Universitat d'Alacant
Universidad de Alicante

ellis | UNIT ALICANTE

From Interpretability Methods to Interpretable Models

Julien Colin

Thesis presented in fulfillment of the requirements
for the degree of Doctor of Philosophy by the

UNIVERSITY OF ALICANTE

With international mention

DOCTOR OF INFORMATICS

Advised by:
Nuria Oliver, *ELLIS Alicante*
Miguel Ángel Lozano Ortega, *University of Alicante*

The research presented in this thesis has been financed by the ELLIS unit Alicante Foundation with funding from a nominal grant from the Regional Government of Valencia in Spain (Convenio Singular signed with Generalitat Valenciana, Conselleria de Industria, Turismo, Innovación y Comercio, Dirección General de Innovación), from a grant by the Banco Sabadell Foundation, and from Intel via RESUMAIS, the Center of Scientific Excellence in Responsible AI. Views and opinions expressed are however those of the author(s).

Universitat d'Alacant
Universidad de Alicante

ellis | UNIT ALICANTE

From Interpretability Methods to Interpretable Models

Julien Colin

Tesis presentada para aspirar al título de doctor por la

UNIVERSIDAD DE ALICANTE

Mención de doctor internacional

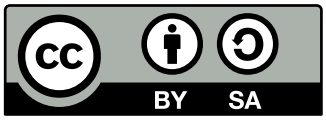
DOCTORADO EN INFORMÁTICA

Dirigida por:
Nuria Oliver, *ELLIS Alicante*
Miguel Ángel Lozano Ortega, *University of Alicante*

La investigación presentada en esta tesis ha sido financiada por la Fundación de la Comunitat Valenciana Unidad ELLIS Alicante con fondos de una subvención nominativa por parte del gobierno de la Comunitat Valenciana en España (Convenio firmado con la Generalitat Valenciana, Conselleria de Industria, Turismo, Innovación y Comercio, Dirección General de Innovación), de la fundación Banco Sabadell y de Intel, en virtud de RESUMAIS, el Centro de Excelencia en Inteligencia Artificial Responsable. Las opiniones expresadas son exclusivas del autor(es).

This document was proudly typeset with LaTeX.

This work is licensed under a Creative Commons "Attribution-ShareAlike 4.0 International" license.



- Licensees may copy, distribute, display and perform the work and make derivative works and remixes based on it only if they give the author or licensor the credits (attribution) in the manner specified by these.
- Licensees may distribute derivative works only under a license identical ("not more restrictive") to the license that governs the original work. (See also copyleft.) Without share-alike, derivative works might be sublicensed with compatible but more restrictive license clauses, *e.g.*, CC BY to CC BY-NC.)

Please see creativecommons.org/licenses/by-sa/4.0/ for greater detail.

Contact Details

Julien Colin
julien@ellisalicante.org, colinjulienlouis@gmail.com

The artist must have something to say, for mastery over form is not his goal but rather the adapting of form to its inner meaning.

—Wassily Kandinsky

Abstract

Deep learning models are increasingly used to make consequential decisions, yet remain largely opaque to the people who rely on them. Over the past decade, explainable AI (XAI) has developed a mature toolbox for generating explanations, but that toolbox has mostly been turned inward: it has been used primarily to compare explanation methods against each other rather than to ask whether the resulting models are actually understandable. This thesis redirects the focus of XAI from evaluating explanations to evaluating the models those explanations are intended to illuminate. It develops a framework for *measuring* and *comparing* the interpretability of vision models, and for *causally testing* whether that interpretability can be deliberately shaped. The framework is built across three chapters, which address three fundamental questions: what representational basis should interpretability be measured over; how can human understanding of models be measured and compared across today's vision foundation models; and can interpretability be improved through targeted interventions rather than simply observed?

The **first chapter** establishes the appropriate representational basis for interpretability. Individual neurons often respond to several unrelated visual patterns, making them an unreliable unit of explanation. Dictionary learning has been proposed as an alternative, but its success has largely been evaluated using technical criteria rather than the human judgments that determine whether a representation is genuinely interpretable. This chapter closes that gap through three large-scale psychophysics experiments (N=481) across two convolutional networks. Dictionary-based features are consistently more interpretable than neurons, with the advantage increasing in deeper layers. Moreover, the choice of representational basis itself can determine whether differences between models are detected at all.

Building on this foundation, the **second chapter** turns to the problem of measurement. Existing protocols for comparing model interpretability rely on forced-choice tasks that allow participants to succeed through elimination rather than genuine understanding, making fair cross-model comparisons difficult. This chapter introduces two behavioral protocols, localizability and semanticity, that evaluate interpretation directly and place models on a common, chance-corrected scale. Applied to six vision transformers, including four foundation models (DINOv2, DINOv3, CLIP, SigLIP), and based on more than 15,000 human responses, the protocols reveal that foundation models are consistently less interpretable than their supervised counterpart. This gap is explained neither by model scale nor by downstream task performance. Instead, feature locality and alignment with human semantic judgments emerge as the strongest predictors of interpretability.

The **third chapter** examines whether these predictive factors can be turned into causal levers for interpretability. Five vision backbones are LoRA-finetuned toward human perceptual judgments and the resulting aligned models are evaluated using the same behavioral measures developed in Chapter 4. The results show that alignment is an unreliable lever

for improving interpretability: its effect is positive across the three vision-only backbones, reaching significance only for DINOv2, and reverses for the two vision-language models. Even when alignment improves interpretability, the underlying mechanism remains inconsistent and only partially understood. Thus, while alignment can serve as a lever for improving interpretability, it does not provide a reliable or universally applicable solution, and the reasons it succeeds for some models while failing for others remain unresolved.

These contributions provide the first behaviorally grounded account of how interpretable current vision foundation models are, why they differ, and how far their interpretability can be deliberately improved. More broadly, this thesis argues that progress in explainable AI requires moving beyond the evaluation of explanations alone and toward the systematic study of whether models themselves become understandable to humans. The framework developed here provides a foundation for that shift, and points toward future models whose representations—and ultimately whose decisions—can be meaningfully understood by the people who depend on them.

Resumen

Los modelos de aprendizaje profundo se utilizan cada vez más para tomar decisiones con consecuencias importantes, pero siguen siendo en gran medida opacos para las personas que dependen de ellos. Durante la última década, la inteligencia artificial explicable (XAI, por sus siglas en inglés) ha desarrollado un conjunto de herramientas para generar explicaciones. Sin embargo, estas herramientas se han utilizado principalmente de forma introspectiva: sobre todo para comparar distintos métodos de explicación entre sí, en lugar de preguntarse si los modelos resultantes son realmente comprensibles. Esta tesis reorienta el foco de la XAI desde la evaluación de las explicaciones hacia la evaluación de los modelos que dichas explicaciones pretenden esclarecer. Para ello, desarrolla un marco destinado a *medir* y *comparar* la interpretabilidad de modelos de visión, así como a *evaluar causalmente* si dicha interpretabilidad puede modificarse de manera deliberada. Este marco se desarrolla a lo largo de tres capítulos, que abordan tres preguntas fundamentales: sobre qué base representacional debe medirse la interpretabilidad; cómo puede medirse la comprensión humana de los modelos y compararse entre los actuales modelos fundacionales de visión; y si la interpretabilidad puede mejorarse mediante intervenciones dirigidas, en lugar de limitarse a observarla.

El **primer capítulo** establece cuál es la base representacional adecuada para estudiar la interpretabilidad. Las neuronas individuales suelen responder a varios patrones visuales no relacionados entre sí, lo que las convierte en unidades de explicación poco fiables. El aprendizaje de diccionarios se ha propuesto como alternativa, pero su eficacia se ha evaluado principalmente mediante criterios técnicos, en lugar de mediante los juicios humanos que determinan si una representación es realmente interpretable. Este capítulo subsana esta carencia mediante tres experimentos de psicofísica a gran escala (N=481) realizados con dos redes neuronales convolucionales. Las características obtenidas mediante aprendizaje de diccionarios son sistemáticamente más interpretables que las neuronas individuales, y esta ventaja aumenta en las capas más profundas. Además, la propia elección de la base representacional puede determinar si las diferencias entre modelos llegan siquiera a detectarse.

Sobre esta base, el **segundo capítulo** aborda el problema de la medición. Los protocolos existentes para comparar la interpretabilidad de distintos modelos se basan en estudios de usuario con tareas de elección forzada que permiten a los participantes acertar por eliminación, en lugar de hacerlo gracias a una comprensión genuina, lo que dificulta realizar comparaciones justas entre modelos. Este capítulo propone dos protocolos conductuales, la localizabilidad y la semanticidad, que evalúan directamente la interpretación y sitúan los modelos en una escala común corregida por el nivel de azar. Aplicados a seis modelos *transformer* de visión, incluidos cuatro modelos fundacionales (DINOv2, DINOv3, CLIP y SigLIP), y basándose en más de 15.000 respuestas humanas, estos protocolos demuestran que los modelos fundacionales son sistemáticamente menos interpretables que su homólogo entrenado de forma supervisada. Esta diferencia no se explica ni por la escala del modelo ni

por su rendimiento en tareas posteriores. En cambio, la localidad de las características y su alineación con los juicios semánticos humanos aparecen como los principales predictores de la interpretabilidad.

El **tercer capítulo** estudia si estos factores predictivos pueden convertirse en mecanismos causales para modificar la interpretabilidad. Cinco codificadores visuales se ajustan mediante LoRA para alinearlos con juicios perceptivos humanos, y los modelos alineados resultantes se evalúan utilizando las mismas medidas conductuales desarrolladas en el Capítulo 4. Los resultados muestran que la alineación constituye un mecanismo poco fiable para mejorar la interpretabilidad: su efecto es positivo en los tres modelos exclusivamente visuales, aunque solo alcanza significación estadística para DINOv2, y se invierte en los dos modelos de visión y lenguaje. Incluso cuando la alineación mejora la interpretabilidad, el mecanismo subyacente sigue siendo inconsistente y solo se comprende parcialmente. Por tanto, aunque la alineación puede servir como mecanismo para mejorar la interpretabilidad, no proporciona una solución fiable ni universalmente aplicable, y siguen sin resolverse las razones por las que funciona en algunos modelos y fracasa en otros.

En conjunto, estas contribuciones proporcionan el primer análisis, basado en medidas conductuales, de hasta qué punto son interpretables los actuales modelos fundacionales de visión, por qué difieren entre sí y en qué medida puede mejorarse deliberadamente su interpretabilidad. De manera más general, esta tesis sostiene que el progreso en inteligencia artificial explicable requiere ir más allá de la evaluación de las explicaciones y avanzar hacia el estudio sistemático de si los propios modelos llegan a ser comprensibles para los seres humanos. El marco desarrollado aquí proporciona una base para este cambio de perspectiva y apunta hacia futuros modelos cuyas representaciones —y, en última instancia, cuyas decisiones— puedan ser comprendidas de manera significativa por las personas que dependen de ellos.

Acknowledgments

A PhD is a life experience, one in which curiosity, stress, and ambition force us to grow. Fortunately, this is not a journey I had to walk alone, and for that I am grateful to the mentors and friends who helped me along the way.

First, I would like to express my deepest gratitude to my supervisor Nuria Oliver, who was always there to mentor me these past four years, through the good and the bad moments, always supporting me regardless of how tight the deadlines were and sharing my boundless optimism.

I am equally grateful to Thomas Serre, not only for his mentorship during the PhD, but also for taking me in for an internship close to six years ago when I had little Deep Learning experience, and officially kickstarting my interpretability journey.

In a similar spirit, I would like to thank Francois Dehez for taking me for an internship during my first year at University, introducing me to research, and sharing valuable advice to set myself up to do a good PhD, which has greatly shaped some of the decisions leading me here today.

I want to thank all the amazing people I have met and worked with over the years. You have been a great source of fun, valuable advice, and inspiration.

Thank you, Thomas, for the kindness you showed me in taking me on as a Master's student, even though you did not yet feel ready to mentor someone. Your infinite drive and positive energy, your openness in sharing your knowledge and insights, and your curiosity have all been a great source of inspiration. I am deeply grateful that our paths crossed, and for the years of collaboration, insightful conversations, occasional memes, and friendship that grew from that first experience.

Thank you, Remi, for your kindness and your generosity in sharing valuable insights whenever I needed them.

Thank you, Victor, for showing me that meetings can be incredibly fun while staying equally productive.

Thank you, Lakshmi, for not only being a great mentor, but also a friend and a brother. I truly appreciate your guidance, in and out of academia.

My relatively short time in the US was made special by the people I had the chance to meet, and I feel very fortunate for the friendships that grew from it. Thank you also to Tiantian, Lore, Alex, Alekh, Ivan, Jeason, and all the other amazing people I met in and around the lab, who made those months so memorable.

I want to thank all the amazing people at ELLIS Alicante for sharing this journey with

me. Thank you, Adrian, for the endless conversations about science, football or rap. Thank you, Aditya, for always being available to give advice, feedback or just discuss anything and everything.

I was able to focus on doing science because of the incredible people who supported me in all the other aspects of the PhD. Thank you, Cristina, Rebeca, Miguel Angel and Nicole.

I want to thank all my friends and family for being there for me all those years. Your support and encouragement were instrumental to me, leading up to the PhD and during the PhD, and I know I will continue to trust and rely on them. Thank you from the bottom of my heart.

Thank you to Antoine and David. Our friendship has been a constant source of inspiration and motivation to do better and to be better. I am deeply grateful for the mindset we have grown and refined together over the years, for giving me motivation when mine was low, and for reminding me to slow down and take care of myself when I needed to hear it. I know this journey would have been much harder without your constant support and friendship.

Thank you, Nicolas, for giving me a space where I could take my mind off the PhD. Thank you for somehow always being awake at 3 a.m. when I was still working, needed to unwind, and inevitably ended up talking about novels with you. More generally, all the carefree fun we have had over the years has helped bring some much-needed balance to my life.

Thank you, Thomas, for being a true friend for more than half my life, and for always being there for me like family.

Thank you, maman and Philippe, for always believing in me, and for everything you have done to make it possible for me to pursue my dreams. I am eternally grateful for that.

Thank you, Maxime. I grew up looking up to you, and I know that a big part of who I am today was shaped by having you as my big brother. Years later, I keep learning from you and finding inspiration in the person you are and the person you strive to be. Thank you for also being an anchor, and for always being someone I can turn to when I feel lost and need your advice.

Finally, to Mathilde, my life partner, I want to express my infinite gratitude. Almost four years ago, you made the courageous decision to follow me to Alicante, to a country whose language you did not speak, while looking for your very first job. I deeply admire the courage and decisiveness with which you stepped so far outside your comfort zone so that we could live this adventure together, and I will always be grateful that you did. You have been my pillar through all these years, in the PhD and far beyond it, hyping me up through the highs, supporting me through the lows, through all the weekends I was working and all the nights I was not sleeping. I was able to put so much into this PhD because I knew you always had my back; you were always there. These last four years have been the most stressful, but also the best years of my life. Over that time, I have grown immensely, and I am proud to say that a large part of that is thanks to you. I am leaving this PhD not only with a thesis, but also as a better person. This journey was difficult, as you know better than anybody, and you lived every part of it by my side. For that, I feel that this achievement belongs to you as much as it does to me. As this adventure comes to an end, I find myself more excited than ever for all the ones we still have ahead of us.

Contents

Abstract ix

Resumen xi

Acknowledgments xiii

1 Introduction 1
- 1.1 The interpretability problem . . . 1
- 1.2 A timely question . . . 3
- 1.3 From explaining models to reducing their opacity . . . 5
- 1.4 Outline and Contributions . . . 6

2 Background: The Interpretability Toolbox 11
- 2.1 Attribution Methods . . . 11
- 2.2 Feature Visualization Methods . . . 13
- 2.3 Concept-Based Methods . . . 14
- 2.4 A Mature but Incomplete Toolbox . . . 15

3 Where to look for interpretability? 17
- 3.1 Introduction . . . 17
- 3.2 Related work . . . 19
- 3.3 Methodology . . . 20
- 3.4 Results . . . 25
- 3.5 Conclusion . . . 29

4 How to measure and compare interpretability across models? 31
- 4.1 Introduction . . . 31
- 4.2 Related Work . . . 33
- 4.3 Method . . . 34
- 4.4 Results . . . 37
- 4.5 Discussion . . . 41
- 4.6 Conclusion . . . 42

5 Can we engineer interpretability? 45
- 5.1 Introduction . . . 45
- 5.2 Related Work . . . 46
- 5.3 Methods . . . 47

5.4 Results . . . 50
5.5 Conclusion . . . 55

6 Conclusion and Perspective **57**
6.1 Conclusion . . . 57
6.2 Perspective . . . 58

I Appendix **61**

A Appendix of Where to look for interpretability? **63**
A.1 Selection of dictionary elements . . . 63
A.2 Experimental confounds . . . 64
A.3 Examples of trials from Experiment II . . . 65
A.4 Sensitivity analysis . . . 67

B Appendix of How to measure and compare interpretability across models? **69**
B.1 Additional Related Work . . . 69
B.2 Broader impact. . . . 69
B.3 Details about the control experiment . . . 71
B.4 Scaling model interpretability . . . 73
B.5 Overview of our feature extraction and explanation pipeline . . . 74
B.6 SAE training, reconstruction, and stability . . . 75
B.7 Distribution of feature reuse across models . . . 76
B.8 Psychophysics experiment . . . 77
B.9 Quadrant of features . . . 80
B.10 Plausibility robustness . . . 83
B.11 Detailed results for the psychophysics experiments. . . . 85
B.12 Additional results on locality of representations . . . 87
B.13 Additional results on alignment with human similarity judgments . . . 88

C Appendix of Can we engineer interpretability? **89**
C.1 LoRA fine-tuning on THINGS . . . 89
C.2 Direct matching across independently trained SAEs . . . 90
C.3 Sensitivity to the co-firing cutoff . . . 91
C.4 Image-level co-firing. . . . 92
C.5 Sensitivity to the similarity threshold . . . 93
C.6 Informative-feature loss across backbones. . . . 94
C.7 Accuracy-change buckets and interpretability. . . . 95
C.8 Joint-dictionary correspondence and feature turnover . . . 96
C.9 Paired interpretability changes for matched features . . . 97

II Resumen en Español **99**

Bibliography **113**

Chapter 1

Introduction

1.1 The interpretability problem

Since AlexNet's breakthrough in 2012 [110], deep learning models have advanced at a remarkable pace, with their capabilities growing faster than our ability to understand them. This interpretability gap matters most where these models are used to make consequential decisions, including medical image interpretation [11, 37, 98], autonomous driving [68, 97], or robotic control [105]. The models are accurate enough to be deployed in real-life scenarios, yet remain largely opaque to their users. Understanding them nonetheless matters for a variety of reasons, such as identifying failures before they cause harm [36, 70], meeting emerging legal expectations of explanation [101], and learning from them when they are used to make new scientific discoveries [55, 100]. As models grow more capable and more widely deployed, the gap between what they do and what we understand about them widens [87], making interpretability a pressing need [103]. These pressures are beginning to acquire regulatory relevance: interpretability is increasingly named in certification frameworks—most explicitly by EASA, which lists "AI explainability" as a trustworthiness building block for aviation [53], and by the EU AI Act, whose high-risk provisions require systems transparent enough for humans to interpret and oversee their outputs [52]—an expectation that, while not yet a hard requirement, rises in proportion to risk.

At its strongest, interpretability asks for more than an explanation related to an individual prediction. It asks whether a person can form a useful mental model of the system: whether its representations can be inspected, its behavior anticipated on new cases, and its failures predicted before they occur. How to construct models that support this kind of human understanding remains an open question.

Interpretability vs Explainability. The two terms are often used interchangeably. Even EASA's own framework lists "AI explainability" [53] as the trustworthiness building block for the kind of understanding we describe above, but they focus on different things, and part of the field's long-running disagreement over what "interpretability" even means [121] stems from this conflation. In this thesis, we use *explainability* to describe a property of a system: whether it can produce an explanation –e.g., a saliency map, a concept score, a prototype match, a natural-language rationale– meant to convey the basis for an output or the content of a representation. We use *interpretability* for a property of the interaction between that system and a person: whether the explanation, once produced, actually leaves them with a

correct understanding [45, 134]. Explainability is a claim about a mechanism and can be verified by inspecting the system, whereas interpretability is a claim about a mental state and can only be verified by testing the person. An explanation can be available, even faithful to the model's computation, and still fail to be interpretable if it is not amenable for a person to parse it [95], or if it is comprehensible for a single prediction but never resolves into a mental model of the system's broader behavior. Explainability is therefore necessary but not sufficient for interpretability.

Two routes to achieve interpretable models. Existing approaches broadly follow two routes to achieve interpretability. The first builds transparency directly into the architecture or learning process of the model, yielding models that are *explainable-by-design* [162][1] by means of a variety of methods, including prototype networks that classify by comparison to visualizable exemplars [29], concept bottleneck models that route predictions through human-named concepts [108], or self-explaining networks that emit an explanation as part of the forward pass [6, 16]—trading architectural flexibility for transparency.

The second route, commonly referred to as *post-hoc explainable AI* (XAI) [73], starts with an already trained opaque model and develops methods to explain its predictions or internal representations after the fact. In practice, however, the boundary between these two paradigms is less clear than it first appears. For example, concept-bottleneck models and post-hoc concept discovery via sparse coding can converge on similar concepts [160]. More fundamentally, both approaches share the same limitation. While they guarantee that an explanation is *available*, neither guarantees that the resulting model is actually *understood* by a person. Whether either approach succeeds in fostering genuine human understanding remains a largely and open empirical question.

In the case of explainable-by-design models, evaluation typically focuses not on whether people understand the model, but on proxies for understanding. For example, in concept-based models, this often means testing whether the intended human-labeled concepts are recovered. Yet recent work suggests that these learned representations may instead function as metonymic proxies, with each recovered concept encoding information about multiple other concepts [7]. Recovering the intended concepts, therefore, does not necessarily imply that the model's reasoning is transparent to humans.

Post-hoc XAI fares somewhat better, largely because it has received more attention from the research community. A broad range of methods has been proposed and evaluated. In the case of computer vision, the focus of this thesis, there are three main families of XAI—namely attribution, feature visualization, and concept-based methods (discussed in more details in Chapter 2)—and each has drawn considerably more human evaluation than explainable-by-design models [3, 17, 27, 31, 60, 80, 106, 145, 158, 167, 173, 203]. However, these evaluations have primarily focused on comparing XAI methods against one another, rather than on comparing the interpretability of the underlying models, a gap Section C.12 revisits. As a result, we still know little about how interpretable different models are, or whether the field is making meaningful progress towards that goal. More fundamentally, deciding whether explainable-by-design or post-hoc XAI approaches are ultimately the more promising path presupposes a way to measure model understanding in the first place. That is precisely the

[1] The term generally used in the literature is *interpretable-by-design*. Given the distinction above, we find *explainable-by-design* more precise: these methods guarantee that an explanation is available, not that the model is understood.

capability the field lacks today.

Measuring human understanding. Part of why evidence that people genuinely understand a model—rather than merely find its explanations plausible or recognizable—is so hard to obtain is the lack of consensus on how to measure model interpretability. We find a variety of characterizations in the literature [45, 122] that generally only share the notion that interpretability is ultimately human-centric: an explanation has value if a person can understand it [133]. Measuring such an understanding is its own problem. Self-report is insufficient given that people are reliably overconfident about their own understanding of a system [19, 161]. A useful measure must instead test what that understanding enables. In science, understanding is often demonstrated through successful generalization: Newtonian mechanics is valuable as an account of planetary motion because it predicts trajectories beyond the observations from which it was derived. Interpretability requires an analogous functional test. A substantial literature in psychology argues that explanations support learning by helping people extract rules that generalize beyond previously observed cases [123–127, 192, 195]. Several interpretability protocols operationalize this idea directly. They test whether an explanation helps a person predict a model's decision on an unseen input [31], infer which stimuli will activate an internal feature [17, 203], identify biases the model has learned [158, 173], distinguish correct from incorrect predictions [106], or improve human and human–AI team performance [5, 145, 166, 184]. Which of these behaviors provides the most informative and representative measure of model understanding remains unresolved.

This thesis focuses on general-purpose pretrained deep learning models. The emphasis is intentional, as these models are the ones most widely deployed in the critical applications and, consequently, those for which reducing opacity is most pressing. Nevertheless, the framework developed in this thesis for *measuring*, *comparing*, and *shaping* interpretability is not tied to whether a model was designed to be inherently transparent or is explained post-hoc.

1.2 A timely question

Although model interpretability remains underexplored, several developments are converging to facilitate its study: the increasing maturity of post-hoc XAI methods, the convergence of computer vision models around shared architectures, and the emergence of methods for comparing representations at progressively finer levels of granularity. Hence, we believe that the field has reached the necessary maturity to shift from developing XAI tools to comparing the interpretability of models.

Imperfect tools can support cumulative science. Post-hoc XAI methods have received significant attention, yet, they remain imperfect: saliency maps can be insensitive to randomizing the model's weights or training labels [2], attributions can be fragile to imperceptible input perturbations [71], and different methods can disagree with one another on the same prediction [107, 188]. However, that is not a problem in itself: the use of imperfect but informative tools to build cumulative understanding of a complex system is a recurring practice in science, and systems neuroscience is a clear precedent. Decades of single-electrode recordings and hand-crafted stimuli, crude by modern standards, already

recovered a durable core of insight about visual processing, from orientation-selective units in early visual cortex [92] to highly selective neurons further up the ventral stream [78, 156]. That picture later proved incomplete: because single-electrode recordings sample one neuron at a time, they favored a serial, feedforward view of vision, until anatomical and physiological evidence showed the cortex to be densely recurrent, with feedback shaping responses at every stage [64, 112]. Nonetheless, the new findings did not overturn the earlier findings so much as enrich them, and rapid feedforward processing remains a real, central part of vision science.

The same principle motivates the use of current XAI methods, and in this thesis, we argue in Chapter 2 that today's post-hoc XAI toolbox is informative enough to begin building a cumulative account of model representations. Studying models with today's XAI methods is also what will tell us where those tools themselves still need to improve.

A converging model landscape. A second development is architectural convergence of computer vision models. For over a decade, convolutional networks (CNNs) dominated computer vision, and especially ResNet50 [83], which is still among the most widely used vision backbones. The field has since increasingly converged around transformer architectures, even as the objectives used to train these models have become more diverse. The same broad architectural template now supports supervised Vision Transformer [46], language-supervised contrastive models such as CLIP and its sigmoid-loss successors [157, 199], the self-supervised DINO family, now in its third generation [25, 152, 170], masked image modeling (MAE) [82], emerging predictive world-model objectives such as I-JEPA [8], and their extension to video [9, 12, 189, 193]. This convergence makes comparisons across models substantially more informative. With the architecture held fixed, differences in their interpretability can be attributed more cleanly to their training data and learning objectives rather than being confounded entirely by architecture. And because it is an architecture nearly everyone in the field now builds on, whatever those differences turn out to be is unusually broadly relevant.

Model comparison at progressively finer scales. Even with comparable architectures, studying differences between models is difficult because each representation is expressed in its own idiosyncratic basis. Existing methods address this problem at several levels of resolution. At the coarse level, early work compared models through aggregate quantities, such as the number of unique concept detectors a model contains [13, 65], or an average human interpretability score across a sample of its features [32, 204]. Such comparisons can establish whether models differ overall, but their aggregate nature provides limited power to isolate the specific factors responsible for those differences. At a finer level, a recent body of work has proposed methods such as crosscoders and universal sparse autoencoders that learn concept spaces shared across models [76, 102, 119, 186]. By supporting comparison at the level of individual features, they facilitate the identification of driver of interpretability. In this thesis, we adopt the same increasing granularity in the analysis.

1.3 From explaining models to reducing their opacity

Despite this favorable convergence of mature XAI tools and shared architectures, little work has systematically leveraged post-hoc methods to study the models' internal representations and compare them to one another.

Models as the unit of analysis. Detailed studies have mapped circuits and visual features within individual architectures [26, 148, 149], identified multimodal neurons in CLIP [75], decomposed its representations across attention heads [67], reverse-engineered the role of register tokens in vision transformers [38]. Collectively, these studies demonstrate the explanatory power of the XAI toolbox when it is used to probe models themselves rather than to simply generate explanations.

Studying a model in this kind of detail is also what reveals the limitations of the interpretability toolbox itself. A prime example comes from the circuit-level mapping of individual neurons in the studies cited above [149]. When pursued at sufficient scale, this fine-grained analysis uncovered neurons that were *polysemantic*, responding to mixtures of seemingly unrelated concepts rather than to a single identifiable pattern, a property later formalized as superposition [50]. This finding undermines a neuron-centric account of a model's representations almost by construction: if individual neurons need not correspond to coherent concepts, then neurons themselves cannot serve as a reliable unit of interpretation. This realization motivated a shift toward dictionary learning as an alternative representational basis. By learning a sparse, overcomplete code over a layer's activations, methods such as sparse autoencoders aim to recover directions that are individually more monosemantic than the neurons from which they are derived [18]. These methods have since been used to expose interpretable, causally testable features in modern vision models [62, 176]. Yet this progress has not resolved the question of the appropriate unit of interpretation so much as displaced it. The same body of work that established the case for sparse representations has also revealed their instability: learned features can split or merge as dictionary width changes [18], and autoencoders trained on identical activations with different seeds or hyperparameters can recover markedly different, non-canonical sets of directions [59, 113].

Superposition makes the neuron an unreliable unit of analysis; the absence of a stable, identifiable dictionary makes the sparse alternative an improvement rather than a resolution. Which of the two—or under what conditions either—a model comparison should be built on, and whether that choice can itself change the comparison's outcome, is, at this point in the literature, still an open question.

Moreover, comparative studies remain relatively scarce. Existing work suggests that interpretability does not necessarily improve with model scale or capability [13, 204, 205], while automated concept labeling, representational alignment, and shared feature dictionaries have been used to examine concept coverage and representational convergence across models [44, 47, 132, 137, 147, 186]. At the same time, psychophysical evaluation protocols provide mature methodologies for assessing whether humans actual understand model explanations [17, 27, 60, 106, 158, 184, 203]. However, these human evaluations have been used primarily to compare XAI methods rather than the models themselves. Systematically applying such evaluations across models to ask *which model humans understand best* remains a largely unexplored direction, which this thesis aims to advance by establishing a framework to compare the interpretability of modern vision models.

Models as the object of intervention. Describing a model and comparing it to others leaves the model itself unchanged. Such approaches quantify the gap between what a model encodes and what humans can understand, but they do not attempt to reduce it. A complementary line of research instead asks whether this gap can be narrowed by intervening on the model's representations rather than on the tools used to inspect it. However, only a small number of these efforts explicitly optimize for human understanding.

One recent proposal in NLP frames the problem as a vocabulary mismatch: models already encode many concepts that they lack the means to express, suggesting that interpretability can be improved by teaching them a richer conceptual vocabulary [87]. Vision, however, offers no direct analogue, as visual representations lack a discrete vocabulary to extend. Instead, computer vision has pursued representation-level interventions that encourage the latent space as a whole to become more interpretable, for example by promoting robustness to imperceptible input perturbations [190, 194]. A related and increasingly active line of work seeks to align model representations with human perception. Representations that better match human similarity judgments have been shown to improve few-shot learning, robustness, and cross-task transfer [140, 178]. Building on the observation that interpretability is ultimately bounded by human cognitive [111], several recent works have further argued that perceptual alignment may serve as a route towards more interpretable models [61, 139, 181].

What this cycle rarely produces, however, is a direct test of the hypothesis that motivated it. Rather than evaluating whether an intervention makes a model more understandable to humans, studies typically report improvements on the intervention's own optimization criterion. In this thesis, we argue that interpretability ultimately demands behavioral evidence: whether people can better predict, generalize from, or reason about the resulting model. Consequently, the field has accumulated a growing collection of candidate principles for interpretability that are plausible, controllable, and widely adopted, but have seldom been evaluated against the very criterion they are intended to satisfy.

1.4 Outline and Contributions

Closing the gap identified above proceeds through three questions: where to look for interpretability, addressed in Chapter 3; how to measure whether models are understood by humans and compare that understanding across today's vision foundation models, addressed in Chapter 4; and to what degree such understanding can be engineered rather than simply observed, addressed in Chapter 5. These chapters trace a coherent path from identifying the appropriate basis for studying interpretability, to measuring it across models, to acting on what those measurements reveal. Figure 1 provides a visual overview of this progression. The thesis also develops a broader perspective on how the field should approach model interpretability, introduced in the preceding sections. We now turn to each of these contributions in greater detail, as well as to the scientific publications upon which they are grounded.

Introduction & Background These two chapters serve two purposes. Chapter 1 contextualizes where the field of explainable AI currently stands and argues for where it should go next, from a focus on making models explainable, to studying and comparing how in-

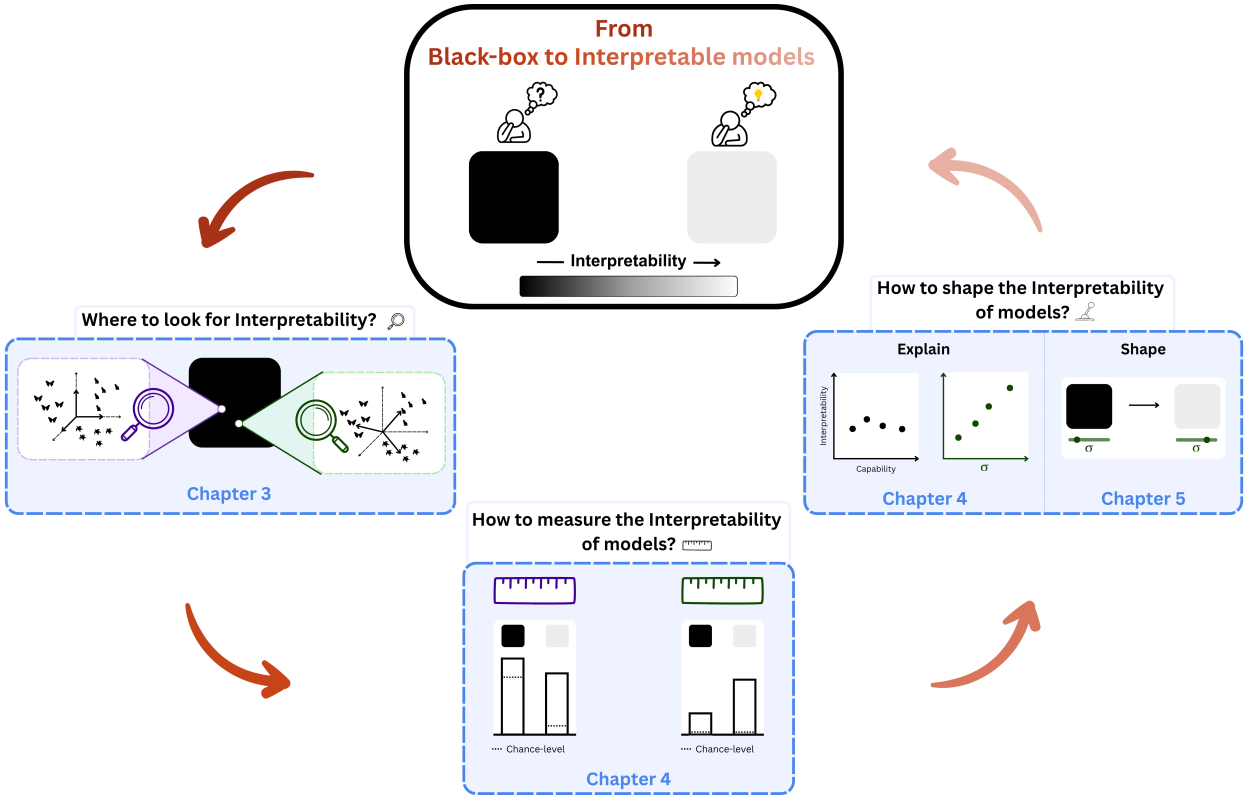


Figure 1. Overview of the thesis. The thesis progresses from identifying an appropriate representational basis for interpretability, to measuring and comparing interpretability fairly across models, and finally to testing whether it can be deliberately shaped, with the broader goal of moving from opaque black-box models toward models that are more interpretable to humans.

terpretable those models actually are to humans, setting the stage for the rest of the thesis as a first step toward that shift. Chapter 2 describes the specific XAI toolbox the thesis builds on, and argues that this toolbox, though imperfect, is mature enough to support the model-centric study the chapters that follow undertake.

- **Julien Colin**, Nuria Oliver, Thomas Serre. *From Interpretability Methods to Interpretable Models.* ECCV Workshop on Explainable Computer Vision, 2026.

Where to look for interpretability? Chapter 3 addresses the first question on which the rest of the thesis depends: *what unit of a representation should interpretability be measured over*? It revisits the unresolved issue introduced in §C.12: because of superposition, individual neurons are an unreliable basis for explanation, motivating alternatives based on learning dictionaries, such as sparse autoencoders [18] and matrix factorization [60]. Yet these alternatives have largely been validated only using explainability-oriented criteria (§C.10), including reconstruction fidelity and cross-seed stability, with few studies asking the decisive question of interpretability itself: which representation is more interpretable to people?

To answer this question, the chapter presents three large-scale online psychophysics experiments (N=481) using an adapted activation-prediction protocol [17, 203]. Participants first recognized, and then generalized, the shared visual pattern underlying a unit's maximally

activating images, allowing neuron-based and dictionary-based representations to be compared head-to-head. The experiments were conducted on ResNet50 [83] and VGG16, two convolutional networks that have long served as testbeds for interpretability research [13, 57, 65]. Focusing on these models isolates the basis question before extending it to the less well understood transformer-based foundation models discussed in §C.11.

Dictionary-based representations prove consistently more interpretable than neurons, with increasing interpretability in deeper layers. This finding is particularly consequential because these are the layers most heavily probed and reused in practice. The degree to which learned dictionary elements align with individual neurons also varies across models, meaning that the effect of representation basis is itself model dependent. Consequently, comparisons performed at the neuron level, the implicit assumption underlying many aggregate comparisons discussed in §C.11 [13, 65, 204], can mask differences between models that become apparent only when representations are compared in a learned dictionary basis, as demonstrated for ResNet50. These findings provide the first human-grounded evidence favoring dictionary-based representations over neurons and show that the representational basis can itself determine the outcome of an interpretability comparison. This observation motivates the methodological choices carried forward into the model comparisons of Chapter 4.

- **Julien Colin**, Lore Goetschalckx, Thomas Fel, Victor Boutin, Thomas Serre, Nuria Oliver. *Choosing the Right Basis for Interpretability: Psychophysical Comparison Between Neuron-based and Dictionary-based Representations.* Transactions on Machine Learning Research (TMLR), 2026.

How to measure and compare interpretability of models? Chapter 3 establishes where to look for interpretability; Chapter 4 asks how it should be measured. The existing model-comparison protocols discussed in §C.11 [32, 204] cannot be directly applied to this new representational basis. We show that their forced-choice design allows participants to succeed through elimination alone, with performance bounded by a model-specific chance level, making fair cross-model comparisons impossible regardless of the underlying representation.

To address this limitation, the chapter introduces two complementary protocols that replace the categorization task with identification tasks. The first, *localizability*, asks whether an observer can predict *where* a feature will activate in a novel image. The second, *semanticity*, asks whether an observer can describe *what* the feature represents. Localizability scores are calibrated against model-specific chance levels, placing models on a common scale. We additionally match functionally analogous features across models, to allow for a fairer and finer comparison across models (§C.11).

