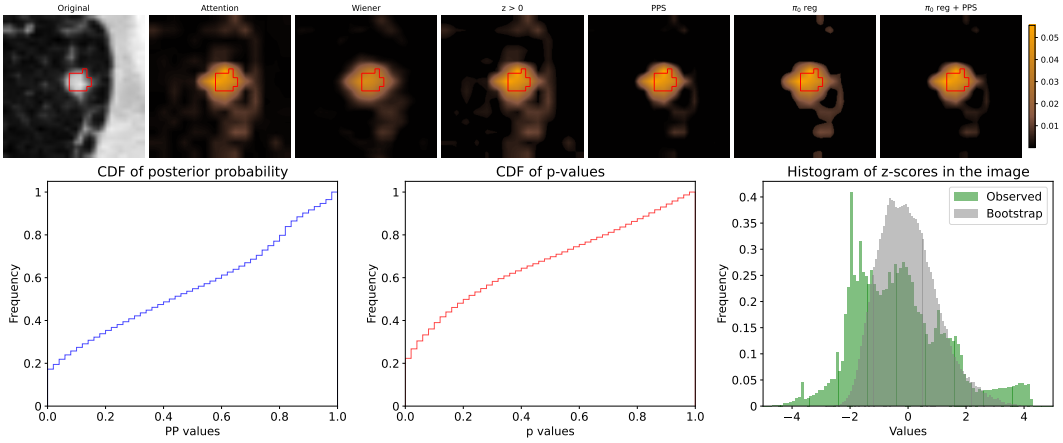

Figure S35: Example of regularization via different methods for the attention map derived from the DeiT ViT for image `LIDC-IDRI-0454_1`. Same panel layout as in Figure S29.

F.2 OTHER METRIC DISTRIBUTIONS

In this section we demonstrate the metric evaluation results for various regularization methods computed by DeiT ViT on DUTS (left column) and LIDC-IDRI (right column). Figure S36 shows the TP50 improvement ratio vs. the total non-zero pixel fraction (analogously to Figure 2), Figure S37 shows the ER ratio improvement vs. the AGM contrast improvement (analogously to Figure 3) and Figure S38 shows the percentages of non-zero pixels inside and outside of ROI (analogously to Figure S16).

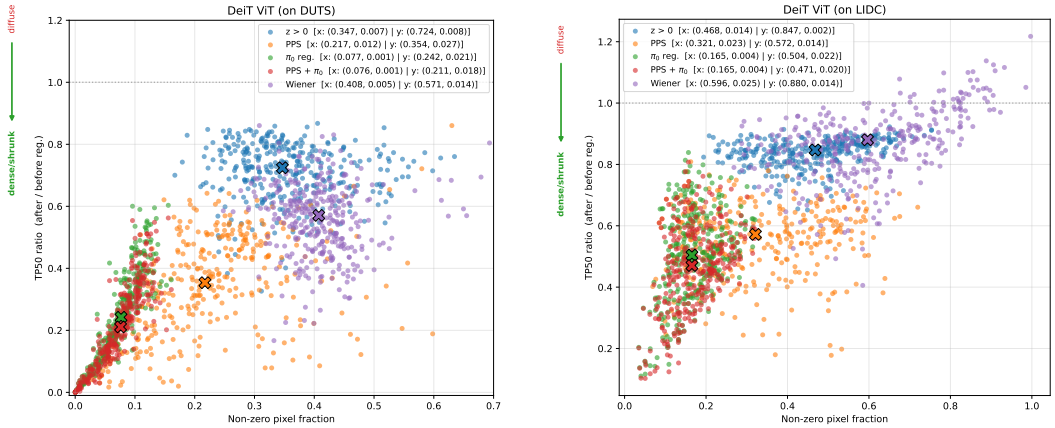

Figure S36: The ratio of TP50 metric after and before regularization (y-axis) vs. the fraction of non-zero attention scores (x-axis) for DeiT ViT on DUTS (left) and LIDC-IDRI (right) for different methods: z-thresholding (blue), PPS (orange), π_0-thresholding (green), PPS_{π_0} (red) and Wiener (purple). Each dot corresponds to the results for one image. The median and variance for each distribution dimension are reported next to each legend entry: [x: (median, var), . . .]. The intersection of the corresponding median values for each method are shown with the cross signs of the respective color.

