# Comprehensive Information Bottleneck for Unveiling Universal Attribution to Interpret Vision Transformers

## Supplementary Material

| Model | Base model | Large model | Search space |
|---|---|---|---|
| Start Layer | 4 | 8 | {0,1,2,...,23 } |
| End layer | 12 | 24 | {1,2,3,...,24 } |
| Iteration | 10 | | {1,5,10,20} |
| Optimizer | Adam | | {SGD, Adam} |
| Batch size | 10 | | {1,5,10,20} |
| Learning rate | 1 | | {0.1,0.5,1,10} |
| Trade-off parameter $\beta$ | 10 | | {0.1,1,10,100} |

Table A. **Hyperparameters selected in CoIBA**. Except for the start and end layer index, we unify the hyperparameters among the base and large models.

## A. Derivative of Eq. 4

$$
\begin{aligned}
&I[Z_l; Z_{l-1}] \\
&= E_{Z_{l-1}}[D_{KL}[P(Z_l|Z_{l-1})||P(Z_l)]] \\
&= \int_{Z_{l-1}} p(z_{l-1}) \left( \int_{Z_l} p(z_l|z_{l-1}) \log \frac{p(z_l|z_{l-1})}{p(z_l)} dZ_l \right) dZ_{l-1} \\
&= \int_{Z_l} \int_{Z_{l-1}} p(z_l, z_{l-1}) \log \frac{p(z_l|z_{l-1})}{q(z_l)} dZ_l dZ_{l-1} \\
&\quad - \int_{Z_l} \int_{Z_{l-1}} p(z_l, z_{l-1}) \log \frac{p(z_l)}{q(z_l)} \\
&= \int_{Z_l} \int_{Z_{l-1}} p(z_l, z_{l-1}) \log \frac{p(z_l|z_{l-1})}{q(z_l)} dZ_l dZ_{l-1} \\
&\quad - \int_{Z_l} \left( \int_{Z_{l-1}} p(z_{l-1}|z_l) \right) p(z_l) \log \frac{p(z_l)}{q(z_l)} \\
&= E_{Z_{l-1}}[D_{KL}[P(Z_l|Z_{l-1})||Q(Z_l)]] \\
&\quad - D_{KL}[P(Z_l)||Q(Z_l)] \\
&\leq E_{Z_{l-1}}[D_{KL}[P(Z_l|Z_{l-1})||Q(Z_l)]] \, .
\end{aligned}
\tag{A}
$$

## B. Relationship between $I[Y; Z_L]$ and Cross-Entropy Loss

Computation of the cross entropy $H(Y; \hat{Y})$ between the label $Y$ and predicted label $\hat{Y}$ can be considered as the conditional cross entropy $H(Y; \hat{Y}|Z_L)$ because $Z_L$ determines the prediction. Here, ViT utilizes the imputed representation $Z_L$ to predict the label $\hat{Y}$. Such that, the conditional cross entropy can be divided into conditional entropy and KL divergence:

$$
\begin{aligned}
H(Y; \hat{Y}|Z_L) &= H(Y|Z_L) - H(Y) + D_{KL}[Y|Z_L||\hat{Y}|Z_L] \\
&= H(Y|Z_L) + D_{KL}[Y||\hat{Y}|Z_L] \, .
\end{aligned}
\tag{B}
$$

| Model | Setting | Low-confident | | High-confident | |
|---|---|---|---|---|---|
| | | 0-20 | 20-40 | 60-80 | 80-100 |
| ViT-B ($s$) | 2-12 | 1.20/23.86 | 2.57/35.02 | 8.63/56.98 | 17.30/75.66 |
| | 4-12 | 1.27/24.40 | 2.66/33.17 | 8.70/55.48 | 16.83/75.64 |
| | 6-12 | 1.11/21.45 | 2.83/31.99 | 9.61/54.62 | 18.58/75.24 |
| | 8-12 | 1.53/15.44 | 3.48/29.37 | 10.58/51.42 | 22.37/72.65 |
| DeiT-B ($s$) | 2-12 | 0.67/22.08 | 2.44/33.16 | 9.82/54.42 | 17.54/66.12 |
| | 4-12 | 0.62/19.90 | 2.33/34.34 | 9.49/54.49 | 16.63/66.29 |
| | 6-12 | 0.67/18.82 | 2.57/32.14 | 9.14/54.00 | 15.62/66.02 |
| | 8-12 | 0.79/17.47 | 3.23/30.26 | 10.37/52.13 | 17.42/64.38 |
| ViT$^\dagger$-B ($s$) | 2-12 | 1.30/35.20 | 4.17/45.11 | 12.41/64.28 | 22.70/75.36 |
| | 4-12 | 1.22/31.93 | 3.53/42.82 | 11.94/64.42 | 20.58/75.56 |
| | 6-12 | 1.23/32.07 | 3.69/41.54 | 12.20/62.82 | 21.37/74.93 |
| | 8-12 | 1.50/27.70 | 4.30/36.59 | 13.79/60.44 | 23.49/73.12 |
| ViT-B ($e$) | 4-6 | 4.17/13.04 | 3.92/25.56 | 13.12/47.93 | 27.88/62.34 |
| | 4-8 | 2.85/11.94 | 3.14/28.94 | 10.09/52.79 | 19.18/74.24 |
| | 4-10 | 2.81/17.34 | 2.85/29.77 | 9.04/54.17 | 16.86/75.66 |
| | 4-12 | 1.27/24.40 | 2.66/33.17 | 8.70/55.48 | 16.83/75.64 |
| DeiT-B ($e$) | 4-6 | 1.49/14.19 | 5.66/25.92 | 18.60/44.53 | 36.85/50.18 |
| | 4-8 | 1.30/18.34 | 3.76/18.34 | 14.12/40.51 | 25.10/50.75 |
| | 4-10 | 0.86/20.92 | 3.32/33.47 | 11.49/52.65 | 19.40/64.56 |
| | 4-12 | 0.62/19.90 | 2.33/34.34 | 9.49/54.49 | 16.32/66.29 |
| ViT$^\dagger$-B ($e$) | 4-6 | 2.89/17.21 | 7.27/34.18 | 22.30/55.61 | 40.42/63.98 |
| | 4-8 | 2.14/25.92 | 5.27/39.92 | 15.11/60.99 | 25.95/73.52 |
| | 4-10 | 1.21/29.02 | 4.17/41.50 | 12.87/63.30 | 22.19/75.31 |
| | 4-12 | 1.22/31.93 | 3.53/42.82 | 11.94/64.42 | 20.58/75.56 |

