CONTEXT SENSITIVITY IMPROVES HUMAN-MACHINE VISUAL ALIGNMENT

Frieda Born[1,2,3,*], **Tom Neuhäuser**[1,2,*], **Lukas Muttenthaler**[1,4,5,6,†], **Brett D. Roads**[7,†], **Bernhard Spitzer**[3,8], **Andrew K. Lampinen**[9], **Matt Jones**[9,10], **Klaus-Robert Müller**[1,2,9,11,12], **Michael C. Mozer**[9]

[1]Technische Universität Berlin [2]BIFOLD [3]Max Planck Institute for Human Development
[4]Aignostics [5]Helmholtz Munich [6]Technical University of Munich
[7]University College London [8]Technische Universität Dresden [9]Google DeepMind
[10]University of Colorado Boulder [11]Korea University [12]Max Planck Institute for Informatics

ABSTRACT

Modern machine learning models typically represent inputs as fixed points in a high-dimensional embedding space. While this approach has been proven powerful for a wide range of downstream tasks, it fundamentally differs from the way humans process information. Because humans are constantly adapting to their environment, they represent objects and their relationships in a highly context-sensitive manner. To address this gap, we propose a method for context-sensitive similarity computation from neural network embeddings, applied to modeling a triplet odd-one-out task with an anchor image serving as simultaneous context. Modeling context enables us to achieve up to a 15% improvement in odd-one-out accuracy over a context-insensitive model. We find that this improvement is consistent across both original and "human-aligned" vision foundation models.

1 INTRODUCTION

Modern neural networks typically represent images and other entities as fixed points in a high-dimensional embedding space. Representing information this way has proven useful for downstream tasks that require generalization (Bengio et al., 2013; Dosovitskiy et al., 2021), e.g., benchmarks such as ImageNet, where modern neural networks tend to outperform humans (Vasudevan et al., 2022; Beyer et al., 2020; Shankar et al., 2020). However, this paradigm, reminiscent of the platonic ideal of a one-to-one-mapping between entities and representations (e.g., Huh et al., 2024), rests on the assumption that context does not affect how information is best encoded.

In contrast, human perception and representation are inherently adaptive, dynamically adjusting to complex and ever-changing environments composed of semantically related objects (Alvarez, 2011; Brady & Alvarez, 2015). This dynamic nature of real-world perception suggests that the human cognitive system is optimized not for the isolated processing of entities but for the integration of complex yet structured information that exploits long-term regularities of the environment and supports flexible responses to changing contexts. Representations adapt dynamically and on a moment-to-moment basis depending on what requires attention (Maljkovic & Nakayama, 1994), the simultaneous context (e.g., Bonner & Epstein, 2021; Tversky, 1977; Simonson, 1989), decoy effects (Huber et al., 1982), goals, (e.g., Molinaro & Collins, 2023), and recent experiences (also known as sequential dependencies; Mozer et al., 2007; Fischer & Whitney, 2014; Howard & Kahana, 2002; Akrami et al., 2018; Raviv et al., 2012; Fritsche et al., 2017). Whereas rapid adaptation to the surrounding environment and its context is fundamental to human perception and cognition, it is a property that current vision machine learning models typically lack. Yet these models are increasingly being used as *autoraters*, i.e., as a proxy for costly human data collection (e.g., Ouyang et al., 2022; Christiano et al., 2017; Stiennon et al., 2020). The growing use of such systems underscores the need for model representations that adapt to changes in context. This adaptation necessitates

[*]Equal contributions as first author.
[†]Equal contributions as second author.

a conceptual understanding of the world that is human-like (e.g., Hebart et al., 2020; Cichy et al., 2019; Peterson et al., 2018; Attarian et al., 2020) and, hence, inherently contextual. An alternative approach would be to remove contextual effects from human judgments (e.g., Mozer et al., 2011).

In this work, we take a step forward in building context-sensitive models of human behavior. We utilize a large-scale behavioral dataset of human similarity judgments where participants had to judge the similarity of images with a contextual reference image (Roads & Love, 2021). We derive predictions of human behavior from a diverse range of neural network models, by computing similarity from their embeddings. We compare models that do not consider contextual information to ones that incorporate it in a manner inspired by cognitive science theory. In addition to evaluating behavioral predictions, we examine how context influences the resulting (context-sensitive) representational spaces, showing how feature salience in decision making can shift in meaningful ways (see Fig. 1). Through this work, we demonstrate how cognitively inspired interventions in machine learning systems can hold rich potential to address the need for improving the alignment between machine learning models and human representations and behavior.

