

## APPENDIX / SUPPLEMENTAL MATERIAL

## NOMENCLATURE

$M$	A perturbation mask
$\mathbb{V}$	The variance over the perturbed space
$\odot$	Hadamard product
$f_R$	A reasonable concept extraction function
$Q$	A string of class question
$R$	A set of reasons that define a <i>concept explanation</i>
$\ell$	A loss function
$\hat{f}$	A concept-based explainer
$\mathbb{E}$	An expectation function
$\mathcal{A}$	CNN feature map with activations $a \in \mathcal{A}$
$\mathcal{A}'$	CNN feature map with activations $a' \in \mathcal{A}'$ produced by $\mathcal{J}'$
$\mathcal{G}^n$	A set of axioms $g_i, (i = 1, 2, \dots, n)$
$\mathcal{J}$	A concept-based explainer's reducer
$\mathcal{J}'$	A concept-based explainer's inverse reducer
$\mathcal{P}$	Concept activation vectors (CAVs)
$\mathcal{S}$	Concepts
$\mathcal{W}$	Concept importance measure
$\mathcal{X}$	Set of training images $x$
$\mathcal{X}_p^y$	A set of <i>prototypes</i> $x_p \in \mathcal{X}_p^y$ belonging to class $y$
$\mathcal{Y}$	Set of training labels $y$
$\phi$	A homogeneity measure
$C$	The classifier layer of a CNN model
$c$	Feature map channel
$c'$	Reduced feature map channel
$CONV$	A convolutional layer
$E$	Conv-ReLU layer of a CNN model
$f$	A CNN model
$L$	A subset of literals $L \in L^T$
$m$	Feature map dimension, $h \times w$
$P(\cdot)$	Probability distribution
$ReLU$	Rectified linear units activation function
$t$	Trainable linear weights of classifier $C(\cdot)$

## A PRELIMINARIES

### A.1 CONCEPT IMPORTANCE:

**Recap:** We consider a CNN  $f(\cdot)$  constituting the Conv-ReLU  $E(\cdot)$  and the classifier  $C(\cdot)$  where  $f(\cdot) = C(E(\cdot))$ . For a classification task  $f(\cdot) : x \in \mathcal{X} \rightarrow y \in \mathcal{Y}$ ,  $E(x) = \mathcal{A}$  where  $\mathcal{A} \in \mathbb{R}^{m \times c}$  ( $m = (h, w)$ ). Given the explainer  $\hat{f}\{\mathcal{J}, \mathcal{J}'\}$ , the *reducer*  $\mathcal{J}(\cdot)$  achieve  $\mathcal{A} \approx \mathcal{SP}$  i.e. the concepts  $\mathcal{S} \in \mathbb{R}^{m \times c'}$  and CAVs  $\mathcal{P} \in \mathbb{R}^{c \times c'}$ , where  $c'$  is a user-defined number of concepts such that  $c' \ll c$ .

**Theorem 2** (Concept Importance). *This quantifies the sensitivity of a  $f(x)$  to a given CAV  $p \in \mathcal{P}$  calculated as the directional derivative Kim et al. (2018); Fel et al. (2023); Kim et al. (2023) of  $C(\cdot)$  w.r.t  $\mathcal{P}$  in  $\mathcal{A}$ .*

This could be computed as the TCAV score Kim et al. (2018):

$$\frac{\partial C_{l,y_i}}{\mathcal{P}_l} = \lim_{\epsilon \rightarrow 0} \frac{h_{l,y_i}(\mathcal{A}_l^D + \epsilon \mathcal{P}_l) - h_{l,y_i}(\mathcal{A})}{\epsilon}, \quad (1)$$

where the estimated concept importance  $\mathcal{W}$  for CAV  $\mathcal{P}$  is  $\mathcal{W} = \mathcal{P} \cdot t$ , following a global average pooling of  $\mathcal{A}$ .

Proposed in Fel et al. (2023), concept importance could also be estimated as the Total Sobol Indices—a measure of a concept’s contribution and its interactions (of any order) with any other concepts to the CNN’s output variance:

$$\mathcal{W}_{y_i} = \frac{\mathbb{E}_{\mathbf{M}_{\sim i}}(\mathbb{V}_{M_i}(\mathbf{Y} | \mathbf{M}_{\sim i}))}{\mathbb{V}(\mathbf{Y})}, \quad (2)$$

where  $\mathbf{M} = (M_1, \dots, M_r) \sim \mathcal{U}([0, 1]^r)$  is a perturbation mask containing independent and identically distributed random variables,  $\mathbf{Y} = C((\mathcal{S} \odot \mathbf{M})\mathcal{P})$  is the label predictions of the perturbed activation,  $\odot$  denotes the Hadamard product,  $\mathbb{V}(\cdot)$  is the variance over the perturbed space, and  $\mathbb{E}(\cdot)$  is the expectation over the perturbation space.

### A.2 PROTOTYPE SELECTION:

Numerous approaches for prototype selection have been proposed Kim et al. (2016); Fel et al. (2023); Snell et al. (2017); Koh & Liang (2017); Ma et al. (2023); Singh & Yow (2021); they follow a similar principle: Given a set of literals  $L$  in an instance  $x$  that forms a class question, a prototype  $x'$  is an instance  $x' \in \mathcal{X}$  such that its set of literals  $L'$  is maximally representative of a class  $y \in \mathcal{Y}$ , satisfying:

$$x_p = \arg \max_{x \in \mathcal{X}_y} \mathbb{E}_{x' \sim P(\mathcal{X}_y)} \phi(L, L'). \quad (3)$$

where:

- $\mathbb{E}_{x' \sim P(\mathcal{X}_y)} \phi(L, L')$  denotes the expectation in instances  $x'$  drawn from the probability distribution  $P(\mathcal{X}_y)$ .
- $\phi(L, L')$  is a similarity function between the literals of  $x$  and  $x'$ .
- A *prototype* is a set of literals sourced from other instances that align with the literals of the given instance.
- $x'$  is prototypical of  $x$  if  $\forall L \in \hat{f}(\mathbf{Q})$ ,  $x'$  are similar to  $x$  such that  $f(x) = f(x') = y$ .

