Self-Configurable Information Bottleneck Attribution

Vayangi Ganepola[1,2,*][0009−0005−1619−1686],
Oluwabukola G. Adegboro[1,2,*][0000−0002−2704−8528],
Julia Dietlmeier[2,3][0000−0001−9980−0910],
Mayug Maniparambil[2,3][0000−0002−9976−1920],
Simon P. Wilson[3,4][0000−0003−0312−3586],
Aonghus Lawlor[3,5][0000−0002−6160−4639], Claudia Mazo[2][0000−0003−1703−8964],
and Noel E. O'Connor[2,3][0000−0002−4033−9135]

[1] Research Ireland Centre for Research Training in Machine Learning (ML-Labs)
{vayangi.ganepola2, oluwabukola.adegboro2}@mail.dcu.ie
[2] Dublin City University, Glasnevin, Dublin 9, Ireland
[3] Insight Research Ireland Centre for Data Analytics
[4] Trinity College Dublin, Dublin 2, Ireland
[5] University College Dublin, Belfield, Dublin 4, Ireland
[*]Equal contribution.

Abstract. The Information Bottleneck Attribution (IBA) framework is a leading information-theoretic approach to explaining neural network predictions. However, its effectiveness is hindered by a critical limitation: the subjective, manual selection of the bottleneck layer and the trade-off parameter β. This arbitrary selection strategy compromises reliability and limits broad applicability. We introduce Self-Configurable IBA (SC-IBA), a novel algorithm that automates these crucial parameters by recognizing that the optimal bottleneck is inherently input-specific. Leveraging the Kernelized Information Bottleneck (KIB) formulation via the Hilbert-Schmidt Independence Criterion (HSIC), we reformulate layer selection as a discrete Lagrangian optimization problem. By computing centered channel-wise Gram matrices for each convolutional layer, we define a robust information profit score that maximizes the difference between task-relevant sufficiency and input complexity. This theoretically grounded metric identifies the semantic peak within the feature hierarchy automatically for every input. We validate SC-IBA across diverse domains, including natural images (ImageNet) and decision-critical medical imaging (MSSEG16, Cheng). Across these three distinct datasets, SC-IBA consistently surpasses standard methods, outperforming a strong baseline rigorously tuned via systematic hyperparameter optimization by a significant margin in the vast majority of metric instances. SC-IBA transforms IBA into a robust, theoretically justified, self-optimizing standard for neural network explainability. The code is available at `https://github.com/VishmiVishara/SC-IBA`.

Keywords: Explainable AI · Hilbert-Schmidt Independence Criterion · Information Bottleneck Attribution · Kernelized Information Bottleneck

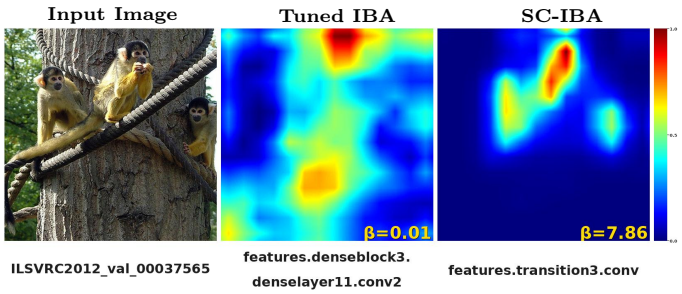

Fig. 1: Attribution maps (ImageNet sample) on DenseNet121, comparing systematically tuned IBA baseline **(middle)**, the best-performing baseline configuration against the proposed SC-IBA **(right)**. While the baseline relies on a static target layer and a fixed β value derived via hyperparameter search on a sample training set, SC-IBA dynamically identifies the most informative layer and adaptively scales β (here, β=7.86) for this specific input, achieving superior target localization.

1 Introduction

Explainable AI (XAI) and transparent decision-making are fundamental to modern computer vision. However, deep neural networks are often criticized for their lack of transparency, making it difficult to understand how they arrive at a particular decision [16]. Post-hoc interpretability seeks to solve this problem by explaining a model's predictions after the fact. These algorithms work by assigning a contribution score to each input feature, essentially highlighting which parts of the input were most important for the model's final prediction. Various attribution methods are used to achieve this goal, including perturbation-based methods such as model-agnostic LIME [15] and SHAP [13], gradient-based techniques such as the Grad-CAM family of algorithms [20], attention-based attribution [4], Information Bottleneck Attribution (IBA) class of algorithms [18], [9], [29] and backpropagation-based Integrated Gradients [25].

Specifically, IBA [18] is an information-theoretic attribution approach that partitions a model by inserting a *bottleneck* to restrict the flow of information through the model. Although, showing strong performance on natural images, the IBA requires an empirical tuning of the layer index and the trade-off parameter β that controls how much information is flowing through the bottleneck layer. Moreover, the original IBA work assumes that this layer selection is stationary for every input image.

In this paper, we propose Self-Configurable Information Bottleneck Attribution (SC-IBA). As illustrated in Figure 1, SC-IBA yields significant qualitative improvements in target localization compared to the standard IBA baseline. Our framework is built upon two fundamental insights: (1) the optimal bottleneck layer is not a static architectural property but an input-dependent dynamic, and (2) the trade-off between task-relevant sufficiency and input redundancy can be

analytically optimized via a Kernelized Information Bottleneck (KIB) objective [28,14].