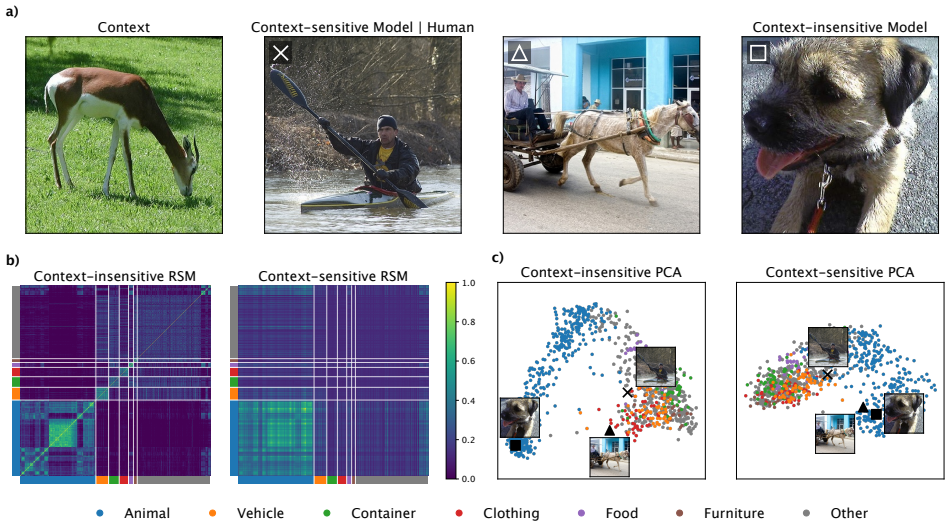


Figure 1: Qualitative comparison of context-insensitive and context-sensitive models. (a) Example triplet, where the human odd-one-out choice was made anchored in a given context. The context-sensitive model matches the human choice, while the context-insensitive model does not, reflecting the ambiguity of the triplet without context. (b) Representational similarity matrices (RSMs) over a reference image set, showing how the context-sensitive model modulates similarities given the context image in a. (c) PCA plot of the reference set and the three triplet images (same symbols as in a). The horse-cart image shifts to the animal cluster for the context-sensitive model.

2 Related Work

Judging similarity. Humans' ability to judge object similarity helps us organize the rich visual information surrounding us. This supports cognitive processes like object recognition, decision-making, and memory formation (Roads & Love, 2021; 2024). Information can be compared in many ways. Thus, similarity is a flexible concept and there is no single definition of what makes things appear alike (Goodman, 1972). Tversky (1977) famously rejected geometric theories of similarity that claim similarity is related to distance in a metric space, given evidence that human similarity judgments are asymmetric and context bound. Nosofsky (1986) showed that geometric similarity can provide excellent alignment with human judgment and downstream classification performance if the dimensions of the representation are rescaled with context-specific weights.

There are different experimental designs to extract human similarity ratings of concepts and study properties of human mental representation (e.g., Hebart et al., 2019; Muttenthaler et al., 2022), such as multi-arrangement tasks (Cichy et al., 2019; King et al., 2019), numerical/Likert-scale pairwise similarity judgments (Peterson et al., 2016; 2018), or odd-one-out triplet tasks where subjects judge

which of three items is least similar to the others (Robilotto & Zaidi, 2004; Hebart et al., 2020; 2019; Muttenthaler et al., 2022; Fukuzawa et al., 1988). Here, we use an odd-one-out triplet task setting in the reference frame of a context image (Roads & Love, 2021, see §3 for details).

Human alignment. The machine-learning community has seen growing interest in improving the alignment between human and neural network representations (e.g., Peterson et al., 2018; 2019; Attarian et al., 2020; Muttenthaler et al., 2023a;b; for a survey, see Sucholutsky et al., 2025). Research in this area has either studied the differences between these representation spaces (Muttenthaler et al., 2023a; Mahner et al., 2025) or aligned the representational structure of neural networks with the way humans represent information (e.g., semantically) to increase generalization and improve performance for downstream tasks, such as few-shot learning, anomaly detection, or object categorization (Muttenthaler et al., 2023b; 2025; Fel et al., 2022; Fu et al., 2023).

3 METHODS

Data. The analyses presented in this work are based on the ImageNet-HSJ dataset introduced by Roads & Love (2021), which consists of large-scale human similarity judgments for images of the ImageNet ILSVRC validation set. The dataset was originally collected in an 8-choose-2 task setting, in which participants view a display of nine images in a grid, with a query image placed in the center, and select two reference images most similar to the query image (Roads & Love, 2021). For our primary analysis, this dataset was restructured into a triplet-with-context task format with the query image as context, an example of which is shown in Fig. 1a. Each triplet trial is composed of the two reference images chosen in a given trial along with one unselected image from the same trial, resulting in 6 triplet trials derived from one 8-choose-2 trial. The task is to predict the odd-one-out *in the context of the query*. For details on data preprocessing and splits, see Appx. C.

