École Polytechnique Fédérale de Lausanne

Harvard Medical School

Beyond Anecdotal Evidence: A Systematic Framework for Evaluating Neuron Interpretability

Ernesto Bocini
Karstlernstrasse 12, 8048 Zürich
ernesto.bocini@epfl.ch

Master Thesis

Prof. Gabriel Kreiman - Kreiman Lab, Harvard Medical School, Boston Children's Hospital
Thesis Advisor

Prof. Antoine Bosselut - NLP LAB, EPFL
Thesis Co-Advisor

August 29, 2025

Acknowledgments

To my family, my constant anchor through every move: from Prato to Bologna, Lund, Lausanne, Boston, and finally Zurich. None of this would have been possible without your support.

To the friends and people who colored my three crazy years at EPFL, thank you for the laughter, late nights, and belief.

To Prof. Gabriel Kreiman, thank you for exceptional supervision, patient guidance, and the many things I learned from working with you.

This thesis was made possible thanks to the generous support of the Hasler Foundation, which fully financed my work at Harvard University and turned a dream into reality.

Lausanne, August 29, 2025 Ernesto Bocini

Karstlernstrasse 12, 8048 Zürich
ernesto.bocini@epfl.ch

Abstract

Mechanistic interpretability research commonly relies on activation selectivity as the primary evaluation criterion when claiming to understand individual neurons. While the limitations of single-metric evaluation are increasingly recognized, systematic frameworks for multi-dimensional neuron-level validation remain underdeveloped. This thesis introduces a compact, multi-axis framework for neuron evaluation that combines *Statistical Selectivity (S)*, *Causal Impact (C)*, *Robustness (R)*, and *Human Consistency (H)* into an equal-weight composite, INTERPSCORE.

We instantiate the framework on CLIP RN50x4 at image block 4/5/ReLU 2 (2,560 channels), evaluating ten high-selectivity neurons with curated natural/control sets, semantics-preserving perturbations (Gaussian noise, mosaics), small-budget PGD, and DEEPDREAM maxima. Human recognizability is measured via crowd annotations on the same stimuli. Causal impact is quantified with bidirectional interventions at the site (ablation, amplification) while verifying attention-pooling parity so that measured embedding shifts reflect the intervention.

INTERPSCORE separates neurons more than selectivity alone: S is saturated with negligible between-neuron variation ($SD \approx 2.3e-4$), whereas InterpScore varies modestly ($SD \approx 0.0716$, $\sim 14\%$ of its mean). A human-free variant, $\text{InterpScore}_{\neg H} = (S + C + R)/3$, predicts human recognizability better than selectivity (explained-variance gain $\Delta R^2 \approx 0.20$). Scores and rankings are numerically stable across seeds and perturbations.

These results move neuron-level claims beyond anecdotes toward a more objective, reproducible basis for assessing "what is more interpretable."

Keywords: mechanistic interpretability; neuron evaluation; causal interventions; robustness; human alignment; CLIP; DeepDream.

Contents

Chapter 1

Introduction

Mechanistic interpretability seeks to reverse-engineer neural networks by identifying human-understandable computational units and algorithms within their learned representations [2, 10, 23]. A central challenge in this endeavor is evaluation: *how do we determine whether our interpretations of neural network components are accurate and meaningful?* Traditional approaches have relied heavily on activation selectivity, measuring how consistently a neuron responds to specific input categories [1, 10, 23]. Approaches that had originally drawn inspiration from neuroscience findings of highly selective "concept cells", such as neurons that respond specifically to Jennifer Aniston [26], which demonstrated that individual biological neurons can exhibit remarkable selectivity for specific concepts. A neuron that fires strongly for images of dogs and weakly for other animals might be labeled a "dog detector", with this interpretation supported primarily by its selective activation pattern.

The community has increasingly recognized that activation selectivity alone provides insufficient evidence for robust assessment. Recent frameworks and benchmarks formalize this concern from multiple angles: multi-dimensional evaluation standards in MIB [20], systematic validation for neuron explanations [22], and comprehensive assessments in sparse-feature work [18]. Yet, despite progress at the circuit and feature levels, *neuron-level* evaluation remains comparatively underdeveloped. Neurons are the atomic units from which circuits and features are composed; without reliable validation at this granularity, higher-level analyses rest on a weak empirical foundation.

A concrete illustration motivates the gap. A CLIP neuron may exhibit high selectivity for a visually coherent concept (e.g., Arabic text), suggesting interpretability at first glance. However, such selectivity does not answer whether the unit has causal impact on downstream representations, whether its behavior is robust to non-semantic perturbations, or whether humans consistently recognize the claimed concept. In short, single-axis evidence leaves central questions unresolved.

This paper addresses that gap by proposing a simple, multi-axis evaluation for neuron-level

interpretability. We integrate four complementary dimensions: *Selectivity (S)*, *Causality (C)*, *Robustness (R)*, and *Human Consistency (H)*. Our guiding question is: can a compact evaluation move mechanistic interpretability beyond anecdotes toward a more objective and discriminative measure at the neuron level?

We study CLIP RN50x4 and intervene at `image_block_4/5/ReLU_2` (2,560 neurons), evaluating ten high-selectivity neurons. Three empirical findings support the need for multi-axis assessment. First, relative to selectivity alone, the composite substantially increases dispersion across neurons: S is saturated (~ 1.00) with negligible between-neuron variation (SD $\approx 2.3 \times 10^{-4}$), whereas *InterpScore* varies modestly (SD ≈ 0.0716, $\sim$14% of its mean), revealing differences that a single metric obscures. Second, a human-free variant of the composite (averaging S, C, and R) aligns more strongly with human recognizability than selectivity alone, explaining notably more variance in human scores ($\Delta R^2 \approx 0.20$). Third, the metrics and resulting rankings are numerically stable under standard seeds and benign perturbations, providing a reproducible basis for comparison.

We release protocols and implementation details for systematic neuron evaluation and provide specific-neurons visualization tool in the spirit of Microscope [10] to facilitate careful inspection of the 2,560 neurons. Our aim is methodological rather than tool-centric: to replace anecdotal judgments with a compact, reproducible procedure that surfaces meaningful differences among neurons and clarifies what counts as "more interpretable."

Main Research Question (MRQ). Can a naive, multi-axis evaluation move mechanistic interpretability beyond anecdotes toward a more *objective* and *discriminative* measure at the neuron level? We study CLIP RN50x4 at `image_block_4/5/ReLU_2` (2,560 channels).

We make four contributions toward principled, neuron-level evaluation:

- **Framework.** A compact, four-axis evaluation for individual neurons: Selectivity (S), Causality (C), Robustness (R), and Human Consistency (H), summarized by an equal-weight composite (*InterpScore*).
- **Empirical validation at a fixed site.** On ten high-selectivity neurons in CLIP RN50x4 at `image_block_4/5/ReLU_2`, the composite yields substantially greater dispersion across neurons than selectivity alone (**Evidence 1**).
- **Alignment with human recognizability.** A human-free variant (InterpScore$\neg H = (S+C+R)/3$) aligns better with human scores than selectivity alone and explains more variance (gain $\Delta R^2 \approx 0.20$; **Evidence 2**).
- **Stability & tooling.** Metrics and rankings are numerically stable across standard seeds/benign perturbations (**Evidence 3**). We also provide enhanced Microscope-style inspections to support systematic analysis (used as a validation aid rather than a central claim).

Chapter 2

Multi-Dimensional Framework for Neuron Interpretability

The transition from compelling anecdotal discoveries toward systematic evaluation represents a natural maturation of the interpretability field. Early striking examples, from artificial neurons detecting visual concepts [1, 10] to learned features like the "Golden Gate Bridge feature" [2], have shaped our understanding of interpretable representations. However, while such examples provide important intuitions, they may not represent typical behavior and can mask critical failure modes invisible to activation-based assessment. Current interpretability evaluation typically relies on activation-based analysis: identify computational units showing high selectivity for specific concepts, then interpret their function based on activation patterns. Beyond neuron selectivity, standard attribution and concept-based baselines include gradient- and cam-based saliency, randomized masking, and concept activation methods, which we use as reference points for scope and claims [9, 14, 25, 28–30]. Our work extends the systematic evaluation principles established by recent frameworks [18, 20, 22] specifically to neuron-level assessment, addressing fundamental limitations through multi- dimensional evaluation.

Figure 2.1 presents our complete framework for systematic multi-dimensional neuron interpretability assessment.

2.1 Model and intervention site

Following [10], we analyze **CLIP RN50x4**. The backbone follows the ResNet family [12], and the attention pooling we target builds on transformer-style multi-head attention [31]. We intervene at the last convolutional block before attention pooling (`image_block_4/5/ReLU_2`). This layer represents the final output of the deepest ResNet block before attention processing, making it ideal

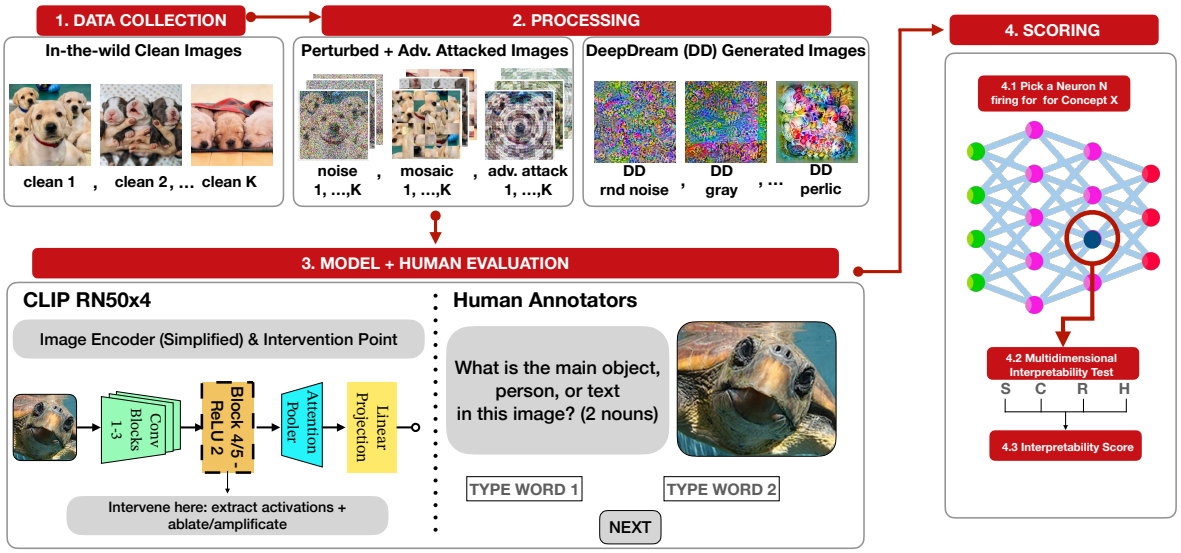

Figure 2.1: **Framework overview**. **(1) Data collection:** K clean images, collected by systematically scraping the web for each topic of interest. *Default K = 30.* **(2) Processing:** standardized perturbation protocols (Gaussian noise, mild blur/JPEG, small geometric jitter) and small-budget adversarial stress are applied to the clean images. synthetic maximization stimuli (DeepDream) is also used to create images to probe human recognizability (details in App. B.3). **(3) Model + Human Evaluation:** We extract neuron activations from CLIP **RN50x4** at `image_block_4/5/ReLU_2` (2.560 channels) for all image conditions from steps 1-2, while human annotators evaluate the exact same images for concept recognition; causal testing involves ablating ($\lambda = 0$) and amplifying ($\lambda = 2$) individual neuron activations to measure impact on final 640-dimensional embeddings. **(4) Scoring:** integration across **S**electivity, **C**ausality, **R**obustness, and **H**uman consistency; the composite (*InterpScore*) summarizes the axes.

for causality testing as it captures high-level semantic features while preserving all downstream computation.