SC-IBA offers four primary advantages over existing attribution methods. First, it is a distribution-agnostic approach that operates without restrictive Gaussian assumptions. Second, it replaces subjective, manual trial-and-error with a Lagrangian optimization, ensuring reproducibility across diverse architectures. Third, by utilizing the trace of centered, channel-wise Gram matrices, we maintain strict alignment with the information-theoretic definitions of sufficiency and complexity. Fourth, the framework dynamically identifies the "Semantic Peak" for every individual image, accounting for variations in feature hierarchies that global selection methods overlook. Our contributions are:

- We introduce **SC-IBA**, a framework that automates the selection of the bottleneck layer and the trade-off parameter β through Lagrangian optimization of the Information Bottleneck functional.
- We propose the **Information Profit Score (IPS)**, a non-parametric metric derived from the Hilbert-Schmidt Independence Criterion (HSIC) [8]. Using channel-wise, weight-aligned kernels, the IPS identifies the optimal representation depth without manual hyperparameter tuning.
- We demonstrate that SC-IBA consistently outperforms standard IBA baselines across both natural and medical imaging domains. Furthermore, SC-IBA directly surpasses a heavily optimized, tuned baseline by a substantial margin in nearly all evaluation metrics, proving the superiority of our instance-adaptive derivation over brute-force heuristic search.

2 Related Works

Attribution-based methods, a key component of XAI, provide an overview into a model's decision making process by assigning a relevance score to each input variable e.g. pixels, and quantifies its contribution to a specific output [18]. These relevance scores can be visualized as saliency heatmaps displayed over the input's pixels on images, hence highlighting distinct areas useful for the network's prediction. These methods fall into three broad categories: gradient-based, activation-based, and information-theoretic.

Gradient-based attribution techniques estimate feature importance by computing the gradients of the model's output with respect to the input features. Saliency Maps [22] computes the gradients of the class score using a single backpropagation pass; Integrated Gradients [26] and SmoothGrad [23] average gradients across multiple input perturbations to reduce noise, while Guided Backpropagation [24] suppresses negative gradients for sharper maps.

Activation-based methods use intermediate feature maps to localize relevant regions. GradCAM [19] uses class-specific gradients at the last convolutional layer to produce coarse activation maps. Guided GradCAM [21] combines GradCAM with GuidedBP for higher resolution and class discriminative visualizations, and EigenCAM [3] visualizes principal components of intermediate features.

Information bottleneck methods apply the Information Bottleneck Principle [28] to feature attribution. The original IBA [18] and its variants - InputIBA [31], InverseIBA [11], ML-IBA [11], IBISA [6], and CoIBA [9] - intervene at different layers or resolutions to suppress irrelevant information. However, these approaches still rely on manually selecting the bottleneck layer and the trade-off parameter β for all images. Our work addresses this limitation by proposing SC-IBA, which automatically selects both the bottleneck layer and β per input sample.

3 Method

3.1 Original IBA framework

For each model layer l, IBA calculates a new representation, called the bottleneck representation Z_l as shown in Eq.1. This is a combination of the original activation R_l and the independent noise ϵ_l i.e. random noise sampled from a Gaussian distribution. The noise term ϵ_l has the same statistical properties as the original activation, helping to maintain the general characteristics of the data while restricting unnecessary information flow.

$$Z_l = \theta_l R_l + (1 - \theta_l)\epsilon_l. \tag{1}$$

This mixing is controlled by a parameter called the damping ratio θ_l, which is a value between 0 and 1. The goal of IBA is to find the perfect balance for θ_l at each layer. It does this by solving an optimization problem [18]:

$$\max_{\theta_l} I[Z_l; Y] - \beta I[R_l, Z_l]. \tag{2}$$

Maximizing $I[Z_l; Y]$ keeps as much information as possible about the final label Y in the bottleneck representation Z_l. This step ensures that the information that is most relevant for prediction passes through the bottleneck. Minimizing $I[R_l, Z_l]$ means removing as much information as possible that is shared between the original activation R_l and the bottleneck representation Z_l. This is where compression takes place - it forces the bottleneck to only hold the most critical information, removing everything else.

The hyperparameter β controls the trade-off between these two goals. A higher β means the model prioritizes compressing the information (removing more details), while a lower β means it prioritizes keeping the information relevant to the final prediction. In summary, IBA finds the minimal amount of information at each layer that is still sufficient to make a correct prediction. This minimal information is then used to create the attribution map.

3.2 Motivation

Using IBA involves two crucial manual choices. First, one has to choose a layer l to analyze. The user must select an intermediate layer of the neural network to