Reliable similarity metrics include Jaccard similarity, Cosine similarity, and Kernel-based similarity metrics Xie et al. (2016).

**Lemma 2** (Prototype Stability Condition). *Let  $\mathcal{X}_y$  be the set of instances belonging to class  $y$ , and let  $x_p$  be a prototype selected for  $x \in \mathcal{X}_y$ . If  $\phi(L, L')$  is a valid similarity function, then  $x_p$  satisfies:*

$$\mathbb{E}_{x' \sim P(\mathcal{X}_y)} \phi(L, L') \geq \delta, \quad \text{for some threshold } \delta > 0. \quad (4)$$

*Proof.* Since  $x_p$  maximizes  $\mathbb{E}_{x' \sim P(\mathcal{X}_y)} \phi(L, L')$ , we have:

$$\mathbb{E}_{x' \sim P(\mathcal{X}_y)} \phi(L, L') = \sup_{x' \in \mathcal{X}_y} \phi(L, L'). \quad (5)$$

As similarity functions are bounded (i.e.,  $0 \leq \phi(L, L') \leq 1$ ), there exists a threshold  $\delta$  such that:

$$\sup_{x' \in \mathcal{X}_y} \phi(L, L') \geq \delta. \quad (6)$$

Thus,  $x_p$  is a stable prototype.  $\square$

**Theorem 3** (Prototype Consistency Theorem). *Let  $\mathcal{X}_y$  be a set of instances belonging to class  $y$ , and let  $\mathcal{X}_p^y = \{x_1, x_2, \dots, x_m\}$  be a set of selected prototypes such that for each prototype  $x_p$ , the following holds:*

$$\phi(L_p, L_q) \geq \tau, \quad \forall x_p, x_q \in \mathcal{X}_p^y. \quad (7)$$

*Then, the prototype selection process is consistent if:*

$$\forall x \in \mathcal{X}_y, \exists x_p \in \mathcal{X}_p^y \text{ such that } \phi(L, L_p) \geq \delta. \quad (8)$$

*where  $\tau$  is an intra-prototype similarity threshold and  $\delta$  is a class-wise similarity threshold.*

*Proof.* From Lemma 1, we have:

$$\mathbb{E}_{x' \sim P(\mathcal{X}_y)} \phi(L, L') \geq \delta. \quad (9)$$

Since prototype selection is an optimization problem constrained by  $\delta$ , there exists at least one prototype  $x_p$  satisfying  $\phi(L, L_p) \geq \delta$ , ensuring consistency.  $\square$

**Theorem 4** (Optimal Prototype Selection as a Bounded Optimization Problem). *Given a dataset  $\mathcal{X}_y$  and a prototype function  $g(x, x') = \phi(L, L')$ , prototype selection can be formulated as:*

$$\max_{\mathcal{X}_p^y \subseteq \mathcal{X}_y} \sum_{x \in \mathcal{X}_y} \max_{x' \in \mathcal{X}_p^y} g(x, x'), \quad (10)$$

*subject to  $|\mathcal{X}_p^y| \leq k$ .*

*Proof.* The objective function maximizes the similarity of all instances to their closest prototype. Given that similarity functions are bounded and the dataset is finite, this problem is a constrained maximization solvable via sub-modular optimization techniques Bachem et al. (2017).  $\square$

## B FORMALIZATION OF AXIOMATIC FOUNDATIONS

The axiomatic foundations for concept-based explanations of CNNs proposed in the study are conceptually formulated and justified based on five principles of CNN behaviour and logic. Let  $f : \mathcal{X} \rightarrow \mathcal{Y}$  be a CNN classifier, and let  $\hat{f} : X \rightarrow \mathcal{S}$  be an explanation mapping, where  $\mathcal{S}$  is the concept space. An explanation is valid if it satisfies a set of principles, including the principles of Human Alignment, Causality, Consistency, Faithfulness, and Representation  $\{P_1, P_2, P_3, P_4, P_5\}$ , respectively.

We derive each principle from first principles under the assumption of orthogonality:

$$P_i \perp P_j \quad \forall i \neq j,$$

i.e., no principle can be derived as a corollary of another, and each encodes an irreducible requirement.

## B.1 INTERPRETABILITY

**Necessity:** Even if explanations are grounded in internal features, they are useless unless interpretable to humans. That is, explanations must satisfy  $P_1$ .

**Derivation:** Let  $\iota : \mathcal{S} \rightarrow \mathcal{H}$  be the interpretability mapping, where  $\mathcal{H}$  is the space of human-recognizable concepts. Then  $\forall s \in \hat{f}(x), \iota(s) \neq \emptyset$ .

**Orthogonality:**

- Cannot be reduced to Representation ( $P_5$ ), since features may correspond to uninterpretable dimensions.
- Independent of Faithfulness ( $P_4$ ), because perfect predictive preservation does not guarantee interpretability.

A concept explainer  $\hat{f}$  satisfies the *Interpretability* axiom if the concept explanations can be intuitively understood, i.e., satisfies  $P_1$ . In practical terms, for *dog* classification task (as shown in Figure 1), the explanation should focus on high-level, meaningful concepts such as *Ear*, *Eye*, *Nose*, and *Cheek*, which are both simple and relevant to the prediction, without introducing unnecessary complexity.