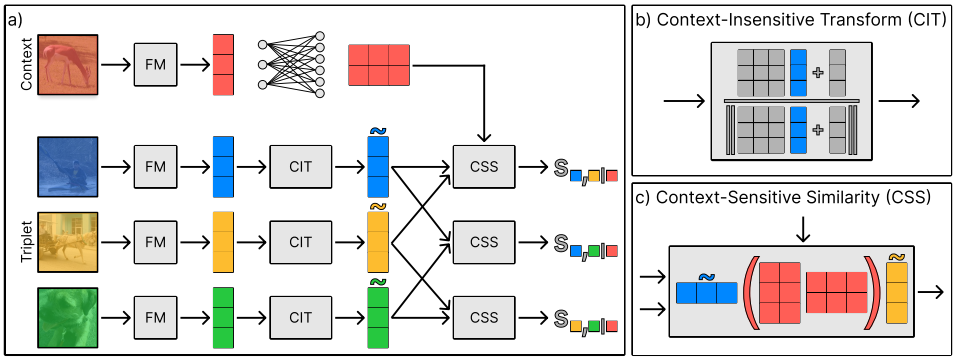


Figure 2: Method overview. A foundation model (FM) extracts embeddings for the context and triplet images. The triplet embeddings undergo a context-insensitive transform (CIT), and the context embedding is mapped to a matrix that is used to compute pairwise, context-sensitive similarities (CSS) between the transformed triplet embeddings. The trainable parameters consist of those of the context-insensitive transform and the neural network that maps the context embedding to the matrix.

Model. We propose a context-sensitive model that takes as input a triplet of images along with a context image and produces an odd-one-out decision over the triplet conditioned on the context. The model is shown in Fig. 2a. The model processes each image independently through a foundation model (FM) to obtain an embedding. A *context-insensitive transform* is then applied to improve alignment with human similarity judgments for our triplet-with-context task. A neural network, consisting of a single fully connected linear layer, maps the context-image embedding to a low-rank matrix, which is then used to calculate context-dependent similarity scores between pairs of image embeddings. Having obtained this *context-sensitive similarity* for each image pair in the triplet, we identify the pair with the greatest similarity and deem the remaining image to be the odd-one-out.

Feature extraction. To embed triplet and context images, we extract features using FMs from Muttenthaler et al. (2025) trained with three different pretraining objectives: supervised (ViT; Dosovitskiy et al., 2021), image/text contrastive (SigLIP; Zhai et al., 2023), and self-supervised (DINOv2;

Oquab et al., 2024). For each, Muttenthaler et al. (2025) provides *human-aligned* versions, obtained via fine-tuning to better match human similarity judgments, learning a static transformation of the embedding space. Our approach adds an explicit context-sensitive component, enabling similarities to dynamically shift with context rather than rely on fixed representations.

Context-insensitive transform. Although we use FMs fine-tuned for human-alignment, this may not generalize to the triplet-with-context task. To improve alignment on our task, we apply an affine transform followed by normalization to the embedding $x_i \in \mathbb{R}^d$ of image i:

$$\tilde{x}_i = \frac{Wx_i + b}{\|Wx_i + b\|},$$

where $W \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^d$ are learned parameters. Normalization is motivated by prior work indicating that the vector norm provides no useful categorical information (Scott et al., 2021).

Context-sensitive similarity. To compute the context-sensitive similarity $s_{i,j|c} \in \mathbb{R}$ between images i and j conditioned on context image c, we define:

$$s_{i,j|c} = \tilde{x}_i^\top A_c \, \tilde{x}_j,$$

where $A_c \in \mathbb{R}^{d \times d}$ is a context-dependent symmetric positive semidefinite matrix. To enforce this constraint, we parameterize $A_c = B_c^\top B_c$, with $B_c \in \mathbb{R}^{r \times d}$ and $r \ll d$. The matrix B_c is the output of a neural network applied to the context image embedding. This yields a context-dependent kernel that corresponds to the dot product between projected embeddings $B_c\tilde{x}_i, B_c\tilde{x}_j \in \mathbb{R}^r$.

Odd-one-out decision. Given a triplet of images $\{p, q, r\}$ and a context image c, the model chooses image $k \in \{p, q, r\}$ to be the odd-one-out according to the softmax probability for the similarity of the pair of non-chosen images $\{i, j\} = \{p, q, r\} \setminus \{k\}$, as $\Pr(k \mid \{p, q, r\}, c) \propto \exp\big(s_{i,j|c}\big)$.

Alignment loss and identity regularization. We adapt the approach of Muttenthaler et al. (2023b) to the context-sensitive setting. Similarly, we train the model via maximum-likelihood estimation with respect to the model parameters θ, consisting of the affine transform and the neural network parameters, to predict the human odd-one-out choice. For a set of images s, we denote the human-chosen oddball as k_s^*, leading to a loss defined over n training instances:

$$\mathcal{L}(\theta) = -\frac{1}{n}\sum_{s=1}^{n} \log \Pr(k_s^* \mid \{p_s, q_s, r_s\}, c_s) + \lambda_1 \|W - I\|_F^2 + \frac{\lambda_2}{n}\sum_{s=1}^{n} \|A_{c_s} - I\|_F^2,$$

where the latter two terms preserve the pretrained representational structure. Specifically, the second term aims to keep the context-independent transformation close to the original representation and the third term attempts to keep the context-sensitive similarity close to the cosine similarity.