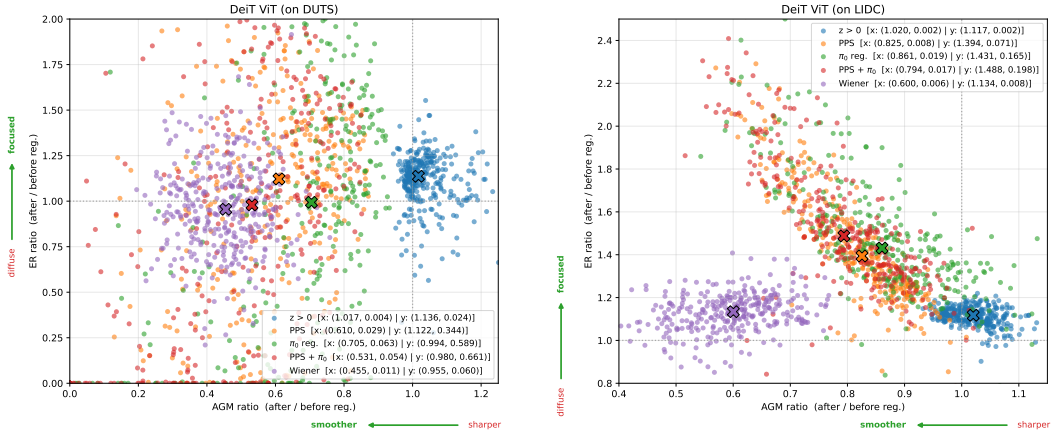

Figure S37: The ratios of two metric values after and before regularization: ER ratio (y-axis) vs. AGM contrast (x-axis) for DeiT ViT on DUTS (left) and LIDC-IDRI (right) for different methods: z-thresholding (blue), PPS (orange), π_0-thresholding (green), PPS_{π_0} (red) and Wiener (purple). The panel arrangement and graphical conventions follow those of Figure S36.

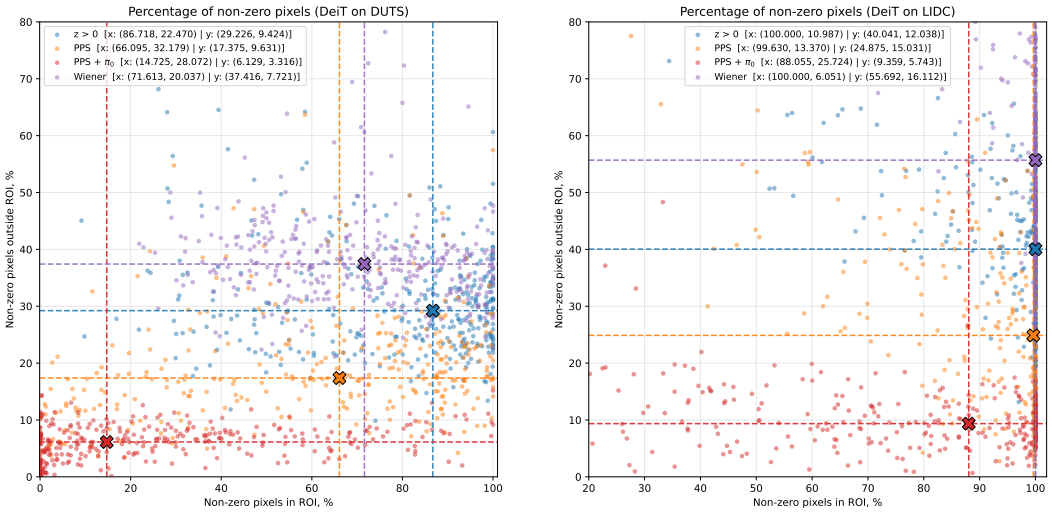

Figure S38: The percentage of non-zero pixels after regularization via different methods in ROI (x-axis) vs. outside of ROI (y-axis) for DeiT ViT on DUTS (left) and LIDC-IDRI (right). The panel arrangement and graphical conventions follow those of Figure S16.