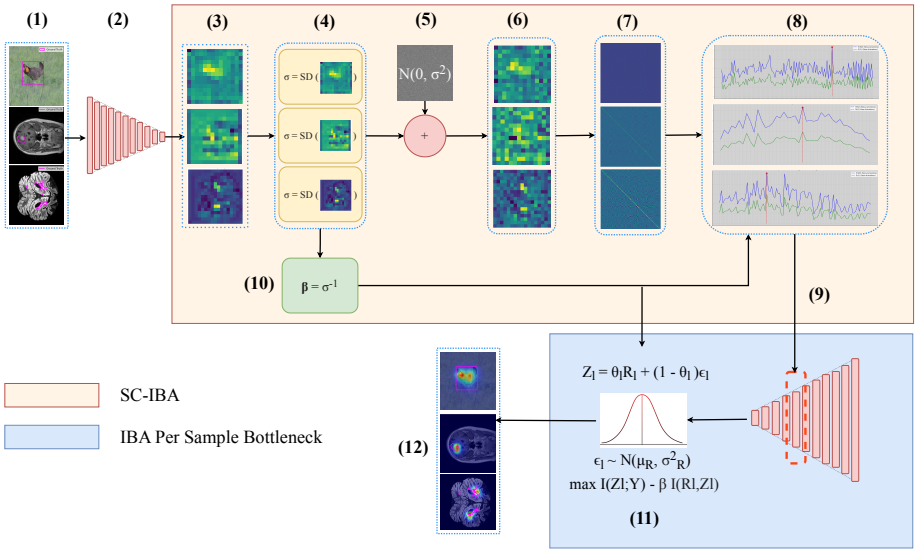

Fig. 2: The concept of the proposed SC-IBA. Our processing pipeline includes the following steps: (1) input images; (2) classification CNN e.g. DenseNet121, modified ResNet50 and InceptionV3; (3) get activations per layer; (4) compute σ (SD of activations); (5) generate Gaussian noise using σ; (6) apply Gaussian noise to activations; (7) compute $\mathbf{K}_Z$, $\mathbf{K}_R$, $\mathbf{K}_Y$ kernels; (8) compute IPS; (9) select the optimal layer with the highest IPS; (10) compute β; (11) apply per-sample bottleneck as in original IBA [18]; (12) output the attribution maps.

serve as the bottleneck. The quality of the attribution map is highly dependent on this choice. A layer that is too shallow may not contain high-level semantic information, while a layer that is too deep might have already lost crucial spatial details. Second, the β parameter controls the trade-off between fidelity to the original model's output and the simplicity of the explanation. Choosing the right β is a delicate balance, and a single value may not be optimal for different inputs or different layers.

3.3 Proposed self-configurable IBA

We outline the theoretical framework of SC-IBA (Figure 2) by adopting the standard definition of layer activations $R_l \in \mathbb{R}^{C\times H\times W}$ for each layer l. Here, C denotes the number of channels, while H and W represent the spatial dimensions. To isolate the most robust features, we introduce isotropic Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ to obtain perturbed activations $N_l = R_l + \epsilon$. This stochastic perturbation acts as an information filter; as the noise suppresses non-essential fluctuations, the remaining representation highlights the fundamental structural features utilized by the model. To analyze these representations, we reshape the

tensor N_l into a matrix $\mathbf{A} \in \mathbb{R}^{C\times(H\times W)}$. We then construct a positive semi-definite (PSD), symmetric channel-wise Gram matrix $\mathbf{M} \in \mathbb{R}^{C\times C}$, representing the pairwise inner products between feature maps:

$$\mathbf{M} = \mathbf{A}\mathbf{A}^T. \tag{3}$$

To ensure scale invariance across varying network depths, we normalize $\mathbf{M}$ by its Frobenius norm $||\mathbf{M}||_F = \sqrt{\sum_{i,j} |\mathbf{M}_{ij}|^2}$. The normalized matrix $\hat{\mathbf{M}} = \mathbf{M}/||\mathbf{M}||_F$ removes dependencies on the absolute magnitude of the feature space. Finally, we center the matrix to obtain the bottleneck kernel $\mathbf{K}_Z$:

$$\mathbf{K}_Z = \mathbf{H}\hat{\mathbf{M}}\mathbf{H}, \tag{4}$$

where $\mathbf{H} = \mathbf{I} - \frac{1}{C}\mathbf{1}\mathbf{1}^T$ is the centering matrix.

We reformulate the IBA framework [18] by replacing the variational upper bound with a KIB objective [28,14]. By substituting Shannon Mutual Information with the Hilbert-Schmidt Independence Criterion (HSIC) [8], we leverage a non-parametric measure of dependence. Unlike variational approaches that assume Gaussian activation distributions, our HSIC-based formulation effectively captures non-linear dependencies without the bias of parametric assumptions - a critical advantage given the non-normality induced by ReLU activations and architectural bottlenecks. For two centered kernels $\mathbf{K}_U$ and $\mathbf{K}_V$:

$$HSIC(\mathbf{K}_U, \mathbf{K}_V) = \frac{1}{(C-1)^2}\mathrm{Tr}(\mathbf{K}_U\mathbf{K}_V). \tag{5}$$

Note that because $\mathbf{K}_U$ and $\mathbf{K}_V$ are pre-centered, the estimator simplifies to the trace of their product. To identify the optimal layer for attribution, we define and maximize the Information Profit Score (IPS) via the Lagrangian form:

$$IPS = \underbrace{HSIC(\mathbf{K}_Z, \mathbf{K}_Y)}_{\text{Sufficiency}} - \beta \cdot \underbrace{HSIC(\mathbf{K}_Z, \mathbf{K}_R)}_{\text{Complexity}} \tag{6}$$

where $\mathbf{K}_R$ is the centered channel-wise Gram matrix of the original (clean) activations. This formulation pinpoints the "Semantic Peak" - the hierarchical point where task-relevant information is most efficiently separated from input-level redundancy. The target kernel $\mathbf{K}_Y$ is constructed from the weights $\mathbf{W}$ of the succeeding layer. This choice is motivated by the fact that the succeeding layer's weights dictate which features from the current layer are preserved and transformed for the final prediction. By computing the channel-wise Gram matrix $\mathbf{K}_Y = \mathbf{W}_{flat}\mathbf{W}_{flat}^T$, we define the target structure as the feature-correlation manifold required for downstream operations. This allows the sufficiency term in Eq. 6 to measure the alignment between the layer's internal correlations and the model's decision logic.