Table B. **Ablation study on departure ($s$) and arrival ($e$) layers.** The comparisons on interpolating the hyper-parameters from start to end layers. The hyper-parameters used in CoIBA are filled with a light gray color. We compare the quantitative results of the discrepancy between insertion/deletion scores with different intervals of confidence scores yielded by the model. The better-qualified attribution map yields a higher discrepancy in insertion/deletion.

| Model | *SA* | *FFN* | *Blocks* |
|---|---|---|---|
| ViT-B-16/224 | 12.81/62.47 | 15.71/57.56 | 17.53/56.31 |
| DeiT-B-16/224 | 11.59/53.86 | 13.41/51.12 | 13.53/52.39 |
| ViT$^\dagger$-B-16/224 | 15.90/64.87 | 17.74/62.07 | 19.52/61.97 |

Table C. **Quantitative comparison of the results produced by placing bottleneck into various operations.** This experiment shows the correctness of the performance when inserting bottlenecks into various operations, including self-attention (SA), feed-forward network (FFN), and block between SA and FFN. We compare 6,000 images randomly sampled from the IN-1k validation dataset. ViT$^\dagger$ denotes the model trained with CLIP.

Since mutual information between the bottleneck variable of $L$-th layer $Z_L$ and the label $Y$ can be expressed as:

$$
I[Y; Z_L] = H(Y) - H(Y|Z_L), \tag{C}
$$

where $H(Y)$ and $H(Y|Z_L)$ denote the entropy of $Y$ and conditional entropy of $Y$ conditioned on $Z_L$. Thus, we can relate mutual information and conditional cross entropy as:

| Model | Accuracy | Com. | | | | BI | Cor. | Con. | mX |
|---|---|---|---|---|---|---|---|---|---|
| | | CSDC | PC | DC | D | | SD | TS | |
| Chefer-LRP | 97.6 | 91.1 | 91.2 | 89.4 | 89.7 | 99.8 | 73.9 | 95.8 | 86.6 |
| Generic | 97.6 | 91.0 | 90.8 | 89.6 | 89.6 | 99.8 | 74.2 | 98.5 | 87.6 |
| IIA | 97.6 | 89.2 | 87.6 | 88.0 | 90.7 | 99.8 | 76.4 | 98.6 | 84.1 |
| ViT-CX | 97.6 | 56.9 | 36.2 | 41.6 | 83.8 | 99.8 | 78.3 | 57.7 | 66.8 |
| IBA | 97.6 | 96.0 | 97.8 | 94.4 | 91.8 | 99.8 | 76.9 | 71.7 | 80.8 |
| Beyond | 97.6 | 87.8 | 84.8 | 84.8 | 84.1 | 99.8 | 75.8 | 92.9 | 84.5 |
| CoIBA | 97.6 | 93.5 | 94.2 | 91.4 | 91.3 | 99.8 | 79.0 | 98.2 | **89.8** |

Table D. **Numeric detailed results of FunnyBirds experiment.** We provide the detailed numeric results for the reported FunnyBirds experiment. The mean explainability score (mX) is obtained by averaging Com., Cor., and Con. scores. The completeness score (Com.) is obtained by averaging CSDC, PC, DC, and D scores.

$$
\begin{aligned}
I[Y; Z_L] &= H(Y) - H(Y|Z_L) \\
&= H(Y) + D_{KL}[Y||\hat{Y}|Z_L] - H(Y; \hat{Y}|Z_L) \\
&\geq -H(Y; \hat{Y}|Z_L)\,.
\end{aligned} \tag{D}
$$

Here, we omit $H(Y)$ since it is constant. As $D_{KL}[Y||\hat{Y}|Z_L] \geq 0$, minimization of the conditional cross entropy increases the mutual information.

## C. Experimental Settings

### C.1. Model

We detail the settings for the experiments. We utilize the timm library, which is a publicly accessible open-source framework. We present all models by {*name*}–{*depth*}–{*patch size*}/{*image resolution*}, *e.g.*, ViT-B-16/224. All the included ViT models are pre-trained with ImageNet-21k. The models belonging to the DeiT family are pre-trained with IN-1k, including DeiT3. We leverage Swin-B with the settings of window size 7 and patch size 4. For Swin2-B, we utilize the model with the settings including the input resolution of 256 and window size 16. We denote the ViT$^*$ and ViT$^\dagger$ as the models trained with massive regression and CLIP, respectively.

### C.2. Hyperparameters Selected in CoIBA

We provide quantitative comparisons against various hyperparameter settings. CoIBA includes the departure $s$ and arrival $e$ layers and trade-off parameter $\beta$ as a hyperparameter to set. The overall hyperparameter settings chosen in CoIBA are illustrated in Tab. A. We discuss the setting of trade-off hyperparameter $\beta$ in Sec. D.7.1 regarding the out-of-distribution problem. We insert the bottleneck into the preceding operation of the self-attention layer *i.e.*, normalization layer.

**Departure and Arrival Layers** We empirically select the hyperparameter, which broadly yields the best correctness performance. Tab. B illustrates the quantitative comparisons against various hyperparameter settings. As shown in the results, including earlier layers for ViT yields an increased performance in insertion/deletion. However, our chosen hyperparameter shows enhanced performance in ViT pre-trained with CLIP and DeiT models including DeiT3. Referring to these results, we set 4 and 12 as departure and arrival layers, respectively, for base models. For large models, we set 8 and 24 as departure and arrival layers, respectively.

**Operation** We compare the correctness scores by inserting bottlenecks into the three types of operations: self-attention (SA), feed-forward network (FFN), and intermediate blocks between SA and FFN (Block). Tab. C shows the quantitative results for these settings. As shown in the results, inserting the bottleneck before the SA layer yields the best performance compared to inserting it before other operations.