**Definition 1** (Interpretability of a Concept Explanation). *A concept explainer  $\hat{f}$  satisfies Interpretability for a class question  $Q$  if for every explanation  $L \in \hat{f}(Q)$ , the following holds:*

$$\hat{f}(Q) = \{L_1, L_2, \dots, L_n\}, \quad \forall L_i \in \hat{f}(Q), \quad L_i \neq \emptyset \quad \text{and} \quad L_i \in \mathbf{R}, \quad (11)$$

where  $\mathbf{R}$  is the set of reasonable concepts.

**Lemma 3** (Minimal Explanation Set). *If  $\hat{f}$  satisfies Interpretability, then the concept explanation  $L$  consists of a minimal set of concepts that are both successful and simple. Thus, the size of  $L$  is minimized subject to the success and reasonableness conditions. Formally,*

$$L = \arg \min_{L_i} |L_i| \quad \text{s.t.} \quad L_i \in \mathbf{R}, \quad \hat{f}(Q) \text{ is successful and simple.} \quad (12)$$

**Theorem 5** (Interpretability Guarantees Human Alignment). *If  $\hat{f}$  satisfies Interpretability, then the concept explanations use simple high-level concepts that humans can understand. Specifically, there exists a function that quantifies the human-understandability of the concepts (in this case, we assume the concept importance  $\mathcal{W}$ ), ensuring that:*

$$\sum_{l \in L} \mathcal{W}(l) \approx \sum_{l \in L^T} \mathcal{W}(l), \quad (13)$$

where  $L^T$  is the ground-truth set of concepts, and  $\mathcal{W}$  is a relevance function that measures how understandable each concept is.

*Proof.* The *Interpretability* of  $\hat{f}$  guarantees that the explanations are minimal, intuitive, and contain only relevant concepts. The relevance function  $\mathcal{W}$  quantifies the understandability of the concepts. By minimizing the number of concepts, the explanation remains clear without losing essential meaning. Hence, the theorem holds.  $\square$

**Corollary 1** (Interpretability score). *The Interpretability score  $g_1$  is defined as a binary value:*

$$g_1 \in \{\text{True}, \text{False}\} \quad (14)$$

where:

- *True:* The explanation is interpretable.
- *False:* The explanation is not interpretable.

## B.2 RELEVANCE

**Necessity:** Explanations must refer to causally relevant concepts, not correlations. That is, explanations must satisfy  $P_2$ .

**Derivation:** Let  $do(\cdot)$  denote an intervention in Pearl’s do-calculus Correa & Bareinboim. Then, for an instance  $x$  and associating set of concepts  $\mathcal{S} = \hat{f}(x): \forall s \in \mathcal{S}, P(f(x) \mid do(s = 0)) \neq P(f(x))$ .

Using do-calculus, for a concept  $s \in \hat{f}(x)$ :

$$P(f(x) \mid do(s = 0)) \neq P(f(x)).$$

This ensures that removing the concept alters the prediction distribution.

**Orthogonality:**

- Not implied by Faithfulness ( $P_4$ ): a faithful surrogate may still rely on spurious correlations.
- Not implied by Representation ( $P_1$ ): an interpretable internal feature may still be non-causal.

The *Relevance* of a concept explanation implies that the explanation should highlight only the Causal concepts. Consider the case of a concept explainer  $\hat{f}$  used to explain the classification of a *dog*. The explanation should highlight only the relevant concepts, such as *Ear* and *Eye*, which contribute to the classification of the instance as a dog. It should not include irrelevant concepts, such as *Background*, which do not contribute to the prediction.

**Definition 2** (Relevance of a Concept Explanation). *A concept explainer  $\hat{f}$  satisfies the Relevance axiom if, for any class question  $Q$  and for any explanation  $L \in \hat{f}(Q)$  that explains an instance decision  $f(x)$ , the mutual information between the reasonable concept predictions  $f_R(L)$  and the CNN’s prediction  $f(x)$  is maximized. This ensures that the explanation contains concepts that causally influence the model’s prediction:*

$$\mathbf{g}_9 = \text{Mutual Information}(\hat{f}(x), f(x)) \text{ where } 0 < \mathbf{g}_9 \leq 1. \quad (15)$$

Additionally, for a literal  $l \in L$ , removing or altering  $l$  should significantly change the prediction  $f(x)$ :

$$\exists l \in L, f(x) \neq f(x \setminus l) \quad (16)$$

This ensures that the concepts used in the explanation are not only correlated with the prediction but also have a causal influence on it.

**Lemma 4** (Maximization of Mutual Information). *The mutual information  $\mathbf{g}_9$  between the set of reasonable concept predictions  $f_R(L)$  and the CNN’s prediction  $f(x)$  is defined as:*

$$I(f_R(L); f(x)) = \mathbb{E}_{p(x)} \left[ \log \frac{p(f_R(L), f(x))}{p(f_R(L))p(f(x))} \right] \quad (17)$$

Maximizing this mutual information ensures that the concepts provided by the explainer are causally relevant to the CNN’s decision.

**Theorem 6** (Relevance Guarantees Causality). *If a concept explanation  $\hat{f}$  satisfies the Relevance axiom, then for any class question  $Q$ , and for any explanation  $L \in \hat{f}(Q)$  that explains an instance decision  $f(x)$ , the mutual information between the concept explanation and the CNN’s prediction is maximized. Moreover, removing or altering a literal  $l$  in  $L$  significantly changes the model’s prediction:*

$$\exists l \in L, f(x) \neq f(x \setminus l) \quad (18)$$

This ensures that the explanation reflects true causal reasoning rather than spurious correlations.