This site has spatial size 9×9 at input resolution 288×288 and channel width $2,560$. Let $a_N(x) \in \mathbb{R}^{9\times9}$ denote the post-ReLU activation map of neuron N. Downstream, features are reshaped and pooled as

$$(1,2,560,9,9) \rightarrow (1,2,560,81) \rightarrow (1,81,2,560)$$

and passed through a single-layer multi-head attention pooler (40 heads × 64-d) followed by a linear projection to a 640-dimensional image embedding.[1]

[1] CLIP ResNets replace GAP with a single-layer multi-head attention pooler [31]: a summary token attends over the 81 spatial tokens, and the first output token is projected to the 640-d embedding.

2.2 Preprocessing and notation

Images follow the CLIP preprocessing. Let $\mathscr{D}$ be the evaluation set, $X \subset \mathscr{D}$ a concept subset, and $\bar{X} = \mathscr{D} \setminus X$. We define a scalar response via global max pooling $a_N(x) = \max_{h,w} A_N(x)_{h,w}$, where $A_N(x) \in \mathbb{R}^{9\times 9}$ denote the post-ReLU activation map of neuron N for image x. Expectations below are over the indicated sets.

2.3 Statistical Selectivity (S)

We quantify separability between activations on *clean* images of concept X versus non-X using an unbiased effect size and map it to a bounded score. Let $\mathscr{D}_{\text{clean}} \subset \mathscr{D}$ denote the clean subset. Write $a_N(x)$ for neuron N's activation on image x (defined as above) and let $n_X, n_{\bar{X}}$ be sample sizes; $\mu_X, \mu_{\bar{X}}$ the sample means; and $s^2_X, s^2_{\bar{X}}$ the sample variances (unbiased, $ddof$=1) of $a_N(x)$ for $x \in X \cap \mathscr{D}_{\text{clean}}$ and $x \in \bar{X} \cap \mathscr{D}_{\text{clean}}$, respectively. We form the pooled standard deviation

$$s_p = \sqrt{\frac{(n_X - 1)s^2_X + (n_{\bar{X}} - 1)s^2_{\bar{X}}}{n_X + n_{\bar{X}} - 2}},$$

Cohen's $d = (\mu_X - \mu_{\bar{X}})/s_p$, and Hedges' correction $J = 1 - \frac{3}{4(n_X + n_{\bar{X}}) - 9}$. We follow conventional effect-size practice for standardized mean differences [6, 13]. Our selectivity score is then

$$S(N, X) = \Phi\left(\frac{Jd}{\sqrt{2}}\right) \in [0, 1], \tag{2.1}$$

where Φ is the standard normal CDF. This yields $S = 0.5$ when there is no separation and increases monotonically as activations on X exceed those on $\bar{X}$. (Under an equal-variance normal model, this is equivalent to ROC-AUC).

2.4 Causal Impact (C)

We measure functional relevance by the embedding shift induced by bidirectional interventions at the site. Let $E(x) \in \mathbb{R}^{640}$ be the baseline embedding and $E^\lambda(x)$ the embedding when scaling neuron N's activation map by factor λ (elementwise at the site; all else unchanged). For clean exemplars of X (level 5), define

$$\Delta_\lambda(x) = \frac{\|E^\lambda(x) - E(x)\|_2}{\|E(x)\|_2}, \qquad C_{\text{raw}}(N, X) = \tfrac{1}{2}\Big(\mathbb{E}_{x\in X}\big[\Delta_0(x)\big] + \mathbb{E}_{x\in X}\big[\Delta_2(x)\big]\Big). \tag{2.2}$$

For aggregation and comparability with other components, we report a bounded variant obtained via a monotone exponential remap:

$$\tilde{C}(N,X) \;=\; 1-\exp\big(-C_{\text{raw}}(N,X)\big) \in [0,1).$$

In practice, to reduce runtime we approximate the expectation over $x \in X$ by averaging over a small random subset of clean exemplars (k images; k=30 by default).

Note (attention-path reconstruction as validation). Our interventions are executed by re-running the image attention-pooling subgraph outside the host framework; even tiny mismatches (e.g., token ordering, reshape/transpose, Q/K/V projections, parameter loading, broadcasting, or mixed-precision effects) could contaminate $\Delta_\lambda(x)$ and be mistaken for causal impact. To preserve internal validity, before measuring $C(N,X)$ we therefore *certify fidelity* of the subgraph: we re-run the path with λ=1 and require the resulting embedding to match the model's native forward pass within a tight tolerance (max-abs difference $\le 10^{-6}$). Items failing this parity test are excluded from C; the reconstruction serves only to ensure that measured shifts stem from the λ-scaling intervention itself. Appendix B.1 details operator/weight mapping, numerical settings, coverage statistics, and additional checks (round-trip λ=1 and monotonicity for $\lambda \in \{0,2\}$).

2.5 Robustness (R)

Robustness captures stability under semantically preserving perturbations and small adversarial stress, restricted to items that remain human-recognizable under our QC protocol (`soft_correct=1`, details in Appendix A.2). Let X_{clean} be the clean images of concept X (level 5). We partition perturbed data into a benign set $X^{\text{rec}}_{\text{benign}}$ (levels 1-2) and an adversarial set $X^{\text{rec}}_{\text{adv}}$ (level 3). Synthetic maximization stimuli (DeepDream; level 4) are excluded from R and analyzed separately (Sec. 2.8).

Denote the mean absolute activations

$$A_N^{\text{clean}}(X) = \mathbb{E}_{x\in X_{\text{clean}}}\big[\,|a_N(x)|\,\big], \quad A_N^{\text{benign}}(X) = \mathbb{E}_{x\in X^{\text{rec}}_{\text{benign}}}\big[\,|a_N(x)|\,\big], \quad A_N^{\text{adv}}(X) = \mathbb{E}_{x\in X^{\text{rec}}_{\text{adv}}}\big[\,|a_N(x)|\,\big].$$

With $r_{\text{ben}} = \frac{A_N^{\text{benign}}(X)+\varepsilon}{A_N^{\text{clean}}(X)+\varepsilon}$ and $r_{\text{adv}} = \frac{A_N^{\text{adv}}(X)+\varepsilon}{A_N^{\text{clean}}(X)+\varepsilon}$ (numerical $\varepsilon = 10^{-8}$), we use a symmetric multiplicative stability function

$$\sigma(r) = \exp\big(-|\log r|\big) = \min(r, 1/r) \in (0,1],$$

and define

$$R^{\text{inv}}(N,X) = \sigma(r_{\text{ben}}), \qquad R^{\text{adv}}(N,X) = \sigma(r_{\text{adv}})$$

Finally:

$$R(N,X) = \tfrac{1}{2}\big(R^{\text{inv}}(N,X) + R^{\text{adv}}(N,X)\big) \tag{2.3}$$

Thus $R = 1$ when perturbed means match the clean mean, and R decreases smoothly and symmetrically for multiplicative drops or spikes.

This perspective aligns with the view that models often lean on non-robust, yet highly predictive, features [**ilyas2019featuresnotbugs**].

2.6 Human Consistency (H)

Human recognizability is measured from blinded annotations over a curated pool of evaluation items for each (N, X). We first compute a neuron-specific activation threshold from the non-X *clean* distribution:

$$\tau_N = Q_{0.95}\big(a_N(x) \,\big|\, x \in \bar{X} \cap \mathscr{D}_{\text{clean}}\big),$$

i.e., the 95$^{\text{th}}$ percentile of activations on clean images that do not belong to concept X. We then form the selection set

$$\mathscr{S}(N, X) = \big\{x : \text{ground-truth}(x) = X,\ a_N(x) > \tau_N\big\} \cup \mathscr{V}_N^{\text{DD}},$$

where the first term collects top-activating natural images for X across levels (clean, benign perturbations, small adversarial stress), and $\mathscr{V}_N^{\text{DD}}$ are the neuron's DeepDream maximization stimuli (Sec. 2.8). Each item $i \in \mathscr{S}(N, X)$ receives a blinded binary label $h_i \in \{0, 1\}$ under our QC protocol (1 = depicts X; 0 = otherwise), and we report

$$H(N, X) = \frac{1}{|\mathscr{S}(N, X)|} \sum_{i \in \mathscr{S}(N,X)} h_i \in [0, 1], \tag{2.4}$$

with H=1 indicating perfect agreement. When $\mathscr{S}(N, X)$ is empty we set $H(N, X) = 0$ by convention. Full annotation and QC details (blinding, minimum dwell time, and compensation) are in App. A.2.

2.7 Interpretability-Score (InterpScore)

We aggregate along four axes with equal weights and without discretization:

$$\text{InterpScore}(N, X) = \tfrac{1}{4}\big(S(N, X) + C(N, X) + R(N, X) + H(N, X)\big). \tag{2.5}$$

All metrics are computed per neuron; aggregate statistics (e.g., coefficients of variation across neurons) are reported in the Results Section.