## D. Additional Quantitative Results

### D.1. FunnyBirds Experiment

FunnyBirds assessment measures the faithfulness of an attribution map with a comprehensive metric, including three metrics: completeness (Com.), correctness (Cor.), and contrastivity (Con.). First, to assess Com., FunnyBirds assesses the controlled synthetic data check (CSDC), preservation check (PC), deletion check (DC), and distractibility (D). These four metrics are averaged to compute the Com. score. Second, evaluating Cor. includes a single deletion check (SD) which measures the correlation between part importance and the predicted scores of the targeted class. Finally, Con. measures target sensitivity to directly measure the sensitivity to a target class by assessing whether parts of different classes are correctly identified as their respective class from a single image. The overall score is obtained by averaging Com., Cor. and Con. scores. Tab. D shows the full numeric results of the FunnyBirds experiment. As shown in the results, CoIBA In particular, the outperforming of CoIBA in terms of target sensitivity compared to IBA

| Variant | Model | Chefer-LRP | Generic | IIA | ViT-CX | IBA | Beyond | CoIBA |
|---|---|---|---|---|---|---|---|---|
| SS* | ViT-B-16/224 (MAE) | - | 24.44/42.80 | 24.53/43.63 | 20.01/45.80 | 15.66/48.90 | 16.23/48.09 | **13.77/53.42** |
| | ViT-B-16/224 (Dino) | - | 7.62/50.42 | 7.46/50.52 | 14.09/45.55 | 8.91/49.45 | 8.12/50.23 | **6.83/53.68** |
| | BeiTv1-B-16/224 | - | 24.82/47.04 | 25.34/47.50 | 21.58/54.06 | 13.96/59.87 | 19.75/51.82 | **13.51/62.45** |
| IN-A | ViT*-B-16/224 | - | 2.46/25.99 | 2.68/25.37 | 4.41/23.97 | 2.34/25.52 | 2.27/26.29 | **1.82/32.62** |
| | ViT*-L-16/224 | - | 3.99/32.74 | 3.95/32.29 | 4.40/33.64 | 3.19/34.86 | 3.10/35.36 | **2.53/41.07** |
| | ViT†-B-16/224 | - | 2.58/28.46 | 2.44/28.60 | 9.67/10.98 | 2.49/27.33 | 2.53/28.72 | **1.97/36.15** |
| | ViT†-L-16/224 | - | 4.19/37.82 | 4.01/38.13 | 7.49/34.45 | 4.27/35.77 | 4.16/38.12 | **3.25/42.74** |
| | EVA-L-14/196 | - | 7.58/39.96 | 8.79/38.48 | 6.60/41.71 | 4.50/44.68 | 4.54/43.57 | **3.65/50.41** |
| IN-R | ViT*-B-16/224 | - | 7.72/36.70 | 8.41/36.13 | 9.18/34.34 | 6.56/37.85 | 6.79/37.58 | **5.14/43.56** |
| | ViT*-L-16/224 | - | 13.21/38.11 | 13.44/37.32 | 12.20/40.84 | 8.66/43.91 | 9.23/44.66 | **7.04/49.77** |
| | DeiT3-B-16/224 | - | 5.91/37.55 | 5.92/37.57 | 9.06/33.85 | 5.41/37.53 | 5.55/38.81 | **4.36/43.87** |
| | DeiT3-L-16/224 | - | 7.24/40.83 | 7.25/40.74 | 10.95/36.54 | 6.46/41.27 | 6.92/42.56 | **5.13/46.82** |
| | EVA-L-14/196 | - | 19.49/46.54 | 19.73/45.96 | 14.01/51.17 | 11.36/54.73 | 12.73/53.00 | **9.29/59.97** |

Table E. **Quantitative feature importance assessment on insertion ↑ / deletion ↓.** We denote SS* as ViT trained with self-supervised learning [3, 4, 7] and ViT† as ViT trained with CLIP. ViT† and ViT* denote the ViTs trained with CLIP and massive regularization methods. We additionally include EVA [6] for the comparison. We underline the state-of-the-art performance among the baselines.

demonstrates the ability of CoIBA in class-discriminative ability.

## D.2. Insertion/Deletion

We provide additional results of insertion/deletion in a wide range of models and datasets. For the deletion test, we leverage the image filled with zero pixels as a baseline, indicating a non-informative image. For the insertion test, we blur the input image using the 2D Gaussian blurring method with kernel size 51 and sigma 50. To demonstrate the generalizability of CoIBA, we include the model pre-trained with self-supervised learning and the results of IN-A and IN-R datasets in Tab. E. As shown in the results, CoIBA provides attribution maps with remarkable correctness scores compared to the baselines. In particular, the correctness performance in IN-A and IN-R indicate that CoIBA provides attribution maps regardless of the difficulty of input samples.

## D.3. Difficulty-aware Analysis

We provide the additional quantitative results of difficulty-aware analysis. First, we provide the confidence scores computed by the model per sample for different datasets including IN-1k, IN-A, and IN-R. Second, we include the quantitative results to demonstrate that CoIBA consistently improves the correctness of resulting attribution maps for various confident samples.

### D.3.1. Distribution of Confidence Scores

Depending on the type of pre-trained parameters of the models and datasets, the resulting confidence score per sample for the model is diversified. We provide the distribution per sample about confidence scores in Fig. B. As shown in the figure, massive samples in IN-1k lead the model to output a high confidence score. In contrast to this, IN-R and IN-A include numerous samples difficult to model. Thus, amplified correctness scores in IN-A and IN-R demonstrate the capability of CoIBA in generating explanations with a high correctness score compared to the baselines.

### D.3.2. Quantitative Results

In addition to the results presented in Sec. 4.5, we provide the additional results in Fig. A. For IN-1k results, we include the ViT pre-trained with massive regression and CLIP and DeiT3 models. For IN-A results, we include the same models to provide the results. As shown in the results,

## D.4. Sensitivity-N

The sensitivity-N [2] evaluates feature attribution assessment by measuring the correlation. The sensitivity is measured between the sum of attributions corresponding to the mask indices and the drop in model confidence caused by the rest of the feature subsets. We leverage the Pearson correlation coefficient (PCC) to measure the correlation. To compute the sensitivity, we compute the condition over 100 different indices and average them over 1,000 image samples. We generate indices from 1 to 80% of the number of pixels. Loosely speaking, sensitivity-N extends the *summation over delta* and *completeness*. Thus, the attribution maps yielding the high sensitivity provide a faithful explanation. We compare the sensitivity-N of the original and enhanced ViT architecture. The quantitative results of the sensitivity-N become visually distinguishable as shown in Fig. C. Compared to the existing approaches, CoIBA provides faithful attribution maps after the $10^3$ pixels are removed.