The IPS objective (Eq. 6) generalizes the Information Bottleneck (IB) framework (Eq. 2) by substituting traditional Mutual Information (MI) with HSIC-based dependence measures. This transition is motivated by the inherent challenges of estimating MI in deep feature maps, where variational bounds often

introduce significant noise. By leveraging the kernel-based properties of HSIC within the Reproducing Kernel Hilbert Space (RKHS), the IPS criterion captures high-dimensional dependencies directly. This allows for the *deterministic* identification of the optimal layer and β, offering a mathematically rigorous alternative to the stochastic and computationally intensive brute-force optimization of the standard IBA objective. To the best of our knowledge, we are the first to formulate this selection criterion by utilizing channel-wise Gram matrices to represent the latent feature distributions, allowing for a more granular capture of spatial-semantic information across the network's depth. Critically, despite the formal symbolic similarities between Eq. 6 and Eq. 2, the implementation of IPS is fundamentally distinct: it replaces the iterative, stochastic optimization of a variational proxy with a direct, algebraic calculation of kernel alignment.

We specifically utilize a linear kernel because deep CNN layers act as non-linear feature extractors that project raw data into a manifold where task-relevant information is more linearly accessible [30]. In this high-level feature space, the relationship between latent representations and the target output is often approximately linear, rendering complex non-linear kernels (e.g., RBF or Gaussian) redundant. Furthermore, the linear kernel avoids hyperparameter nesting - such as the sensitive tuning of the RBF bandwidth - thereby improving the reproducibility and computational efficiency of the SC-IBA framework in clinical settings. By using a linear kernel, we are essentially performing Linear HSIC, which is mathematically equivalent to measuring the Hilbert-Schmidt norm of the cross-covariance operator. This approach is a principled method for measuring information alignment in deep representations, as demonstrated by its relationship to Centered Kernel Alignment (CKA) [12], and provides a bias-free measure of dependency without the need for manual kernel scaling [8].

β Selection To maintain local adaptability, the noise σ is set to the standard deviation of the layer's activations, while β is adjusted inversely to σ to balance the compression pressure.

4 Datasets and Implementation Details

The experimental validation of the proposed approach spans two domains: natural and medical (Figure 3). ImageNet is selected as the natural image dataset, while Cheng et al. [5] (brain tumor) and MSSEG16 (Multiple Sclerosis (MS) lesion) datasets [7] are selected as the medical imaging datasets. Specifically in MSSEG16, based on the lesion availability in the mask, two classes are defined as lesion-positive and lesion-negative. The train set consists of 23,532 slices, comprising 12,700 (54.0%) lesion-positive and 10,832 (46.0%) lesion-negative slices. The validation set contains 2,915 slices with 1,487 (51.0%) of lesion-positive and 1,428 (49.0%) lesion-negative slices. The test set consists of 6,037 slices with 3,744 (62.0%) lesion-positive and 2,298 (38.0%) lesion-negative slices. More details are provided in the supplementary file[6].

[6] https://github.com/VishmiVishara/SC-IBA/tree/main/supplementary

Different techniques using three state-of-the-art CNN architectures and three datasets are implemented to test the generalizability of the proposed method. All experiments are performed using the PyTorch framework on a single NVIDIA GeForce RTX 5090 GPU (32GB).

We use the pretrained DenseNet121 architecture [10] for experiments with the ImageNet dataset. A pretrained ResNet50 model architecture (named **SEA-ResNet50**) is adapted and used to conduct experiments on the Cheng dataset [1], achieving a standard test accuracy of 98.36% (AUC: 0.9931). We create a custom architecture, **MedicalInceptionV3**, for MS to classify lesion and non-lesion 2D slices. This model adapts the standard InceptionV3 [27] architecture for single channel medical data and integrates an attention mechanism to enhance the feature representation, achieving a test accuracy of 90.31% (AUC: 0.9657) on the MSSEG16 dataset. More details on models' configuration, training, and classification performance are provided in the supplementary file.

5 Attribution Methods for Comparison

We systematically benchmark our dynamic optimal bottleneck layer and β selection approach (Section 3) against multiple baseline configurations. These range from standard static parameter formulation, which replicates dynamics of the per-sample bottleneck of the original IBA [18] to strongly optimized variants with rigorously tuned hyperparameters (target layer and β).