With a more reliable metric established, the chapter returns to the question posed in §C.10: how interpretable are the vision models increasingly being deployed in high-stakes scenarios? We apply these protocols to six vision transformers from the architectural family surveyed in §C.11—two supervised ViTs and four foundation models (DINOv2, DINOv3, CLIP, SigLIP). We find that foundation models are consistently less interpretable than supervised baselines, and this gap is not explained by either model scale or downstream task performance, reinforcing the concern raised in §C.12. The two protocols also produce strongly correlated rankings, indicating that they measure a shared dimension of interpretability. Interpretability is best predicted by feature locality and coarse-grained semantic alignment with humans,

precisely the types of representation-level properties surveyed in §C.12 and the levers that Chapter 5 examines causally.

- **Julien Colin**, Lore Goetschalckx, Nuria Oliver, Thomas Serre. *Capability ≠ Interpretability: Human Interpretability of Vision Foundation Models.* arXiv, 2026.
- **Julien Colin**, Nuria Oliver, Thomas Serre. *Can Human Perceptual Similarity Alignment Improve Interpretability in Visual Representations?.* Conference on Cognitive Computational Neuroscience (CCN) (2026)

Can we engineer interpretability? Chapter 3 and Chapter 4 establish where and how interpretability should be measured; Chapter 5 closes the loop by returning to the growing gap introduced in §C.10 and asking whether it can be narrowed, rather than just measured. One promising route, surveyed in §C.12, consists of intervening directly on pretrained models by aligning their representations with human perceptual or semantic judgments. However, existing evidence for this approach is largely correlational and has not been evaluated against the human judgments that §C.10 argues should serve as the ultimate arbiter for interpretability. This chapter performs that test directly.

We LoRA-finetune the five vision backbones from Chapter 4 toward human perceptual judgments, and then evaluate the resulting models using localizability. We find that alignment does not consistently improve interpretability. Its effect is positive across the three vision-only backbones but reaches significance only for DINOv2, while reversing for the two vision-language models. This contrast, enabled by the convergence around shared architectures discussed in §C.11, provides a setting in which the conditions under which alignment succeeds or fails can be studied systematically.

To understand how alignment reshapes individual features, we further match features between each backbone and its aligned counterpart with the shared-dictionary methods introduced in §C.11 [76, 102, 119, 186]. This enables a more granular analysis of the mechanisms underlying any changes in interpretability. Even where alignment helps, its effects are not uniform: DINOv2's improvement is distributed across all its features, while DINOv3's gains are concentrated among the features most strongly modified by alignment. In vision-only backbones, increases in concept locality partially explain this difference, though they do not account for why alignment improves interpretability in one model but not the other. When alignment harms interpretability, the mechanism remains even less clear: SigLIP exhibits the largest decline of any backbone, yet none of the candidate mechanisms we examine explain this effect.

Altogether, these results show that perceptual alignment is a plausible but unreliable route toward more interpretable models. The intervention can improve interpretability, but it does not do so consistently, and the mechanisms governing when it succeeds, fails, or even reverses remain unresolved.

- **Julien Colin**, Nuria Oliver, Thomas Serre. *Does Human-Alignment Benefit Interpretability?.* ICLR Workshop on Representational Alignment (Re^4-Align) (2026)
- **Julien Colin**, Thomas Serre, Nuria Oliver. *Testing Human-Perceptual Alignment as a Lever for Interpretability in Vision Models.* (2026)

Conclusion and Perspective Chapter 6 brings together the thesis's three contributions into a unified framework for identifying, measuring, and causally testing principles of model interpretability. It then revisits the open questions raised in Chapter 1, using the findings developed throughout the thesis to chart the next steps toward models whose representations—and ultimately whose decisions—can be meaningfully understood by humans.

Chapter 2

Background: The Interpretability Toolbox

The case for studying models rather than methods rests on a premise: the methods must be good enough to build on. This chapter argues that they are. We review the interpretability toolbox through three established families of methods in computer vision. Each answers a different question: attribution methods identify *where* the evidence for a computation lies, feature visualization methods characterize *what* drives that computation, and concept-based methods identify the latent components through which information is represented. Although none provides a complete account of how a model reasons, each has matured enough that, taken together, they serve as reliable instruments for studying and comparing models.

Shared notation. Let $\mathcal{X} \subseteq \mathbb{R}^{H \times W \times C}$ denote the space of natural images, with $H \times W$ spatial locations and C color channels. Whenever a method requires a reference set of images, we draw on a fixed collection $\mathcal{D} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \subset \mathcal{X}$; expectations over $\mathcal{D}$ are understood as empirical averages. Let $f : \mathcal{X} \to \mathbb{R}^K$ denote a neural network producing K output scores, and let $f_t(\boldsymbol{x})$ denote its scalar score for a target output $t \in 1, \ldots, K$.

We write $h_\ell(\boldsymbol{x}) \in \mathbb{R}^{S_\ell \times d_\ell}$ for the representation of the model at layer ℓ, where S_ℓ is the number of spatial positions for convolutional networks or tokens for vision transformers, and d_ℓ is the dimensionality of each representation vector. A scalar target within this representation is specified by a functional

$$\rho : \mathbb{R}^{S_\ell \times d_\ell} \to \mathbb{R}, \qquad a_\rho(\boldsymbol{x}) = \rho\left(h_\ell(\boldsymbol{x})\right), \tag{1}$$

which may select a neuron, average a channel over spatial locations, or measure activation along a concept direction. When the target is the model output itself, we write $a_\rho(\boldsymbol{x}) = f_t(\boldsymbol{x})$.

2.1 Attribution Methods

Given a model, an input, and a target computation, attribution methods ask: *which parts of the input drove that computation?* This is perhaps the most fundamental interpretability question in computer vision, and one of the first to be systematically addressed by XAI [171]. Although attribution methods were introduced to relate pixels to class scores, the question they ask is general: it extends to any pair of input and scalar target, including internal representations on either side [41, 115].

Formally, an attribution method scores the contribution of each entry of a vector $\boldsymbol{v}$ to the target a_ρ, producing $\Phi(a_\rho, \boldsymbol{v}) \in \mathbb{R}^{|\boldsymbol{v}|}$. When $\boldsymbol{v} = \boldsymbol{x}$, this recovers the familiar heatmap over the $H \times W$ pixel locations (i, j); when $\boldsymbol{v}$ is instead an internal representation, the same construction scores each latent dimension of that representation. A very large number of methods instantiate this idea: gradient-based methods [169, 171, 174, 182] exploit differentiability to compute an exact local sensitivity; perturbation-based methods [58, 154, 158, 172] estimate importance by perturbing the input, without requiring access to gradients; and activation-based methods [28, 165] combine internal feature maps directly. We illustrate each family below: first with Saliency, from which Gradient×Input directly derives; then with RISE as a perturbation-based alternative; and finally with a brief discussion of Grad-CAM as a representative activation-based method.

Saliency [171], one of the earliest attribution methods, takes the absolute gradient of the target as importance:

$$\Phi_{\text{Sal}}(a_\rho, \boldsymbol{v}) = \left|\nabla_{\boldsymbol{v}} a_\rho(\boldsymbol{x})\right|, \tag{2}$$

Intuitively, it shows which coordinates of $\boldsymbol{v}$ would most change the target under a small perturbation, in either direction.

Gradient×Input [169] additionally weights this local sensitivity by the value of the input itself:

$$\Phi_{\text{GI}}(a_\rho, \boldsymbol{v}) = \boldsymbol{v} \odot \nabla_{\boldsymbol{v}} a_\rho(\boldsymbol{x}). \tag{3}$$

where $\odot$ denotes elementwise multiplication. Taking $\boldsymbol{v}$ to be an internal representation rather than the image has been shown to be particularly faithful [57], which is why we use it throughout the thesis to score the importance of concepts for the model's decisions.

RISE [154] estimates the effect of retaining or removing image regions. It samples random binary masks $\mathbf{M} \sim \mathcal{P}_{\mathbf{M}}$ over spatial locations, applied identically across channels, and computes

$$\Phi_{\text{RISE}}(a_\rho, \boldsymbol{x})_{ij} = \frac{\mathbb{E}_{\mathbf{M}}\left[a_\rho(\boldsymbol{x} \odot \mathbf{M})\, \mathbf{M}_{ij}\right]}{\mathbb{E}_{\mathbf{M}}[\mathbf{M}_{ij}]} = \mathbb{E}[a_\rho(\boldsymbol{x} \odot \mathbf{M}) \mid \mathbf{M}_{ij} = 1], \tag{4}$$

the model's expected output averaged over every mask that keeps location (i, j) visible, regardless of what else is masked out. A location scores high when revealing it alone is, on average, enough to drive a strong response. Since it treats the model as a black box, RISE needs no gradients at all. This makes it a fairer default when comparing architectures: in vision transformers, gradients pass through attention and normalization layers that make them a less reliable guide to pixel-level importance than in convolutional networks, so a perturbation-based measure puts every model on the same footing.

Finally, activation-based methods such as **Grad-CAM** [164] instead combine internal feature maps using gradients. Their standard formulation is tied to convolutional feature maps, whereas Gradient×Input and RISE apply consistently to both convolutional networks and vision transformers. We therefore use the latter two methods in the remainder of the thesis.

Attribution methods have received particularly sustained scrutiny. Early sanity checks revealed that several popular methods behaved like edge detectors and were insensitive to model or label randomization [2]; subsequent work exposed shortcomings in reliability [107],

robustness to manipulation [43, 71], evaluation methodology [89, 196], and the very definition of faithfulness [95]. This criticism has progressively refined both the methods and their evaluation, favoring approaches with stronger axiomatic [128, 182] or statistical [10, 58] foundations. Although imperfect, attribution maps demonstrably help humans: they support predicting model behavior on unseen inputs [31] and improve performance in visual-search and human–AI teaming tasks [106, 184]. Their outputs and limitations are therefore sufficiently characterized for the narrower purpose required here: identifying and comparing what contributes to model predictions or internal activations.

2.2 Feature Visualization Methods

Attribution methods explain *where* evidence is located, but leave the user to infer *what* feature is detected there. Feature visualization methods address this complementary question by retrieving or synthesizing inputs that strongly activate a selected model component [129, 143, 144, 150]. The two options are themselves complementary: retrieving natural images grounds the interpretation in what actually occurs in the data, but only ever shows a correlate of what drives the target; synthesizing an image instead searches directly for what causes a target activation, at the cost of leaving the distribution of natural images altogether. We therefore use both forms of visualization in this thesis.

Top-example visualization [17, 198] retrieves the most activating natural examples.

Given a target a_ρ and the reference dataset $\mathcal{D}$, its m most activating examples are

$$\text{TopExamples}_m(a_\rho; \mathcal{D}) = \underset{\mathcal{S}\subseteq\mathcal{D},\ |\mathcal{S}|=m}{\arg\max} \sum_{\boldsymbol{x}\in\mathcal{S}} a_\rho(\boldsymbol{x}). \tag{5}$$

Synthetic feature visualization [129, 143, 144, 150] searches directly for an input that maximizes the target:

$$\boldsymbol{x}^\star \in \underset{\boldsymbol{x}\in\Omega}{\arg\max}\, \mathbb{E}_{\tau\sim\mathcal{T}}\left[a_\rho(\tau(\boldsymbol{x}))\right] - \lambda\Psi(\boldsymbol{x}), \tag{6}$$

where $\mathcal{T}$ is a distribution of transformations (translations, crops, rescalings) encouraging robust solutions, Ψ is an image prior weighted by $\lambda \geq 0$, and $\Omega \subseteq \mathcal{X}$ is a hard constraint on the search space. Different methods instantiate this objective through different choices of target, parameterization, prior, and optimizer. Historically, this flexibility introduced a trade-off: without a strong prior, suppressing adversarial-looking patterns required heavy transformation robustness, making optimization slow and difficult to scale; with a strong generative prior instead, scaling became possible but at the risk of producing unfaithful images [141–143].

MACO [56] addresses these failure modes by constraining the Fourier magnitude of the synthesized image and optimizing only its phase, inspired by psychophysics findings that human object recognition is driven primarily by phase [22, 74, 79, 99, 151]. This halves the number of parameters being optimized and eliminates the need for a separate prior, allowing the method to scale to modern convolutional networks and vision transformers [56, 80].

Let $\mathcal{F}$ denote the (per-channel) Fourier transform, and let

$$\boldsymbol{r}_\mathcal{D} = \frac{1}{N}\sum_{\boldsymbol{x}\in\mathcal{D}} |\mathcal{F}(\boldsymbol{x})| \tag{7}$$

be a reference magnitude spectrum estimated from natural images. The image is parameterized by its phase alone, $\boldsymbol{x}(\boldsymbol{\varphi}) = \mathcal{F}^{-1}(\boldsymbol{r}_{\mathcal{D}} \odot e^{i\boldsymbol{\varphi}})$, optimized in place of the pixels themselves:

$$\boldsymbol{\varphi}^\star \in \arg\max_{\boldsymbol{\varphi}} \mathbb{E}_{\tau\sim\mathcal{T}} \left[a_\rho(\tau(\boldsymbol{x}(\boldsymbol{\varphi})))\right], \qquad \boldsymbol{x}^\star = \boldsymbol{x}(\boldsymbol{\varphi}^\star). \tag{8}$$

Natural and synthetic visualizations have both been evaluated by how well they allow humans to infer what a unit encodes, with encouraging results [17, 32, 203, 204]. A complementary line of work replaces visual examples with natural-language descriptions [14, 86, 147]. Together, these methods provide multiple imperfect but useful interfaces for characterizing the content of an internal representation.

2.3 Concept-Based Methods

Attribution and feature visualization can be applied to any target, but historically that target was often an individual neuron. This is too restrictive when individual neurons respond to a mixture of unrelated visual patterns [13, 65, 150]. Concept-based methods therefore seek a more appropriate basis in which to study the representation.

TCAV [104] replaced the neuron basis with directions estimated from labeled examples of predefined concepts. Its explanations are limited to the concepts the investigator chooses to test, which may not correspond to those the model actually uses. ACE [72] and ICE [200] therefore automated concept discovery from the representations themselves, a line continued by the two methods used in this thesis. Both describe representations as combinations of r reusable directions learned from data. Let

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1^\mathsf{T} \\ \vdots \\ \boldsymbol{a}_n^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{n\times d_\ell} \tag{9}$$

collect n representations extracted from images, crops, patches, or spatial positions at layer ℓ, where d_ℓ is the dimensionality of the representation at that layer, with each row $\boldsymbol{a}_i \in \mathbb{R}^{d_\ell}$ obtained through spatial pooling when required.

CRAFT [60] discovers concepts through non-negative matrix factorization. For non-negative activations $\boldsymbol{A} \in \mathbb{R}_+^{n\times d_\ell}$, it solves

$$(\boldsymbol{Z}^\star, \boldsymbol{D}^\star) = \arg\min_{\boldsymbol{Z}\geq 0,\ \boldsymbol{D}\geq 0} \frac{1}{2} \left\|\boldsymbol{A} - \boldsymbol{Z}\boldsymbol{D}^\mathsf{T}\right\|_F^2, \qquad \boldsymbol{Z} \in \mathbb{R}_+^{n\times r}, \quad \boldsymbol{D} \in \mathbb{R}_+^{d_\ell\times r}, \tag{10}$$

with $r < d_\ell$. The columns of the dictionary $\boldsymbol{D}$ define r concept directions in activation space, while each row of the codes $\boldsymbol{Z}$ gives the non-negative coefficients with which those concepts reconstruct an observed representation.

Sparse autoencoders [20, 59, 69] learn a similar dictionary through a parameterized encoder and decoder. For a representation $\boldsymbol{a} \in \mathbb{R}^{d_\ell}$, an autoencoder computes a code $\boldsymbol{z}$ and a reconstruction $\widehat{\boldsymbol{a}}$,

$$\boldsymbol{z} = \sigma(\boldsymbol{W}_{\text{enc}}(\boldsymbol{a} - \boldsymbol{b}_{\text{dec}}) + \boldsymbol{b}_{\text{enc}}), \tag{11}$$

$$\widehat{\boldsymbol{a}} = \boldsymbol{W}_{\text{dec}}\boldsymbol{z} + \boldsymbol{b}_{\text{dec}}, \tag{12}$$

where $\boldsymbol{W}_{\text{enc}} \in \mathbb{R}^{r \times d_\ell}$ and $\boldsymbol{W}_{\text{dec}} \in \mathbb{R}^{d_\ell \times r}$ with, typically, $r > d_\ell$; $\boldsymbol{b}_{\text{enc}} \in \mathbb{R}^r$ and $\boldsymbol{b}_{\text{dec}} \in \mathbb{R}^{d_\ell}$ are bias terms; and σ is a nonlinearity chosen so that $\boldsymbol{z}$ stays sparse while $\widehat{\boldsymbol{a}}$ reconstructs $\boldsymbol{a}$.

TopK sparse autoencoders [69] enforce that sparsity directly rather than through a penalty on σ: only the k largest activations survive,

$$\boldsymbol{z} = \text{TopK}_k(\text{ReLU}(\boldsymbol{W}_{\text{enc}}(\boldsymbol{a} - \boldsymbol{b}_{\text{dec}}) + \boldsymbol{b}_{\text{enc}})) \,, \tag{13}$$

with the rest set to zero. Since exactly k units are ever active, its parameters $\theta = (\boldsymbol{W}_{\text{enc}}, \boldsymbol{b}_{\text{enc}}, \boldsymbol{W}_{\text{dec}}, \boldsymbol{b}_{\text{dec}})$ can be learned by minimizing reconstruction error alone, $\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{a}_i - \widehat{\boldsymbol{a}}_i\|_2^2$, with no separate sparsity penalty needed. The columns of $\boldsymbol{W}_{\text{dec}}$ form an overcomplete dictionary of feature directions—the analogue of D in Equation (10)—and $\boldsymbol{z}$ specifies the small subset used to reconstruct each representation.

These formulations make the relationship between concept methods explicit. CRAFT imposes a non-negative, low-rank decomposition, whereas TopK sparse autoencoders learn an often overcomplete nonlinear encoding in which exactly k features are active for each input. Both describe a representation as a combination of reusable directions and can therefore be viewed through the broader lens of dictionary learning [130]; their different assumptions provide different lenses on the same representation rather than a uniquely correct decomposition [57, 88]. Despite these assumptions, concept-based methods already support concrete forms of model analysis: direct interventions on representations [42], identifying which concepts are recruited by different tasks [62], exposing systematic blind spots in other models [15], and helping human observers understand model predictions [60]. They therefore provide a sufficiently mature basis for comparing not only which concepts different models encode, but also how explicitly and intelligibly those concepts are represented.

2.4 A Mature but Incomplete Toolbox

Table 1. The interpretability toolbox as used in this thesis. Each method serves as a fixed instrument applied across models and chapters.

Method	Role in this thesis	Chapters
Gradient×Input (3)	Concept-importance scores for the model	3–5
RISE (4)	Feature explanation	4–5
Top-example visualization (5)	Feature explanation	3–5
MACO (8)	Feature explanation	4–5
CRAFT (10)	Concept extraction	3
TopK sparse autoencoder (13)	Concept extraction	4–5

The three families answer complementary questions. Attribution methods identify where the evidence for a computation is located; feature visualization methods characterize the inputs that drive it; and concept-based methods determine the representational basis in which it can be analyzed. None of these outputs should be mistaken for a complete causal account of a model's reasoning, and every method introduces assumptions about what constitutes an explanation. Nevertheless, their limitations are increasingly well understood, their formulations are explicit, and their utility has been demonstrated empirically. Most importantly

for the present work, applying the same methods across models makes it possible to study differences in the models themselves rather than continually treating the interpretability method as the sole object of improvement. Table 1 summarizes how each instrument is used in the remainder of this thesis.

Chapter 3

Where to look for interpretability?

The remainder of this thesis studies interpretability through the internal representations of models: first by measuring and comparing how understandable different models are, and then by testing whether that understanding can be improved. Although we can already intervene on models, determining whether an intervention has improved interpretability requires deciding what in their representations should be measured, and how. This chapter addresses the first question by establishing which elements of a representation should serve as the units of analysis.

3.1 Introduction

A central goal of explainable AI (XAI) in computer vision is to identify the visual elements that drive the decisions of deep neural networks (DNNs) [13, 24, 72, 104, 135, 165]. Doing so requires recovering the visual patterns that systematically influence internal activations and ultimately shape model outputs.

This goal is shared with the study of biological vision, where decades of work have sought the "preferred stimulus" of individual neurons in the visual cortex [92, 156]. Early neuroscience-inspired XAI similarly focused on neuron-level analyses [13, 201], alongside methods that synthesize maximally activating images for individual units [e.g., 51, 150, 198].

In this context and for the remainder of this chapter, we use the term *neuron* to refer to the unit of a layer: an individual neuron in fully-connected ReLU layers and a channel (feature map) in convolutional layers.

At the same time, neuroscience has increasingly shifted from single-neuron selectivity to population codes (see [49]), reflecting the view that neural representations are sparse and distributed rather than purely local [81, 155]. A similar shift is emerging in XAI because neuron axes can be a poor explanatory basis under the "superposition'' hypothesis [50, 57]: when a model encodes more patterns than it has units, individual neurons may respond to mixtures of unrelated patterns—often referred to as polysemantic neurons.

To address this challenge, recent work applies dictionary learning methods [18, 21, 34, 59, 60, 69, 197] to learn a dictionary basis over layer activations, ideally yielding directions that correspond to single, human-recognizable patterns. From an interpretability perspective, a learned dictionary basis is appealing because it aims to replace those polysemantic neuron axes with more monosemantic directions (Fig. 2).

Despite growing interest in dictionary-based representations, direct human evidence that

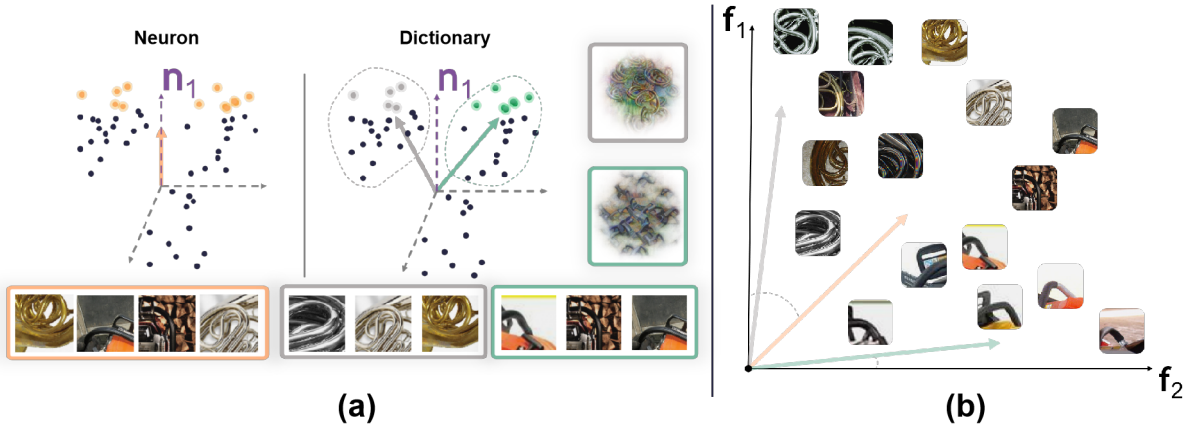


Figure 2. **(a)** • **Neuron (axis) basis** versus • **Dictionary basis**. Individual neurons may respond to multiple unrelated visual patterns (bottom), whereas dictionary learning aims to recover feature directions that isolate simpler patterns. **(b)** Dictionary learning yields a new basis over activations whose elements ideally correspond to single patterns. The interpretability hope is that these patterns align with the set of human-recognizable concepts $S = \{f_1, f_2, ..., f_n\}$.

they provide a better foundation for interpretability than neuron-based representations remains limited. This chapter aims to fill this gap. Furthermore, model comparisons remain underexplored and have largely relied on neuron-based evaluations [204, 205]; if models differ in their degree of superposition, conclusions drawn from neuron-based analyses may be unreliable.

Following the functional view of understanding introduced in Chapter 1—that an explanation is interpretable only insofar as it lets a person generalize beyond the cases they were shown—we operationalize interpretability via *visual coherence*: a basis is more interpretable to the extent that humans can reliably identify the common visual pattern in its maximally activating images and generalize that pattern to new images[2]. We additionally require that the directions evaluated within a basis be demonstrably involved in the model's decisions.

We conducted three large-scale psychophysics experiments (N=481 participants, 16,835 responses) comparing the interpretability of neuron-based and dictionary-based representations in ResNet50 and VGG16. Our main contributions are:

- We identify a semantic confound in widely used human-evaluation protocols [17, 204] and introduce a control that mitigates it.
- Across experiments, we find that dictionary-based representations are consistently more interpretable than neuron-based representations, with the advantage increasing in deeper layers.
- By quantifying axis-alignment[3], we find that ResNet50 exhibits lower axis-alignment of its dictionary elements than VGG16, indicating more distributed representations which is consistent with a greater degree of superposition than VGG16, obscuring cross-model differences in neuron-based analyses that become apparent only in dictionary-based comparisons, revealing ResNet50 to be the more interpretable model.

[2]In the remainder of this chapter, we use *interpretability* to refer specifically to this notion of human-recognizable visual coherence.

[3]Through the chapter, axis-alignment refers to the alignment of the representation with neuron axes.

- We empirically demonstrate that the choice of representational basis materially affects model comparisons, cautioning against drawing conclusions from neuron-level analyses alone.

3.2 Related work

From neuron-based to dictionary-based representations. Early work in explainable AI (XAI) for computer vision developed attribution methods [165, 182, 198] to explain individual predictions. These approaches primarily address the *where* question, that is, identifying which pixels or regions are most influential for a given output. However, they often fall short on the *what* question [31, 104]: the visual patterns or factors the model relies on to make its decision. This limitation motivated the development of feature visualization methods [143, 150], which aim to characterize what neurons or layers represent. A consistent finding is that single neurons can respond to multiple, visually distinct patterns [18, 23, 143], echoing similar observations in natural language processing [50].

These polysemantic responses suggest that neuron axes may not align with the underlying factors of variation in a model's representations. Under the superposition hypothesis [50, 57], models can encode more patterns than they have neurons, causing individual units to act as mixtures of multiple patterns. In that regime, interpreting single neurons may be no more principled than interpreting arbitrary directions in the activation space, motivating the search for alternative bases that better isolate interpretable elements. This perspective has spurred increased interest in learning dictionary bases over latent activations via dictionary learning and related concept-extraction methods [60, 72]. In parallel, sparse autoencoders (SAEs) have emerged as a promising approach for discovering more monosemantic directions in deep networks [18, 21, 34, 35, 59, 69, 197]. In this chapter, we compare the suitability of the neuron basis versus a learned dictionary basis as a foundation for interpretability in computer vision models.

Human-evaluation of interpretability. Since a primary goal of XAI is to make model behavior understandable to people, interpretability ultimately requires human-centered evaluation. Psychophysics-style experiments provide a direct benchmark for assessing whether an explanation is comprehensible and usable to human observers, complementing automated or proxy metrics.

[17] were the first to quantify the interpretability of deep neural network representations using psychophysics experiments. Their protocol visualized unit selectivity by contrasting maximally and minimally activating natural stimuli. In particular, they evaluated neuron-based representations (single-unit activations) in an Inception V1 [183] trained on ImageNet [40]. In each trial, participants viewed sets of maximally and minimally activating images for a given neuron and then selected which of two query images also strongly activated that neuron (see Fig. 3). They concluded that, from a human-centered perspective, natural exemplars are more effective than synthetically generated feature visualizations [150].

[203] proposed a variation of this task to investigate if humans gained causal insights from those visualizations. Participants predicted the effect of an intervention (*e.g.*, occluding an image region) on unit activation. They found that synthetic feature visualizations could improve performance, but provided limited advantage over natural exemplar images.

Interestingly, in a similar vein but in NLP, [18] compared the interpretability of 162 elements drawn from both neuron-based and dictionary-based representations. In their study, a single author rated each element using examples sampled across its activation range. They reported that dictionary-based elements were substantially more interpretable than individual neurons. Our work differs in domain (vision rather than language), experimental design, and scale (a single rater versus 16,835 behavioral responses from 481 participants).

Comparative analyses of model interpretability. [204] were the first to compare the interpretability of models using human evaluations. They extended the work of [17] to a broader range of computer vision architectures, including ResNet50 [83], and concluded that increasing model scale does not enhance interpretability. While we both build on the experimental protocol introduced by [17], their focus is on scaling these original insights across architectures. In contrast, we adapt the same protocol to address a different research question: which type of representation (neuron-based or dictionary-based) is more interpretable to humans. To the best of our knowledge, this is the first study that compares models using dictionary-based representations.

To further scale the evaluation of model interpretability, [205] automated the human evaluation protocol introduced by [17]. In each trial, human judgments were approximated by computing pairwise perceptual similarities between queries and explanations, which were then used by a binary classifier to predict the correct query. The resulting metric was found to correlate strongly with previous results [204]. This work shows that although most models contain a substantial number of interpretable units, differences across models primarily arise from the prevalence of highly uninterpretable units. Such units are often attributed to superposition, whose impact may vary across models. Consequently, comparing models solely based on neuron-based representations may disproportionately penalize certain architectures, potentially yielding misleading conclusions. Our work compares models using both neuron-based and dictionary-based representations and demonstrates that the choice of representational basis can substantially impact comparative outcomes.

3.3 Methodology

In this section, we first provide an overview of the technical methods used in our experiments (named Experiments I, II and III), followed by a description of the psychophysical experiments conducted to compare the interpretability of neuron-based versus dictionary-based representations.

3.3.1 Technical methods

Models Experiments I and II described below used a ResNet50 [83] and Experiment III used a VGG16 [172], both sourced from the Torchvision [131] library and pre-trained on ImageNet-1k [40]. We focus on convolutional neural networks (CNNs) because they remain widely used in practical computer vision applications, making our findings broadly relevant, and they produce the positive activations required by our dictionary learning method.

Dictionary-based representations. To compute dictionary-based representations for the psychophysics experiments (see Section 3.3.2), we employed CRAFT [60], a dictionary

learning method based on Non-negative Matrix Factorization (NMF). Specifically, given a model $f : \mathcal{X} \to \mathcal{A}$ that maps from an input space $\mathcal{X} \subseteq \mathbb{R}^d$ to an activation space $\mathcal{A} \subseteq \mathbb{R}^p$ (*i.e.*, any layer of the network), we compute the activations $\mathbf{A} = f(\mathbf{X}) \in \mathbb{R}^{n \times p}$, where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times d}$ represents a set of n input data points. Each row $_i \geq \mathbf{0}$ of $\mathbf{A}$ contains the non-negative activations for a given data point $\mathbf{x}_i$, due to the use of ReLU activations. NMF approximates $\mathbf{A}$ as:

$$(\mathbf{Z}^\star, \mathbf{D}^\star) = \underset{\mathbf{Z} \geq \mathbf{0},\ \mathbf{D} \geq \mathbf{0}}{\arg\min} \left\| \mathbf{A} - \mathbf{Z}\mathbf{D}^\top \right\|_F ,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Here, $\mathbf{Z} \in \mathbb{R}^{n \times k}$ are the **codes**, and $\mathbf{D} \in \mathbb{R}^{p \times k}$ forms the **dictionary** with k elements. Both $\mathbf{Z}$ and $\mathbf{D}$ are constrained to have non-negative entries and tend to be sparse due to the properties of NMF. The dictionary matrix $\mathbf{D}$ provides a new set of basis vectors aligned with the activation patterns of the neural network, while $\mathbf{Z}$ contains the coefficients representing the original activations $\mathbf{A}$ in terms of these basis vectors.

We use NMF rather than sparse autoencoders (SAEs) because NMF respects the geometry of post-ReLU activations: both the dictionary and codes are constrained to be non-negative, whereas SAEs can learn negative dictionary atoms corresponding to directions that cannot exist in the activation space [57]. NMF is also more stable, avoiding the hyperparameter sensitivity inherent in SAE training [59].

Element importance. Let $\mathbf{v}$ be a vector from an intermediate representation, either a dictionary element in $\mathbf{D}$ or a neuron axis in $\mathcal{A}$. Let $\mathbf{x}$ be an input that strongly activates $\mathbf{v}$, and let $f(\mathbf{x})$ denote the corresponding logit score. To assess the importance of $\mathbf{v}$, we measure how sensitive $f(\mathbf{x})$ is to perturbations of $\mathbf{v}$ using Gradient×Input [169], which has been shown to provide a faithful importance measure in latent spaces [57]:

$$\mathrm{GI}(\mathbf{x}, \mathbf{v}) \;=\; \mathbf{v} \odot \frac{\partial f(\mathbf{x})}{\partial \mathbf{v}} \,. \tag{14}$$

The more important $\mathbf{v}$ is for the model's decision at $\mathbf{x}$, the larger $\mathrm{GI}(\mathbf{x}, \mathbf{v})$.

Axis-alignment of dictionary elements. If a model contains polysemantic neurons, then a learned dictionary is expected to recover directions that are not well-approximated by any single neuron axis. To quantify how closely each dictionary element aligns with the neuron (coordinate) basis—or equivalently, how distributed it is across neurons—we use the sparsity measure of [90]. We note that axis-alignment measures the structure of the dictionary matrix $\mathbf{D}$—specifically, how concentrated each basis vector is on individual neuron axes—and should not be conflated with the sparsity of the codes $\mathbf{Z}$ or with the orthogonality of the representation. In the non-negative setting enforced by NMF and ReLU, lower axis-alignment indicates that a recovered feature direction is distributed across multiple neurons rather than concentrated on a single neuron. Such distributed representations are often associated with superposition, but axis-alignment does not provide a direct measure of the number of encoded features or the degree to which those features overlap.

Let $D \in \mathbb{R}^{p \times k}$ denote the learned dictionary for a layer with p neurons, and let $\mathbf{d}_\ell = D_{:,\ell} \in \mathbb{R}^p$ be the ℓ-th dictionary element expressed in the neuron basis. We define the average

axis-alignment score as

$$\mathcal{H}(D) \;=\; \frac{1}{k}\sum_{\ell=1}^{k}\frac{\sqrt{p}-\|\mathbf{d}_\ell\|_1/\|\mathbf{d}_\ell\|_2}{\sqrt{p}-1}\,. \tag{15}$$

Here $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the ℓ_1 and ℓ_2 norms. For nonnegative vectors, this index lies in $[0,1]$: it is 0 for maximally dense vectors (mass evenly spread across coordinates) and approaches 1 for one-hot vectors (mass concentrated on a single neuron axis). Thus, higher values indicate that dictionary elements are closer to individual neuron axes, *i.e.,* more coordinate-sparse or axis-aligned, whereas lower values indicate elements that combine many neuron axes.

3.3.2 Psychophysics experiments

Experimental protocol

Interpretability is ultimately a human-centered concept. To examine how the choice of representational basis shapes human judgments, we conducted three large-scale online psychophysics experiments comparing units expressed either in the neuron basis or in a learned dictionary basis. Throughout, we use *visual pattern* (or *visual feature* in the perceptual sense) to refer to recurring, human-recognizable regularities in images, and we reserve *dictionary element* for a learned direction (dictionary element) in the activation space. Across all experiments, we evaluate how easily observers can identify the *consistent visual pattern* associated with a unit (neuron axis or dictionary element) from its maximally activating stimuli.

Because standard interpretability evaluations do not easily scale to large numbers of units [31], we adapted the psychophysics protocol of [17] to measure a unit's *visual coherence*, or equivalently, its lack of ambiguity, as a proxy for interpretability.

Each participant was assigned to one of two between-subject conditions: the neuron-basis condition or the dictionary-basis condition. After completing a practice session of 9 trials, participants performed 40 trials of the same task, with each trial corresponding to a different unit (a specific neuron axis or a dictionary element, depending on the condition).

In each trial, participants were shown two panels of 9 reference images, one on each side of the screen, separated by two *query* images in the center (Fig. 3). The right panel displayed images selected to strongly activate the target unit (maximally activating stimuli). The left panel displayed images intended to contrast with the right panel (the exact selection procedure depends on the experiment; see Section 3.3.2). Participants were asked to select the query image that they believed matched the reference images on the right, *i.e.*, the query image that shared the same consistent visual pattern as the maximally activating reference set.

Intuitively, when the maximally activating images are visually coherent, participants should more reliably select the correct query. Visual coherence suggests the unit's activation is driven by a consistent, recognizable pattern. Conversely, if the maximally activating images are heterogeneous (consistent with polysemanticity), the task becomes more ambiguous and accuracy should drop. We therefore summarize performance as the proportion of correct responses per unit, which serves as our primary measure of visual coherence and hence interpretability.

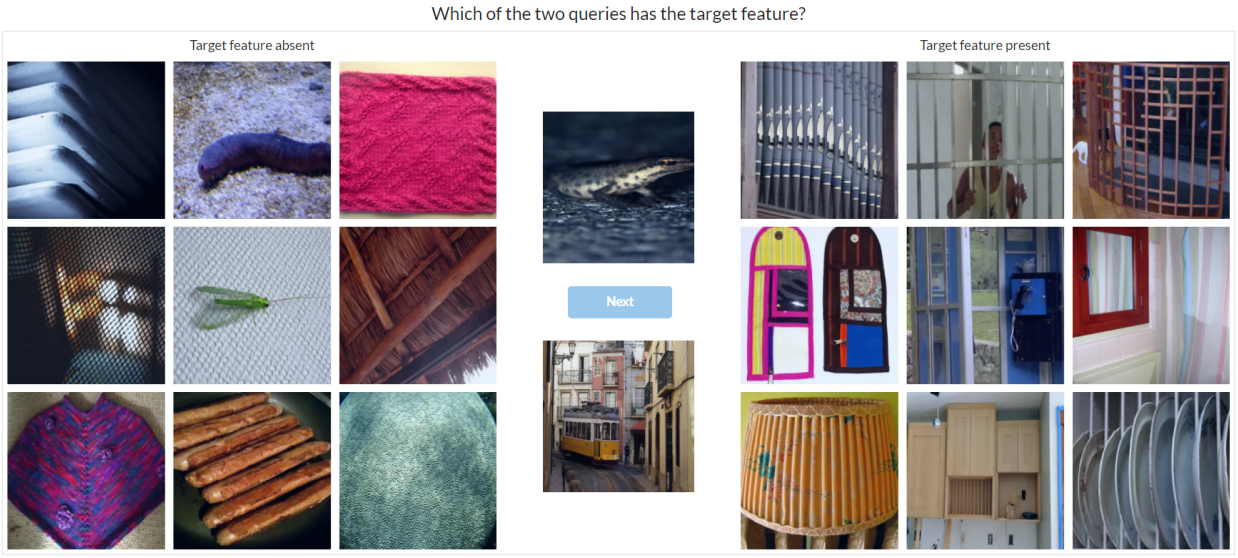


Figure 3. Illustration of a trial. Example of a trial in our study corresponding to Experiment I, dictionary-based condition for a unit located in *layer*2.0.*bn*2. Two panels of 9 reference images are located on the left and right-hand side of the display, separated by 2 query images in the center. Participants were asked to select the query image they believed shared the same visual pattern as the reference images displayed on the right panel, corresponding to maximally activating stimuli. The less ambiguous this shared pattern is, the more visually coherent the set of images and the more likely participants are to select the correct query. In this case, the correct query is the bottom image depicting a yellow tram.

