

VISIONLOGIC: FROM NEURON ACTIVATIONS TO CAUSALLY GROUNDED CONCEPT RULES FOR VISION MODELS

006 **Anonymous authors**

007 Paper under double-blind review

ABSTRACT

013 While concept-based explanations improve interpretability over local attributions,
 014 they often rely on correlational signals and lack causal validation. We introduce
 015 VISIONLOGIC, a novel neural-symbolic framework that produces faithful, hi-
 016 erarchical explanations as global logical rules over causally validated concepts.
 017 VISIONLOGIC first learns activation thresholds to convert neuron activations into
 018 a reusable predicate vocabulary and induces class-level logical rules from these
 019 predicates. It then grounds predicates to visual concepts via ablation-based causal
 020 tests with iterative region refinement, ensuring that discovered concepts correspond
 021 to features that are causal for predicate activation. Across different vision architec-
 022 tures such as CNNs and ViTs, it produces interpretable concepts and compact rules
 023 that largely preserve the original model’s predictive performance. In our large-scale
 024 human evaluations, VISIONLOGIC’s concept explanations significantly improve
 025 participants’ understanding of model behavior over prior concept-based methods.
 026 VISIONLOGIC bridges neural representations and symbolic reasoning, providing
 027 more trustworthy explanations suited for safety-critical applications.

1 INTRODUCTION

031 Deep learning-based vision models have achieved remarkable success across numerous tasks, yet
 032 their *black-box* nature remains a major obstacle to trustworthy AI. This challenge has only intensified
 033 with the transition from Convolutional Neural Networks (CNNs) (LeCun et al., 1998; He et al.,
 034 2016) to Vision Transformers (ViTs) (Dosovitskiy et al., 2021), which introduce greater architectural
 035 complexity and opacity. To address this issue, a wide range of interpretability methods have been
 036 proposed. Among them, concept-based explanations (Nguyen et al., 2016; Bau et al., 2017; Kim
 037 et al., 2018; Ghorbani et al., 2019; Fel et al., 2023) have attracted particular interest because they
 038 uncover high-level semantic concepts rather than low-level attribution maps such as Grad-CAM and
 039 its extensions (Selvaraju et al., 2017; Wang et al., 2020; Jiang et al., 2021).

040 However, existing concept-based methods rely almost entirely on *correlational evidence* without
 041 *causal validation*, leading to potentially unfaithful or misleading explanations (Lopez-Paz et al., 2017;
 042 Zhang et al., 2023). For instance, TCAV (Kim et al., 2018) uses linear classifiers in activation space,
 043 while ACE (Ghorbani et al., 2019) applies clustering—both purely correlational approaches that lack
 044 rigorous causal foundations. As a result, these methods often conflate dataset biases with genuine
 045 model reasoning, such as associating the concept *pasture* with the class *cow*, a classic case where
 046 correlation fails to imply causation (Wu et al., 2023). Consequently, the concepts themselves may be
 047 spurious, leaving a fundamental methodological gap: the absence of *principled causal validation* for
 048 robust, interpretable concepts.

049 We address this gap with VISIONLOGIC, a novel neural-symbolic framework for generating faithful,
 050 interpretable explanations via global logical rules defined over causally validated concepts. Our
 051 approach operates in two stages. First, we transform high-level neuron activations into abstract
 052 predicates by learning activation thresholds and derive logical rules that approximate the model’s
 053 class-level decision making. These predicates provide an intermediate symbolic representation that
 captures key aspects of the model’s internal reasoning while remaining flexible and generalizable
 across input images. By converting neuron activations into predicates, we not only abstract the

054 model’s computations but also create a structured foundation for reasoning about causally relevant
 055 concepts at a higher semantic level.
 056

057 Second, we ground these predicates into high-
 058 level visual concepts using ablation-based causal
 059 tests. For each image, we start with an initial
 060 bounding box likely to influence the predicate,
 061 perturb the region with random noise or similar
 062 masking strategies, and check whether this flips
 063 the predicate’s truth value. A transition from ac-
 064 tivation to deactivation provides causal evidence
 065 that the region is critical for the predicate. We
 066 then propose an efficient algorithm to iteratively
 067 refine the bounding box for more precise localiza-
 068 tion. For further refinement, segmentation meth-
 069 ods such as Mask R-CNN (He et al., 2017) or
 070 SAM (Kirillov et al., 2023) are used to validate
 071 the intersection of the segmentation and refined
 072 box. Finally, the refined regions are consolidated
 073 across images within the same class to form con-
 074 sistent, causally validated visual concepts.
 075

076 The final result is a set of causally validated con-
 077 cepts and global logical rules that collectively pro-
 078 vide transparent, faithful explanations of model
 079 behavior. For example, VISIONLOGIC discovers
 080 concepts such as *Beak* and *Claw* for the class
 081 *House Finch* in ImageNet (Deng et al., 2009),
 082 which are combined to form the logical rule: $Beak \wedge Claw \Rightarrow House Finch$, as illustrated in Figure 1.
 083 This rule reveals how the vision model leverages these causally validated visual concepts to make
 084 class-level predictions, providing interpretable insights into the model’s decision-making process.
 085

086 Our extensive human studies confirm that the concepts discovered by VISIONLOGIC significantly
 087 enhance understanding of the model’s decision-making process compared to prior concept-based
 088 methods. Experiments on both CNNs and ViTs further demonstrate that VISIONLOGIC maintains
 089 strong predictive performance, achieving over 90% top-5 test accuracy on covered images, while
 090 providing explanations that are both causally grounded and human-understandable. *To the best of
 091 our knowledge, VISIONLOGIC is the first framework to deliver both causally validated concepts and
 092 interpretable logic-rule explanations.* We envision this as an important step toward bridging the gap
 093 between complex neural network representations and human-interpretable causal reasoning, offering
 094 trustworthy insights for high-stakes applications. Our contributions are summarized as follows:
 095

- 096 • We propose VISIONLOGIC, a novel neural-symbolic framework that learns activation thresh-
 097 olds to form predicate-based abstractions and extracts logical rules over causally validated
 098 high-level visual concepts, bridging symbolic reasoning with neural representations.
- 099 • We develop an efficient, iterative refinement algorithm that precisely localizes causally
 100 relevant image regions using bounding-box adjustment and segmentation masks, ensuring
 101 accurate and consistent concept discovery.
- 102 • We conduct a large-scale human evaluation demonstrating significant improvements over
 103 state-of-the-art concept-based explanation methods in understanding the model’s decision-
 104 making process through causally validated concepts and human-aligned interpretability.
- 105 • We empirically show that VISIONLOGIC largely retains the discriminative power of vision
 106 models with compact rules, and its grounded concepts are highly interpretable to humans,
 107 thereby providing a strong tool for understanding decision-making in vision models.

108 2 RELATED WORK

109 Existing approaches to interpreting vision models increasingly focus on *concept-based methods*,
 110 which aim to link internal representations to human-interpretable concepts rather than pixel-level

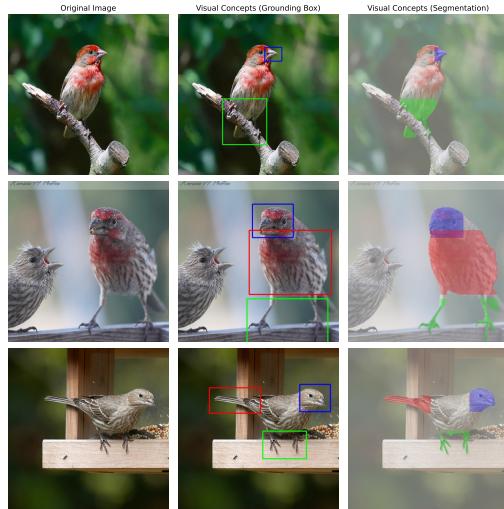


Figure 1: Causally validated concepts discovered by VISIONLOGIC, highlighted with bounding boxes and colored overlays.

108 attributions (Selvaraju et al., 2017; Chattopadhyay et al., 2018; Wang et al., 2020). Early studies
 109 explore the *hidden semantics* of neural networks by visualizing what individual neurons or layers
 110 encode. For instance, Nguyen et al. (2016) generate synthetic images that maximally activate specific
 111 neurons, while Mahendran & Vedaldi (2015) reconstruct inputs from intermediate feature maps to
 112 reveal the information preserved at different network depths. Network Dissection (Bau et al., 2017)
 113 further provides a systematic framework to quantify the alignment between hidden units and semantic
 114 concepts across diverse architectures. Although these methods offer valuable qualitative insights into
 115 model representations, they remain largely descriptive and lack systematic approaches to associate
 116 neurons with semantically meaningful concepts.

117 Subsequent work introduces *concept discovery methods* to establish explicit links between model
 118 activations and human-interpretable concepts. NET2VEC (Fong & Vedaldi, 2018) trains linear
 119 predictors on activation patterns to align feature maps with semantic labels, while TCAV (Kim et al.,
 120 2018) uses linear classifiers to score concept importance in activation space. However, all these
 121 approaches rely almost entirely on *correlational evidence* without any form of *causal validation*. As
 122 a result, they may capture spurious correlations between concepts and model decisions—for example,
 123 a concept might appear predictive of a class simply because both frequently co-occur in the training
 124 data, even if it plays no causal role in the model’s reasoning (Wu et al., 2023).

125 More recent methods have attempted to refine concept discovery. For instance, ACE (Ghorbani et al.,
 126 2019) applies unsupervised clustering to extract concepts directly from activation patterns, while
 127 ICE (Zhang et al., 2021) improves upon ACE by introducing invertible concept-based explanations
 128 and leveraging Non-Negative Concept Activation Vectors to enhance interpretability and fidelity.
 129 CRAFT (Fel et al., 2023) further integrates sensitivity analysis into concept scoring to better measure
 130 how concept perturbations affect model predictions. Nevertheless, all these approaches still rely
 131 on unsupervised discovery techniques such as clustering or matrix factorization, which provide no
 132 causal guarantees (Lopez-Paz et al., 2017; Zhang et al., 2023). Consequently, the identified concepts
 133 themselves may be spurious, leaving a fundamental methodological gap: the lack of principled causal
 134 validation for robust and trustworthy interpretability in deep vision models.

135 3 THE VISIONLOGIC FRAMEWORK

137 VISIONLOGIC explains deep vision models by replacing the final decision layer with an interpretable
 138 program over *causally validated* concepts. The framework proceeds in three stages: (i) derive binary
 139 predicates from neuron activations, (ii) compose these predicates into class-wise rules and define an
 140 inference score, and (iii) ground predicates to image regions via *occlusion ablation*.

142 3.1 DERIVING PREDICATES FROM NEURON ACTIVATIONS

144 We begin with the final-layer activations $\mathbf{Z}(x) \in \mathbb{R}^d$ for an input x . For class c , the network computes
 145 the logit $\mathbf{F}^c(x) = \mathbf{W}^c \mathbf{Z}(x) + \mathbf{b}^c$, with $\mathbf{W}^c \in \mathbb{R}^d$ and $\mathbf{b}^c \in \mathbb{R}$. The term $\mathbf{W}_j^c \mathbf{z}_j(x)$ measures how
 146 much channel j pushes toward class c on this input. We sort channels by this per-example contribution
 147 so that rank 1 is the largest. Averaged over examples of class c , the *representativeness* of channel j is
 148 captured by its expected rank; the most representative channel minimizes:

$$149 \quad j^* = \arg \min_{j \in J} \mathbb{E}_{x \in X_c} [\mathcal{R}^c(\mathbf{W}_j^c \mathbf{Z}_j(x))]. \quad (1)$$

151 Since \mathbf{b}^c adds a constant to $\mathbf{F}^c(x)$, it cannot change the within-example ordering of contributions
 152 $\{\mathbf{W}_j^c \mathbf{Z}_j(x)\}_j$. The ranking is thus bias-invariant (Geng et al., 2022). For clarity, all classwise
 153 statistics in Section 3 use only training examples correctly classified by the base model, preventing
 154 predicate learning from contamination by misclassified instances.

155 We convert real-valued activations to *binary predicates* $p_j(x) \in \{0, 1\}$ that serve as logical atoms.
 156 Rather than fixing ad hoc thresholds, we learn per-channel thresholds T_j and *sharpness* $s_j > 0$ (note:
 157 temperature = $1/s_j$), which define a differentiable gate during training and a Boolean at test time:

$$159 \quad \tilde{p}_j(x) = \sigma(s_j(\mathbf{z}_j(x) - T_j)), \quad p_j(x) = \mathcal{I}(\mathbf{z}_j(x) \geq T_j). \quad (2)$$

160 The relaxed gate \tilde{p}_j enables gradient-based learning of (T_j, s_j) ; after training, we harden to p_j . We
 161 call a predicate *invalid* if its learned threshold makes $p_j(x) \equiv 0$ on all data; otherwise it is *valid*.

162 Note that some activation functions such as GELU (Hendrycks & Gimpel, 2016) can produce positive
 163 and negative responses with distinct semantics. To allow a single channel to encode two features, we
 164 define branch-specific predicates:

$$166 \quad p_{j,+}(x) = \mathcal{I}(\mathbf{z}_j(x) \geq T_{j,+}), \quad p_{j,-}(x) = \mathcal{I}(\mathbf{z}_j(x) \leq T_{j,-}), \quad (3)$$

167 trained with the same relaxation as Eq. 2 using branch-specific $(T_{j,\pm}, s_{j,\pm})$.