The original IBA framework [18] selects a layer from the middle of the network i.e. `conv3_4` for ResNet50 and `Mixed_5b` for InceptionV3 and uses this same layer to generate attributions for all images. Since our SEA-ResNet50 and MedicalInceptionV3 differ in depth, we cannot use these specific layer names directly. Therefore, we establish four fixed layer baseline configurations. The first three are fixed-layer approximations: (1) one using the middle layer of our respective model, and (2 & 3) two baselines by selecting a layer at an equivalent proportional depth based on ResNet-50 *(Resnet-50 Ratio)* and InceptionV3 *(InceptionV3 Ratio)*. We determine the target layer by calculating the original IBA layer's selection ratio and applying it to our own model's total layer count N to find the corresponding layer. For example in the *Resnet-50 Ratio*, we calculate the ratio of the `conv3_4` layer's index to the total number of layers in standard ResNet50, multiplying this fraction by N. To ensure a fair comparison with the original IBA, we keep the β value static ($\beta = 10$) for these fixed-layer experiments.

For our fourth baseline, we implement a global hyperparameter search using Optuna [2] with a Random Sampler search strategy. The sampler selects layer–β pairs uniformly at random. First, we estimate IBA statistics for all available convolutional layers using the training set. The calibrator then tests sampled layer–β pairs on the validation split to find the best combination. The best pair is fixed, and final attribution evaluation is performed on the test set using those values. This provides a strong validation-driven baseline for comparison with SC-IBA. Configuration details are provided in the supplementary file.

5.1 Evaluation Metrics

The proposed SC-IBA, baselines and attribution methods are evaluated with degradation, bounding box, and weakly-supervised segmentation metrics (e.g. Dice). To effectively capture the result performance, we report the same metrics used in the original IBA paper (detailed in the supplementary file). The degradation integral score (e.g. Deg. 8×8 score) measures completeness by progressively degrading the most salient pixels salient regions [17] and measuring the change in model confidence, using a patch-based perturbation scheme. Bounding box (ratio top in bbox), measures the localization accuracy by identifying the top most salient pixels. Then it calculates the ratio of these specific pixels that fall inside the ground truth object bounding box. A high score means the heatmap's peak attributions are correctly positioned on the object [18].

6 Results and Discussion

The proposed method is compared against the original IBA baselines, and other state-of-the-art attribution methods, with quantitative results summarized in Tables 1, 2 and 3. To provide a comprehensive and transparent evaluation of our dynamic approach, we explicitly report both the absolute performance differences and the relative percentage gains of SC-IBA compared to the IBA baselines. Meanwhile, the ***Absolute diff. (Δ)*** reflects the direct magnitude of the increase relative to the *highest-performing* baseline, which is selected based on Dice and Deg. 8×8 metrics for MSSEG16 and Cheng and the Deg. 8×8 metric for the ImageNet. Selected qualitative results are presented in Figure 3 for each dataset.

ImageNet: As presented in Table 1, our SC-IBA achieves the state-of-the-art performance, outperforming all four tested IBA baselines and other gradient and activation-based methods. It achieves the highest mean degradation integral score of 0.4395 and the highest mean ratio top in bbox of 0.7814. These results suggest that the attributions generated by SC-IBA are not only more accurately localized, the highest number of top pixels in the attribution map is inside the targeted object's bounding box. Notably, it has also achieved the lowest standard deviation (± 0.1759) among IBA-based methods in ratio top in bbox metric, indicating highly consistent performance. Among IBA baselines, *Tuned Baseline* is the strongest baseline, and our approach demonstrates a substantial **55.80**% increase (Δ= +0.1574) in the Deg. 8×8 metric and a **28.69%** improvement (Δ=+0.1742) in the localization metric over this best-performing baseline. Furthermore, SC-IBA demonstrates vastly superior computational efficiency, requiring only 4 h 28 min and 5sec of GPU runtime compared to the tuned baseline's 17 h 33 min and 17 sec.

MSSEG2016: This robust performance extends seamlessly to the medical domain. Table 2 demonstrates SC-IBA's superior performance, particularly in segmentation overlap and localization accuracy. It achieves a state-of-the-art Dice score of 0.2061 and a bbox score of 0.3365, outperforming all the IBA baselines and competing methods in these metrics. Notably, SC-IBA also maintains

Table 1: Quantitative results on ImageNet (DenseNet121). SC-IBA outperforms all baselines and competing methods. Metrics include Degradation Integral Score (Deg. 8×8), and Ratio Top in Bounding Box (bbox). Best, second-best, and third-best results are highlighted in **bold**, underline, and *italics*, respectively.

Method	**Deg. 8×8↑**	**bbox ↑**
Middle Layer	0.1381 ± 0.2137	0.3952 ± 0.3673
ResNet-50 Ratio	0.1188 ± 0.1990	0.3700 ± 0.3706
InceptionV3 Ratio	0.2770 ± 0.2618	0.6072 ± 0.3008
Tuned Baseline	0.2821 ± 0.2765	0.6010 ± 0.2725
SC-IBA (Ours)	**0.4395 ± 0.2662**	**0.7814 ± 0.1759**
Absolute Diff. (Δ)	**+0.1574**	**+0.1742**
EigenCAM	0.1547 ± 0.3811	0.6065 ± 0.3132
GradCAM	0.3043 ± 0.2751	*0.6470 ± 0.2743*
GuidedBP	*0.3682 ± 0.2531*	0.6435 ± 0.2062
Guided GradCAM	0.3844 ± 0.2675	0.6765 ± 0.2224
Saliency	0.2124 ± 0.2645	0.5768 ± 0.2501