4 Experimental Results

Quantitative Results. To evaluate the impact of context-sensitive modeling, we assessed odd-one-out accuracy on our triplet-with-context task. In this analysis, we considered three differently trained FMs (i.e., DINOv2, SigLIP, ViT), in addition to their human-aligned variants obtained via AligNet fine-tuning (Muttenthaler et al., 2025). For each of the six FMs, we compare the FM baseline to our context-sensitive model and to a context-insensitive model, which removes the context-sensitive similarity computation and retains only the context-insensitive transform. Experimental details are provided in Appx. D. Across all conditions, both the context-insensitive and context-sensitive models improved over the FM baseline, with the context-sensitive model yielding the largest gains. Paired bootstrap analyses on held-out test trials confirmed that all stepwise improvements were reliable, with 95% confidence intervals excluding zero in every comparison (see Appx. E for further details). Fig. 3 presents six sets of bars indicating odd-one-out accuracy on a held-out test set of data (for detailed performances across models, see Appx. G.1). For each set, the FM baseline (blue bar) consistently shows a weaker odd-one-out prediction accuracy than a variant with only a context-insensitive transform (orange bar). Critically, incorporating a context-sensitive modulation of the embeddings (red bar) further improves accuracy. These stepwise gains are observed across all three FMs and both original and human-aligned representations.

Qualitative Results. To assess the effect of context-sensitive modeling, we visualize representational similarity matrices (RSMs; Kriegeskorte et al., 2008) and PCAs of the representation spaces

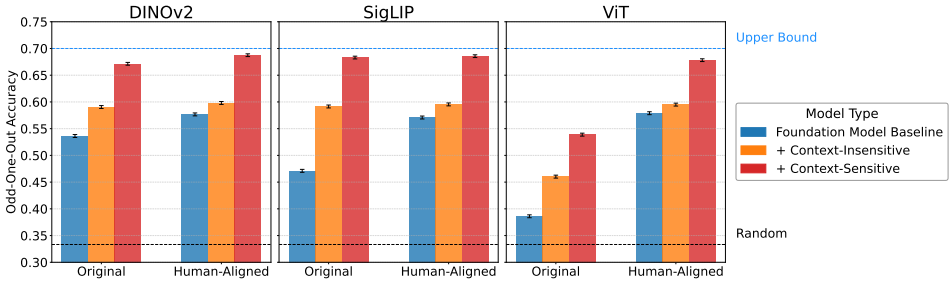


Figure 3: Odd-one-out accuracy on our triplet-with-context task for original and human-aligned DINOv2, SigLIP, and ViT bases. The context-sensitive models consistently outperform both FM baselines and context-insensitive transforms. Error bars denote 95% confidence intervals computed over held-out test trials. See Appx. F on upper bound estimation.

produced by the context-insensitive and context-sensitive models for the example triplet and context shown in Fig. 1a; for more examples, see Appx. G.2. For PCA, we interpret the projected embeddings as the context-sensitive representations. As shown in Fig. 1a, the context-insensitive model selects the dog as the odd-one-out. Consistently, the context-insensitive RSM (Fig. 1b, left) assigns higher similarity to the canoe and horse-cart images, both showing vehicles, than to the dog and horse-cart images, both containing animals. In the corresponding PCA plot (Fig. 1c, left), the dog clusters with other animal images and is separated from the vehicle-related images, including the horse-cart. Given the context image of a deer, humans select the canoe as the odd-one-out, a choice mirrored by the context-sensitive model (Fig. 1a). In the context-sensitive RSM (Fig. 1b, right), similarities between images containing animals are upweighted, while similarities between images of vehicles are downweighted. In the corresponding PCA plot (Fig. 1c, right), this modulation pushes the horse-cart image toward the animal cluster, reflecting its reinterpretation in the given context.

5 Discussion

We introduce a psychologically plausible, context-sensitive modulation of neural network representations, and demonstrate that it improves the degree of alignment with human judgments. Our approach draws inspiration from cognitive theories of similarity (e.g., Tversky's features of similarity; Tversky, 1977), introducing selective feature re-weighting in the presence of contextual cues. Qualitative analyses highlight a context-induced reshaping of the embedding space. Notably, our model re-weights arbitrary directions in the embedding space whereas human dimensional attention is restricted to predetermined axes (Kruschke, 1993). Quantitatively, our context-sensitive modulation leads to substantial improvements in predicting human odd-one-out judgments across a range of state-of-the-art FMs. Interestingly, combining context-sensitive modulation with human-alignment via fine-tuning (Muttenthaler et al., 2025) yields remarkably similar performance levels across different FMs. Relative benefits from both human alignment and contextual modulation seem to be largest for supervised models, suggesting that this kind of pretraining benefits the most from re-organizing the representation space reflecting context. These results indicate the potential of cognitively inspired modelling to better reflect human behavior.

