Making the Classification Explanation Faithful to the Confidence Score

Supplementary Material

The supplementary is organized as follows.

- Appendix A: Algorithm description of the MHE.
- Appendix B: Supplementary experimental details.
  - Appendix B.1: Ablation study and analysis of the MHE method and the PNN metric.
  - Appendix B.2: Analysis of metrics involved in quantitative experiments.
  - Appendix B.3: Analysis of the new metric PNN.
  - Appendix B.4: Runtime and analysis of MHE.
  - Appendix B.5: Verification experiments for two assumptions related to MHE.
- Appendix C: Supplementary qualitative experiments and analysis.
- Appendix D: Supplementary quantitative experiments and analysis (datasets: ImageNet, CUB-200-2011, VOC2012).

A. MHE Algorithm Description

A.1. MHE

The proposed MHE framework is described as Algorithm 1.

Algorithm 1 Metropolis-Hastings Explainer

Require: Input image I, classifier f, target class c, total iterations n, burn-in period k, deviation threshold ε
Ensure: Explanation map $S^c(I)$

Initialize:
Generate random mask m_0
$m_{acc}[0] \leftarrow m_0$, $i \leftarrow 0$
for $t = 1$ **to** n **do**
  if $t \leq k$ **then**
    // Burn-in period: skip acceptance
    continue
  end if
  Proposal Sampling:
  Sample $m_j \sim P(i, j)$ based on $m_{acc}[i]$
  Score Computation:
  Compute $dis_j \leftarrow f(I \odot m_j)$
  Compute $dis_i \leftarrow f(I \odot m_{acc}[i])$
  Acceptance Probability:
  $\alpha(i, j) \leftarrow \min\left(1, \frac{e^{-dis_j} \cdot Q(j, i)}{e^{-dis_i} \cdot Q(i, j)}\right)$
  Acceptance/Rejection:
  Sample $x \sim \mathcal{U}[0, 1]$
  if $x < \alpha(i, j)$ **and** $dis_j < \varepsilon$ **then**
    $m_{acc}[i + 1] \leftarrow m_j$
    $i \leftarrow i + 1$
  else
    Retain $m_{acc}[i]$
  end if
end for
Aggregation:
$S^c(I) \leftarrow \sum_{i=0}^{N} m_{acc}[i] \cdot f_c(I \odot m_{acc}[i])$
return $S^c(I)$

A.2. MHE-e

The proposed MHE-e variant is described in Algorithm 2. MHE-e replaces the distance metric dis_i with the confidence score f_i^c of the mask m_i corresponding to class c. With each iterative operation, MHE-e endeavors to capture the positively contributing regions associated exclusively with class c.

A.3. MHE-pro

The proposed MHE-pro variant is detailed in Algorithm 3. MHE-pro mandates that each mask generation step leverages the explanation map obtained from the preceding MHE module as prior knowledge. The explanation result generated by the final MHE module constitutes the ultimate explanation.

B. Supplementary for The Experiment

Regarding the experimental setup, it is necessary to provide additional clarification. The input category prompts for the CLIP model are specified as ["a dog", "a cat", "a bird", "a car", "a person", "a tree"].

B.1. Ablation Study

This section discusses the ablation studies of the MHE framework and the PNN metric. The former includes the acceptance threshold, the number of rounds, the vector distance, and the mask variation. The latter includes threshold a and threshold b.

The ablation study on the MHE acceptance threshold is presented in Fig. 1(a). As the MHE acceptance threshold ε increases, the Del, Ins, AD, AI, PG, and EBPG metrics gradually improve, whereas the PNN and dis metrics gradually degrade before stabilization. This trend occurs because a larger MHE acceptance threshold results in a greater disparity between the explanation score and the original score, thereby diminishing performance in explaining the confidence score. We hypothesize that when the acceptance threshold decreases, MHE tends to search for specific combinations of positive and negative contribution re-

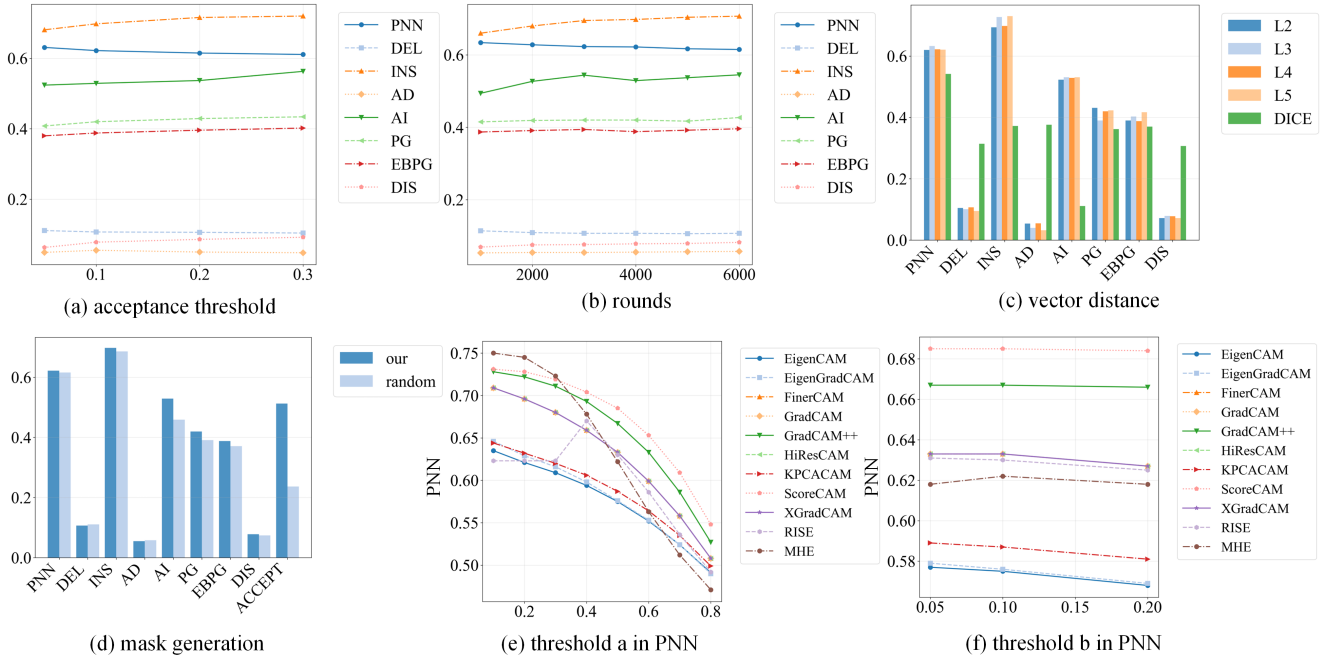


Figure 1. Ablation studies were conducted on the MHE acceptance threshold, the number of rounds, the dis_i computation method, the mask variation strategy, and the PNN metric thresholds a and b. As depicted in Figure (a), an increase in the acceptance threshold ε gradually enhances the localization capability while slightly attenuating the capacity to explain the confidence score. This suggests that a portion of MHE's confidence explanation capacity is being effectively redirected towards explaining the positive contribution regions. In Figure (b), MHE's performance gradually stabilizes as the number of rounds increases. Since the AI metric takes only two discrete values, and MHE's confidence fluctuates around the original score during stabilization, the AI metric exhibits noticeable fluctuations. Figure (c) shows the subtle impact of different norm selections in dis_i on MHE. As illustrated in Figure (d), utilizing our designed mask generation strategy not only improves performance on conventional metrics but also significantly enhances the acceptance efficiency, which is crucial for explaining challenging scenarios. In Figure (e), as threshold a increases, the proportion of artificially defined positive contribution regions decreases, resulting in reductions in both P_{dec} and N_{inc}. This occurs because, as a rises, a portion of the true positive contribution regions is reclassified from the defined interval $[a, 1]$ to $[b, a)$ (the artificially defined negative contribution area). In Figure (f), increasing threshold b designates a larger range of regions as non-contributing. Threshold b has a minimal impact on the performance of most explanation methods.

gions that strictly conform to the original score. Conversely, when the acceptance threshold increases, it may accept diverse region combinations. Moreover, since the confidence score is utilized as a weighting factor, samples containing a larger proportion of positive contribution regions attain higher weights, which subsequently leads to the improvement observed in the aforementioned multiple metrics.