*Proof.* The mutual information between the concept explanation and the model’s prediction is a measure of the dependency between the two. Maximizing this mutual information ensures that the explanation contains concepts that directly influence the model’s prediction. Furthermore, altering or removing a literal from the explanation should result in a change in the model’s output, indicating a causal relationship between the literal and the prediction.  $\square$

**Corollary 2** (Bounded Causal Influence Score). *The Relevance score  $g_2$  is bounded within the range:*

$$0 < g_2 \leq 1 \quad (19)$$

*A score closer to 1 indicates that the explanation contains concepts that causally Relevance the prediction, while a score closer to 0 indicates that the explanation is merely correlated with the prediction without a true causal Relevance.*

### B.3 COHERENCE

**Necessity:** Explanations must be stable across similar inputs, ensuring reproducibility. That is, explanations must satisfy  $P_3$ .

**Derivation:** Let  $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$  be a metric on the input space, and  $d_{\mathcal{S}} : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}^+$  a metric in the concept space. Then:  $\forall x, x' \in \mathcal{X}, \quad d_{\mathcal{X}}(x, x') \leq \epsilon \implies d_{\mathcal{S}}(\hat{f}(x), \hat{f}(x')) \leq \delta$ , for small  $\epsilon, \delta > 0$ .

**Orthogonality:**

- Not implied by Faithfulness ( $P_4$ ): faithful but unstable explanations are unreliable..
- Not implied by Human Alignment ( $P_1$ ): an interpretable explanation may vary arbitrarily.

The *Coherence* of a *concept explanation* ensures that the explanation is internally consistent and logically structured. In other words, the set of concepts provided in the explanation must not be contradictory, and each explanation should be logically interconnected. A concept explainer  $\hat{f}$  satisfies the *Coherence* axiom if, for any class question  $Q$ , the explanations generated for different instances should not use the same concept literals that conflict with each other.

**Definition 3** (Coherence of a Concept Explanation). *A concept explainer  $\hat{f}$  satisfies Coherence for a class question  $Q$  if, for every explanation  $L \in \hat{f}(Q)$ , the following condition is met:*

- *The set of concepts in  $L$  should not contain literals that can explain conflicting or contradictory predictions. That is, for any pair of instances  $x$  and  $x'$ , the literals in  $L$  should not be shared between the explanations for  $f(x)$  and  $f(x')$ .*

*Formally, for any class question  $Q$ :*

$$\forall L \in \hat{f}(Q), \forall x, x' \quad \text{if } f(x) \neq f(x'), \quad L(x) \cap L(x') = \emptyset. \quad (20)$$

*where  $L(x)$  represents the explanation for instance  $x$  and  $L(x')$  represents the explanation for instance  $x'$ .*

**Lemma 5** (Coherence Function). *The coherence of a concept explanation is measured by the consistency of the explanation literals. Adapted from Piersol (2010), the coherence function is defined as:*

$$g_5 = \text{Coherence function}(\hat{f}(x), f(x)) = \frac{|G_{f\hat{f}}(s)|^2}{G_{ff}(s)G_{\hat{f}\hat{f}}(s)}, \quad (21)$$

*where  $G_{ff}(s)$ ,  $G_{f\hat{f}}(s)$  and  $G_{\hat{f}\hat{f}}(s)$  are the one-sided cross-spectral density function and the power spectral density for  $f(x)$  and  $\hat{f}(x)$  for instances  $x$  and  $x'$ , respectively. The coherence score  $g_5$  ranges from 0 to 1, where a higher score indicates a more coherent explanation.*

**Theorem 7** (Coherence Guarantees Consistency). *If a concept explainer  $\hat{f}$  satisfies the Coherence axiom, then for any pair of instances  $x$  and  $x'$ , the explanation literals should not be shared between contradictory predictions, ensuring that the explanation is logically consistent. Formally:*

$$L(x) \cap L(x') = \emptyset \quad \text{for } f(x) \neq f(x'). \quad (22)$$

*Therefore, the explanations for different instances should be logically interconnected and consistent with the model's decision-making process.*

*Proof.* The coherence of a concept explanation ensures that the literals in an explanation are logically connected. If two explanations share contradictory literals, this would violate the principle of coherence. Therefore, the theorem holds by ensuring that  $L(x)$  and  $L(x')$  do not share conflicting concepts when the predictions for  $x$  and  $x'$  differ.  $\square$

**Corollary 3** (Bounded Coherence). *The Coherence score  $g_3$  is bounded within the range:*

$$0 < g_3 \leq 1 \quad (23)$$

*A score closer to 1 indicates that the explanation is highly coherent, whereas a score closer to 0 indicates low coherence.*

#### B.4 FIDELITY

**Necessity:** Explanations must preserve the CNN’s original decision logic. That is, explanations must satisfy  $P_4$ .

**Derivation:** For an explanation  $\hat{f}(\cdot)$  and a prediction function  $f(\cdot)$ :  $\forall x \in \mathcal{X}, f(x) = f'(x \mid \hat{f}(x))$ , where  $f'(\cdot)$  is the surrogate function reconstructed from the explanations.

**Orthogonality:**

- Independent of Representation ( $P_5$ ): surrogate could match outputs without grounding in CNN internals.
- Independent of Causality ( $P_2$ ): causal concepts may not cover the full predictive logic.

The *Fidelity* of a concept explanation ensures that the explanation must faithfully represent how the model arrived at its decision. In practical terms, for a *dog* classification task (as shown in Figure 1), the explanation should contain concepts such as *Ear*, *Eye*, etc., which accurately reflects the model’s decision to classify the instance as a *dog*.

**Definition 4** (Fidelity of a Concept Explanation). *A concept explainer  $\hat{f}$  satisfies Fidelity for a class question  $Q$  if, for every explanation  $L \in \hat{f}(Q)$ , the following holds:*

- *The explanation  $L$  is equivalent to the ground truth for the class question  $Q$ .*
- *That is, the explanation  $\hat{f}(x)$  should reflect the decision-making process of the model  $f(x)$  as closely as possible.*

Formally,

$$L = \text{ground truth}(Q) \quad \text{and} \quad \hat{f}(x) = f(x), \quad (24)$$

where the set of concepts in  $L$  corresponds to the true concepts used by the model  $f(x)$  for prediction.