2.8 Synthetic maximization stimuli (DeepDream)

We use DeepDream optimization [19] to generate *maximally activating stimuli* for each neuron. Related feature-visualization and inversion approaches include early gradient-ascent visualizations, optimization-based inversions, and interactive atlas views [4, 8, 17, 21, 32]. These stimuli are included *within the experimental pool* evaluated by human annotators; they probe the boundary of recognizability and thus directly affect the **H (human consistency)** axis. Neurons that exploit non-semantic regularities typically score lower on H when judged on DeepDream images. We instantiate multiple initializations (gray, random/structured noise, gradient and Perlin patterns) and a 4-octave pyramid (72→288 px; scale factor 1.4). Parameterization and optimization details (step sizes, iteration schedules, and regularizers) are provided in App. B.3. We do *not* fold DeepDream into the robustness $R(\cdot)$, which is defined on semantically preserving transforms of natural images. The DeepDream condition serves as a particularly revealing interpretability litmus test: if synthetic images that maximally activate the neuron are unrecognizable to humans, this indicates the neuron responds to "syntactic tricks in pixel space" rather than meaningful semantic content.

Chapter 3

Experimental Setup

3.1 Model and Neuron Selection

We evaluate our framework using OpenAI CLIP RN50x4 [27] vision encoder with 288x288 pixel input resolution, targeting the `image_block_4/5/ReLU_2` layer (2,560 channels). This model provides a well-studied architecture with established interpretability research, enabling direct comparison with existing findings including the foundational multimodal neurons analysis [10]. Our neuron selection combines multiple approaches: integration of neurons from existing interpretability literature, statistical pre-screening using concept versus control image sets, and diversity sampling across semantic categories. To address the critical reproducibility gap created by the unavailability of the original Microscope visualization tool [10], we recreated and enhanced this foundational infrastructure. Our enhanced version provides systematic exploration capabilities with improved statistical analysis tools and includes Lucid-generated feature visualizations [24](Appendix B.4).

Table 3.1: **Selected neurons** from CLIP RN50x4 [27] at layer `image_block_4/5/Relu_2`. Neuron IDs indicate channel indices within the 2,560-dimensional feature map. Concepts span diverse semantic categories to test framework generalizability. Each concept name it's a link to the Enhanced Microscope-style visualization tool (recreating [10])

#	Concept	ID	Category
1	Trump	89	Political figures
2	Arabic Alphabet	479	Text/language
3	Puppies	355	Animals
4	Sailboat	363	Objects
5	Fire	297	Natural elements
6	Australia	513	Geography
7	Droplets	967	Phenomena
8	Raised Hand	1116	Gestures
9	Mushroom	1157	Biological forms
10	Fashion Model	1424	Human figures

The tool is publicly available at `https://microscope-clip.streamlit.app/` with a live demo accessible at `https://github.com/ernestoBocini/rebuilt-microscope-CLIP.git`, enabling researchers to immediately explore

CLIP neuron representations (detailed in Appendix B.2).

Our selection spans multiple semantic categories and can be observed in Table 3.1

3.2 Dataset Construction and Evaluation Infrastructure

For each concept, we construct carefully curated datasets with manual verification to ensure quality and consistency, supporting all four evaluation dimensions as outlined in Figure 2.1.

Concept Images: Target concept images are manually verified to belong to the specific topic. We ensure sufficient diversity in presentation, context, and visual characteristics, with matched control sets avoiding concept contamination while maintaining similar visual complexity.

Perturbation Protocol: Our robustness evaluation employs four perturbation types designed to test different aspects of interpretability stability: (1) **Gaussian noise** at various levels (σ = 0.1, 0.2, 0.3) testing basic robustness to pixel-level corruption, (2) **Mosaics** made by shuffling portions of the image at different ratios (25%, 50%, 75%) to test spatial disruption tolerance, (3) **Adversarial attacks** in line with canonical PGD/first-order formulations and related attacks [3, 11, 16], with small perturbations testing optimized perturbation resistance, and (4) **DeepDream synthetic images** optimized for specific neurons using gradient ascent techniques, providing maximally activating synthetic stimuli for semantic coherence testing. Complete implementation details are provided in Appendix B.3.

Critically, we only evaluate activation robustness on perturbed images that humans still recognize as containing the target concept through systematic validation, ensuring robustness evaluation aligns with human perception rather than arbitrary mathematical transformations.

Human Evaluation Infrastructure: Our human consistency validation employs rigorous experimental design using Prolific Academic for controlled crowd-sourcing with English-speaking participants (age 18–65, normal vision). Comprehensive quality control includes attention checks distributed throughout experiments, response time monitoring to detect careless participation, and consistency validation through repeated items. We collect 20-30 participants per image condition (total of 110 participants), providing sufficient statistical power for reliable assessment. Complete participant demographics, quality control measures, and inter-rater reliability statistics are detailed in Appendix A.2.

Chapter 4

Results

We address the MRQ using three lines of evidence on CLIP RN50x4 at `image_block_4/5/ReLU_2` (2,560 channels): *discrimination*, *alignment*, and *stability*. Unless noted, metrics are computed per neuron and aggregated across the ten high-selectivity units. Complete results can be observed in 4.1. A worked example showcasing the approach is in **??**

4.1 Evidence 1: Discrimination

As shown in Fig. 4.2, activation selectivity is saturated (~ 1.00) with negligible between-neuron variation (SD $\approx 2.3 \times 10^{-4}$), whereas the composite *InterpScore* varies more across neurons (SD ≈ 0.0716, $\sim$14% of its mean). This contrast indicates that relying on Selectivity alone can mask between-neuron differences that *InterpScore* makes evident.

Supporting validation. At the neuron level, we observe a causality–robustness trade-off (Spearman's $\rho = -0.76$, $p = 0.011$, $n = 10$). Correlation with the full framework crosses the strong-correlation threshold ($r = 0.9$) using three axes, and pairwise comparisons yield large effect sizes (Cohen's $d > 2$) for most component pairs; see Appendix for full analyses.

4.2 Evidence 2: Alignment with human recognizability

To avoid circularity, we exclude the human term and compute the human-free variant

$$\text{InterpScore}\neg H = (S + C + R)/3.$$

Table 4.1: **Multi-dimensional interpretability assessment** of ten high-selectivity neurons. Individual component scores reveal distinct interpretability profiles: Selectivity (S) measures statistical separability between concept and non-concept images; Causality (C) quantifies functional impact on downstream embeddings via ablation/amplification; Robustness (R) evaluates consistency under semantically-preserving perturbations; Human Consistency (H) measures agreement with human concept recognition. The composite InterpScore integrates all dimensions with equal weighting. Standard errors from bootstrap resampling (1000 samples). Notable cases include Arabic Alphabet (perfect selectivity but zero causality) and Raised Hand (moderate selectivity but strong human consistency).

Concept	S	C	R	H	InterpScore
Fire	**1.000**±0.003	0.216±0.015	0.377±0.025	**0.871**±0.018	**0.616**±0.007
Raised Hand	0.999±0.008	0.177±0.020	0.234±0.035	0.840±0.017	0.563±0.010
Trump	**1.000**±0.003	0.279±0.022	0.239±0.038	0.613±0.026	0.533±0.009
Sailboat	**1.000**±0.004	0.136±0.016	0.161±0.032	0.834±0.019	0.533±0.008
Droplets	**1.000**±0.005	0.157±0.018	0.313±0.028	0.634±0.024	0.526±0.009
Mushrooms	**1.000**±0.004	0.084±0.012	0.168±0.022	0.776±0.020	0.507±0.008
Arabic Alphabet	**1.000**±0.003	0.207±0.008	**0.478**±0.027	0.252±0.034	0.484±0.012
Fashion Model	**1.000**±0.006	**0.310**±0.025	0.067±0.042	0.440±0.031	0.454±0.011
Australia	**1.000**±0.007	0.102±0.019	0.314±0.033	0.328±0.036	0.436±0.013
Puppies	**1.000**±0.005	0.198±0.023	0.040±0.040	0.204±0.038	0.361±0.014

This variant aligns more strongly with human recognizability than Selectivity alone: Spearman's $\rho = 0.33$ (p=0.347) and Kendall's $\tau = 0.29$ (p=0.291) between InterpScore$\neg H$ and H. A regression comparison shows an explained-variance gain of $\Delta R^2 \approx 0.20$ for $H \sim$ InterpScore$\neg H$ over $H \sim S$. These results indicate that the multi-axis structure captures human-salient variation that a single metric misses.

4.3 Evidence 3: Stability

We assess numerical and ranking stability across seeds/perturbations and verify implementation parity within a small tolerance (e.g., $\leq 10^{-6}$). A two-seed analysis shows small per-metric deviations and high rank consistency (Table 4.2).

Table 4.2: Two-seed stability (ten neurons). Mean and max absolute changes across neurons.

Metric	Mean $\lvert\Delta\rvert$	Max $\lvert\Delta\rvert$
Selectivity (S)	0.0010	0.0030
Causality (C)	0.0062	0.0181
Robustness (R)	0.0044	0.0129
InterpScore$\neg H$	0.0031	0.0102
InterpScore	0.0028	0.0091

Ranking stability: Kendall's τ is 0.92 for InterpScore$\neg H$ and 0.93 for InterpScore (bootstrap 95% CIs: 0.78-0.98 and 0.80-0.98); ICC(1,k) for InterpScore is 0.97 (95% CI 0.92-0.99).

Summary. Together, the three evidences, *better discrimination, better human alignment without circularity*, and *numerical/ranking stability*, directly answer the MRQ and establish a reproducible basis for comparing "what is more interpretable."