## D.5. Localization Assessment

To assess the localization ability of CoIBA compared to the baselines, we compare the effective heat ratio (EHR) [9] of the attribution maps yielded by different methods. This EHR measures whether the explanation highlights object-focused attribution. To this end, EHR computes the As shown in Tab. F, the attribution maps provided by CoIBA correctly highlight the foreground object. We include the ViT and DeiT models additionally fine-tuned to concentrate

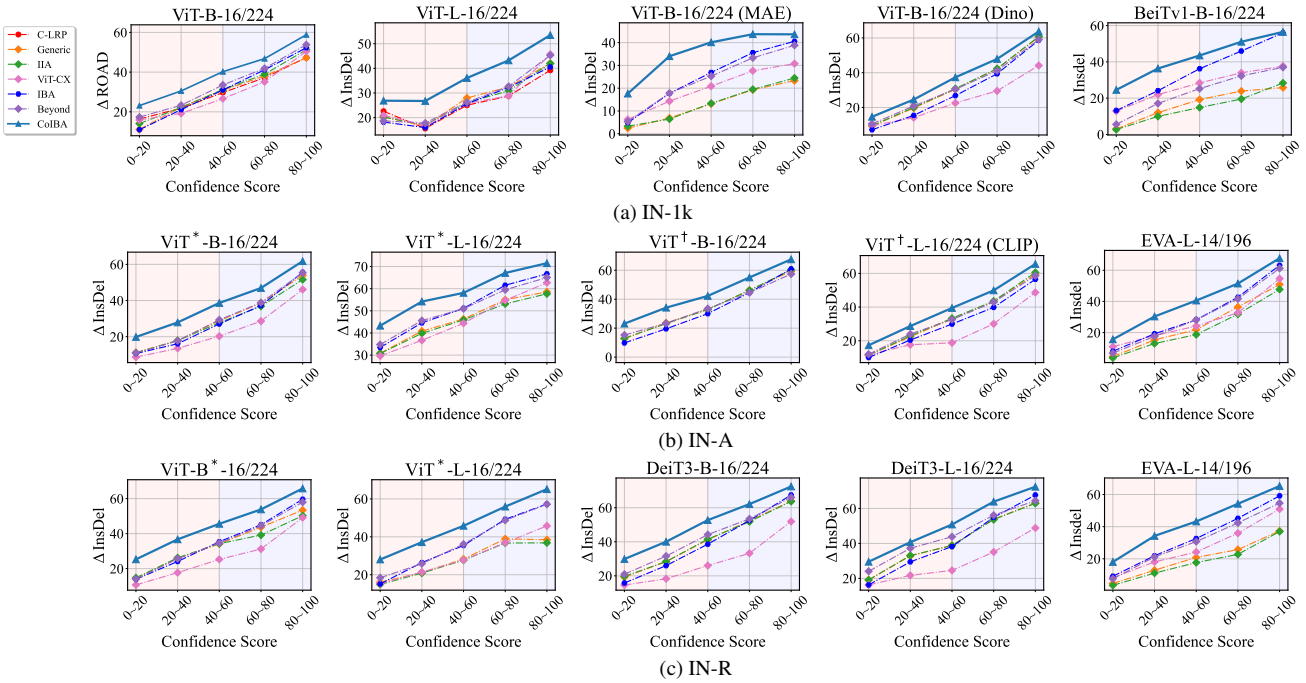



Figure A. **Difficulty-aware correctness assessment on insertion/deletion** We measure differences in insertion/deletion (ΔInsDel↑) scores (higher is better). We fill the regions including low-confident samples and high-confident samples with red and blue, respectively, based on the prediction made by the model.

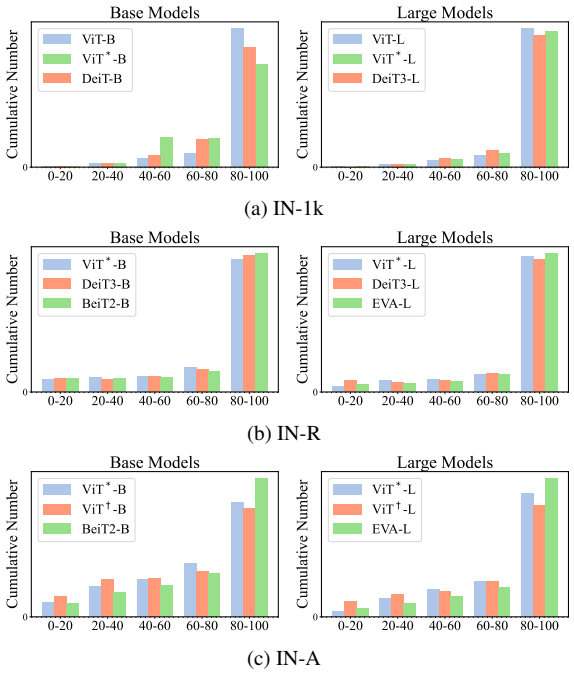



Figure B. **Cumulative number of confidence scores per sample.** We compare the cumulative number of samples per different confidence scores computed by the model. We include three datasets including IN-1k, IN-A, and IN-R for the experiment.

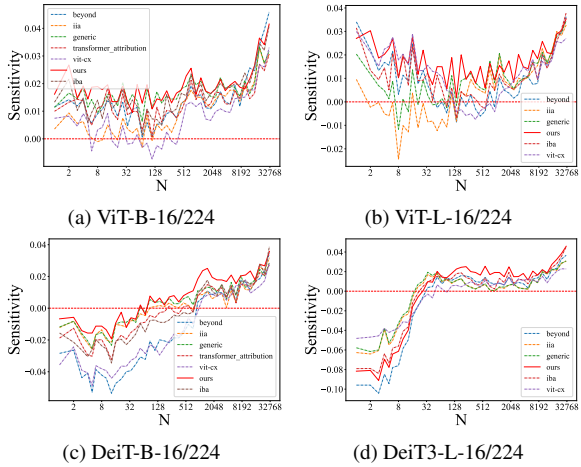



Figure C. **Quantitative results of Sensitivity-N.** We compare the results of Sensitivity-N against the base and large models. We include the ViT and DeiT models. We plot the performance of CoIBA with a solid line with red color.