**Lemma 6** (Relative Accuracy of Fidelity). *If a concept explainer  $\hat{f}$  satisfies Fidelity, then the concept explanation faithfully reflects the model’s decision with relative accuracy. The relative accuracy of the concept explanation  $\hat{f}(x)$  is given by:*

$$g_4 = \text{Relative Accuracy}(\hat{f}(x), f(x)) = \frac{|L \cap \text{ground truth}(Q)|}{|L \cup \text{ground truth}(Q)|} \quad (25)$$

where  $|L|$  denotes the cardinality (number of elements) of the set  $L$ , and  $|\cdot|$  represents the size of the set.

**Theorem 8** (Fidelity Ensures Faithfulness). *If  $\hat{f}$  satisfies Fidelity, then the concept explanation  $L$  is a faithful reflection of the model’s decision  $f(x)$ , i.e., the explanation contains the same or similar concepts that were used by the model to make the prediction. Formally, for any class question  $Q$ , the following holds:*

$$L = \text{ground truth}(Q) \quad \text{and} \quad \hat{f}(x) = f(x), \quad (26)$$

ensuring that the concept explainer  $\hat{f}$  produces explanations that align with the true reasoning of the model.

*Proof.* The *Fidelity* of a concept explanation requires that the explanation accurately reflects the decision process of the model. Thus, the set of concepts in the explanation  $L$  must match the ground truth of the decision-making process of the model. The relative accuracy defined in the lemma ensures that the explanation is a faithful reflection of the model’s decisions. Hence, the theorem holds.  $\square$

**Corollary 4** (Bounded Fidelity). *The Fidelity score  $g_4$  is bounded within the range:*

$$0 < g_4 \leq 1 \quad (27)$$

*A score closer to 1 indicates Fidelity of explanations for an instance, whereas a score closer to 0 indicates low Fidelity.*

## B.5 SANITY

**Necessity:** Explanations must connect to the CNN’s internal mechanisms; otherwise, they are post-hoc stories. That is, explanations must satisfy  $P_5$ . If  $\hat{f}$  is independent of internal features  $E(x)$ , then explanations are arbitrary.

**Derivation:** Let  $E(\cdot) : \mathcal{X} \rightarrow \mathbb{R}^k$  be the CNN representation function up to a latent layer. Then for any explanation  $\hat{f}(x)$ ,  $\forall s \in \hat{f}(x)$ ,  $\exists g_s : \mathbb{R}^k \rightarrow \{0, 1\}$  such that  $g_s(E(x)) = 1$ .

**Orthogonality:**

- Cannot be replaced by Human Alignment ( $P_1$ ), because interpretability alone does not ensure internal grounding.
- Cannot be derived from Faithfulness ( $P_4$ ), since perfect output alignment does not prove link to internal features.

The *Sanity* of a concept explanation implies that the concepts used in the explanation should be representative of the internal CNN mechanisms under dynamic conditions, such as slight modifications in the input that do not change the prediction of the underlying class. Consider a CNN that classifies an image of a *dog*. If the image’s background colour is altered or if the image is rotated, the model should still classify the image as a dog, and the explanation should remain similar. For example, concepts such as *Ear* or *Eye* should still appear in the explanation after such transformations. This ensures that the explanation is stable under adversarial conditions.

**Definition 5** (Sanity of Concept Explanation). *A concept explainer  $\hat{f}$  satisfies the Sanity axiom if, for any two class questions  $Q = \langle T, C, y \rangle$  and  $Q' = \langle T', C', y' \rangle$  such that  $T \neq T'$ ,  $C = C'$ , and  $y = y'$ , the concept explanations  $L \in \hat{f}(Q)$  and  $L' \in \hat{f}(Q')$  must be disjoint:*

$$L \cap L' = \emptyset \quad \text{if} \quad x \neq x', \quad \hat{f}(x) \neq \hat{f}(x') \quad (28)$$

*The sanity score  $g_{10}$  is computed as:*

$$g_{10} = \text{Sanity}(\hat{f}(x) \parallel \hat{f}(x')) \equiv g_i^x \parallel g_i^{x'} \quad (29)$$

*In the case where  $Q = Q'$ , the result is indeterminate, denoted by  $\star$ .*

**Lemma 7** (Sanity of Concept Explanations under Adversarial Conditions). *If a concept explainer  $\hat{f}$  satisfies the Sanity axiom, then for any two class questions  $Q$  and  $Q'$  with the same class prediction  $y$ , and differing inputs  $x$  and  $x'$ , the explanations must remain disjoint, i.e., the concept explanations for  $x$  and  $x'$  should not overlap:*

$$L \cap L' = \emptyset \quad \text{if} \quad x \neq x', \quad \hat{f}(x) \neq \hat{f}(x') \quad (30)$$

*This ensures that adversarial changes to  $x$  do not drastically alter the concept explanations for the same class prediction.*

**Theorem 9** (Representation of Explanations under Input Modifications). *Let  $x$  and  $x'$  be two instances with the same class prediction  $y$  and different representations, such as changes in background colour or rotation. If a concept explainer  $\hat{f}$  satisfies the Sanity axiom, then modifying  $x$  to  $x'$  should not result in drastic changes to the explanation, i.e., the concept explanation should remain stable:*

$$\hat{f}(x) \approx \hat{f}(x') \quad (31)$$

*Specifically, minor transformations, such as changes in background colour or image rotation, should not alter the essential concepts used in the explanation.*



*Proof.* The proof of the sanity axiom hinges on the concept of explanation stability under adversarial conditions. In practice, small modifications like image rotation or background colour changes do not affect the class label of the model. Therefore, the concept explanation provided by  $\hat{f}$  should remain the same or only change slightly, while maintaining stability in the set of concepts that explain the prediction.  $\square$

**Corollary 5** (Sanity Score). *The Sanity score  $g_5$  is defined as a binary or indeterminate value:*

$$g_5 \in \{\text{True}, \text{False}, \star\} \quad (32)$$

where:

- *True: The explanation is stable under adversarial conditions.*
- *False: The explanation is unstable under adversarial conditions.*
- *$\star$ : The result is indeterminate if  $Q = Q'$ .*

A *True* value indicates that the concept explanation is robust and reliable, while a *False* value indicates that the explanation is sensitive to small changes in the input.