The ablation study on the number of MHE rounds is shown in Fig. 1(b). As the number of rounds increases, the Del and Ins metrics gradually improve, PNN slightly deteriorates, AI first increases and then decreases, while the other metrics remain stable. The AD metric encourages the explanation's confidence score to lie within the range $[f_c(I) - \beta, 1]$ (where β is a very small positive number). Conversely, the dis metric encourages the score to reside within $[f_c(I) - \beta, f_c(I) + \beta]$. The minimal changes in dis and AD indicate that starting from 1000 rounds, the confidence score of MHE explanations stabilizes within the range $[f_c(I) - \beta, f_c(I) + \beta]$. The AI metric, which evaluates whether the explanation's confidence score exceeds the original score, takes only two values: 0 or 1. Given that MHE's confidence stabilizes within $[f_c(I) - \beta, f_c(I) + \beta]$, this naturally leads to pronounced fluctuations in the AI metric within that range. The PG and EBPG metrics are influenced by the overall distribution, suggesting that from 1000 rounds onward, the overall distribution of MHE explanations remains stable. It is generally accepted that the explanations of perturbation-based methods gradually stabilize as the number of rounds and samples increases. Based on the prior analysis that MHE searches for specific combinations of positive and negative regions, we analyze that MHE continuously approaches this specific combination, resulting in a gradual decrease in PNN and convergence to a stable level. In conjunction with the finding in Sec 4.3 that MHE-pro outperforms MHE in terms of PNN, this indicates that increasing the scale aids in enhancing the PNN

Algorithm 2 MHE-e Explainer

Require: Input image I, classifier f, target class c, total iterations n, burn-in period k, deviation threshold ε
Ensure: Explanation map $S^{c+}(I)$

Initialize:
Generate random mask m_0
$m_{acc}[0] \leftarrow m_0$, $i \leftarrow 0$
for $t = 1$ **to** n **do**
  if $t \leq k$ **then**
    // Burn-in period: skip acceptance
    continue
  end if
  Proposal Sampling:
  Sample $m_j \sim P(i, j)$ based on $m_{acc}[i]$
  Score Computation:
  Compute $f_j^c \leftarrow f(I \odot m_j)$
  Compute $f_i^c \leftarrow f(I \odot m_{acc}[i])$
  Acceptance Probability:
  $\alpha(i, j) \leftarrow \min\left(1, \frac{f_j^c \cdot Q(j, i)}{f_i^c \cdot Q(i, j)}\right)$
  Acceptance/Rejection:
  Sample $x \sim \mathcal{U}[0, 1]$
  if $x < \alpha(i, j)$ **then**
    $m_{acc}[i + 1] \leftarrow m_j$
    $i \leftarrow i + 1$
  else
    Retain $m_{acc}[i]$
  end if
end for
Aggregation:
$S^c(I) \leftarrow \sum_{i=0}^{N} m_{acc}[i] \cdot f_i^c$
return $S^{c+}(I)$

Algorithm 3 MHE-pro Explainer

Require: Input image I, target class c
Ensure: Explanation map $S^c(I)$

Initialize: Generate random mask m^0
Procedure: k cascaded MHE modules
$k_{size} \leftarrow [5, 7, 9, 11]$ ▷ Predefined kernel sizes
for $j = 1$ **to** k **do**
  $m^j = \text{MHE}(m^{j-1}, k_{size}[j], 1000)$ ▷ Modified MHE module, default 1000 rounds
  procedure of modified MHE module:
    1. Warm-up phase
    2. Clamp m^{j-1} values to range $[0.1, 0.9]$
    3. Sample $m_i \sim P(j - 1, i)$ based on m^{j-1}
    4. Subsequent operations follow original MHE
end for
$S^c(I) = m^k$
return $S^c(I)$

performance of the explanations.

An ablation study on the computation of the distance dis_i in MHE is presented in Fig. 1(c). In most scenarios, the metric results corresponding to different norms are relatively close. This indicates that the choice of norm has limited impact on MHE's localization capability and confidence faithfulness. In contrast, the DICE distance performs poorly in nearly all scenarios.

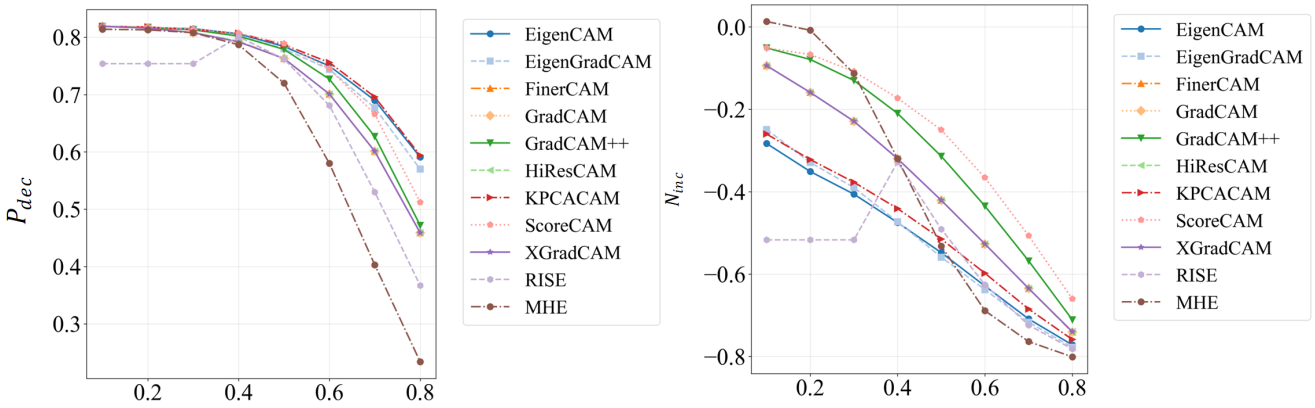
The ablation study on the mask variation strategy employed in MHE is presented in Fig. 1(d). Here, the ACCEPT metric refers to the sample acceptance rate, defined as the ratio of accepted samples to the total number of iterations. Our designed mask method exhibits varying degrees of advantage across different scenarios, with the exception of the dis metric, particularly demonstrating significant enhancement in the acceptance rate. For the dis metric, the performance difference between the two tested methods is minimal.

The ablation study on the PNN metric threshold a is presented in Fig. 1(e). As the threshold a increases, the PNN score of most explanation methods decreases (i.e., performance degrades). This observation indicates that for the artificially designated positive contribution region $[a, 1]$, as a increases, the actual positive contribution regions within this interval progressively shift into $[b, a)$, consequently leading to an overall reduction in the PNN metric score.

The variations of the sub-metrics of the PNN metric corresponding to the ablation study on threshold a are illustrated in Fig. 2. When $a < 0.3$, MHE exhibits the best performance in both P_{dec} and N_{inc}, suggesting that MHE possesses the most effective negative and positive contribution regions within the intervals $[0.1, 0.3]$ and $[0.3, 1]$, respectively.

Two potential reasons may account for the decrease in P_{dec}: a reduction in key positive contribution regions, or the introduction of key negative contribution regions. As illustrated in Fig. 2(a), when a increases, the interval corresponding to P_{dec} shrinks, indicating that the first reason is the primary influencing factor. MHE experiences the most rapid decline in the $a \in [0.4, 1]$ interval, implying that the positive contribution regions in MHE explanations are predominantly concentrated within this range. In contrast, the CAM-family methods and RISE maintain relatively high P_{dec} values even at $a = 0.8$, suggesting that they still retain a substantial number of effective positive contribution regions within the $[0.8, 1]$ interval.

The decrease in N_{inc} can be attributed to two factors: first, an insufficient number of negative contribution regions; second, the newly introduced positive contribution regions diminishing the effect of the negative contributions. As depicted in Fig. 2(b), when a increases, the interval corresponding to N_{inc} expands, suggesting that the second reason is the dominant factor. MHE maintains the highest N_{inc}



(a) threshold a in PNN

(b) threshold a in PNN

Figure 2. The variation trends of the P_{dec} and N_{inc} metrics in the ablation study on the PNN threshold a are illustrated in Figures (a) and (b). As the ranges of the two regions change, both P_{dec} and N_{inc} gradually decrease. This suggests that the studied explanation methods do not structure their results in strict accordance with the partitioning concept underlying PNN; instead, these methods predominantly prioritize the relative importance of the regions.

when $a < 0.3$, demonstrating that MHE possesses the most effective negative contribution regions in the $a \in [0.1, 0.3]$ interval. MHE shows the fastest decline in the $a \in [0.3, 0.6]$ interval, indicating that the positive contribution regions within this range can most effectively counteract the negative contribution regions. Integrating the previous analysis from P_{dec} regarding the concentration of positive contribution regions in MHE and other methods, since key positive contribution regions still exist in the $a > 0.8$ range, the N_{inc} values for CAM-family methods and RISE at $a = 0.8$ do not reach the minimum. Conversely, for MHE in the $a \in [0.6, 0.8]$ interval, the rate of decline in N_{inc} significantly slows, suggesting that the capacity of positive contributions within $[b, a]$ to diminish negative contributions is weakening, i.e., the actual importance of the two types of regions is becoming comparable. This further corroborates that MHE explains the original confidence score by balancing the importance of positive and negative contribution regions.