Our current work has limitations (see Appx. B). Future approaches ought to incorporate additional sources of contextual bias, including sequential dependencies (e.g., Mozer et al., 2010; 2011; Jones et al., 2013; 2022; Akrami et al., 2018; Fritsche et al., 2017), internal goals or motivational states (e.g., Molinaro & Collins, 2023) and expand to vision–language models capable of visual reasoning across multiple images (Zhao et al., 2024). Although the odd-one-out task treats the image comparisons as symmetric, other human judgment tasks require asymmetric comparison to a reference image. In such cases, Tversky's observation of asymmetries in similarity ought to come into play to further improve human-machine alignment.

Acknowledgments. We thank Maria Eckstein for helpful feedback and insightful comments on this work. We also thank Lorenz Linhardt for helpful discussions and contributions to the conceptualization of the project. Finally, we thank Anouk Bielefeldt and Hannah Louisa Boldt for their assistance with research support tasks.

REFERENCES

Athena Akrami, Charles D Kopec, Mathew E Diamond, and Carlos D Brody. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature*, 554(7692):368–372, 2018.

George A Alvarez. Representing multiple objects as an ensemble enhances visual cognition. *Trends in cognitive sciences*, 15(3):122–131, 2011.

Maria Attarian, Brett D Roads, and Michael C Mozer. Transforming neural network visual representations to predict human judgments of similarity. *arXiv preprint arXiv:2010.06512*, 2020.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Michael F Bonner and Russell A Epstein. Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature communications*, 12(1):4081, 2021.

Timothy F Brady and George A Alvarez. Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of vision*, 15(15):6–6, 2015.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. volume 30, 2017.

Radoslaw M Cichy, Nikolaus Kriegeskorte, Kamila M Jozwik, Jasper JF van den Bosch, and Ian Charest. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *Neuroimage*, 194:12–24, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. volume 35, pp. 9432–9446, 2022.

Jason Fischer and David Whitney. Serial dependence in visual perception. *Nature neuroscience*, 17 (5):738–743, 2014.

Matthias Fritsche, Pim Mostert, and Floris P de Lange. Opposite effects of recent history on perception and decision. *Current Biology*, 27(4):590–595, 2017.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9f09f316a3eaf59d9ced5ffaefe97e0f-Paper-Conference.pdf.

Kazuyoshi Fukuzawa, Motonobu Itoh, Sumiko Sasanuma, Tsutomu Suzuki, Yoko Fukusako, and Tohru Masui. Internal representations and the conceptual operation of color in pure alexia with color naming defects. *Brain and Language*, 34(1):98–126, 1988.

Nelson Goodman. Seven strictures on similarity. In Nelson Goodman (ed.), *Problems and projects*. Bobbs-Merrill, 1972.

Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10):e0223792, 2019.

Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185, 2020.

Marc W Howard and Michael J Kahana. A distributed representation of temporal context. *Journal of mathematical psychology*, 46(3):269–299, 2002.

Joel Huber, John W. Payne, and Christopher Puto. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, 9(1):90–98, 1982.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Matt Jones, Tim Curran, Michael C Mozer, and Matthew H Wilder. Sequential effects in response time reveal learning mechanisms and event representations. *Psychological review*, 120(3):628, 2013.

Matt Jones, Tyler R Scott, Mengye Ren, Gamaleldin Fathy Elsayed, Katherine Hermann, David Mayo, and Michael Curtis Mozer. Learning in temporally structured environments. In *The Eleventh International Conference on Learning Representations*, 2022.

Marcie L King, Iris IA Groen, Adam Steel, Dwight J Kravitz, and Chris I Baker. Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, 197:368–382, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

John K Kruschke. Human category learning: Implications for backpropagation models. *Connection Science*, 5(1):3–36, 1993.

Florian P Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*, 7(6):848–859, 2025.

Vera Maljkovic and Ken Nakayama. Priming of pop-out: I. role of features. *Memory & cognition*, 22(6):657–672, 1994.

Gaia Molinaro and Anne GE Collins. Intrinsic rewards explain context-sensitive valuation in reinforcement learning. *PLoS Biology*, 21(7):e3002201, 2023.

Michael C Mozer, Sachiko Kinoshita, Michael Shettel, et al. Sequential dependencies in human behavior offer insights into cognitive control. volume 1, pp. 180–193. 2007.

Michael C Mozer, Harold Pashler, Matthew Wilder, Robert V Lindsey, Matt Jones, and Michael N Jones. Improving human judgments by decontaminating sequential dependencies. *Advances in neural information processing systems*, 23, 2010.

Michael C Mozer, Benjamin Link, and Harold Pashler. An unsupervised decontamination procedure for improving the reliability of human judgments. *Advances in neural information processing systems*, 24, 2011.

Lukas Muttenthaler, Charles Y Zheng, Patrick McClure, Robert A Vandermeulen, Martin N Hebart, and Francisco Pereira. Vice: Variational interpretable concept embeddings. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 33661–33675. Curran Associates, Inc., 2022.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=ReDQ1OUQR0X.

Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50978–51007. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9febda1c8344cc5f2d51713964864e93-Paper-Conference.pdf.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, 2025.

Robert M Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1):39, 1986.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. Adapting deep network features to capture psychological representations. In Anna Papafragou, Daniel Grodner, Daniel Mirman, and John Trueswell (eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, pp. 2093–2098. Cognitive Science Society, 2016.

Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42 (8):2648–2669, 2018.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9617–9626, 2019.

Ofri Raviv, Merav Ahissar, and Yonatan Loewenstein. How recent history affects perception: the normative approach and its heuristic approximation. *PLoS computational biology*, 8(10): e1002731, 2012.

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*, pp. 3547–3557, 2021.

Brett D Roads and Bradley C Love. Modeling similarity and psychological space. *Annual Review of Psychology*, 75(1):215–240, 2024.

Rocco Robilotto and Qasim Zaidi. Limits of lighness identification for real objects under natural viewing conditions. *Journal of Vision*, 4(9):9–9, 2004.

Tyler R Scott, Andrew C Gallagher, and Michael C Mozer. von mises-fisher loss: An exploration of embedding geometries for supervised learning, 2021.

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pp. 8634–8644. PMLR, 2020.

Itamar Simonson. Choice based on reasons: The case of attraction and compromise effects. *Journal of consumer research*, 16(2):158–174, 1989.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize with human feedback. *arXiv preprint arXiv:2009.01325*, 2020.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine Hermann, Kerem Oktar, Klaus Greff, Martin N Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas O'Connell, Thomas Unterthiner, Andrew Kyle Lampinen, Klaus Robert Muller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Hiq7lUh4Yn. Survey Certification, Expert Certification.

Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Vijay Vasudevan, Benjamin Caine, Raphael Gontijo Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. When does dough become a bagel? analyzing the remaining mistakes on imagenet. *Advances in Neural Information Processing Systems*, 35:6720–6734, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning, 2024. URL https://arxiv.org/abs/2406.12742.

A Disclosure on the Use of Generative Artificial Intelligence

Generative artificial intelligence was used to improve grammar and wording and to assist with writing experiment code. The submission is primarily human authored and the authors hold responsibility for all content of the submission.

B Limitations

One limitation of the present study concerns the behavioral dataset used for training and evaluation. In the original ImageNet-HSJ dataset introduced by Roads & Love (2021), each trial was presented to only a single participant. As a result, our reconstructed dataset likewise contains one participant judgment per trial, meaning that models are trained on individual responses rather than on a distribution of judgments reflecting consensus in object similarity across participants. Although this data collection strategy is common in large-scale similarity datasets, including the THINGS dataset (Hebart et al., 2019), it may introduce additional variability in the training signal. A second limitation concerns the construction of the triplet-with-context dataset itself. Triplets were derived by pairing the two images selected as most similar to the query with each unselected image. While this reconstruction provides a practical way to derive contextual triplet comparisons and increases dataset size, it may introduce statistical regularities related to the dataset construction. In particular, the conversion implicitly hypothezises that the two selected items are more similar to each other than to unselected images, an assumption not explicitly required by the original behavioral task.

C Data Preprocessing and Splits

We filtered out trials in which two or more triplet images were in the same ImageNet class, ensuring that trivial choices were excluded. For the training and evaluation of our models, we created training, validation, and test splits on the trial level (of the original 8-choose-2 task), stratified by participant, such that all participants proportionally contributed an equal amount of trials to each of the splits. The resulting splits included N = 1,003,759 trials for training, N = 125,646 trials for validation, and N = 125,710 trials for testing.

D Experimental Details

In all experiments, the base variants of the FMs were used to extract 768-dimensional representations. All context-sensitive and context-insensitive models were trained for 50 epochs using stochastic gradient descent. The learning rate and batch size were set to 0.001 and 128, respectively. A grid search was conducted to find the best values for the following hyperparameters:

- the rank $r \in \{16, 32\}$ of the matrix B_c,
- the regularization strength $\lambda_1 \in \{0.0001, 0.001\}$ for the context-insensitive transform,
- the regularization strength $\lambda_2 \in \{10^{-5}, 10^{-4}, 10^{-3}\}$ for the context-sensitive similarity,
- and the softmax temperature $\tau \in \{1.0, 5.0, 7.5\}$.

The hyperparameter configuration with the highest validation accuracy was selected, and the corresponding model was evaluated on a held-out test set.

E Statistical Analysis of Model Comparisons

To assess the reliability of performance differences between models, we conducted paired bootstrap analyses for all pairs of models shown in Fig. 3. For each pair, we report the mean difference in accuracy (Δ) along with 95% confidence intervals (CI) computed over held-out test trials.

Across all FMs, both the context-insensitive and context-sensitive models improved over the FM baselines, and the context-sensitive models consistently show higher prediction accuracy compared to the context-insensitive models. All reported confidence intervals exclude zero, indicating reliable differences.