169 A channel can be informative even when it is not the largest contributor on an example. We therefore
 170 require *both* high contribution rank and sufficient activation magnitude. We adopt a *single*, class-
 171 agnostic threshold T_j shared across classes so that each predicate simply denotes “feature present”
 172 for neuron j ; this deliberately *handles polysemanticity by reuse*—the same predicate may activate in
 173 multiple classes (Elhage et al., 2022). For class c and input x , let $u_j^c(x) = \mathbf{W}_j^c \mathbf{z}_j(x)$ and $\mathcal{R}^c(u_j^c(x))$
 174 be its within-example rank (1 is best). We define

$$175 \quad p_{j,\leq k}(x) = \mathcal{I}(\mathcal{R}^c(u_j^c(x)) \leq k \wedge \mathbf{z}_j(x) \geq T_j), \quad k \in \{1, 2, 3\}. \quad (4)$$

177 During training, we replace the non-differentiable rank test with a *soft top- k* weight $w_j^c(x) \in$
 178 $[0, 1]$ computed with *SoftSort* (Prillo & Eisenschlos, 2020), which approximates the indicator
 179 $\mathcal{I}(\mathcal{R}^c(u_j^c(x)) \leq k)$. The relaxed gate is $\tilde{p}_{j,\leq k}(x) = w_j^c(x) \cdot \sigma(s_j(\mathbf{z}_j(x) - T_j))$. We instantiate
 180 a small set of rank windows $k \in \{1, 2, 3\}$ and impose *structured sparsity* (group lasso) across
 181 $\{p_{j,\leq 1}, p_{j,\leq 2}, p_{j,\leq 3}\}$ so the optimizer selects *at most one* k per channel (i.e., the most predictive
 182 window), preventing predicate proliferation. The top- k gate, combined with the shared threshold T_j ,
 183 suppresses spurious activations and keeps the predicate vocabulary compact and easy to learn.

185 **Learning objective.** Given the predicate vector $P(x) = [p_1(x), \dots, p_m(x)]^\top$, we train a
 186 lightweight linear head $f_{\text{rule}}(x) = \mathbf{W}_{\text{rule}} P(x) + \mathbf{b}_{\text{rule}}$ with the base network frozen. *At test time we*
 187 *do not use* f_{rule} ; it serves solely to learn stable thresholds. The head supplies calibrated logits and,
 188 crucially, gradients that help place thresholds during learning:

$$189 \quad \min_{\Theta_{\text{pred}}, \Theta_{\text{rule}}} \underbrace{\mathcal{L}_{\text{teach}}(f_{\text{rule}}(x), f_{\text{nn}}(x))}_{\text{distill the frozen teacher}} + \underbrace{\lambda_T \|T - T^{(0)}\|_2^2 + \lambda_s \sum_j (s_j - 1)^2}_{\text{threshold/temperature stability}} + \underbrace{\lambda_{\text{use}} \Omega(\tilde{P})}_{\text{compact predicate set}}, \quad (5)$$

194 where $\Theta_{\text{pred}} = \{T, s\}$, $\Theta_{\text{rule}} = \{\mathbf{W}_{\text{rule}}, \mathbf{b}_{\text{rule}}\}$, and \tilde{P} collects the relaxed gates $\tilde{p}_j(x) = \sigma(s_j(\mathbf{z}_j(x) -$
 195 $T_j))$. The distillation loss $\mathcal{L}_{\text{teach}}$ is the *Kullback–Leibler divergence* between the teacher and rule-head
 196 predictive distributions. The per-channel seed $T^{(0)}$ is initialized at a high percentile of $\mathbf{z}_j(x)$ over
 197 influential training examples (we use the 0.8-quantile; influential = *SoftSort* top- k by contribution
 198 with $k = 3$). We initialize \mathbf{W}_{rule} with classwise normalized predicate frequencies (mean-centered
 199 by the global frequency per predicate) and set \mathbf{b}_{rule} from class priors; values are then refined
 200 during learning. The stability terms keep thresholds near $T^{(0)}$ and temperatures near 1, preventing
 201 degenerate always-on/off gates. The compactness penalty $\Omega(\tilde{P})$ is a group-lasso over the rank variants
 202 $\{p_{j,\leq 1}, p_{j,\leq 2}, p_{j,\leq 3}\}$ for each channel, which selects a single rank regime and prevents predicate
 203 proliferation. More implementation details and hyperparameters are provided in the Appendix C.2.

204 *Observation.* Empirically, the learned thresholds T_j often align with the $k=1$ specialization of the
 205 rank-aware predicate in Eq. 4: restrict to correctly classified examples from the class $c^*(j)$ where
 206 neuron j is most representative (i.e., has the lowest expected contribution rank), keep those instances
 207 $x \in X_{c^*(j)}$ with $p_{j,\leq 1}^{(c^*)}(x) = 1$ (i.e., $\mathcal{R}^{c^*(j)}(u_j^{c^*(j)}(x)) = 1$), and observe that T_j tends to be close
 208 to the minimum $\mathbf{z}_j(x)$ over that subset. With *SoftSort* providing the soft top- k proxy and structured
 209 sparsity selecting a single k per channel, training recovers this heuristic in a data-driven way.

211 3.2 LEARNING LOGICAL RULES AND AN INFERENCE SCORE

213 Given the learned predicate vocabulary P , we induce symbolic rules and a *rank-based* inference
 214 score to explain the base model’s predictions. For class c , we evaluate all predicates on each training
 215 example $x \in X_c$ that the base model classifies correctly, obtaining a binary vector $P(x) \in \{0, 1\}^m$.
 Each distinct vector defines a conjunctive clause that requires exactly the predicates that are true

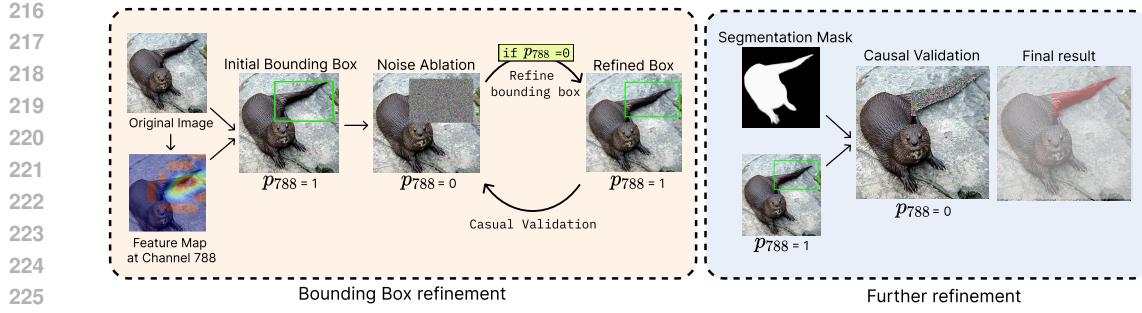


Figure 2: **Grounding predicates to visual concepts.** The orange panel illustrates how a bounding box is iteratively refined to capture the candidate region that causally influences an example predicate p_{788} via noise ablation. The blue panel shows a further refinement step using segmentation masks.

and forbids those that are false. Joining all such clauses with disjunctions yields a DNF that exactly captures the training patterns of class c :

$$\forall x, \left(\bigvee_{v \in \mathcal{V}_c} \left(\bigwedge_{i: v_i = 1} p_i(x) \wedge \bigwedge_{i: v_i = 0} \neg p_i(x) \right) \right) \Rightarrow \text{Label}(x) = c, \quad (6)$$

where \mathcal{V}_c is the set of unique predicate patterns observed in X_c . Because exact matching is brittle on unseen data, we summarize class *characteristic strength* with a rank profile.

We build a class profile by counting predicate appearances across the class- c clauses in Eq. 6 and sorting predicates by frequency to obtain a ranking $\mathcal{R}^c(p_i)$ (lower rank = more characteristic of c). For a test input x with active predicates $P(x)$, we compute the explanation score:

$$S(x, c) = \frac{1}{|P(x)|} \sum_{p_j \in P(x)} \mathcal{R}^c(p_j), \quad (7)$$

and predict with $\hat{c}(x) = \arg \min_c S(x, c)$. Intuitively, the chosen class is the one whose characteristic predicates best explain those active on x . Because predicates use a *single*, class-agnostic threshold T_j per neuron, a predicate can fire in multiple classes (i.e., polysemy), but disambiguation comes from the class profiles $\{\mathcal{R}^c\}$: the same predicate typically has different ranks across classes, so the score $S(x, c)$ separates them.

Observation. Recall we initialize \mathbf{W}_{rule} with classwise normalized predicate frequencies (mean-centered by each predicate’s global frequency). Since the class profile ranking $\mathcal{R}^c(\cdot)$ is induced by these frequencies, the resulting weights are, up to a positive monotone transform, equivalent to using inverse ranks (i.e., $\mathbf{W}_{\text{rule}}^{c,i} \propto 1/\mathcal{R}^c(p_i)$). Consequently, maximizing $f_{\text{rule}}^c(x)$ is monotone-equivalent to minimizing $S(x, c)$. A formal proof is given in Claim 1 (see Appendix C.2).

3.3 GROUNDING PREDICATES TO VISION CONCEPTS

The final step grounds abstract predicates in the input space, linking logical atoms to interpretable visual features. Our binary predicate formulation enables principled causal reasoning—a key advantage over existing concept-based methods that rely on statistical correlations without establishing genuine causal relationships between visual features and model decisions.

Given an image x and predicate p_j , we test whether a region is necessary for $p_j(x)$ by replacing that region with random noise¹ to obtain x' , then recomputing $p_j(x')$. A flip from activation to deactivation indicates that the region is causally important for $p_j(x)$. To find such regions efficiently, we initialize a bounding box intended to deactivate p_j (typically large and image-covering). For CNNs, the initialization can be seeded from the feature map associated with p_j ; for ViTs, it is aligned to the patch grid. We then iteratively refine the box until p_j reactivates, following a procedure similar

¹Other replacements, e.g., blurring, mean-filling, or blacking out, are also effective; in our experiments, blurring often performs best. We use random noise here for clarity and consistency in visualization.

270
 271 Table 1: Utility scores in three application scenarios. The Utility benchmark measures how well
 272 explanations help users identify general rules that transfer to unseen instances. During training,
 273 participants are shown images along with the explanations and model predictions, and are asked to
 274 infer the underlying decision rules. At test time, the benchmark evaluates participants’ accuracy in
 275 predicting the model’s output on novel images. Higher Utility scores indicate that the explanations
 276 provide more useful information for understanding the model’s behavior on new samples. For each
 277 scenario, the first and second best results are in **bold** and underlined respectively. *Our proposed*
 278 *method VISIONLOGIC achieves consistently higher Utility scores than prior approaches, with*
 279 *statistically significant improvements in the explanatory model’s behavior across all three scenarios.*

Session n°	Husky vs. Wolf				Otter vs. Beaver				Kit Fox vs. Red Fox			
	1	2	3	Utility	1	2	3	Utility	1	2	3	Utility
Baseline	65.7	68.6	70.3	1.00	84.4	90.3	92.2	<u>1.00</u>	84.1	89.0	84.1	1.00
Control	55.3	63.6	70.0	0.92	85.1	88.3	92.9	<u>1.00</u>	80.8	79.2	79.2	0.93
ACE	60.4	71.1	74.6	<u>1.01</u>	80.4	85.7	90.5	0.96	80.6	83.2	76.2	0.93
CRAFT	55.5	60.8	65.3	0.89	86.3	90.9	90.9	<u>1.00</u>	76.8	81.8	76.8	0.92
VISIONLOGIC	74.8	90.0	91.0	1.25	96.8	98.4	99.2	1.10	84.1	84.5	82.9	<u>0.98</u>

286
 287 to Geng et al. (2024); in some cases, only a few refinement steps are sufficient. The full algorithm
 288 is provided in Appendix A. Because the refinement is stochastic, multiple runs may yield different
 289 boxes; we retain the smallest successful box to improve precision. We also verify *sufficiency* by
 290 constructing an image with random noise everywhere except the candidate region and checking
 291 whether p_j remains activated. This provides causal evidence that the candidate region influences p_j .
 292

293 To better match object boundaries, we add a segmentation-based refinement step using off-the-shelf
 294 methods such as Mask R-CNN (He et al., 2017) or SAM (Kirillov et al., 2023). We intersect
 295 the segmentation mask with the refined box and repeat the intervention to confirm the expected
 296 predicate flip, thereby strengthening causal validity. Figure 2 illustrates the workflow. Finally, we
 297 aggregate validated regions across multiple images of the same class to form consistent, causally
 298 supported visual concepts, establishing a robust link between the concepts and p_j and ensuring that
 299 the explanations faithfully reflect the model’s decision-making.