Unit and stimuli selection

Unit selection. Following [204], we construct a representative sample of units across different layers of the neural networks of interest by adopting their proposed sampling procedure. Specifically, we sampled neurons by first selecting a network layer uniformly at random from the set of layers of interest, and then selected a neuron uniformly at random from within the chosen layer. This two-stage sampling scheme was used instead of sampling uniformly over all neurons to avoid bias toward later layers, which typically contain more units in CNNs. We repeated this procedure to obtain 80 units in total, spanning multiple layers of the network [204]. Because CRAFT requires non-negative activations [60], we mapped neurons from convolutional layers to their counterparts in the subsequent batch normalization layer (*e.g.*, *layer*1.1.*conv*1 neuron 53 -> *layer*1.1.*bn*1 neuron 53), which are followed by ReLU and thus yield non-negative post-ReLU activations. For VGG16 [172], we randomly selected 80 neurons across the ReLU layers. In total, we evaluated 80 units per model, distributed across 43 layers for ResNet50 and 11 layers for VGG16.

Stimuli selection for the neuron-based condition. For each of the selected neurons, we identified a representative set of 2,900 images from the validation set of ImageNet ILSVRC 2012 [163]: the 2,500 most strongly activating images and the 400 least strongly activating images. Following [17], we illustrated each neuron's selectivity through both maximally activating stimuli (images likely to contain the relevant visual pattern; see Fig.3, right panel)

and minimally activating stimuli (images unlikely to contain it; see Fig.3, left panel). For the maximally activating panel, we selected a random sample of 9 images from the top 150; for the minimally activating panel, we uniformly sampled 9 images from the bottom 20. We created 10 different trials per neuron following this procedure to ensure image independence across results.

Stimuli selection for the dictionary-based condition. Our hypothesis is that applying dictionary learning to neuron-based representations allows us to recover alternative directions capturing the distributed structure that is not aligned with individual neurons. Such structure is often associated with superposition and may provide a more interpretable basis for analysis than the original neuron basis. To test this, we derived a complementary dictionary-based representation for each neuron in the neuron-based condition. Specifically, for each neuron in the neuron-based condition, we selected its top N=300 maximally activating images, *i.e.*, those that most strongly activated it. We then applied CRAFT to this set of images to learn a dictionary of k=10 elements over the activations of that neuron's layer (we select N and k following the recommendations in [60]). From these 10 dictionary elements, we selected the one that was most frequently maximally activated across the 300 images (see Appendix A.1 for more details). Finally, we ranked the 2,900 images according to their activation along the chosen dictionary element to obtain the set of stimuli that illustrate the corresponding visual pattern.

Experiments

In this section, we describe each of the three psychophysics experiments conducted in this work. Each experiment consists of 1,600 trials (80 units $\times$ 2 conditions $\times$ 10 trials per unit).

Experiment I. This experiment is an adaptation of the methodology proposed by [17], conducted on a ResNet-50, with two conditions: neuron-based versus dictionary-based representations. An illustration of a trial from Experiment I can be found in Fig. 3, and the experimental protocol is described in Section 3.3.2.

Experiment II. The main objective of Experiment II was to control for a potential semantic confound uncovered in Experiment I. If the reference images that contain or do not contain the pattern of interest belong to distinct semantic categories, then it may be possible to solve the task through simple semantic grouping. Fig. 12 illustrates this phenomenon: in the depicted example, it is easier to solve the trial by inferring that the pattern of interest is *not* about a monkey than to identify the actual visual pattern present in the reference images. In such cases, correctly solving the trial tells us little about the interpretability of the unit. To mitigate this semantic confound, we proceeded as follows. Given a set of reference images containing the pattern of interest, we extracted their semantic labels from ImageNet and searched the 400 minimally activating images for a set of 9 images matching these semantic labels. When ImageNet labels were insufficient, we expanded the search by moving upward through the WordNet [63] hierarchy. After up to 4 iterations, all but one unit were successfully controlled. This unit was excluded from both the neuron-based and dictionary-based conditions.

Experiment III. To compare the interpretability across models, Experiment III applied the protocol from Experiment II to a VGG16.

Participants

A total of 481 participants were recruited for Experiments I, II, and III through the Prolific [4] online platform. All participants were native English speakers who reported no visual impairments and completed the study on a laptop or desktop computer (not mobile devices). They provided informed consent electronically and were compensated $2.75 for their time, corresponding to $15 USD per hour (approximately 10–13 minutes). The protocol was approved by the Institutional Review Board (IRB) of an institution affiliated with the authors. Based on the power analysis of [204], a minimum of 60 participants per condition (*i.e.*, 120 participants per experiment) was needed to obtain statistically robust results ($p < 0.05$). This prior was used given the closely matched experimental design. A sensitivity analysis in Appendix A.4 further shows that our retained sample sizes were adequate to detect medium-sized dictionary-vs-neuron effects. Participants were required to: (1) succeed in at least 5 of the 9 practice trials, (2) correctly answer at least 4 of the 5 catch trials (attentiveness tests) randomly inserted throughout the experiment, and (3) complete the experiment within 3 standard deviations of the mean completion time for that experiment.

Trial assignment. Each participant was assigned to a single condition (neuron-level or dictionary-level), making `Representation` a between-subject factor. For each participant, we (i) randomly sampled 40 of the 80 units and (ii) for each sampled unit, randomly drew 1 of its 10 trials. Participants therefore performed 40 trials, each corresponding to a different unit. No participant saw the same unit twice. Each experiment comprised 80 units x 2 conditions x 10 trials per unit = 1,600 unique trials. The 10 independent trials per unit (drawn from different random image samples within the top-150 pool) were distributed across different participants, such that each trial was seen by an approximately equal number of participants. `Depth` and, across experiments, `Model` were within-subject factors, as each participant saw units from multiple layers of one model.

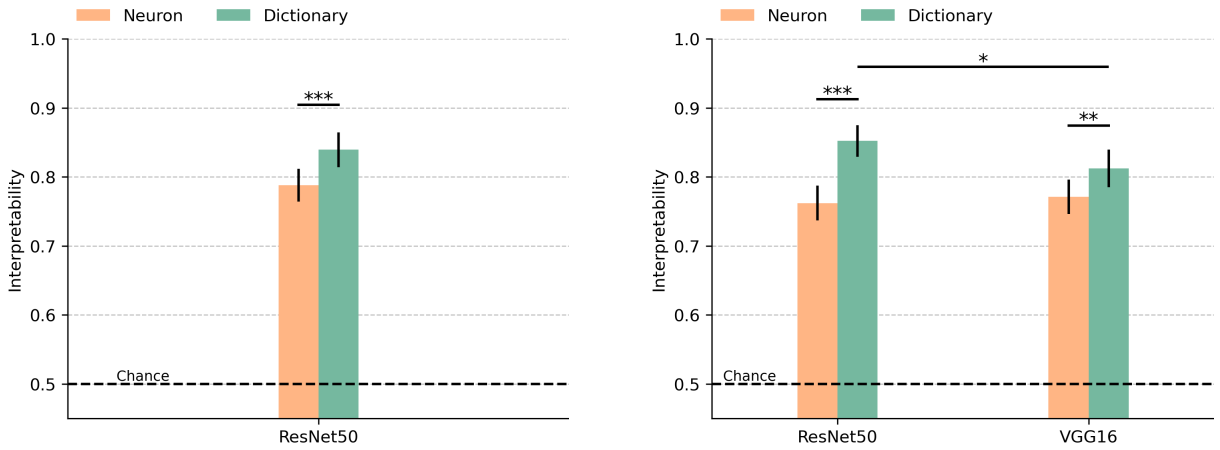
3.4 Results

3.4.1 Dictionary elements are at least as decision-relevant as individual neurons

Interpretability aims to enable humans to understand the decision-making processes of machine learning models. Accordingly, it is essential to first verify that the elements under investigation genuinely influence model decisions. We therefore assess the relative importance of dictionary elements relative to neuron axes, confirming that the former are at least as informative as the latter for driving model outputs.

We quantify the importance of each element used in our psychophysics experiments by measuring the sensitivity of model decisions to perturbations of that element across its 300 most activating images (see Eq. 14). For each matched pair, we compute the difference in importance between the dictionary element and its corresponding neuron. We find no

[4] www.prolific.com



(a) Experiment I. **(b)** Experiment II and III.

Figure 4. Results for (a) Experiment I, and (b) Experiments II (ResNet50) and III (VGG16). Given a unit and a set of images illustrating it, we assess how visually coherent participants find this set of images, or equivalently, how unambiguous the underlying pattern is. Specifically, we measure the proportion of trials in which participants correctly identified the query image belonging to this set of images and refer to this value as *Interpretability*. Across all experiments, participants found it significantly easier to identify the consistent pattern in the dictionary-based condition than in the neuron-based condition. Additionally, ResNet50 is more interpretable than VGG16 in the dictionary-based condition but not in the neuron-based condition.

significant difference in ResNet50 ($M = 0.0002, SD = 0.0018$), $t(77) = 0.93$, $p = 0.36$, while dictionary elements are slightly more important than neurons in VGG16 ($M = 0.002, SD = 0.005$), $t(76) = 3.34$, $p = 0.001$ (One-sample t-test against 0).

These results show that dictionary elements have similar influence on model outputs as their corresponding neurons, attesting to the relevance of the selected elements for interpretability and supporting the fairness of our comparisons by matching elements of comparable importance.

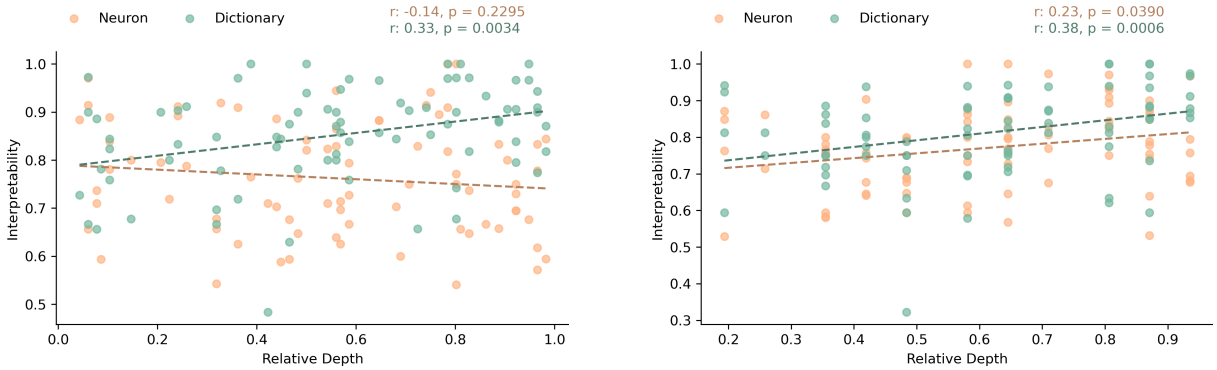
3.4.2 Interpretable features should be understandable by humans

In this section, we summarize the results obtained from analyzing the responses from 133, 130, and 129 participants who successfully completed Experiments I, II, and III.

Replication of previous findings. The experimental protocol employed in Experiment I for the neuron-based condition is the same as that described by [204]. Thus, we first assess the extent to which our results replicate previous findings. For ResNet50, [204] report an average task performance of $83.0\% \pm 2.0$ [5]. In our experiment, we obtain an average performance of $78.8\% \pm 1.5$. Given the similarity of these results, and considering that the specific selected units are not exactly the same (as described in Section 3.3), we conclude that Experiment I reproduces previously reported findings for neuron-based representations. This result also provides external validation of our experimental protocol.

[5] Values inferred from Fig. 3 in [204]

Human performance is superior in the dictionary-based condition. Based on our main hypothesis that dictionary-based representations constitute a better basis for interpretability than neuron-based representations, we predicted that participants would perform better in the dictionary-based condition. Results across all three experiments are illustrated in Fig. 4. In Experiment I, the average performance across participants in the dictionary-based condition was ● $83.5\% \pm 1.4$ [6] when compared to ● $78.8\% \pm 1.5$ in the neuron-based condition. A Mann-Whitney U test revealed that participants performed significantly better in the dictionary-based condition $z = 3.12$, $p < .001$. This result was corroborated in both Experiments II (see examples of trials in Figs. 13–15) and III, with a mean participant performance of ● $85.1\% \pm 1.4$ *vs.* ● $76.2\% \pm 1.6$, $z = 4.83$, $p < .001$ and a mean performance of ● $81.5\% \pm 1.5$ *vs.* ● $77.1\% \pm 1.6$, $z = 2.46$, $p = 0.01$, respectively. In subsequent analyses, we focus on the results from Experiments II and III.



(a) ResNet50. **(b)** VGG16.

Figure 5. Correlation between human performance and the relative depth of the layer where the feature is extracted from for (a) ResNet50 and (b) VGG16. For both models, we only find a significant correlation between interpretability and depth in the dictionary-based condition, after correcting for multiple comparisons (Bonferroni correction).

The deeper the layer, the more prominent the benefits of dictionary-based representations. Interestingly, our results suggest that the benefits of dictionary-based representations increase with relative layer depth in both models. Specifically, we observe a significant positive correlation between *interpretability* and *depth* for the dictionary-based condition only: Experiment I: $r = 0.35$, $p = 0.001$; Experiment II: $r = 0.33$, $p = 0.003$ for ResNet50 and $r = 0.38$, $p < 0.001$ for VGG16. In contrast, the correlations for the neuron-based condition are weaker and not statistically significant after Bonferroni correction: Experiment I: $r = 0.09$, $p = 0.4$; Experiment II: $r = -0.14$, $p = 0.23$ for ResNet50 and $r = 0.24$, $p = 0.04$ for VGG16. These patterns are illustrated in Fig. 5a and 5b, corresponding to Experiments II and III, respectively. To directly test this interaction, we fit a mixed-effects model on the pooled data from all three experiments: `Interpretability` $\sim$ `Representation` $\times$ `Depth` $\times$ `Model + (1|Subject) + (1|Experiment)`. The Experiment random effect was estimated at zero variance (boundary-singular fit, as expected with only three levels), and refitting without it yielded essentially identical fixed-effect estimates. A Type-III ANOVA confirms the predicted `Representation` $\times$ `Depth` interaction,

[6] The values reported correspond to a 95% confidence interval.

$F(1,\ 15320.6) = 11.59$, $p < 0.001$. Post-hoc slope contrasts (*emtrends*), averaged over `Model`, show that the depth slope is significantly steeper for the dictionary basis than for the neuron basis (dictionary: 0.159, 95% CI $[0.122, 0.195]$; neuron: 0.068, 95% CI $[0.030, 0.105]$; contrast $= 0.091$, $z = 3.41$, $p < 0.001$). This directly confirms—without relying on the absence of a significant correlation—that the interpretability advantage of dictionary-based representations grows with layer depth. The three-way `Representation` × `Depth` × `Model` interaction did not reach significance ($F = 3.12$, $p = 0.078$), so we do not interpret the per-model slope differences as statistically significant, although the descriptive pattern (ResNet50 contrast: 0.138, $p < 0.001$; VGG16 contrast: 0.044, $p = 0.354$) is consistent with our results.

3.4.3 Comparison of model interpretability

While evaluating the interpretability of a specific machine learning model is important for validating its transparency for downstream use cases, it is equally important to understand what makes certain models more interpretable than others. Doing so requires fair model comparisons. In this section, we report three main findings obtained by comparing ResNet50 and VGG16.

ResNet50 exhibits lower axis-alignment than VGG16. Superposition arises when a model attempts to represent more patterns than it has neurons, resulting in patterns that are encoded by populations of neurons rather than by single units. As the number of encoded patterns increases, individual dictionary elements are expected to be less aligned with any single neuron axis. As summarized in Fig. 6, we quantify the alignment of learned dictionary elements extracted with CRAFT with the neuron basis using Eq. 15. We find that ResNet50 exhibits lower average axis-alignment than VGG16 (0.40 *vs.* 0.45, $z = 6.01, p < .001$, Mann–Whitney U test), indicating that its dictionary elements are more distributed across neurons. This finding is consistent with a greater degree of representational sharing across neurons in ResNet50, a pattern commonly linked to superposition. Consequently, neuron-level analyses may obscure structure that only becomes visible in a dictionary-based basis, potentially affecting interpretations of model interpretability.

ResNet50 is more interpretable than VGG16. Within the Rashomon set of similarly accurate models—*e.g.*, ResNet50 achieves 76% accuracy compared to 72% for VGG16 on ImageNet, some models are expected to be more interpretable than others. A common assumption in mechanistic interpretability is that sparser representations should be more interpretable.

Based on the theoretical intuition that neural networks encode patterns in superposition and that interpretability improves when dictionary elements are sparse, we hypothesized that ResNet50 would exhibit greater interpretability than VGG16 in the dictionary-based condition. Our results corroborate this intuition: ResNet50 achieves significantly higher interpretability than VGG16 ($85.1\% \pm 1.4$ *vs.* $81.5\% \pm 1.5$, $z = 2.14, p = 0.016$) in the dictionary-based condition. These findings contradict prior work [205], which reported higher interpretability for VGG16 with 89.27% *vs.* 87.40% for ResNet50, underscoring the critical role of the representational basis when evaluating model interpretability.

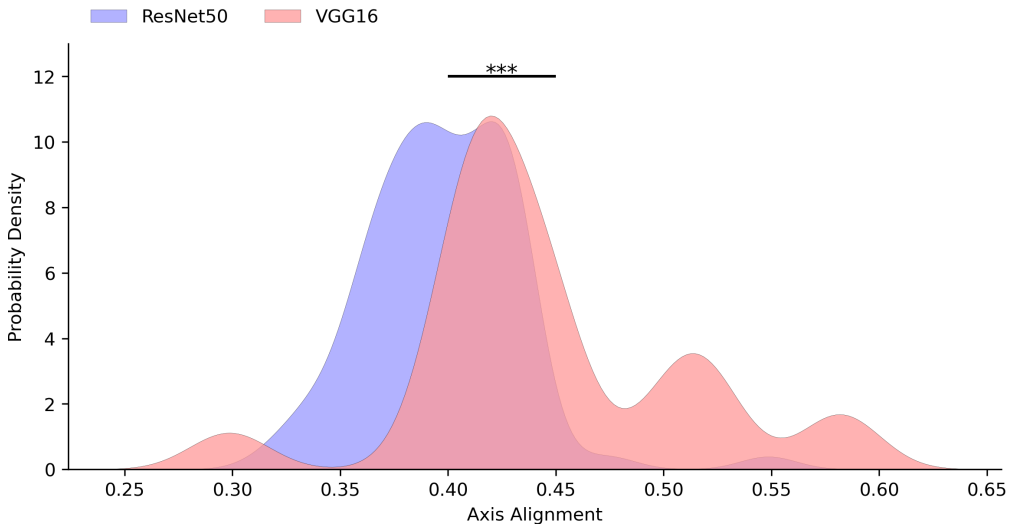


Figure 6. Axis-alignment of dictionary elements. We measure the axis-alignment of dictionary elements recovered by CRAFT for both models using the sparsity measure of [90] (see Eq. 15). Higher values indicate stronger alignment with individual neuron axes, *i.e.*, the recovered dictionary elements can be well approximated by single neurons, whereas lower values indicate more distributed representations across neurons. We find that ResNet50 exhibits significantly lower axis-alignment than VGG16: 0.40 *vs.* 0.45, $z = 6.01, p < .001$ (Mann Whitney U test). This difference is consistent with greater representational distributedness in ResNet50 a pattern commonly linked to superposition.

The choice of basis matters when comparing model interpretability. Fig.4b depicts the average interpretability of ResNet50 and VGG16 for each representational basis (neuron-based vs. dictionary-based). While ResNet50 achieves higher interpretability than VGG16 in the dictionary-based condition, this advantage disappears when interpretability is assessed using neuron-based representations ($z = 0.56, p = 0.71$). These results suggest that a model's apparent interpretability critically depends on the choice of representational basis. Importantly, they challenge prior claims about model interpretability [204], and highlight the need to consider dictionary-based representations when evaluating and comparing models.

3.5 Conclusion

In this work, we have investigated the suitability of neuron-based and dictionary-based representations as a basis for interpretability in two convolutional neural networks, ResNet50 and VGG16. Across three large-scale psychophysics experiments, we find converging evidence that the choice of representational basis significantly influences human interpretability: dictionary-based representations are consistently more interpretable than neuron-based representations, with the advantage increasing in deeper layers. Furthermore, we find models rely no less on dictionary elements than on neuron axes to make their decisions.

We also observe that the degree of representational distributedness in models, as measured by axis-alignment of dictionary elements, influences their apparent interpretability. ResNet50

exhibits lower axis-alignment than VGG16, indicating more distributed representations. This leads to comparable interpretability when assessed using neuron-based representations but superior interpretability when evaluated using dictionary-based representations. This finding is particularly notable given that ResNet50 is the model with higher classification accuracy.

These findings demonstrate that the choice of representational basis matters when comparing models, suggesting a need to reinterpret prior results. Overall, our results highlight that dictionary-based representations not only constitute a superior basis for interpretability but also that comparing models using neuron-based representations alone can lead to misleading conclusions.

Limitations. Our study is not without limitations. First, the methodology proposed by [17] and followed by [204] uses the entire ImageNet validation set to cover a broad range of stimuli and thereby increase the likelihood of identifying images representative of a neuron's selectivity. While this motivation is sound, neurons selective to class-specific patterns (*e.g.*, fish scales) will often be maximally activated by images from the corresponding classes (*e.g.*, fish) and minimally activated by images from unrelated classes (*e.g.*, dogs). Participants may therefore solve the task using class semantics rather than the neuron's specific selectivity. Experiment II represents an initial attempt to mitigate this confound, but manual inspection suggests only partial success. Even after matching semantic labels, low-level statistics such as dominant color or average texture may correlate with class identity and provide additional shortcuts, while expansion through the WordNet hierarchy occasionally matched related but non-identical categories. More broadly, these residual cues expose a limitation of the forced-choice design itself: participants can succeed by identifying any difference between the candidate images, rather than by forming a positive understanding of what the feature represents. Although the protocol remains informative for the controlled comparison of representational bases pursued here, it appears less suited to placing different models on a common scale of interpretability.

Second, this work focused on whether individual neurons reflect mixtures of pattern detectors in superposition. This framing enabled a fair one-to-one comparison between neuron-based and dictionary-based elements, providing evidence that dictionary learning can yield a more interpretable representational basis. It does not, however, establish the interpretability of dictionary elements recovered directly from the full activation space of a layer, which is the setting required to characterize a model's representation more broadly.

Finally, our experiments were limited to two convolutional models. This was sufficient to demonstrate that the choice of representational basis can alter the outcome of an interpretability comparison, but not to determine how interpretability varies across the modern vision models now in widespread use.

Taken together, these limitations inform the continuation of this work in Chapter 4, which extends the analysis from a controlled comparison of representational bases to the measurement and comparison of full-layer representations across widely used vision models.

Chapter 4

How to measure and compare interpretability across models?

Chapter 3 established dictionary-based features as a more appropriate unit for studying interpretability than individual neurons. It also raised concerns about whether existing forced-choice protocols provide a sufficiently reliable basis for comparing models. This chapter takes up that measurement question directly, asking how the interpretability of dictionary-based features can be assessed fairly across models.

4.1 Introduction

The dominant paradigm in modern computer vision is to pretrain large foundation models and fine-tune them for specific tasks. Built on the vision transformer [46], models like CLIP [157], SigLIP [199], DINOv2 [152], and DINOv3 [170] already serve as general-purpose visual backbones across an ever-expanding range of downstream tasks, including classification [30], segmentation [168], object tracking [54, 191], and robotic perception [105]. The same models are increasingly deployed in high-stakes settings—clinical decision support [11, 37, 98, 136], autonomous driving [68, 97]—where a user often needs to understand what a model is actually responding to. How interpretable, then, are the features that today's leading vision models have learned?

Surprisingly little work compares vision models on this dimension, because existing evaluation protocols cannot place different models on a common, reliable scale of feature interpretability (Section 4.2).

To fill this gap, we introduce a model evaluation framework comprising two psychophysics protocols that quantify two complementary dimensions of feature understanding: feature *localizability* — can an observer predict *where* a feature fires on a novel image? — and feature *semanticity* — can an observer accurately describe *what* a feature represents? Together they form a comprehensive evaluation (see Fig. 7, bottom): a feature can fire in a predictable location yet defy verbal description, or be easy to name yet hard to point to. The proposed framework has three properties that support a fair cross-model comparison: (i) it replaces the polysemantic neuron basis used in prior work with monosemantic directions recovered via sparse autoencoders [35, 69], so each tested unit corresponds to a single concept; (ii) it evaluates *functionally analogous* features across models by anchoring feature selection to a shared set of input images (Fig. 7, top), rather than sampling features randomly from

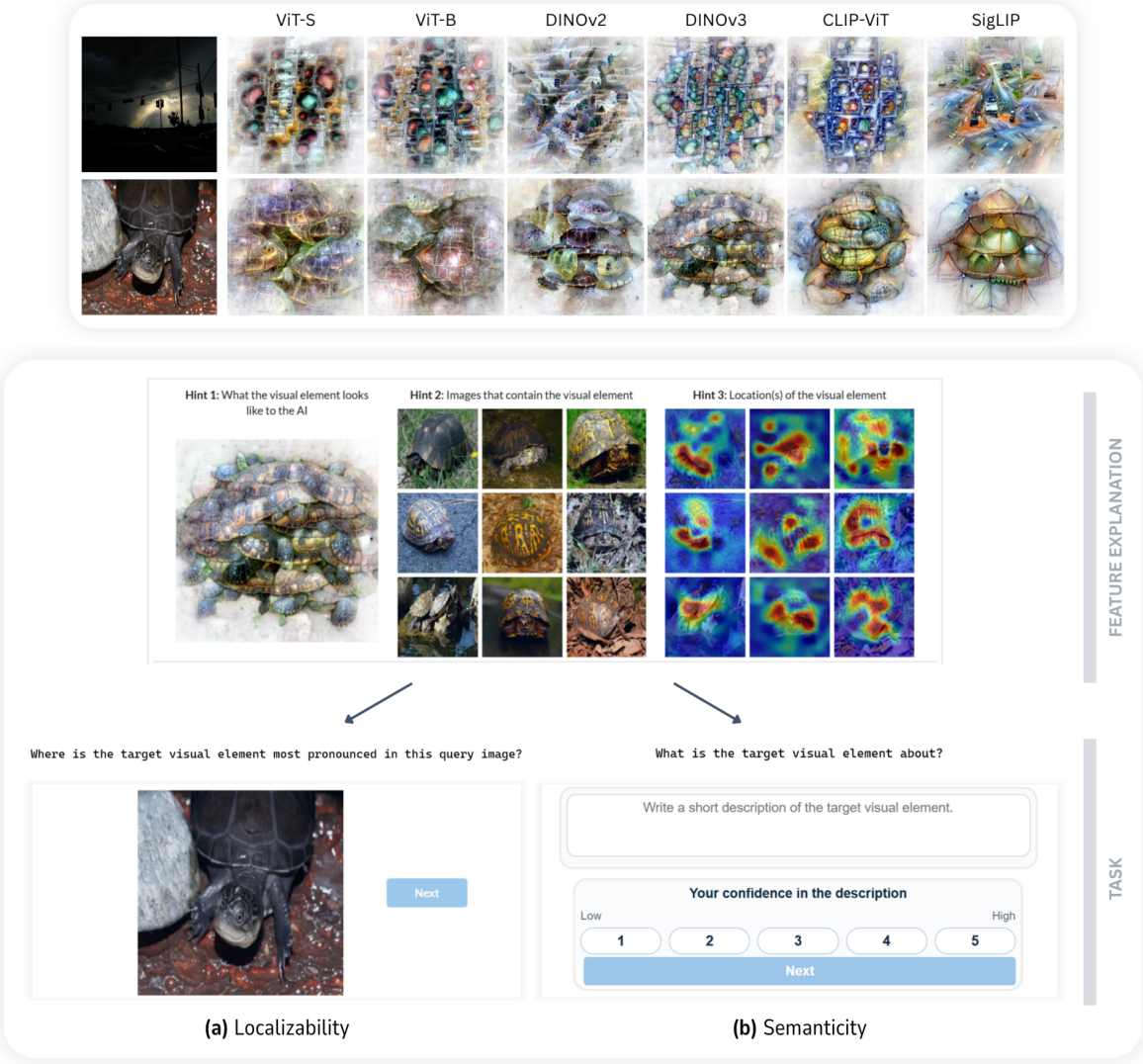


Figure 7. Models trained under different objectives learn different internal representations and, consequently, different features. Which of these features are more interpretable to humans remains an open question. **Top:** the most important feature recovered by each of the six models we study, for the prediction of *traffic light* (first row) and *mud turtle* (second row). **Bottom:** Our framework, illustrated for DINOv3. Given a feature and its visual explanation—a synthetic visualization, highly activating images, and corresponding heatmaps showing where the feature fires (top panels)—participants either click where they think the feature fires on a novel image (left, *localizability*) or describe what the feature represents in free text (right, *semanticity*).

each model; and (iii) it uses calibrated scores to a common chance level across features and models, so scores are directly comparable. Furthermore, the evaluation is scaled to ∼6,000 features across six models at a fraction of the cost of previous protocols [204].

We apply the framework to six vision transformers—two supervised ViTs and four foundation models (DINOv2 [152], DINOv3 [170], CLIP [157], SigLIP [199])—and collect more than 15,000 behavioral responses, of which the 13,400 from the 377 participants who passed our pre-specified quality checks enter the analyzes. The headline finding is sobering: every

foundation model we tested produces features that are *less* interpretable than those of the supervised baselines that came before, and downstream task performance does not predict the gap. Two representational properties do: feature *locality* and *coarse-grained* semantic alignment with human similarity judgments. DINOv3—whose training objective explicitly promotes local features—is the encouraging counterexample, suggesting that the gap is not architectural destiny.

Our contributions are: **(1)** a framework for measuring and comparing the human interpretability of vision models, with two complementary psychophysics protocols—*localizability* and *semanticity*; **(2)** a human study across six vision transformers showing that every foundation model we tested generally produces features that are *less* interpretable than those of the supervised baselines that came before, yet this gap is not a consequence of their greater capability: downstream performance does not explain it; and **(3)** the identification of two representational properties that predict interpretability—*locality* of feature activations and *coarse-grained* semantic alignment with humans.

4.2 Related Work

Sparse autoencoders and monosemanticity. Under the superposition hypothesis [50], networks encode more features than they have neurons by superimposing concepts onto shared activation channels, producing polysemantic neurons. Sparse autoencoders (SAEs) recover a dictionary of monosemantic directions and have emerged as the dominant remedy [35, 59, 69, 185]. We extract features via SAEs to avoid the polysemy that complicates neuron-level analysis.

Evaluation of model interpretability. Bau, Zhou, Khosla, et al. [13] introduced Network Dissection, to our knowledge the first framework to compare interpretability across models (architectures and training regimes). They score every convolutional unit from four CNNs against pixel-level masks from a set of human-labeled concepts (Broden), and models are ranked by the number of units that align with a concept above a given threshold. The approach is bounded by Broden's vocabulary, restricted to the neuron basis, and conflates concept alignment with interpretability—which, as Section 4.4.2 shows, do not necessarily coincide.

Borowski, Zimmermann, Schepers, et al. [17] pioneered psychophysics-based evaluation of DNN interpretability: participants viewed maximally and minimally activating images for a given neuron, then identified which of two novel query images also strongly activated it. Zimmermann, Klein, and Brendel [204] scaled this protocol across nine vision and vision-language architectures (80 features per model, 720 in total—a small fraction of the thousands a typical model encodes) and concluded that more capable models are not necessarily more interpretable. Zimmermann, Klindt, and Brendel [205] later automated the evaluation by replacing human judgments with pairwise perceptual similarity. Their metric correlates with human results, but cross-model differences arise primarily from a tail of highly uninterpretable units linked to superposition. Applying a similar protocol to CNNs, Colin, Goetschalckx, Fel, et al. [32] showed that reliance on a neuron basis can mask cross-model differences because the impact of superposition is itself model-specific. Two methodological limitations, therefore, prevent fair cross-model comparison in this lineage: exclusive reliance on the polysemantic neuron basis and a forced-choice task whose process-of-elimination

shortcuts produce model-specific chance levels. Our framework addresses both limitation and extends model coverage to several foundation models that postdate prior studies.

4.3 Method

4.3.1 Current protocols are ill-suited for model interpretability.

Prior protocols [17, 204] frame the task as a contrastive forced-choice: participants view maximally and minimally activating images for a feature, then select which of two novel query images also strongly activates it (see Fig. 16 in Appendix B.3 for an example trial). While intuitive, this design has a fundamental confound: participants can identify the correct answer by understanding what the feature is not about—*e.g.*, not a corn detector (Fig. 16)—rather than genuinely interpreting the feature itself.

To test whether this confounding effect is consequential, we performed a control experiment on two models from Zimmermann, Klein, and Brendel [204] (ResNet-50 and ViT-B32). We ran two versions of a given trial: the original version (Fig. 16) and a variation where we replaced the maximally activating images with images ranking 25,000 out of 50,000 in each feature's activation distribution, thus carrying little to no information about the feature's visual content (Fig. 17). In this case, participants were expected to perform at chance (50%). Participants nevertheless performed significantly above the 50% chance level for both models. For ResNet-50, accuracy was 60.0% across 2,480 trials from 62 participants (binomial test, 95% CI $[58.1\%, 62.0\%]$, $p < .001$). For ViT-B/32, accuracy was 53.6% across 1,840 trials from 46 participants (95% CI $[51.3\%, 55.9\%]$, $p = .001$). The protocol therefore produces significantly inflated and *model-dependent baselines* even when the maximally activating images carry little information about the feature.

This model-specific baseline inflation makes raw accuracy scores incomparable across architectures. Zimmermann, Klein, and Brendel [204] reported ResNet-50 as more interpretable than ViT-B32 (83% vs. 80%). However, once each score is expressed relative to its model-specific baseline, the ranking *reverses*: ViT-B32 is 27% above its baseline while ResNet-50 is only 23% above its own. Any protocol that does not account for this artifact risks drawing erroneous conclusions.

4.3.2 Framing interpretability as identification.

Overview. We address the limitations described above with a framework comprising of two complementary protocols that operationalize interpretability: *localizability*, *i.e.*, can a participant predict *where* a feature fires on a novel image?; and *semanticity*, *i.e.*, can a participant accurately describe *what* a feature represents? Both protocols share the same feature extraction pipeline (Fig. 18 in Appendix B.5) and the same visual explanations presented to participants (Fig. 7). They differ only in the task and scoring function. Because there is no distractor to exploit, participants must rely on genuine understanding of the feature's visual content to perform. Localizability scores are mapped to a common chance-anchored scale (Section 4.3.2), enabling direct comparison across features, models and protocols. Semanticity requires no chance-level correction: images and descriptions vary across models, but they are scored in a shared embedding space, making cross-model comparisons fair.

By evaluating the interpretability of features, we aim to derive a more general score representative of model interpretability. Whereas Zimmermann, Klein, and Brendel [204] studied features across multiple layers, our focus on foundation models leads us, moving forward, to examine only the last latent representation of each model, *i.e.*, the output of the penultimate layer for supervised models, the output of the last layer for DINO models, and the output of the vision encoder for VLMs.

Feature extraction. For each model, we extract features from its vision encoder. Because individual neurons in vision models are often polysemantic [50], we use sparse autoencoders (SAEs) to recover a dictionary of monosemantic directions. TopK SAEs [69] (K=38) are trained on activations from the ImageNet training split (≈1.28M images), using all patch tokens per image (256 for patch size 14 in DINOv2, 196 for patch size 16 in all other models). Each SAE uses an expansion factor of ×10, yielding 3,840 features for ViT-S and 7,680 features from all the other models. Reconstruction quality, dead-feature rates, and stability across independent SAE seeds are reported in Appendix B.6.

Feature relevance and linear probing. For each model, we train a linear classification head on ImageNet [39] on top of the frozen backbone, from which we derive (i) per-feature importance scores that quantify each feature's contribution to the classifier's predictions, computed as Gradient input [169], shown to be particularly faithful in latent space [57]; and (ii) a downstream accuracy baseline used in Section 4.4.1.

Feature selection. Given that current computer vision models are expected to encoded thousands of features, we aim to scale the number of features evaluated. Building on Zimmermann, Klein, and Brendel [204], we first ran a pilot with 80 features and 10 trials per feature to obtain reliable per-feature interpretability estimates and a model interpretability score. We then evaluated the depth-vs-breadth tradeoff: we found that the number of features evaluated is more important than the number of repetitions per feature to obtain stable model-level estimates (see Appendix B.4). We therefore adopted this design in the main experiment.

To obtain a representative subset of features, we aim to select features that are both diverse and relevant to the model's decisions. To this end, we randomly sample one image for each of the 559 superordinate classes underlying the 1,000 ImageNet labels, yielding a fixed set of 559 images. For each image, we select the three visual features that rank as the most important for the model's prediction on that image. During selection, we exclude features that activate on more than 10% of patch tokens, following Sun, Stolfo, Engels, et al. [180]—in practice, this removes about 7-90 candidate features per model. This procedure yields a total of $559 \times 3 = 1{,}677$ trials per model, but because the same feature can be selected for multiple images, this results in 500–1,300 unique features evaluated per model (see Appendix B.7).

Feature explanation. Both protocols present participants with the same visual explanation of a feature (Fig. 7- top): (a) a synthetic feature visualization synthesized via MACO [56] that maximally activates the feature (left panel in Fig. 7); (b) nine highly activating natural images (middle panel in Fig. 7); and (c) their corresponding RISE heatmaps [154] indicating where the feature fires on each image (right panel in Fig. 7). Participants use these three sources of information to form an understanding of the feature before solving the task.

Localizability (*where*). Each *localizability* trial additionally presents (d) a novel query image (bottom, left in Fig. 7) and participants are asked to click on the region of the image where they believe the feature appears. Each click is scored against the RISE heatmap of the query image (smoothed with a Gaussian filter). Since an attribution map is only a proxy ground truth, we test whether the resulting comparison depends on this choice by rescoring the original human clicks with HSIC [146], an independent black-box attribution method. The broad model ordering is preserved, indicating that the results presented below are not specific to RISE; full results are reported in Appendix B.9.1.

Let v be the heatmap value at the clicked location and $p = \widehat{F}(v)$ its empirical cumulative distribution function (ECDF) over the heatmap pixels. Because heatmap distributions vary across features (*e.g.*, sparse vs. diffuse activations), raw percentiles are not directly comparable across features or models. We therefore introduce a *chance-anchored normalization* that pins chance to $s = 0.5$. Writing $p_\mu = \widehat{F}(\mu)$ for the percentile rank of the mean activation μ, we map p_μ to 0.5 and the extremes to 0 and 1:

$$s(v) = \begin{cases} 0.5 - 0.5 \cdot \dfrac{p_\mu - p}{p_\mu}, & \text{if } p < p_\mu, \\ 0.5 + 0.5 \cdot \dfrac{p - p_\mu}{1 - p_\mu}, & \text{otherwise.} \end{cases}$$

By pinning chance ($s = 0.5$) to the mean activation, the score carries the same interpretation across features and models — a click above the mean scores above chance, below the mean scores below—regardless of whether the feature activates focally or diffusely. We refer to this score as the *localizability* score.

Semanticity (*what*). Each *semanticity* trial asks participants to (d) write a short free-text description of the feature (bottom right in Fig. 7) and rate their confidence on a 5-point Likert scale. We evaluate two complementary properties. Description *plausibility* measures whether the proposed interpretation is compatible with the feature's visual evidence, while descriptions *agreement* measures whether independent participants converge on a similar interpretation. *Confidence* helps distinguish ambiguity, reflected by low agreement and low confidence, from confident misinterpretation, reflected by low plausibility despite high confidence.