In summary, MHE exhibits the most effective irrelevant contribution regions in $[0, 0.1]$, negative contributions dominate in $[0.1, 0.2]$, positive contributions dominate in $[0.2, 0.6]$, and the importance of both types of regions becomes comparable in $[0.6, 1]$. For CAM-family methods, negative contribution regions dominate in $[0.1, 0.2]$, while the importance of positive contributions gradually increases in $[0.2, 1]$. RISE falls between these two patterns. This aligns with our earlier analysis: CAM-family methods and RISE tend to explain positive contribution regions (S^{c+}), whereas MHE explains the original confidence score (S^c) by balancing the importance of both types of contribution regions.

According to Fig. 1(e&f) , for all methods, PNN is insensitive to b but sensitive to a. Concurrently, Fig. 2(a) shows that across all explanation methods, the importance score interval from 0.5 to 1 (starting from the $[0.4, 0.5]$ range) begins to exhibit significant positive contribution. Therefore, it is reasonable to uniformly set $a = 0.5$.

Furthermore, PNN can be employed to analyze the type and intensity of contributions from different regions within the explanation maps produced by other explainers. For instance, when gradually increasing a, the newly added region $[a_{old}, a_{new}]$ reveals the following: a faster decline in P_{dec} indicates a stronger positive contribution; a slower decline in N_{inc} indicates a stronger negative contribution.

We provide a guideline for hyperparameter settings. As shown in Fig. 2, the P_{dec} values for almost all methods begin to decrease at $a = 0.4$, indicating that the positive contribution within the selected interval $[a, 1]$ starts to strengthen. The N_{inc} values for most methods start to accelerate their decline from $a = 0.4$, indicating that the negative contribution within this interval is being increasingly offset by the positive contribution. Therefore, $a = 0.4$ serves as an effective threshold for current common methods, with a recommended range of $[0.3, 0.5]$. For future explanation methods, if the intervals of positive and negative contributions are unclear, one can also investigate the positive and negative effects of each interval by performing an ablation study using the PNN metric, as illustrated in Fig. 2,. The ablation study reveals that the threshold b has a negligible impact. Furthermore, since a low importance score implies a minor contribution, b can be set to a small value within $[0, 0.2]$, for instance, $b = 0.1$.

Concerning the acceptance threshold ε, this metric pri-

marily influences confidence faithfulness. An excessively large value (> 0.3) compromises faithfulness, while an overly small value (< 0.05) increases the difficulty of sampling valid samples. Hence, it is recommended to choose ε within the range $[0.05, 0.2]$. A value of $\varepsilon = 0.1$ is sufficient for MHE. The discussion above implicitly includes the Softmax operation.

β is employed to prevent the sampling process from halting prematurely due to an excessively small dis_i. It is typically set to a small value, with a suggested range of $[0.05, 0.2]$.

The mask size affects the sampling space. Within a budget of fewer than 4000 iterations, a size not exceeding 9 is recommended. For MHE-pro (with 1000 iterations per scale), a size not exceeding 14 is suggested.

Regarding the number of explanation iterations, MHE with 1000 iterations performs adequately for simple scenarios. For more challenging scenarios (hard to sample valid instances), 4000 iterations are sufficient.

The default configurations are suitable for most scenarios: PNN ($b = 0.1, a = 0.5$) and MHE ($\varepsilon = 0.1, \beta = 0.1$, mask size (8,8), 4000 iterations).

B.2. Analysis on Metrics

The Del metric commences by progressively removing pixels from the explanation area starting with the highest importance score, subsequently computes the variation curve of class confidence scores during this removal process, and ultimately calculates the area under this curve. In contrast, the Ins metric continuously inserts the explanation regions, also starting with the highest importance score.

$$PG = \frac{\#\text{Hits}}{\#\text{Hits} + \#\text{Misses}}, \tag{1}$$

$$EBPG = \frac{\sum L_{(i,j)\in \text{bbox}}}{\sum L_{(i,j)\in \text{bbox}} + \sum L_{(i,j)\notin \text{bbox}}} \tag{2}$$

Here, #Hits denotes the number of images where the location corresponding to the top-1 importance score in the explanation result falls within a predefined range, while #Misses represents the number of images where it does not. $L^c(i, j)$ signifies the importance score of the pixel at coordinate (i, j) in the explanation map for class c.

DEL and INS emphasize whether the region possessing the highest importance score corresponds to positively contributing features; if so, these metrics yield superior values. PG assesses whether the region with the top-1 importance score falls within a predefined area. EBPG measures the proportion of the explanatory region that lies within a predefined area. In our experiments, the predefined areas for PG and EBPG were the bounding boxes provided in the respective datasets.

$$\text{Average Drop} = \sum_{i=1}^{N} \frac{\max(0, Y_i - O_i)}{Y_i} \times 100, \tag{3}$$

$$\text{Average Increase} = \sum_{i=1}^{N} \frac{Sign(Y_i < O_i)}{N} \times 100 \tag{4}$$

Here, Y_i^c refers to the confidence score of the original image, O_i^c corresponds to the confidence score derived from the explanation map, and N indicates the total number of images.

The AD metric encourages explanation confidence scores that are greater than or equal to the original score, or at least not significantly lower. The AI metric solely evaluates whether the explanation's confidence score exceeds the original score. Consequently, both AD and AI inherently favor explanations that emphasize positively contributing regions S^{C+}. Given that the confidence score when explaining S^C typically fluctuates around the original confidence, the AI metric is less suitable for evaluating S^C explanations. Conversely, the dis metric is employed to assess the discrepancy between the explanation's and the original confidence scores, defined as the absolute difference between these two scores.

B.3. Analysis on PNN Metrics

For the PNN metric, an increase in P_{dec}^c indicates a stronger positive contribution effect resulting from masking the region. Conversely, an increase in N_{inc}^c signifies a stronger negative contribution effect resulting from masking the region. When N_{mid}^c is close to 0, it suggests that the masked region exerts weak effects regarding both positive and negative contributions. A desirable scenario is one where both P_{dec}^c and N_{inc}^c are high, while N_{mid}^c approximates 0. If the explanation $S^{c'}$ assigns importance scores greater than a to a significant number of truly negative regions, P_{dec}^c will decrease, potentially becoming negative. Similarly, if many positive regions are misclassified as negative, masking these regions would lead to a reduced N_{inc}^c due to an insufficient positive contributions. Conversely, the removal or addition of neutral contribution regions should ideally not affect the detection of class c.

Due to its reliance on manually set thresholds for a simple distinction among the three contribution regions, the PNN metric is primarily suitable for explanation methods that distinctly map the three region types to corresponding importance intervals. For instance, RISE, which uses confidence scores as weights, tends to assign lower importance scores to low-confidence samples containing more negatively contributing regions, thus making it suitable for this metric. However, when explaining the complete contribution region S^c, which comprehensively considers both positive and negative regions, there are scenarios where some

negatively contributing regions might be assigned importance scores similar to positive ones. In such cases, the metric may not be entirely appropriate.

B.4. Runtime

The average time required to explain one image on a V100 GPU is shown in Tab. 1, RISE and Sobol support setting batch size. The slightly longer time for MHE stems from its inability to parallelize the evaluation of an arbitrary number of mask samples like RISE (due to the Markov chain in MHE), and the fact that MHE does not incorporate MCMC-related parallel optimizations by default.

Table 1. The average time required to explain one image (in seconds).

RISE (batch1)	RISE (b100)	Sobol (b1)	Sobol (b100)	MHE	GradCAM
50.792	4.161	79.204	2.412	61.518	0.116

As an MCMC process, potential optimizations for MHE include: (1) MHE is inherently parallelizable, which could significantly reduce its runtime. For instance, instead of generating only one test mask sample per iteration and attempting to accept/reject/transition once, the process can be adjusted to generate and evaluate k test mask samples per iteration, accept l ($l \leq k$) samples (for computing the explanation map), but transition only to the best among the l samples for the subsequent acceptance process. This optimization strategy is analogous to RISE, leveraging the parallel computing capability of GPUs to evaluate multiple test samples concurrently. (2) Classic MCMC optimizations, such as using momentum to accelerate convergence. For example, the mask generation scheme inherent to MHE (adaptive mask mutation probability) serves a similar purpose as momentum optimization: both aim to rapidly skip mask samples with low confidence faithfulness and primarily search for mask samples with medium or high confidence faithfulness.