the lowest standard deviation (±0.1705) in bbox metric, confirming its consistency in small-lesion localization. Among the IBA baselines, the *ResNet-50 Ratio* is the most competitive. Nevertheless, our SC-IBA still maintains a significant improvement of **33.05%** (Δ=+0.0512) in the Dice, **6.58%** (Δ=+0.0274) in Degradation (integral score) and a **40.50%** (Δ=+0.0970) in bbox over this strongest baseline.While Guided GradCAM shows a marginal edge in the Degradation (integral score), SC-IBA fundamentally outperforms all other techniques in every metric tied directly to spatial overlap and precise target localization (Dice score and bbox metric). The generally high standard deviations across all methods highlight the inherent complexity and high variance of the MSSEG16 dataset. Despite this challenge, the substantial quantitative gains confirm that SC-IBA provides a much more reliable, dense, and medically relevant attribution map compared to traditional CAM or gradient-based methods. This robust performance is paired with significantly lower computational overhead. SC-IBA completes the evaluation in 3 h 47 min and 7 sec, whereas the tuned baseline requires over a full day. (1 day 1 h, 46 min and 21 sec).

Cheng Brain Tumor: Experimental results in Table 3 reveal that SC-IBA outperforms all four standard baselines except the Ratio Top in Bounding Box (bbox) result for the Tuned Baseline. While it underperforms the heavily optimized tuned baseline in bbox, its dice score of 0.1771 ± 0.2387 and degradation score of 0.1166 ± 0.1282 surpasses the Tuned Baseline's dice with a gain of **3.27%** (Δ=+0.0056) and degradation score with a gain of **60.16%** (Δ=+0.0438). This divergence highlights a fundamental distinction in task complexity and metric evaluation. Multi-class tumor differentiation on Cheng requires a holistic analysis of complex, diffuse structural cues and surrounding anatomical landmarks. Because SC-IBA optimizes for a minimal, high-sufficiency representation, it prunes

Table 2: Quantitative results on the MSSEG16 dataset. Metrics include Dice Similarity Coefficient (Dice), Degradation Integral Score (Deg. 8×8), and Ratio Top in Bounding Box (bbox). Best, second-best, and third-best results are highlighted in **bold**, underline, and *italics*, respectively.

Method	**Dice ↑**	**Deg. 8×8 ↑**	**bbox ↑**
Middle Layer	0.1116 ± 0.1291	0.3572 ± 0.1990	0.2040 ± 0.1847
ResNet-50 Ratio	0.1549 ± 0.1283	0.4164 ± 0.2050	0.2395 ± 0.2001
InceptionV3 Ratio	0.0642 ± 0.1130	0.2910 ± 0.2242	0.1352 ± 0.1735
Tuned Baseline	0.1130 ± 0.1295	0.3580 ± 0.1986	0.2053 ± 0.1855
SC-IBA (Ours)	**0.2061 ± 0.1779**	0.4438 ± 0.2108	**0.3365 ± 0.1705**
Absolute Diff. (Δ)	**+0.0512**	**+0.0274**	**+0.0970**
EigenCAM	0.0838 ± 0.1076	0.4246 ± 0.1892	0.1633 ± 0.1914
GradCAM	0.0965 ± 0.1118	*0.4433 ± 0.1551*	0.1579 ± 0.1865
GuidedBP	0.1336 ± 0.1329	0.4082 ± 0.1915	*0.2430 ± 0.1245*
Guided GradCAM	*0.1389 ± 0.1360*	**0.4521 ± 0.1695**	0.2493 ± 0.1392
Saliency	0.0962 ± 0.1041	0.4080 ± 0.1922	0.1984 ± 0.1282

away this broad non-focal context. While this aggressive background filtering yields lower overlap scores (bbox) compared to the looser, broader maps of the tuned baseline, SC-IBA remains highly competitive in dice and degradation. In terms of computational cost, SC-IBA requires substantially less runtime (27 min) compared to the tuned baseline (3 h 6 min 43 sec). This proves that SC-IBA successfully pinpoints the precise, high-density pixels driving the model's actual decisions, establishing superior faithfulness even when it contradicts human-annotated boundaries, while maintaining significantly lower computational overhead.

β **Ablation Study:** To isolate the impact of our dynamic parameter selection, we conduct an ablation study on the trade-off parameter β while keeping the layer selection fixed to the optimal choice identified by the IPS criterion. We compare fixed β values evaluated across a logarithmic scale $\{0.01, \ldots, 100\}$ against our analytically derived $\beta = 1/\sigma$. As shown in Table 4, this dynamic selection consistently outperforms fixed heuristics in both the Degradation Integral and Bbox Ratio metrics. This confirms that $1/\sigma$ optimally adapts the sufficiency-complexity trade-off for each specific feature map, providing a superior bottleneck compared to global constant values.

Figure 4 illustrates the performance trade-off between faithfulness (Degradation 8x8 integral) and localization (Ratio top in bbox) across the ImageNet, MSSEG16, and Cheng datasets. As shown by its placement in the top right region of each subplot, SC-IBA consistently achieves a superior balance compared to various baseline techniques. Furthermore, on the medical datasets, SC-IBA maintains highly competitive Dice scores (represented by marker size) alongside this optimal trade-off. Across three distinct datasets and multiple CNN architec-

Table 3: Quantitative results on the Cheng Brain Tumor dataset. Metrics include Dice Similarity Coefficient (Dice), Degradation Integral Score (Deg. 8×8), and Ratio Top in Bounding Box (bbox). Best, second-best, and third-best results are highlighted in **bold**, underline, and *italics*, respectively.