To measure description *plausibility*, we extract a 96×96 crop around the peak RISE activation from each of the nine most-activating images and encode these crops and the participant's description using CLIP ViT-B/16 [157]. The score is the mean cosine similarity between the text embedding and the nine images embeddings. We validate that the results are stable across six alternative vision-language encoders ($\rho = .74$–$.94$, all $p < .001$) and when using full images rather than crops; see Appendix B.10.

To measure descriptions *agreement*, we embed all descriptions of features with at least two responses using a popular sentence-encoder, MPNet [175], and compute their mean pairwise cosine similarity. High agreement indicates a clear, shared interpretation, whereas low agreement suggests that the feature is ambiguous.

Because agreement requires multiple descriptions per feature, we evaluate a representative subset. For each model, we divide features into localizability deciles and sample uniformly within each bin, yielding approximately 30 features per decile and 300 per model, leading to approximately 900 *semanticity* trials per model. Due to the CLIP modality gap, *plausibility*

scores range from 0.13 to 0.35, with a chance baseline of 0.19, defined as the average similarity between image and text embeddings drawn from different features.

Participants. Across both experiments, we recruited about 440 participants via Prolific[7]. All were native English speakers without reported visual impairments and completed the study on a laptop or desktop. Each provided informed consent and received $2.95 (∼$16/hr; 10–13 minutes). The protocol was approved by the IRB of an author-affiliated institution. We analyze about 13,400 behavioral responses from the 377 participants who (i) passed at least 4 of 6 practice trials, (ii) answered all 4 attentiveness catch trials correctly, and (iii) completed the experiment within 3 SDs of the mean duration. We define one *behavioral response* as one participant's answer to one trial: a click in the localizability experiment or a free-text description, together with its confidence rating, in the semanticity experiment.

4.4 Results

We evaluated six vision transformers: two supervised baselines (`ViT-S/16` and `ViT-B/16`) and four foundation models—`DINOv2 ViT-B/14`, `DINOv3 ViT-B/16`, `CLIP ViT-B/16`, and `SigLIP ViT-B/16`.

Table 2. Localizability, Semanticity—plausibility and agreement—, and confidence scores across six vision transformers. Supervised models consistently outperform foundation models on both dimensions. Best result per row is in bold, second best is underlined.

	Supervised		**Foundation**			
	ViT-S	ViT-B	DINOv2	DINOv3	CLIP	SigLIP
Localizability (↑)	<u>80.3</u>	**86**	71	80	79.7	71.4
Semanticity (plausibility) (↑)	**0.274**	<u>0.273</u>	0.259	0.260	0.266	0.253
Semanticity (agreement) (↑)	<u>0.49</u>	**0.52**	0.43	0.46	0.44	0.44
Confidence (↑)	3.43	<u>3.61</u>	**3.68**	3.41	3.51	3.38

Foundation models are less interpretable than supervised ones. We measure two dimensions of interpretability: *localizability* and *semanticity*, the latter quantified through description *plausibility* (P) and *agreement* (A)[8], as summarized in Table 2. Scores differ significantly across models (L: Kruskal–Wallis, $H(5) = 118.40$, $p < .001$; P: $H(5) = 295.36$, $p < .001$; A: $H(5) = 30.3$, $p < .001$). As seen in the table, foundation models are consistently less interpretable than their supervised counterparts (Dunn's test and Tukey HSD's test, $p < .001$ vs. ViT-B), regardless of whether they were trained with language supervision (CLIP, SigLIP) or pure self-supervised objectives (DINOv2, DINOv3). Moreover, while we expect *localizability* and *semanticity* to quantify two different dimensions of interpretability, in practice we find that they are highly correlated (L-P: $r = 0.84$, $p = 0.036$; L-A: $r = 0.82$, $p = 0.046$; P-A: $r = 0.80$, $p = 0.059$). Qualitative examples spanning the four quadrants of *localizability* and *plausibility* are shown in Appendix Fig. 23.

[7] www.prolific.com

[8] We report the median per model because the scores are not normally distributed; see Appendix B.11.

Models can feel more interpretable than they are. Average confidence is fairly homogeneous across models ($3.38 - 3.68$, Table 2, bottom row), but its alignment with interpretability is not (P, $\rho = 0.34$; A, $\rho = 0.30$). While DINOv2 elicits the highest confidence and ranks near the bottom on interpretability (last on *localizability* and *agreement*, second-to-last on *plausibility*), SigLIP performs poorly on all fronts. Based on these results, a model can feel more interpretable than it is—see Appendix Fig. 22 for examples of features that are confidently but inaccurately named (and vice versa), a caution worth keeping in mind given that DINOv2 features have been treated as a reference for interpretable representations [48].

4.4.1 Alignment with performance

A natural question is whether interpretability and capability are linked: models with richer, more semantically structured representations might be both more capable on downstream tasks and more legible to humans. We measured the Spearman correlation between each model's interpretability scores and its performance on three vision tasks: ImageNet-1k classification, semantic segmentation, and perceptual grouping. We report those results in Fig. 8.

ImageNet-1k classification. Foundation models lead on ImageNet top-1 accuracy but trail on interpretability. Across the six models, neither *localizability* ($\rho = -0.48$, $p = 0.333$) nor *semanticity* (P: $\rho = -0.60$, $p = 0.209$; A: $\rho = -0.48$, $p = 0.334$) correlate significantly with classification. DINOv2 and SigLIP match CLIP on accuracy yet score roughly 8 points lower in *localizability*.

Semantic segmentation. Segmentation performance (mIoU) on ADE20K[202]—using a linear probe[9]—does not correlate significantly with either *localizability* ($\rho = -0.54$, $p = 0.273$) or with *semanticity* (P: $\rho = -0.77$, $p = 0.072$; A: $\rho = -0.69$, $p = 0.133$).

Perceptual grouping. On a perceptual grouping benchmark[10] [117]—a more constrained form of segmentation that probes the models' ability to bind individual object instances—we again find no significant correlation with either *localizability* ($\rho = 0.12$, $p = 0.820$) nor *semanticity* (P: $\rho = 0.09$, $p = 0.867$; A: $\rho = -0.12$, $p = 0.817$).

Across all three benchmarks, task performance does not predict interpretability. This dissociation suggests that the representational properties that make a model useful for downstream tasks are largely orthogonal to those that make its features interpretable to humans.

Locality of the representation. Beyond task performance, we ask whether *locality*, *i.e.*, the degree to which a feature's activations concentrate on confined image regions, predicts interpretability: a focal activation supplies a concrete spatial anchor (an object part, a texture), while a diffuse one leaves the participant to guess what the activating regions share.

[9] We follow the DINOv3 methodology but force all models to an input size of 224×224, since supervised ViTs from `timm` cannot take flexible input sizes. The same pipeline at 512×512 reproduces DINOv3 results.

[10] Perceptual grouping refers to the object-instance binding benchmark of Li, Salehi, Ungar, et al. [117]: the ability to determine which patches or regions belong to the same object instance, measured through the benchmark's binding-accuracy protocol.

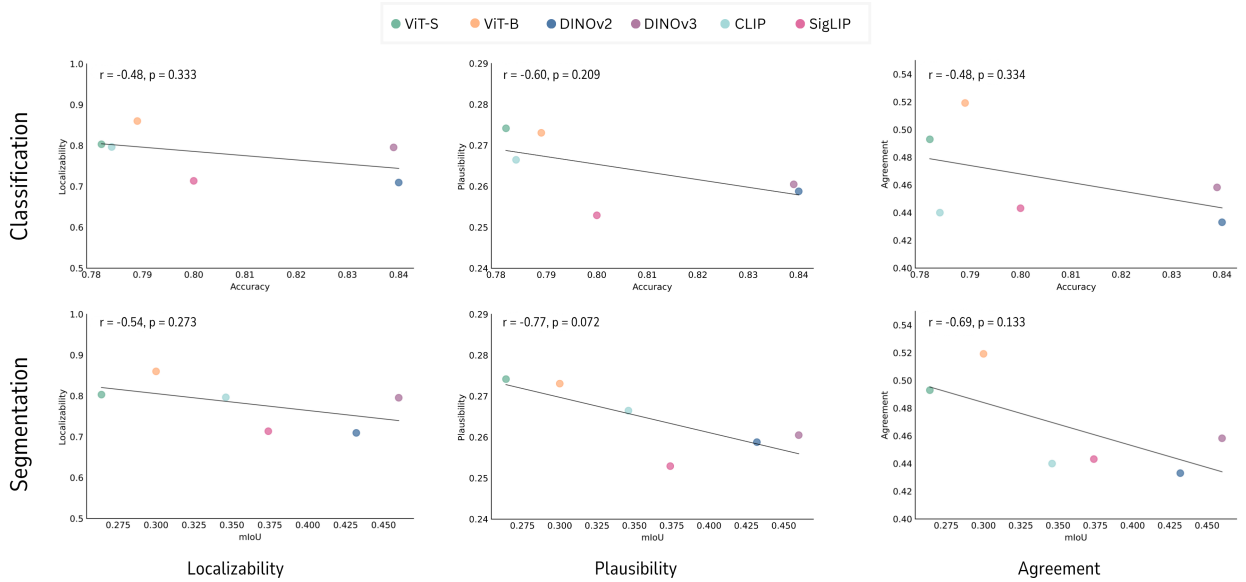


Figure 8. Interpretability is uncorrelated with downstream task performance. Each panel plots an interpretability score against a capability benchmark across the six models. Top row: Classification (ImageNet-1k top-1 accuracy [39]); bottom row: Segmentation (ADE20K semantic segmentation [202]). Columns: *Localizability*, *Plausibility*, and *Agreement*. Spearman ρ and p-values are shown in each panel. Correlations are non-significant for all three interpretability measures across both benchmarks, suggesting that interpretability and downstream task performance capture largely orthogonal properties of a visual representation.

We quantify locality with the Hoyer metric applied to each feature's RISE heatmaps [90]:

$$H(\mathbf{x}) = \frac{(\sqrt{n} - \|\mathbf{x}\|_1/\|\mathbf{x}\|_2)}{(\sqrt{n} - 1)}$$

with n the number of pixels, and where $H = 0$ corresponds to uniform activation across the image and $H = 1$ to all mass on a single pixel. We average H across each feature's 9 most activating images and then across the model´s features. In contrast to task performance, locality correlates strongly with *localizability* ($\rho = 0.91$, $p = 0.011$; Fig. 9, top left) and *semanticity* (P: $\rho = 0.99$, $p < .001$; Fig. 9, top middle; A: $\rho = 0.82$, $p = 0.048$; Fig. 9, top right). See Appendix B.12 for additional results.

4.4.2 Alignment with human perception

Next, we consider human alignment, *i.e.*, the degree to which a model aligns with some properties of human perception, as alignment has been shown to improve representation quality over a wide range of downstream tasks [139, 140, 178, 181]. Because not all measures of alignment capture the same properties of human perception [4], we study 3 kinds of alignment: alignment with human (a) visual strategies; (b) coarse-grained similarity judgments (semantic similarity); and (c) fine-grained similarity judgments (lower-level visual similarity).

Alignment with human visual strategies. Strategy alignment—the overlap between model-driven and human-attended image regions [61, 120]—does not correlate significantly

with *localizability* ($\rho = 0.14$, $p = 0.798$) or *semanticity* (P: $\rho = 0.21$, $p = 0.696$; A: $\rho = 0.30$, $p = 0.564$).

Alignment with human similarity judgments. Human similarity judgments operate at multiple levels of abstraction [139]. *Fine-grained* similarity captures within-category distinctions (*e.g.*, how similar two butterfly species are to each other), while *coarse-grained* similarity captures broad semantic structure across categories (*e.g.*, how similar a buffalo is to a spider relative to grass). We measure alignment at the different levels of similarity using three datasets: THINGS similarity [84, 85], Levels [139], and NIGHTS [66, 181].

Alignment with *fine-grained* similarity does not correlate with neither *localizability* (NIGHTS: $\rho = -0.41$, $p = 0.418$; Levels: $\rho = 0.20$, $p = 0.705$), *plausibility* (NIGHTS: $\rho = -0.47$, $p = 0.347$; Levels:$\rho = 0.26$, $p = 0.617$) or *agreement* (NIGHTS: $\rho = -0.33$, $p = 0.523$; Levels:$\rho = -0.06$, $p = 0.915$).

In contrast, alignment with *coarse-grained* similarity is generally lower and shows a strong correlation with interpretability (Fig. 9, bottom), which is significant for the THINGS similarity dataset for both *localizability* and *plausibility*: *localizability* (THINGS: $\rho = 0.84$, $p = 0.036$, Levels:$\rho = 0.71$, $p = 0.110$), *plausibility* (THINGS: $\rho = 0.85$, $p = 0.030$; Levels $\rho = 0.6$, $p = 0.205$), and *agreement* (THINGS: $\rho = 0.85$, $p = 0.03$; Levels $\rho = 0.6$, $p = 0.21$). See Appendix B.13 for additional results.

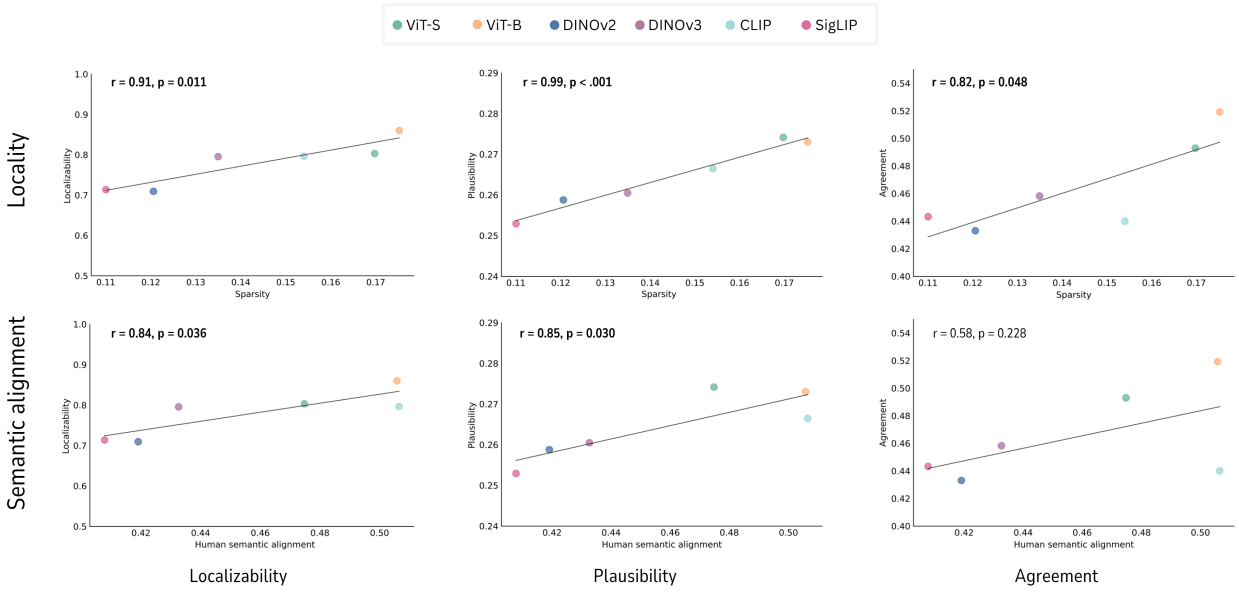


Figure 9. Locality and coarse-grained semantic alignment track interpretability. Each panel plots an interpretability score against a representational property across the six models. Top row: Locality ((mean Hoyer sparsity over feature heatmaps); bottom row: Human semantic alignment (coarse-grained alignment with human similarity judgments on THINGS [85]). Columns: *Localizability*, *Plausibility*, and *Agreement*. Spearman ρ and p-values per panel. Locality is the strongest predictor of interpretability, consistent with focal activations providing a clearer visual anchor for understanding a feature. Coarse-grained alignment also tracks interpretability, particularly for *localizability* and *plausibility*, suggesting that interpretable representations are not only local but also organized in ways that better reflect human categorical similarity judgments.

This dissociation between granularity levels suggests that what matters for interpretability

is not perceptual fidelity to fine visual detail, but rather the degree to which a model organizes its features around the same high-level semantic categories that structure human perception. A feature that groups concepts according to what humans consider broadly similar is more likely to be a feature that humans can understand.

4.5 Discussion

Two protocols, three measures, one interpretability. The two protocols differ in nearly every respect: one asks participants to point, the other to write; one is scored against a spatial heatmap, the other against a language or vision–language embedding. That they yield concordant rankings and surface the same predictors is unlikely to be a coincidence—we appear to be measuring a stable property of the representations rather than an artifact of any one task, and either protocol may suffice on its own in future work.

Foundation models are less interpretable than their supervised counterparts. The gap appears in both vision-only SSL models (DINOv2, DINOv3) and vision-language models (CLIP, SigLIP), so neither pretraining signal is the cause. The common factor is foundation-style pretraining itself. One reading is that these models reach—and on many benchmarks exceed—human performance in part by developing *superhuman* features: representations effective for the task but not the kind a human observer can readily make sense of. The very generality that makes foundation models capable may also push their representations away from a human-interpretable basis. This gap is easy to miss without direct evaluation—DINOv2, often treated as a reference for interpretable representations [48], elicits the highest rater confidence yet ranks among the least interpretable models on *localizability* and *semanticity*.

Interpretability is orthogonal to task performance. Across every benchmark we examined, capability and interpretability are uncorrelated. The relationship is neither a tradeoff nor a free lunch: how well a model classifies, segments, or groups says little about whether a human can interpret its features. The leaderboards by which the field currently judges progress are silent on interpretability, a property independently desirable in high-stakes settings (*e.g.*, clinical decision support or autonomous driving) where these same backbones are increasingly considered. Interpretability has to be designed for explicitly; it will not arrive as a byproduct of capability, and prior evidence suggests that scale alone does not guarantee it either [204].

Locality is the proximate signature of an interpretable representation. The locality of feature activations is the strongest predictor of interpretability we identify. Foundation models tend to develop diffuse features that fire across larger, less focused regions, blending local content with broader scene context. This may help on downstream tasks but leaves a human observer without a clear visual anchor. DINOv3 is the encouraging exception: one of the most capable models in our set and also among the most interpretable foundation models, a pattern consistent with its training objective explicitly promotes local features. This finding suggests that locality-aware objectives can partially closes the interpretability gap,

though whether they do so without capability cost requires additional direct experimental verification.

Coarse-grained semantic alignment, not perceptual fidelity, predicts interpretability. Of the alignment measures we considered, only coarse-grained alignment with human similarity judgments tracks interpretability. A model can attend to the same image regions humans do, yet still encode what it sees there in features humans cannot make sense of. Fine-grained perceptual alignment has been found to be uniformly high and largely shared across models [77]. Coarse-grained alignment, on the other hand, is consistently lower, leaving real room for improvement. Interpretability appears driven less by perceptual fidelity than by how a representation organizes the visual world at the categorical level. Training signals built from human odd-one-out judgments [85] seem like a promising direction to close part of the gap.

4.6 Conclusion

We introduced a framework for measuring and comparing the human interpretability of vision models, built around two complementary psychophysics protocols: *localizability* (where a feature fires) and *semanticity* (what it represents). Across six vision transformers and more than 15,000 behavioral responses, foundation models, whether trained with language supervision or self-supervised objectives, were consistently *less* interpretable than their supervised counterparts—a surprising result for models that have surpassed those same baselines on most other axes. The gap is not a capability tradeoff: interpretability is uncorrelated with downstream task performance on every benchmark we examined, and the two structurally different protocols agree on this conclusion and on the predictors that drive it. Two properties track interpretability: the *locality* of a feature's activations and *coarse-grained* semantic alignment with human similarity judgments. Together, they point to a concrete recipe for building more interpretable vision models—training objectives that promote local feature activations and align representations with the coarse semantic structure of human perception [139]—and DINOv3, whose objective explicitly promotes locality, suggests that capability and interpretability need not trade off when locality is an explicit training objective.

Limitations We acknowledge several limitations. First, our model-level analyses span only six architectures, which limits statistical power and makes it difficult to disentangle correlated properties of representations (*e.g.*, sparsity and coarse-grained alignment may co-vary across architectures). Second, almost all models are evaluated at ViT-B scale. This controls for scale across architectures, but whether the interpretability gap between supervised and foundation models persists or reverses at larger scales remains an open question. Third, we focus on a specific set of representational properties: downstream task performance, locality, and alignment with human similarity judgments. Other factors, such as training data or robustness, may also shape interpretability. Finally, on external vision-language models to score the *plausibility* of human descriptions. This is imperfect: using an external model introduces its own biases and cannot provide a direct ground-truth measure of whether a description captures the feature. However, directly testing free-form descriptions against

the evaluated vision foundation models is not currently possible, since these models do not provide a reliable way to judge whether a natural-language description matches one of their internal visual features. We therefore use external image–text similarity as a practical proxy, keeping the scorer fixed across all evaluated models so that its biases are held constant in the comparisons.

More fundamentally, the relationships identified in this chapter remain correlational. Locality and coarse-grained semantic alignment are both associated with interpretability across models, but those models differ in many other respects. The present comparison therefore cannot determine whether deliberately increasing either property would make a representation easier for humans to understand.

With the unit of analysis established in Chapter 3 and a framework for measuring interpretability developed here, the thesis now has the tools needed to evaluate the success of an intervention. Chapter 5 closes the loop by causally testing one of the candidate properties identified in this chapter: alignment with human semantic judgments.

Chapter 5

Can we engineer interpretability?

Having established where and how interpretability should be measured, we now turn to whether it can be improved. Chapter 4 found that alignment with human semantic judgments tracks interpretability across models, but could not determine whether this relationship is causal. This chapter tests that possibility directly by intervening on model representations and evaluating the resulting changes using the same behavioral framework.

5.1 Introduction

Understanding why and how deep neural networks produce the predictions they do has been a question since long before today's foundation models existed: early visualization and analysis work on opaque networks aimed precisely at opening up the "black box" [51, 198]. Models have grown enormously in scale and capability since then, but that growth has not been matched by a corresponding gain in interpretability. In fact, recent evidence points the other way, with larger, more capable models proving harder to interpret than their smaller predecessors [33, 204]. Nevertheless, these are the models increasingly used in high-risk scenarios, where opacity is not something that can simply be accepted. Working out how to make models interpretable is therefore a more pressing question now than it has ever been.

As described in Chapter 1, there are two approaches to achieve interpretable models: building transparency directly into the architecture, or explaining a fixed, already-trained representation post-hoc via attribution, feature visualization, and concept-based methods [103] (see Related Work for the by-design literature and its costs). Because the models that are deployed in high-stakes settings are typically general-purpose pretrained networks with no built-in transparency constraints, this thesis follows the post-hoc route. However, comparatively less attention has been given to a complementary question: can the pretrained models be modified so that their representations become easier for humans to understand?

Rather than improving the XAI tools used to inspect an opaque model, this approach seeks to reduce the opacity of the model being inspected.

As Chapter 1 discusses, one line of work pursues this intervention route through representation-wide changes in both NLP and computer vision (see Related Work). A related and increasingly active version of this idea aligns model representations with human perceptual judgments and has also been proposed as a lever on interpretability directly [61, 139, 181]. The evidence for that specific claim, however, is so far only correlational: across a range of pre-

trained models, alignment with human semantic judgments tracks interpretability without establishing which one drives the other [33].

In this chapter, we test this claim directly, moving beyond correlation to a controlled experiment. We take the five pretrained ViT-B backbones introduced in Chapter 4—three vision-only models and two vision-language models—and fine-tune each to better align its representational geometry with the semantic structure derived from human similarity judgments [84]. Adopting the same analysis pipeline, we then extract sparse features from each aligned model, re-evaluate their interpretability, and compare each model with its aligned counterpart. We assess whether improving alignment with human semantic representations leads to a corresponding increase in the interpretability of a model's internal features.

The results do not admit a simple yes-or-no conclusion. Though alignment consistently improves agreement with human semantic representations, it does so at a measurable cost to classification accuracy, and its effect on interpretability depends on the model's pretraining: it improves interpretability for vision-only backbones but reduces it for the vision-language models. We do not yet have a definitive explanation for this divergence. Instead, this chapter demonstrates that the observed correlation between semantic alignment and interpretability is genuine but conditional, refining rather than resolving our understanding of the mechanisms that link the two.

5.2 Related Work

Interpretability by architectural design. A body of work seeks to build interpretability directly into a model's architecture, constraining predictions to pass through human-interpretable intermediate representations rather than reconstructed after the fact [162]. Concept bottleneck models, for example, make predictions through a layer of human-defined concepts [108]; prototype-based networks instead classify inputs by comparing them to a learned set of representative examples [29]; and self-explaining neural networks incorporate a similar constraint directly into the prediction function [6]. B-cos networks follow a different strategy, replacing conventional linear layers with a transformation that promotes alignment between weights and inputs, allowing faithful attribution maps to emerge directly from the forward pass [16].

Despite their methodological differences, these approaches share a common principle: transparency is imposed through the model's architecture. Each prediction is explicitly grounded in a small set of human-interpretable units, making the resulting explanation faithful by construction rather than an approximation that is inferred post-hoc. Achieving this level of transparency, however, typically requires concept- or prototype-level annotations collected in advance, commits the model to a fixed explanatory vocabulary, and may reduce predictive performance or representational flexibility.

While the resulting models are commonly described as *interpretable-by-design*, we argue that the label *explainable-by-design* is more precise. Their primary guarantee is that every prediction can be accompanied by a faithful explanation expressed in pre-defined human-interpretable terms. Whether those explanations are sufficient to understand the model's internal representations or broader computational strategy remains a separate question.

Interpretability through explanation methods. A larger body of work instead treats the trained model as opaque and focuses on developing methods to explain its behavior

post-hoc. Attribution methods estimate the contribution of individual input features to a prediction or intermediate computation [58, 154, 158, 165, 169, 171, 172, 174, 182]. Feature visualization takes a different approach, characterizing a neuron, feature or direction in representation space by synthesizing or retrieving inputs that maximally activate it [51, 56, 80, 129, 141, 143, 144, 150]. Concept-based methods extend this idea by probing or decomposing representations with respect to human-defined or automatically discovered concepts rather than individual units [20, 21, 59, 60, 69, 72, 104, 200].

Unlike interpretable-by-design approaches, these methods are largely agnostic to the model architecture or training procedure, making them broadly applicable across existing systems. Their effectiveness, however, ultimately depends on the representations they seek to explain. If the model's internal features are poorly aligned with concepts that humans can readily understand, then even faithful explanation methods may yield explanations that are difficult to interpret. Improving the interpretability of the representations themselves, rather than only the methods used to analyze them, is the complementary problem addressed in this chapter.

Acting directly on the representation. A third line of work intervenes directly on the representation, independent of any specific explanation method. In NLP, neologism learning enables models to express concepts they already encode internally but lack the vocabulary to describe, treating the mismatch between human and machine language as something that can be explicitly reduced [87]. In vision, where representations are continuous rather than grounded in a discrete vocabulary, comparable interventions instead reshape the representation itself. For example, training for robustness to human-imperceptible perturbations has been shown to encourage models to learn features that are more visually recognizable to humans [190, 194].

A closely related literature asks whether models represent the world in a manner that resembles human perception. One line of work measures the correspondence between model representations and human similarity judgments, behavior, or neural activity [61, 138, 159, 179]. A second line moves beyond evaluation by using human judgments as a supervisory training signal to directly reshape a model's learned representations, yielding improvements in few-shot learning, robustness, and feature transfer across tasks [140, 178, 181].

Motivated in part by the observation that interpretability is ultimately constrained by human cognitive capacity [1, 111], several of these works have proposed that representations more closely aligned with human semantic organization should also be more interpretable [61, 139, 181]. To date, however, the evidence for this claim has remained correlational: across a range of pretrained models, representations that better match human semantic judgments also tend to exhibit greater interpretability [33]. This chapter turns that observation into a causal test by asking whether explicitly increasing alignment with human semantic judgments leads to a corresponding improvement in the interpretability of a model's learned features.

5.3 Methods

We next describe the methodology used to (i) align vision models with human semantic organization, (ii) evaluate whether this alignment changes model interpretability and task performance, and (iii) characterize how alignment reshapes learned representations. Specifically, we measure changes in interpretability using human behavioral judgments, assess

whether alignment alters task competence, quantify representational similarity across models, and analyze feature-level properties associated with interpretability, including locality and reliance on high-frequency image information.

Aligning models with human semantic organization. We align vision backbones with human semantic organization using odd-one-out judgments from THINGS [84]. We apply Low-Rank Adaptation (LoRA) [91] to fine-tune the attention projections of each model while keeping all remaining backbone parameters frozen. The alignment objective follows the contrastive triplet loss introduced by Muttenthaler, Linhardt, Dippel, et al. [140], which encourages the model's representation geometry to better reflect human judgments of visual similarity. During training, a temporary linear projection head is added to compute the objective and discarded afterward. All subsequent analyses—including sparse autoencoder training and interpretability evaluation—operate directly on the aligned backbone representations without additional task-specific heads. Full implementation and optimization details are provided in Appendix C.1.

Human interpretability evaluation We evaluate the interpretability of the features derived from the aligned models using the localizability task introduced in Chapter 4. Localizability measures whether a human observer can infer where a feature will activate in a novel image after inspecting examples of the feature's behavior.

Each feature is explained to participants through three complementary sources of evidence (Fig. 7- top): (a) a synthetic feature visualization generated using MACO [56] that maximally activates the feature (left panel in Fig. 7); (b) nine highly activating natural images (middle panel in Fig. 7); and (c) their corresponding RISE heatmaps [154] showing the spatial regions that drive feature activation (right panel in Fig. 7). Participants use these three sources of information to infer the feature's visual pattern and then predict its activation location in a previously unseen image. Predictions are compared with the feature's actual activation map, producing a score that reflects the extent to which a feature's behavior can be understood and generalized by humans.

Chapter 4 introduced localizability and nameability and found them to be strongly correlated. We therefore focus on localizability in this chapter, applying the same evaluation pipeline to both pretrained models (evaluated in Chapter 4) and their human-aligned counterparts, and comparing aggregate interpretability scores before and after alignment.

Task-performance evaluation. To quantify whether alignment affects conventional model competence, we train linear probes on frozen backbone representations for two downstream tasks: ImageNet-1k classification (top-1 validation accuracy) and ADE20K semantic segmentation (mean intersection-over-union). This isolates the effect of representation alignment from changes in task-specific optimization.

Representational similarity across models. To determine whether alignment makes different backbones converge toward more similar representations, we perform representational similarity analysis (RSA) [109]. For each model, we construct a representational dissimilarity matrix (RDM) over a fixed image subset using pairwise cosine distances between image representations. We then quantify similarity between models by computing the Spearman correlation between their RDMs. Pairwise correlations are compared before and

after alignment to measure whether alignment reduces representational differences across architectures.

Individual and shared sparse dictionaries. To characterize how alignment changes individual features, we learn sparse dictionaries over both pretrained and aligned representations. For each backbone, we independently train a TopK sparse autoencoder [20] on the pretrained and aligned representations (dictionary size 7680, expansion factor 10). In addition, we train a joint TopK sparse autoencoder across each pretrained/aligned pair (dictionary size 6144) using a translation loss that encourages the dictionary to provide a shared basis across both representations.

This shared dictionary enables the comparison of features across models without requiring independently learned dictionaries to discover identical features. Rather than allowing each SAE to represent features in an arbitrary basis, the joint dictionary identifies features that both representations can express, providing a common space for analyzing how alignment modifies individual features.

Choosing a basis for cross-model correspondence. Aggregate localizability scores reveal whether alignment changes interpretability at the model level, but they do not reveal which features change or how individual features evolve. To address this limitation, we identify correspondences between features in pretrained and aligned representations.

We initially attempted to match independently trained SAE features directly, but this produced few reliable correspondences because independently learned dictionaries do not share a common basis (Appendix C.2). We therefore use the jointly trained dictionary as an intermediate representation space for establishing feature correspondences.

Defining cross-model features matches. We match each individual-SAE features to its closest dimension in the shared dictionary using cosine similarity between decoder directions. A candidate correspondence is accepted when similarity exceeds 0.7, a value commonly used in prior work [96, 114, 153]; results at other thresholds are reported in Appendix C.5. We additionally require that the shared feature activates in the other representation. Otherwise, the feature is treated as lacking a correspondence. Features present only before alignment are categorized as *dropped*, while features appearing only after alignment are categorized as *introduced* (Appendix C.8).

For matched features, we compute a co-firing score that measures the extent to which the matched dimensions activate together relative to a shuffled null distribution. Because conclusions based on a fixed threshold are sensitive to the choice of cutoff—varying the threshold from 0.3 to 0.7 changes the ranking of persistence groups for two of the five backbones (Appendix C.3)—we analyze co-firing as a continuous measure.

We additionally characterize persistence patterns using directional precision, recall, and firing rates on each side of the alignment boundary. A Gaussian mixture model (GMM; $k = 3$, selected using silhouette score) over these quantities identifies the qualitative persistence regimes reported below. Following prior work on cross-model features correspondence [186], we use patch-level co-firing as our primary measure and report an image-level alternative in Appendix C.4.

Locality. For each informative feature, we compute a *locality* score based on the Hoyer sparsity of its RISE activation heatmap, averaged over images in which the feature is active. Higher scores indicate more spatially concentrated activation patterns. We examine locality both as a potential contributor to human interpretability and as a representational property that may change as a consequence of alignment.

High-frequency reliance. For each informative feature, we measure sensitivity to high-frequency image information by progressively blurring its top-nine activating images and tracking the retained peak activation relative to the original image. The area under this curve provides a summary of robustness to the removal of high-spatial-frequency content: larger values indicate reduced dependence on fine-scale information, whereas smaller values indicate greater reliance on high-frequency details.

Because sensitivity to natural image texture is strongly associated with high-frequency information and spatial-frequency sensitivity contributes to conventional shape–texture bias [118, 177], we interpret this measure as a proxy for texture reliance. However, blur also removes fine contours and other non-textural structures. Therefore, this metric does not directly measure texture bias or isolate shape–texture trade-offs. For simplicity, we refer to it throughout as the *texture-reliance proxy*.

These analyses allow us to evaluate both the effectiveness and the mechanisms of perceptual alignment. We first ask whether alignment improves human interpretability while preserving downstream task performance. We then examine whether alignment changes the geometry of the learned representations, before turning to feature-level analyses that characterize which features are preserved, introduced, or removed, and how these changes relate to locality and reliance on high-frequency image information. This progression moves from the behavioral effects of alignment to the representational mechanisms that may underlie them.

5.4 Results

We first establish whether alignment successfully shifts model representations toward human semantic organization and whether this shift comes at a cost to downstream performance. We then ask whether these representational changes translate into improved human interpretability. Finally, we investigate the mechanisms underlying successes and failures by examining changes in feature vocabulary, locality, feature persistence, and reliance on high-frequency image information.

All the experiments described in this section correspond to aligning five ViT-B/16 backbones: supervised ViT-B, DINOv2, DINOv3, CLIP ViT-B/16, and SigLIP.

5.4.1 Alignment reshapes representations toward human semantic organization

Alignment improves agreement with human semantic structure. Alignment fine-tuning substantially improves agreement with human semantic organization. Across all five backbones, fine-tuning increases THINGS odd-one-out accuracy and improves between-class alignment on the held-out Levels benchmark (Table 3, first and second rows). These changes

are specific to the targeted coarse semantic structure: alignment decreases both within-class and class-boundary alignment across all backbones (Table 3, third and fourth rows), indicating that the intervention selectively reshapes the representation geometry rather than uniformly increasing similarity.

Alignment drives convergence across architectures. Optimizing different backbones toward the same human similarity structure also makes their representations more similar. Representational similarity increases for all ten model pairs after fine-tuning, with a mean increase of $\Delta\rho = +0.14$ (95% CI $[0.11, 0.17]$; Wilcoxon signed-rank test, $W = 0$, $p = 0.002$). Thus, optimizing different backbone models toward the same human similarity structure produces not only the intended semantic reorganization, but also a broader convergence toward more similar internal representations across models.

Table 3. Alignment scores before and after fine-tuning on THINGS and Levels. Values are percentages, reported as baseline → aligned.

	Metric	ViT-B	DINOv2	DINOv3	CLIP	SigLIP
	THINGS	48.6 → 60.1	45.1 → 60.6	44.4 → 60.5	50.7 → 60.6	48.8 → 60.7
Levels	Between-class	54.7 → 66.3	49.7 → 70.1	47.6 → 69.7	59.7 → 70.1	58.0 → 70.7
Levels	Within-class	57.8 → 53.4	65.2 → 57.2	64.3 → 56.6	62.1 → 56.7	63.5 → 57.9
Levels	Class-border	94.9 → 86.6	93.0 → 86.1	90.9 → 86.6	91.0 → 85.9	92.9 → 86.9

Alignment introduces task-dependent performance trade-offs. We next test whether these representational changes affect downstream task performance. Linear-probe ImageNet-1k accuracy decreases across all five backbones, with declines ranging from 0.6 to 3.4 points (mean −1.8; Fig. 10, left). In contrast, the effect on semantic segmentation yields a mixed result (Fig. 10, right): three backbones improve, with DINOv2 gaining up to +2.3 mIoU, while two decline, with the largest decrease observed for SigLIP (−4.0 mIoU). Thus, alignment consistently reduces classification performance under linear probing, but does not produce a uniform loss of downstream capability.

5.4.2 Alignment does not reliably improve human interpretability

Does alignment reliably improve interpretability? *No, it does not.* Alignment produces model-dependent changes in interpretability rather than a consistent improvement. Interpretability increases directionally for all three vision-only backbones but decreases for both vision-language backbones (Table 4). The observed interpretability gains range from +0.7 points for ViT-B/16 to +7.1 points for DINOv2, while interpretability losses reach −1.9 points for CLIP and −4.5 points for SigLIP. However, only the improvement for DINOv2 is statistically significant (Kruskal–Wallis test, $H(1) = 6.91$, $p = 0.009$). The evidence therefore supports a model-specific interpretability gain rather than a general benefit of alignment.

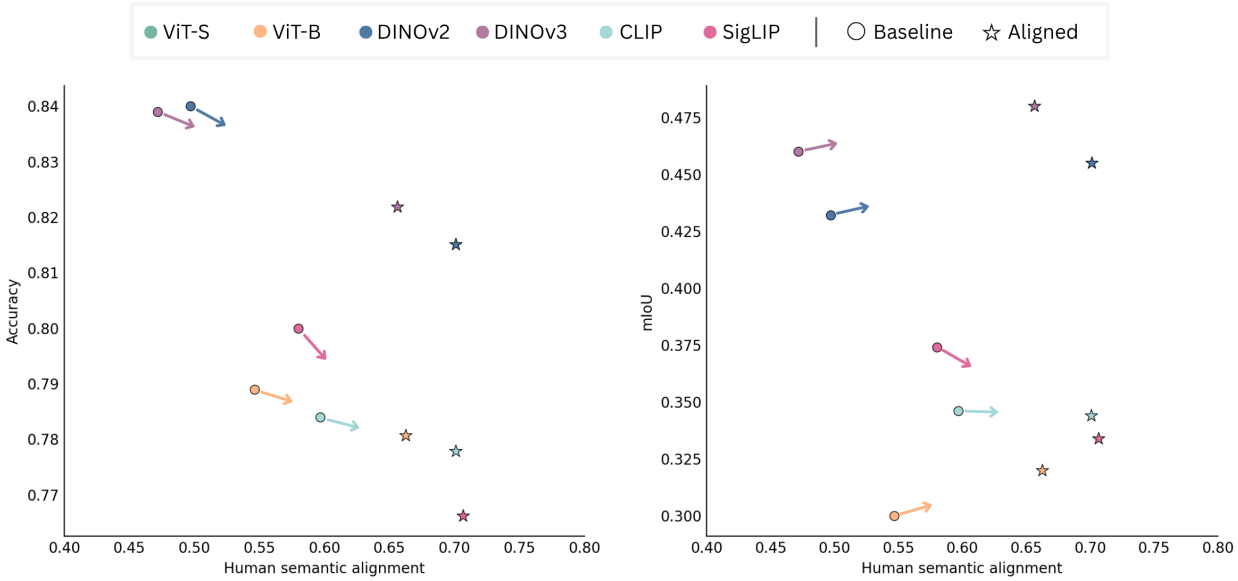


Figure 10. Effect of alignment fine-tuning on downstream performance. Human semantic alignment (Levels, *between-class*) is plotted against ImageNet-1k linear-probe accuracy (left) and segmentation mIoU (right). Circles denote pretrained baselines, stars denote aligned models, and colors identify the different backbones. Alignment consistently increases semantic alignment, while classification accuracy decreases across models and segmentation performance changes in both directions.