4 EXPERIMENTS

4.1 HUMAN EVALUATION OF CAUSALLY GROUNDED CONCEPTS

300
 301 **Setup.** We evaluate the practical utility of our proposed VISIONLOGIC, alongside prior concept-
 302 based methods ACE and CRAFT, in terms of *concept explanations* using the human-in-the-loop
 303 framework of Colin et al. (2022), which assesses how well explanations help participants understand
 304 a model’s behavior across three real-world scenarios: (1) detecting bias in AI decisions (using Husky
 305 vs. Wolf classification), (2) identifying novel model strategies that are non-obvious to untrained
 306 observers (using Otter vs. Beaver classification²), and (3) understanding failure cases (using Kit Fox
 307 vs. Red Fox classification). Each scenario adopts the meta-predictor paradigm: during the training
 308 phase, participants study example images paired with explanation images and model outputs, then
 309 predict the model’s output on unseen images without access to the corresponding explanations.
 310

311 In total, 531 participants were recruited from Prolific (Prolific, 2024), in which 465 passed screening
 312 criteria aligned with Colin et al. (2022). We designed questionnaires for five conditions: (1) baseline
 313 (no explanation), (2) control (bottom-up saliency maps (Simonyan et al., 2014)), (3) ACE (Ghorbani
 314 et al., 2019), (4) CRAFT (Fel et al., 2023), and (5) VISIONLOGIC (ours), where ACE and CRAFT
 315 are considered state-of-the-art. To ensure fair comparison on the adapted datasets, all methods were
 316 re-run under identical experimental settings. Following prior work (Colin et al., 2022; Fel et al.,
 317 2023), we report the *utility score*, defined as participants’ average test accuracy across the three
 318 sessions, normalized by the baseline; higher values indicate more effective human understanding of
 319 the model. Additional experimental details, including the participant recruitment process and the
 320 number of images used in each session and phase, are provided in Appendix F.1.
 321

322
 323 ²This task replaces the legacy “Leaves” dataset from Colin et al. (2022) that is no longer available.

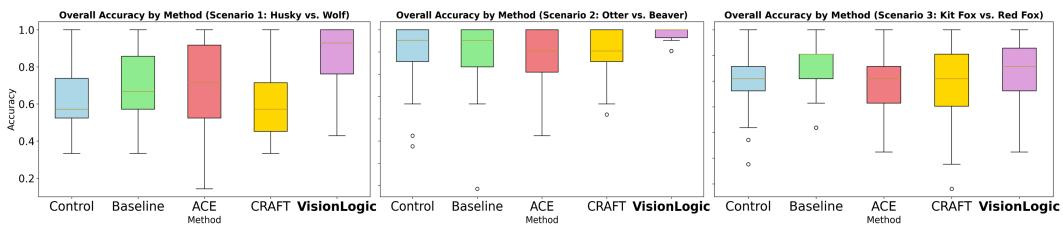


Figure 3: Data distribution for each explanation method in three scenarios. Our method VISIONLOGIC consistently enhances human understanding of model behavior over prior concept-based methods.

Table 2: Evaluation of VISIONLOGIC’s rule-based explanations across vision models. The first column lists the models; the remaining columns report results for the proposed metrics. The #Valid metric gives the number of *valid* predicates; the following parentheses show the total predicate count.

Model	#Valid	Complexity	Coverage (%)	Fidelity (%)	Top-1 Acc. (%)	Top-5 Acc. (%)
ResNet	1944 (2048)	9.49	83.48	75.42	69.27	93.53
ConvNet	1303 (2048)	33.75	83.77	86.07	80.34	97.23
ViT	1465 (1536)	42.63	80.48	87.55	80.70	97.38
Swin	1460 (1536)	53.58	88.64	78.60	72.83	91.26

Results. Table 1 shows that VISIONLOGIC consistently outperforms ACE and CRAFT across all three scenarios. In the first two scenarios, VISIONLOGIC achieves utility scores significantly higher than 1, demonstrating clear benefits provided by our method to the participants when assisting them to infer model predictions. In the third scenario, we observe the same trend as prior work (Fel et al., 2023) where no existing method provides more effective information than Baseline. Nevertheless, our method shows substantial improvement over ACE and CRAFT; a utility score of 0.98 suggests that causally grounded concepts provide actionable guidance for understanding model failures.

To rigorously examine the performance gain of our method over Baseline, Control, ACE, and CRAFT in the evaluation shown in Table 1, we perform statistical tests on the collected data. We initially follow the procedure of Colin et al. (2022): an analysis of variance (ANOVA; (Scheffe, 1999)) followed by Tukey’s honestly significant difference (HSD) test for pairwise comparisons (Tukey, 1949). However, we noticed that the assumption checks for normality and homogeneity, which are crucial to the validity of parametric test results (Öztuna et al., 2006), are absent from prior works (Fel et al., 2023). To this end, we performed a complete statistical testing procedure with assumption checks. Figure 3 shows that the data is highly skewed as test accuracy is upper bounded by 1.0, hence violating the normality assumption. Therefore, non-parametric statistics are more appropriate for testing whether there is any statistically significant difference between the five conditions listed.

A Kruskal-Wallis test (McKight & Najab, 2010) with a null hypothesis of “*All condition distributions are identical*” and an alternative hypothesis of “*At least one condition distribution differs*” at a significance level of 0.05 rejects the null hypothesis with $p = 3.4 \times 10^{-5}$, 8.83×10^{-4} for the first two scenarios. We then utilize Dunn’s test (Dinno, 2015) to analyze pairwise differences with Bonferroni correction (Weisstein, 2004) applied. In the first scenario, VISIONLOGIC is shown to be *significantly* better than all other conditions with $p = 3.03 \times 10^{-2}$, 4.00×10^{-4} , 2.41×10^{-2} , $p < 0.001$, respectively, suggesting that our method is effective in helping participants detect biases in the model. In the second scenario, VISIONLOGIC is shown to be *significantly* more effective than ACE and CRAFT with $p = 4.5 \times 10^{-3}$ and 3.09×10^{-2} respectively, supporting its improvement over prior methods on the task of identifying unobvious visual clues. Full test details, including the statistical tests involved, assumption checks, corrections, and test statistics are reported in Appendix F.2.

4.2 EVALUATING LOGICAL RULE-BASED EXPLANATIONS

Setup. We evaluate the global logical rule-based explanations generated by VISIONLOGIC on deep vision models. To demonstrate generalizability across architectures, we cover four representative backbones: ResNet-50 (He et al., 2016), ConvNeXt-Base (Liu et al., 2022), ViT-B (Dosovitskiy et al., 2021), and Swin-T (Liu et al., 2021). VISIONLOGIC learns thresholds and logical rules on the full ImageNet-1k training set (Deng et al., 2009) (1,281,167 images) and is evaluated on the

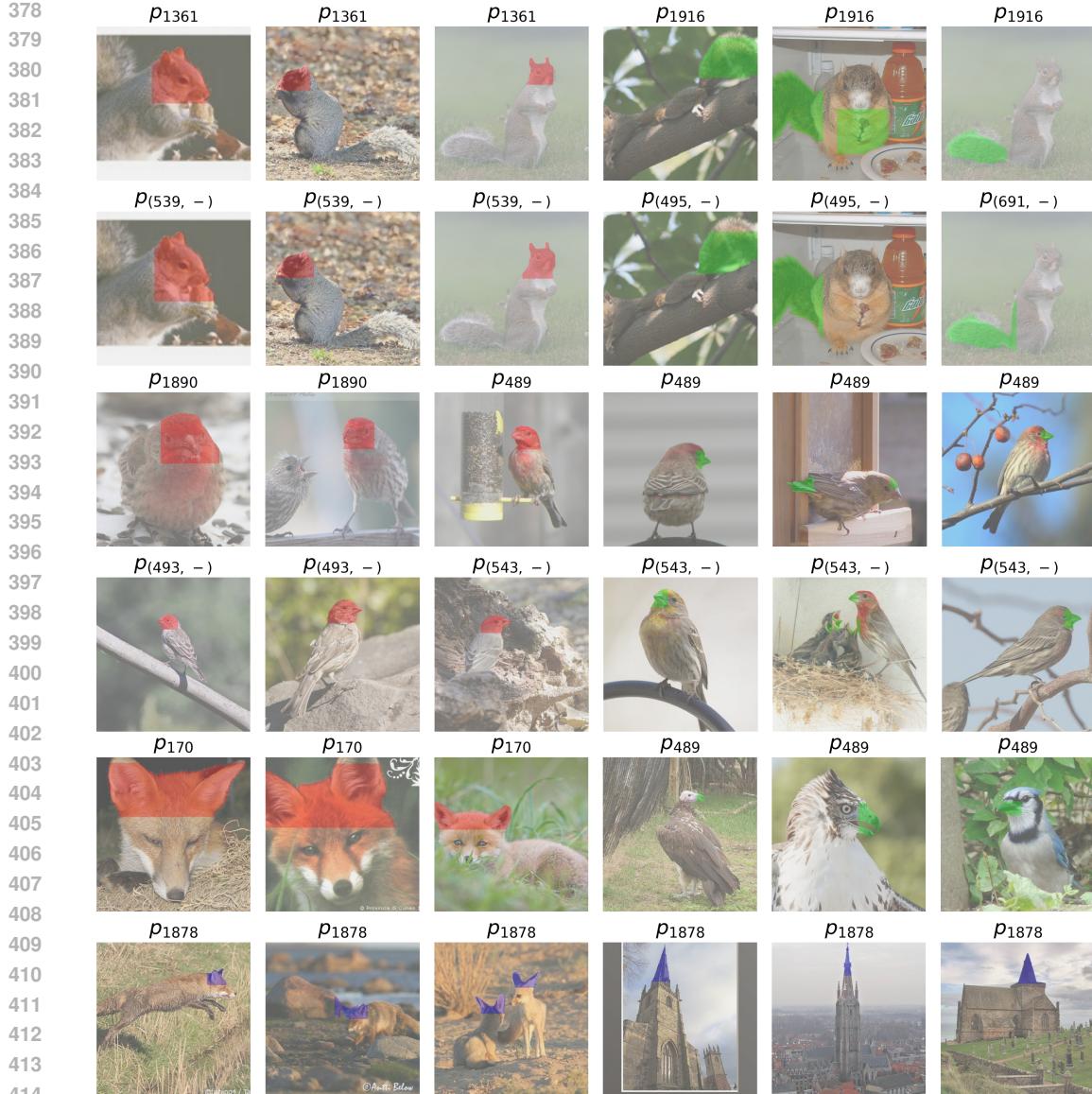


Figure 4: Hidden predicates grounded in visual concepts. In each image, the predicate appears above the frame, and the colored region in the image highlights the concept identified by VISIONLOGIC.

50,000-image validation set. To our knowledge, no prior method extracts *explicit, global* logical rules at this scale on modern backbones. Existing rule-extraction approaches (e.g., Cohen (1995); Zilke et al. (2016); Zarlenga et al. (2021); Hemker et al. (2023)) have only been applied to small datasets or shallow architectures, and have not scaled to state-of-the-art vision models. And the closest related effort (Jiang et al., 2024) relies on *indirect* evidence (e.g., counting sub-explanations) rather than explicitly extracting rules. *Thus, VISIONLOGIC provides, to the best of our knowledge, the first explicit, global interpretable rules for large vision models such as CNNs and ViTs.*

Results. Accordingly, we assess how well VISIONLOGIC preserves the base model’s decisions and discriminative power, as well as the complexity of per-sample explanations, using metrics defined in Appendix D.1. Table 2 reports results across different base models under these metrics. Trained on the ImageNet training set, VISIONLOGIC exhibits strong generalization to the unseen validation set. VISIONLOGIC attains high *coverage* (80–89%) across all backbones while maintaining strong *accuracy* on covered images (76–88%). On covered, labeled images, VISIONLOGIC’s rule-based predictions maintain competitive top-1/top-5 accuracy (ConvNeXt: 80.34%/97.23%; ViT: 80.70%/97.38%), indicating that the symbolic rules retain much of the base models’ discriminative signal.

432 For a model-wise comparison, CNNs tend to produce shorter explanations than Transformers. For
 433 example, ResNet requires, on average, 9.49 predicates per image, whereas Swin requires 53.58, even
 434 though ResNet exposes more valid predicates overall (1,944 vs. 1,460). This aligns with recent
 435 findings (Jiang et al., 2024): CNNs exhibit more “disjunctive” (rule-like) behavior, while ConvNeXt
 436 and Transformers appear more “compositional,” reflected in their longer clauses. Even 50 predicates,
 437 however, remain far more compact and interpretable than the thousands of hidden neurons driving
 438 base model decisions. Nevertheless, the rules may still appear complex to humans, and developing
 439 simplification methods that preserve predictive performance is an important direction for future work.
 440

441 4.3 QUALITATIVE ANALYSIS OF VISUAL CONCEPTS

442 We present human-interpretable visual concepts encoded by predicates discovered in both ResNet and
 443 ViT. Each concept judgment is based on consistent visual inspection across many instances; Figure 4
 444 shows representative cases. The 2nd and 4th rows are from ViT, the others from ResNet. Sampled
 445 concepts include *squirrel head*, *squirrel tail/paws*, *bird head*, *bird beak*, *fox ears*, and *church tops*.
 446

447 **Polysemanticity.** We observe a many-to-many relationship between predicates and concepts (poly-
 448 semanticity). A single predicate can deactivate when either of two distinct concepts is ablated,
 449 implying that one neuron-predicate may encode multiple concepts (within or across classes). Con-
 450 versely, one concept can be encoded by multiple predicates, so masking its region can deactivate
 451 several predicates simultaneously. For example, in the 5th image of row 1, both the tail and paws of
 452 the squirrel are captured by p_{1916} (ResNet). Predicate p_{1878} simultaneously encodes *fox ears* and
 453 *church tops*; while semantically different, both share a triangular geometry, which may explain the
 454 reuse. Predicate p_{489} (rows 3 and 5) frequently encodes *bird beak* across species (*jay*, *vulture*, *kite*,
 455 etc.), and recurs in local explanations for bird images, indicating an influential role in inference.