the object [5]. The results show that CoIBA provides the attribution maps that correctly highlight the human-labeled bounding box. In particular, according to the results of ViT-B-8/224, when the granularity of the attributions is increased as the patch size of self-attention is decreased, the CoIBA the significant performance in localization. These results demonstrate that the overall attribution maps pro-

| Model | Non-finetuned | | Finetuned [5] | |
|---|---|---|---|---|
| | ViT-B | DeiT-B | ViT-B | DeiT-B |
| Chefer-LRP | 0.277 | 0.252 | 0.297 | 0.274 |
| Generic | 0.229 | 0.203 | 0.285 | 0.232 |
| IIA | 0.256 | 0.239 | 0.293 | 0.271 |
| ViT-CX | 0.240 | 0.212 | 0.229 | 0.210 |
| IBA | **0.297** | 0.235 | **0.310** | 0.250 |
| Beyond | 0.252 | 0.216 | 0.263 | 0.208 |
| CoIBA | 0.248 | **0.269** | 0.253 | **0.281** |

(a) Non-finetuned model vs finetuned model [5]

| Model | ViT* | | ViT† | |
|---|---|---|---|---|
| | ViT-B | ViT-L | ViT-B | ViT-L |
| Chefer-LRP | - | - | - | - |
| Generic | 0.200 | 0.103 | 0.247 | 0.226 |
| IIA | 0.107 | 0.110 | 0.237 | 0.234 |
| ViT-CX | 0.232 | 0.221 | 0.212 | 0.234 |
| IBA | 0.226 | 0.191 | 0.258 | 0.235 |
| Beyond | 0.286 | 0.218 | 0.225 | 0.205 |
| CoIBA | **0.300** | **0.265** | **0.305** | **0.304** |

(b) Varaints of training strategy

| Model | Patch size | | Depth | |
|---|---|---|---|---|
| | 8 | 32 | ViT-L | DeiT3-L |
| Chefer-LRP | - | - | 0.236 | - |
| Generic | 0.169 | 0.143 | 0.173 | 0.297 |
| IIA | 0.201 | 0.150 | 0.176 | 0.304 |
| ViT-CX | 0.187 | 0.236 | 0.212 | 0.207 |
| IBA | 0.211 | 0.246 | 0.213 | 0.245 |
| Beyond | 0.202 | 0.243 | 0.201 | 0.207 |
| CoIBA | **0.272** | **0.263** | **0.247** | **0.334** |

(c) Variants of patch size and depth

Table F. **Quantitative visual evaluation results of EHR.** This metric evaluates the localization capability of the feature attribution methods. We omit patch size and input resolution for simplicity. For the variants of patch size, we utilize ViT-B-8/224 and ViT-B-32/224. The robust [5] indicates the model is further fine-tuned to focus on the foreground object. We denote ViT pre-trained with massive regularization and CLIP as Reg and CLIP.

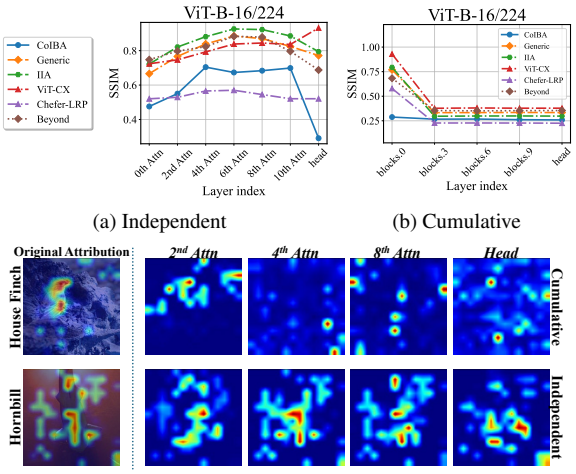
| Model | Acc. | $\beta = 0.01$ | $\beta = 0.1$ | $\beta = 1$ | $\beta = 10$ | $\beta = 100$ |
|---|---|---|---|---|---|---|
| ViT-B | 81.8 | 100.0/86.6 | 100.0/90.2 | 98.8/88.8 | 15.2/13.4 | 1.4/0.6 |
| DeiT-B | 82.0 | 100.0/90.1 | 100.0/93.0 | 99.8/94.5 | 36.6/43.3 | 0.1/1.0 |

Table G. **Quantitative comparison on various $\beta$ settings.** We report top-1 accuracy per trade-off hyper-parameter setting $\beta$. We divide correctly/incorrectly predicted samples.

duced by CoIBA correctly highlight the foreground object in addition to the increased correctness performance.

## D.6. Sanity Check

The sanity check [1] confirms whether the produced explanations are sensitive to the model parameter. This experiment measures the similarity of the attributions produced with non-randomized and randomized model parameters. We include two tests: cumulatively or independently randomizing the parameters of each layer. For the cumulative parameter randomization, we randomize the model parameters after the 0, 3, 6, and 9 layers and measure the similarity




(a) Independent (b) Cumulative

(c) Qualitative Results

Figure D. **An analysis of parameter randomization test for sanity check.** We utilize ViT-B-16/224 as the baseline architecture. We measure SSIM for the attribution maps produced from different indexed layers after randomization. The Fig. D (a) and D (b) represent the comparisons of independent and cumulative parameter randomization tests, respectively.

| Model | IBA | CoIBA |
|---|---|---|
| ViT-B-16/224 | 0.060 | 0.062 |
| ViT-L-16/224 | 0.168 | 0.174 |

Table H. **Comparisons of computational cost (sec) required to generate an attribution map.**

of the attributions produced. We select each attention layer with interval 2 in an independent parameter randomization test. We report quantitative results in Fig. D with exemplary results. We utilize 1,000 images randomly sampled from the IN-1k validation dataset to measure the similarity between attributions utilizing the SSIM metric. The diminished similarity in the results indicates that the attributions are consistently obfuscated as the parameters of each layer are randomized. Therefore, CoIBA is sensitive to the model parameters, leading to yield faithful attributions to explain the decision-making process. Furthermore, CoIBA is fairly sensitive to all the layers within the model parameters as the SSIM score results in uniformity across the layers.

## D.7. Discussion

In this section, we provide further results included in Sec. 4.6. We provide the quantitative results to show the effectiveness of the variational upper bound. After that, we provide the out-of-distribution caused by overly defined trade-off hyperparameter $\beta$.