B.5. Verification Experiment for Assumptions

To validate Assumption 1, which posits that samples with high confidence play a crucial role in explaining positive contribution regions (S^{c+}), we partitioned all samples generated during the MHE sampling process into five disjoint intervals based on their classification confidence. An explanation map was then computed for each interval, and the positive effect of each resulting map was evaluated. A trend where explanation maps derived from higher-confidence intervals exhibit stronger positive contribution would validate Assumption 1. As presented in Table 2, as the confidence associated with the explanation map increases, its positive contribution exhibits an overall increasing trend.

Table 2. Verifying the positive effect of features corresponding to different confidence on classification.

Confidence Interval	[0,0.2]	[0.2,0.4]	[0.4,0.6]	[0.6,0.8]	[0.8,1]
Del↓	0.142	0.149	0.135	0.138	0.092
INS↑	0.663	0.643	0.662	0.678	0.747
AD↓	0.146	0.126	0.107	0.092	0.066
AI↑	0.260	0.260	0.312	0.333	0.510

To validate Assumption 2, which posits that samples with confidence close to the original image's confidence are essential for confidence faithfulness (S^c), we partitioned all samples into six disjoint intervals based on their confidence distance dis_i. If samples with smaller confidence differences yield explanation maps whose classification confidence more closely resembles that of the original image, Hypothesis 2 would be confirmed. As shown in Table 3, as the confidence disparity increases, confidence faithfulness decreases markedly.

Table 3. Verifying the effect of features corresponding to various confidence difference on confidence faithfulness.

dis_i Interval	[0,0.1]	[0.1,0.3]	[0.3,0.5]	[0.5,0.7]	[0.7,0.9]	[0.9,1]
$dis \downarrow$	0.069	0.096	0.143	0.220	0.350	0.368

C. Qualitative Experiment Analysis

In Fig. 3, the CAM series exhibits higher confidence scores on ResNet50 and VGG16, suggesting a primary focus on positively contributing regions. Their performance on the other three models is unstable, which we attribute to the requirement that the model's output and the final layer weights maintain a linear relationship. RISE produces confidence scores closer to the original on VGG16 and ViT but tends to explain positively contributing regions in other scenarios. Excluding VGG16 and ViT, MHE generates values closer to the original than RISE. Both RISE and MHE-e overlook most negatively contributing regions on DINO, leading to inflated explanation scores. MHE-pro yields excessively low confidence on ResNet50 and VGG16; we analyze this might be caused by insufficient search, missing some key positive contributions, as the explanation maps from MHE-pro lack highlighting in the bird's back region. An interesting observation is that explanation methods focusing more on positive regions, like RISE and MHE-e, produce seemingly random, extensive highlighted areas on CLIP. Our analysis suggests this might stem from CLIP's enhanced fault tolerance towards images, facilitated by its

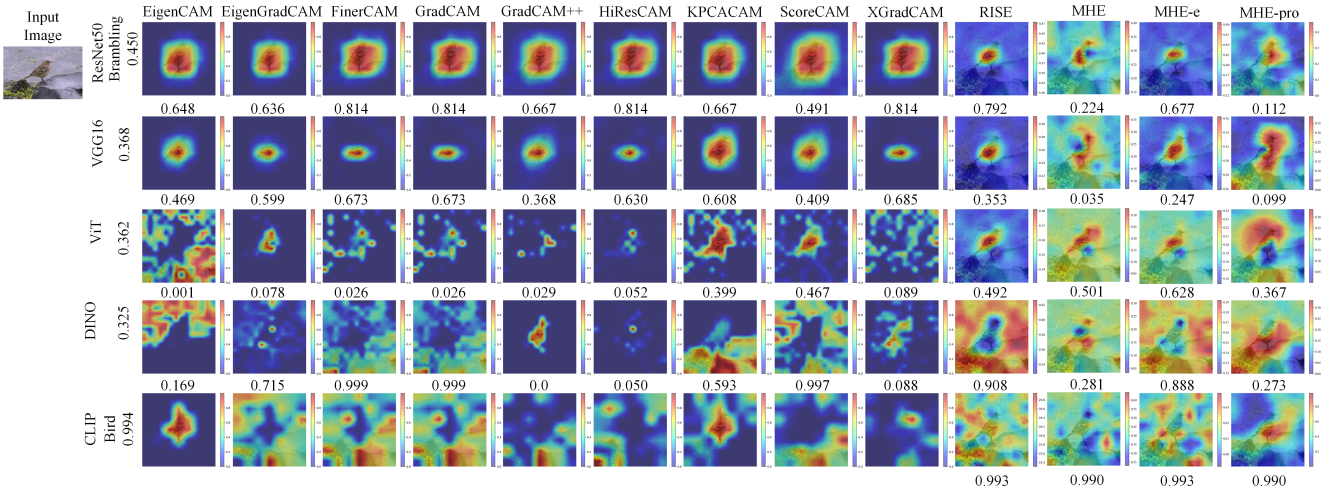


Figure 3. Explanation maps for various explanation methods are analyzed. The CAM series demonstrates stable explanatory capability only on ResNet50 and VGG16. Explanation methods such as RISE and MHE-e, which focus more on positively contributing regions, are prone to being misled by the powerful detection capability of CLIP, consequently highlighting numerous and extensive areas. In contrast, MHE and MHE-pro, with their greater emphasis on explaining the original confidence scores, produce fewer highlighted regions on CLIP, resulting in superior readability.

text input, a phenomenon also reflected in the DEL and INS experiments on CLIP in Sec. D.1. This fault tolerance in CLIP causes different masked image patches to maintain high confidence scores during the iteration process of RISE and MHE-e, misleading them to highlight numerous areas. For CLIP, MHE produces fewer highlighted regions, thereby improving readability, a characteristic further enhanced by MHE-pro's multi-scale search. Since MHE and MHE-pro emphasize maintaining confidence scores similar to the original, they are less susceptible to being misled by CLIP's high-confidence detections.

As shown in Fig. 4, the interpretation quality of the CAM series exhibits significant instability across different models. RISE demonstrates consistent capability in explaining target classes; however, in most scenarios, it fails to maintain faithfulness to the original confidence scores. MHE outperforms RISE by reflecting the original confidence scores more accurately, while simultaneously maintaining compact highlighted regions. MHE-pro further enhances this capability through multi-scale search. In most cases, MHE-e effectively focuses on explaining regions with only positive contributions.

The qualitative results above indicate that negative features are not consistent across images and may even be absent for a given object (e.g., when both MHE and MHE-e yield high and similar classification confidence). As shown in Fig. 3 and Fig. 4, negative regions are not necessarily meaningful, as the model may capture non-semantic information (e.g., texture).

D. Quantitative Experiment Analysis

D.1. ImageNet

The performance on the Del and Ins metrics for the ImageNet dataset is shown in Tab. 4. All explanation methods exhibit anomalous behavior on CLIP, with both Del and Ins scores being excessively high. The direct cause is that the class confidence score does not change significantly when the regions of highest importance in the explanation are either inserted or deleted. We hypothesize that this is likely due to CLIP's dual capability in text and visual understanding, allowing it to compensate for masked positive visual information via its textual modality. Consequently, in deletion experiments, the model can still identify the class even after removing parts of the image, leading to inflated Del scores. This implies that for explaining multimodal models like CLIP, reliance solely on high confidence scores is insufficient—a characteristic inherent to MHE. Across most scenarios, RISE achieves the best performance. The performance of MHE and MHE-pro differs significantly from RISE, as they prioritize explaining the original confidence score, leading them to assign high importance to both significant positive and negative contribution regions. Since MHE-e focuses specifically on positive contribution regions, its performance gap with RISE remains relatively small.