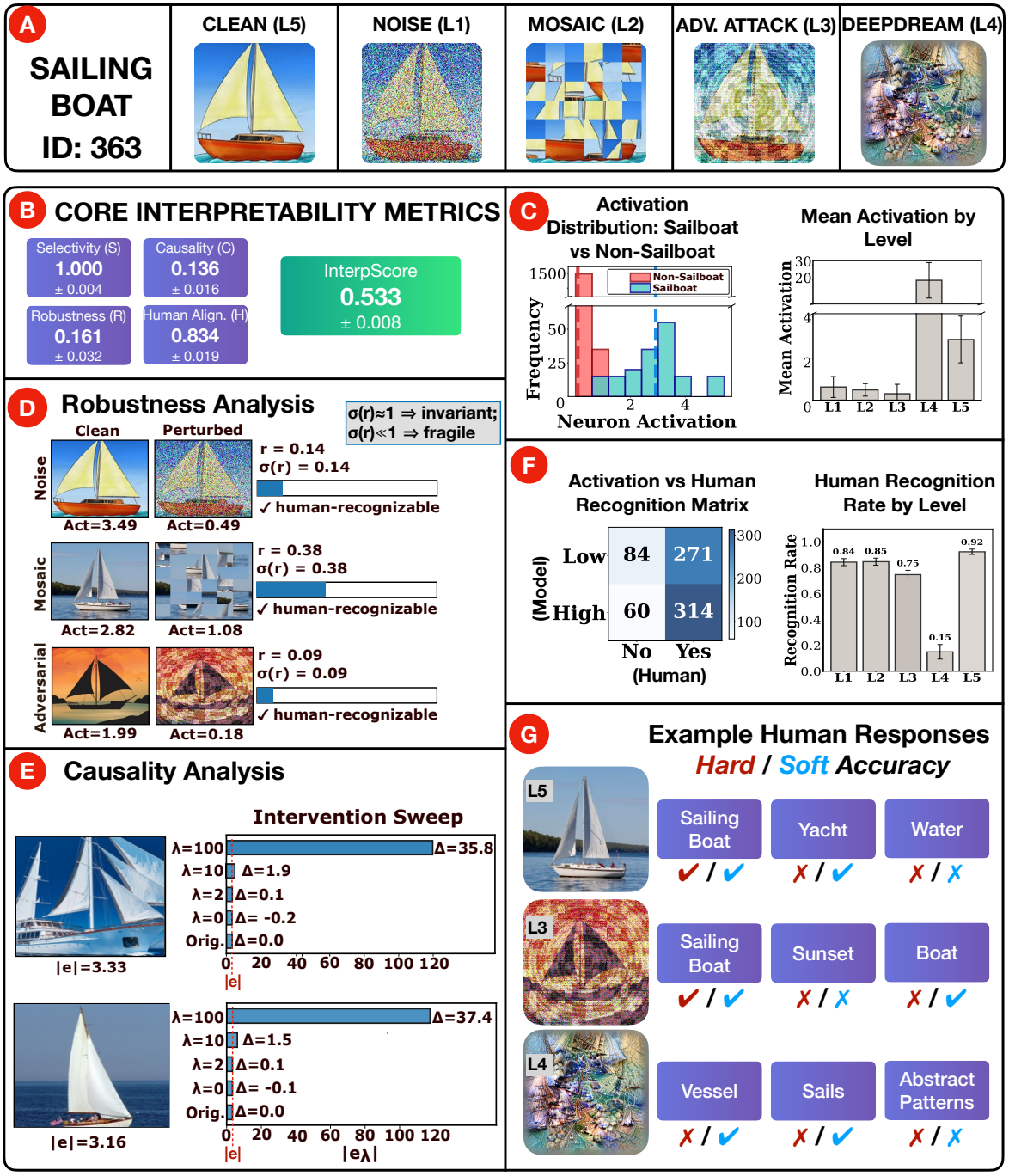

Figure 4.1: **Worked example: Sailboat neuron (#363). (A)** Stimuli: **L1** noise, **L2** mosaic, **L3** adversarial, **L4** DeepDream, **L5** clean. **(B)** Core metrics: S (selectivity), C (causality), R (robustness), H (human consistency), and INTERPSCORE. **(C)** Selectivity: target vs. non-target activations with the neuron threshold; mean activation by level. **(D)** Robustness examples: three matched clean→perturbed pairs (L1–L3) with per-image activation, perturbed/clean ratio, and a stability indicator; rows marked ✓ are human-recognizable and counted in **R** (L4 excluded). **(E)** Causality examples: for two images, horizontal bars show embedding norms for *Baseline*, *Ablate* (λ=0), and *Amplify*; inline labels give the relative embedding change, and **C** aggregates ablation with small amplification. **(F)** Human consistency: activation (high/low via the threshold) vs. recognition (no/yes) with rates by level. **(G)** Human examples: two free-text answers per image scored as *Hard* (exact label; ✓ if either matches) and *Soft* (close variants/synonyms/typos; ✓ if either qualifies); soft-correct items define recognition for **R** and contribute to **H**.

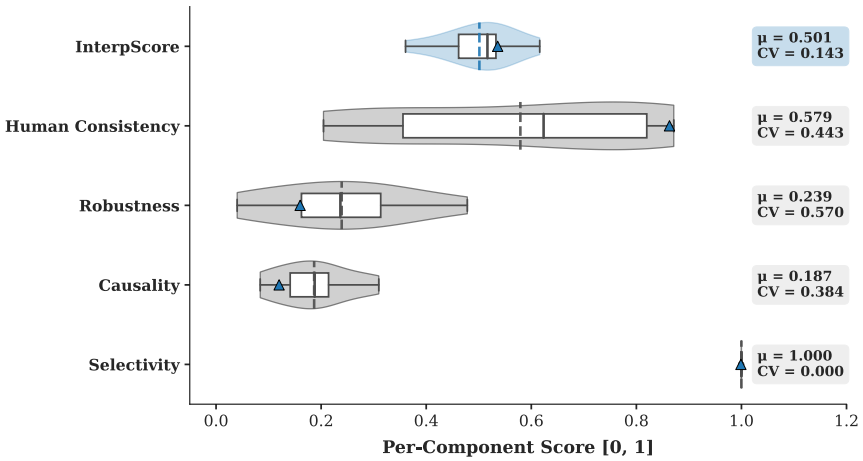

Figure 4.2: **Selectivity is near ceiling; other components and InterpScore show variability.** Each violin shows the distribution across 10 neurons for one component (S, C, R, H, InterpScore). Box = IQR; solid line = median; dashed line = mean (μ); points = individual neurons (N=10). Labels display μ and CV (σ/μ) across neurons: S ($\mu \approx 1.000$, CV ≈ 0.000), C ($\mu \approx 0.187$, CV ≈ 0.384), R ($\mu \approx 0.239$, CV ≈ 0.570), H ($\mu \approx 0.579$, CV ≈ 0.443), InterpScore ($\mu \approx 0.501$, CV ≈ 0.143). ▲ marks Sailboat (#363) as in Fig. 4.1.

Chapter 5

Discussion and Implications

5.1 Methodological Contributions

Our framework integrates *statistical selectivity*, *causal impact*, *robustness*, and *human consistency* into uniform protocols at a fixed site in CLIP RN50x4 (`image_block_4/5/ReLU_2`, 2,560 channels), summarized by an equal-weight composite (*InterpScore*). Three evidences motivate this design: better separation between neurons, a human-free variant aligns more strongly with human recognizability than selectivity alone ($\Delta R^2 \approx 0.20$), and values/rankings are numerically stable under standard seeds/perturbations. *Granularity.* The neuron level is a deliberate choice: interventions are local and well-posed; (S, C, R, H) are portable across layers/architectures; and unit-level measures provide a practical substrate for circuit/feature analyses without presupposing monosemanticity. Attention-path reconstruction is used as validation (parity within a small tolerance, e.g., $\leq 10^{-6}$), and Microscope-style visualization supports systematic inspection.

5.2 Practical Applications

The protocols drop into existing interpretability workflows. In research settings, they enable principled triage, separating neurons whose high selectivity masks weak causal or human alignment from units with balanced evidence, while reporting dispersion (e.g., CV across neurons) to detect ceiling effects. For safety-oriented uses, the per-axis tuple (S, C, R, H) acts as a compact checklist before relying on a unit: selective, behavior-relevant, robust to benign variation, and human-recognizable. Equal weights provide a neutral baseline; application-specific weights can be pre-registered as sensitivity analyses.

5.3 Limitations and Future Directions

This study is intentionally scoped: one architecture (CLIP RN50x4), one intervention site (`image_block_4/5/ReLU_2`), and ten high-selectivity neurons. Equal weighting in *InterpScore* is a pragmatic default. Future work broadens scale (larger neuron sets, additional layers), tests cross-architecture generality (e.g., ViT-based CLIP), develops more efficient causal measures with the locality and clarity of C, and explores semi-automated supplements to H that preserve human grounding. At larger scales, unit-level results can be composed into circuit/feature evaluations, using these neuron-level protocols as the empirical base.

Chapter 6

Conclusion

This paper asks whether a compact, multi-axis evaluation can move neuron-level mechanistic interpretability beyond anecdotes toward an objective, discriminative, and reproducible basis. On CLIP RN50x4 at `image_block_4/5/ReLU_2` (2,560 channels), the answer is affirmative. Relative to activation selectivity alone, the four-axis evaluation: Selectivity (S), Causality (C), Robustness (R), and Human Consistency (H), and its equal-weight composite (*InterpScore*) (i) increase discrimination among neurons, (ii) align more closely with human recognizability when evaluated without the human term (gain $\Delta R^2 \approx 0.20$ over $H \sim S$), and (iii) remain numerically and rank-wise stable under standard seeds and benign perturbations.

Practically, the result is a simple procedure that surfaces differences that a single metric obscures and makes "what is more interpretable" an empirical question. *InterpScore* should be read as an operational summary rather than a label: reporting the per-axis tuple (S, C, R, H) alongside the composite (with intervals) and the dispersion across neurons (e.g., CV) preserves diagnostic value. In our case studies, this separation explains why highly selective units can lack functional or human relevance, while moderately selective units can rise on the strength of robustness and recognizability.

The choice to work at the neuron level is deliberate. Interventions are local and well-posed at this granularity; the axes (S, C, R, H) are portable across layers and architectures; and unit-level measurements provide a practical substrate for building circuit- and feature-level claims. Implementation checks rely on numerical parity within tolerance (e.g., $\leq 10^{-6}$), and we avoid ad-hoc discretization throughout. DeepDream stimuli are included within H to probe recognizability, while R is reserved for semantically preserving transforms of natural images.

Looking forward, the immediate priorities are scale and breadth: larger neuron sets, additional layers, and cross-architecture replications (e.g., ViT-based CLIP and language models). Methodologically, there is room for more efficient causal measures with the same locality and clarity as C, and for semi-automated supplements to H that retain human grounding. We view this work as establishing

a minimal, portable unit of measurement for interpretability: once unit-level properties are gauged consistently, claims at the circuit and feature levels can rest on clearer empirical foundations.

Bibliography

[1] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. "Network dissection: Quantifying interpretability of deep visual representations". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6541–6549.

[2] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning". In: *Transformer Circuits Thread* (2023). https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[3] Nicholas Carlini and David Wagner. "Towards Evaluating the Robustness of Neural Networks". In: *IEEE S&P*. 2017.

[4] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. "Activation Atlas". In: *Distill* 4.3 (2019), e15.

[5] Jacob Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46.

[6] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Lawrence Erlbaum, 1988.

[7] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. CRC Press, 1994.

[8] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Visualizing Higher-Layer Features of a Deep Network". In: *University of Montreal* (2009). Tech. Rep. 1341.