Interpretability changes are unrelated to classification costs. One possibility is that changes in interpretability simply reflect the downstream performance trade-offs introduced above. However, across image categories, interpretability changes are unrelated to changes in classification accuracy (Spearman $\rho = 0.014$, $p = 0.63$). The relationship also remains negligible within every backbone ($|\rho| \leq 0.09$; none significant; Appendix C.7). Therefore, reduced classification performance does not explain the observed interpretability changes.

Alignment concentrates predictive support into fewer features. Following Chapter 4, we define a feature as *informative* if it is among the three features contributing most strongly to the model's top prediction for at least one image from the psychophysics stimulus set. Alignment reduces the size of this informative feature vocabulary across all backbones: aligned SAEs identify between 1.2× fewer informative features for ViT-B/16 and 2.4× fewer for DINOv3 (App. C.6). Thus, alignment consistently concentrates predictive support into a smaller subset of features. However, this reduction alone cannot distinguish whether alignment removes features from the representation or reorganizes them into a more compact and efficient feature structure.

Does feature turnover explain the change? A natural explanation for the reduced feature vocabulary is that alignment improves interpretability by removing poorly understood features and introducing more interpretable ones. To test this hypothesis, we compare features lost during alignment—baseline features without an aligned-side correspondent—with features introduced by alignment—aligned features without a baseline-side correspondent (Appendix C.8). Across all five backbones, introduced features are slightly more inter-

Table 4. Aggregate localizability scores, multiplied by 100, before and after alignment across five vision transformers. Only DINOv2 changes significantly.

	Supervised	**Foundation**			
	ViT-B	DINOv2	DINOv3	CLIP	SigLIP
Baseline (↑)	86.0	71.0	79.6	79.7	71.4
Aligned (↑)	86.7	78.1*	83.2	77.8	66.9

*$p < 0.05$.

pretable than dropped features with the median differences ranging between +0.02 and +0.07. However, none of these differences reaches statistical significance (Kruskal–Wallis, all $p > 0.13$). Thus, although alignment changes the composition of the feature vocabulary, we find no evidence that it systematically improves interpretability by replacing less interpretable features with more interpretable ones, including in SigLIP, where alignment produces the largest overall decline.

5.4.3 Mechanisms underlying interpretability changes

Locality tracks interpretability. Feature locality increases significantly after alignment in exactly the three vision-only backbones (ViT-B/16 and DINOv2 $p < .001$, DINOv3 $p = 0.001$), whereas CLIP and SigLIP show no significant change ($p = 0.43$ and $p = 0.10$, respectively).. This separation is sharper than the pattern observed for interpretability itself.

Across all informative features, locality is the strongest correlate of interpretability among the mechanisms considered. Pooling all informative features across backbones yields $\rho = 0.33$ ($p < .001$, $n = 6695$), and the association remains positive and significant within every model ($\rho = 0.17$ to 0.39, all $p < .001$). Locality therefore tracks interpretability both across and within models.

Nevertheless, locality explains only part of the observed pattern. SigLIP's interpretability declines despite unchanged locality. Moreover, DINOv2 and DINOv3 exhibit different outcomes despite similar locality trends: DINOv3 even shows a stronger locality–interpretability correlation ($\rho = 0.36$ versus 0.29), yet only DINOv2 exhibits a significant aggregate interpretability gain.

Feature persistence reveals different alignment responses. Continuous co-firing analysis provides a more detailed view of how individual features evolve under alignment. DINOv2's significant improvement is *diffuse*: co-firing strength does not predict which features become more interpretable ($\rho = 0.09$, $p = 0.21$). By contrast, DINOv3's nonsignificant aggregate improvement is *concentrated* among features that persist less strongly across alignment ($\rho = -0.29$ to -0.34, $p = 0.002$). An image-level measure of co-firing preserves this overall contrast, although DINOv2 shows a weak association at this coarser scale (Appendix C.4). Feature locality independently reproduces the DINOv3 association ($\rho = -0.19$ to -0.27, $p = 0.004$). A complementary paired analysis finds no significant uniform shift among confidently matched features, although DINOv2 shows the largest positive mean and median change across the five backbones (Appendix C.9).

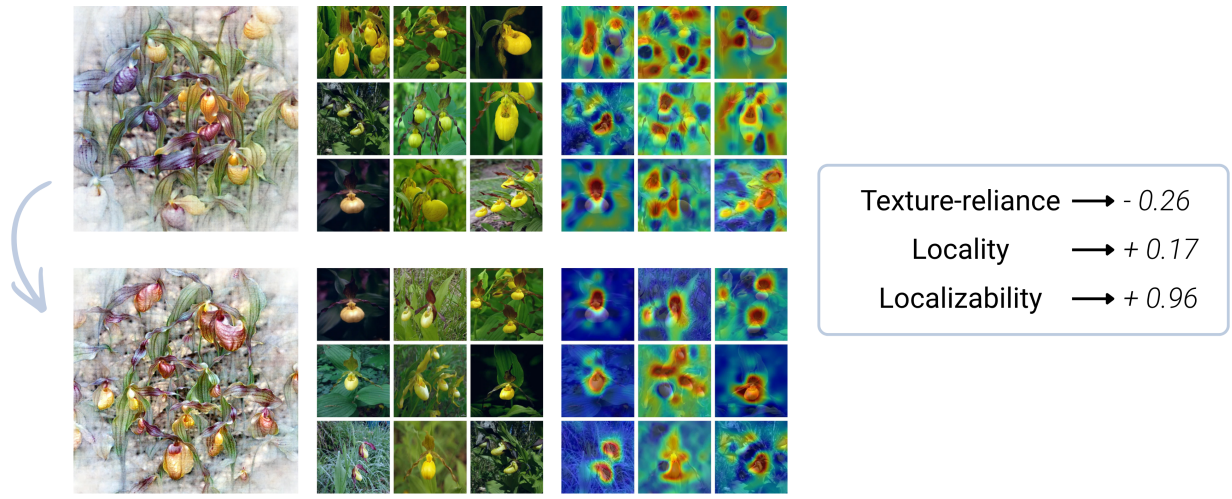


Figure 11. Illustration of a feature strongly affected by alignment. We illustrate a feature from DINOv2 before alignment (top) and after alignment (bottom) through its synthetic visualization, top-activating images, and corresponding heatmaps. Its localizability increases from 0.019 to 0.981, its locality from 0.063 to 0.229, and its texture reliance drops by 0.264.

Decomposing co-firing into precision and recall further separates features into commonly preserved, narrowed, and broadened features. DINOv3's interpretability gain is concentrated among commonly preserved features (+0.08) and reverses among broadened features (−0.06), although the overall regime difference does not reach significance ($p = 0.28$). We therefore treat this decomposition as an exploratory characterization rather than a confirmed mechanism.

Interpretability changes do not differ between these regimes ($p = 0.28$); in DINOv3, commonly preserved features barely change (+0.005) while broadened features decline (−0.06).

Alignment reduces texture reliance independently of interpretability. The texture-reliance proxy decreases significantly after alignment in all five backbones (Kruskal–Wallis, all $H(1) > 37$, all $p < .001$), with the largest reduction for CLIP (−0.14). Unlike locality, however, texture reliance does not track interpretability: the pooled correlation between the underlying blur-robustness score and interpretability is negligible ($\rho = -0.046$, $p = 0.0002$, $n = 6695$) and inconsistent in sign across backbones ($\rho = +0.05$ to -0.20). Reduced texture reliance is therefore a robust, general consequence of alignment, not an explanation for the interpretability changes.

Texture reliance nevertheless tracks feature persistence: its median decrease is −0.01 for commonly-preserved features, −0.05 for narrowed features, and −0.10 for broadened features (Kruskal–Wallis, $H(2) = 141.7$, $p < .001$). Thus, alignment consistently reduces high-frequency reliance, but this change is not the mechanism driving interpretability improvements.

5.4.4 Why do vision-language models fail to improve?

SigLIP exhibits the largest raw interpretability decline of any backbone, but the effect is not significant in either the full feature population ($p = 0.21$) or the persistence-matched analysis ($p = 0.15$; Appendix C.9). CLIP's smaller decline is also not significant. None of the

candidate mechanisms accounts for these trends: neither model shows a co-firing association or a significant locality shift, and SigLIP's features are predominantly commonly preserved (88%), arguing against large-scale restructuring.

The decline also cannot be attributed simply to feature turnover, as DINOv3 loses more features than SigLIP (58% versus 37%; Appendix C.6).

5.5 Conclusion

This chapter provides the first direct causal test of whether aligning vision models with human semantic judgments makes their representations more interpretable. Fine-tuning five pretrained ViT-B backbones successfully produces the intended semantic shift: models become better aligned with human similarity judgments and move toward more similar representational geometries. However, this representational alignment comes with trade-offs, including reduced linear-probe classification accuracy and a contraction of the informative feature vocabulary. More importantly, semantic alignment does not translate into a uniform improvement in human interpretability. Only DINOv2 exhibits a statistically significant gain, while the other vision-only models improve only directionally and both vision-language models move in the opposite direction.

The mechanisms underlying these changes are similarly non-uniform. DINOv2's improvement is distributed broadly across its representation, whereas DINOv3's directional improvement is concentrated among specific subsets of features. Increased feature locality provides one partial explanation: locality rises in the three vision-only models and is consistently associated with human interpretability at the feature level. Yet locality alone cannot account for the full pattern, as it fails to explain the difference between DINOv2 and DINOv3 or the declines observed in CLIP and SigLIP. Alignment also consistently reduces reliance on high-frequency image information, but this change appears orthogonal to interpretability. Finally, feature turnover and feature persistence analyses provide no evidence that improvements arise simply from replacing less interpretable features with more interpretable ones. The negative response of SigLIP remains unresolved: its decline cannot be attributed to changes in locality, feature persistence, feature turnover, or the loss of informative features.

Together, these findings refine the correlational relationship identified in Chapter 4. Human-semantic alignment is neither a guaranteed route to interpretability nor a single mechanism that uniformly reshapes models toward human understanding. Instead, alignment acts as a model-dependent intervention whose effects depend on how a particular representation reorganizes under the pressure of human semantic structure. The results therefore shift the question from whether human-semantic alignment can improve interpretability in principle to which representational changes make that improvement possible. Answering this question will require interventions that target not only semantic alignment itself, but the specific representational properties through which alignment can support human understanding.

Limitation One key limitation of our results concerns the intervention itself. We use LoRA rather than full fine-tuning, which limits how extensively alignment can reshape the representation and may therefore underestimate its potential effect on interpretability. Early experiments with full fine-tuning achieved similar levels of alignment but caused substantially larger losses in downstream performance. We therefore chose LoRA as a compromise:

an intervention intended to improve interpretability must preserve a reasonable level of capability. Still, a less constrained intervention might produce larger or qualitatively different representational changes.

We also evaluate interpretability only through localizability. Chapter 4 found localizability and semanticity to produce strongly correlated model rankings, supporting the use of the more scalable protocol here. However, the two measures are not identical, and alignment may affect the ability to predict *where* a feature activates differently from the ability to describe *what* it represents. Repeating the evaluation with semanticity could therefore reveal changes not captured by our current results.

Identifying shared features across representations also remains difficult. Our analysis uses a shared sparse autoencoder to establish a common basis and co-firing to measure feature persistence. These tools provide a practical approximation, but not a canonical correspondence between features, as illustrated by the sensitivity of some results to the co-firing cutoff. The feature-level analyses should therefore be viewed as an initial characterization of how alignment reorganizes representations. More robust correspondence methods will be needed for a finer account of the mechanisms involved.

Finally, the five backbones considered here are insufficient to explain why alignment affects the two vision–language models differently from the vision-only models. The present comparison cannot determine whether the declines observed for CLIP and SigLIP arise from language supervision, contrastive training, properties of their initial representations, or interactions between these factors and the intervention. Additional vision–language and contrastively trained vision-only models—e.g., I-JEPA—will be needed to distinguish these possibilities.

These limitations constrain the generality and mechanistic interpretation of the results, but not their central implication: under the conditions tested here, increasing alignment with human semantic judgments does not reliably improve interpretability. With the representational basis, measurement framework, and first causal test now established, Chapter 6 brings these contributions together and considers what they reveal—and leave unresolved—about the principles that make model representations understandable to humans.

Chapter 6

Conclusion and Perspective

6.1 Conclusion

Explainable AI emerged from the goal of making machine learning models understandable to humans, and over the past fifteen years has developed an increasingly sophisticated set of tools for probing their internal computations. Those XAI tools are now sufficiently mature to revisit the question that motivated their development. Rather than viewing interpretability as a property of the explanations we produce, this perspective treats it as a property of the models themselves, one that can be measured, compared, and steered.

The central premise of this thesis follows from that shift in perspective. Instead of focusing on improving the instruments used to inspect opaque models, we sought to identify principles that reduce the opacity of the models themselves, making their learned representations intrinsically more interpretable to humans. Several such principles have been proposed in the literature, to date, none has been rigorously established as both measurable and causally linked to interpretability. The primary contribution of this thesis is to provide the conceptual and methodological framework needed to identify, test and validate such principles.

To determine which principles govern interpretability, one must compare the interpretability of many models with sufficient precision to isolate the factors that distinguish more interpretable representations from less interpretable ones. Such a comparison requires an evaluation framework that is both principled and fine-grained, yet no such framework previously existed. Moreover, because interpretability is fundamentally a property of a model's representations, the evaluation must rest on an appropriate representational basis. Whether that basis should consist of individual neurons or distributed directions, however, remained an open question. Before any candidate principle could be rigorously tested, the representational basis first had to be established, and a reliable evaluation built upon it.

These prerequisites, together with the causal test they ultimately enable, correspond to the three main contributions of this thesis. Chapter 3 resolves the representational question by demonstrating that distributed directions, rather than individual neurons, provide an appropriate basis on which the remainder of the thesis is built. Chapter 4 develops the evaluation metric and experimental framework, and applies them to a broad set of contemporary vision foundation models. The comparisons enabled by this framework reveal the candidate principles that systematically distinguish models with more interpretable representations from those with less interpretable ones. Chapter 5 then takes one such candidate, namely the alignment with human semantics, and subjects it to a causal test. The re-

sults provide insufficient evidence that alignment reliably improves interpretability: gains are limited to the vision-only models, significant only for DINOv2, and reversed for the two vision-language models. Alignment can therefore improve interpretability, but its effects are model-dependent, can be negative, and remain only partially understood. Human semantic alignment is a promising principle, but its benefits are contingent on the representations to which it is applied.

6.2 Perspective

The thesis resolves several open questions regarding model interpretability while raising new ones. This section reflects on the research directions that emerged throughout the PhD: avenues to which this work directly contributes, questions for which it establishes a baseline or point of comparison, and broader challenges that remain only partially understood. Several of these perspectives revisit open questions introduced in Chapter 1, now reconsidered through the lens of the empirical findings developed throughout this thesis. While these results narrow the space of possible explanations and interventions, they do not fully resolve the deeper question of how to build representations (and ultimately models) that are genuinely understandable to humans.

Two routes to interpretable models. This thesis studies general-purpose pretrained deep learning models for computer vision and asks to what extent the representations they learn are understandable to humans. It does not examine the explainable-by-design models introduced in Chapter 1. As discussed there, these two routes toward interpretability have rarely been compared empirically, particularly with respect to the understanding they afford human observers. The framework developed in this thesis provides one side of that comparison while making the other readily accessible, since it is not specific to general-purpose models. Instead, it evaluates the interpretability of learned representations independently of how those representations were obtained. Consequently, it can be applied with minimal adaptation to concept bottleneck models and other architectures designed around explicit human concepts. Applying this framework across both families of models would provide a direct test of a fundamental question in interpretability research: whether understandable models are better achieved by constraining models to represent human concepts from the outset, or by shaping the internal representations learned by otherwise general-purpose systems. Neither approach is guaranteed to succeed, and behavioral evaluation will be essential for determining which route produces representations that humans can genuinely understand.

More candidates to test. This thesis establishes a framework for testing principles of interpretability, but causally evaluates only one candidate: alignment with human semantic judgments. Several other principles identified throughout the thesis remain open for investigation. One particularly promising candidate is locality. Unlike semantic alignment, locality is less straightforward to manipulate after training: DINOv3 achieves locality through properties of its pretraining objective, rather than through a simple fine-tuning intervention. A key future challenge is therefore to develop interventions that directly control the spatial or conceptual granularity of learned features while preserving other properties of the representation.

One possible route is alignment with human visual strategies. Because human visual processing relies strongly on localized and object-centered representations, aligning models with human strategies can induce more local features as a consequence [120]. However, such an alignment also changes many other properties of the representation, making it difficult to isolate the contribution of locality itself. Another candidate principle is robustness. Beyond the work discussed in Chapter 1, recent evidence suggests that robust models encode features that better resemble those recovered from neural recordings in visual cortex [94]. If features that humans themselves use are easier for humans to understand, robustness may provide another route toward more interpretable representations. Testing such principles systematically would transform interpretability research from identifying correlations into discovering reliable design principles.

A related open question is which aspects of model development determine these properties: architecture, training objective, data distribution, or some interaction between them. The contrasting results obtained for DINOv2, DINOv3, CLIP, and SigLIP suggest that the same intervention can produce different outcomes depending on the structure of the representation being modified. Understanding these dependencies will require moving beyond comparisons between isolated models toward controlled studies that vary one component at a time: training objectives, supervision signals, architectural constraints, and alignment procedures. Such comparisons may reveal whether interpretable representations arise from particular ingredients of learning systems or from broader principles that transcend specific architectures.

***The* principle.** The candidate principles considered here share a common theme: they are different forms of alignment with human organization. However, they are not interchangeable. For example, models that best align with human semantic judgments, visual strategies, and neural responses are not necessarily the same models [4]. Just as the neuroscience literature asks which dimensions of alignment are necessary for models to resemble the brain, interpretability research must ask which dimensions are necessary for humans to understand models.

This does not require models to become human-like in any complete sense. Human cognition itself contains many specialized structures that are neither necessary nor desirable for artificial systems. Instead, models may only need to align with humans along a subset of dimensions that make their internal organization accessible to human reasoning. Identifying those dimensions is likely to be central to building models that are understandable by design.

An important implication of this view is that interpretability may not be reducible to any single property of a representation. The results of this thesis already point in this direction: semantic alignment can improve without reliably improving human understanding, while locality correlates with interpretability but does not fully determine it. A mature science of interpretability may therefore require moving beyond ranking models along a single axis and instead characterizing the specific representational properties that support different forms of human understanding.

Measuring human understanding. This thesis began with the observation that defining and measuring understanding remains an open challenge, and it does not resolve that question. Instead, it adopts a practical operationalization of understanding: simulatability, or the ability of humans to predict model behavior, and introduces an additional behavioral

protocol whose limitations are discussed in Section 4.6.

A broader limitation is that these evaluations capture only a static snapshot of understanding. In practice, understanding a model is an iterative process: users inspect explanations, form hypotheses about the model, test those hypotheses on novel cases or failures, and revise their understanding accordingly. Interpretability should therefore be viewed not only as a property of a representation, but also as a property of the interaction between a model and the humans attempting to understand it. Early work has begun to formalize interpretability as a dynamic human–AI interaction process [187]. Developing evaluation protocols that capture this interactive loop remains an important challenge.

From interpretable representations to interpretable decisions. All the analyses described in this thesis focus on representations rather than decisions. This choice reflects a foundational assumption: although decisions are ultimately what affect people, they emerge from the representations on which models operate. If we view a model as a two-stage process—mapping inputs into representations and then mapping representations into decisions—then interpretable decisions require interpretable intermediate representations. In this sense, interpretable representations are a necessary, though not sufficient, foundation for interpretable models.

The emergence of general-purpose representations makes this direction increasingly important. A future goal is to develop broadly reusable representations that are interpretable by construction and can support many downstream tasks. Building on such representations, the next step is to study interpretable decision layers: sparse linear models, logical models, concept-based predictors, and other approaches developed in the explainable-by-design literature [162]. The challenge will be to preserve the flexibility of general-purpose representations while making the final decision process transparent and controllable.

Structural insight is the tractable place to start. Every result in this thesis rests on human validation, and we take human evaluation to be the final arbiter of interpretability. It need not, however, be the first step. The space of possible interventions, architectures, and learning objectives is too large for exhaustive behavioral evaluation. Structural analysis therefore provides an essential first stage: identify meaningful differences between models and representations, then evaluate only the differences most likely to affect human understanding. The tools required for such comparison are rapidly improving [119, 186], as discussed in Chapter 1. They provide a natural path toward extending the agenda developed in this thesis: not only asking which models are more interpretable, but identifying which features they converge on and which they do not—across architectures, objectives, and alignment procedures. This returns to the convergence question raised in Chapter 1 and connects the thesis to a broader line of research on representational convergence in deep learning [77, 93]. If models trained under different regimes converge on the same features, those features are where an account of interpretability should begin.

From the Pioneer plaques to the Voyager Golden Record, humanity has long sought a shared language with other intelligences. Perhaps interpretability is only the latest expression of that enduring quest—to be understood, and to understand.

Part I

Appendix

Appendix A

Appendix of Chapter 3

A.1 Selection of dictionary elements

Our hypothesis is that applying dictionary learning to neuron-based representations enables the recovery of directions corresponding to multiple patterns encoded in superposition, yielding representations that are more interpretable than individual neurons. To evaluate this hypothesis, for each neuron in the neuron-based condition, we aim to recover the underlying patterns it encodes by learning a complementary dictionary over neuron activations.

Concretely, following recommendations of Fel, Picard, Bethune, et al. [60], we learn a dictionary with $k = 10$ elements. This choice of k is intended to be sufficiently large to capture the meaningful patterns encoded by a neuron while satisfying $k > s$, where s denotes the (unknown) number of latent patterns encoded in superposition. Unused dictionary elements are expected to capture residual noise.

From the learned dictionary, we select a single element to compare directly with the original neuron. A key methodological question is how to select this dictionary element. We consider three strategies: (a) random selection; (b) selection based on predictive importance; and (c) selection based on relevance, measured by activation frequency. We adopt strategy (c), as it introduces the least amount of bias into the comparison.

Random selection (a) is undesirable because, unless the model is in an extreme superposition regime, any given neuron is expected to encode at most $s < k$ different patterns in superposition. Consequently, randomly selecting a dictionary element could lead to an element that captures noise rather than a meaningful pattern, which would unfairly disadvantage dictionary-based representations. On the other hand, selecting elements based on their predictive importance (b) would bias the comparison in favor of dictionary-based representations, since the neurons themselves were not selected according to their importance for the model's predictions. In contrast, selecting the dictionary element with the highest activation frequency (strategy (c)) provides a principled and symmetric criterion that reflects how consistently a pattern is expressed in the data, without privileging either representation.

A.2 Experimental confounds

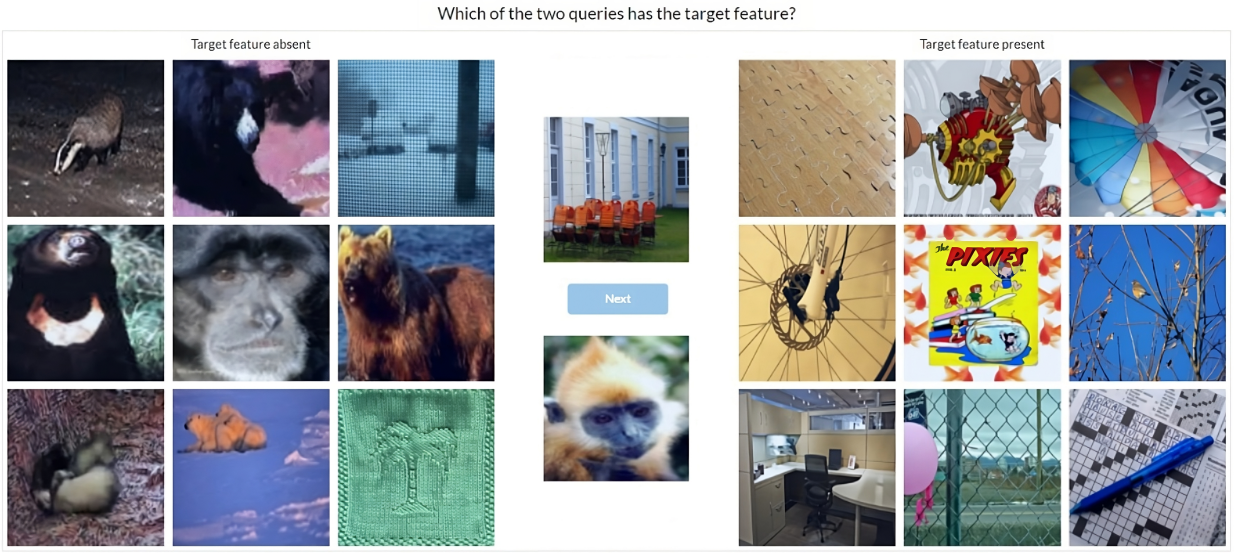
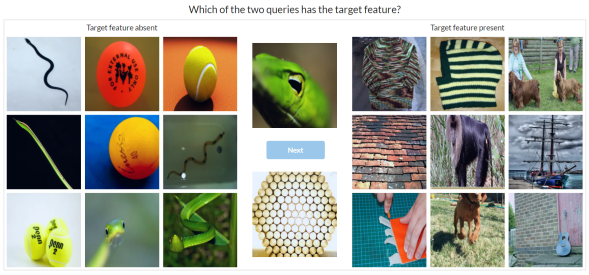
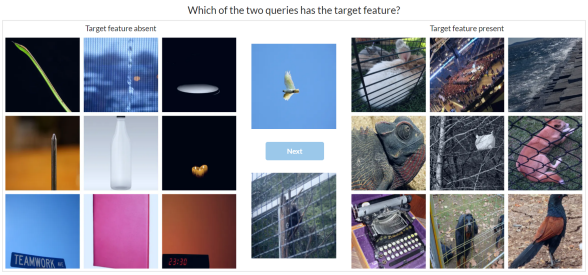


Figure 12. Illustration of the role of semantics. Example of a trial from Experiment I in the neuron-based representation condition. In this case, the task can be trivially solved by relying on semantic grouping. By observation of the minimally activating stimuli (left panel), it is easy to conclude that the neuron of interest is not a monkey detector, yet, it is hard to articulate what visual pattern is captured by the neuron (images in the right panel).

A.3 Examples of trials from Experiment II



(a) Neuron-based trial

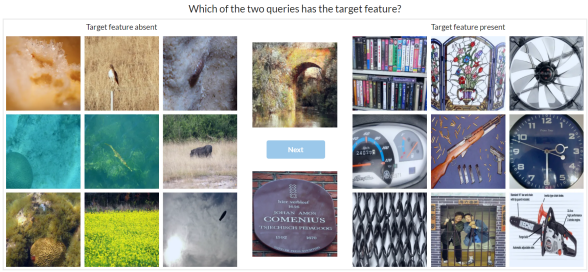


(b) Dictionary-based trial

Figure 13. Layer2. This figure illustrates a trial used to assess the features encoded in layer2.0 either by the neuron 52 (a) or at least partially through the neuron 52 (b).

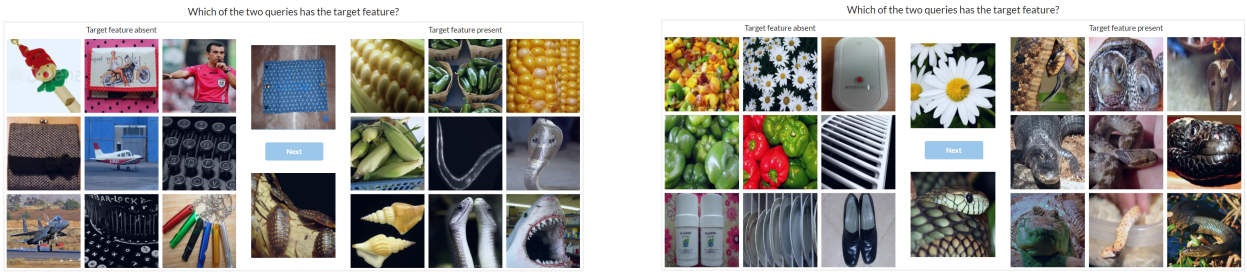


(a) Neuron-based trial



(b) Dictionary-based trial

Figure 14. Layer3. This figure illustrates a trial used to assess the features encoded in layer3.1 either by the neuron 957 (a) or at least partially through the neuron 957 (b).



(a) Neuron-based trial

(b) Dictionary-based trial

Figure 15. Layer4. This figure illustrates a trial used to assess the features encoded in layer4.2 either by the neuron 259 (a) or at least partially through the neuron 259 (b).

A.4 Sensitivity analysis

To complement the sample-size prior from Zimmermann, Klein, and Brendel [204], we performed a sensitivity analysis based on our retained sample sizes. For the main dictionary-vs-neuron comparisons, each experiment can be approximated as a two-group comparison with approximately 65 participants per condition. With $\alpha = 0.05$ and 80% power, this design is sensitive to medium-sized effects: Cohen's $d = 0.49$ or larger. The observed dictionary-vs-neuron effects were d = 0.56, d = 0.94, and d = 0.45 for Experiments I–III, respectively. Thus, our individual experiments were sensitive to medium-sized dictionary-vs-neuron differences, and the observed effects were generally of this magnitude. For the analyses pooling across experiments, we rely on the mixed-effects models reported in the main text, which account for the structure of the data by including participant-level variability and experiment/model structure.

Appendix B

Appendix of Chapter 4

B.1 Additional Related Work

Representational alignment with human perception. A growing body of work investigates how well DNN representations mirror human perceptual judgments [179], with parallel evidence that aligning model feature attributions with human importance maps improves object recognition [61]. Alignment with human similarity judgments has been shown to improve transfer [140, 178], though benefits are task-dependent: alignment helps retrieval but may hurt discriminative classification [181]. Crucially, Muttenthaler, Greff, Born, et al. [139] showed that this trade-off depends on granularity: coarse-grained and fine-grained alignment affect downstream tasks differently. While prior work treats alignment as a driver of task performance, we treat it as a potential correlate of *feature interpretability*—a connection that, to our knowledge, has not been investigated.

Automated feature scoring with vision-language models. Hernandez, Schwettmann, Bau, et al. [86] proposed labeling visual neurons with natural-language descriptions and measuring label quality via retrieval accuracy. Bills, Cammarata, Mossing, et al. [14] extended this idea to language model neurons, using GPT-4 to generate explanations scored by how well they predict held-out activations. CLIP-Dissect [147] matched each neuron to the most CLIP-similar concept from a large vocabulary, enabling scalable, open-vocabulary labeling without human annotation. Our *semanticity* score builds on this spirit—participants describe a feature in free text and we measure CLIP cosine similarity between their description and feature-centered crops—while retaining the validity guarantee of human behavioral data.

B.2 Broader impact.

This chapter contributes to human-centered evaluation of model interpretability. By testing whether people can localize and describe decision-relevant features, our protocol adds to a broader effort to assess whether vision models are transparent enough for settings where human oversight matters, including critical applications. However, interpretability evaluations should not be taken as evidence of safety or reliability on their own. A model whose features appear more interpretable may still fail under distribution shift or behave unpredictably in downstream applications. Our results should therefore be viewed as one component of

a broader evaluation ecosystem, complementing robustness, fairness, calibration, and task-specific validation.

B.3 Details about the control experiment



Figure 16. Example of a trial from Zimmermann, Klein, and Brendel [204]. In this protocol, the task is to pick the query image that matches the reference images illustrating the feature (left panel), by leveraging reference images that both highly activate the feature (right panel) and images that minimally activate it (left panel). In this example, the task can be trivially solved by relying on semantic grouping. By observation of the minimally activating stimuli (left panel), it is easy to conclude that the neuron of interest is not a corn detector, yet, it is hard to articulate what visual pattern is captured by the neuron (images in the right panel).



Figure 17. Instantiation of the trial from Fig. 16 in our control experiment (Sect. B.3). In this protocol, we replace the images that highly activate the feature with images that carry little to no information about the feature's visual content. If the protocol only quantify the degree to which participant understand the feature, we expect them to perform at chance (50%) in this setup.

B.4 Scaling model interpretability

Model-level interpretability is a property of a model's full feature population. Under a fixed annotation budget, we therefore face a choice: cover more features (*breadth*) or measure each feature more carefully (*depth*). This pilot asks which choice better approximates the population score.

What we do. We collect interpretability judgments on 80 DINOv2 neurons; this defines our reference score of 0.5555. We then use bootstrapping (200 resamples per design) to simulate two cheaper alternatives: subsampling fewer units (breadth) or reducing the number of images and trials per unit (depth). For each design we report the bootstrap mean and SD.

Strategy	**Setting**	**Bootstrap score**	**Δ vs. ref.**
Reference	All 80 units (full data)	0.5555	—
Breadth	10 units	0.5586 ± 0.0291	$+0.0031$
Breadth	30 units	0.5541 ± 0.0142	-0.0014
Breadth	60 units	0.5552 ± 0.0057	-0.0003
Depth	1 image, 1 trial	0.5475 ± 0.0376	-0.0080
Depth	1 image, 3 trials	0.5484 ± 0.0278	-0.0071
Depth	10 images, 1 trial	0.5509 ± 0.0083	-0.0046

Table 5. Breadth vs. depth under a fixed budget. Bootstrap estimates of cheaper designs against the full-data reference. Breadth approaches the reference more quickly than depth.

What we find. Breadth converges quickly to the reference: 60 units already match it within ± 0.006. On the other hand, depth converges more slowly. Even 3 trials on a single image (± 0.0278) is less stable than 30 units alone (± 0.0142).

Why this points to scaling features. Depth makes our estimate of the units we picked more precise; breadth makes it more representative. Because the 80-unit reference is itself a stand-in for a much larger feature population, only breadth moves the estimate toward what we actually care about. We therefore allocate the main experiment's budget across as many features as possible rather than across repeated judgments on a small set of features.

B.5 Overview of our feature extraction and explanation pipeline

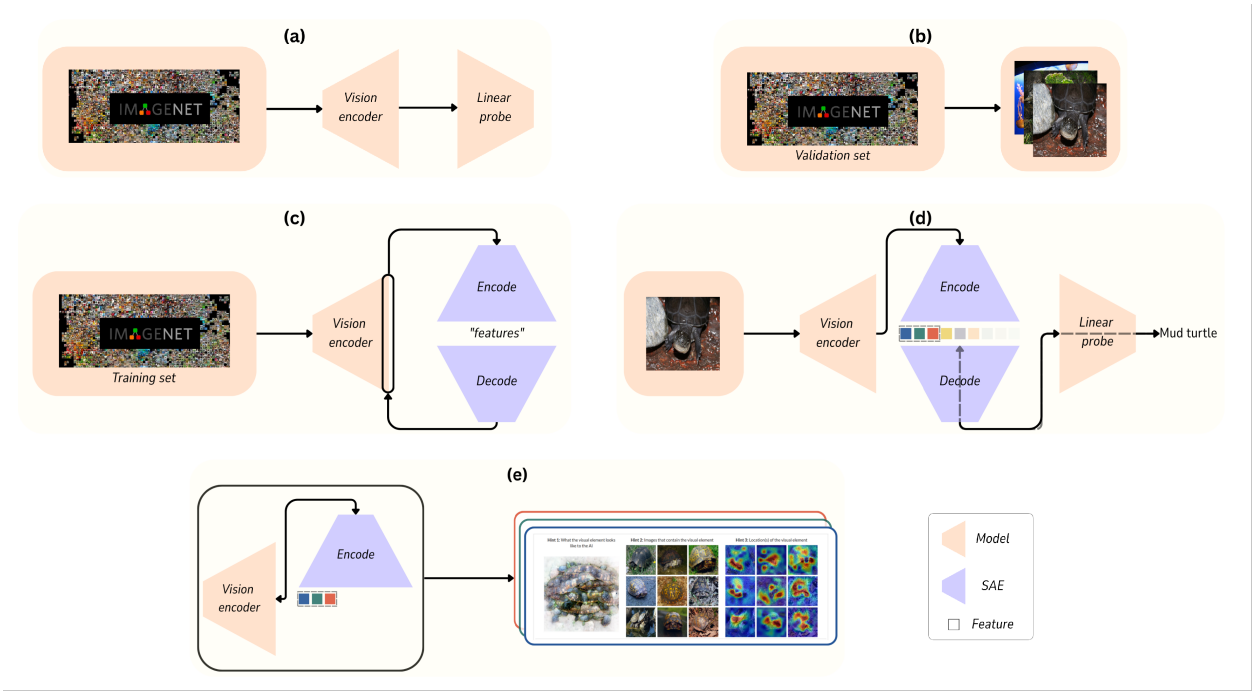


Figure 18. Pipeline for extracting and explaining vision model features. (a) A linear probe is trained on ImageNet over the frozen backbone to derive per-feature importance and an accuracy baseline. (b) A representative subset of ImageNet images is selected to anchor evaluation across all models. (c) A TopK SAE is trained on patch-token activations to recover monosemantic features. (d) The most important features are selected per image, using Gradient×Input. (e) Each feature is explained via a MACO visualization [56], nine top-activating images, and RISE heatmaps [154]. See Section 4.3.2 for full details.

Table 6. Reconstruction quality and dead-feature rates for the sparse autoencoder trained on each model.

Model	Reconstruction (R^2)	Dead features
ViT-S	0.7147	1.0%
ViT-B	0.7152	0.0%
CLIP	0.7257	2.5%
DINOv3	0.8211	0.0%
DINOv2	0.8019	0.0%
SigLIP	0.8963	0.0%

Table 7. Stability of SAE features across random seeds and its effect on median localizability. Stable features are those whose maximum cosine similarity with a feature in the independently trained dictionary exceeds 0.5.

Model	Stable features	All features	Stable only	Δ
ViT-S	64.6%	0.803	0.802	−0.001
ViT-B	80.4%	0.860	0.864	+0.004
CLIP	50.1%	0.797	0.764	−0.033
SigLIP	76.6%	0.714	0.743	+0.029
DINOv2	83.7%	0.710	0.698	−0.012
DINOv3	50.7%	0.796	0.791	−0.005

B.6 SAE training, reconstruction, and stability

Reconstruction quality does not track human interpretability in the direction required for it to explain the supervised–foundation pattern. ViT-S and ViT-B have nearly identical reconstruction (R^2 = .7147 and .7152) yet differ by 5.7 localizability points. Likewise, DINOv2 and DINOv3 differ only modestly in reconstruction but by approximately 9 localizability points. Dead-feature rates also do not explain the ordering: CLIP has the largest dead-feature fraction yet is the highest-scoring foundation model on semanticity.