The performance on the PG and EBPG metrics for the ImageNet dataset is presented in Tab. 5. In most scenarios, the localization capability of the CAM series is significantly superior. The MHE series demonstrates an advantage only

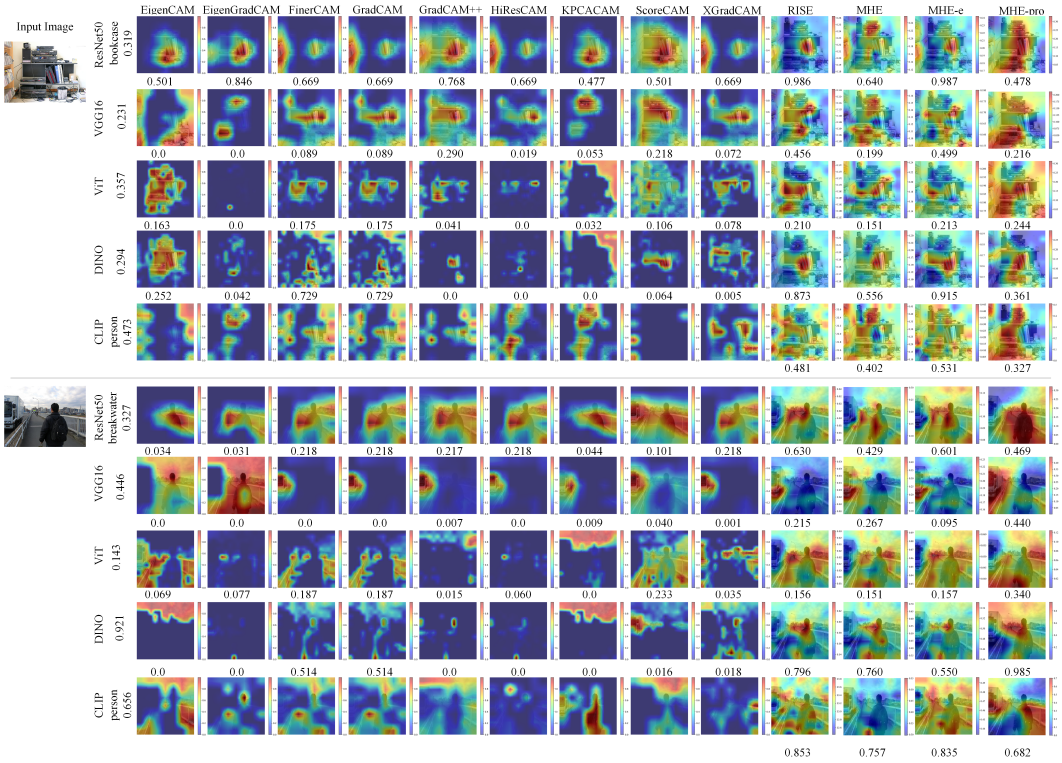


Figure 4. Supplement to qualitative experiments.

Table 4. The DEL and INS performance on ImageNet.

Method	VGG16		ResNet50		VIT		DINOv1		CLIP	
	Del↓	Ins↑	Del↓	Ins↑	Del↓	Ins↑	Del↓	Ins↑	Del↓	Ins↑
GradCAM(2017)	0.084	0.636	0.112	0.706	0.306	0.523	0.145	0.376	23.748	23.437
GradCAM++(2018)	0.099	0.636	0.122	0.718	0.167	0.637	**0.091**	0.475	24.237	23.075
ScoreCAM(2020)	0.103	0.642	0.122	0.718	0.262	0.561	0.167	0.340	24.025	23.216
HiResCAM(2020)	0.096	0.600	0.137	0.670	0.229	0.569	0.158	0.351	23.704	23.133
EigenCAM(2020)	0.367	0.337	0.141	0.666	0.269	0.529	0.183	0.334	23.667	23.217
EigenGradCAM	0.272	0.433	0.127	0.687	0.261	0.583	0.160	0.346	23.917	23.202
XGradCAM(2020)	0.105	0.633	0.122	0.718	**0.117**	0.654	0.106	0.485	23.775	23.548
KPCACAM(2024)	0.087	0.607	0.123	0.684	0.197	0.612	0.134	0.457	23.612	23.522
FinerCAM(2025)	0.105	0.633	0.122	0.718	**0.117**	0.654	0.106	0.485	23.776	23.549
RISE(2018)	**0.070**	**0.688**	**0.085**	**0.755**	0.184	**0.700**	0.102	**0.588**	0.347	0.575
Sobol(2021)	0.066	0.590	0.083	0.670	0.164	0.625	0.118	0.399	0.372	0.514
MHE	0.086	0.628	0.107	0.700	0.203	0.663	0.149	0.453	0.329	0.616
MHE-e	0.074	0.661	0.093	0.732	0.207	0.685	0.111	0.570	**0.307**	**0.670**
MHE-pro	0.130	0.643	0.167	0.712	0.250	0.653	0.175	0.454	0.362	0.616

in a minority of cases, such as with the CLIP model. Consistent with the discussion regarding Del and Ins, MHE and MHE-pro consider both positive and negative contribution regions. Negative regions have a higher probability of lying outside the bounding boxes defined for PG and EBPG, resulting in their suboptimal performance on these metrics.

In summary, on the Del and Ins metrics, which emphasize positive contribution regions, RISE performs best, with MHE-e demonstrating comparable effectiveness. On the PG and EBPG metrics, which assess whether explanations fall within specific predefined regions, the CAM series holds a distinct advantage.

Table 5. The PG and EBPG performance on ImageNet.

Method	VGG16		ResNet50		VIT		DINOv1		CLIP	
	PG↑	EBPG↑	PG↑	EBPG↑	PG↑	EBPG↑	PG↑	EBPG↑	PG↑	EBPG↑
GradCAM(2017)	0.431	0.419	0.476	0.445	0.384	0.392	0.400	0.411	0.373	0.363
GradCAM++(2018)	0.418	0.403	0.472	0.440	0.431	0.425	0.435	**0.428**	0.407	0.403
ScoreCAM(2020)	0.436	0.407	0.472	0.440	0.395	0.398	0.409	0.405	0.401	0.403
HiResCAM(2020)	0.440	**0.433**	0.488	0.465	0.382	0.403	0.427	0.414	0.413	0.410
EigenCAM(2020)	0.371	0.385	0.496	**0.474**	0.377	0.398	0.413	0.408	0.367	0.391
EigenGradCAM	0.397	0.405	0.477	0.469	0.416	0.399	0.436	0.420	0.412	0.407
XGradCAM(2020)	**0.466**	0.429	**0.507**	0.462	**0.453**	**0.451**	0.411	0.416	0.390	0.397
KPCACAM(2024)	0.420	0.420	0.487	0.444	0.392	0.405	0.402	0.404	0.398	0.397
FinerCAM(2025)	0.438	0.410	0.472	0.440	0.424	0.427	0.394	0.399	0.384	0.387
RISE(2018)	0.427	0.413	0.431	0.413	0.332	0.413	0.430	0.413	0.118	0.408
Sobol(2021)	0.438	0.429	0.423	0.429	0.435	0.430	0.482	0.445	0.428	0.407
MHE	0.421	0.396	0.423	0.389	0.420	0.410	0.432	0.391	0.417	0.412
MHE-e	0.440	0.413	0.442	0.413	0.420	0.413	**0.459**	0.413	**0.432**	0.412
MHE-pro	0.417	0.422	0.424	0.421	0.414	0.419	0.431	0.423	0.404	**0.415**

For detailed data, please refer to Tabs. 6 and 7. The CAM series demonstrates superior performance in the P^c_{dec} metric, indicating its enhanced capability in identifying regions of positive contribution. Conversely, the MHE framework exhibits more robust results in the PNN, N^c_{inc}, and N^c_{mid} metrics, suggesting its proficiency in simultaneously capturing all three region types.

D.2. CUB-200-2011

For detailed data, please refer to Tabs. 8 to 12. Overall, the performance of various explanation methods on the CUB dataset is largely consistent with the observations made on the ImageNet dataset.

For the Del and Ins metrics on the CUB dataset, the trends resemble those on the ImageNet dataset but differ from the performance on VOC. RISE performs excellently in most scenarios. The MHE series only shows favorable results on the Ins metric of the CLIP model.

For the PG and EBPG metrics on the CUB dataset, KPCACAM delivers superior performance in the majority of scenarios.

D.3. VOC2012

For detailed data, please refer to Tabs. 13 to 17. Overall, the performance of various explanation methods on the VOC dataset is largely consistent with the observations made on the ImageNet dataset.

For the AD, AI, and dis metrics on the VOC dataset, the CAM series achieves the best performance specifically on the CLIP model. MHE-e and MHE-pro attain the top results in most scenarios, respectively. RISE also reaches optimal performance in certain scenarios.

For the Del and Ins metrics on the VOC dataset, black-box approaches dominate, with RISE and the MHE series each exhibiting the best performance in nearly half of the scenarios.

For the PG and EBPG metrics on the VOC dataset, the CAM series maintains an advantage in localization capability. Specifically, FinerCAM and GradCAM demonstrate consistent performance, which is expected since FinerCAM can be enhanced based on other CAM methods, with GradCAM selected as its foundation here.