[9] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. "Towards Automatic Concept-based Explanations". In: *NeurIPS*. 2019.

[10] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. "Multimodal neurons in artificial neural networks". In: *Distill* 6.3 (2021), e30.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *ICLR*. 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition”. In: *CVPR*. 2016.

[13] Larry V Hedges and Ingram Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, 1985.

[14] Been Kim et al. “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)”. In: *ICML*. 2018.

[15] Klaus Krippendorff. “Computing Krippendorff’s Alpha Reliability”. In: *Departmental Papers (ASC), University of Pennsylvania* (2011).

[16] Aleksander Madry et al. “Towards Deep Learning Models Resistant to Adversarial Attacks”. In: *ICLR*. 2018.

[17] Aravindh Mahendran and Andrea Vedaldi. “Understanding Deep Image Representations by Inverting Them”. In: *CVPR*. 2015.

[18] Aleksandar Makelov, George Lange, and Neel Nanda. “Towards principled evaluations of sparse autoencoders for interpretability and control”. In: *arXiv preprint arXiv:2405.08366* (2024).

[19] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. “Inceptionism: Going deeper into neural networks”. In: *Google research blog* 17.6 (2015), pp. 1–10.

[20] Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, et al. “Mib: A mechanistic interpretability benchmark”. In: *arXiv preprint arXiv:2504.13151* (2025).

[21] Anh Nguyen, Jason Yosinski, and Jeff Clune. “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images”. In: *CVPR*. 2015.

[22] Tuomas Oikarinen, Ge Yan, and Tsui-Wei Weng. “Evaluating neuron explanations: A unified framework with sanity checks”. In: *arXiv preprint arXiv:2506.05774* (2025).

[23] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. “Zoom in: An introduction to circuits”. In: *Distill* 5.3 (2020), e00024–001.

[24] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. “Feature visualization”. In: *Distill* 2.11 (2017), e7.

[25] Vitali Petsiuk, Abir Das, and Kate Saenko. “RISE: Randomized Input Sampling for Explanation of Black-box Models”. In: *BMVC*. 2018.

[26] R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. “Invariant visual representation by single neurons in the human brain”. In: *Nature* 435.7045 (2005), pp. 1102–1107.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[28] Ramprasaath Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. In: *ICCV*. 2017.

[29] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. “Learning Important Features Through Propagating Activation Differences”. In: *ICML*. 2017.

[30] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic Attribution for Deep Networks”. In: *ICML*. 2017.

[31] Ashish Vaswani et al. “Attention Is All You Need”. In: *NeurIPS*. 2017.

[32] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. *Understanding Neural Networks Through Deep Visualization*. 2015. arXiv: 1506.06579.

Appendix A

METHODOLOGY SUPPLEMENTS

A.1 Statistical Methods and Validation

A.1.1 Bootstrap Procedures

Bootstrap analysis with 1000 resamples (following the bootstrap framework of Efron and Efron–Tibshirani [7]) establishes high precision for all measurements (standard errors ≤ 0.007) Fig. A.1, with particularly tight intervals for the InterpScore enabling reliable interpretability assessment. Confidence intervals are computed using scipy.stats.bootstrap with bias-corrected percentile method. Standard errors are calculated consistently from the bootstrap distribution to avoid double-sampling artifacts.

This follows the bootstrap framework of Efron and Efron–Tibshirani [7].

A.1.2 Effect Size Calculations

Cohen's d computed for all pairwise component comparisons using paired-samples formula (mean difference divided by standard deviation of differences) to account for within-subject design. P-values corrected for multiple comparisons using Benjamini-Hochberg false discovery rate procedure. Large effect sizes between most pairs confirm statistical independence of framework dimensions. Fig. A.2.

Figure A.1: **Selectivity exhibits minimal uncertainty while other components demonstrate substantial variability.** Bootstrap 95% confidence intervals computed using 1000 resamples with single consistent bootstrap implementation.

A.1.3 Power Analysis

We conducted a two-panel power analysis using corrected statistical methods to evaluate study sensitivity and observed effect magnitudes. Panel (a) shows prospective power curves calculated with proper two-tailed test formulas for paired t-tests, illustrating detection capability across sample sizes. Our study with $n = 10$ neurons (highlighted in orange) achieves 80% power to detect large effects ($d \geq 0.8$) and moderate power for medium effects ($d \approx 0.5$).

Panel (b) presents observed effect sizes between component pairs with 95% bootstrap confidence intervals, using paired Cohen's d formula (mean difference divided by standard deviation of differences) to avoid post-hoc power calculation pitfalls. The largest effect was observed between Selectivity and Causality ($d = 11.34$), followed by Selectivity-Robustness ($d = 5.58$). Selectivity-Human Consistency showed a moderate effect ($d = 1.64$), while comparisons among Causality, Robustness, and Human Consistency yielded smaller effects ranging from $d = 0.32$ to $d = 1.40$.

All effect sizes involving Selectivity exceed Cohen's large effect threshold ($d \geq 0.8$), while non-Selectivity comparisons fall below this threshold, confirming that selectivity captures fundamentally different interpretability aspects and supporting our multi-dimensional framework's necessity. Fig. A.3

A.1.4 Inter-Component Correlation Analysis

Pearson correlation coefficients computed between all component pairs reveal weak inter-component relationships, supporting framework independence. P-values calculated using t-distribution with

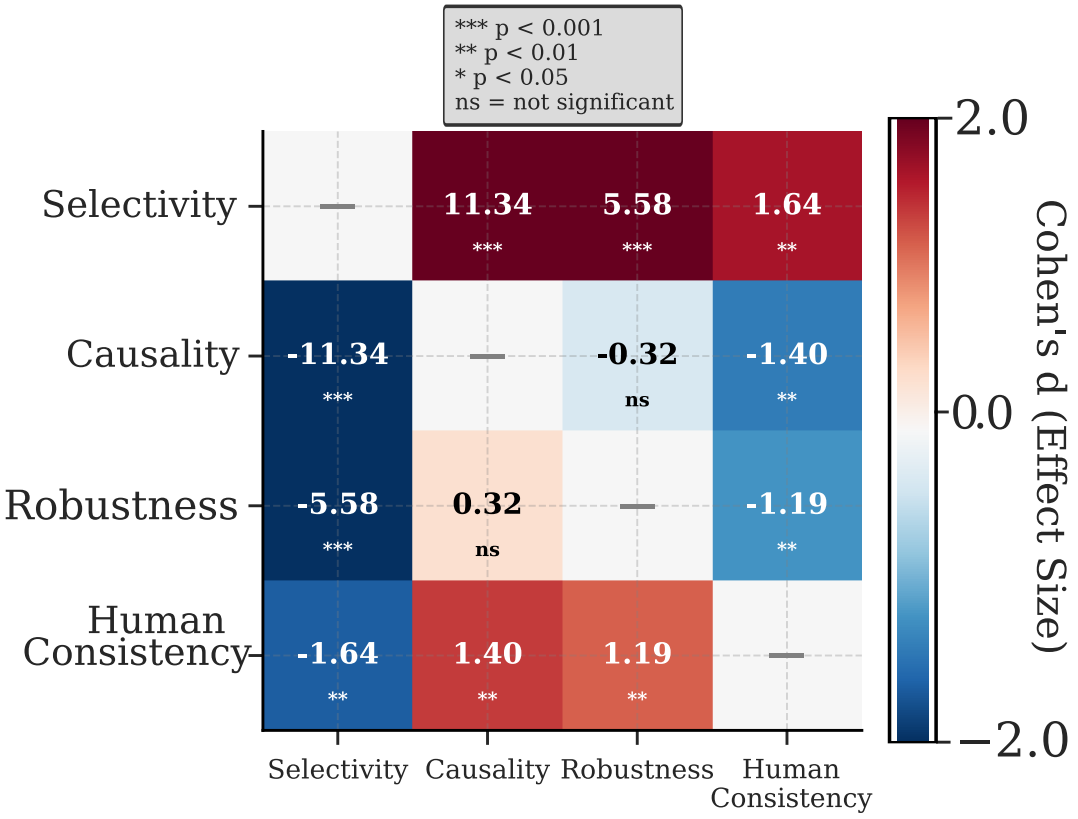

Figure A.2: **Selectivity differs dramatically from other components with large effect sizes.** Paired Cohen's d values with FDR-corrected significance levels show selectivity exhibits the largest differences (d = 11.34 vs causality, d = 5.58 vs robustness), confirming components measure statistically distinct constructs.

degrees of freedom correction, though significance testing is omitted from visualization due to small sample size limitations. Figure A.4 presents the complete correlation structure.

Selectivity shows minimal correlations with other dimensions: $r = 0.106$ with Causality, $r = 0.044$ with Robustness, and $r = -0.014$ with Human Consistency. Among non-selectivity components, correlations remain weak to moderate: Causality-Robustness ($r = -0.130$), Causality-Human Consistency ($r = -0.196$), and Robustness-Human Consistency ($r = 0.033$). All correlations fall below $|r| = 0.2$, indicating minimal shared variance between components and confirming that each dimension captures distinct aspects of neural interpretability.

A.1.5 Framework Dimensionality Analysis

Dimensionality Scaling

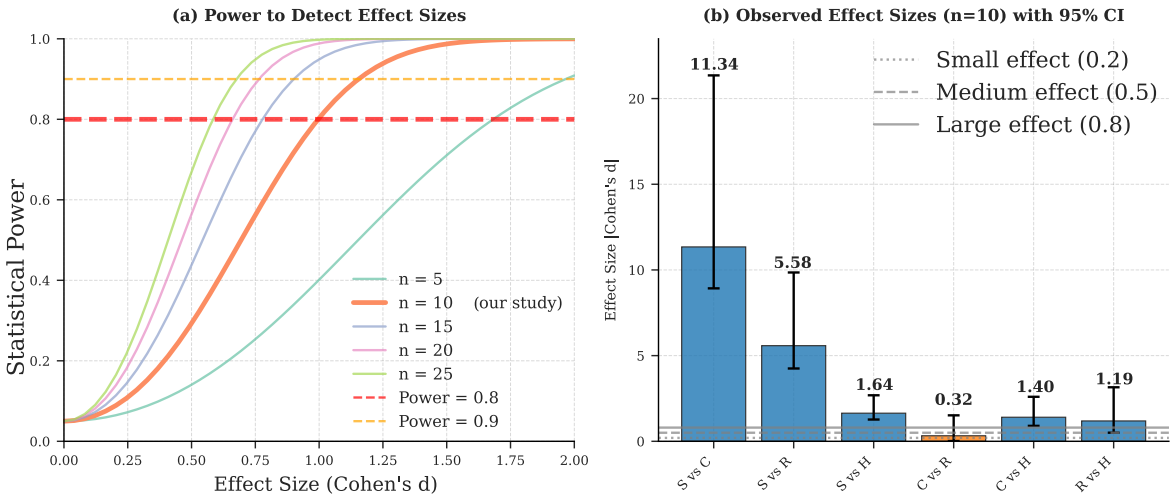

Figure A.3: **Selectivity comparisons yield extremely large effect sizes while other component pairs show smaller differences.** (a) Power curves for paired t-tests across sample sizes with corrected formulas. (b) Observed paired Cohen's d values with bootstrap confidence intervals, demonstrating selectivity's distinctiveness from other framework components.