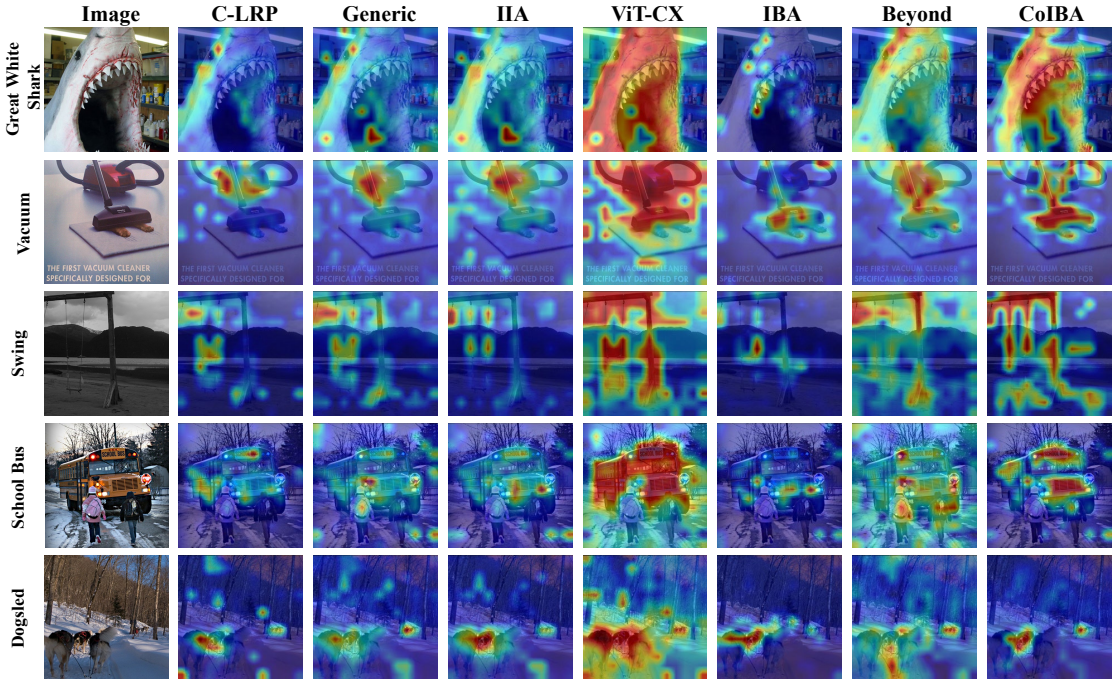
### D.7.1. Out of Distribution and Over-compression

The trade-off hyperparameter $\beta$ controls the degree of compression related to the relevancy term in Eq. (8) The relevant information is suppressed by setting excessive trade-off hyperparameter $\beta$. In contrast to this, setting this hy-

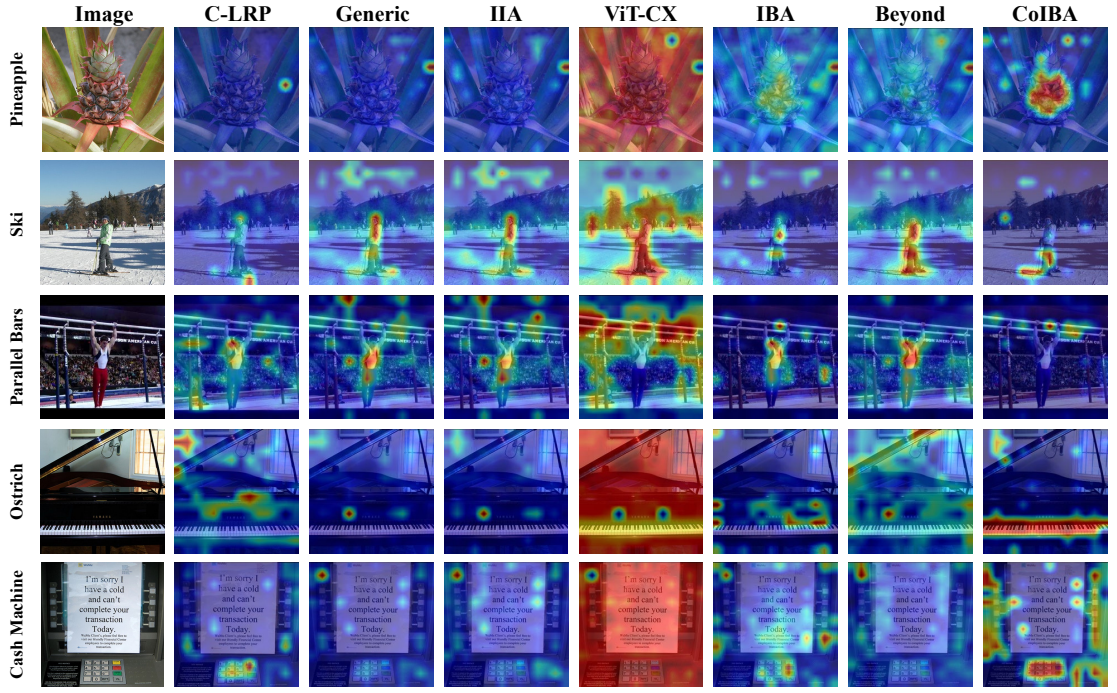
perparameter overly small leads to leaving irrelevant information in the activations. To empirically choose $\beta$, we analyze whether the decision made by the model is corrupted per compressing with different trade-off parameter settings. As shown in Tab. G, setting $\beta$ larger than $1$ leads to over-compression, vanishing the relevant information as well. In contrast to this, setting $\beta$ smaller than $1$ diminishes the correctness performance of a resulting attribution map. These results demonstrate the reasonability of our choice $\beta = 1$ in dealing with the trade-off between relevancy and compression term.

## E. Computational Cost

We report the computational cost while generating a single attribution map for an input sample. We utilize NVIDIA A6000 GPU to measure the computational time. As shown in Tab. H, CoIBA requires a similar computational cost compared to IBA. For example, IBA and CoIBA consume 0.06 and 0.062 (sec) for computing the attribution map. Therefore CoIBA requires a significantly small computational cost, compared to dealing with a specific layer, Therefore, a significantly small computational cost required by CoIBA demonstrates that CoIBA significantly amplifies the correctness of the resulting attribution map while requiring a small computational cost.