Table 6. The full PNN performance on ImageNet (VGG16, ResNet50 and VIT).

Method	VGG16				ResNet50				VIT			
	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓
GradCAM(2017)	0.595	0.690	-0.515	0.203	0.667	0.779	-0.314	0.129	0.482	0.515	-0.674	0.430
GradCAM++(2018)	0.543	0.688	-0.639	0.334	0.633	0.763	-0.421	0.178	0.440	0.506	-0.730	0.576
ScoreCAM(2020)	0.559	0.690	-0.601	0.293	0.633	0.763	-0.421	0.178	0.458	0.603	-0.730	0.583
HiResCAM(2020)	0.568	0.733	-0.531	0.362	0.587	0.787	-0.515	0.338	0.558	0.724	-0.517	0.414
EigenCAM(2020)	0.483	**0.747**	-0.702	0.629	0.575	0.784	-0.548	0.363	0.542	**0.726**	-0.550	0.466
EigenGradCAM	0.500	0.741	-0.680	0.563	0.576	0.780	-0.559	0.342	0.468	0.616	-0.707	0.569
XGradCAM(2020)	0.553	0.694	-0.609	0.321	0.633	0.763	-0.421	0.178	0.618	0.668	-0.379	0.201
FinerCAM(2025)	0.553	0.694	-0.609	0.321	0.633	0.763	-0.421	0.178	0.618	0.668	-0.379	0.201
KPCACAM(2024)	0.610	0.701	-0.470	0.182	0.685	**0.789**	-0.250	0.113	0.621	0.594	-0.345	0.144
RISE(2018)	0.631	0.713	-0.423	0.134	0.630	0.762	-0.491	0.123	**0.648**	0.559	-0.221	0.097
Sobol(2021)	0.510	0.662	-0.706	0.407	0.503	0.670	-0.761	0.394	0.491	0.275	-0.647	0.172
MHE	0.619	0.696	-0.475	0.123	0.622	0.720	-0.532	0.079	0.616	0.455	-0.295	0.079
MHE-e	0.616	0.707	-0.496	0.130	0.614	0.743	-0.561	0.113	0.610	0.458	-0.306	0.102
MHE-pro	**0.697**	0.715	**-0.162**	**0.070**	**0.693**	0.746	**-0.218**	**0.064**	0.642	0.460	**-0.184**	**0.067**

Table 7. The full PNN performance on ImageNet (DINOv1 and CLIP).

Method	DINOv1				CLIP			
	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓
GradCAM(2017)	0.503	0.463	-0.490	0.460	0.116	-1.623	1.152	1.950
GradCAM++(2018)	0.493	0.478	-0.506	0.509	0.155	-1.379	0.928	1.776
ScoreCAM(2020)	0.499	0.443	-0.494	0.452	0.114	-1.584	1.128	1.976
HiResCAM(2020)	0.545	**0.480**	-0.385	0.371	0.220	-1.489	1.292	1.702
EigenCAM(2020)	0.533	0.476	-0.410	0.402	0.172	-1.540	1.207	1.804
EigenGradCAM	0.493	0.460	-0.503	0.492	0.133	-1.529	1.202	2.008
XGradCAM(2020)	0.506	0.392	-0.464	0.400	0.180	-1.865	1.372	1.606
FinerCAM(2025)	0.506	0.392	-0.464	0.400	0.180	-1.869	1.374	1.607
KPCACAM(2024)	0.512	0.431	-0.448	0.423	0.140	-1.595	1.479	2.185
RISE(2018)	0.610	0.427	-0.129	0.246	0.628	**0.375**	-0.143	0.090
Sobol(2021)	0.505	0.344	-0.459	0.357	0.603	0.336	-0.213	0.110
MHE	0.611	0.405	-0.206	0.146	0.609	0.294	-0.179	0.071
MHE-e	0.605	0.417	-0.155	0.237	0.613	0.316	-0.142	0.110
MHE-pro	**0.640**	0.419	**-0.096**	**0.125**	**0.629**	0.293	**-0.083**	**0.064**

Table 8. The AD, AI and dis performance on CUB.

Method	VGG16			ResNet50			VIT			DINOv1			CLIP		
	AD↓	AI↑	dis↓	AD↓	AI↑	dis↓	AD↓	AI↑	dis↓	AD↓	AI↑	dis↓	AD↓	AI↑	dis↓
GradCAM(2017)	0.516	0.191	—	0.367	0.284	0.237	0.406	0.263	0.246	0.629	0.234	0.461	0.034	0.375	—
GradCAM++(2018)	0.370	0.265	0.250	0.292	0.297	0.192	0.670	0.094	—	0.811	0.106	—	0.067	0.189	—
ScoreCAM(2020)	0.355	0.253	0.234	0.247	0.317	0.163	0.252	0.401	0.208	0.732	0.153	—	0.026	0.625	1.722
HiResCAM(2020)	0.528	0.194	—	0.367	0.284	0.237	0.582	0.154	0.319	0.920	0.038	0.561	0.076	0.140	—
EigenCAM(2020)	0.843	0.038	0.531	0.466	0.188	0.273	0.656	0.091	0.358	0.691	0.117	0.427	0.029	0.619	1.687
EigenGradCAM	0.753	0.078	0.479	0.454	0.202	0.273	0.754	0.077	0.411	0.914	0.038	0.553	0.033	0.403	1.354
XGradCAM(2020)	0.506	0.205	—	0.367	0.284	0.237	0.701	0.098	0.397	0.747	0.116	0.464	0.036	0.480	1.617
KPCACAM(2024)	0.464	0.187	0.288	0.437	0.205	0.257	0.465	0.159	0.261	0.555	0.153	0.341	0.013	**0.724**	1.330
FinerCAM(2025)	0.516	0.191	—	0.367	0.284	0.237	0.406	0.263	0.246	0.629	0.234	0.461	0.034	0.375	—
RISE(2018)	0.102	0.655	0.222	0.122	0.635	0.233	**0.082**	**0.660**	0.188	**0.080**	0.767	0.244	0.003	0.555	**0.008**
Sobol(2021)	0.561	0.155	0.330	0.536	0.175	0.333	0.345	0.230	0.201	0.446	0.265	0.284	0.005	0.540	0.011
MHE	0.141	0.534	0.136	0.138	0.522	0.134	0.152	0.471	0.122	0.112	0.637	0.135	0.003	0.574	**0.008**
MHE-e	0.091	**0.662**	0.213	0.099	**0.664**	0.225	0.094	0.618	0.178	0.062	**0.801**	0.236	**0.002**	0.624	0.009
MHE-pro	**0.075**	0.594	**0.097**	**0.092**	0.558	**0.105**	0.124	0.485	**0.100**	**0.080**	0.670	**0.113**	0.004	0.500	0.009

Table 9. The DEL and INS performance on CUB.

Method	VGG16		ResNet50		VIT		DINOv1		CLIP	
	Del↓	Ins↑	Del↓	Ins↑	Del↓	Ins↑	Del↓	Ins↑	Del↓	Ins↑
GradCAM(2017)	0.041	0.575	0.043	0.529	0.053	0.471	0.142	0.526	27.319	26.811
GradCAM++(2018)	**0.026**	0.549	0.036	0.505	0.173	0.349	0.147	0.449	26.862	26.783
ScoreCAM(2020)	0.029	0.530	0.040	0.489	0.092	0.443	0.104	0.546	25.844	27.123
HiResCAM(2020)	0.036	0.575	0.043	0.529	0.056	0.460	**0.079**	0.575	27.303	26.639
EigenCAM(2020)	0.283	0.221	0.046	0.479	0.128	0.358	0.182	0.408	25.661	26.884
EigenGradCAM	0.214	0.310	0.042	0.488	0.119	0.402	0.127	0.436	27.226	26.681
XGradCAM(2020)	0.039	0.582	0.043	0.529	0.134	0.384	0.166	0.412	26.445	26.855
KPCACAM(2024)	0.029	0.521	0.043	0.481	0.077	0.422	0.122	0.455	25.237	26.941
FinerCAM(2025)	0.041	0.575	0.043	0.529	0.053	0.471	0.142	0.526	27.319	26.811
RISE(2018)	**0.026**	**0.653**	**0.029**	**0.615**	**0.070**	**0.568**	0.090	**0.677**	0.631	0.955
Sobol(2021)	0.022	0.513	0.030	0.506	0.058	0.440	0.081	0.548	0.630	0.938
MHE	0.028	0.538	0.039	0.509	0.088	0.477	0.106	0.544	0.631	**0.965**
MHE-e	0.028	0.636	0.037	0.597	0.084	0.563	0.096	0.676	**0.621**	**0.965**
MHE-pro	0.048	0.546	0.057	0.509	0.107	0.456	0.125	0.548	0.659	**0.965**

Table 10. The PG and EBPG performance on CUB.