## C MORE EXPERIMENTS

### C.1 CLASS QUESTIONS AND CLASS EXPLANATIONS

We present more random *class questions* and *class explanations* in this section. It would be futile to present all possible *concept explanations* for all the classes in all the datasets using all the pre-trained models. However, the results presented provide a more nuanced understanding of our work’s contributions.

#### C.2 *How do the class explanations for the ICE Zhang et al. (2021) explainer compare for different CNNs for an instance featuring an Eagle and Macaw?*

The query seeks to understand how the *class questions*  $Q_4 - Q_9$  compare helps draw a meaningful conclusion. This results in 6 unique class questions (3 class explanations for 2 classes), namely:

- $Q_4$ : *What concept explanations support a ResNet18’s decision for an Eagle?*
- $Q_5$ : *What concept explanations support a ResNet18’s decision for a Macaw?*
- $Q_6$ : *What concept explanations support a ResNet50’s decision for an Eagle?*
- $Q_7$ : *What concept explanations support a ResNet50’s decision for a Macaw?*
- $Q_8$ : *What concept explanations support an Inceptionv3’s decision for an Eagle?*
- $Q_9$ : *What concept explanations support a Inceptionv3’s decision for a Macaw?*

Figure A7, A8, and A9 presents the *class explanations* for  $Q_4 - Q_7$  at  $c' = 32$  for ResNet18, ResNet50, and Inceptionv3, respectively, for an *instance* featuring an Eagle and Macaw. The red-bound pixels represent concepts. Each explanation includes a unique ID, five prototypes, and a  $\mathcal{W}$  score sorted in descending order by  $\mathcal{W}$  scores. All three CNNs achieved over 96% *classification accuracy* ( $Acc$ ) for the Eagle and Macaw classes. The proposed axiomatic foundations are shown in rectangular boxes below each class explanation. At the bottom of each figure are the global explanation plots for each class. The black dotted lines represent the feature map predictions (ground truth) while the blue dots represent the CAV predictions.

Each *class explanation* shown in Figures A7, A8, and A9 is grounded in the proposed axiomatic framework ( $g_1$  to  $g_5$ ). For clarity and interpretability, only the top three local *concept explanations*, identified by their unique IDs, representative *prototypes*, and corresponding  $\mathcal{W}$  scores, are displayed. These concept-level insights are directly linked to the axiomatic properties, reinforcing trust in the CNN’s decision-making process and offering a transparent lens into the model’s internal logic.

A qualitative comparison across the three figures reveals that similar concepts yield visually similar prototypes, regardless of the underlying CNN model. This alignment underscores the axiom

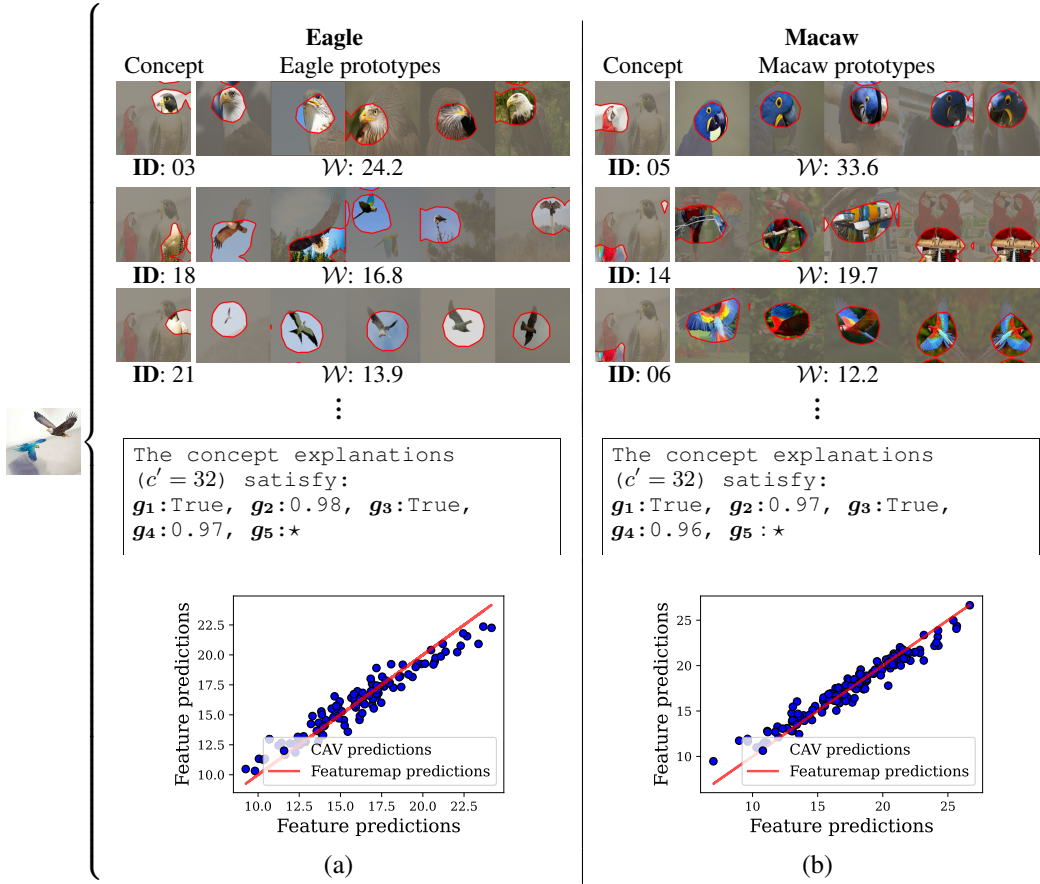


Figure A7: The three top *concept explanations* for an *instance* featuring an Eagle and a Macaw, evaluated across their axiomatic foundations ( $g_1, \dots, g_{10}$ ) for the *class question*,  $Q_4$ . (a) ICE Zhang et al. (2021) on ResNet18 for Eagle and (b) ICE Zhang et al. (2021) on ResNet18 for Macaw. (*Best viewed in colour*).