456 We also find cases where a single concept is captured by multiple predicates. In class *fox* (rows 5–6),
 457 both p_{1878} and p_{170} encode *fox ears*. A predicate may also attend to multiple instances of the same
 458 concept within an image: in row 6, image 3 (two foxes), p_{1878} captures both pairs of ears. Such
 459 behaviors highlight the role of predicates as global concept detectors rather than merely local ones.
 460

461 **Top-ranked predicates encode global structure.** Beyond the local and modular concepts discussed
 462 above, some predicates capture *global* object structure. Figure 10 (Appendix E) illustrates two classes
 463 (*church*, *squirrel*), where p_{908} , $p_{(498, -)}$, p_{219} , and $p_{(312, +)}$ (ResNet/ViT) encode the entire church
 464 or squirrel. Causally, masking any single part does not deactivate these predicates, but masking the
 465 whole object does. These global predicates are typically top-ranked within their classes and tend to be
 466 more class-specific, whereas local predicates are more frequently shared across classes. This pattern
 467 holds for both CNNs and Transformers and suggests a potential avenue for further rule simplification.
 468

469 **CNNs vs. Transformers.** Similar concepts are found in both model families (see rows 1–2, where
 470 both capture the same concept on the same image). A key difference is that Transformers tend to
 471 involve *more* predicates per concept, whereas CNNs yield sparser, more distinct encodings. We
 472 hypothesize three factors: (1) CNN backbones expose a larger overall predicate set than ViTs;
 473 (2) activation functions (ReLU vs. GELU) induce different predicate sparsity via positive/negative
 474 branches; and (3) convolution versus attention imposes different inductive biases on concept formation.
 475 Validating these hypotheses and conducting in-depth analysis of learned predicates and rule structures,
 476 particularly differences between CNN- and ViT-based models, remains a direction for future work.
 477

478 5 CONCLUSION

479 We introduce VISIONLOGIC, a novel neural-symbolic framework that produces faithful, hierarchical
 480 explanations as global logical rules over *causally validated* concepts, directly addressing the core
 481 limitation of prior concept-based methods that heavily rely on correlational statistics. VISIONLOGIC
 482 (i) learns activation thresholds to convert neuron activations into a reusable predicate vocabulary, (ii)
 483 induces compact, class-level logical rules over learned predicates that approximate the base model’s
 484 predictions, and (iii) grounds predicates to visual concepts via ablation-based causal tests with iterative
 485 refinement. Across CNNs and ViTs, VISIONLOGIC largely retains models’ discriminative power with
 486 compact rules. In human studies, it explains model behavior better than state-of-the-art concept-based
 487 methods, yielding clearer and more useful explanations. By unifying neural representations with
 488 symbolic reasoning, VISIONLOGIC offers trustworthy, actionable insight for high-stakes applications.
 489

486 REFERENCES
487

488 David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection:
489 Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer*
490 *Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 3319–3327.
491 IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.354. URL <https://doi.org/10.1109/CVPR.2017.354>.

493 Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-
494 cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018*
495 *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, 2018.

496 William W. Cohen. Fast effective rule induction. In Armand Priedis and Stuart Russell
497 (eds.), *Machine Learning, Proceedings of the Twelfth International Conference on Machine*
498 *Learning, Tahoe City, California, USA, July 9-12, 1995*, pp. 115–123. Morgan Kaufmann,
499 1995. doi: 10.1016/B978-1-55860-377-6.50023-2. URL <https://doi.org/10.1016/B978-1-55860-377-6.50023-2>.

500 Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What I cannot predict, I do not
501 understand: A human-centered evaluation framework for explainability methods. In Sanmi
502 Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances*
503 *in Neural Information Processing Systems 35: Annual Conference on Neural Information*
504 *Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December*
505 *9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/13113e938f2957891c0c5e8df811dd01-Abstract-Conference.html.

506 Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale
507 hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*,
508 pp. 248–255. Ieee, 2009.

509 Alexis Dinno. Nonparametric pairwise multiple comparisons in independent groups using dunn's
510 test. *The Stata Journal*, 15(1):292–300, 2015.

511 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas
512 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit,
513 and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale.
514 *arXiv preprint arXiv:2010.11929*, 2021.

515 Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec,
516 Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger B. Grosse, Sam McCandlish,
517 Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models
518 of superposition. *CoRR*, abs/2209.10652, 2022. doi: 10.48550/ARXIV.2209.10652. URL
519 <https://doi.org/10.48550/arXiv.2209.10652>.

520 Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, and Alek-
521 sander Madry. Robustness library. In *GitHub repository*, 2019. URL <https://github.com/MadryLab/robustness>.

522 Thomas Fel, Agustin Martin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien
523 Colin, Rémi Cadène, and Thomas Serre. CRAFT: concept recursive activation factorization for
524 explainability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023,*
525 *Vancouver, BC, Canada, June 17-24, 2023*, pp. 2711–2721. IEEE, 2023. doi: 10.1109/CVPR52729.
526 2023.00266. URL <https://doi.org/10.1109/CVPR52729.2023.00266>.

527 Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by
528 filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and*
529 *pattern recognition*, pp. 8730–8738, 2018.

530 Chuqin Geng, Xiaojie Xu, Haolin Ye, and Xujie Si. Scalar invariant networks with zero bias. *CoRR*,
531 abs/2211.08486, 2022. doi: 10.48550/ARXIV.2211.08486. URL <https://doi.org/10.48550/arXiv.2211.08486>.

540 Chuqin Geng, Zhaoyue Wang, Haolin Ye, Saifei Liao, and Xujie Si. Learning minimal NAP
 541 specifications for neural network verification. *CoRR*, abs/2404.04662, 2024. doi: 10.48550/
 542 ARXIV.2404.04662. URL <https://doi.org/10.48550/arXiv.2404.04662>.

543

544 Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-
 545 based explanations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Flo-
 546 rence d’Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural In-
 547 formation Processing Systems 32: Annual Conference on Neural Information Process-
 548 ing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp.
 549 9273–9282, 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html>.

550

551 Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1.
 552 MIT Press, 2016.

553

554 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image
 555 recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,
 556 pp. 770–778, 2016.

557

558 Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the
 559 IEEE international conference on computer vision*, pp. 2961–2969, 2017.

560

561 Konstantin Hemker, Zohreh Shams, and Mateja Jamnik. Cgexplain: Rule-based deep neural network
 562 explanations using dual linear programs. In Hao Chen and Luyang Luo (eds.), *Trustworthy
 563 Machine Learning for Healthcare - First International Workshop, TML4H 2023, Virtual Event,
 564 May 4, 2023, Proceedings*, volume 13932 of *Lecture Notes in Computer Science*, pp. 60–72.
 Springer, 2023. doi: 10.1007/978-3-031-39539-0_6. URL https://doi.org/10.1007/978-3-031-39539-0_6.

562

563

564 Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint
 565 arXiv:1606.08415*, 2016.

566

567 Mingqi Jiang, Saeed Khorram, and Li Fuxin. Comparing the decision-making mechanisms by
 568 transformers and cnns via explanation methods. In *Proceedings of the IEEE/CVF Conference on
 569 Computer Vision and Pattern Recognition*, pp. 9546–9555, 2024.

570

571 Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam:
 572 Exploring hierarchical class activation maps for localization. *IEEE Trans. Image Process.*, 30:
 573 5875–5888, 2021. doi: 10.1109/TIP.2021.3089943. URL <https://doi.org/10.1109/TIP.2021.3089943>.

572

573

574

575 Zhenchao Jin, Bin Liu, Qi Chu, and Nenghai Yu. Isnet: Integrate image-level and semantic-level
 576 context for semantic segmentation. In *2021 IEEE/CVF International Conference on Computer
 577 Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 7169–7178. IEEE, 2021.
 578 doi: 10.1109/ICCV48922.2021.00710. URL <https://doi.org/10.1109/ICCV48922.2021.00710>.

579

580 Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas,
 581 and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept
 582 activation vectors (TCAV). In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the
 583 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2673–
 584 2682. PMLR, 2018. URL <http://proceedings.mlr.press/v80/kim18d.html>.

585

586 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete
 587 Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings
 588 of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

589

590 Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to
 591 document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

592

593 Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo.
 Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the
 594 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

594 Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining
 595 Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern*
 596 *Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 11966–11976. IEEE, 2022.
 597 doi: 10.1109/CVPR52688.2022.01167. URL <https://doi.org/10.1109/CVPR52688.2022.01167>.
 598 2022.01167.

599 David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou.
 600 Discovering causal signals in images. In *2017 IEEE Conference on Computer Vision and Pattern*
 601 *Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 58–66. IEEE Computer Society,
 602 2017. doi: 10.1109/CVPR.2017.14. URL <https://doi.org/10.1109/CVPR.2017.14>.
 603

604 Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu.
 605 Towards deep learning models resistant to adversarial attacks. In *6th International Conference on*
 606 *Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference*
 607 *Track Proceedings*. OpenReview.net, 2018. URL <https://openreview.net/forum?id=rJzIBfZAb>.
 608

609 Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting
 610 them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.
 611 5188–5196, 2015.

612 Patrick E McKnight and Julius Najab. Kruskal-wallis test. *The corsini encyclopedia of psychology*, pp.
 613 1–1, 2010.

614 Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the
 615 preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural*
 616 *information processing systems*, 29, 2016.

617 Derya Öztuna, Atilla Halil Elhan, and Ersöz Tüccar. Investigation of four different normality tests
 618 in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical*
 619 *Sciences*, 36(3):171–176, 2006.

620 Sebastian Prillo and Julian Martin Eisenschlos. Softsort: A continuous relaxation for the argsort
 621 operator. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020,*
 622 *13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp.
 623 7793–7802. PMLR, 2020. URL <http://proceedings.mlr.press/v119/prillo20a.html>.
 624

625 Prolific. Prolific. <https://www.prolific.com>, 2024. Participants were recruited using
 626 Prolific (www.prolific.com). Accessed: April 24, 2025.

627 Henry Scheffe. *The analysis of variance*. John Wiley & Sons, 1999.

628 Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh,
 629 and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localiza-
 630 tion. *International Journal of Computer Vision*, 128(2):336–359, 2017.

631 Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh,
 632 and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based local-
 633 ization. *Int. J. Comput. Vis.*, 128(2):336–359, 2020. doi: 10.1007/S11263-019-01228-7. URL
 634 <https://doi.org/10.1007/s11263-019-01228-7>.

635 Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks:
 636 Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun
 637 (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada,*
 638 *April 14-16, 2014, Workshop Track Proceedings*, 2014. URL <http://arxiv.org/abs/1312.6034>.

639 John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pp. 99–114,
 640 1949.

641 Haofan Wang, Mengnan Du, Fan Yang, and Zifan Zhang. Score-cam: Score-weighted visual
 642 explanations for convolutional neural networks. *Proceedings of the IEEE/CVF Conference on*
 643 *Computer Vision and Pattern Recognition Workshops*, pp. 24–25, 2020.

648 Eric W Weisstein. Bonferroni correction. <https://mathworld.wolfram.com/>, 2004.
649

650 Shirley Wu, Mert Yüksekgönül, Linjun Zhang, and James Zou. Discover and cure: Concept-
651 aware mitigation of spurious correlation. In Andreas Krause, Emma Brunskill, Kyunghyun
652 Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference
653 on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of
654 *Proceedings of Machine Learning Research*, pp. 37765–37786. PMLR, 2023. URL <https://proceedings.mlr.press/v202/wu23w.html>.
655

656 Mateo Espinosa Zarlenga, Zohreh Shams, and Mateja Jamnik. Efficient decompositional rule
657 extraction for deep neural networks. *CoRR*, abs/2111.12628, 2021. URL <https://arxiv.org/abs/2111.12628>.
658

659 Kexuan Zhang, Qiyu Sun, Chaoqiang Zhao, and Yang Tang. Causal reasoning in typical computer
660 vision tasks. *CoRR*, abs/2307.13992, 2023. doi: 10.48550/ARXIV.2307.13992. URL <https://doi.org/10.48550/arXiv.2307.13992>.
661

662 Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein.
663 Invertible concept-based explanations for CNN models with non-negative concept activation
664 vectors. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third
665 Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Sym-
666 posium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February
667 2-9, 2021*, pp. 11682–11690. AAAI Press, 2021. doi: 10.1609/AAAI.V35I13.17389. URL
668 <https://doi.org/10.1609/aaai.v35i13.17389>.
669

670 Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep
671 features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern
672 Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2921–2929. IEEE Computer
673 Society, 2016. doi: 10.1109/CVPR.2016.319. URL <https://doi.org/10.1109/CVPR.2016.319>.
674

675 Jan Ruben Zilke, Eneldo Loza Mencía, and Frederik Janssen. Deepred–rule extraction from deep
676 neural networks. In *Discovery Science: 19th International Conference, DS 2016, Bari, Italy,
677 October 19–21, 2016, Proceedings 19*, pp. 457–473. Springer, 2016.
678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

 702 **A THE BOUNDING BOX LOCALIZATION ALGORITHM**
 703

 704 Algorithm 1 presents the detailed procedure for using bounding box localization to identify regions
 705 that significantly contribute to the computation of predicate $p_j(x)$. While the MASKREGION function
 706 in our implementation can replace the highlighted region with random noise³, it also supports
 707 alternative masking strategies such as blurring, mean-fill, white-fill, and black-fill. We use blurring
 708 by default, as it performs most consistently in our experiments and aligns with our goal of non-
 709 destructive ablation. The GETFEATUREMAPREGION function thresholds the feature map at 15% of
 710 its maximum activation value, which results in connected segments of surviving pixels. It then draws
 711 a bounding box around the single largest segment (Zhou et al., 2016; Selvaraju et al., 2020). This
 712 yields a coarse initial estimate, which serves as a starting point for our iterative refinement algorithm,
 713 allowing it to converge toward more accurate and compact solutions.