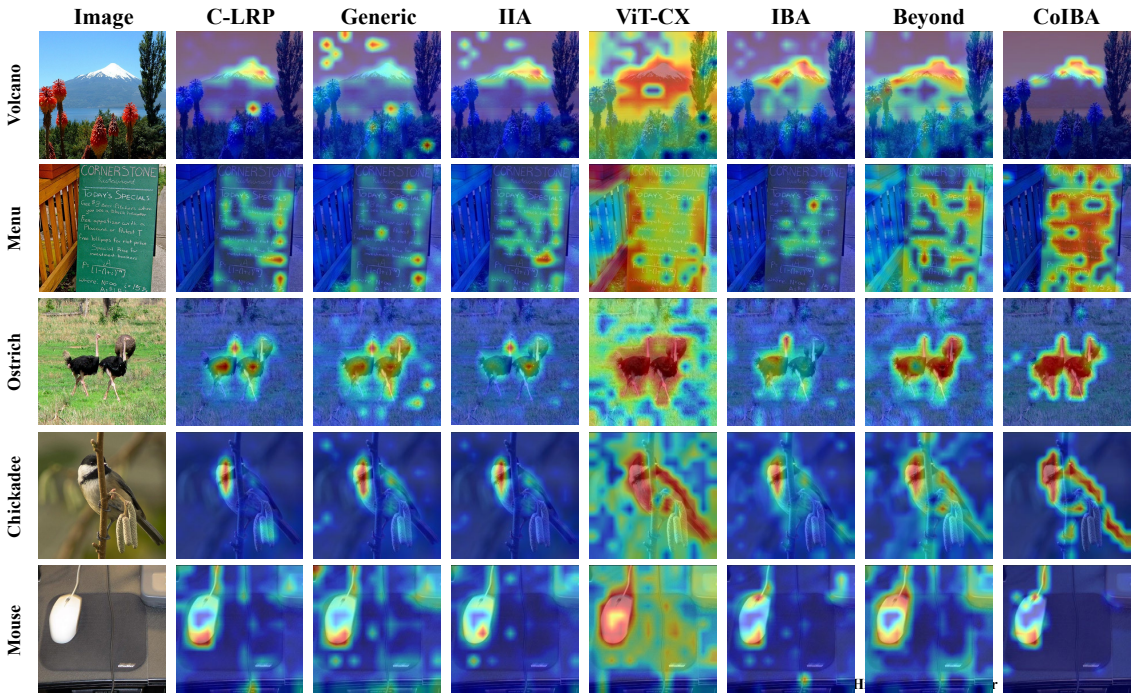



(a) ViT-B-16/224

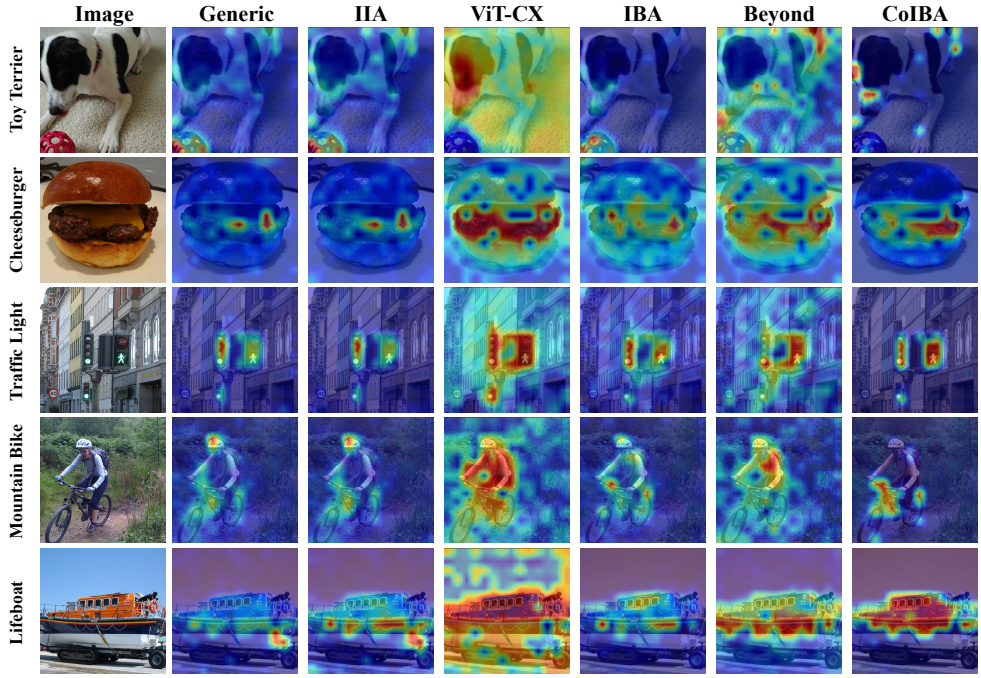



(b) ViT-L-16/224

Figure E. **Visualized attribution maps for IN-k produced from ViT. C-LRP indicates the Chefer-LRP method.**