Method	VGG16		ResNet50		VIT		DINOv1		CLIP	
	PG↑	EBPG↑	PG↑	EBPG↑	PG↑	EBPG↑	PG↑	EBPG↑	PG↑	EBPG↑
GradCAM(2017)	0.332	0.298	0.393	0.343	0.350	0.330	0.298	0.279	0.277	0.268
GradCAM++(2018)	0.349	0.323	0.400	0.344	0.214	0.250	0.287	0.301	0.247	0.219
ScoreCAM(2020)	0.359	0.323	0.383	0.337	0.323	0.295	**0.309**	0.305	**0.350**	0.322
HiResCAM(2020)	0.331	0.288	0.393	0.343	0.355	**0.348**	0.299	0.330	0.277	0.273
EigenCAM(2020)	0.257	0.263	0.404	0.373	0.291	0.286	0.321	0.295	0.329	0.321
EigenGradCAM	0.264	0.288	**0.405**	**0.376**	0.297	0.270	0.307	0.308	0.273	0.262
XGradCAM(2020)	0.333	0.295	0.393	0.343	0.283	0.278	0.340	0.288	0.290	0.287
KPCACAM(2024)	**0.362**	**0.345**	0.416	0.374	**0.364**	0.326	0.360	**0.333**	0.377	**0.343**
FinerCAM(2025)	0.332	0.298	0.393	0.343	0.350	0.330	0.298	0.279	0.277	0.268
RISE(2018)	0.307	0.274	0.314	0.274	0.294	0.273	0.332	0.274	0.012	0.272
Sobol(2021)	0.295	0.336	0.380	0.350	0.305	0.353	0.390	0.384	0.200	0.305
MHE	0.307	0.273	0.317	0.274	0.299	0.274	0.341	0.272	0.257	0.271
MHE-e	0.305	0.273	0.317	0.273	0.313	0.273	0.332	0.273	0.283	0.273
MHE-pro	0.332	0.291	0.328	0.287	0.329	0.287	0.354	0.291	0.241	0.267

Table 11. The full PNN performance on CUB (VGG16, ResNet50 and VIT).

Method	VGG16				ResNet50				VIT			
	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓
GradCAM(2017)	0.545	0.616	-0.523	0.370	0.595	0.580	-0.354	0.251	0.549	0.414	-0.408	0.260
GradCAM++(2018)	0.590	0.618	-0.388	0.282	0.630	0.588	-0.239	0.200	0.503	0.414	-0.503	0.394
ScoreCAM(2020)	0.600	0.621	-0.360	0.259	0.647	**0.590**	-0.186	0.167	0.576	0.408	-0.313	0.214
HiResCAM(2020)	0.537	0.615	-0.544	0.387	0.595	0.580	-0.354	0.251	0.497	0.382	-0.510	0.387
EigenCAM(2020)	0.501	**0.629**	-0.584	0.541	0.583	0.586	-0.377	0.293	0.550	**0.516**	-0.411	0.353
EigenGradCAM	0.513	0.627	-0.562	0.499	0.580	0.587	-0.394	0.293	0.500	0.443	-0.506	0.435
XGradCAM(2020)	0.547	0.617	-0.516	0.367	0.595	0.580	-0.354	0.251	0.502	0.433	-0.508	0.415
FinerCAM(2025)	0.545	0.616	-0.523	0.370	0.595	0.580	-0.354	0.251	0.549	0.414	-0.408	0.260
KPCACAM(2024)	0.587	0.625	-0.373	0.318	0.595	0.587	-0.338	0.274	0.585	0.509	-0.341	0.243
RISE(2018)	0.600	0.606	-0.383	0.224	0.583	0.565	-0.414	0.236	0.603	0.382	-0.176	0.190
Sobol(2021)	0.543	0.603	-0.516	0.373	0.532	0.577	-0.542	0.377	0.537	0.300	-0.390	0.223
MHE	0.613	0.598	-0.395	0.138	0.602	0.554	-0.412	0.134	0.588	0.336	-0.273	0.123
MHE-e	0.596	0.594	-0.399	0.216	0.583	0.549	-0.406	0.228	0.593	0.356	-0.212	0.179
MHE-pro	**0.678**	0.616	**-0.127**	**0.099**	**0.662**	0.577	**-0.157**	**0.109**	**0.629**	0.393	**-0.145**	**0.102**

Table 12. The full PNN performance on CUB (DINOv1 and CLIP).

Method	DINOv1				CLIP			
	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓
GradCAM(2017)	0.493	0.472	-0.529	0.480	-0.025	0.147	-1.721	1.553
GradCAM++(2018)	0.488	0.524	-0.548	0.534	-0.023	1.273	-2.161	2.228
ScoreCAM(2020)	0.493	0.511	-0.538	0.509	0.254	0.330	-0.192	1.869
HiResCAM(2020)	0.480	0.552	-0.581	0.573	-0.016	1.492	-2.142	2.429
EigenCAM(2020)	0.528	**0.557**	-0.479	0.437	0.571	1.294	0.496	1.934
EigenGradCAM	0.478	0.542	-0.581	0.569	0.024	-0.027	-1.330	1.524
XGradCAM(2020)	0.484	0.482	-0.565	0.498	0.086	0.491	-1.148	1.915
FinerCAM(2025)	0.493	0.472	-0.529	0.480	-0.025	0.147	-1.721	1.553
KPCACAM(2024)	0.554	0.546	-0.420	0.358	0.833	1.427	1.315	1.575
RISE(2018)	0.592	0.457	-0.250	0.246	**0.676**	**0.412**	**-0.022**	**0.008**
Sobol(2021)	0.502	0.386	-0.532	0.342	0.623	0.303	-0.176	0.011
MHE	0.591	0.432	-0.342	0.137	0.624	0.251	-0.124	**0.008**
MHE-e	0.589	0.443	-0.261	0.239	0.639	0.315	-0.111	0.009
MHE-pro	**0.660**	0.507	**-0.092**	**0.114**	0.620	0.249	-0.139	0.009

Table 13. The AD, AI and dis performance on VOC.

Method	VGG16			ResNet50			VIT			DINOv1			CLIP		
	AD↓	AI↑	dis↓	AD↓	AI↑	dis↓	AD↓	AI↑	dis↓	AD↓	AI↑	dis↓	AD↓	AI↑	dis↓
GradCAM(2017)	0.550	0.218	0.248	0.416	0.303	0.221	0.522	0.230	0.247	0.627	0.185	0.418	0.008	0.757	1.353
GradCAM++(2018)	0.458	0.264	0.203	0.354	0.321	0.188	0.747	0.097	—	0.874	0.049	0.524	**0.007**	0.768	—
ScoreCAM(2020)	0.513	0.195	0.203	0.384	0.286	0.183	0.273	0.440	0.187	0.727	0.126	0.467	0.004	**0.889**	2.001
HiResCAM(2020)	0.548	0.224	0.248	0.416	0.303	0.221	0.799	0.072	0.351	0.951	0.018	0.559	0.015	0.673	—
EigenCAM(2020)	0.863	0.046	0.342	0.720	0.102	0.312	0.738	0.104	0.327	0.820	0.061	0.487	0.011	0.751	1.450
EigenGradCAM	0.826	0.056	0.327	0.649	0.148	0.298	0.868	0.041	0.378	0.946	0.022	0.557	0.008	0.822	1.775
XGradCAM(2020)	0.523	0.237	0.239	0.416	0.303	0.221	0.780	0.070	0.343	0.872	0.041	0.517	0.008	0.791	1.510
KPCACAM(2024)	0.709	0.116	0.278	0.682	0.125	0.296	0.740	0.093	0.321	0.850	0.051	0.508	0.010	0.751	1.385
FinerCAM(2025)	0.550	0.218	0.248	0.416	0.303	0.221	0.522	0.230	0.247	0.627	0.185	0.418	0.008	0.757	1.353
RISE(2018)	0.086	0.750	0.238	**0.099**	0.729	0.250	0.181	0.547	0.178	**0.066**	**0.768**	0.241	0.070	0.477	0.079
Sobol(2021)	0.715	0.105	0.268	0.704	0.080	0.295	0.485	0.18	0.200	0.666	0.115	0.396	0.082	0.345	0.073
MHE	0.139	0.636	0.150	0.139	0.608	0.146	0.215	0.431	0.130	0.121	0.578	0.141	0.075	0.442	0.071
MHE-e	**0.060**	**0.803**	0.229	0.074	**0.768**	0.241	0.177	**0.562**	0.177	0.058	0.765	0.230	0.017	0.747	0.094
MHE-pro	0.100	0.672	**0.116**	0.103	0.664	**0.125**	**0.162**	0.484	**0.104**	0.093	0.650	**0.125**	0.062	0.436	**0.061**

Table 14. The DEL and INS performance on VOC.