Algorithm 1: BOUNDING BOX LOCALIZATION ALGORITHM

 714 **Input:** Input image x , predicate p_j , model M , shrink factor λ , max attempts max_attempt
 715 **Output:** Critical region R for predicate p_j
 716
 717 1 **Function** InitialGuess(x, p_j)
 718 2 **if** $isCNN(M)$ **then**
 719 3 $R \leftarrow GetFeatureMapRegion(p_j)$ /* Initialize using feature map if model
 720 4 is CNN */
 721 5 **else**
 722 6 $R \leftarrow LargeCentralBox(x)$ /* Use a large central box as default */
 723 7 **return** R
 724 7 **Function** RefineRegion(R, λ)
 725 8 // Extract current region dimensions
 726 8 $x, y, h, w \leftarrow R$
 727 8 // Generate a random sub-box with approximately λ times the area of
 728 9 R
 729 9 $R_{new} \leftarrow GenerateNewBox(R, \lambda)$
 10 **return** R_{new}
 730 11 **Function** LocateCriticalSection($x, p_j, M, \lambda, max_attempt$)
 731 12 $R \leftarrow InitialGuess(x, p_j)$ /* Start with an initial region guess */
 732 13 **while** True **do**
 733 14 $x' \leftarrow MaskRegion(x, R)$ /* Replace region R in x with noise */
 734 15 $p_{masked} \leftarrow M(x')$
 735 16 $p_{crop} \leftarrow M(CropRegion(x, R))$ /* Run model on cropped region alone */
 736 17 **if** $p_{masked} < \tau$ **and** $p_{crop} \geq \tau$ **then**
 737 18 refined \leftarrow False
 738 19 **for** $i \leftarrow 1$ **to** max_attempt **do**
 739 20 $R_{new} \leftarrow RefineRegion(R, \lambda)$ /* Try a smaller random sub-region */
 740 21 $x'' \leftarrow MaskRegion(x, R_{new})$
 741 22 $p_{masked_new} \leftarrow M(x'')$
 742 23 $p_{crop_new} \leftarrow M(CropRegion(x, R_{new}))$
 743 24 **if** $p_{masked_new} < \tau$ **and** $p_{crop_new} \geq \tau$ **then**
 744 25 $R \leftarrow R_{new}$
 745 26 refined \leftarrow True
 746 27 **break** /* Accept this refined region and continue */
 747 28 **if** not refined **then**
 748 29 **break** /* Stop if no better sub-region found */
 749 30 **else**
 750 31 **break** /* Initial region fails to meet constraints */
 751 32 **return** R /* Return final critical region */

³Random-noise masks may introduce out-of-distribution artifacts and yield unstable model outputs.

756 **B LLM USAGE**
757758 Large Language Models (LLMs) were used as a general-purpose assistive tool in the preparation
759 of this work. Specifically, LLMs supported tasks such as refining the clarity of writing, suggesting
760 alternative phrasings, and checking the consistency of technical terminology. They were **not** used
761 for generating research ideas, conducting experiments, or producing original scientific contributions.
762 All substantive research decisions, analysis, and results presented in this paper are the responsibility
763 of the authors. The authors have carefully reviewed and verified all LLM-assisted text to ensure
764 accuracy and originality.
765766 **C EXPERIMENTAL SETUP**
767768 All experiments are conducted on Ubuntu 22.04 LTS with an AMD EPYC™ 7532 (32 cores), 128 GB
769 RAM, and a single NVIDIA A100 (40 GB). We use pretrained ImageNet-1k models to extract
770 final-layer activations for all 1,281,167 training images. These activations are used to learn valid
771 predicates and thresholds. We then compute class profiles and evaluate the induced logical rules on
772 the ImageNet validation set (50,000 images) using the metrics in Appendix D.1. Hyperparameter
773 choices and implementation details for the bounding-box algorithm (Algorithm 1) and predicate
774 learning are discussed in the following subsections.
775776 **C.1 BOUNDING BOXES**
777778 We localize predicate-supporting regions via iterative noise ablation using Algorithm 1.
779780 **Initialization.** For CNNs, we first form an activation heatmap from the last-layer feature map. We
781 binarize it at 15% of its maximum intensity, which results in connected pixel segments, and then draw
782 the tightest axis-aligned bounding box around the single largest segment (Zhou et al., 2016; Selvaraju
783 et al., 2020). If no pixels survive, we fall back to a centered box covering 90% of the image area. For
784 ViTs, we initialize with a centered box covering 90% of the image, aligned to the patch grid.
785786 **Refinement.** At each iteration, we propose up to 10 random shrinks of the current box using a shrink
787 factor $\lambda = 0.9$ (uniformly sampling aspect ratio and position within the shrunken envelope). A
788 proposal is *accepted* if ablating its region (replace with noise) flips the target predicate from active to
789 inactive, $p_j(x) = 1 \rightarrow p_j(x') = 0$; otherwise it is rejected. We repeat until no accepted proposal exists.
790 We run 5 independent trials per predicate (different random seeds) and keep the smallest accepted
791 box across trials.
792793 **Sufficiency check.** To verify that the retained region is sufficient to trigger the predicate, we paste
794 the final box back into a noise canvas and confirm re-activation, $p_j(\hat{x}) = 1$.
795796 **Further refinement.** We further refine the box using off-the-shelf segmentation models—SAM
797 (Kirillov et al., 2023), Mask R-CNN (He et al., 2017), or ISNet (Jin et al., 2021). We intersect
798 the predicted mask with the current box and then re-validate causality using the same ablation
799 and sufficiency checks. Empirically, SAM and Mask R-CNN often produce fine-grained, part-
800 level segments (e.g., an animal’s ear or leg). While this does not affect the correctness of our
801 pipeline, it can occasionally be cumbersome because it yields multiple disjoint regions to handle. By
802 contrast, ISNet focuses on foreground–background separation and is therefore better aligned with our
803 goal of isolating the entire foreground object (e.g., the whole car or animal) before computing the
804 mask–box intersection. We hence mostly use ISNet with the default setting and hyperparameters in
805 our experiments.
806807 **C.2 THRESHOLD LEARNING PROBLEM**
808809 **Selecting influential examples.** For each channel j , we score contributions $u_j^c(x) = \mathbf{W}_j^c \mathbf{z}_j(x)$ per
810 class c and per example x . Influential examples for j are those where j is among the *SoftSort* top- k
811 contributors (Eq. 4) with $k = 3$. Concretely, we apply *SoftSort* to the vector $\{u_\ell^c(x)\}_{\ell=1}^d$ to obtain a
812 differentiable top- k mask $w_j^c(x) \in [0, 1]$, and declare x influential for j if j belongs to the top- k set
813 under this mask (i.e., $w_j^c(x)$ is one of the k largest values). *Intuitively, this selects examples where*
814 *channel j provides strong, class-relevant evidence, yielding stable, noise-resistant threshold seeds.*
815

810 **Seed threshold and sharpness.** The seed threshold and sharpness are
 811

$$812 \quad T_j^{(0)} = \text{Quantile}_{0.8} \left\{ z_j(x) \mid x \text{ is influential for } j \right\}, \quad s_j^{(0)} = 1.$$

813
 814 We initialize $T_j^{(0)}$ at a high percentile (0.8) of $z_j(x)$ over influential examples to anchor “feature
 815 present” in the activation tail, gain robustness to outliers, and start from a conservative, compact
 816 predicate set. We set $s_j^{(0)} = 1$ (logistic slope) and constrain $s_j \in [0.5, 5]$ during training for numerical
 817 stability.

818 **Objective and optimization.** We optimize Eq. 5 with Adam; learning rates are 10^{-3} for $\{T, s\}$ and
 819 5×10^{-4} for $\{\mathbf{W}_{\text{rule}}, \mathbf{b}_{\text{rule}}\}$, using batch size 512 and early stopping on validation KL (patience 5).
 820 Regularization uses

$$821 \quad \lambda_T = 1.0, \quad \lambda_s = 0.1 \text{ (keeps } s_j \approx 1\text{)}, \quad \lambda_{\text{use}} = 5 \times 10^{-3},$$

822 with a group lasso over $\{p_{j,\leq 1}, p_{j,\leq 2}, p_{j,\leq 3}\}$ per channel to select a single rank window. We clip
 823 T_j to the empirical range of z_j (per channel, class-agnostic) and add a small $\epsilon = 10^{-6}$ inside the
 824 sigmoid to avoid saturation in mixed precision.

825 **Schedule and convergence.** We train for up to 30 epochs; \mathbf{W}_{rule} and \mathbf{b}_{rule} are warm-started from
 826 classwise normalized predicate frequencies (mean-centered by each predicate’s global frequency),
 827 then jointly refined with $\{T, s\}$. Convergence is declared when validation KL improves by $< 10^{-4}$
 828 or when the patience budget is exhausted.

829 **Hardening and test-time prediction.** After training, we harden $\tilde{p}_j(x) = \sigma(s_j(z_j(x) - T_j))$ to
 830 $p_j(x) = \mathcal{I}(z_j(x) \geq T_j)$ and discard f_{rule} . Test-time prediction uses the symbolic rank-based score:

$$831 \quad \hat{c}(x) = \arg \min_c S(x, c),$$

832 i.e., the class whose characteristic predicates best explain those active on x (Section 3.2).

833 **Notes and ablations.** Empirically, the learned thresholds T_j often align with the $k=1$ specialization
 834 of the rank-aware predicate (cf. Observation), even though we train with candidate windows $k \in$
 835 $\{1, 2, 3\}$ and let structured sparsity select one per channel. In practice, $k=1$ is chosen for the majority
 836 of channels, with $k=2$ or $k=3$ retained for a minority of polysemantic channels that contribute reliably
 837 without being the single top contributor. Performance and rule sparsity are stable for $k \in \{2, 3, 4\}$;
 838 using $k=3$ as the candidate ceiling slightly improves recall of informative channels while keeping
 839 the predicate vocabulary compact. Using per-class thresholds harms transfer and readability; the
 840 class-agnostic T_j yields stable, reusable predicates and defers disambiguation to the class profiles
 841 $\{\mathcal{R}^c\}$ via $S(x, c)$.

842 **Claim 1** (Monotone equivalence of rule head and rank score). *Initialize \mathbf{W}_{rule} from classwise
 843 normalized predicate frequencies (mean-centered by each predicate’s global frequency), which
 844 induces the same ordering over predicates as the class profile ranks $\mathcal{R}^c(\cdot)$. If, during training, the
 845 (training-only) head weights are set to a strictly decreasing affine function of rank, e.g.,*

$$846 \quad \mathbf{W}_{\text{rule}}^{c,i} = \alpha_c - \beta \mathcal{R}^c(p_i), \quad \beta > 0, \quad (8)$$

847 and a class-independent bias \mathbf{b}_{rule} is used (or scores are compared after subtracting class-specific
 848 constants), then

$$849 \quad \arg \max_c f_{\text{rule}}^c(x) = \arg \min_c S(x, c).$$

850 *Proof.* Let $A(x) = \{i : p_i(x) = 1\}$ be the active predicates and $m = |A(x)|$. Using Eq. 8 with a
 851 class-independent bias,
 852

$$853 \quad f_{\text{rule}}^c(x) = \sum_{i \in A(x)} (\alpha_c - \beta \mathcal{R}^c(p_i)) + \mathbf{b}_{\text{rule}} = m \alpha_c - \beta \sum_{i \in A(x)} \mathcal{R}^c(p_i) + \mathbf{b}_{\text{rule}}.$$

854 Since m , α_c , and \mathbf{b}_{rule} do not affect the ordering induced by the sum of ranks, maximizing $f_{\text{rule}}^c(x)$
 855 over c is equivalent to minimizing $\sum_{i \in A(x)} \mathcal{R}^c(p_i)$. By definition,

$$856 \quad S(x, c) = \frac{1}{m} \sum_{i \in A(x)} \mathcal{R}^c(p_i),$$

857 so $\arg \max_c f_{\text{rule}}^c(x) = \arg \min_c S(x, c)$. □

864
865

C.3 ADDITIONAL RESULTS ON THRESHOLD LEARNING

866

867 Figure 5–8 show the empirical distributions of learned per-channel thresholds $\{T_j\}$ across four
868 architectures.

869

870 **(i) Transformers and ConvNeXt are bimodal and near-symmetric.** For ViT and Swin (Figure 7, 8),
871 and for ConvNeXt (Figure 6), thresholds concentrate in two tight modes near ± 1 . This is consistent
872 with (a) sign-aware predicates ($p_{j,+}$ and $p_{j,-}$) and (b) LayerNorm/GELU producing roughly zero-
873 mean, unit-scale channel responses, so “feature present” naturally anchors away from 0 on both
874 branches.

875

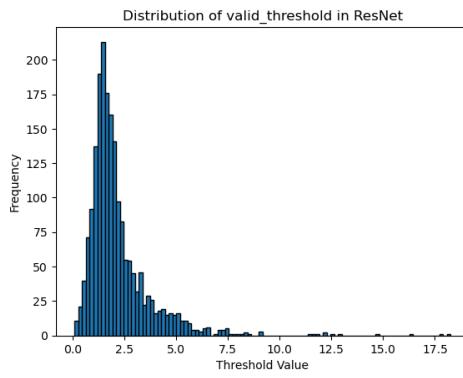
876 **(ii) ResNet is strictly positive and right-skewed.** For ResNet (Figure 5), thresholds are nonnegative
877 and exhibit a heavy right tail. This aligns with ReLU activations (no negative branch) and localized
878 high-energy features that occasionally require larger cutoffs; large T_j outliers are infrequent but
879 present.