DINOv2. For the original representation, the context-insensitive model improved over the baseline by $\Delta = 0.054$ [0.052, 0.057], and the context-sensitive model by $\Delta = 0.135$ [0.132, 0.138]. The additional gain of the context-sensitive over the context-insensitive model was $\Delta = 0.081$ [0.078, 0.083]. For the human-aligned representation, improvements were $\Delta = 0.021$ [0.019, 0.023] (context-insensitive vs. baseline), $\Delta = 0.111$ [0.108, 0.114] (context-sensitive vs. baseline), and $\Delta = 0.089$ [0.087, 0.092] (context-sensitive vs. context-insensitive).

SigLIP. For the original representation, improvements were $\Delta = 0.120$ [0.117, 0.123] (context-insensitive vs. baseline), $\Delta = 0.212$ [0.209, 0.215] (context-sensitive vs. baseline), and $\Delta = 0.092$ [0.089, 0.094] (context-sensitive vs. context-insensitive). For the human-aligned representation, improvements were $\Delta = 0.025$ [0.022, 0.026], $\Delta = 0.115$ [0.112, 0.118], and $\Delta = 0.090$ [0.088, 0.093], respectively.

ViT. For the original representation, improvements were $\Delta = 0.074$ [0.071, 0.077] (context-insensitive vs. baseline), $\Delta = 0.153$ [0.149, 0.156] (context-sensitive vs. baseline), and $\Delta = 0.079$ [0.075, 0.082] (context-sensitive vs. context-insensitive). For the human-aligned representation, improvements were $\Delta = 0.016$ [0.014, 0.018], $\Delta = 0.099$ [0.097, 0.102], and $\Delta = 0.083$ [0.081, 0.086], respectively.

F Upper Bound Estimation

Each 8-choose-2 trial, and thus each triplet-with-context trial, is observed by exactly one participant and therefore yields only a single response. Estimating an upper bound on odd-one-out accuracy requires multiple responses per trial. To address this, we identify trials by their ImageNet class labels rather than image identities. As a result, multiple image-level trials are mapped to the same class-level trial, yielding multiple responses for a subset of trials.

For each trial in this subset, we estimate the empirical distribution of odd-one-out responses. Let this distribution be $p = (p_1, p_2, p_3)$, where p_i denotes the probability that the i-th image in the triplet is selected as the oddball. The optimal strategy is to predict the image with the highest probability, yielding an expected odd-one-out accuracy of $\max_i p_i$.

We estimate the upper bound on odd-one-out accuracy as the average of $\max_i p_i$ across the subset of trials with multiple responses.