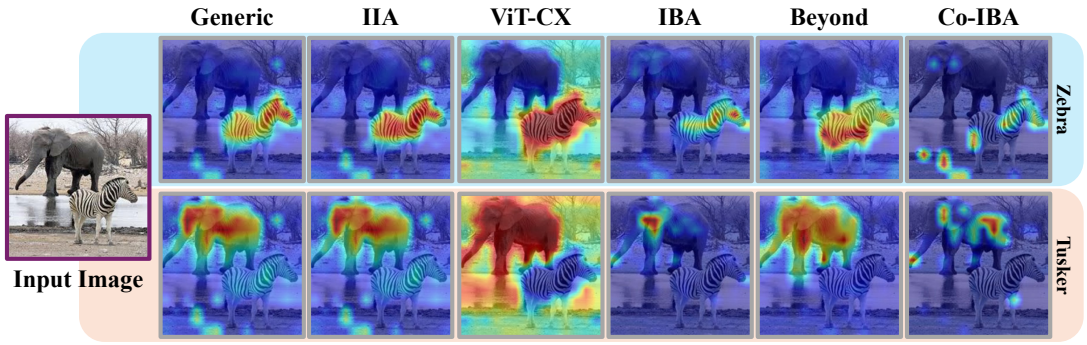



(a) DeiT-B-16/224

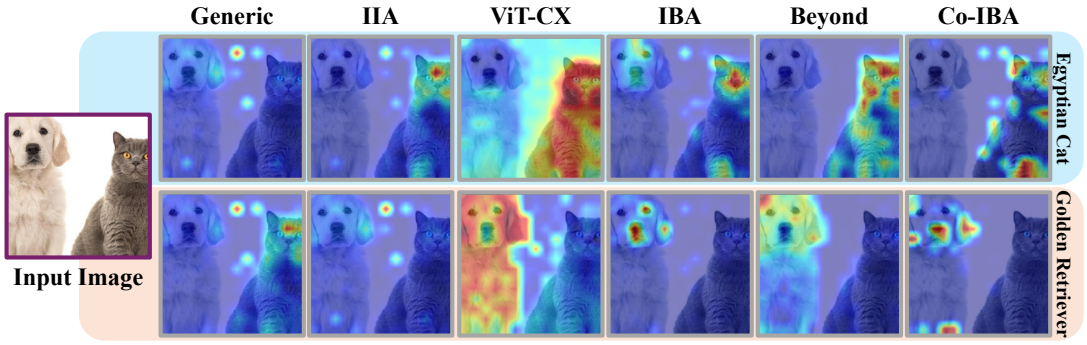



(b) DeiT3-B-16/224

Figure F. **Visualized attribution maps for IN-k produced from DeiT-B and DeiT3-L.**

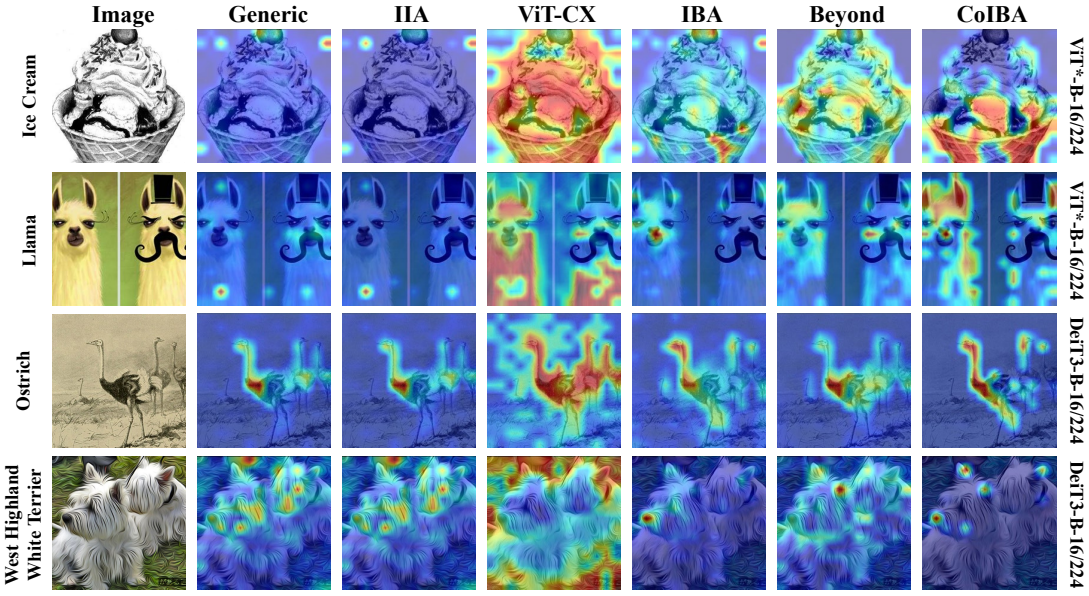



(a) Class-discriminative ability – ViT-B-16/224



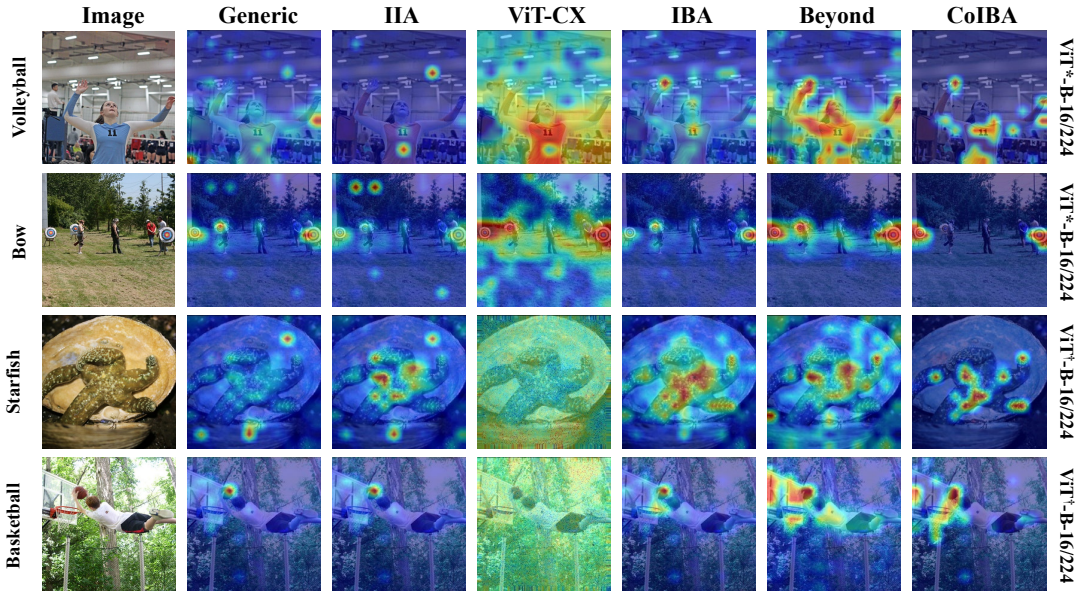

(b) Class-discriminative ability – ViT*-B-16/224

Figure G. **Comparisons of class-discriminative ability from choosing different targeted classes.**




(a) IN-R

Figure H. **Visualized attribution maps of IN-R.**




(a) IN-A

Figure I. **Visualized attribution maps of IN-A.**

## References


[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31, 2018. 5

[2] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks, 2018. 3

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1, 3

[5] Hila Chefer, Idan Schwartz, and Lior Wolf. Optimizing relevance maps of vision transformers improves robustness. *Advances in Neural Information Processing Systems*, 35:33618–33632, 2022. 4, 5

[6] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 3

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1, 3

[8] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3981–3991, 2023. 2, 6

[9] Yang Zhang, Ashkan Khakzar, Yawei Li, Azade Farshad, Seong Tae Kim, and Nassir Navab. Fine-grained neural network explanation by identifying input features with predictive information. *Advances in Neural Information Processing Systems*, 34:20040–20051, 2021. 1, 2, 3, 5