We independently train a second SAE for each model using a different random seed. For every original feature, we compute its maximum cosine similarity to any feature in the second dictionary and treat similarities above 0.5 as a match [186].

Restricting the analysis to stable features changes model localizability by at most .033. The only ordering change is a swap between CLIP and DINOv3, whose original medians differ by only .001; both supervised models remain above all four foundation models.

B.7 Distribution of feature reuse across models

Table 8. Number of unique SAE features retained after filtering dense features for each model. Differences reflect varying degrees of feature reuse across architectures.

	Supervised		**Foundation**			
	ViT-S/16	ViT-B/16	DINOv2	DINOv3	CLIP	SigLIP
# Unique features	1012	1305	1119	1018	545	896

B.8 Psychophysics experiment

The experiment proceeded in three stages. Participants first read a welcome screen describing the task they were about to perform (Fig. 19). They then completed a short practice session to familiarize themselves with the task, followed by the main experiment. On each trial, a feature was presented through its visualization, top activating images, and corresponding heatmaps; depending on the condition, participants were asked either to click on the location where they expected the feature to appear in a new image (*localizability*, Fig. 20) or to provide a free-text description of the feature (*semanticity*, Fig. 21).

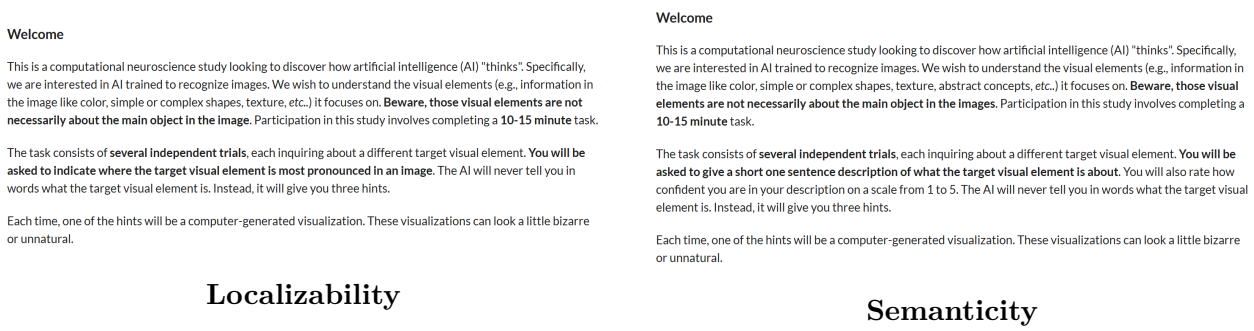


Figure 19. Welcome screens shown to participants before the psychophysics experiments. Left: instructions for the *localizability* task. Right: instructions for the *semanticity* task.

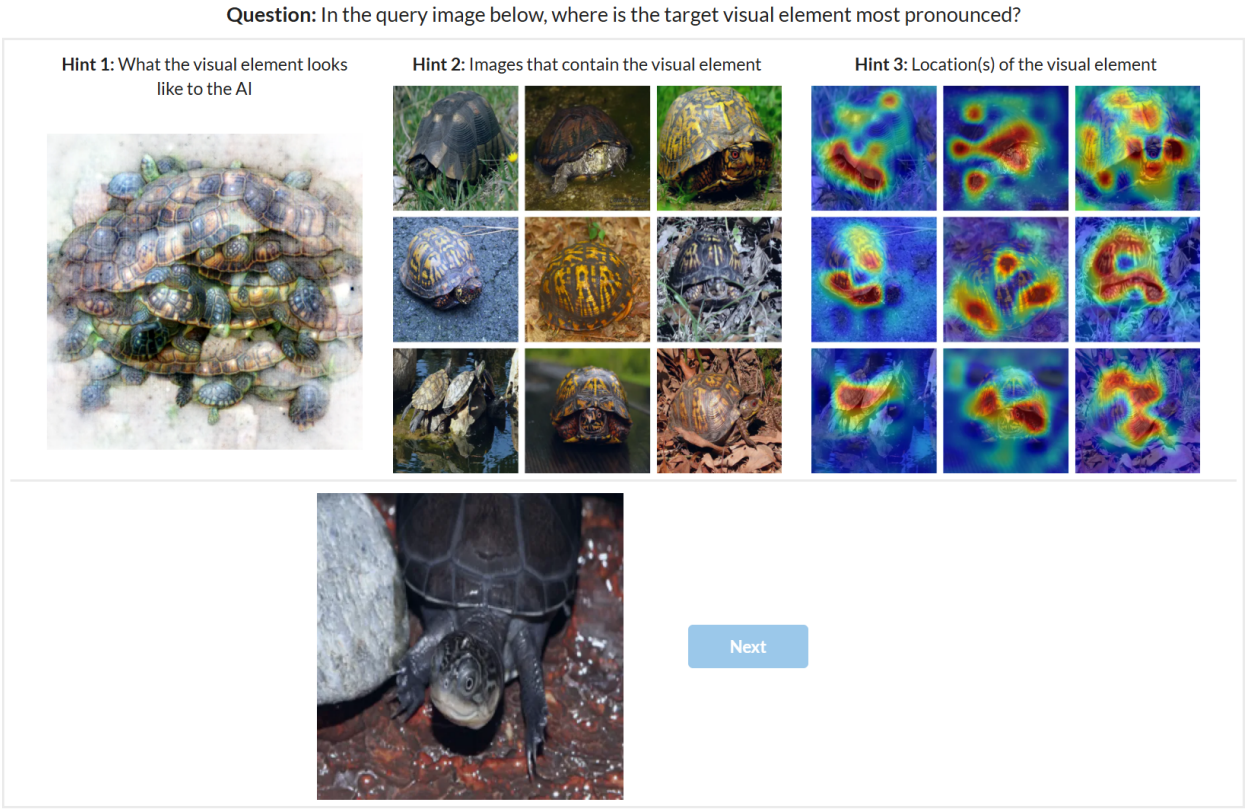


Figure 20. Screenshot of a trial for the *localizability* experiment for DINOv3. A feature is explained through 3 panels at the top: (left) feature visualization by means of a maximally activating synthetic image, (middle) a set of images that highly activate the features, with their associated heatmaps (right) highlighting where the feature is located on those images. Participants were asked to click on a location where they expected the feature to be present on a new image (bottom). The more interpretable the feature, the more likely they are to correctly identify an area where the feature is present in the image.

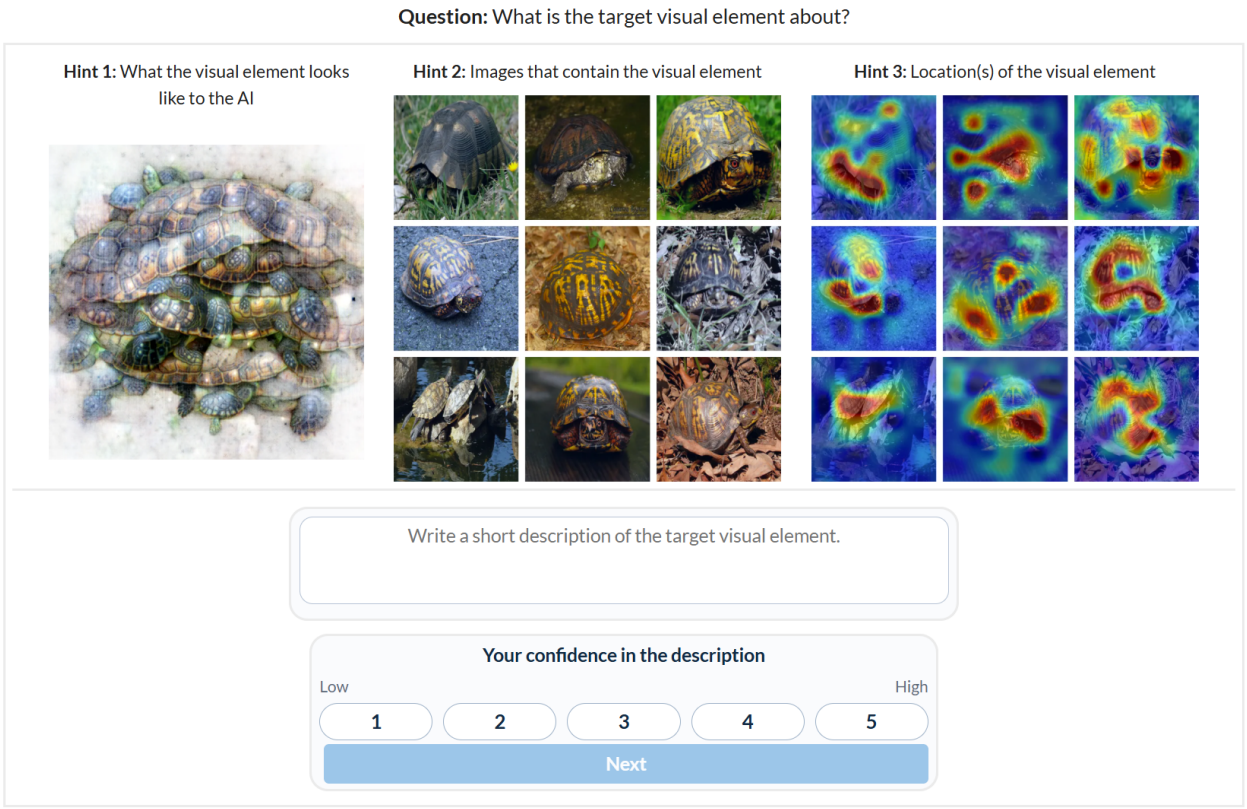


Figure 21. Screenshot of a trial for the *semanticity* experiment for DINOv3. A feature is explained through 3 panels at the top: (left) feature visualization by means of a maximally activating synthetic image, (middle) a set of images that highly activate the features, with their associated heatmaps (right) highlighting where the feature is located on those images. Participants were asked to give a text-free description of what the feature is about (bottom). The more interpretable the feature, the more accurate their description.

B.9 Quadrant of features

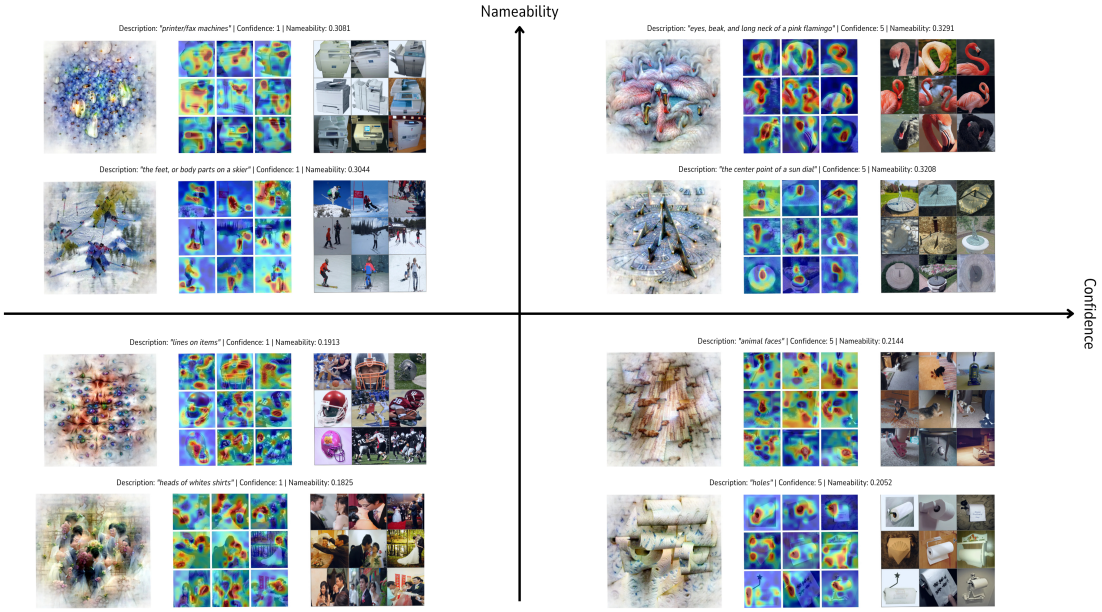


Figure 22. Relationship between *semanticity* and confidence at the feature level. Examples of features from each of the four quadrants defined by human-rated confidence (x-axis) and *semanticity* (y-axis). For each feature we show, from left to right: the feature visualization, heatmaps for top activating images, and the top activating images. The header above each example reports the human-provided description together with its confidence and *semanticity* scores. Features in the upper-right quadrant are both confidently and accurately named, whereas features in the lower-left are neither confidently nor accurately named.

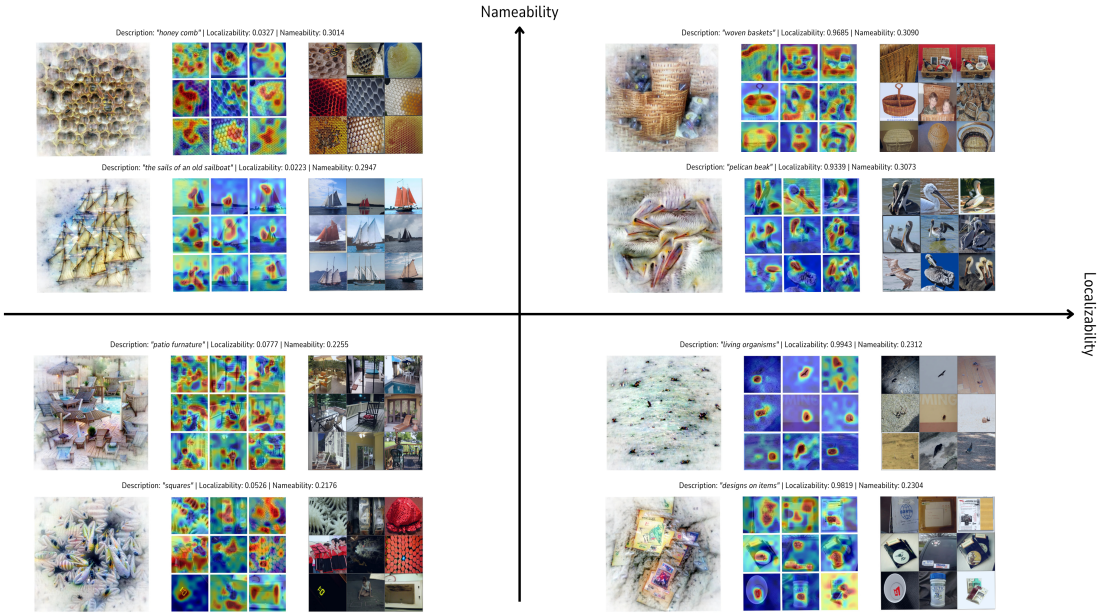


Figure 23. Relationship between *semanticity* and *localizability* at the feature level. Examples of features from each of the four quadrants defined by *localizability* (x-axis) and *semanticity* (y-axis). For each feature we show, from left to right: the feature visualization, heatmaps for top activating images, and the top activating images. The header above each example reports the human-provided description together with its *localizability* and *semanticity* scores. Features in the upper-right quadrant are both accurately localized and named, whereas features in the lower-left are not accurately localized or named.

B.9.1 Robustness to the choice of attribution method

Our localizability score uses RISE heatmaps as a ground truth for evaluating participant clicks. To test whether the results depend on this choice, we repeat the analysis with HSIC [146], an independent perturbation-based attribution method.

Agreement between RISE and HSIC. We first treat the maximum of each HSIC heatmap as a proxy participant click and score it against the corresponding RISE heatmap using the same *localizability* as in the main experiment. HSIC peaks consistently fall within high-scoring regions of the RISE maps (median scores 0.889–0.949). Scores differ across models (Kruskal–Wallis, $H = 123.2$, $p < .001$), but the effect is negligible ($\epsilon^2 = .012$), reflecting near-ceiling agreement throughout.

HSIC as ground truth. Then, we rescore the original participant clicks against HSIC rather than RISE heatmaps. Table 9 compares the resulting aggregate *localizability*. Although absolute scores decrease under HSIC, the broad model ordering is retained and remains highly correlated with the original ranking (Spearman $\rho = .714$, $p = .111$, $n = 6$).

Overall, these analyses indicate that the main localizability pattern is not specific to RISE.

Table 9. *Localizability* under RISE and HSIC. *Localizability* obtained by scoring the same participant clicks against heatmaps generated by either attribution method.

Model	**RISE**	**HSIC**	**Δ**
ViT-S	0.803	0.692	−0.111
ViT-B	0.860	0.738	−0.122
DINOv2	0.710	0.649	−0.061
DINOv3	0.796	0.724	−0.072
CLIP	0.797	0.722	−0.075
SigLIP	0.714	0.578	−0.136

B.10 Plausibility robustness

Crop vs full image. For a given feature, we evaluate both the similarity of a general description—the name of the main object, and a more specific one which we believe encapsulates the feature better. Descriptions that go beyond simply naming the main object in the image generally receive a higher score when using crops (see Fig. 24 for an illustration).

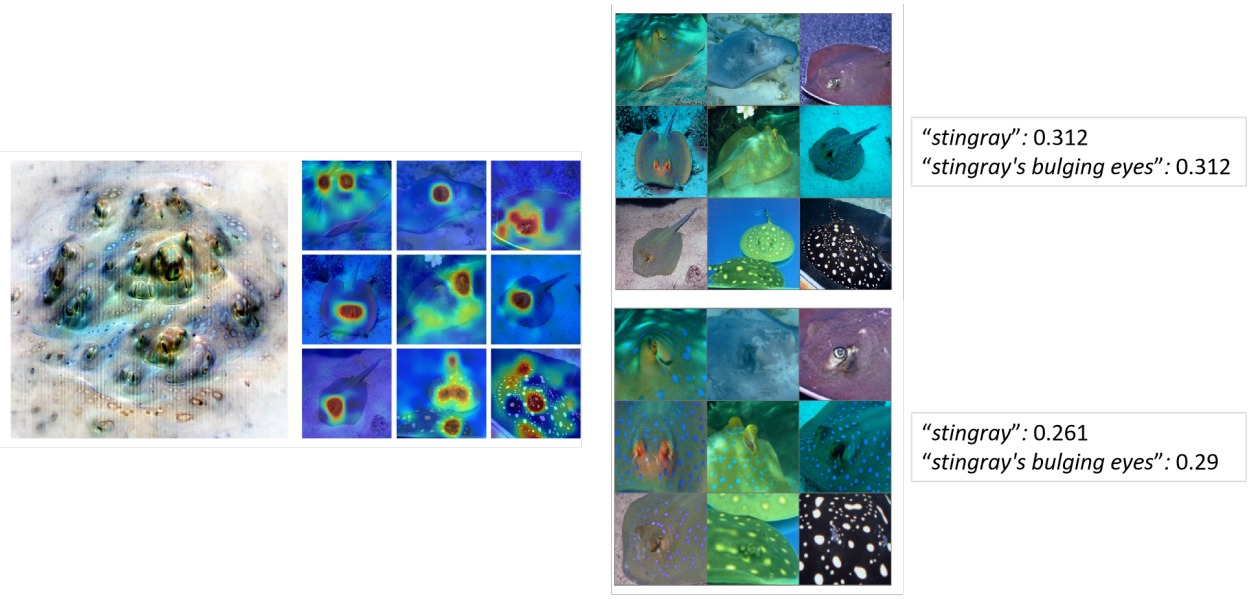


Figure 24. For a given feature, we evaluate the similarity of two descriptions, a general one consisting of the name of the main object, and a more specific one which we believe encapsulates the feature better. Using crops allows us to reward descriptions that go beyond simply naming the main object in the image.

Regardless, Table 10 highlights that results are fairly robust to the type of image used: *full images vs. crops*. We also rescore the 4,470 trials using 64, 128, and 160 pixel crops in addition to the 96 pixel crops reported in the main text. Relative to the 96 pixel baseline, pooled trial-level correlations are $\rho = .910$, .970, and .945, respectively. The weakest per-model correlation is still .843 for CLIP ViT-B/16 at 64 pixels, and the model ranking remains unchanged.

Table 10. Semanticity scores computed using the full image (bottom row) versus crops (top row) around the feature location. Best result per row is in bold, second best is underlined.

	Supervised		**Foundation**			
	ViT-S/16	ViT-B/16	DINOv2	DINOv3	CLIP	SigLIP
Crops (↑)	**0.274**	<u>0.273</u>	0.259	0.260	0.266	0.253
Full image (↑)	<u>0.276</u>	**0.277**	0.256	0.258	0.260	0.250

VLM scorer. The main-text plausibility score is computed with CLIP ViT-B/16 [157]. To verify this choice does not drive the reported model ranking, we re-score every participant description with six additional vision-language encoders spanning different architectures

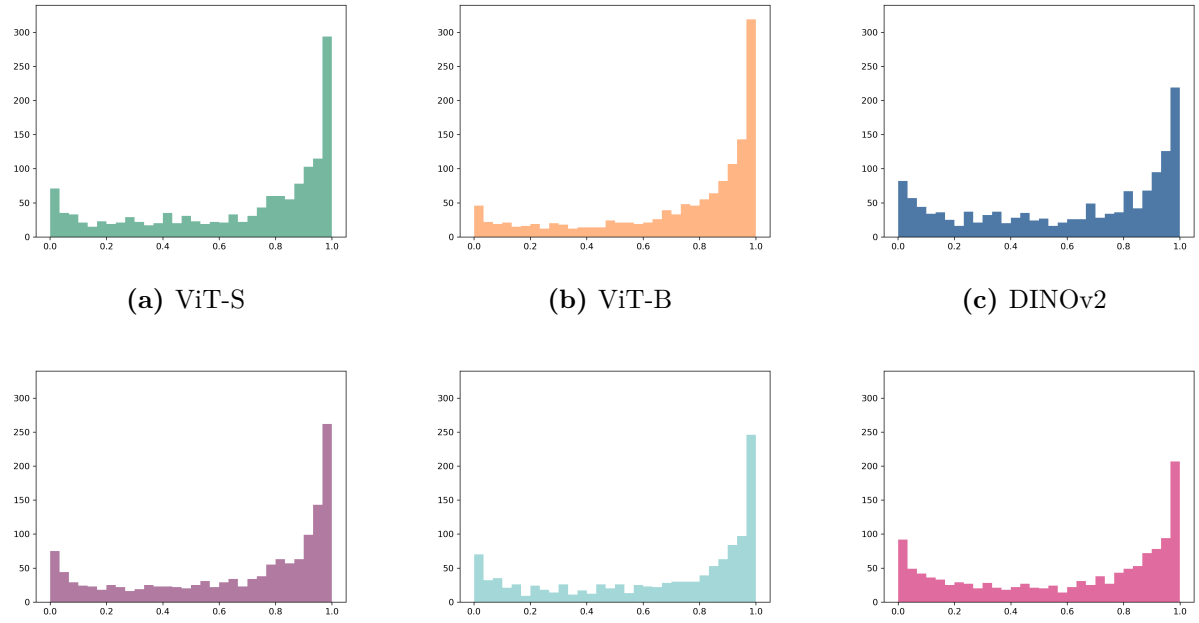
and training objectives: CLIP ViT-L/14, OpenCLIP ViT-B/16 (LAION-2B), OpenCLIP ConvNeXt-B, SigLIP, SigLIP2, and BLIP (ITM). Each encoder is used to compute the same image-text cosine similarity score used for the main-text results ($n = 4470$), and correlated (Spearman) against the CLIP ViT-B/16 score at the trial level. Table 11 reports the resulting per-backbone median scores and correlations. Four of the six alternative encoders closely replicate the main-text ranking ($\rho = .83$–$.94$, all $p < .001$); SigLIP2 and BLIP correlate slightly more weakly ($\rho = .74$–$.75$). Regardless, across all VLM scorers, among ViT-B, the supervised one stands at the top and SigLIP at the bottom.

Table 11. *Plausibility* scores per model, scored with CLIP ViT-B/16 (main text) and six alternative vision-language encoders, using the 9-crop feature-visualization images. ρ is the trial-level Spearman correlation with the CLIP ViT-B/16 score ($n = 4470$ trials, all $p < .001$). Bold is best, italic is second best, underline is worst.

Encoder	ViT-S	ViT-B	DINOv2	DINOv3	CLIP	SigLIP	ρ
CLIP ViT-B/16	**0.274**	*0.273*	0.259	0.260	0.269	<u>0.253</u>	–
CLIP ViT-L/14	**0.225**	*0.224*	0.205	0.205	0.213	<u>0.194</u>	0.907
OpenCLIP ViT-B/16 (LAION-2B)	*0.238*	**0.239**	0.222	0.221	0.226	<u>0.205</u>	0.870
OpenCLIP ConvNeXt-B	**0.356**	*0.354*	0.337	0.338	0.344	<u>0.320</u>	0.860
SigLIP (ViT-B/16)	**0.087**	*0.082*	0.045	0.043	0.045	<u>0.021</u>	0.791
SigLIP2 (SO400M)	**0.061**	*0.054*	0.027	0.030	0.031	<u>0.012</u>	0.706
BLIP	*0.391*	**0.396**	0.372	0.370	0.316	<u>0.264</u>	0.699

B.11 Detailed results for the psychophysics experiments.

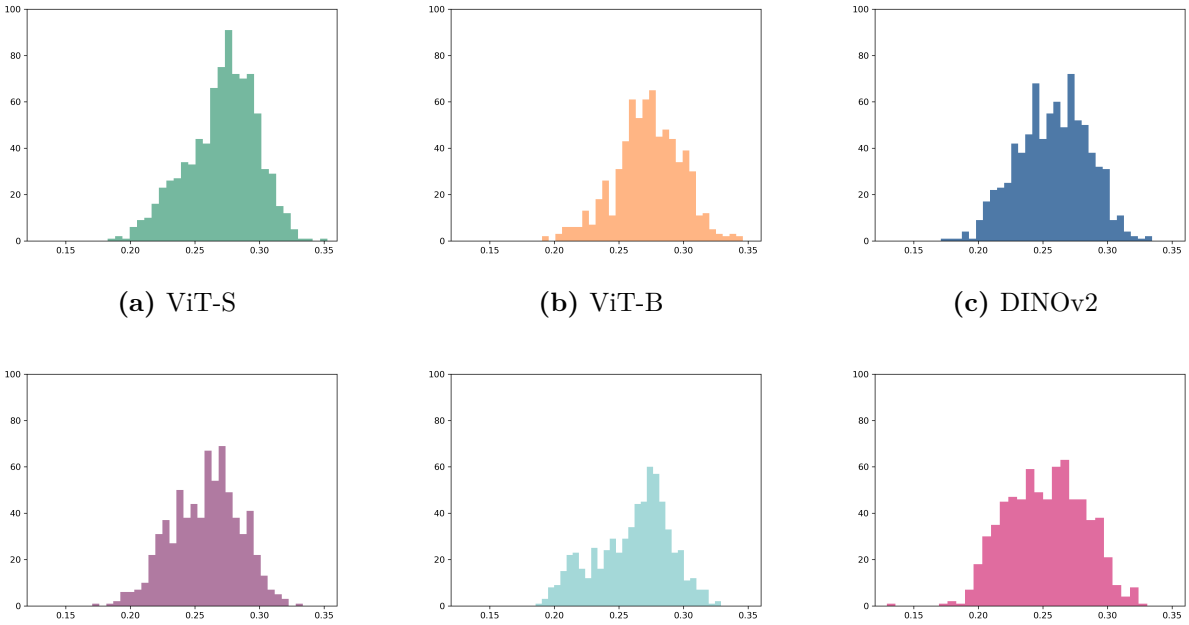
Across all protocols, interpretability scores were non-normally distributed (Fig. 25, Fig. 26, and Fig. 26) for every model (Shapiro–Wilk test, $p < .001$). We therefore report the median score per model as a more representative measure of central tendency.



(a) ViT-S **(b)** ViT-B **(c)** DINOv2

(d) DINOv3 **(e)** CLIP **(f)** SigLIP

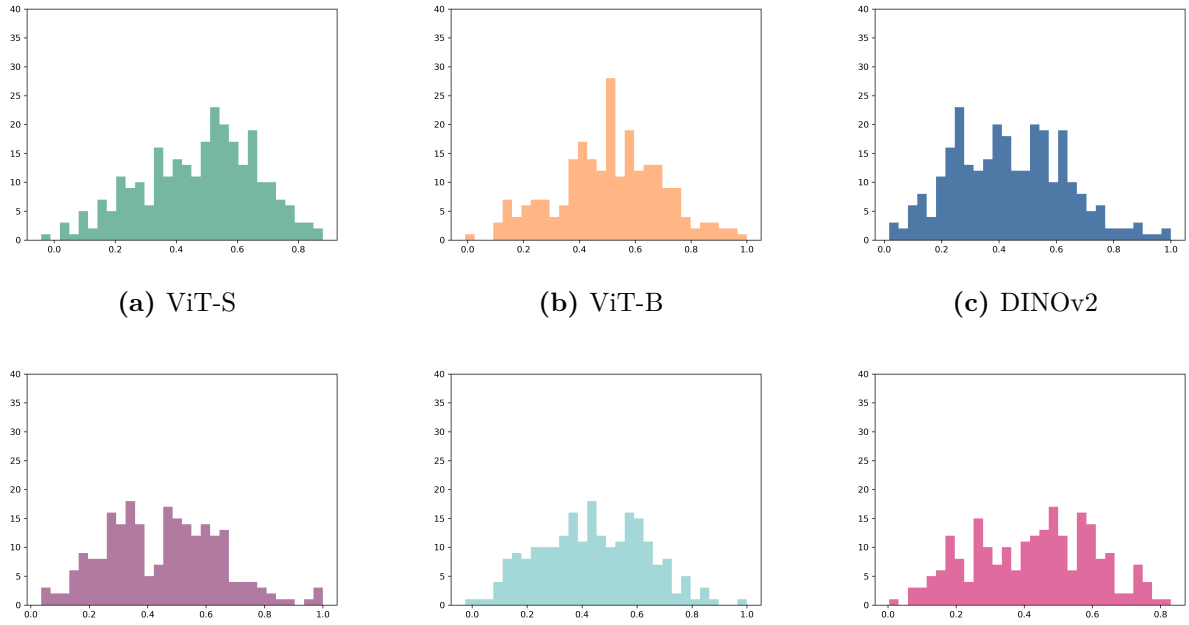
Figure 25. Distribution of feature *localizability* scores across models. Each panel shows the distribution of *localizability* scores for one model.



(a) ViT-S **(b)** ViT-B **(c)** DINOv2

(d) DINOv3 **(e)** CLIP **(f)** SigLIP

Figure 26. Distribution of feature *semanticity* scores across models. Each panel shows the distribution of *semanticity* scores for one model.



(a) ViT-S **(b)** ViT-B **(c)** DINOv2

(d) DINOv3 **(e)** CLIP **(f)** SigLIP

Figure 27. Distribution of feature *semanticity* scores across models. Each panel shows the distribution of *semanticity* scores for one model.

B.12 Additional results on locality of representations

A feature can be considered local if it fires on a few grouped pixels, i.e., (i) *few* pixels are active and (ii) those active pixels are *spatially grouped*. The Hoyer sparsity used in Section 4.4.1 primarily targets the first component. As a complementary proxy for the latter, we use the **Kolmogorov complexity** [116] of each feature's RISE heatmaps, approximated by their JPEG compressibility: the size of the losslessly compressed heatmap, normalized by its uncompressed size. Grouped activations form contiguous, redundant regions and compress well; scattered activations do not. Lower values therefore indicate more spatially grouped heatmaps. We average over the heatmaps of each feature's nine most activating images and then over all features for a model.

Results. Across the six models, compressibility correlates across all three interpretability measures in the expected direction—interpretable models tend to have more compressible heatmaps—although only significant for *agreement*: *localizability* ($\rho = -0.63$, $p = 0.178$), *plausibility* ($\rho = -0.95$, $p = 0.052$), and *agreement* ($\rho = -0.88$, $p = 0.021$).

Limitation. Compressibility alone does not distinguish a heatmap with a single active pixel and a heatmap where every pixel takes the same value, as both are highly compressible. The score, therefore, captures the presence and, to some extent, the number of activation groups, but not whether those groups are spatially confined. We therefore read it as a useful complement to Hoyer sparsity: suggestive of grouping, but, on its own, an incomplete proxy for locality.

B.13 Additional results on alignment with human similarity judgments

The Levels dataset [139] provides human similarity judgment at three different levels, by constructing triplets using 3 images from 1 class to 3 different classes. For completeness, in this section we report results from the third triplet category beyond the coarse- and fine-grained ones reported in Section 4.4.2. In this *border-level* category, two images share the same ImageNet class while the third belongs to a different ImageNet class, probing a granularity that sits between fine and coarse similarity.

The Spearman correlations with border-level alignment fall between the coarse and fine-grained results, with positive but non-significant trends across all three measures: *localizability* ($\rho = 0.58$, $p = 0.223$), *plausibility* ($\rho = 0.72$, $p = 0.11$), and *agreement* ($\rho = 0.42$, $p = 0.402$). The pattern is consistent with interpretability tracking categorical alignment more closely than within-category alignment.

Appendix C

Appendix of Chapter 5

C.1 LoRA fine-tuning on THINGS

Each backbone is fine-tuned on the full THINGS triplet training set using Low-Rank Adaptation (LoRA) [91]. For supervised ViT-B/16, DINOv2, DINOv3, and CLIP ViT-B/16, the training representation concatenates the CLS token with the mean of the patch tokens, yielding a 1536-dimensional vector. Because the SigLIP vision encoder we use has no CLS token, we instead use its 768-dimensional mean-pooled patch representation. A linear projection head maps these representations to 1024 dimensions during training. The head is discarded afterward, and all downstream analyses use the corresponding backbone output without projection.

LoRA adapters use rank $r = 8$ and scaling $\alpha = 1.0$ and are inserted into the attention query, key, and value projections of every transformer block. For DINOv2 and DINOv3, these projections are implemented separately, resulting in $12 \times 3 = 36$ adapted matrices per backbone. For supervised ViT-B/16, CLIP ViT-B/16, and SigLIP, the adapters target the fused `qkv` projection in each of the 12 transformer blocks. All other parameters are frozen.

We optimize the contrastive triplet objective of Muttenthaler, Linhardt, Dippel, et al. [140], which applies a softmax over the three pairwise similarities in each odd-one-out triplet. Training uses temperature $\tau = 1.0$, Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, no weight decay, and a peak learning rate of 10^{-4}. The learning rate increases linearly over the first 500 steps and then follows cosine decay. We train for 20,000 steps.

C.2 Direct matching across independently trained SAEs

Before introducing the shared dictionary, we attempted to match features directly between the independently trained baseline and aligned SAEs. Because the two dictionaries do not share a coordinate system, decoder directions cannot be compared directly. Instead, for each feature, we searched the other SAE for the feature with the highest null-adjusted co-firing score among all 7680 candidates and accepted the match when this score exceeded 0.5.

This procedure yielded few confident correspondences, particularly for the foundation models (Table 12). The independently trained dictionaries are not constrained to decompose their representations into comparable features, so even related features may be expressed by different combinations of SAE dimensions. We therefore discarded direct SAE-to-SAE matching and use the jointly trained dictionary described in the main Methods.

Table 12. Confident correspondences recovered by direct matching across independently trained SAEs. Percentages are relative to the number of informative features considered for each backbone.

Backbone	Informative features	Confident matches
ViT-B/16	1029	311 (30.2%)
CLIP ViT-B/16	417	40 (9.6%)
DINOv2	667	58 (8.7%)
DINOv3	426	51 (12.0%)
SigLIP	561	66 (11.8%)

C.3 Sensitivity to the co-firing cutoff

Full results behind the rejected cutoff-based attempt. Note this is a different population from dropped/introduced (Methods, no correspondent at all): every feature here already has a correspondent (passed the similarity threshold); the question is only whether it persists strongly or weakly. At each co-firing threshold $c \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, matched features above it count as *shared*, matched features below it split into *unique-baseline* and *unique-aligned*, and we take the median interpretability score of each group. Bold marks which group ranks highest at that cutoff.

Table 13. Group with the highest median interpretability under different co-firing cutoffs. S denotes shared features, U_{B} unique-baseline features, and U_{A} unique-aligned features.

Backbone	c**=0.3**	c**=0.4**	c**=0.5**	c**=0.6**	c**=0.7**
ViT-B/16	U_{A}	U_{A}	U_{A}	U_{A}	U_{A}
CLIP ViT-B/16	U_{A}	U_{A}	U_{A}	S	S
DINOv2	S	S	S	U_{A}	U_{A}
DINOv3	U_{A}	U_{A}	U_{A}	U_{A}	U_{A}
SigLIP	S	S	S	S	S

The inferred ordering changes with the cutoff for ViT-B/16, CLIP, DINOv2, and SigLIP. Only DINOv3 is stable, consistently favoring unique-aligned features.

Thus, for two of the five backbones, the conclusion about whether strongly or weakly persisting features are more interpretable depends on an arbitrary threshold. We therefore treat co-firing as a continuous quantity in the main analysis rather than dividing features using a fixed cutoff.

C.4 Image-level co-firing.

Our primary co-firing measure requires matched features to activate at the same spatial locations. As a broader alternative, image-level co-firing only requires them to activate somewhere within the same image, allowing for shifts in where or through which visual cues a feature is expressed.

Repeating the persistence–interpretability analysis at the image level preserves the direction of the association for four of the five backbones; ViT-B/16 is close to zero at both scales. DINOv3's pattern becomes stronger ($\rho = -0.34$, $p < .001$, compared with $\rho = -0.29$, $p = 0.002$ at the patch level). DINOv2 shows a weak association at the image level ($\rho = -0.11$, $p = 0.026$), and no association at the patch level ($\rho = -0.02$, $p = 0.68$).

C.5 Sensitivity to the similarity threshold

All feature correspondences rest on one threshold: an individual-SAE feature is matched to its closest shared-dictionary dimension only when the cosine similarity between their decoder directions exceeds 0.7. We use 0.7 because it is the value most commonly adopted to decide that two SAE latents represent the same feature [96, 114, 153]. It remains a convention rather than a calibrated quantity, and we do not know whether a lower or a higher value is more meaningful here. We therefore keep the standard value as the default and, for completeness, repeat every correspondence-based analysis at 0.5, 0.6, and 0.8.

Alignment scores, task performance, aggregate interpretability, locality, and texture reliance do not depend on feature correspondence and are unaffected. Table 14 reports the correspondence-based results. The image-level association between co-firing and interpretability in DINOv3, its counterpart with locality, the absence of a patch-level association in DINOv2, and the larger texture-reliance decrease among broadened features keep their direction and significance at every threshold. The patch-level association between co-firing and interpretability in DINOv3 does not survive a lower threshold: it is significant only at 0.7. Introduced and dropped features do not differ significantly at any threshold, except for CLIP at 0.6 and 0.8, where introduced features score lower, and the DINOv3 difference between persistence regimes is not significant at any threshold.

Table 14. Correspondence-based results across similarity thresholds. Entries are Spearman ρ or median differences in interpretability or texture reliance; $^*p < 0.05$, uncorrected. The number of features entering each analysis shrinks as the threshold rises. The default threshold is 0.7.