880

881 **(iii) Conservativeness and the $k=1$ heuristic.** Across all models, thresholds sit well away from
882 0, indicating a conservative notion of “feature present.” Qualitatively, many T_j align with our $k=1$
883 heuristic (Section 3.1): for a given channel j , T_j tends to be close to the minimum activation among
884 correctly classified examples where j is top-1 by contribution for its most representative class. This
885 matches the view that training recovers a data-driven, architecture-stable cutoff.

886

887



888

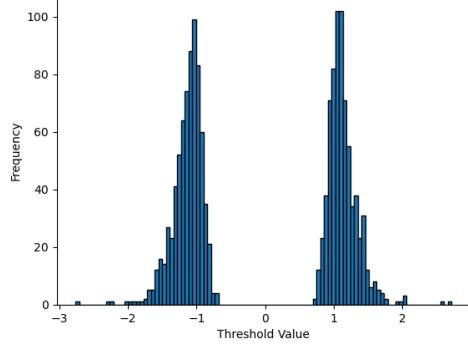
889

890 Figure 5: Distribution of learned thresholds
891 for valid predicates in ResNet.

892

893

894 Distribution of valid_threshold in ViT



895

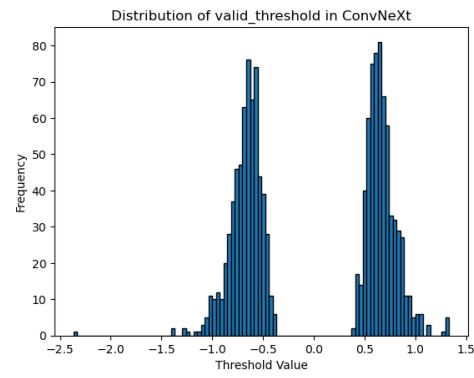
896

897 Figure 7: Distribution of learned thresholds
898 for valid predicates in ViT.

899

900

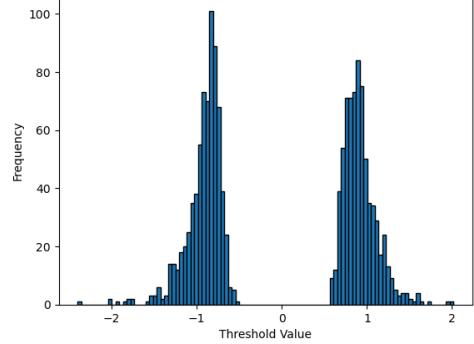
901 Distribution of valid_threshold in ConvNeXt

902 Figure 6: Distribution of learned thresholds
903 for valid predicates in ConvNeXt.

904

905

906 Distribution of valid_threshold in Swin

907 Figure 8: Distribution of learned thresholds
908 for valid predicates in Swin.

909

910

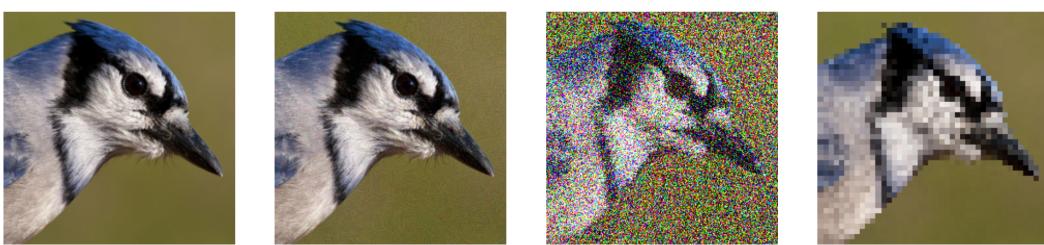
911

912

913

918 D ADDITIONAL DETAILS ON ASSESSING RULE-BASED EXPLANATIONS
919920 D.1 MERTICS
921922 For completeness, we report the metrics used to evaluate global logical rule-based explanations:
923

- 924 • **Number of valid predicates:** Count of predicates that take both `True` and `False` values
925 on the evaluation set under the learned thresholds.
- 926 • **Explanation complexity:** Average number of predicates in the selected explanation per
927 image (equivalently, the average number of literals in the satisfied DNF clause).
- 928 • **Coverage:** Fraction of images for which VISIONLOGIC returns a valid explanation (e.g., at
929 least one rule is satisfied or a score-based explanation is produced).
- 930 • **Fidelity (covered):** Among covered images, the percentage for which VISIONLOGIC’s
931 predicted class matches the base model’s predicted class.
- 932 • **Top-1 accuracy (covered):** Among covered images with ground-truth labels, the percentage
933 for which VISIONLOGIC’s top-1 class equals the ground-truth class.
- 934 • **Top-5 accuracy (covered):** Among covered images with ground-truth labels, the percentage
935 for which VISIONLOGIC’s top-5 class equals the ground-truth class.

937 D.2 PROBING ROBUSTNESS OF VISION MODELS
938939 Although VISIONLOGIC is trained exclusively on positive examples, it still correctly identifies a
940 non-trivial fraction of images misclassified by the neural networks. This is evident from the gap
941 between *Fidelity (covered)* and *Top-1 accuracy (covered)*: VISIONLOGIC can match erroneous model
942 predictions with rules, offering insights into misclassification causes and helping probe robustness
943 under perturbations and adversarial settings.944 We investigate the success of adversarial attacks through the lens of local explanations generated
945 by VISIONLOGIC. From a logical rule perspective, misclassification typically occurs when (a) the
946 top-ranked predicates of the ground truth class are deactivated, and (b) predicates associated with
947 other classes become active. While these often co-occur, we define (a) as the root cause—denoted
948 as a *Type A* cause—if it alone can alter the prediction without (b). Similarly, we define a *Type B*
949 cause when (b) alone is sufficient to induce misclassification. These two causes are attack-agnostic,
950 enabling us to understand the underlying logic behind different attacks.951 We provide concrete examples in Figure 9. The original image follows the rule $p_{669} \wedge p_{844} \wedge p_{489} \Rightarrow$
952 “*jay*”. After the PGD (Madry et al., 2018) attack, the rule becomes $p_{1220} \wedge p_{489} \wedge p_{537} \wedge p_{844} \Rightarrow$ “*tray*”,
953 which is a *Type B* cause, as introducing the new predicate p_{1220} significantly increases the explanation
954 score for the class “*tray*”. The Gaussian Noise attack yields $p_{2032} \wedge p_{1074} \wedge p_{2028} \dots \wedge p_{669} \wedge p_{844} \wedge$
955 $p_{489} \Rightarrow$ “*badger*”, introducing many new predicates, which is a *Type B* cause. The Pixelation attack
956 results in $p_{2028} \wedge p_{376} \wedge p_{1940} \dots \wedge p_{211} \Rightarrow$ “*black grouse*”, deactivating all original predicates, thus
957 exhibiting a *Type A* cause.
958959
960 Figure 9: This displays the original image, followed by the images after PGD, Gaussian noise, and
961 Pixelate attacks from left to right.
962
963964 To conduct a systematic evaluation, we randomly sample 1,000 images from the ImageNet validation
965 set, applying a successful attack to each image once using Projected Gradient Descent (PGD) (Madry
966

972 et al., 2018), Gaussian noise (Goodfellow et al., 2016), and Pixelate attacks (Engstrom et al., 2019),
 973 respectively.
 974

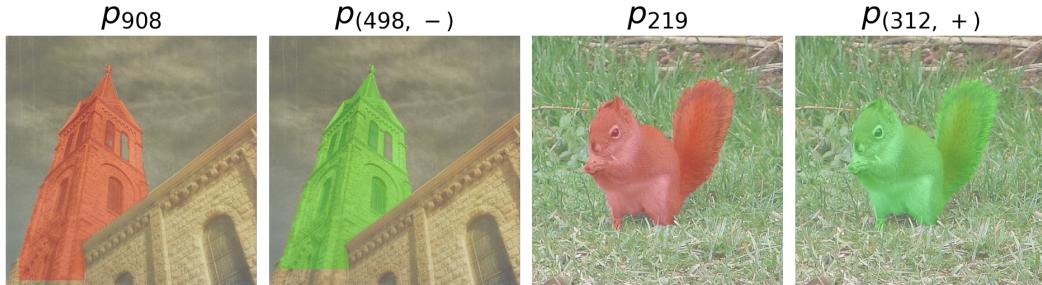
975
976 **Table 3: Statistics of root causes under various adversarial attacks.**

977 978 979 Attack	980 ResNet		981 ConvNet		982 ViT		983 Swin	
	984 Type A	985 Type B	986 Type A	987 Type B	988 Type A	989 Type B	990 Type A	991 Type B
Gaussian	79	921	116	884	149	851	109	891
Pixelate	214	786	329	671	267	733	298	702
PGD	377	623	561	439	653	347	612	388

984 Table 3 presents the statistics of root causes across different adversarial attacks. The Gaussian noise
 985 attack tends to affect all active predicates simultaneously, often activating new predicates, which is
 986 primarily explained by Type B causes. The Pixelate attack operates similarly but has a higher chance
 987 of deactivating existing predicates, leading to more Type A causes, as it dilutes the fine-grained
 988 details of regions attended by the predicates. Finally, the PGD attack is the most efficient and selects
 989 either Type A or Type B causes to achieve the quickest result. For ResNet, which generates shorter
 990 explanations, it favors introducing new predicates. For transformers, which use more predicates
 991 in their explanations, it leans more toward deactivations. We aim to investigate more attacks and
 992 defenses in future work.
 993

994

E ADDITIONAL OBSERVATIONS ON PREDICATES

1006 Figure 10: Top-ranked predicates often capture global structure. Red: ResNet; green: ViT.
 1007

1008 **Top-ranked predicates encode global structure.** Beyond the local and modular concepts discussed
 1009 above, some predicates capture *global* object structure. Figure 10 illustrates two classes (*church*,
 1010 *squirrel*), where p_{908} , $p_{(498, -)}$, p_{219} , and $p_{(312, +)}$ (ResNet/ViT) encode the entire church or squirrel.
 1011 Causally, masking any single part does not deactivate these predicates, but masking the whole object
 1012 does. These global predicates are typically top-ranked within their classes and tend to be more
 1013 class-specific, whereas local predicates are more frequently shared across classes. This pattern holds
 1014 for both CNNs and Transformers and suggests a potential avenue for further rule simplification.
 1015

1016 **Predicates robustly identify visual concepts across variations in appearance.** The human
 1017 visual system can recognize objects belonging to the same concept despite significant differences in
 1018 appearance, such as color and shape. Interestingly, our learned predicates exhibit a similar capability.
 1019 For example, in the third row of Figure 4, the concept “*bird’s head*” appears from both the front
 1020 and the side, yet p_{1890} consistently attends to it. Similarly, p_{1878} captures varying forms, sizes, and
 1021 angles of “*church tops*”.

1022 **Sensitivity to location changes.** Some predicates appear to be location-invariant, while others are
 1023 not, as illustrated in Figure 11. Predicate p_{170} in the top row remains active even when the “*fox’s ears*”
 1024 are shifted across the third, fourth, and fifth images. In contrast, predicate p_{1916} in the bottom row,
 1025 which encodes “*squirrel’s tail*”, is deactivated once the tail moves away from its original position.



Figure 11: Sensitivity to spatial changes. The predicate encoding “fox’s ears” appears to be location-invariant, whereas the predicate encoding “squirrel’s tail” is sensitive to positional shifts.

F HUMAN EVALUATION

F.1 EXPERIMENT SETUP

We describe the evaluation setup in detail as follows.

Participants. We recruited participants from Prolific (Prolific, 2024), a popular online platform for human evaluation of research projects. The users recruited have high qualifications as we requested a task acceptance rate greater than 98%. Our questionnaire is approved by the Social Sciences, Humanities & Education REB at University X, and all users provided informed consent before they started the experiment. The questionnaire is designed to be completed in 8-10 minutes and users who passed the screening received USD\$ 2.00 upon completion.

Statistics. For the Husky vs. Wolf scenario, $n = 161$ participants passed screening and filtering, respectively $n = 31, 25, 40, 35, 30$ for control, baseline, ACE, CRAFT and VISIONLOGIC.

For the Otter vs. Beaver scenario, $n = 114$ participants passed screening and filtering, respectively $n = 22, 22, 27, 25, 18$ for control, baseline, ACE, CRAFT, and VISIONLOGIC.

For the Kit Fox vs. Red Fox scenario, $n = 190$ participants passed screening and filtering, respectively $n = 35, 35, 45, 40, 35$ for control, baseline, ACE, CRAFT, and VISIONLOGIC.

Study design. We followed the experimental design proposed by Colin et al. (2022), which quantitatively assesses to what degree a concept-based explanation method can help a human observer understand the behavior of an AI model. Each participant is only tested on a single condition (control, baseline, ACE, CRAFT, or VISIONLOGIC) to avoid possible experimental confounds. *For fairness, we compare only the causally grounded concepts produced by VISIONLOGIC, excluding any rule-structure information, against the four other conditions.*

The screening, training, and testing phases are exactly as described in Colin et al. (2022). After granting consent (shown in Figure 12), participants are presented with detailed instructions and the study goal: learning to predict which class *an AI model will predict* for a given image. They start with a practice session that shows simple images along with their explanations and model prediction, followed by some unseen images without explanations, where they are expected to answer the model’s prediction correctly. Participants who failed this session were not allowed to proceed with the study. They are then tested on their understanding of the study goal by a short quiz (exactly formulated as in Colin et al. (2022)). Again, participants who failed this session were not allowed to proceed.