Framework performance scales systematically with dimensionality: single-component assessment shows poor correlation with comprehensive evaluation (median $r = 0.5$), two-dimensional combinations achieve good correlation (median$r = 0.85$), while three-dimensional subsets reach strong correlation threshold ($r > 0.9$).

Minimum Subset Analysis

The optimal three-dimensional combination (Selectivity + Robustness + Human Consistency, $r = 1.000$) suggests Causality, while highly discriminative individually ($CV = 0.589$), introduces complexity that may not always improve overall assessment.

A.2 Human Evaluation Protocol

A.2.1 Participant Demographics and Recruitment

Sample Size: 110 participants recruited via Prolific Academic platform **Demographics:**

- Age: 18-65 years (M = 32.4, SD = 8.7)
- Gender: 52% female, 47% male, 1% other/prefer not to say

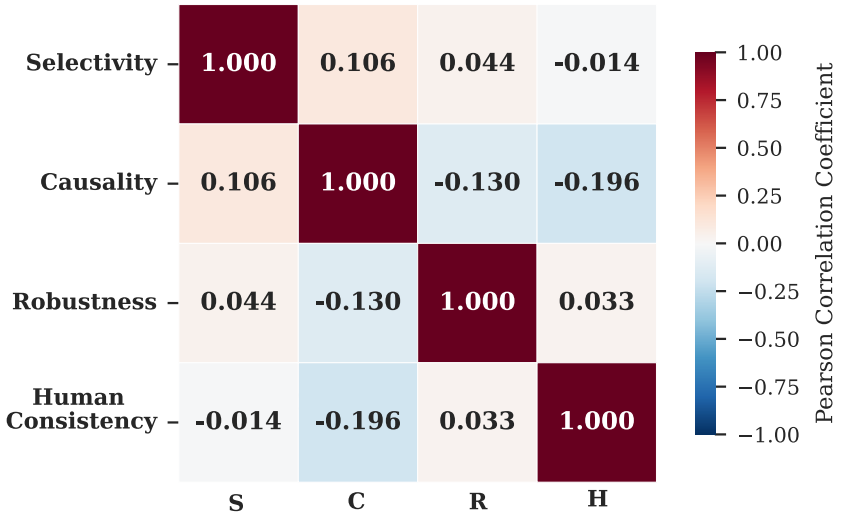

Figure A.4: **Framework components demonstrate weak inter-correlations supporting dimensional independence.** Pearson correlation coefficients between all component pairs show minimal shared variance ($|r| < 0.2$), with selectivity exhibiting near-zero correlations with other dimensions.

- Education: 78% college-educated, 15% graduate degree, 7% high school
- Geography: 62% UK, 23% US, 15% other English-speaking countries
- Vision: 100% normal or corrected-to-normal (self-reported)

Inclusion Criteria:

- English speaker
- Age 18-65 years
- Normal or corrected vision
- Prolific approval rate > 95%
- Previous study completion rate > 90%

A.2.2 Experimental Interface Design

After a warmup phase, participants viewed images in randomized order and answered: "Does this image contain [CONCEPT]?" with two open answer boxes where to insert text. Interface features:

- **Image presentation:** 512x512 pixels, 3-second minimum viewing time
- **Response recording:** Two registered open text responses

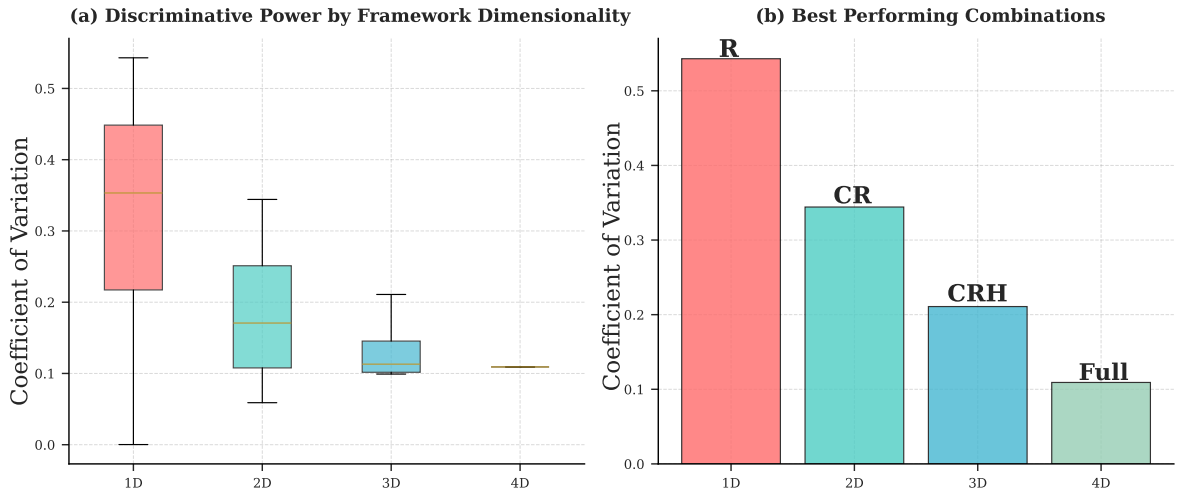

Figure A.5: Framework dimensionality analysis showing: (a) discriminative power by framework dimensionality and (b) best performing component combinations.

A.2.3 Quality Control Implementation (QC Ptotocol)

Attention Checks:

- Obvious positive cases (e.g., clear fire images for fire concept)
- Obvious negative cases (e.g., clear puppies images for fire concept)
- Expected accuracy > 95%, participants < 80% excluded
- Result: no participants excluded
- Checks on the inputs: text was real-time checked to not be the same in both answers, to be at least 3 characters long, to not have repetitions of characters, and to be all upper case.

Response Time Analysis:

- Median response time: 10.2 seconds per image
- Responses < 0.5s flagged as too fast (0.0% of trials)
- Responses > 30s flagged as attention lapses (1.8% of trials)
- Flagged responses excluded from analysis

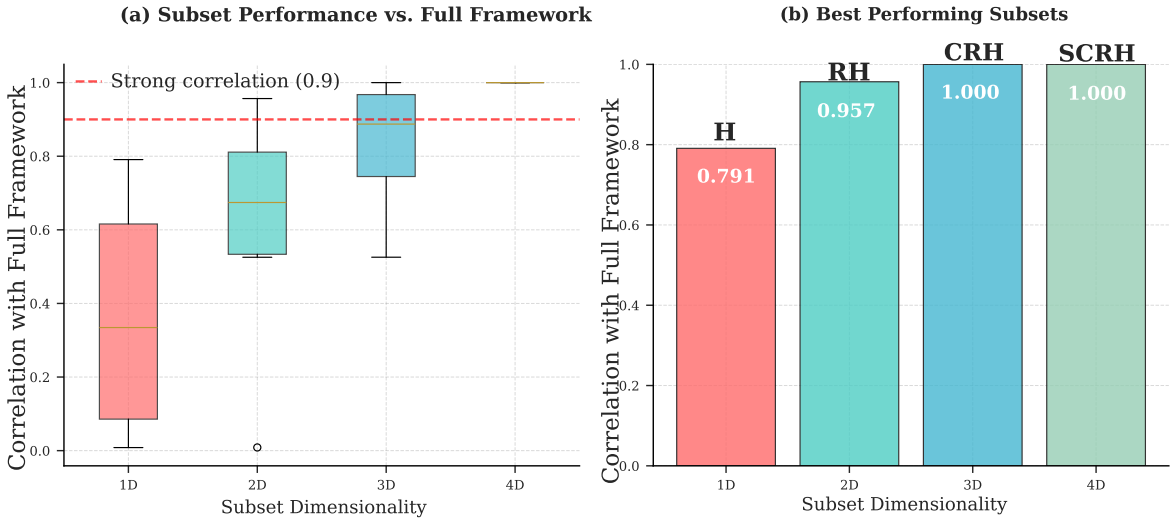

Figure A.6: Minimum subset analysis for reliable interpretability assessment showing correlation with full framework by subset dimensionality. Results establish three dimensions as minimum viable framework.

A.2.4 Hard vs. Soft Accuracy Metrics

We computed two distinct accuracy metrics to capture different aspects of participant performance in the image recognition task:

Hard Accuracy represents exact string matching between participant responses and ground truth labels. A response is considered hard correct only if it contains an exact lexical match to the ground truth concept (case-insensitive). For example, if the ground truth is “dog”, only responses containing exactly “dog” would be marked as hard correct.

Soft Accuracy employs a more lenient evaluation that accounts for semantic similarity and common variations in responses. This metric considers responses correct if they meet any of the following criteria:

- Exact match (similarity = 1.0)
- Partial containment between response and ground truth (similarity = 0.95)
- Synonym matching using a predefined dictionary of common concept variations (similarity = 0.9)
- Sequence-based string similarity above a threshold of 0.7 using the Ratcliff-Obershelp algorithm

The soft accuracy metric handles several common response variations that would be penalized under hard accuracy:

- Multi-word responses (e.g., “donald trump” vs. “trump”)
- Plural forms (e.g., “hands” vs. “hand”)
- Synonymous terms (e.g., “automobile” vs. “car”)
- Comma-separated multiple responses (e.g., “bed,bedroom” vs. “bed”)
- Minor spelling variations and typos

Focus on Soft Accuracy: We primarily report soft accuracy results because this metric provides a more ecologically valid assessment of participant understanding. In real-world image recognition tasks, multiple valid labels often exist for the same visual concept, and exact string matching fails to capture semantically correct responses that use alternative but equivalent terminology. Soft accuracy better reflects whether participants successfully identified the core concept in the image, regardless of minor linguistic variations in their response formulation.