Method	**Dice ↑**	**Deg. 8×8 ↑**	**bbox ↑**
Middle Layer	0.1032 ± 0.2108	0.0679 ± 0.1284	0.1951 ± 0.2877
ResNet-50 Ratio	0.0123 ± 0.0731	0.0068 ± 0.0445	0.0262 ± 0.1201
InceptionV3 Ratio	0.0182 ± 0.0933	0.0059 ± 0.0507	0.0416 ± 0.1518
Tuned Baseline	0.1715 ± 0.2187	0.0728 ± 0.1284	0.3259 ± 0.3051
SC-IBA (Ours)	*0.1771 ± 0.2387*	0.1166 ± 0.1282	0.2626 ± 0.3006
Absolute Diff. (Δ)	**+0.0056**	**+0.0438**	**-0.0633**
EigenCAM	0.1017 ± 0.1370	0.1113 ± 0.1102	0.3003 ± 0.3137
GradCAM	0.2082 ± 0.2466	**0.1318 ± 0.1098**	*0.3141 ± 0.3179*
GuidedBP	**0.2177 ± 0.2644**	0.1016 ± 0.1193	**0.3320 ± 0.2302**
Guided GradCAM	0.1492 ± 0.2381	0.0602 ± 0.1123	0.2821 ± 0.2776
Saliency	0.1063 ± 0.1393	*0.1155 ± 0.1233*	0.1888 ± 0.1685

Table 4: SC-IBA β ablation results on the ImageNet. Metrics include Degradation Integral Score (Deg. 8×8) and Ratio Top in Bounding Box (bbox).

β **Value**	**Deg. 8×8 ↑**	**bbox ↑**
0.01	0.2692 ± 0.2759	0.6117 ± 0.2498
0.1	0.3298 ± 0.2828	0.6572 ± 0.2344
1	0.3596 ± 0.2799	0.6812 ± 0.2532
10	0.4367 ± 0.2677	0.7780 ± 0.1810
100	0.4251 ± 0.2708	0.7696 ± 0.1846
$1/\sigma$	**0.4395 ± 0.2662**	**0.7814 ± 0.1759**

tures, SC-IBA consistently outperforms standard IBA baselines. Remarkably, it also surpasses a heavily optimized, tuned baseline by substantial margin achieving superior faithfulness across all datasets and better localization in most. While alternative explainability methods may excel in specific metrics, SC-IBA establishes a new state-of-the-art benchmark for the IBA framework. Ultimately, these results demonstrate that instance-adaptive analytical derivation can successfully outperform brute-force heuristic tuning.

7 Conclusion

We introduce SC-IBA, a self-configuring, distribution-agnostic framework that transforms Information Bottleneck-based attribution from a manual heuristic into a rigorous, automated standard. Reformulating bottleneck selection as a discrete Lagrangian optimization problem, we define the IPS - a non-parametric metric derived from the HSIC, to analytically identify the "Semantic Peak" of a

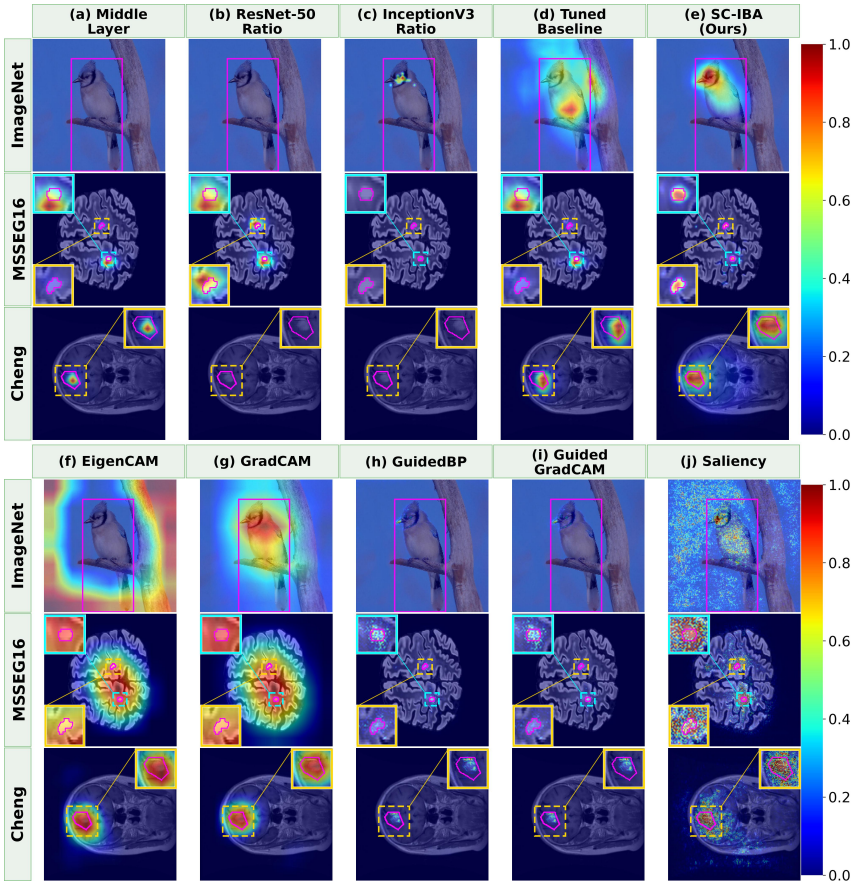

Fig. 3: Selected qualitative results showing the heatmaps of the implemented methods using different model architectures and datasets. Our SC-IBA shows superior focusing ability on the salient image parts.