After the above two screening phases, participants went through 3 sessions of training phases (5, 10, and 15 images, respectively, with explanations and model decisions) followed by a testing phase (7 new images without explanation). The answers for the testing phases are collected to compute the main result. Figure 13 shows an example of the training phase. As described in Colin et al. (2022), we implemented a reservoir (Figure 14) for training images as a reference point for participants during the testing phase. The last test image of each session is an image from the reservoir, and participants who incorrectly answered the last test question were filtered out.

1080 Learn to predict AI model decisions by studying explanations

1081 **AI Explanation Study Instructions**

1082 **Your Goal:** Learn to predict what an AI model will output by studying its explanations.

1083 **Important Guidelines:**

1084 • Focus on the AI's behavior - not what you think is "correct"

1085 • Study the explanations carefully - they show what the AI is looking at

1086 • During training: You'll see images + explanations + AI predictions

1087 • During testing: You'll see only images and must predict what the AI will say

1088 • Use the reference panel - you can look back at training examples during tests

1089 • If you do not see an explanation image - you are in the control group, please try your best to predict AI's decision

1090 **Payment:** You must pass all screening questions and attention checks to receive payment.

1091 **Participant Consent Form**

1092 **Explainable AI: human interpretation of machine learning systems**

1093 **Invitation to Participate**

1094 You are invited to participate in a short online research study conducted by researchers at the [REDACTED]. The purpose of this study is to better understand how people interpret the explanations provided by machine learning models.

1095 Your participation is entirely voluntary.

1096 **What Will You Be Asked to Do?**

1097 If you agree to participate:

1098 • You will complete a short series of visual tasks.

1099 • The study will take approximately 5–7 minutes.

1100 **Compensation**

1101 You will be compensated [REDACTED], based on the estimated time commitment and in accordance with fair-pay guidelines (approx. \$15/hour equivalent). Partial payment may be issued if you begin but do not complete the full task.

1102 **Risks and Benefits**

1103 There are no known risks beyond those of normal online activity. While there are no direct personal benefits, your participation will help improve understanding of how people interact with machine learning explanations, which may benefit future AI system design.

1104 **Voluntary Participation and Withdrawal**

1105 Participation is voluntary. You may withdraw at any time by closing your browser. If you do not complete the task, your data will not be saved. If you complete and submit the task, your data will be anonymized and cannot be withdrawn, as it will no longer be linked to you.

1106 **Confidentiality**

1107 • No personally identifying information will be collected.

1108 • All data will be stored securely and analyzed in aggregate.

1109 • Results may be published or presented, but individual participants will not be identifiable.

1110 **Questions or Concerns**

1111 If you have any questions about the study, you may contact the student researcher at: [REDACTED]

1112 If you have questions about your rights as a research participant, please contact the [REDACTED]

1113 **Consent Statement**

1114 By checking the box below and proceeding to the task, you confirm:

1115 • You are 18 years of age or older;

1116 • You have read and understood the information above;

1117 • You voluntarily consent to participate in this study.

1118

1119 I consent to participate in this study

1120 I do not consent to participate in this study

1121

1122

1123

Figure 12: The online questionnaire begins with a consent form.

F.2 STATISTICAL TEST RESULTS FOR SIGNIFICANCE

We provide statistical test results in Figure 15, 16, 17, 18.

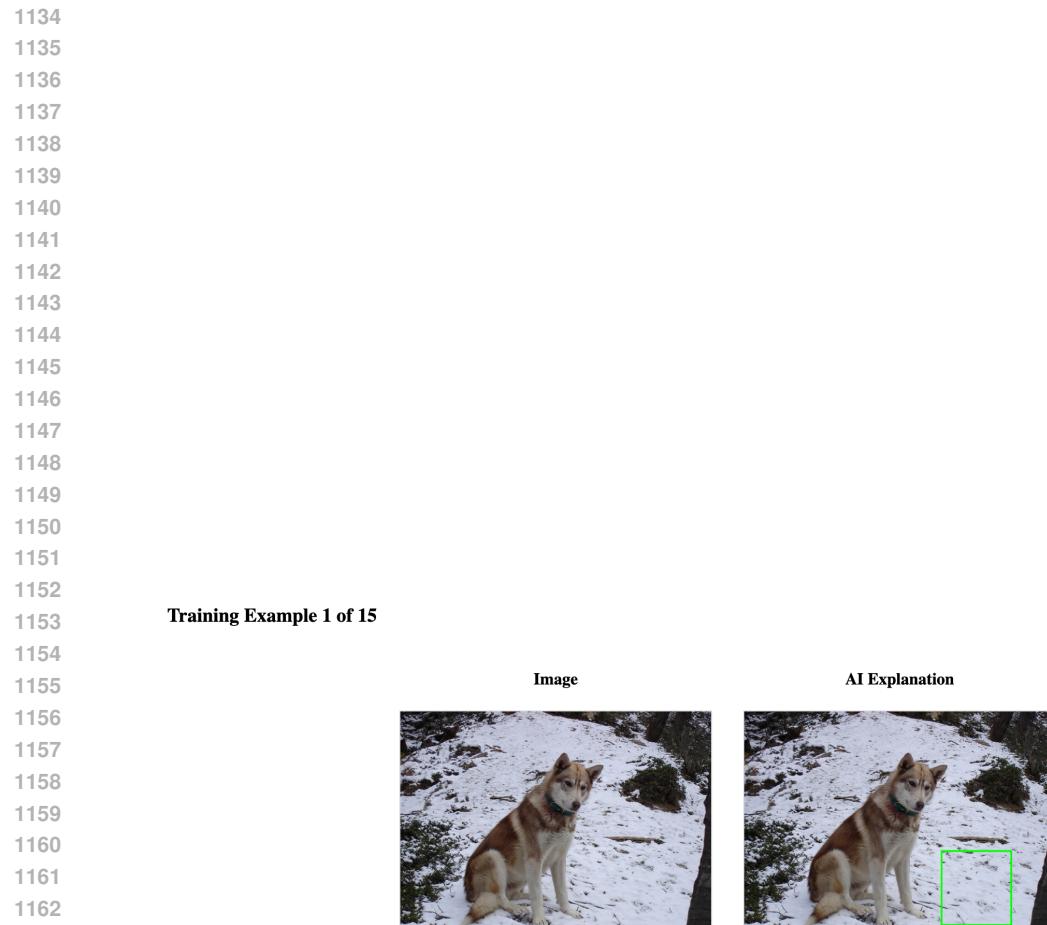


Figure 13: The online questionnaire displaying an example of the training session with the original image, the explanation, and the model prediction.

1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

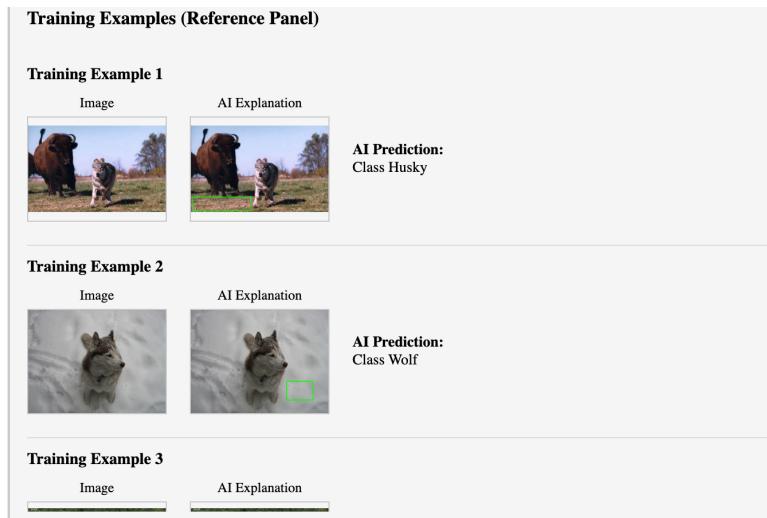
1229

1230

1231

1232

Figure 14: The online questionnaire displaying the testing session with a reservoir containing all examples during the training phase.



Test Session 1

Now predict what the AI will say for these images. You can refer to the training examples above.

Test Question 1 of 7:



- Model predicts Husky
- Model predicts Wolf

```

1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254 Starting analysis for EXP1
1255 =====
1256 Loading exp1 data files...
1257 =====
1258 ✓ Control : 31 participants loaded from Control_exp1_individual_results.csv
1259 ✓ Baseline: 25 participants loaded from Baseline_exp1_individual_results.csv
1260 ✓ ACE   : 40 participants loaded from ACE_exp1_individual_results.csv
1261 ✓ CRAFT  : 35 participants loaded from CRAFT_exp1_individual_results.csv
1262 ✓ LOGIC  : 30 participants loaded from LOGIC_exp1_individual_results.csv
1263
1264 Total participants loaded: 161
1265 Methods found: ['ACE', 'Baseline', 'CRAFT', 'Control', 'LOGIC']
1266 =====
1267 NORMALITY TEST (Shapiro-Wilk)
1268 =====
1269 H0: Data is normally distributed
1270 H1: Data is not normally distributed
1271 Significance level: 0.05
1272 -----
1273 Control | n=31 | W=0.9372 | p=0.0689 | Normal
1274 Baseline | n=25 | W=0.9521 | p=0.2790 | Normal
1275 ACE    | n=40 | W=0.9404 | p=0.0357 | Not Normal
1276 CRAFT   | n=35 | W=0.9376 | p=0.0473 | Not Normal
1277 LOGIC   | n=30 | W=0.8189 | p=0.0001 | Not Normal
1278
1279 Overall normality assumption: ✗ VIOLATED
1280 =====
1281 HOMOGENEITY OF VARIANCE TEST (Levene's Test)
1282 =====
1283 H0: All groups have equal variances
1284 H1: At least one group has different variance
1285 Significance level: 0.05
1286 -----
1287 Levene's statistic: 1.3810
1288 p-value: 0.2430
1289 Homogeneity assumption: ✓ MET
1290
1291 Group variances:
1292 Method | n | Variance | SD | SE
1293 -----
1294 Control | 31 | 0.0334 | 0.1827 | 0.0328
1295 Baseline | 25 | 0.0385 | 0.1962 | 0.0392
1296 ACE    | 40 | 0.0593 | 0.2435 | 0.0385
1297 CRAFT   | 35 | 0.0317 | 0.1780 | 0.0301
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2298
2299
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2398
2399
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2498
2499
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2598
2599
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2698
2699
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2789
2790
2791
2792
2793
2794
2795
2796
2797
2798
2798
2799
2799
2800
2801
2802
2803
2804
2805
2806
2807
2808
2809
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2859
2860
2861
2862
2863
2864
2865
2866
2867
2868
2869
2869
2870
2871
2872
2873
2874
2875
2876
2877
2878
2879
2879
2880
2881
2882
2883
2884
2885
2886
2887
2888
2889
2889
2890
2891
2892
2893
2894
2895
2896
2897
2898
2898
2899
2899
2900
2901
2902
2903
2904
2905
2906
2907
2908
2909
2909
2910
2911
2912
2913
2914
2915
2916
2917
2918
2919
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2939
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2949
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969
2969
2970
2971
2972
2973
2974
2975
2976
2977
2978
2979
2979
2980
2981
2982
2983
2984
2985
2986
2987
2988
2989
2989
2990
2991
2992
2993
2994
2995
2996
2997
2998
2998
2999
2999
3000
3001
3002
3003
3004
3005
3006
3007
3008
3009
3009
3010
3011
3012
3013
3014
3015
3016
3017
3018
3019
3019
3020
3021
3022
3023
3024
3025
3026
3027
3028
3029
3029
3030
3031
3032
3033
3034
3035
3036
3037
3038
3039
3039
3040
3041
3042
3043
3044
3045
3046
3047
3048
3049
3049
3050
3051
3052
3053
3054
3055
3056
3057
3058
3059
3059
3060
3061
3062
3063
3064
3065
3066
3067
3068
3069
3069
3070
3071
3072
3073
3074
3075
3076
3077
3078
3079
3079
3080
3081
3082
3083
3084
3085
3086
3087
3088
3089
3089
3090
3091
3092
3093
3094
3095
3096
3097
3098
3098
3099
3099
3100
3101
3102
3103
3104
3105
3106
3107
3108
3109
3109
3110
3111
3112
3113
3114
3115
3116
3117
3118
3119
3119
3120
3121
3122
3123
3124
3125
3126
3127
3128
3129
3129
3130
3131
3132
3133
3134
3135
3136
3137
3138
3139
3139
3140
3141
3142
3143
3144
3145
3146
3147
3148
3149
3149
3150
3151
3152
3153
3154
3155
3156
3157
3158
3159
3159
3160
3161
3162
3163
3164
3165
3166
3167
3168
3169
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3179
3180
3181
3182
3183
3184
3185
3186
3187
3188
3189
3189
3190
3191
3192
3193
3194
3195
3196
3197
3198
3198
3199
3199
3200
3201
3202
3203
3204
3205
3206
3207
3208
3209
3209
3210
3211
3212
3213
3214
3215
3216
3217
3218
3219
3219
3220
3221
3222
3223
3224
3225
3226
3227
3228
3229
3229
3230
3231
3232
3233
3234
3235
3236
3237
3238
3239
3239
3240
3241
3242
3243
3244
3245
3246
3247
3248
3249
3249
3250
3251
3252
3253
3254
3255
3256
3257
3258
3259
3259
3260
3261
3262
3263
3264
3265
3266
3267
3268
3269
3269
3270
3271
3272
3273
3274
3275
3276
3277
3278
3279
3279
3280
3281
3282
3283
3284
3285
3286
3287
3288
3289
3289
3290
3291
3292
3293
3294
3295
3296
3297
3298
3298
3299
3299
3300
3301
3302
3303
3304
3305
3306
3307
3308
3309
3309
3310
3311
3312
3313
3314
3315
3316
3317
3318
3319
3319
3320
3321
3322
3323
3324
3325
3326
3327
3328
3329
3329
3330
3331
3332
3333
3334
3335
3336
3337
3338
3339
3339
3340
3341
3342
3343
3344
3345
3346
3347
3348
3349
3349
3350
3351
3352
3353
3354
3355
3356
3357
3358
3359
3359
3360
3361
3362
3363
3364
3365
3366
3367
3368
3369
3369
3370
3371
3372
3373
3374
3375
3376
3377
3378
3379
3379
3380
3381
3382
3383
3384
3385
3386
3387
3388
3389
3389
3390
3391
3392
3393
3394
3395
3396
3397
3398
3398
3399
3399
3400
3401
3402
3403
3404
3405
3406
3407
3408
3409
3409
3410
3411
3412
3413
3414
3415
3416
3417
3418
3419
3419
3420
3421
3422
3423
3424
3425
3426
3427
3428
3429
3429
3430
3431
3432
3433
3434
3435
3436
3437
3438
3439
3439
3440
3441
3442
3443
3444
3445
3446
3447
3448
3449
3449
3450
3451
3452
3453
3454
3455
3456
3457
3458
34
```

```

1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308 LOGIC | 30 | 0.0299 | 0.1730 | 0.0316
1309 =====
1310 NON-PARAMETRIC ANALYSIS
1311 =====
1312 KRUSKAL-WALLIS TEST (Non-parametric)
1313 =====
1314 H0: All group distributions are identical
1314 H1: At least one group distribution differs
1314 Significance level: 0.05
1315 -----
1316 H-statistic: 25.8440
1316 p-value: 0.000034
1316 Degrees of freedom: 4
1316 Result: ✓ SIGNIFICANT
1317
1318 → Significant differences found between method distributions.
1318 → Proceeding to Dunn's post-hoc test...
1319
1320 =====
1321 DUNN'S TEST (Non-parametric post-hoc)
1322 =====
1323 Pairwise p-values (Bonferroni corrected):
1323 ACE Baseline CRAFT Control LOGIC
1324 ACE 1.0000 1.0000 0.5956 1.0000 0.0241
1324 Baseline 1.0000 1.0000 1.0000 1.0000 0.0303
1325 CRAFT 0.5956 1.0000 1.0000 1.0000 0.0000
1326 Control 1.0000 1.0000 1.0000 1.0000 0.0004
1326 LOGIC 0.0241 0.0303 0.0000 0.0004 1.0000
1327
1328 Pairwise Comparisons Summary (Bonferroni corrected,  $\alpha = 0.05$ ):
1328 =====
1329 ACE vs Baseline | p: 1.0000
1330 ACE vs CRAFT | p: 0.5956
1330 ACE vs Control | p: 1.0000
1331 ACE vs LOGIC | p: 0.0241 ***
1331 Baseline vs CRAFT | p: 1.0000
1332 Baseline vs Control | p: 1.0000
1332 Baseline vs LOGIC | p: 0.0303 ***
1333 CRAFT vs Control | p: 1.0000
1334 CRAFT vs LOGIC | p: 0.0000 ***
1334 Control vs LOGIC | p: 0.0004 ***
1335
1336
1337
1338
1339
1340
1341 Figure 16: Complete statistical test results for scenario 1, page 1.
1342
1343
1344
1345
1346
1347
1348
1349