A.2.5 Inter-Rater Agreement and Corruption Level Analysis

Inter-rater agreement varied substantially across corruption levels, with highest agreement for adversarial attacks (L3: 0.801) and clean images (L5: 0.790), while DeepDream generated images (L4) showed notably low agreement (0.209). Soft correct scores consistently exceeded hard correct scores across all levels, with clean images achieving the highest accuracy (hard: 0.9, soft: 0.95) and progressive degradation toward more corrupted levels.

For context, inter-rater reliability in similar annotation settings is commonly summarized with coefficients such as Cohen’s κ and Krippendorff’s α [5, 15].

A.2.6 Trump Neuron Analysis

The Trump neuron (Neuron 89) demonstrates how multi-dimensional evaluation reveals interpretability characteristics beyond statistical selectivity alone.

The neuron exhibits perfect statistical selectivity (S = 1.000) with strong separation between Trump and non-Trump images (Cohen’s d = 8.36). Human recognition remains stable across naturalistic corruptions (64-66% for L1-L3) but drops to zero at L4, which corresponds to DeepDream-generated synthetic images. This complete recognition failure occurs because DeepDream optimization creates images that maximally activate the neuron through low-level visual patterns that appear as abstract, psychedelic imagery rather than recognizable Trump-related content.

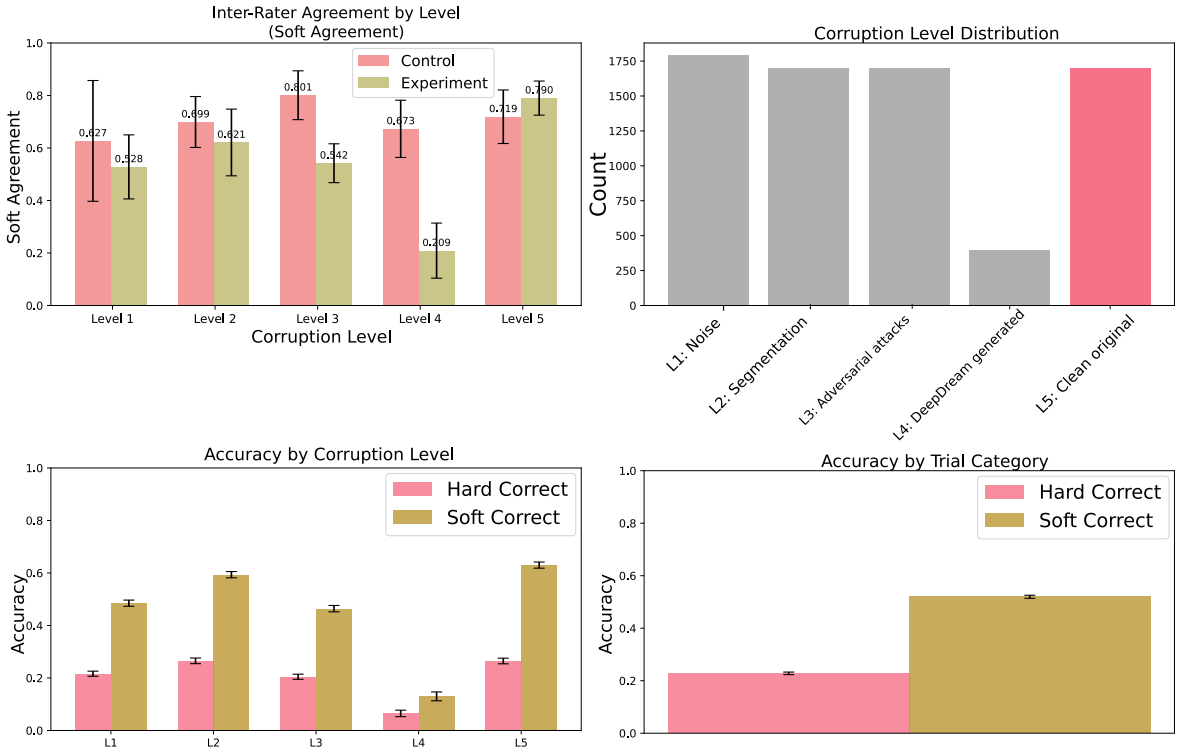

Figure A.7: Inter-rater agreement and accuracy analysis across corruption levels. **Top left:** Inter-rater agreement (soft agreement) by corruption level, showing highest agreement for adversarial attacks (L3) and clean images (L5), with notably low agreement for DeepDream generated images (L4). **Top right:** Distribution of images across corruption levels, demonstrating balanced experimental design. **Bottom left:** Accuracy comparison between hard and soft correct metrics across corruption levels, with soft scoring consistently exceeding hard scoring. **Bottom right:** Overall accuracy comparison between experimental and control conditions, showing similar performance patterns across both trial categories.

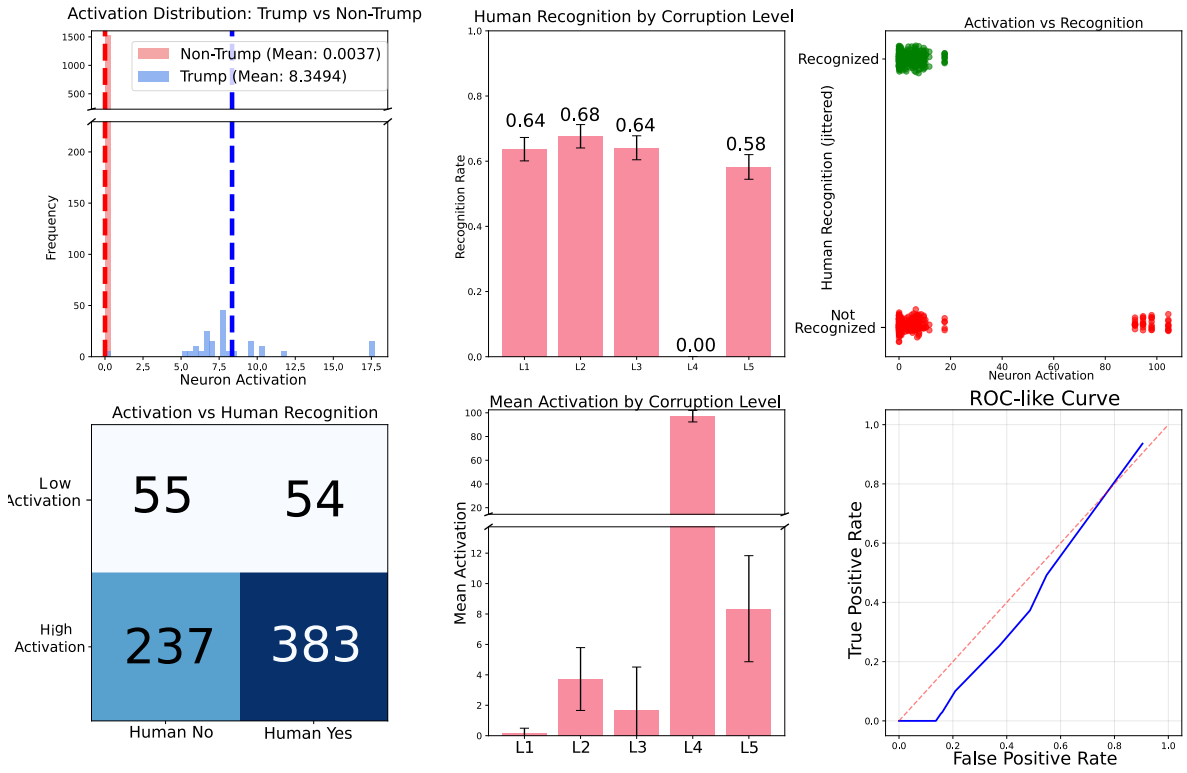

Figure A.8: Multi-dimensional analysis of Trump neuron (Neuron 89). Top row: (left) Activation distribution comparing Trump vs. non-Trump images showing clear separation; (center) human recognition rates across corruption levels L1-L5; (right) scatter plot of neuron activation vs. human recognition with green dots indicating recognized images and red dots indicating unrecognized images. Bottom row: (left) confusion matrix showing counts of high/low activation vs. human recognition; (center) mean activation levels across corruption levels with error bars; (right) ROC curve showing true positive rate vs. false positive rate across activation thresholds.

The activation vs. recognition scatter plot reveals two distinct clusters: green dots (recognized images) at moderate activation levels, and a prominent cloud of red dots (unrecognized images) at high activation values in the top-right. These red dots represent the DeepDream synthetic images, they achieve the highest neuron activations but remain completely unrecognizable to humans, illustrating the disconnect between optimal neuron stimulation and semantic interpretability.

Appendix B

TECHNICAL IMPLEMENTATION

B.1 Complete Graph Surgery

Overview We intervene at a single neuron in the CLIP RN50x4 image encoder and forward the exact downstream path to measure embedding changes. This section specifies the architecture context, intervention operators, and validation.

B.1.1 CLIP Architecture Context

Hierarchy (RN50x4).

- Input: 288×288×3.
- ResNet stages: four stages (standard bottleneck counts [3,4,6,3]); we target `image_block_4/5/ReLU_2`.
- Activation at site: $(1,2,560,9,9)$.
- Attention pooling: flatten to $81\times2,560$, learned positional encodings, multi-head attention (40 heads × 64 dim), projection to a 640-d image embedding.

B.1.2 Intervention and Causality Measure

Let $A \in \mathbb{R}^{C\times H\times W}$ be the activation tensor at the site and $E(x) \in \mathbb{R}^{640}$ the baseline embedding.

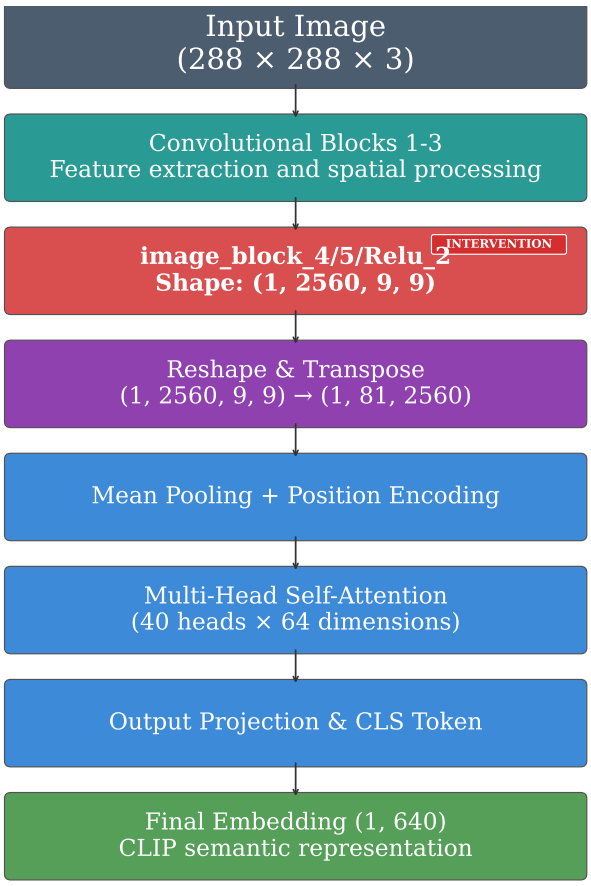

Figure B.1: CLIP RN50x4 image encoder and our intervention point `image_block_4/5/ReLU_2`. The site lies just upstream of the attention pooling head, enabling precise neuron manipulation while preserving all downstream computation.