$g_1$ , ensuring that explanations remain *Interpretable*. The middle-tier axioms ( $g_2$  to  $g_4$ ) evaluate essential properties such as *Relevance*, *Coherence*, and *Fidelity*. High scores across these dimensions indicate that the generated explanations are faithful to the model and internally consistent and human-aligned. Crucially, the *Sanity* axiom  $g_5$  assesses the explainer’s robustness under perturbations. It outputs boolean values (*True* or *False*) when comparing the behaviour of any of the axioms  $g_2$  to  $g_4$  across multiple *instances* for the same *class explanation*, and returns  $\star$  (indeterminate) when applied to single-instance evaluations. This design reflects its focus on comparative reliability.

Beyond individual concept quality, Figures A7, A8, and A9 enable a comparative analysis across models. Despite variations in architecture, all models show consistent alignment between concept-level predictions and corresponding feature map outputs, as evidenced by the global explanation plots. This cross-model agreement supports the model-agnostic nature of the proposed framework and suggests that the identified concepts capture fundamental decision patterns shared across different CNNs. These plots serve as a quantitative proxy for faithfulness. Specifically, lower dispersion and higher correlation between the CAV predictions and ground truth indicate greater faithfulness. The reduced sparsity of CAV activations signals strong alignment with model outputs, thereby validating that the proposed concept explanations are informative but also accurate and robust.

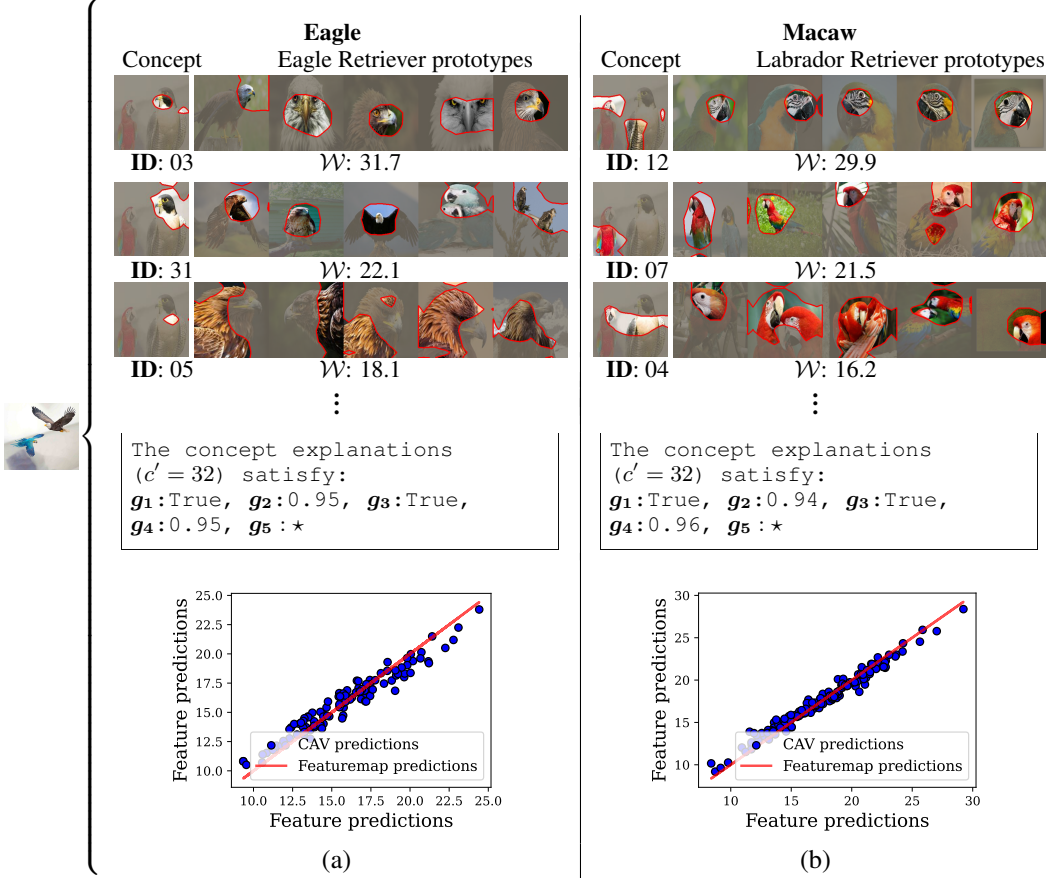


Figure A8: The three top *concept explanations* for an *instance* featuring an Eagle and a Macaw, evaluated across their axiomatic foundations ( $g_1, \dots, g_{10}$ ) for the *class question*,  $Q_4$ . (a) ICE Zhang et al. (2021) on ResNet50 for Eagle and (b) ICE Zhang et al. (2021) on ResNet50 for Macaw. (*Best viewed in colour*).