network's feature hierarchy. This ensures the attribution process adaptively captures task-relevant information while filtering out input-level redundancy per image. By eliminating human bias and manual trial-and-error, SC-IBA establishes a scalable, mathematically grounded foundation for the next generation of self-optimizing diagnostic and interpretability tools. Although currently evaluated on convolutional networks, the model-agnostic nature of SC-IBA makes it readily tailorable to Vision Transformers, presenting a compelling direction for future work.

Acknowledgments. This work was funded by Taighde Éireann – Research Ireland under Grant number 18/CRT/6183 (Ganepola, Adegboro) and partially supported by the Insight Research Ireland Centre for Data Analytics. Our sincere thanks to Ms. Vicky Flanagan for her invaluable support in securing our computational infrastructure.

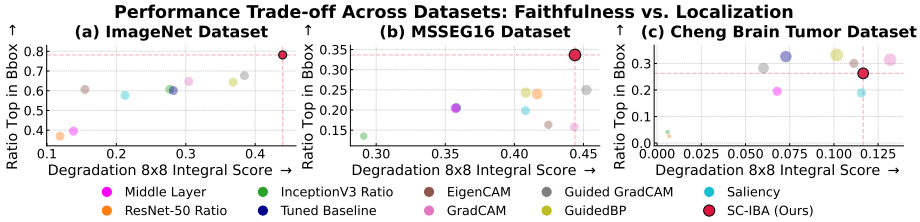

Fig. 4: Faithfulness vs. Localization Trade-off: Comparison of SC-IBA against baselines across three datasets. The x-axis shows faithfulness (Degradation 8x8 integral) and the y-axis shows localization (Ratio Top in BBox). Bubble sizes represent Dice scores for the medical datasets (b, c), while ImageNet (a) uses uniform markers.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this paper.

References

1. Adegboro, O., Ganepola, V., Dietlmeier, J., Mazo, C., O'Connor, N.E.: XAIMed-Net: towards explainable brain tumour detection in 2D T1-weighted CE-MRI images using transfer learning. In: IET Conference Proceedings CP887. vol. 2024, pp. 194–201. IET (2024)
2. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019)
3. Bany Muhammad, M., Yeasin, M.: Eigen-CAM: Visual explanations for deep convolutional neural networks. SN Computer Science **2**(1), 47 (2021)
4. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. ICCV pp. 397–406 (2021)
5. Cheng, J.: Brain tumor dataset (2017). `https://doi.org/10.6084/m9.figshare.1512427.v5`
6. Coelho, B., Cardoso, J.S.: Information bottleneck with input sampling for attribution. Neurocomputing **638**(C) (2025)
7. Commowick, O., et al.: Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset. NeuroImage **244** (2021)
8. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with Hilbert-Schmidt norms. International Conference on Algorithmic Learning Theory (2005)
9. Hong, J.H., Kim, H.J., Jeon, K.S., Lee, S.W.: Comprehensive information bottleneck for unveiling universal attribution to interpret vision transformers. CVPR (2025)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR. pp. 4700–4708 (2017)
11. Khakzar, A., Zhang, Y., Mansour, W., Cai, Y., Li, Y., Zhang, Y., Kim, S.T., Navab, N.: Explaining covid-19 and thoracic pathology model predictions by identifying informative input features. In: MICCAI. pp. 391–401. Springer (2021)

12. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International Conference on Machine Learning (ICML). pp. 3519–3529. PMLR (2019)
13. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. NIPS pp. 4768 – 4777 (2017)
14. Ma, W.D.K., Lewis, J., Kleijn, W.B.: The HSIC bottleneck: Deep learning without back-propagation. AAAI (2020)
15. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. KDD '16: Proceedings of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining pp. 1135 – 1144 (2016)
16. Salahuddin, Z., Woodruff, H.C., Chatterjee, A., Lambin, P.: Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Computers in Biology and Medicine **140** (2022)
17. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., , M¨uller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems **28**(11), 2660–2673 (2016)
18. Schulz, K., Sixt, L., Tombari, F., Landgraf, T.: Restricting the flow: Information bottlenecks for attribution. ICLR (2020)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision (IJCV) (2019)
21. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-CAM: Why did you say that? NIPS (2016)
22. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. ICLR (2014)
23. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: SmoothGrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
24. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. ICLR (2015)
25. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. ICML (2017)
26. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
27. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016)
28. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. In: proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing. pp. 368–377 (1999)
29. Wang, Y., Rudner, T.G.J., Wilson, A.G.: Visual explanations of image-text representations via multi-modal information bottleneck attribution. NeurIPS (2023)
30. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: ICML. pp. 12310–12320. PMLR (2021)
31. Zhang, Y., Khakzar, A., Li, Y., Farshad, A., Kim, S.T., Navab, N.: Fine-grained neural network explanation by identifying input features with predictive information. Advances in Neural Information Processing Systems **34**, 20040–20051 (2021)