Result	Backbone	0.5	0.6	0.7	0.8
Aligned features with a correspondent (%)	ViT-B/16	82	71	53	31
	CLIP ViT-B/16	64	56	45	28
	DINOv2	95	84	63	34
	DINOv3	75	54	28	9
	SigLIP	86	80	71	58
Introduced − dropped interpretability	ViT-B/16	−0.013	+0.014	+0.017	+0.009
	CLIP ViT-B/16	−0.074	−0.094*	−0.059	−0.068*
	DINOv2	+0.140	+0.143	+0.079	+0.062
	DINOv3	+0.026	+0.006	+0.019	+0.013
	SigLIP	+0.029	+0.075	+0.085	+0.057
Co-firing vs. interpretability, patch (ρ)	DINOv2	−0.006	−0.029	−0.021	+0.015
	DINOv3	−0.050	−0.078	−0.289*	−0.084
Co-firing vs. interpretability, image (ρ)	DINOv2	−0.087*	−0.097*	−0.114*	−0.031
	DINOv3	−0.175*	−0.225*	−0.339*	−0.348*
Co-firing vs. locality, image (ρ)	DINOv3	−0.200*	−0.234*	−0.265*	−0.351*
Texture change, broadened − preserved	all	−0.091*	−0.090*	−0.091*	−0.068*
Interpretability change, preserved − broadened	DINOv3	+0.041	+0.054	+0.065	+0.215

C.6 Informative-feature loss across backbones.

Table 15 reports how the number of informative features changes after alignment. All five aligned models contain fewer informative features than their baseline counterparts, but the reduction varies substantially across backbones. SigLIP loses 37% of its informative features, whereas DINOv3 loses the most (58%) despite showing a positive directional change in interpretability. The decline observed for SigLIP therefore cannot be explained simply by unusually strong feature loss or survivorship.

Table 15. Number of informative features before and after alignment. Percentage loss is computed relative to the baseline count.

Backbone	Baseline	Aligned	Loss
ViT-B/16	1259	1029	18%
CLIP ViT-B/16	551	417	24%
SigLIP	896	561	37%
DINOv2	1119	667	40%
DINOv3	1018	426	58%

C.7 Accuracy-change buckets and interpretability.

To complement the continuous correlation reported in the main text, we divide the 559 curated image categories separately for each backbone into groups according to their change in classification accuracy after alignment. These groups contain the categories whose accuracy decreased most, changed least, or increased most. Table 16 reports the mean interpretability change within each group, measured as aligned minus baseline localizability.

Table 16. Mean interpretability change within model-specific groups of classification-accuracy change. Positive values indicate increased localizability after alignment.

Backbone	Accuracy decreased	Accuracy stable	Accuracy increased
ViT-B/16	+0.005	+0.012	+0.008
CLIP ViT-B/16	+0.032	−0.006	−0.010
DINOv2	+0.025	+0.002	+0.056
DINOv3	−0.014	+0.044	+0.019
SigLIP	−0.071	+0.007	−0.004

If interpretability directly followed classification accuracy, interpretability changes should become progressively more positive from the decreased to the increased groups. No such ordering appears consistently across backbones. For example, CLIP shows its largest interpretability increase among categories whose accuracy decreased most, whereas DINOv3 shows its largest increase in the stable group. These bucketed results agree with the continuous image-category analysis reported in the main text, which finds no association between accuracy and interpretability changes when pooling all backbones (Spearman $\rho = 0.014$, $p = 0.47$) and only weak, nonsignificant associations within individual backbones ($|\rho| \leq 0.08$).

C.8 Joint-dictionary correspondence and feature turnover

The feature-turnover analysis uses the jointly trained dictionary to determine whether each individual-SAE feature has a correspondent in the other representation. A feature is assigned to its closest shared-dictionary dimension by decoder-direction cosine similarity. We retain the candidate when the similarity exceeds 0.7 and the same shared dimension activates in the other model. features failing either condition are treated as having no cross-model correspondent.

A baseline feature without an aligned-side correspondent is classified as *dropped*; an aligned feature without a baseline-side correspondent is classified as *introduced*. Table 17 reports the resulting populations used in the feature-turnover analysis.

Table 17. Dropped and introduced informative features identified through joint-dictionary matching. Percentages are relative to the corresponding baseline or aligned feature population.

Backbone	Baseline	Dropped	Aligned	Introduced
ViT-B/16	1259	309 (24.5%)	1029	487 (47.3%)
CLIP ViT-B/16	551	270 (49.0%)	417	230 (55.2%)
DINOv2	1119	365 (32.6%)	667	253 (37.9%)
DINOv3	1018	511 (50.2%)	426	304 (71.4%)
SigLIP	896	360 (40.2%)	561	180 (32.1%)

Depending on the backbone and matching direction, between 24.5% and 71.4% of informative features have no confident correspondent in the other representation.

C.9 Paired interpretability changes for matched features

As a complementary feature-level analysis, we directly compare the interpretability of features that can be confidently matched between the baseline and aligned representations. For each matched pair, we compute the change in localizability, $\Delta = s_{\text{aligned}} - s_{\text{baseline}}$, and test whether the median change differs from zero using a paired Wilcoxon signed-rank test. The pairing is performed at the feature level; baseline and aligned features were evaluated by independent participant samples, making these estimates comparatively noisy and best interpreted as exploratory.

Table 18. Paired interpretability changes for confidently matched features. Positive values indicate higher localizability after alignment. The Wilcoxon signed-rank test is applied to the per-feature differences $\Delta = s_{\text{aligned}} - s_{\text{baseline}}$.

Backbone	n **pairs**	**Mean** Δ	**Median** Δ	W	p
ViT-B/16	322	−0.006	+0.000	25271	0.66
CLIP ViT-B/16	87	+0.031	−0.009	1766	0.53
DINOv2	206	+0.052	+0.031	9168	0.08
DINOv3	61	+0.005	−0.020	929	0.91
SigLIP	193	+0.038	+0.025	8179	0.13

None of the paired changes reaches statistical significance. DINOv2 shows the largest mean and median increase and comes closest to significance ($p = 0.08$), whereas ViT-B/16 and DINOv3 remain close to zero. The paired analysis therefore provides no evidence of a uniform interpretability shift among matched features. For DINOv2, the positive but nonsignificant distribution is consistent with an improvement spread broadly across features; for DINOv3, the near-zero average supports the interpretation that its directional aggregate change is concentrated among particular features rather than shared uniformly across matched pairs.

Parte II

Resumen en Español

Implementación Eficaz de la IA Fiable

C.10 El problema de la interpretabilidad

Desde el avance decisivo de AlexNet en 2012 [110], los modelos de aprendizaje profundo han progresado a un ritmo extraordinario, y sus capacidades han crecido más rápido que nuestra capacidad para comprenderlos. Esta brecha de interpretabilidad resulta especialmente relevante cuando estos modelos se utilizan para tomar decisiones con consecuencias importantes, como en la interpretación de imágenes médicas [11, 37, 98], la conducción autónoma [68, 97] o el control robótico [105]. Los modelos son suficientemente precisos como para desplegarse en escenarios reales, pero siguen siendo en gran medida opacos para sus usuarios. Sin embargo, comprenderlos es importante por diversos motivos, como identificar fallos antes de que causen daños [36, 70], satisfacer las crecientes exigencias legales en materia de explicación [101] y aprender de ellos cuando se utilizan para realizar nuevos descubrimientos científicos [55, 100]. A medida que los modelos adquieren mayores capacidades y se despliegan de forma más generalizada, aumenta la brecha entre lo que son capaces de hacer y lo que comprendemos de ellos [87], haciendo de la interpretabilidad una necesidad cada vez más urgente [103]. Estas presiones empiezan además a adquirir relevancia regulatoria: la interpretabilidad aparece cada vez con mayor frecuencia en los marcos de certificación—de manera especialmente explícita en la EASA, que incluye la "explicabilidad de la IA" como uno de los pilares de la confianza en la aviación [53], y en la Ley de Inteligencia Artificial de la Unión Europea, cuyas disposiciones relativas a sistemas de alto riesgo exigen que estos sean suficientemente transparentes para que las personas puedan interpretar y supervisar sus resultados [52]—una expectativa que, aunque todavía no constituye un requisito estricto, aumenta proporcionalmente con el riesgo.

En su sentido más exigente, la interpretabilidad requiere algo más que una explicación asociada a una predicción individual. Requiere que una persona pueda construir un modelo mental útil del sistema: que pueda inspeccionar sus representaciones, anticipar su comportamiento ante casos nuevos y prever sus fallos antes de que se produzcan. Cómo construir modelos que permitan este tipo de comprensión humana sigue siendo una cuestión abierta.

Explicabilidad frente a interpretabilidad. Ambos términos se utilizan a menudo de manera intercambiable. Incluso el propio marco de la EASA incluye la 'explicabilidad de la IA" [53] como el pilar de confianza correspondiente al tipo de comprensión descrito anteriormente. Sin embargo, ambos conceptos hacen referencia a aspectos distintos, y parte del prolongado desacuerdo en torno a qué significa realmente la 'interpretabilidad" [121] proviene precisamente de esta confusión. En esta tesis utilizamos *explicabilidad* para describir una propiedad de un sistema: si este puede producir una explicación –por ejemplo, un mapa

de saliencia, una puntuación asociada a un concepto, una correspondencia con un prototipo o una justificación en lenguaje natural– destinada a comunicar la base de una salida o el contenido de una representación. Utilizamos *interpretabilidad*, en cambio, para describir una propiedad de la interacción entre dicho sistema y una persona: si la explicación, una vez proporcionada, permite que esa persona alcance una comprensión correcta [45, 134]. La explicabilidad constituye una afirmación sobre un mecanismo y puede verificarse inspeccionando el sistema, mientras que la interpretabilidad constituye una afirmación sobre un estado mental y solo puede verificarse evaluando a la persona. Una explicación puede estar disponible, e incluso ser fiel al funcionamiento del modelo, y aun así no ser interpretable si no resulta adecuada para que una persona pueda procesarla [95], o si resulta comprensible para una predicción concreta pero nunca llega a integrarse en un modelo mental del comportamiento más general del sistema. Por tanto, la explicabilidad es necesaria, pero no suficiente, para la interpretabilidad.

Dos vías para lograr modelos interpretables. Los enfoques existentes siguen, a grandes rasgos, dos vías para lograr interpretabilidad. La primera incorpora la transparencia directamente en la arquitectura o en el proceso de aprendizaje del modelo, dando lugar a modelos *explicables por diseño* [162][11] mediante diversos métodos, entre ellos redes de prototipos que clasifican comparando las entradas con ejemplos visualizables [29], modelos de cuello de botella conceptual que hacen pasar las predicciones a través de conceptos definidos por humanos [108], o redes autoexplicativas que generan una explicación como parte de su propia pasada hacia delante [6, 16]—intercambiando flexibilidad arquitectónica por transparencia.

La segunda vía, denominada habitualmente *inteligencia artificial explicable post hoc* (XAI) [73], parte de un modelo opaco ya entrenado y desarrolla métodos para explicar sus predicciones o sus representaciones internas a posteriori. En la práctica, sin embargo, la frontera entre ambos paradigmas es menos nítida de lo que podría parecer inicialmente. Por ejemplo, los modelos de cuello de botella conceptual y el descubrimiento post hoc de conceptos mediante codificación dispersa pueden converger hacia conceptos similares [160]. Más fundamentalmente, ambos enfoques comparten la misma limitación. Aunque garantizan que una explicación esté *disponible*, ninguno garantiza que una persona llegue realmente a *comprender* el modelo resultante. Determinar si cualquiera de estas dos vías consigue fomentar una auténtica comprensión humana sigue siendo, en gran medida, una cuestión empírica abierta.

En el caso de los modelos explicables por diseño, la evaluación suele centrarse no en determinar si las personas comprenden el modelo, sino en utilizar indicadores indirectos de dicha comprensión. Por ejemplo, en los modelos basados en conceptos, esto suele consistir en comprobar si se recuperan los conceptos etiquetados por humanos que se pretendía representar. Sin embargo, trabajos recientes sugieren que estas representaciones aprendidas pueden funcionar en realidad como sustitutos metonímicos, de modo que cada concepto recuperado codifica información sobre múltiples conceptos adicionales [7]. Por tanto, recuperar los conceptos previstos no implica necesariamente que el razonamiento del modelo sea transparente para los seres humanos.

La XAI post hoc se encuentra en una situación algo más favorable, en gran medida porque

[11] El término utilizado habitualmente en la literatura es *interpretable por diseño*. Dada la distinción anterior, consideramos que *explicable por diseño* es más preciso: estos métodos garantizan que exista una explicación, no que el modelo sea comprendido.

ha recibido una mayor atención por parte de la comunidad investigadora. Se ha propuesto y evaluado una amplia variedad de métodos. En el caso de la visión por computador, que constituye el foco de esta tesis, existen tres grandes familias de métodos de XAI—atribución, visualización de características y métodos basados en conceptos, que se describen con mayor detalle en el Capítulo 2—y todas ellas han sido objeto de una evaluación humana considerablemente mayor que los modelos explicables por diseño [3, 17, 27, 31, 60, 80, 106, 145, 158, 167, 173, 203]. Sin embargo, estas evaluaciones se han centrado principalmente en comparar métodos de XAI entre sí, en lugar de comparar la interpretabilidad de los modelos subyacentes, una carencia sobre la que vuelve la Sección C.12. Como consecuencia, seguimos sabiendo poco acerca de hasta qué punto son interpretables distintos modelos o de si el campo está realizando progresos significativos hacia ese objetivo. Más fundamentalmente, decidir si los modelos explicables por diseño o los enfoques de XAI post hoc constituyen en última instancia la vía más prometedora presupone disponer primero de una forma de medir la comprensión del modelo. Y esa es precisamente la capacidad de la que el campo carece actualmente.

Medición de la comprensión humana. Una de las razones por las que resulta tan difícil obtener evidencia de que las personas comprenden realmente un modelo—en lugar de limitarse a considerar plausibles o reconocibles sus explicaciones—es la falta de consenso acerca de cómo medir la interpretabilidad de los modelos. En la literatura encontramos diversas caracterizaciones [45, 122] que, en general, solo comparten la idea de que la interpretabilidad es en última instancia un concepto centrado en el ser humano: una explicación tiene valor si una persona puede comprenderla [133]. Medir dicha comprensión constituye un problema en sí mismo. Las medidas de autoinforme no son suficientes, ya que las personas tienden de manera sistemática a sobreestimar su propia comprensión de un sistema [19, 161]. Una medida útil debe, en cambio, evaluar aquello que dicha comprensión permite hacer.

En ciencia, la comprensión suele demostrarse mediante la capacidad de generalizar correctamente: la mecánica newtoniana resulta valiosa como explicación del movimiento planetario porque permite predecir trayectorias más allá de las observaciones a partir de las cuales fue formulada. La interpretabilidad requiere una prueba funcional análoga. Una amplia literatura en psicología sostiene que las explicaciones favorecen el aprendizaje al ayudar a las personas a extraer reglas que se generalizan más allá de los casos observados previamente [123-127, 192, 195]. Diversos protocolos de interpretabilidad operacionalizan esta idea de manera directa. Evalúan si una explicación ayuda a una persona a predecir la decisión de un modelo ante una entrada no observada previamente [31], inferir qué estímulos activarán una característica interna [17, 203], identificar sesgos aprendidos por el modelo [158, 173], distinguir predicciones correctas de incorrectas [106], o mejorar el rendimiento humano y el de los equipos humano–IA [5, 145, 166, 184]. Qué comportamiento proporciona la medida más informativa y representativa de la comprensión de un modelo sigue siendo una cuestión abierta.

Esta tesis se centra en modelos de aprendizaje profundo preentrenados de propósito general. Este énfasis es deliberado, ya que estos modelos son los que se despliegan con mayor frecuencia en aplicaciones críticas y, por tanto, aquellos para los que reducir la opacidad resulta más urgente. No obstante, el marco desarrollado en esta tesis para *medir*, *comparar* y *moldear* la interpretabilidad no depende de si un modelo ha sido diseñado para ser intrínsecamente transparente o si se explica mediante métodos post hoc.

C.11 Una cuestión especialmente pertinente en este momento

Aunque la interpretabilidad de los modelos sigue siendo un ámbito poco explorado, varios desarrollos están convergiendo para facilitar su estudio: la creciente madurez de los métodos de XAI post hoc, la convergencia de los modelos de visión por computador en torno a arquitecturas compartidas y la aparición de métodos para comparar representaciones a niveles de granularidad cada vez más finos. Por ello, sostenemos que el campo ha alcanzado la madurez necesaria para pasar del desarrollo de herramientas de XAI a la comparación de la interpretabilidad de los modelos.

Las herramientas imperfectas pueden sustentar una ciencia acumulativa. Los métodos de XAI post hoc han recibido una atención considerable, pero siguen siendo imperfectos: los mapas de saliencia pueden resultar insensibles a la aleatorización de los pesos del modelo o de las etiquetas de entrenamiento [2], las atribuciones pueden ser frágiles frente a perturbaciones imperceptibles de las entradas [71], y diferentes métodos pueden discrepar entre sí al explicar una misma predicción [107, 188]. Sin embargo, esto no constituye un problema en sí mismo: utilizar herramientas imperfectas pero informativas para construir una comprensión acumulativa de un sistema complejo es una práctica recurrente en la ciencia, y la neurociencia de sistemas proporciona un precedente claro. Décadas de registros con electrodos individuales y estímulos diseñados manualmente, rudimentarios según los estándares actuales, permitieron ya establecer un núcleo duradero de conocimientos sobre el procesamiento visual, desde las unidades selectivas a la orientación en las primeras áreas de la corteza visual [92] hasta neuronas altamente selectivas en regiones más avanzadas de la vía ventral [78, 156]. Posteriormente se demostró que esta imagen era incompleta: dado que los registros con electrodos individuales muestrean una neurona cada vez, favorecieron una concepción serial y feedforward de la visión, hasta que la evidencia anatómica y fisiológica mostró que la corteza presenta una densa recurrencia, con conexiones de retroalimentación que modulan las respuestas en cada etapa [64, 112]. Sin embargo, estos nuevos hallazgos no invalidaron tanto los resultados anteriores como los enriquecieron, y el procesamiento feedforward rápido sigue constituyendo una parte real y central de nuestra comprensión científica de la visión.

El mismo principio motiva el uso de los métodos actuales de XAI y, en esta tesis, argumentamos en el Capítulo 2 que el conjunto actual de herramientas de XAI post hoc es suficientemente informativo como para comenzar a construir una comprensión acumulativa de las representaciones de los modelos. Estudiar los modelos utilizando los métodos de XAI actuales permitirá asimismo identificar en qué aspectos esas propias herramientas todavía deben mejorar.

Un panorama de modelos cada vez más convergente. Un segundo desarrollo es la convergencia arquitectónica de los modelos de visión por computador. Durante más de una década, las redes convolucionales (CNN) dominaron la visión por computador, especialmente ResNet50 [83], que sigue siendo uno de los backbones visuales más utilizados. Desde entonces, el campo ha convergido progresivamente en torno a arquitecturas transformer, incluso a medida que los objetivos de entrenamiento se han diversificado. Una misma plantilla arqui-

tectónica general sustenta actualmente los Vision Transformers supervisados [46], modelos contrastivos supervisados mediante lenguaje como CLIP y sus sucesores basados en pérdida sigmoide [157, 199], la familia DINO auto-supervisada, actualmente en su tercera generación [25, 152, 170], el modelado de imágenes enmascaradas (MAE) [82], nuevos objetivos predictivos de tipo modelo del mundo como I-JEPA [8], y sus extensiones al vídeo [9, 12, 189, 193].

Esta convergencia hace que las comparaciones entre modelos sean sustancialmente más informativas. Al mantener fija la arquitectura, las diferencias en interpretabilidad pueden atribuirse de forma más clara a los datos de entrenamiento y a los objetivos de aprendizaje, en lugar de quedar completamente confundidas con diferencias arquitectónicas. Y dado que se trata de una arquitectura sobre la que actualmente trabaja gran parte del campo, cualesquiera que sean las diferencias observadas tendrán una relevancia especialmente amplia.

Comparación de modelos a escalas progresivamente más finas. Incluso cuando las arquitecturas son comparables, estudiar las diferencias entre modelos resulta difícil porque cada representación se expresa en su propia base idiosincrásica. Los métodos existentes abordan este problema a distintos niveles de resolución. A un nivel grueso, los primeros trabajos compararon modelos utilizando cantidades agregadas, como el número de detectores de conceptos únicos presentes en un modelo [13, 65], o una puntuación media de interpretabilidad humana calculada sobre una muestra de sus características [32, 204]. Estas comparaciones permiten establecer si los modelos difieren globalmente, pero su naturaleza agregada ofrece una capacidad limitada para aislar los factores específicos responsables de dichas diferencias. A un nivel más fino, diversos trabajos recientes han propuesto métodos como los *crosscoders* y los autoencoders dispersos universales, que aprenden espacios conceptuales compartidos entre modelos [76, 102, 119, 186]. Al permitir comparaciones a nivel de características individuales, estos métodos facilitan la identificación de los factores que determinan la interpretabilidad. En esta tesis adoptamos esta misma progresión hacia niveles de análisis cada vez más granulares.

C.12 De explicar los modelos a reducir su opacidad

A pesar de esta convergencia favorable entre herramientas maduras de XAI y arquitecturas compartidas, pocos trabajos han utilizado de manera sistemática métodos post hoc para estudiar las representaciones internas de los modelos y compararlas entre sí.

Los modelos como unidad de análisis. Estudios detallados han cartografiado circuitos y características visuales dentro de arquitecturas individuales [26, 148, 149], identificado neuronas multimodales en CLIP [75], descompuesto sus representaciones entre diferentes cabezas de atención [67] y realizado ingeniería inversa del papel de los tokens de registro en los transformadores de visión [38]. En conjunto, estos estudios demuestran el poder explicativo del conjunto de herramientas de XAI cuando se utiliza para investigar los propios modelos, en lugar de limitarse a generar explicaciones.

Estudiar un modelo con este nivel de detalle es también lo que permite revelar las limitaciones del propio conjunto de herramientas de interpretabilidad. Un ejemplo paradigmático procede del análisis a nivel de circuitos de neuronas individuales en los trabajos anterior-

mente citados [149]. Al llevar este análisis detallado a una escala suficiente, se descubrieron neuronas *polisemánticas*, que respondían a combinaciones de conceptos aparentemente no relacionados en lugar de a un único patrón identificable, una propiedad que posteriormente se formalizó como superposición [50]. Este hallazgo cuestiona casi por construcción una descripción de las representaciones del modelo centrada en las neuronas: si una neurona individual no tiene por qué corresponderse con un concepto coherente, entonces las propias neuronas no pueden constituir una unidad fiable de interpretación.

Este descubrimiento motivó un cambio hacia el aprendizaje de diccionarios como base representacional alternativa. Mediante el aprendizaje de un código disperso y sobredeterminado sobre las activaciones de una capa, métodos como los autoencoders dispersos pretenden recuperar direcciones que sean individualmente más monosemánticas que las neuronas de las que se derivan [18]. Desde entonces, estos métodos se han utilizado para revelar características interpretables y susceptibles de ser evaluadas causalmente en modelos de visión modernos [62, 176]. Sin embargo, este progreso no ha resuelto tanto la cuestión de cuál es la unidad apropiada de interpretación como la ha desplazado. Los mismos trabajos que establecieron las ventajas de las representaciones dispersas han revelado también su inestabilidad: las características aprendidas pueden dividirse o fusionarse a medida que cambia la anchura del diccionario [18], y autoencoders entrenados sobre las mismas activaciones con diferentes semillas o hiperparámetros pueden recuperar conjuntos de direcciones marcadamente diferentes y no canónicos [59, 113].

La superposición convierte a la neurona en una unidad de análisis poco fiable; la ausencia de un diccionario estable e identificable hace que la alternativa dispersa constituya una mejora, pero no una solución definitiva. Cuál de las dos—o en qué condiciones cada una de ellas—debería constituir la base de una comparación entre modelos, y si esta elección puede cambiar por sí misma el resultado de dicha comparación, sigue siendo, en el estado actual de la literatura, una cuestión abierta.

Además, los estudios comparativos siguen siendo relativamente escasos. Los trabajos existentes sugieren que la interpretabilidad no mejora necesariamente con la escala o la capacidad del modelo [13, 204, 205], mientras que el etiquetado automático de conceptos, la alineación representacional y los diccionarios de características compartidos se han utilizado para estudiar la cobertura conceptual y la convergencia representacional entre modelos [44, 47, 132, 137, 147, 186]. Al mismo tiempo, los protocolos de evaluación psicofísica proporcionan metodologías maduras para determinar si los seres humanos comprenden realmente las explicaciones de un modelo [17, 27, 60, 106, 158, 184, 203]. Sin embargo, estas evaluaciones humanas se han utilizado principalmente para comparar métodos de XAI y no los propios modelos. Aplicarlas de manera sistemática a la tarea de determinar *qué modelo comprenden mejor los seres humanos* sigue siendo una dirección de investigación en gran medida inexplorada, que esta tesis pretende impulsar mediante el establecimiento de un marco para comparar la interpretabilidad de los modelos de visión modernos.

Los modelos como objeto de intervención. Describir un modelo y compararlo con otros no modifica el propio modelo. Estos enfoques cuantifican la brecha entre lo que un modelo codifica y aquello que los seres humanos pueden comprender, pero no intentan reducirla. Una línea de investigación complementaria se pregunta, en cambio, si esta brecha puede reducirse interviniendo sobre las representaciones del modelo, en lugar de intervenir sobre las herramientas utilizadas para inspeccionarlo. Sin embargo, solo un pequeño número

de estos trabajos optimiza explícitamente la comprensión humana.

Una propuesta reciente en procesamiento del lenguaje natural plantea el problema como un desajuste de vocabulario: los modelos ya codifican numerosos conceptos que carecen de los medios necesarios para expresar, lo que sugiere que la interpretabilidad podría mejorarse enseñándoles un vocabulario conceptual más rico [87]. Sin embargo, no existe un análogo directo en visión, ya que las representaciones visuales carecen de un vocabulario discreto que pueda ampliarse. En su lugar, la visión por computador ha explorado intervenciones a nivel de representación destinadas a hacer que el espacio latente en su conjunto sea más interpretable, por ejemplo promoviendo robustez frente a perturbaciones imperceptibles de las entradas [190, 194]. Una línea de trabajo relacionada y cada vez más activa pretende alinear las representaciones de los modelos con la percepción humana. Se ha demostrado que las representaciones que se corresponden mejor con los juicios humanos de similitud mejoran el aprendizaje con pocos ejemplos, la robustez y la transferencia entre tareas [140, 178]. Partiendo de la observación de que la interpretabilidad está en última instancia limitada por la cognición humana [111], varios trabajos recientes han argumentado además que la alineación perceptiva podría constituir una vía hacia modelos más interpretables [61, 139, 181].

Sin embargo, lo que este ciclo rara vez proporciona es una prueba directa de la hipótesis que lo motivó. En lugar de evaluar si una intervención hace que un modelo resulte más comprensible para los seres humanos, los estudios suelen limitarse a informar de mejoras en el propio criterio de optimización utilizado por la intervención. En esta tesis sostenemos que la interpretabilidad requiere, en última instancia, evidencia conductual: es necesario determinar si las personas pueden predecir mejor el comportamiento del modelo resultante, generalizar mejor a partir de él o razonar mejor sobre su funcionamiento. En consecuencia, el campo ha acumulado un conjunto creciente de principios candidatos para mejorar la interpretabilidad que son plausibles, controlables y ampliamente adoptados, pero que rara vez se han evaluado según el propio criterio que pretenden satisfacer.

C.13 Estructura y contribuciones

Cerrar la brecha identificada anteriormente requiere abordar tres preguntas: dónde buscar la interpretabilidad, cuestión tratada en el Capítulo 3; cómo medir si los modelos son comprendidos por los seres humanos y comparar dicha comprensión entre los actuales modelos fundacionales de visión, cuestión abordada en el Capítulo 4; y hasta qué punto puede diseñarse deliberadamente dicha comprensión, en lugar de limitarse a observarla, cuestión abordada en el Capítulo 5. Estos capítulos trazan un recorrido coherente desde la identificación de la base apropiada para estudiar la interpretabilidad, pasando por su medición entre distintos modelos, hasta la intervención sobre los factores que dichas mediciones ponen de manifiesto. La tesis desarrolla además una perspectiva más amplia sobre cómo debería abordar el campo la interpretabilidad de los modelos, presentada en las secciones anteriores. A continuación, describimos cada una de estas contribuciones con mayor detalle, así como las publicaciones científicas en las que se fundamentan.

Introducción & **Fundamentos** Estos dos capítulos cumplen dos objetivos. El Capítulo 1 contextualiza el estado actual del campo de la inteligencia artificial explicable y argumenta hacia dónde debería avanzar: desde un énfasis en hacer que los modelos sean explicables hacia el estudio y la comparación de hasta qué punto esos modelos son realmente interpretables para los seres humanos, estableciendo así el marco para el resto de la tesis como un primer paso en esta dirección. El Capítulo 2 describe el conjunto específico de herramientas de XAI sobre el que se construye esta tesis y sostiene que dicho conjunto, aunque imperfecto, ha alcanzado la madurez suficiente para sustentar el estudio centrado en los modelos que desarrollan los capítulos posteriores.

- "‘ **Julien Colin**, Nuria Oliver, Thomas Serre. *From Interpretability Methods to Interpretable Models.* ECCV Workshop on Explainable Computer Vision, 2026. "‘

¿Dónde buscar la interpretabilidad? El Capítulo 3 aborda la primera pregunta de la que depende el resto de la tesis: *¿sobre qué unidad de una representación debería medirse la interpretabilidad*? Retoma la cuestión aún no resuelta introducida en §C.12: debido a la superposición, las neuronas individuales constituyen una base poco fiable para la explicación, lo que ha motivado alternativas basadas en el aprendizaje de diccionarios, como los autoencoders dispersos [18] y la factorización matricial [60]. Sin embargo, estas alternativas se han validado en gran medida únicamente mediante criterios orientados a la explicabilidad (§C.10), entre ellos la fidelidad de reconstrucción y la estabilidad entre diferentes semillas, mientras que pocos estudios han abordado la cuestión decisiva de la propia interpretabilidad: ¿qué representación resulta más interpretable para las personas?

Para responder a esta pregunta, el capítulo presenta tres experimentos de psicofísica en línea a gran escala (N=481) utilizando una adaptación del protocolo de predicción de activaciones [17, 203]. Los participantes primero identificaron y posteriormente generalizaron el patrón visual compartido subyacente a las imágenes que activaban al máximo una unidad, lo que permitió comparar directamente representaciones basadas en neuronas y representaciones basadas en diccionarios. Los experimentos se realizaron sobre ResNet50 [83] y VGG16, dos redes convolucionales que durante mucho tiempo han servido como bancos de pruebas para la investigación en interpretabilidad [13, 57, 65]. Centrarse en estos modelos permite aislar la cuestión de la base representacional antes de extender el análisis a los modelos fundacionales basados en transformers, todavía menos comprendidos, que se analizan en §C.11.

Las representaciones basadas en diccionarios resultan sistemáticamente más interpretables que las neuronas, y la interpretabilidad aumenta en las capas más profundas. Este resultado es especialmente relevante porque son precisamente estas capas las que se analizan y reutilizan con mayor frecuencia en la práctica. El grado de alineación entre los elementos aprendidos del diccionario y las neuronas individuales también varía entre modelos, lo que implica que el efecto de la base representacional depende a su vez del modelo. En consecuencia, las comparaciones realizadas a nivel neuronal, que constituyen la hipótesis implícita de muchas de las comparaciones agregadas discutidas en §C.11 [13, 65, 204], pueden ocultar diferencias entre modelos que solo se hacen visibles cuando las representaciones se comparan en una base de diccionario aprendida, como demostramos para ResNet50. Estos resultados proporcionan la primera evidencia basada en juicios humanos a favor de las representaciones

basadas en diccionarios frente a las neuronas y muestran que la propia base representacional puede determinar el resultado de una comparación de interpretabilidad. Esta observación motiva las decisiones metodológicas utilizadas posteriormente en las comparaciones de modelos del Capítulo 4.

- "‘ **Julien Colin**, Lore Goetschalckx, Thomas Fel, Victor Boutin, Thomas Serre, Nuria Oliver. *Choosing the Right Basis for Interpretability: Psychophysical Comparison Between Neuron-based and Dictionary-based Representations*. Transactions on Machine Learning Research (TMLR), 2026. "‘

¿Cómo medir y comparar la interpretabilidad de los modelos? El Capítulo 3 establece dónde buscar la interpretabilidad; el Capítulo 4 pregunta cómo debería medirse. Los protocolos existentes para comparar modelos descritos en §C.11 [32, 204] no pueden aplicarse directamente a esta nueva base representacional. Mostramos que su diseño de elección forzada permite a los participantes acertar únicamente mediante un proceso de eliminación, con un rendimiento limitado por un nivel de azar específico de cada modelo, lo que imposibilita realizar comparaciones justas entre modelos independientemente de la representación subyacente.

Para abordar esta limitación, el capítulo introduce dos protocolos complementarios que sustituyen la tarea de categorización por tareas de identificación. El primero, la *localizabilidad*, pregunta si un observador puede predecir *dónde* se activará una característica en una imagen nueva. El segundo, la *semanticidad*, pregunta si un observador puede describir *qué* representa dicha característica. Las puntuaciones de localizabilidad se calibran respecto a niveles de azar específicos de cada modelo, situando así todos los modelos en una escala común. Además, emparejamos entre modelos características funcionalmente análogas para permitir una comparación más justa y detallada entre ellos (§C.11).

Una vez establecida una métrica más fiable, el capítulo vuelve a la pregunta planteada en §C.10: ¿hasta qué punto son interpretables los modelos de visión que se despliegan cada vez más en escenarios de alto riesgo? Aplicamos estos protocolos a seis transformadores de visión pertenecientes a la familia arquitectónica descrita en §C.11—dos ViT supervisados y cuatro modelos fundacionales (DINOv2, DINOv3, CLIP y SigLIP). Observamos que los modelos fundacionales son sistemáticamente menos interpretables que los modelos supervisados de referencia, y que esta diferencia no se explica ni por la escala del modelo ni por su rendimiento en tareas posteriores, reforzando así la preocupación planteada en §C.12. Los dos protocolos producen además rankings fuertemente correlacionados, lo que indica que miden una dimensión compartida de la interpretabilidad. Los mejores predictores de la interpretabilidad son la localidad de las características y la alineación semántica de grano grueso con los seres humanos, precisamente el tipo de propiedades a nivel de representación descritas en §C.12 y los mecanismos que el Capítulo 5 estudia causalmente.

- "‘ **Julien Colin**, Lore Goetschalckx, Nuria Oliver, Thomas Serre. *Capability ≠ Interpretability: Human Interpretability of Vision Foundation Models*. arXiv, 2026.
- **Julien Colin**, Nuria Oliver, Thomas Serre. *Can Human Perceptual Similarity Alignment Improve Interpretability in Visual Representations?*. Conference on Cognitive Computational Neuroscience (CCN) (2026) "‘

¿Podemos mejorar deliberadamente la interpretabilidad? Los Capítulos 3 y 4 establecen dónde y cómo debe medirse la interpretabilidad; el Capítulo 5 cierra el ciclo retomando la brecha creciente introducida en §C.10 y preguntando si esta puede reducirse, en lugar de limitarse a medirla. Una vía prometedora, descrita en §C.12, consiste en intervenir directamente sobre modelos preentrenados alineando sus representaciones con juicios perceptivos o semánticos humanos. Sin embargo, la evidencia existente a favor de este enfoque es fundamentalmente correlacional y no se ha evaluado mediante los juicios humanos que §C.10 sostiene que deberían constituir el criterio último de la interpretabilidad. Este capítulo realiza directamente dicha evaluación.

Ajustamos mediante LoRA los cinco backbones visuales del Capítulo 4 para alinearlos con juicios perceptivos humanos y evaluamos posteriormente los modelos resultantes mediante la localizabilidad. Observamos que la alineación no mejora la interpretabilidad de forma consistente. Su efecto es positivo en los tres backbones exclusivamente visuales, pero solo alcanza significación estadística para DINOv2, mientras que se invierte en los dos modelos de visión y lenguaje. Este contraste, posibilitado por la convergencia en torno a arquitecturas compartidas descrita en §C.11, proporciona un marco en el que pueden estudiarse sistemáticamente las condiciones bajo las cuales la alineación funciona o fracasa.

Para comprender cómo la alineación modifica las características individuales, emparejamos además las características de cada backbone con las de su correspondiente versión alineada utilizando los métodos de diccionarios compartidos introducidos en §C.11 [76, 102, 119, 186]. Esto permite realizar un análisis más granular de los mecanismos que subyacen a los cambios en interpretabilidad. Incluso cuando la alineación resulta beneficiosa, sus efectos no son uniformes: la mejora observada en DINOv2 se distribuye entre todas sus características, mientras que las mejoras de DINOv3 se concentran en aquellas características que son modificadas en mayor medida por la alineación. En los backbones exclusivamente visuales, el aumento de la localidad de los conceptos explica parcialmente esta diferencia, aunque no explica por qué la alineación mejora la interpretabilidad en un modelo y no en el otro. Cuando la alineación perjudica la interpretabilidad, el mecanismo resulta todavía menos claro: SigLIP presenta la mayor disminución entre todos los backbones, pero ninguno de los mecanismos candidatos que examinamos permite explicar este efecto.

En conjunto, estos resultados muestran que la alineación perceptiva constituye una vía plausible, pero poco fiable, hacia modelos más interpretables. La intervención puede mejorar la interpretabilidad, pero no lo hace de manera consistente, y siguen sin resolverse los mecanismos que determinan cuándo funciona, cuándo fracasa o incluso cuándo produce el efecto contrario.

- “‘ **Julien Colin**, Nuria Oliver, Thomas Serre. *Does Human-Alignment Benefit Interpretability?*. ICLR Workshop on Representational Alignment (Re^4-Align) (2026)
- **Julien Colin**, Thomas Serre, Nuria Oliver. *Testing Human-Perceptual Alignment as a Lever for Interpretability in Vision Models*. (2026) “‘

Conclusión y perspectivas El Capítulo 6 reúne las tres contribuciones de la tesis en un marco unificado para identificar, medir y evaluar causalmente principios de interpretabilidad de los modelos. A continuación, retoma las preguntas abiertas planteadas en el Capítulo 1

y utiliza los resultados desarrollados a lo largo de la tesis para trazar los próximos pasos hacia modelos cuyas representaciones—y, en última instancia, cuyas decisiones—puedan ser comprendidas de manera significativa por los seres humanos.

Bibliography

[1] Ashraf Abdul et al. “COGAM: measuring and moderating cognitive load in machine learning model explanations”. In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–14 (cit. on p. 47).

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, et al. “Sanity Checks for Saliency Maps”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2018 (cit. on pp. 3, 12, 104).

[3] Julius Adebayo et al. “Post hoc explanations may be ineffective for detecting unknown spurious correlation”. In: *International conference on learning representations*. 2022 (cit. on pp. 2, 103).

[4] Jannis Ahlert, Thomas Klein, Felix A. Wichmann, et al. “How aligned are different alignment metrics?” In: *ICLR 2024 Workshop on Representational Alignment (Re-Align)*. 2024 (cit. on pp. 39, 59).

[5] Yasmeen Alufaisan et al. “Does explainable artificial intelligence improve human decision-making?” In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 8. 2021, pp. 6618–6626 (cit. on pp. 3, 103).

[6] David Alvarez-Melis and Tommi S. Jaakkola. “Towards Robust Interpretability with Self-Explaining Neural Networks”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2018 (cit. on pp. 2, 46, 102).

[7] Ananthu Aniraj et al. “Metonymy in vision models undermines attention-based interpretability”. In: *arXiv preprint arXiv:2605.06095* (2026) (cit. on pp. 2, 102).

[8] Mahmoud Assran, Quentin Duval, Ishan Misra, et al. “Self-supervised Learning from Images with a Joint-embedding Predictive Architecture”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 5563–5574 (cit. on pp. 4, 105).