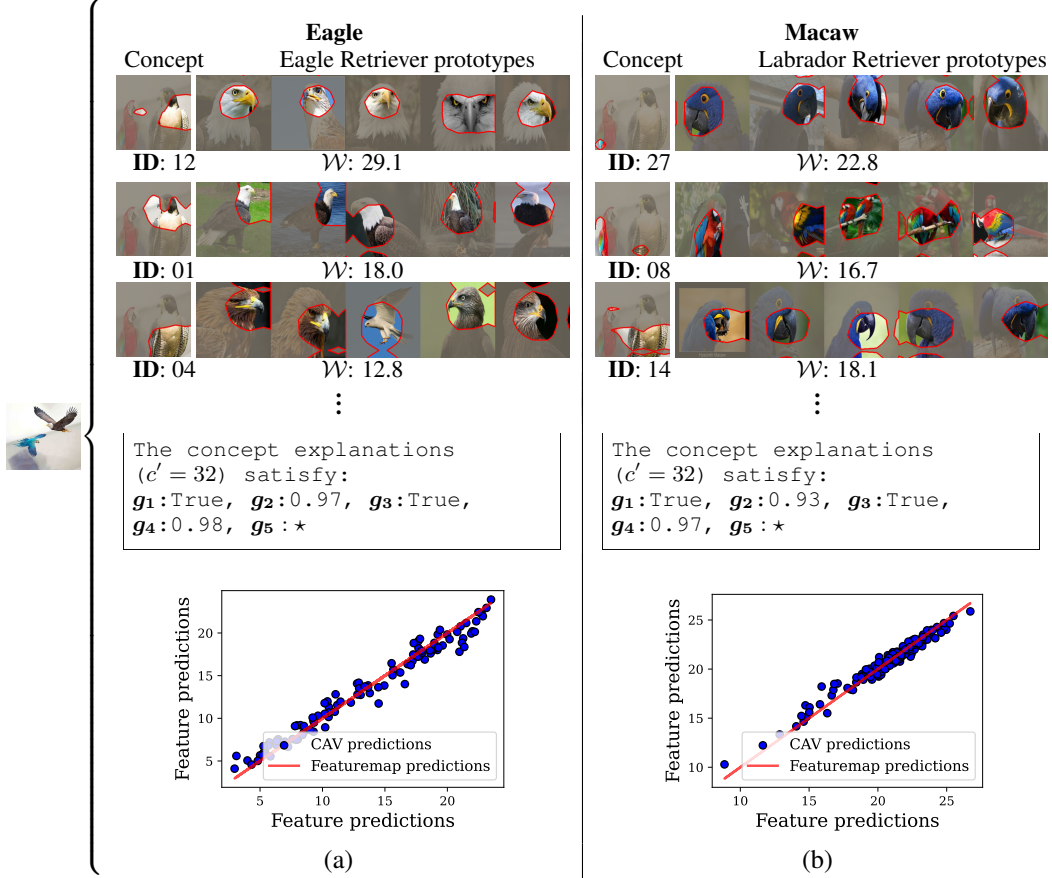


Figure A9: The three top *concept explanations* for an *instance* featuring an Eagle and a Macaw, evaluated across their axiomatic foundations ( $g_1, \dots, g_{10}$ ) for the *class question*,  $Q_4$ . (a) ICE Zhang et al. (2021) on Inceptionv3 for Eagle and (b) ICE Zhang et al. (2021) on Inceptionv3 for Macaw. (Best viewed in colour).

## D BROADER IMPACT

Concept-based explainers are valuable for understanding what a CNN has learned by identifying and visualizing the concepts the model uses for classification. However, their effectiveness is often constrained by the dimensionality reduction methods employed Akpudo et al. (2025a); Zhang et al. (2021); Fel et al. (2023); Ghorbani et al. (2019). Additionally, assigning meaningful labels to identified concepts becomes increasingly challenging as the number of user-defined concepts,  $c'$  grows. This process often requires creating a new ground truth, involving a pre-defined repository of concepts for representation. Despite advancements in automated explainers, confirming identified concepts frequently relies on expert judgment or oracle networks for validation. On the bright side, while our work does not directly solve this problem, it offers sufficient human-centered axiomatic foundations for objectively evaluating the faithfulness of *concept explanations*, providing a quantitative paradigm for gaining trust in the concept explanations.

Designing cohesive, trustworthy, and ethically compliant explainability frameworks requires understanding the interrelationships among the axioms Zhang et al. (2022); Pinhanez et al. (2023). Robust explanations are essential for the practical deployment of AI systems, and methods grounded in well-defined axiomatic foundations enhance the reliability of XAI. However, trade-offs exist; for instance, overly simple explanations may omit critical predictive factors, compromising *Interpretability*, while resolution constraints can weaken visual explanations, raising concerns about an explainer’s *Success*. *Interpretability* and *Simplicity* enable user personalization and foster human-AI collaboration through interactive tools that enhance transparency, interoperability, self-explanation, and real-time integration with minimal disruption. Additionally, *Fidelity*, *Coherence*, *Agreement*, *Relevance*, and *Causality* strengthen credibility and utility, supporting broader objectives such as generalizability, scalability, and efficiency. Finally, Sanity checks improve reliability through ablation studies and multi-level assessments, ensuring transparency, safety, and fairness. Striking a balance between these principles is critical for developing trustworthy explainability frameworks Wang et al. (2023c).

Our proposed axiomatic foundations offer robust quantitative and deterministic checkpoints to evaluate concept-based explainers, ensuring that they are accurate and resilient to adversarial conditions. The study highlights the importance of these axioms in maintaining the reliability and trustworthiness of the explainers. Research has demonstrated the importance of these axioms in enhancing the reliability and interpretability of CNNs, particularly in domains where understanding model decisions is crucial. Rigorous evaluation using the proposed axioms is essential to build trust and transparency in AI systems Wang et al. (2023c). Although the findings validate the sanity of concept-based explainers under adversarial attacks, additional measures may be necessary to protect explainers from such conditions Karimi et al. (2020). This presents both a challenge and a promising research opportunity for developing robust and reliable explainers.