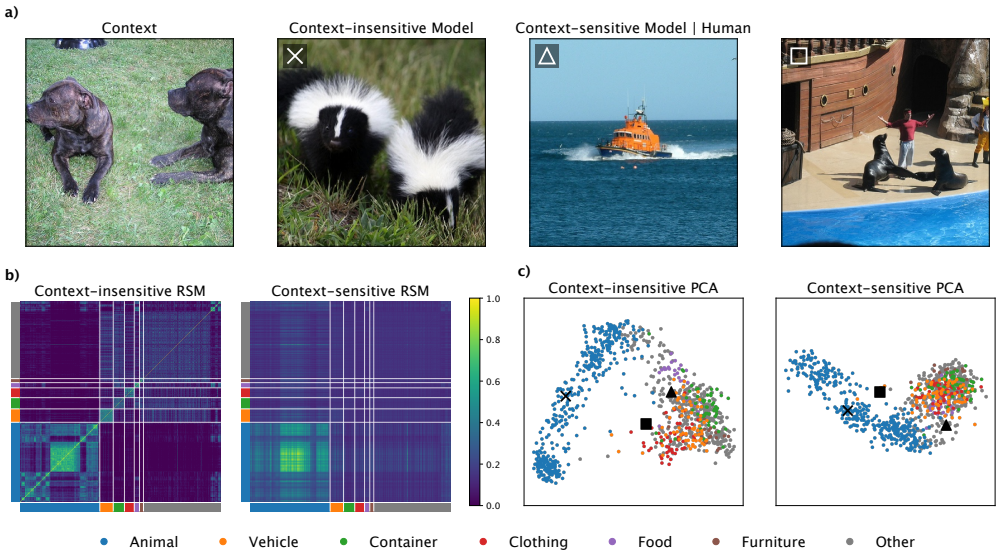
G Supplementary Results

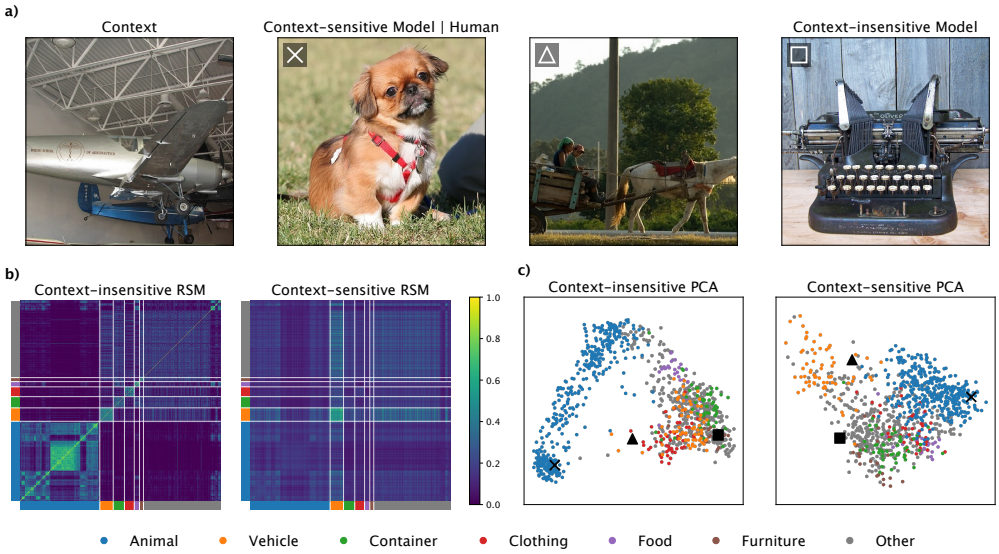
G.1 Quantitative Results

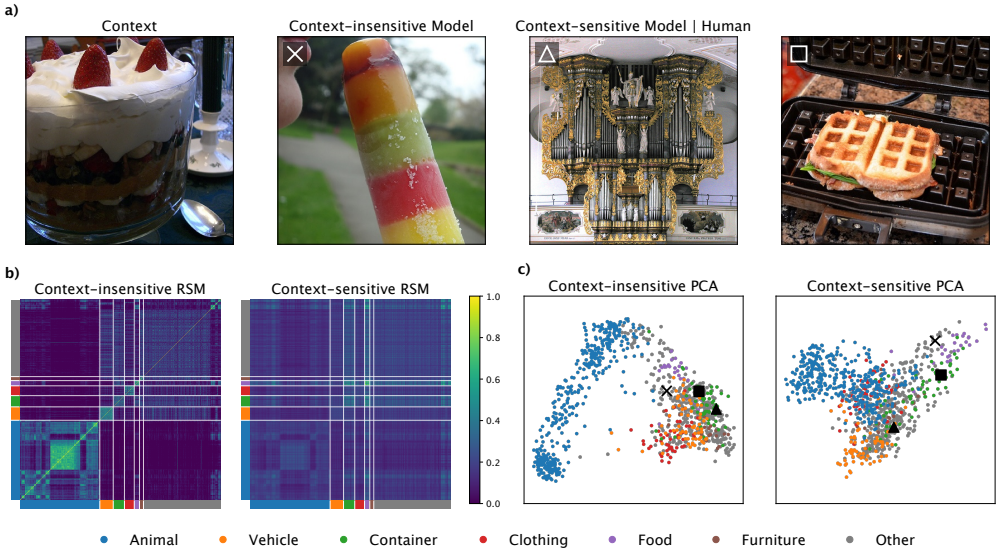
Table 1: Triplet odd-one-out accuracies for original and human-aligned foundation model bases on the triplet-with-context task. We report accuracies for the foundation model baselines (FM), context-insensitive models (CI), and context-sensitive models (CS). Relative improvements are computed with respect to the foundation model baselines and context-insensitive models.

	FM	CI	CS	↑ FM (%)	↑ CI (%)
DINOv2 (original)	0.536	0.590	0.657	22.4	11.2
DINOv2 (human-aligned)	0.577	0.598	0.687	19.2	14.9
SigLIP (original)	0.471	0.582	0.682	44.9	17.4
SigLIP (human-aligned)	0.571	0.595	0.685	20.0	15.0
ViT (original)	0.386	0.460	0.539	39.6	17.1
ViT (human-aligned)	0.579	0.585	0.672	16.1	14.8

G.2 Qualitative Results


a)
Context
Context-insensitive Model
Context-sensitive Model | Human
b)
Context-insensitive RSM
Context-sensitive RSM
1.0
0.8
0.6
0.4
0.2
0.0
c)
Context-insensitive PCA
Context-sensitive PCA
Animal
Vehicle
Container
Clothing
Food
Furniture
Other


a)
Context
Context-sensitive Model | Human
Context-insensitive Model
b)
Context-insensitive RSM
Context-sensitive RSM
1.0
0.8
0.6
0.4
0.2
0.0
c)
Context-insensitive PCA
Context-sensitive PCA
Animal
Vehicle
Container
Clothing
Food
Furniture
Other


a)
Context
Context-insensitive Model
Context-sensitive Model | Human
b)
Context-insensitive RSM
Context-sensitive RSM
1.0
0.8
0.6
0.4
0.2
0.0
c)
Context-insensitive PCA
Context-sensitive PCA
Animal
Vehicle
Container
Clothing
Food
Furniture
Other