[9] Mido Assran et al. “V-jepa 2: Self-supervised video models enable understanding, prediction and planning”. In: *arXiv preprint arXiv:2506.09985* (2025) (cit. on pp. 4, 105).

[10] Sebastian Bach et al. “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation”. In: *Public Library of Science (PloS One)* (2015) (cit. on p. 13).

[11] Mohammed Baharoon, Waseem Qureshi, Jiahong Ouyang, et al. “Evaluating General Purpose Vision Foundation Models for Medical Image Analysis: An Experimental Study of DINOv2 On Radiology Benchmarks”. In: *arXiv preprint arXiv:2312.02366* (2023) (cit. on pp. 1, 31, 101).

[12] Adrien Bardes et al. “Revisiting feature prediction for learning visual representations from video”. In: *arXiv preprint arXiv:2404.08471* (2024) (cit. on pp. 4, 105).

[13] David Bau, Bolei Zhou, Aditya Khosla, et al. “Network Dissection: Quantifying Interpretability of Deep Visual Representations”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6541–6549 (cit. on pp. 4, 5, 8, 14, 17, 33, 105, 106, 108).

[14] Steven Bills, Nick Cammarata, Dan Mossing, et al. “Language Models Can Explain Neurons in Language Models”. In: *OpenAI Blog* (2023) (cit. on pp. 14, 69).

[15] Matyas Bohacek, Thomas Fel, Maneesh Agrawala, et al. “Uncovering Conceptual Blindspots in Generative Image Models Using Sparse Autoencoders”. In: *arXiv preprint arXiv:2506.19708* (2025) (cit. on p. 15).

[16] Moritz Böhle, Mario Fritz, and Bernt Schiele. “B-cos Networks: Alignment Is All We Need for Interpretability”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:2205.10268. 2022 (cit. on pp. 2, 46, 102).

[17] Judy Borowski, Roland S. Zimmermann, Judith Schepers, et al. “Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2021 (cit. on pp. 2, 3, 5, 7, 13, 14, 18–20, 22–24, 30, 33, 34, 103, 106, 108).

[18] Trenton Bricken, Adly Templeton, Joshua Batson, et al. “Towards Monosemanticity: Decomposing Language Models with Dictionary Learning”. In: *Transformer Circuits Thread* (2023) (cit. on pp. 5, 7, 17, 19, 20, 106, 108).

[19] Zana Buçinca et al. “Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems”. In: *Proceedings of the 25th international conference on intelligent user interfaces*. 2020, pp. 454–464 (cit. on pp. 3, 103).

[20] Bart Bussmann, Patrick Leask, and Neel Nanda. “BatchTopK Sparse Autoencoders”. In: *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning* (cit. on pp. 14, 47, 49).

[21] Bart Bussmann et al. “Learning Multi-Level Features with Matryoshka Sparse Autoencoders”. In: *International Conference on Machine Learning*. PMLR. 2025, pp. 6077–6101 (cit. on pp. 17, 19, 47).

[22] Terry Caelli and Paul Bevan. “Visual sensitivity to two-dimensional spatial phase”. In: *JOSA* 72.10 (1982), pp. 1375–1381 (cit. on p. 13).

[23] Nick Cammarata, Shan Carter, Gabriel Goh, et al. “Thread: Circuits”. In: *Distill* (2020) (cit. on p. 19).

[24] Nick Cammarata, Gabriel Goh, Shan Carter, et al. “Curve Detectors”. In: *Distill* (2020) (cit. on p. 17).

[25] Mathilde Caron et al. “Emerging properties in self-supervised vision transformers”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660 (cit. on pp. 4, 105).

[26] Shan Carter et al. “Activation Atlas”. In: *Distill* (2019). https://distill.pub/2019/activation-atlas. DOI: 10.23915/distill.00015 (cit. on pp. 5, 105).

[27] Stephen Casper et al. “Red teaming deep neural networks with feature synthesis tools”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 80470–80516 (cit. on pp. 2, 5, 103, 106).

[28] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, et al. “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2018 (cit. on p. 12).

[29] Chaofan Chen, Oscar Li, Chaofan Tao, et al. “This Looks Like that: Deep Learning for Interpretable Image Recognition”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2019 (cit. on pp. 2, 46, 102).

[30] Christopher Chiu, Maximilian Heil, Teresa Kim, et al. “Fine-grained classification for poisonous fungi identification with transfer learning”. In: *arXiv preprint arXiv:2407.07492* (2024) (cit. on p. 31).

[31] Julien Colin, Thomas Fel, Rémi Cadène, et al. “What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods”. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022), pp. 2832–2845 (cit. on pp. 2, 3, 13, 19, 22, 103).

[32] Julien Colin, Lore Goetschalckx, Thomas Fel, et al. “Choosing the Right Basis for Interpretability: Psychophysical Comparison Between Neuron-based and Dictionary-based Representations”. In: *arXiv preprint arXiv:2411.03993* (2024) (cit. on pp. 4, 8, 14, 33, 105, 109).

[33] Julien Colin, Lore Goetschalckx, Nuria Oliver, et al. “Capability ≠ Interpretability: Human Interpretability of Vision Foundation Models”. In: *arXiv preprint arXiv:2605.20337* (2026) (cit. on pp. 45–47).

[34] Valérie Costa et al. “From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit”. In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2025 (cit. on pp. 17, 19).

[35] Hoagy Cunningham, Aidan Ewart, Logan Riggs, et al. “Sparse Autoencoders Find Highly Interpretable Features in Language Models”. In: *International Conference on Learning Representations (ICLR)*. arXiv:2309.08600. 2024 (cit. on pp. 19, 31, 33).

[36] Alexander D’Amour, Katherine Heller, Dan Moldovan, et al. “Underspecification Presents Challenges for Credibility in Modern Machine Learning”. In: *Journal of Machine Learning Research* 23.226 (2022), pp. 1–61 (cit. on pp. 1, 101).

[37] Tugba Akinci D’Antonoli, Christian Bluethgen, Renato Cuocolo, et al. “Foundation Models for Radiology: Fundamentals, Applications, Opportunities, Challenges, Risks, and Prospects”. In: *Diagnostic and Interventional Radiology* (2025) (cit. on pp. 1, 31, 101).

[38] Timothée Darcet, Maxime Oquab, Julien Mairal, et al. “Vision Transformers Need Registers”. In: *International Conference on Learning Representations (ICLR)*. arXiv:2309.16588. 2024 (cit. on pp. 5, 105).

[39] J. Deng, W. Dong, R. Socher, et al. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009 (cit. on pp. 35, 39).

[40] Jia Deng, Wei Dong, Richard Socher, et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on pp. 19, 20).

[41] Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. “How Important Is a Neuron?” In: *International Conference on Learning Representations (ICLR)*. arXiv:1805.12233. 2019 (cit. on p. 11).

[42] Grégoire Dhimoïla, Thomas Fel, Victor Boutin, et al. “Cross-Modal Redundancy and the Geometry of Vision–Language Embeddings”. In: *The Fourteenth International Conference on Learning Representations (ICLR)*. arXiv:2602.06218. 2026 (cit. on p. 15).

[43] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, et al. “Explanations Can Be Manipulated and Geometry Is to Blame”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019 (cit. on p. 13).

[44] Teresa Dorszewski, Lenka Tětková, Robert Jenssen, et al. “From Colors to Classes: Emergence of Concepts in Vision Transformers”. In: *World Conference on Explainable Artificial Intelligence (xAI)*. Springer. 2025, pp. 28–47 (cit. on pp. 5, 106).

[45] Finale Doshi-Velez and Been Kim. “Towards a Rigorous Science of Interpretable Machine Learning”. In: *arXiv preprint arXiv:1702.08608* (2017) (cit. on pp. 2, 3, 102, 103).

[46] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. “An Image Is Worth 16X16 Words: Transformers for Image Recognition at Scale”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2021 (cit. on pp. 4, 31, 105).

[47] Amil Dravid, Yossi Gandelsman, Alexei A. Efros, et al. “Rosetta Neurons: Mining the Common Units in a Model Zoo”. In: *International Conference on Computer Vision (ICCV)*. arXiv:2306.09346. 2023 (cit. on pp. 5, 106).

[48] Maximilian Dreyer, Erblina Purelku, Johanna Vielhaben, et al. “Pure: Turning polysemantic neurons into pure features by identifying relevant circuits”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 8212–8217 (cit. on pp. 38, 41).

[49] R. Becket Ebitz and Benjamin Y. Hayden. “The population doctrine in cognitive neuroscience”. In: *Neuron* 109.19 (2021), pp. 3055–3068 (cit. on p. 17).

[50] Nelson Elhage, Tristan Hume, Catherine Olsson, et al. “Toy Models of Superposition”. In: *Transformer Circuits Thread* (2022) (cit. on pp. 5, 17, 19, 33, 35, 106).

[51] Dumitru Erhan, Yoshua Bengio, Aaron Courville, et al. “Visualizing higher-layer features of a deep network”. In: *University of Montreal* 1341.3 (2009), p. 1 (cit. on pp. 17, 45, 47).

[52] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. Adopted 13 June 2024; entered into force 1 August 2024. 2024 (cit. on pp. 1, 101).

[53] European Union Aviation Safety Agency. *EASA Artificial Intelligence Concept Paper Issue 2: Guidance for Level 1 & 2 Machine Learning Applications*. Tech. rep. European Union Aviation Safety Agency (EASA), Mar. 2024 (cit. on pp. 1, 101).

[54] Niels G Faber, Seyed Sahand Mohammadi Ziabari, and Fatemeh Karimi Nejadasl. "Leveraging foundation models via knowledge distillation in multi-object tracking: Distilling dinov2 features to fairmot". In: *arXiv preprint arXiv:2407.18288* (2024) (cit. on p. 31).

[55] Alhussein Fawzi, Matej Balog, Aja Huang, et al. "Discovering Faster Matrix Multiplication Algorithms with Reinforcement Learning". In: *Nature* 610.7930 (2022), pp. 47–53 (cit. on pp. 1, 101).

[56] Thomas Fel, Thibaut Boissin, Victor Boutin, et al. "Unlocking Feature Visualization for Deeper Networks with Magnitude Constrained Optimization". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 36. 2023, pp. 37813–37826 (cit. on pp. 13, 35, 47, 48, 74).

[57] Thomas Fel, Victor Boutin, Louis Béthune, et al. "A holistic approach to unifying automatic concept extraction and concept importance estimation". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 54805–54818 (cit. on pp. 8, 12, 15, 17, 19, 21, 35, 108).

[58] Thomas Fel, Remi Cadene, Mathieu Chalvidal, et al. "Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021 (cit. on pp. 12, 13, 47).

[59] Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, et al. "Archetypal SAE: Adaptive and Stable Dictionary Learning for Concept Extraction in Large Vision Models". In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 267. Proceedings of Machine Learning Research. PMLR, 2025, pp. 16543–16572 (cit. on pp. 5, 14, 17, 19, 21, 33, 47, 106).

[60] Thomas Fel, Agustin Picard, Louis Bethune, et al. "CRAFT: Concept Recursive Activation FacTorization For Explainability". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023 (cit. on pp. 2, 5, 7, 14, 15, 17, 19, 20, 23, 24, 47, 63, 103, 106, 108).

[61] Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, et al. "Harmonizing the object recognition strategies of deep neural networks with humans". In: *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), pp. 9432–9446 (cit. on pp. 6, 39, 45, 47, 69, 107).

[62] Thomas Fel, Binxu Wang, Michael A Lepori, et al. "Into the Rabbit Hull: From Task-Relevant Concepts in DINO To Minkowski Geometry". In: *The Fourteenth International Conference on Learning Representations*. 2026 (cit. on pp. 5, 15, 106).

[63] Christiane Fellbaum. "WordNet". In: *Theory and applications of ontology: computer applications*. Springer, 2010, pp. 231–243 (cit. on p. 24).

[64] Daniel J Felleman and David C Van Essen. "Distributed Hierarchical Processing in the Primate Cerebral Cortex". In: *Cerebral Cortex* 1.1 (1991), pp. 1–47 (cit. on pp. 4, 104).

[65] Ruth Fong and Andrea Vedaldi. “Net2Vec: Quantifying and Explaining How Concepts Are Encoded by Filters in Deep Neural Networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8730–8738 (cit. on pp. 4, 8, 14, 105, 108).

[66] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, et al. “DreamSim: Learning new dimensions of human visual similarity using synthetic data”. In: *arXiv preprint arXiv:2306.09344* (2023) (cit. on p. 40).

[67] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. “Interpreting CLIP’S Image Representation via Text-Based Decomposition”. In: *International Conference on Learning Representations (ICLR)*. arXiv:2310.05916. 2024 (cit. on pp. 5, 105).

[68] Haoxiang Gao, Zhongruo Wang, Yaqian Li, et al. “A Survey for Foundation Models in Autonomous Driving”. In: *2025 6th International Conference on Computer Vision and Data Mining (ICCVDM)*. IEEE. 2025, pp. 63–71 (cit. on pp. 1, 31, 101).

[69] Leo Gao, Tom Dupré la Tour, Henk Tillman, et al. “Scaling and Evaluating Sparse Autoencoders”. In: *arXiv preprint arXiv:2406.04093* (2024) (cit. on pp. 14, 15, 17, 19, 31, 33, 35, 47).

[70] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, et al. “Shortcut Learning in Deep Neural Networks”. In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673 (cit. on pp. 1, 101).

[71] Amirata Ghorbani, Abubakar Abid, and James Zou. “Interpretation of Neural Networks Is Fragile”. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2017 (cit. on pp. 3, 13, 104).

[72] Amirata Ghorbani, James Wexler, James Y Zou, et al. “Towards Automatic Concept-based Explanations”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 9273–9282 (cit. on pp. 14, 17, 19, 47).

[73] Leilani H. Gilpin, David Bau, Ben Z Yuan, et al. “Explaining Explanations: An Overview of Interpretability of Machine Learning”. In: *Proceedings of the IEEE International Conference on data science and advanced analytics (DSAA)*. 2018 (cit. on pp. 2, 102).

[74] Evgeny Gladilin and Roland Eils. “On the role of spatial phase and phase correlation in vision, illusion, and cognition”. In: *Frontiers in Computational Neuroscience* 9 (2015), p. 45 (cit. on p. 13).

[75] Gabriel Goh, Nick Cammarata, Chelsea Voss, et al. “Multimodal Neurons in Artificial Neural Networks”. In: *Distill* (2021) (cit. on pp. 5, 105).

[76] Liv Gorton. “Group crosscoders for mechanistic analysis of symmetry”. In: *arXiv preprint arXiv:2410.24184* (2024) (cit. on pp. 4, 9, 105, 110).

[77] Fabian Gröger, Shuo Wen, and Maria Brbić. “Revisiting the Platonic Representation Hypothesis: An Aristotelian View”. In: *arXiv preprint arXiv:2602.14486* (2026) (cit. on pp. 42, 60).

[78] Charles G. Gross, Carlos E. Rocha-Miranda, and David B. Bender. “Visual Properties of Neurons in Inferotemporal Cortex of the Macaque”. In: *Journal of Neurophysiology* 35.1 (1972), pp. 96–111 (cit. on pp. 4, 104).

[79] Nathalie Guyader et al. “Image phase or amplitude? Rapid scene categorization is an amplitude-based process”. In: *Comptes Rendus Biologies* 327.4 (2004), pp. 313–318 (cit. on p. 13).

[80] Chris Hamblin et al. “Feature accentuation: Revealing’what’features respond to in natural images”. In: *arXiv preprint arXiv:2402.10039* (2024) (cit. on pp. 2, 13, 47, 103).

[81] James V. Haxby, M. Ida Gobbini, Maura L. Furey, et al. “Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex”. In: *Science* 293.5539 (2001), pp. 2425–2430 (cit. on p. 17).

[82] Kaiming He, Xinlei Chen, Saining Xie, et al. “Masked Autoencoders Are Scalable Vision Learners”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009 (cit. on pp. 4, 105).

[83] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 (cit. on pp. 4, 8, 20, 104, 108).

[84] Martin N. Hebart, Oliver Contier, Lina Teichmann, et al. “THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior”. In: *eLife* 12 (2023), e82580 (cit. on pp. 40, 46, 48).

[85] Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, et al. “Revealing the multidimensional mental representations of natural objects underlying human similarity judgements”. In: *Nature Human Behaviour* 4.11 (2020), pp. 1173–1185 (cit. on pp. 40, 42).

[86] Evan Hernandez, Sarah Schwettmann, David Bau, et al. “Natural Language Descriptions of Deep Visual Features”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2022 (cit. on pp. 14, 69).

[87] John Hewitt, Robert Geirhos, and Been Kim. “We Can’t Understand AI Using our Existing Vocabulary”. In: *arXiv preprint arXiv:2502.07586* (2025) (cit. on pp. 1, 6, 47, 101, 107).

[88] Sai Sumedh R. Hindupur, Ekdeep S Lubana, Thomas Fel, et al. “Projecting Assumptions: The Duality Between Sparse Autoencoders and Concept Geometry”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Belgrave, C. Zhang, H. Lin, et al. Vol. 38. Curran Associates, Inc., 2025, pp. 11649–11699 (cit. on p. 15).

[89] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, et al. “A Benchmark for Interpretability Methods in Deep Neural Networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019 (cit. on p. 13).

[90] Patrik O. Hoyer. “Non-negative matrix factorization with sparseness constraints”. In: *The Journal of Machine Learning Research (JMLR)* 5.Nov (2004), pp. 1457–1469 (cit. on pp. 21, 29, 39).

[91] Edward J Hu et al. “Lora: Low-rank adaptation of large language models.” In: *Iclr* 1.2 (2022), p. 3 (cit. on pp. 48, 89).

[92] David H. Hubel and Torsten N. Wiesel. “Receptive Fields of Single Neurones in the Cat’s Striate Cortex”. In: *The Journal of physiology* 148.3 (1959), p. 574 (cit. on pp. 4, 17, 104).

[93] Minyoung Huh, Brian Cheung, Tongzhou Wang, et al. "The Platonic Representation Hypothesis". In: *International Conference on Machine Learning (ICML)*. arXiv:2405.07987. 2024 (cit. on p. 60).

[94] Prince Jacob et al. "Parametric control along the encoding axes of IT neurons uncovers hidden differences in model-brain alignment". In: *Conference on Cognitive Computational Neuroscience (CCN)* (2025) (cit. on p. 59).

[95] Alon Jacovi and Yoav Goldberg. "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* (2020) (cit. on pp. 2, 13, 102).

[96] Piotr Jedryszek and Oliver M Crook. "Stable and steerable sparse autoencoders with weight regularization". In: *arXiv preprint arXiv:2603.04198* (2026) (cit. on pp. 49, 93).

[97] Sicong Jiang, Zilin Huang, Kangan Qian, et al. "A Survey on Vision-language-action Models for Autonomous Driving". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 4524–4536 (cit. on pp. 1, 31, 101).

[98] Licheng Jiao, Jiayao Hao, Ruiyang Li, et al. "Foundation Models Meet Medical Image Interpretation". In: *Research* 9 (2026), p. 1024 (cit. on pp. 1, 31, 101).

[99] Olivier R Joubert et al. "Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise". In: *Journal of Vision* 9.1 (2009), pp. 2–2 (cit. on p. 13).

[100] John Jumper, Richard Evans, Alexander Pritzel, et al. "Highly Accurate Protein Structure Prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589 (cit. on pp. 1, 101).

[101] Margot E Kaminski. "The Right to Explanation, Explained". In: *Research Handbook on Information Law and Governance*. Edward Elgar Publishing, 2021 (cit. on pp. 1, 101).

[102] Aly Kassem et al. "Delta-Crosscoder: Robust Crosscoder Model Diffing in Narrow Fine-Tuning Regimes". In: *arXiv preprint arXiv:2603.04426* (2026) (cit. on pp. 4, 9, 105, 110).

[103] Rémi Kazmierczak, Eloïse Berthier, Goran Frehse, et al. "Explainability and Vision Foundation Models: A Survey". In: *Information Fusion* 122 (2025), p. 103184 (cit. on pp. 1, 45, 101).

[104] Been Kim, Martin Wattenberg, Justin Gilmer, et al. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)". In: *International conference on machine learning*. Proceedings of the International Conference on Machine Learning (ICML). 2018 (cit. on pp. 14, 17, 19, 47).

[105] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, et al. "OpenVLA: An Open-Source Vision-Language-Action Model". In: *Conference on Robot Learning*. PMLR. 2025, pp. 2679–2713 (cit. on pp. 1, 31, 101).

[106] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, et al. "HIVE: Evaluating the Human Interpretability of Visual Explanations". In: *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*. 2022 (cit. on pp. 2, 3, 5, 13, 103, 106).

[107] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, et al. "The (UN)Reliability of Saliency Methods". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Vol. 11700. Lecture Notes in Computer Science. Springer, 2019, pp. 267–280 (cit. on pp. 3, 12, 104).

[108] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, et al. "Concept Bottleneck Models". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5338–5348 (cit. on pp. 2, 46, 102).

[109] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. "Representational similarity analysis-connecting the branches of systems neuroscience". In: *Frontiers in systems neuroscience* 2 (2008), p. 249 (cit. on p. 48).

[110] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105 (cit. on pp. 1, 101).

[111] Isaac Lage, Emily Chen, Jeffrey He, et al. "An Evaluation of the Human-Interpretability of Explanation". In: *Workshop on Correcting and Critiquing Trends in Machine Learning, Advances in Neural Information Processing Systems (NIPS)*. 2019 (cit. on pp. 6, 47, 107).

[112] Victor A F Lamme and Pieter R Roelfsema. "The Distinct Modes of Vision Offered by Feedforward and Recurrent Processing". In: *Trends in Neurosciences* 23.11 (2000), pp. 571–579 (cit. on pp. 4, 104).

[113] Patrick Leask et al. "Sparse Autoencoders Do Not Find Canonical Units of Analysis". In: *The Thirteenth International Conference on Learning Representations*. 2025 (cit. on pp. 5, 106).

[114] Patrick Leask et al. "Sparse autoencoders do not find canonical units of analysis". In: *International Conference on Learning Representations*. Vol. 2025. 2025, pp. 53617–53642 (cit. on pp. 49, 93).

[115] Klas Leino, Shayak Sen, Anupam Datta, et al. "Influence-Directed Explanations for Deep Convolutional Networks". In: *IEEE International Test Conference (ITC)*. arXiv:1802.03788. 2018 (cit. on p. 11).

[116] Ming Li, Xin Chen, Xin Li, et al. "The similarity metric". In: *IEEE transactions on Information Theory* 50.12 (2004), pp. 3250–3264 (cit. on p. 87).

[117] Yihao Li, Saeed Salehi, Lyle Ungar, et al. "Does object binding naturally emerge in large pretrained vision transformers?" In: *Advances in Neural Information Processing Systems* 38 (2026), pp. 3394–3423 (cit. on p. 38).

[118] Justin D Lieber et al. "Sensitivity to naturalistic texture relies primarily on high spatial frequencies". In: *Journal of vision* 23.2 (2023), pp. 4–4 (cit. on p. 50).

[119] Jack Lindsey, Adly Templeton, Jonathan Marcus, et al. *Sparse Crosscoders for Cross-Layer Features and Model Diffing*. Transformer Circuits Thread. 2024 (cit. on pp. 4, 9, 60, 105, 110).

[120] Drew Linsley, Dan Shiebler, Sven Eberhardt, et al. "Learning What and Where to Attend". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2019 (cit. on pp. 39, 59).

[121] Zachary C Lipton. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57 (cit. on pp. 1, 101).

[122] Zachary C. Lipton. "The Mythos of Model Interpretability". In: *Workshop on Human Interpretability in Machine Learning, Proceedings of the International Conference on Machine Learning (ICML)*. 2016 (cit. on pp. 3, 103).

[123] Emily G Liquin and Tania Lombrozo. "Motivated to learn: An account of explanatory satisfaction". In: *Cognitive Psychology* 132 (2022), p. 101453 (cit. on pp. 3, 103).

[124] Tania Lombrozo. "Explanatory preferences shape learning and inference". In: *Trends in Cognitive Sciences* 20.10 (2016), pp. 748–759 (cit. on pp. 3, 103).

[125] Tania Lombrozo. "The instrumental value of explanations". In: *Philosophy Compass* 6.8 (2011), pp. 539–551 (cit. on pp. 3, 103).

[126] Tania Lombrozo and Susan Carey. "Functional explanation and the function of explanation". In: *Cognition* 99.2 (2006), pp. 167–204 (cit. on pp. 3, 103).

[127] Tania Lombrozo and Nicholas Z Gwynne. "Explanation and inference: Mechanistic and functional explanations guide property generalization". In: *Frontiers in Human Neuroscience* 8 (2014), p. 700 (cit. on pp. 3, 103).

[128] Scott Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems (NIPS)*. 2017 (cit. on p. 13).

[129] Aravindh Mahendran and Andrea Vedaldi. "Understanding Deep Image Representations by Inverting Them". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015) (cit. on pp. 13, 47).

[130] Julien Mairal, Francis Bach, Jean Ponce, et al. "Sparse Modeling for Image and Vision Processing". In: *Foundations and Trends® in Computer Graphics and Vision* 8.2-3 (2014), pp. 85–283 (cit. on p. 15).

[131] Sébastien Marcel and Yann Rodriguez. "Torchvision the machine-vision package of torch". In: *Proceedings of the 18th ACM international conference on Multimedia*. 2010, pp. 1485–1488 (cit. on p. 20).

[132] Georgii Mikriukov, Gesina Schwalbe, Christian Hellert, et al. "Revealing Similar Semantics Inside CNNS: An Interpretable Concept-Based Comparison of Feature Spaces". In: *Machine Learning and Knowledge Discovery in Databases (ECML PKDD) Workshops*. Springer. 2023, pp. 3–20 (cit. on pp. 5, 106).

[133] Tim Miller. "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: *Artificial Intelligence* 267 (Feb. 2019), pp. 1–38 (cit. on pp. 3, 103).

[134] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38 (cit. on pp. 2, 102).

[135] Ari S Morcos et al. "On the importance of single directions for generalization". In: *International Conference on Learning Representations*. 2018 (cit. on p. 17).

[136] Théo Moutakanni, Piotr Bojanowski, Guillaume Chassagnon, et al. "Advancing human-centric AI for robust X-ray analysis through holistic self-supervised learning". In: *arXiv preprint arXiv:2405.01469* (2024) (cit. on p. 31).

[137] Ahmad Mustapha, Wael Khreich, and Wes Masri. "Inter-model Interpretability: Self-supervised Models as a Case Study". In: *Array* 22 (2024), p. 100350 (cit. on pp. 5, 106).

[138] Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, et al. "Human Alignment of Neural Network Representations". In: *International Conference on Learning Representations (ICLR)*. arXiv:2211.01201. 2023 (cit. on p. 47).

[139] Lukas Muttenthaler, Klaus Greff, Frieda Born, et al. "Aligning machine and human visual representations across abstraction levels". In: *Nature* 647.8089 (2025), pp. 349–355 (cit. on pp. 6, 39, 40, 42, 45, 47, 69, 88, 107).

[140] Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, et al. "Improving neural network representations using human similarity judgments". In: *Advances in Neural Information Processing Systems (NeurIPS)* 36 (2023), pp. 50978–51007 (cit. on pp. 6, 39, 47, 48, 69, 89, 107).

[141] Anh Nguyen, Jeff Clune, Yoshua Bengio, et al. "Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4467–4477 (cit. on pp. 13, 47).

[142] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, et al. "Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks". In: *Advances in neural information processing systems* 29 (2016) (cit. on p. 13).

[143] Anh Nguyen, Jason Yosinski, and Jeff Clune. "Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned by Each Neuron in Deep Neural Networks". In: *arXiv preprint arXiv:1602.03616* (2016) (cit. on pp. 13, 19, 47).

[144] Anh Nguyen, Jason Yosinski, and Jeff Clune. "Understanding Neural Networks via Feature Visualization: A survey". In: *arXiv preprint arXiv:1904.08939* (2019) (cit. on pp. 13, 47).

[145] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. "The effectiveness of feature attribution methods and its correlation with automatic evaluation scores". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021) (cit. on pp. 2, 3, 103).

[146] Paul Novello, Thomas Fel, and David Vigouroux. "Making Sense of Dependence: Efficient Black-box Explanations Using Dependence Measure". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022 (cit. on pp. 36, 81).

[147] Tuomas Oikarinen and Tsui-Wei Weng. "CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2023 (cit. on pp. 5, 14, 69, 106).

[148] Chris Olah, Nick Cammarata, Ludwig Schubert, et al. "An Overview of Early Vision in InceptionV1". In: *Distill* (2020) (cit. on pp. 5, 105).

[149] Chris Olah, Nick Cammarata, Ludwig Schubert, et al. “Zoom in: An Introduction to Circuits”. In: *Distill* (2020) (cit. on pp. 5, 105, 106).

[150] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. “Feature Visualization”. In: *Distill* (2017) (cit. on pp. 13, 14, 17, 19, 47).

[151] Alan V Oppenheim and Jae S Lim. “The importance of phase in signals”. In: *Proceedings of the IEEE* 69.5 (1981), pp. 529–541 (cit. on p. 13).

[152] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. “DINOV2: Learning Robust Visual Features without Supervision”. In: *Transactions on Machine Learning Research* (2024). arXiv:2304.07193 (cit. on pp. 4, 31, 32, 105).

[153] Gonçalo Paulo and Nora Belrose. “Sparse autoencoders trained on the same data learn different features”. In: *arXiv preprint arXiv:2501.16615* (2025) (cit. on pp. 49, 93).

[154] Vitali Petsiuk, Abir Das, and Kate Saenko. “RISE: Randomized Input Sampling for Explanation of Black-box Models”. In: *Proceedings of the British Machine Vision Conference (BMVC)*. 2018 (cit. on pp. 12, 35, 47, 48, 74).

[155] R. Quian Quiroga, Gabriel Kreiman, Christof Koch, et al. “Sparse but Not ‘Grandmother-cell’Coding in the Medial Temporal Lobe”. In: *Trends in cognitive sciences* 12.3 (2008), pp. 87–91 (cit. on p. 17).

[156] R. Quian Quiroga, Leila Reddy, Gabriel Kreiman, et al. “Invariant Visual Representation by Single Neurons in the Human Brain”. In: *Nature* 435.7045 (2005), pp. 1102–1107 (cit. on pp. 4, 17, 104).

[157] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. “Learning Transferable Visual Models from Natural Language Supervision”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PmLR. 2021, pp. 8748–8763 (cit. on pp. 4, 31, 32, 36, 83, 105).

[158] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “ ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier ”. In: *Knowledge Discovery and Data Mining (KDD)*. 2016 (cit. on pp. 2, 3, 5, 12, 47, 103, 106).

[159] Brett D Roads and Bradley C Love. “Enriching imagenet with human similarity judgments and psychological embeddings”. In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2021, pp. 3547–3557 (cit. on p. 47).

[160] Alexandre Rocchi-Henry, Thomas Fel, and Gianni Franchi. “A Geometric Unification of Concept Learning with Concept Cones”. In: *arXiv preprint arXiv:2512.07355* (2025) (cit. on pp. 2, 102).

[161] Leonid Rozenblit and Frank Keil. “The misunderstood limits of folk science: An illusion of explanatory depth”. In: *Cognitive science* 26.5 (2002), pp. 521–562 (cit. on pp. 3, 103).

[162] Cynthia Rudin. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. In: *Nature Machine Intelligence* 1 (2019), pp. 206–215 (cit. on pp. 2, 46, 60, 102).

[163] Olga Russakovsky, Jia Deng, Hao Su, et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115 (2015), pp. 211–252 (cit. on p. 23).

[164] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al. “Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017 (cit. on p. 12).

[165] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017 (cit. on pp. 12, 17, 19, 47).

[166] Julian Senoner et al. “Explainable AI improves task performance in human–AI collaboration”. In: *Scientific reports* 14.1 (2024), p. 31150 (cit. on pp. 3, 103).

[167] Hua Shen and Ting-Hao Huang. “How useful are the machine-generated interpretations to general users? A human evaluation on guessing the incorrectly predicted labels”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 8. 2020, pp. 168–172 (cit. on pp. 2, 103).

[168] Liu Shilong, Zeng Zhaoyang, Ren Tianhe, et al. “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection”. In: *arXiv preprint arXIv:2303.05499* (2024) (cit. on p. 31).

[169] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. “Learning Important Features Through Propagating Activation Differences”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2017 (cit. on pp. 12, 21, 35, 47).

[170] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, et al. “DINOv3”. In: *arXiv preprint arXiv:2508.10104* (2025) (cit. on pp. 4, 31, 32, 105).

[171] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: *Workshop, Proceedings of the International Conference on Learning Representations (ICLR)*. 2013 (cit. on pp. 11, 12, 47).

[172] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-scale Image Recognition”. In: *International Conference on Learning Representations (ICLR)*. arXiv:1409.1556. 2015 (cit. on pp. 12, 20, 23, 47).

[173] Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, et al. “Do Users Benefit from Interpretable Vision? A User Study, Baseline, and Dataset”. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2022) (cit. on pp. 2, 3, 103).

[174] Daniel Smilkov, Nikhil Thorat, Been Kim, et al. “SmoothGrad: removing noise by adding noise”. In: *Workshop on Visualization for Deep Learning, Proceedings of the International Conference on Machine Learning (ICML)*. 2017 (cit. on pp. 12, 47).

[175] Kaitao Song et al. “Mpnet: Masked and permuted pre-training for language understanding”. In: *Advances in neural information processing systems* 33 (2020), pp. 16857–16867 (cit. on p. 36).

[176] Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, et al. "Sparse Autoencoders for Scientifically Rigorous Interpretation of Vision Models". In: *arXiv preprint arXiv:2502.06755* (2025). arXiv:2502.06755 (cit. on pp. 5, 106).

[177] Ajay Subramanian et al. "Spatial-frequency channels, shape bias, and adversarial robustness". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023 (cit. on p. 50).

[178] Ilia Sucholutsky and Tom Griffiths. "Alignment with human representations supports robust few-shot learning". In: *Advances in Neural Information Processing Systems (NeurIPS)* 36 (2023), pp. 73464–73479 (cit. on pp. 6, 39, 47, 69, 107).

[179] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, et al. "Getting Aligned on Representational Alignment". In: *arXiv preprint arXiv:2310.13018* (2023) (cit. on pp. 47, 69).

[180] Xiaoqing Sun, Alessandro Stolfo, Joshua Engels, et al. "Dense SAE Latents Are Features, Not Bugs". In: *Mechanistic Interpretability Workshop at NeurIPS 2025*. 2025 (cit. on p. 35).

[181] Shobhita Sundaram, Stephanie Fu, Lukas Muttenthaler, et al. "When does perceptual alignment benefit vision representations?" In: *Advances in Neural Information Processing Systems (NeurIPS)* 37 (2024), pp. 55314–55341 (cit. on pp. 6, 39, 40, 45, 47, 69, 107).

[182] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2017 (cit. on pp. 12, 13, 19, 47).

[183] Christian Szegedy, Wei Liu, Yangqing Jia, et al. "Going Deeper with Convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9 (cit. on p. 19).

[184] Mohammad Reza Taesiri, Giang Nguyen, and Anh Nguyen. "Visual Correspondence-based Explanations Improve AI Robustness and Human-AI Team Accuracy". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022) (cit. on pp. 3, 5, 13, 103, 106).

[185] Adly Templeton, Tom Conerly, Jonathan Marcus, et al. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet". In: *Transformer Circuits Thread* (2024) (cit. on p. 33).

[186] Harrish Thasarathan, Julian Forsyth, Thomas Fel, et al. "Universal Sparse Autoencoders: Interpretable Cross-model Concept Alignment". In: *Forty-second International Conference on Machine Learning*. 2025 (cit. on pp. 4, 5, 9, 49, 60, 75, 105, 106, 110).

[187] Anja Thieme et al. "Interpretability as a dynamic of human-AI interaction". In: *Interactions* 27.5 (2020), pp. 40–45 (cit. on p. 60).

[188] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, et al. "Sanity Checks for Saliency Metrics". In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2019 (cit. on pp. 3, 104).

[189] Zhan Tong et al. "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training". In: *Advances in neural information processing systems* 35 (2022), pp. 10078–10093 (cit. on pp. 4, 105).

[190] Dimitris Tsipras et al. “Robustness may be at odds with accuracy”. In: *arXiv preprint arXiv:1805.12152* (2018) (cit. on pp. 6, 47, 107).

[191] Narek Tumanyan, Assaf Singer, Shai Bagon, et al. “Dino-tracker: Taming dino for self-supervised point tracking in a single video”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 367–385 (cit. on p. 31).

[192] Ny Vasil, Azzurra Ruggeri, and Tania Lombrozo. “When and how children use explanations to guide generalizations”. In: *Cognitive Development* 61 (2022), p. 101144 (cit. on pp. 3, 103).

[193] Limin Wang et al. “Videomae v2: Scaling video masked autoencoders with dual masking”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 14549–14560 (cit. on pp. 4, 105).

[194] Zifan Wang, Matt Fredrikson, and Anupam Datta. “Robust Models Are More Interpretable Because Attributions Look Normal”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 22625–22651 (cit. on pp. 6, 47, 107).

[195] Joseph J Williams and Tania Lombrozo. “The role of explanation in discovery and generalization: Evidence from category learning”. In: *Cognitive science* 34.5 (2010), pp. 776–806 (cit. on pp. 3, 103).

[196] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, et al. “On the (in)Fidelity and Sensitivity for Explanations”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019 (cit. on p. 13).

[197] Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. “Interpreting CLIP with Hierarchical Sparse Autoencoders”. In: *Forty-second International Conference on Machine Learning*. 2025 (cit. on pp. 17, 19).

[198] Matthew D. Zeiler and Rob Fergus. “Visualizing and understanding convolutional networks”. In: *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*. 2014 (cit. on pp. 13, 17, 19, 45).

[199] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, et al. “Sigmoid Loss for Language Image Pre-training”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2023, pp. 11975–11986 (cit. on pp. 4, 31, 32, 105).

[200] Ruihan Zhang, Prashan Madumal, Tim Miller, et al. “Invertible Concept-based Explanations for CNN Models with Non-negative Concept Activation Vectors”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 2021, pp. 11682–11690 (cit. on pp. 14, 47).

[201] Bolei Zhou, Aditya Khosla, Agata Lapedriza, et al. “Object Detectors Emerge in Deep Scene CNNs”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2016 (cit. on p. 17).

[202] Bolei Zhou, Hang Zhao, Xavier Puig, et al. “Scene parsing through ade20k dataset”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 633–641 (cit. on pp. 38, 39).

[203] Roland S. Zimmermann, Judy Borowski, Robert Geirhos, et al. “How Well Do Feature Visualizations Support Causal Understanding of CNN Activations?” In: *Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021), pp. 11730–11744 (cit. on pp. 2, 3, 5, 7, 14, 19, 103, 106, 108).

[204] Roland S. Zimmermann, Thomas Klein, and Wieland Brendel. "Scale Alone Does Not Improve Mechanistic Interpretability in Vision Models". In: *Advances in Neural Information Processing Systems (NeurIPS)* 36 (2023) (cit. on pp. 4, 5, 8, 14, 18, 20, 23, 25, 26, 29, 30, 32–35, 41, 45, 67, 71, 105, 106, 108, 109).

[205] Roland S. Zimmermann, David Klindt, and Wieland Brendel. "Measuring Per-Unit Interpretability at Scale without Humans". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by A. Globerson, L. Mackey, D. Belgrave, et al. Vol. 37. Curran Associates, Inc., 2024, pp. 48448–48483 (cit. on pp. 5, 18, 20, 28, 33, 106).