Interventions (single index n).

$$\text{ablation:} \quad A'[n,:,:] \leftarrow 0 \tag{B.1}$$

$$\text{amplification:} \quad A'[n,:,:] \leftarrow 2\,A[n,:,:] \tag{B.2}$$

Embedding shift (per image x).

$$R_{\text{int}}(x) \;=\; \frac{\|E_{\text{int}}(x) - E(x)\|_2}{\|E(x)\|_2} \tag{B.3}$$

Causality (per neuron N, concept set X).

$$C(N,X) \;=\; \tfrac{1}{2}\,\mathbb{E}_{x\in X}\big[\,R_{\text{abl}}(x) + R_{\text{amp}}(x)\,\big] \tag{B.4}$$

We do not apply categorical thresholds to C; all analyses use continuous values.

B.1.3 Validation (parity within tolerance)

Table B.1: Parity checks with hooks installed (no-op) and after surgery. Parity is defined as agreement with the original forward pass within numeric tolerance on a held-out set (≥1k images).

Test	Metric	Result	Tolerance
Embedding parity (no-op)	$\max\|E' - E\|$	$\leq 1\times10^{-6}$	absolute
Embedding agreement (no-op)	Pearson r	≥ 0.9999	-
Exact targeting	max change on $m \neq n$	$\leq 1\times10^{-7}$	absolute
Determinism	run-to-run hash match	pass	identical seeds

B.2 Microscope-Style Neuron Browser (Anonymized)

We re-implement and extend the Microscope concept [10] for RN50x4 to support layer `image_block_4/5/ReLU_2` (2,560 neurons), providing: (i) top-k activating natural images (ImageNet; k=100 per neuron), (ii) synthetic feature visualizations (one per neuron), (iii) spatial activation heatmaps (9×9), and (iv) basic statistics (activation distributions, top classes). An anonymized demo and dataset handles are provided in the supplementary repository.[1]

B.3 DeepDream: Maximally Activating Synthesis

Objective and update. For target feature f at layer l,

$$J(a) = \sum_{m,n} z^l_{f,m,n}(a,\theta), \qquad g = \nabla_a J(a), \qquad a_{t+1} = a_t + \eta \frac{g}{\sqrt{\mathbb{E}[g^2]} + \epsilon} \tag{B.5}$$

with step size η, small ϵ (e.g., 10^{-8}), and gradient normalization for stability [19].

Multi-octave schedule. We use K=4 octaves from a base 72×72 to the CLIP input 288×288. Let $s = (288/72)^{1/(K-1)} = 4^{1/3} \approx 1.587$ and $h_i = w_i = \text{round}(72\, s^i)$, yielding $\{72, 114, 181, 288\}$. Each octave runs 2,000 iterations with $\eta \approx 2.0$; the detail image is upsampled and added to the next octave.

[1] Links redacted for double-blind review.

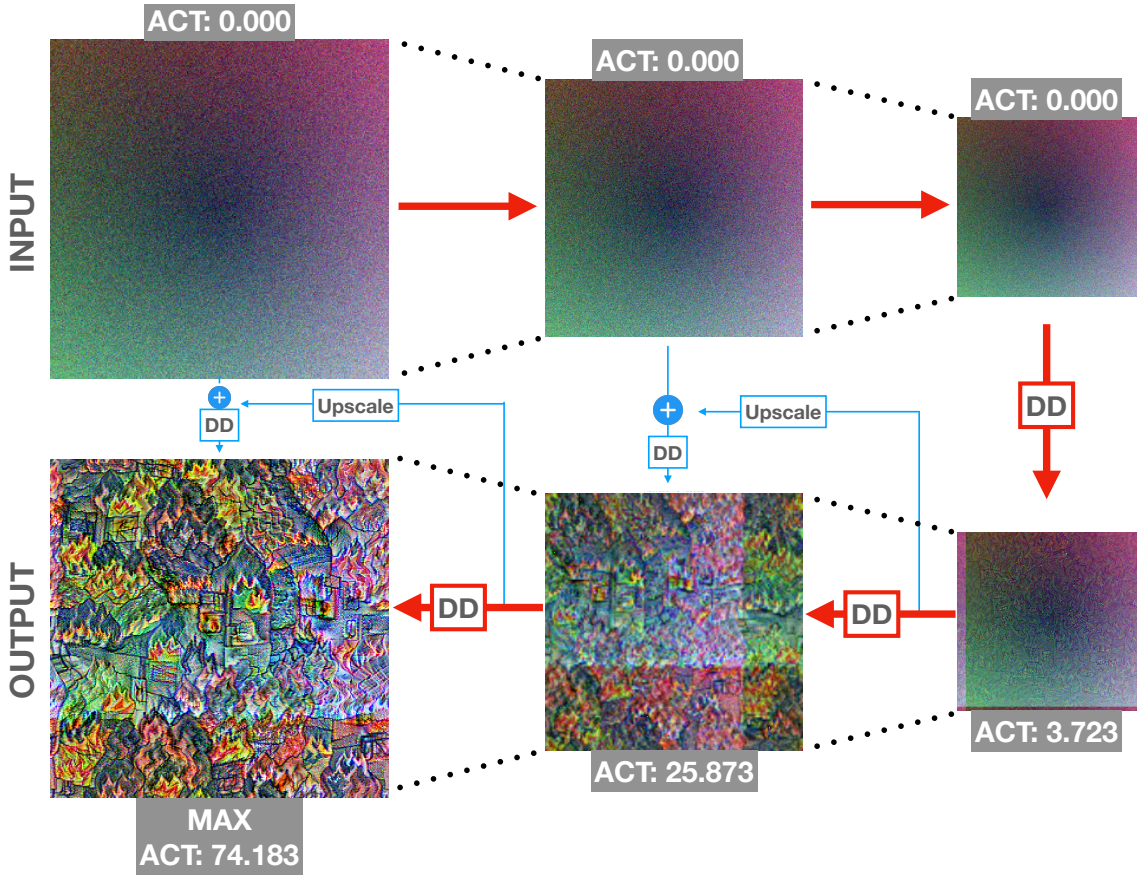

Figure B.2: DeepDream pipeline for a target neuron (e.g., "Fire", ID 297). Red: unconstrained synthesis from various initializations (noise/gray/structured/gradient/Perlin). Blue: constrained (image-conditioned) variant, not used in our main study. This is an actual example from our experiments. *Can you recognize the maximally activating image as containing FIRE?*

Initializations. We test five initializations:

$$\text{Gray: } a_0(x,y,c) = 128 \tag{B.6}$$

$$\text{Uniform noise: } a_0 \sim \mathcal{U}(0,255) \tag{B.7}$$

$$\text{Structured noise: } a_0 = 128 + \sum_{s\in\{4,8,16,32\}} \text{resize}\left(\mathcal{N}(0,30^2), 288\times 288\right) \tag{B.8}$$

$$\text{Gradients: } a_0 = \alpha\,\text{radial}(x,y) + \beta\,\text{linear}_x(x,y) \tag{B.9}$$

$$\text{Perlin-like: } a_0 = 128 + \sum_{o=0}^{3} \frac{50}{o+1} \sin\left(\frac{2\pi 2^o x}{288} + \phi_o\right) \tag{B.10}$$

Regularization. Every 4 steps we clip to $[0, 255]$ and apply random integer shifts $\Delta x, \Delta y \sim \mathcal{U}[-4, 4]$ ("jitter") before back-shifting; this reduces high-frequency artifacts.

Role in our framework. DeepDream acts as a diagnostic for *Human Consistency* (H): if maximally activating synthetic images for a neuron are not recognized by humans as containing the intended concept, we discount that neuron's interpretability signal accordingly.

B.4 Lucid Feature Visualizations

Lucid [24] produces more human-interpretable feature images via diversity objectives, TV/transform regularizers, and preconditioned gradients. We include one Lucid visualization per neuron to complement DeepDream: Lucid favors interpretability (often at lower absolute activation), while DeepDream probes whether maximal activation itself corresponds to human-recognizable content.

Appendix C

Reproducibility

To ensure full reproducibility of our work, we provide open access to all code, data, and implementations used in this study. The complete codebase is organized across two GitHub repositories:

Results, Metrics, and Visualizations: All experimental code, metric implementations, and paper visualizations are available at: `https://github.com/ernestoBocini/Interpretability-Score.git`

Microscope Visualization Tool: The interactive microscope visualization tool for exploring model interpretability is available at: `https://github.com/ernestoBocini/rebuilt-microscope-CLIP`

These repositories contain detailed documentation, installation instructions, and example usage to facilitate replication of our findings and enable further research in this area.

Appendix D

Ethics

D.1 Human Participants

Our human evaluation involved 110 participants recruited through Prolific Academic. Participants were compensated at £7.50/hour and provided informed consent before participating. All data was collected anonymously with no personally identifiable information retained.

The experimental design included quality control measures to prevent participant fatigue, with minimum viewing times and response time monitoring. No participants were excluded based on attention checks, suggesting the task was appropriately designed for the participant population.

D.2 Model and Data Considerations

We use OpenAI's publicly available CLIP RN50x4 model, which was trained on internet-scraped data. This model likely contains biases present in its training data, as evidenced by our analysis of neurons like the "Trump" detector. Our interpretability framework reveals these biases rather than creating them.

For concept evaluation, we manually curated image sets, which introduces potential researcher bias in what we consider "representative" examples of each concept.

D.3 Responsible Applications

This framework is designed to improve AI safety through better interpretability assessment. More systematic evaluation of model components could help identify problematic behaviors and inform safer AI development.

We acknowledge that detailed knowledge of model internals could potentially inform adversarial attacks, but believe the benefits of interpretability research for AI safety outweigh these risks, particularly given the extensive existing literature on model internals.