```

```

1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362 Starting analysis for EXP2
1363 =====
1364 Loading exp2 data files...
1365 =====
1366 ✓ Control : 22 participants loaded from Control_exp2_individual_results.csv
1367 ✓ Baseline: 22 participants loaded from Baseline_exp2_individual_results.csv
1368 ✓ ACE   : 27 participants loaded from ACE_exp2_individual_results.csv
1369 ✓ CRAFT  : 25 participants loaded from CRAFT_exp2_individual_results.csv
1370 ✓ LOGIC  : 18 participants loaded from LOGIC_exp2_individual_results.csv
1371
1372 Total participants loaded: 114
1373 Methods found: ['ACE', 'Baseline', 'CRAFT', 'Control', 'LOGIC']
1374 =====
1375 NORMALITY TEST (Shapiro-Wilk)
1376 =====
1377 H0: Data is normally distributed
1378 H1: Data is not normally distributed
1379 Significance level: 0.05
1380 -----
1381 Control | n=22 | W=0.7308 | p=0.0000 | Not Normal
1382 Baseline | n=22 | W=0.6701 | p=0.0000 | Not Normal
1383 ACE    | n=27 | W=0.8543 | p=0.0014 | Not Normal
1384 CRAFT   | n=25 | W=0.8510 | p=0.0018 | Not Normal
1385 LOGIC   | n=18 | W=0.6087 | p=0.0000 | Not Normal
1386
1387 Overall normality assumption: ✗ VIOLATED
1388 =====
1389 HOMOGENEITY OF VARIANCE TEST (Levene's Test)
1390 =====
1391 H0: All groups have equal variances
1392 H1: At least one group has different variance
1393 Significance level: 0.05
1394 -----
1395 Levene's statistic: 2.5161
1396 p-value: 0.0455
1397 Homogeneity assumption: ✗ VIOLATED
1398
1399 Group variances:
1400 Method | n | Variance | SD | SE
1401 -----
1402 Control | 22 | 0.0230 | 0.1517 | 0.0323
1403 Baseline | 22 | 0.0282 | 0.1678 | 0.0358
1404 ACE    | 27 | 0.0222 | 0.1491 | 0.0287
1405 CRAFT   | 25 | 0.0123 | 0.1111 | 0.0222
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2198
2199
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2298
2299
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2388
2389
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2398
2399
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2498
2499
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2598
2599
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2698
2699
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2789
2790
2791
2792
2793
2794
2795
2796
2797
2797
2798
2799
2799
2800
2801
2802
2803
2804
2805
2806
2807
2808
2809
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2859
2860
2861
2862
2863
2864
2865
2866
2867
2868
2869
2869
2870
2871
2872
2873
2874
2875
2876
2877
2878
2879
2879
2880
2881
2882
2883
2884
2885
2886
2887
2888
2889
2889
2890
2891
2892
2893
2894
2895
2896
2897
2898
2898
2899
2899
2900
2901
2902
2903
2904
2905
2906
2907
2908
2909
2909
2910
2911
2912
2913
2914
2915
2916
2917
2918
2919
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2939
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2949
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969
2969
2970
2971
2972
2973
2974
2975
2976
2977
2978
2979
2979
2980
2981
2982
2983
2984
2985
2986
2987
2988
2989
2989
2990
2991
2992
2993
2994
2995
2996
2997
2998
2998
2999
2999
3000
3001
3002
3003
3004
3005
3006
3007
3008
3009
3009
3010
3011
3012
3013
3014
3015
3016
3017
3018
3019
3019
3020
3021
3022
3023
3024
3025
3026
3027
3028
3029
3029
3030
3031
3032
3033
3034
3035
3036
3037
3038
3039
3039
3040
3041
3042
3043
3044
3045
3046
3047
3048
3049
3049
3050
3051
3052
3053
3054
3055
3056
3057
3058
3059
3059
3060
3061
3062
3063
3064
3065
3066
3067
3068
3069
3069
3070
3071
3072
3073
3074
3075
3076
3077
3078
3079
3079
3080
3081
3082
3083
3084
3085
3086
3087
3088
3089
3089
3090
3091
3092
3093
3094
3095
3096
3097
3098
3098
3099
3099
3100
3101
3102
3103
3104
3105
3106
3107
3108
3109
3109
3110
3111
3112
3113
3114
3115
3116
3117
3118
3119
3119
3120
3121
3122
3123
3124
3125
3126
3127
3128
3129
3129
3130
3131
3132
3133
3134
3135
3136
3137
3138
3139
3139
3140
3141
3142
3143
3144
3145
3146
3147
3148
3149
3149
3150
3151
3152
3153
3154
3155
3156
3157
3158
3159
3159
3160
3161
3162
3163
3164
3165
3166
3167
3168
3169
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3179
3180
3181
3182
3183
3184
3185
3186
3187
3188
3189
3189
3190
3191
3192
3193
3194
3195
3196
3197
3198
3199
3199
3200
3201
3202
3203
3204
3205
3206
3207
3208
3209
3209
3210
3211
3212
3213
3214
3215
3216
3217
3218
3219
3219
3220
3221
3222
3223
3224
3225
3226
3227
3228
3229
3229
3230
3231
3232
3233
3234
3235
3236
3237
3238
3239
3239
3240
3241
3242
3243
3244
3245
3246
3247
3248
3249
3249
3250
3251
3252
3253
3254
3255
3256
3257
3258
3259
3259
3260
3261
3262
3263
3264
3265
3266
3267
3268
3269
3269
3270
3271
3272
3273
3274
3275
3276
3277
3278
3279
3279
3280
3281
3282
3283
3284
3285
3286
3287
3288
3289
3289
3290
3291
3292
3293
3294
3295
3296
3297
3298
3298
3299
3299
3300
3301
3302
3303
3304
3305
3306
3307
3308
3309
3309
3310
3311
3312
3313
3314
3315
3316
3317
3318
3319
3319
3320
3321
3322
3323
3324
3325
3326
3327
3328
3329
3329
3330
3331
3332
3333
3334
3335
3336
3337
3338
3339
3339
3340
3341
3342
3343
3344
3345
3346
3347
3348
3349
3349
3350
3351
3352
3353
3354
3355
3356
3357
3358
3359
3359
3360
3361
3362
3363
3364
3365
3366
3367
3368
3369
3369
3370
3371
3372
3373
3374
3375
3376
3377
3378
3379
3379
3380
3381
3382
3383
3384
3385
3386
3387
3388
3389
3389
3390
3391
3392
3393
3394
3395
3396
3397
3398
3398
3399
3399
3400
3401
3402
3403
3404
3405
3406
3407
3408
3409
3409
3410
3411
3412
3413
3414
3415
3416
3417
3418
3419
3419
3420
3421
3422
3423
3424
3425
3426
3427
3428
3429
3429
3430
3431
3432
3433
3434
3435
3436
3437
3438
3439
3439
3440
3441
3442
3443
3444
3445
3446
3447
3448
3449
3449
3450
3451
3452
3453
3454
3455
3456
3457
3458
3459
3459
3460
3461
3462
3463
3464
3465
3466
3467
3468
3469
3469
3470
3471
3472
3473
3474
3475
3476
3477
3478
3479
3479
3480
3481
3482
3483
3484
3485
3486
3487
3488
3489
3489
3490
3491
3492
3493
3494
3495
3496
3497
3498
3498
3499
3499
3500
3501
3502
3503
3504
3505
3506
3507
3508
3509
3509
3510
3511
3512
3513
3514
3515
3516
3517
3518
3519
3519
3520
3521
3522
3523
3524
3525
3526
3527
3528
3529
3529
3530
3531
3532
3533
3534
3535
3536
3537
3538
3539
3539
3540
3541
3542
3543
3544
3545
3546
3547
3548
3549
3549
3550
3551
35
```

```

1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416 LOGIC | 18 | 0.0011 | 0.0332 | 0.0078
1417 =====
1418 NON-PARAMETRIC ANALYSIS
1419 =====
1420 KRUSKAL-WALLIS TEST (Non-parametric)
1421 =====
1422 H0: All group distributions are identical
1423 H1: At least one group distribution differs
1424 Significance level: 0.05
1425 -----
1426 H-statistic: 13.5628
1427 p-value: 0.008830
1428 Degrees of freedom: 4
1429 Result: ✓ SIGNIFICANT
1430
1431 → Significant differences found between method distributions.
1432 → Proceeding to Dunn's post-hoc test...
1433 =====
1434 DUNN'S TEST (Non-parametric post-hoc)
1435 =====
1436 Pairwise p-values (Bonferroni corrected):
1437 ACE Baseline CRAFT Control LOGIC
1438 ACE 1.0000 1.0000 1.0000 1.0000 0.0045
1439 Baseline 1.0000 1.0000 1.0000 1.0000 0.1556
1440 CRAFT 1.0000 1.0000 1.0000 1.0000 0.0309
1441 Control 1.0000 1.0000 1.0000 1.0000 0.1463
1442 LOGIC 0.0045 0.1556 0.0309 0.1463 1.0000
1443
1444 Pairwise Comparisons Summary (Bonferroni corrected,  $\alpha = 0.05$ ):
1445 =====
1446 ACE vs Baseline | p: 1.0000
1447 ACE vs CRAFT | p: 1.0000
1448 ACE vs Control | p: 1.0000
1449 ACE vs LOGIC | p: 0.0045 ***
1450 Baseline vs CRAFT | p: 1.0000
1451 Baseline vs Control | p: 1.0000
1452 Baseline vs LOGIC | p: 0.1556
1453 CRAFT vs Control | p: 1.0000
1454 CRAFT vs LOGIC | p: 0.0309 ***
1455 Control vs LOGIC | p: 0.1463
1456
1457

```

Figure 18: Complete statistical test results for scenario 2, page 2.