Method	VGG16		ResNet50		VIT		DINOv1		CLIP	
	Del↓	Ins↑	Del↓	Ins↑	Del↓	Ins↑	Del↓	Ins↑	Del↓	Ins↑
GradCAM(2017)	0.029	0.321	0.033	0.378	**0.034**	0.329	0.101	0.550	23.366	23.363
GradCAM++(2018)	0.027	0.289	0.032	0.346	0.093	0.228	0.141	0.380	23.316	23.228
ScoreCAM(2020)	0.036	0.247	0.042	0.307	0.066	0.285	0.138	0.461	23.369	23.279
HiResCAM(2020)	0.028	0.328	0.033	0.378	0.042	0.319	**0.076**	0.504	23.895	22.806
EigenCAM(2020)	0.132	0.123	0.054	0.290	0.104	0.208	0.191	0.317	23.427	22.922
EigenGradCAM	0.100	0.169	0.044	0.318	0.083	0.230	0.158	0.365	23.660	22.951
XGradCAM(2020)	0.028	0.330	0.033	0.378	0.091	0.222	0.171	0.347	23.704	22.968
KPCACAM(2024)	0.041	0.238	0.051	0.295	0.097	0.220	0.200	0.308	23.478	22.823
FinerCAM(2025)	0.029	0.321	0.033	0.378	**0.034**	0.329	0.101	0.550	23.366	23.363
RISE(2018)	**0.021**	**0.396**	**0.022**	**0.447**	0.037	**0.408**	0.101	0.647	0.367	0.663
Sobol(2021)	0.022	0.228	0.023	0.289	0.042	0.289	0.122	0.411	0.421	0.641
MHE	0.030	0.305	0.032	0.356	0.053	0.335	0.137	0.514	0.381	0.679
MHE-e	**0.021**	0.389	0.025	0.440	0.044	0.399	0.107	**0.626**	**0.350**	**0.743**
MHE-pro	0.039	0.302	0.043	0.353	0.068	0.320	0.162	0.512	0.431	0.681

Table 15. The PG and EBPG performance on VOC.

Method	VGG16		ResNet50		VIT		DINOv1		CLIP	
	PG↑	EBPG↑	PG↑	EBPG↑	PG↑	EBPG↑	PG↑	EBPG↑	PG↑	EBPG↑
GradCAM(2017)	0.277	0.264	0.316	0.289	**0.284**	0.275	0.249	0.242	0.226	0.241
GradCAM++(2018)	0.291	0.271	0.321	0.295	0.232	0.248	0.243	0.261	0.216	0.225
ScoreCAM(2020)	**0.295**	0.267	0.328	0.294	0.235	0.252	0.245	0.243	0.263	0.251
HiResCAM(2020)	0.276	0.259	0.316	0.289	0.278	**0.276**	0.270	0.272	**0.264**	0.254
EigenCAM(2020)	0.233	0.240	**0.334**	**0.325**	0.250	0.251	0.269	0.264	0.222	0.232
EigenGradCAM	0.260	0.270	0.328	0.320	0.260	0.263	**0.281**	**0.275**	0.233	0.250
XGradCAM(2020)	0.282	0.261	0.316	0.289	0.253	0.254	0.259	0.252	0.252	0.251
KPCACAM(2024)	0.288	**0.285**	0.333	0.321	0.238	0.244	0.219	0.229	0.250	**0.256**
FinerCAM(2025)	0.277	0.264	0.316	0.289	**0.284**	0.275	0.249	0.242	0.226	0.241
RISE(2018)	0.287	0.251	0.274	0.251	0.264	0.250	0.268	0.250	0.004	0.247
Sobol(2021)	0.240	0.249	0.244	0.251	0.241	0.250	0.224	0.253	0.262	0.230
MHE	0.274	0.247	0.271	0.249	0.273	0.250	0.253	0.236	0.259	0.250
MHE-e	0.272	0.250	0.274	0.251	0.265	0.250	0.273	0.250	0.247	0.250
MHE-pro	0.281	0.264	0.278	0.263	0.260	0.257	0.273	0.257	0.239	0.249

Table 16. The full PNN performance on VOC (VGG16, ResNet50 and VIT).

Method	VGG16				ResNet50				VIT			
	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓
GradCAM(2017)	0.547	0.365	-0.357	0.271	0.570	0.424	-0.333	0.241	0.547	0.365	-0.357	0.273
GradCAM++(2018)	0.563	0.363	-0.327	0.221	0.588	0.424	-0.284	0.198	0.519	0.363	-0.418	0.350
ScoreCAM(2020)	0.563	0.363	-0.329	0.217	0.594	0.425	-0.268	0.188	0.584	0.372	-0.258	0.195
HiResCAM(2020)	0.545	0.362	-0.363	0.273	0.570	0.424	-0.333	0.241	0.506	0.362	-0.426	0.407
EigenCAM(2020)	0.533	**0.381**	-0.370	0.348	0.544	**0.431**	-0.383	0.330	0.545	**0.418**	-0.361	0.332
EigenGradCAM	0.534	0.380	-0.369	0.341	0.548	0.429	-0.375	0.316	0.512	0.384	-0.424	0.399
XGradCAM(2020)	0.551	0.364	-0.352	0.259	0.570	0.424	-0.333	0.241	0.516	0.375	-0.424	0.370
FinerCAM(2025)	0.547	0.365	-0.357	0.271	0.570	0.424	-0.333	0.241	0.547	0.365	-0.357	0.273
KPCACAM(2024)	0.548	0.379	-0.341	0.297	0.550	**0.431**	-0.371	0.312	0.547	0.416	-0.352	0.328
RISE(2018)	0.579	0.369	-0.240	0.234	0.571	0.420	-0.317	0.247	0.596	0.358	-0.197	0.180
Sobol(2021)	0.535	0.344	-0.374	0.297	0.530	0.408	-0.422	0.334	0.537	0.292	-0.388	0.219
MHE	0.599	0.374	-0.229	0.151	0.596	0.420	-0.291	0.147	0.595	0.339	-0.235	0.130
MHE-e	0.583	0.371	-0.229	0.229	0.576	0.415	-0.296	0.241	0.592	0.346	-0.210	0.177
MHE-pro	**0.636**	0.379	**-0.084**	**0.117**	**0.641**	0.429	**-0.097**	**0.128**	**0.631**	0.370	**-0.114**	**0.103**

Table 17. The full PNN performance on VOC (DINOv1 and CLIP).

Method	DINOv1				CLIP			
	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓	PNN↑	P_{dec}↑	N_{inc}↑	N_{mid}↓
GradCAM(2017)	0.495	0.481	-0.540	0.465	0.287	-2.203	2.151	1.516
GradCAM++(2018)	0.489	0.546	-0.559	0.541	0.207	-2.158	1.989	1.798
ScoreCAM(2020)	0.495	0.519	-0.536	0.506	0.129	-2.348	2.123	2.132
HiResCAM(2020)	0.483	0.558	-0.572	0.570	0.279	-1.753	1.727	1.578
EigenCAM(2020)	0.509	0.555	-0.512	0.497	0.138	-2.328	1.575	1.556
EigenGradCAM	0.484	0.553	-0.571	0.564	0.199	-2.033	2.018	1.988
XGradCAM(2020)	0.484	0.525	-0.563	0.541	0.214	-2.066	1.917	1.781
FinerCAM(2025)	0.495	0.481	-0.540	0.465	0.287	-2.203	2.151	1.516
KPCACAM(2024)	0.503	**0.559**	-0.528	0.518	0.149	-2.350	1.598	1.500
RISE(2018)	0.612	0.478	-0.175	0.242	**0.626**	**0.384**	-0.173	0.079
Sobol(2021)	0.495	0.462	-0.534	0.455	0.629	0.362	-0.144	0.076
MHE	0.636	0.506	-0.183	0.145	0.605	0.281	-0.183	0.072
MHE-e	0.613	0.483	-0.185	0.231	0.616	0.302	-0.129	0.093
MHE-pro	**0.657**	0.505	**-0.092**	**0.126**	0.617	0.227	**-0.076**	**0.064**