MISE: a computational model for simulating visual perception and cognitive decision-making via CLIP

Anonymous Author(s)
Affiliation
Address
email

Abstract

Visual perceptual stimuli are essential for human cognitive decision-making. Especially by extracting visual semantic features to represent objects helps to understand the objects in the physical world. Also, it explains the influence of visual stimuli on behavioral decision-making. However, existing methods can already extract interpretable visual features related to decision making from this task without brain imaging data, but they still have some limitations. These methods mainly focus on embedding concepts rather than the embedding of images, and leads to their limited generalization ability. Therefore, we developed a computational model based on CLIP and a linear mapping function to simulate visual perception and cognitive decision-making. The model can learn the semantic representation of visual stimuli and decision-making using the linear mapping function. Experimental results show that our model captures the most interpretable differences in similarity judgment, and obtains 49 object dimensions with high reproducibility and significance, which reflect various semantic features and perceptual attributes of these objects. In addition, human and AI-agent subjects can accurately evaluate objects based on these dimensions, highlighting their interpretability.

1 Introduction

The visual stimuli representation of human visual perception is essential for human cognitive decision-making tasks[17, 16]. Visual perception refers to the process of extracting features, attributes, and semantic information of objects from visual stimuli. In contrast, cognitive decision-making is the process of judging, reasoning and choosing objects or scenes based on perceptual information and prior knowledge. In addition, semantic features are the objects or conceptual and high-level information(e.g., categories, functions, meanings)[1, 5, 24, 9, 29, 28] which play an essential role in human behavioral decision-making. They can help people understand and predict objects or scenes' attributes, states, and changes, thus making rational and effective decisions. However, despite the widely recognized connection between visual perception and cognitive decision-making, there still needs to be more in-depth and systematic on the relationship between the representation of visual perceptual stimuli and human behavior cognitive decision-making tasks[13, 23, 33, 15, 30].

In recent years, visual models have been essential in exploring these problem[23, 33]. Visual models are computational tools for simulating the human visual system, which can be used for various tasks in human vision, such as visual representation, decision-making, and learning. Visual representation is the core part of visual models because it determines the model's ability to capture and encode image information and support subsequent tasks. Visual decision-making is the application part of visual models because it reflects the model's ability to understand and utilize image information and solve practical problems [12, 7].

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

In order to further improve the performance of visual models, many image-text multimodal models have emerged [20, 21, 31, 10, 19]. These models process image and text information simultaneously and use their correlation and complementarity to improve model performance, such as CLIP[26, 25]. Image-text multimodal models can use semantic features provided by text information to enhance visual representation, thereby improving the accuracy and robustness of visual decision-making tasks. At the same time, image-text multimodal models can also use visual features provided by image information to enhance text representation, thereby improving the naturalness and diversity of text generation or understanding tasks. Moreover, those models can also evaluate the model's ability to capture and encode semantic features by comparing the similarity or differences between different modalities, thus inspiring cognitive science research on human visual representation mechanisms and principles.

In order to study the relationship between the representation of visual perceptual stimuli and human behavior cognitive decision-making tasks in depth, we chose the odd-one-out task[13, 6, 32, 12] as an experimental setting. Odd-one-out is a fundamental cognitive task that requires participants to find out the different or mismatched stimulus from a group of stimuli and provide corresponding reasons or explanations.

Although existing methods can extract interpretable visual features related to decision-making from this task without brain imaging data [33, 13, 23], they still have some limitations: These methods mainly focus on conceptual (e.g., objects or categories in images) embeddings rather than image embeddings. Although these methods can generalize in new experimental environments (unseen triplets composed of novel objects), their generalization ability is limited for new objects. For the intuitive interpretability of features, these methods often require new behavioral experiments to explain features.

Therefore, we used AI techniques (i.e., CLIP) and linear mapping functions to construct a model that simulates visual stimuli to behavioral decision-making and applied it to odd-one-out tasks. Experimental results show that our model performs with high accuracy in odd-one-out tasks and can generate reasonable and diverse reasons or explanations. Through in-depth analysis of model performance, we further explore the relationship between the representation of visual perceptual stimuli and human behavior cognitive decision-making tasks.

2 Related work

Interpretable semantic representation refers to a numerical model that can reflect the intrinsic structure and features of semantic concepts, usually using vector space or probabilistic graph and other forms to represent[33]. In order to learn interpretable semantic representation from human behavioral data, oddity task is a common experimental paradigm, which requires subjects to select one of the three stimuli (such as words, images, sounds, etc.) that is most dissimilar to the other two stimuli when given[13, 23]. Oddity task can stimulate different minimal contexts as the basis for grouping objects, thereby highlighting relevant dimensions.

Interpretable semantic representation based on words or images [3, 2]. This aspect of research aims to use oddity task to learn vector space representation of word or image concepts, such as using non-negative matrix decomposition, variational autoencoder[14, 27] , or variational interpretable concept embedding and other methods. These methods can obtain sparse, non-negative, uncertain estimation of semantic representation, and automatically select the dimensions that can best explain the data. Interpretable semantic representation based on multimodal or cross-modal. This aspect of research aims to use oddity task to learn semantic association between different modalities (such as vision, hearing, touch, etc.) or cross-modal (such as text and image), such as using multimodal autoencoder , multimodal deep belief network , or multimodal variational autoencoder and other methods. These methods can obtain semantic representation that can capture multimodal or cross-modal commonality and difference, and realize transformation and generation between different modalities. Interpretable semantic representation based on neural network or brain activity. This aspect of research aims to use oddity task to learn the implicit semantic structure in neural network or brain activity, such as using neural network model , representational similarity analysis , or neural encoding and decoding and other methods. These methods can obtain semantic representation that can predict human behavior or brain activity, and reveal semantic dimensions in neural network or brain activity.

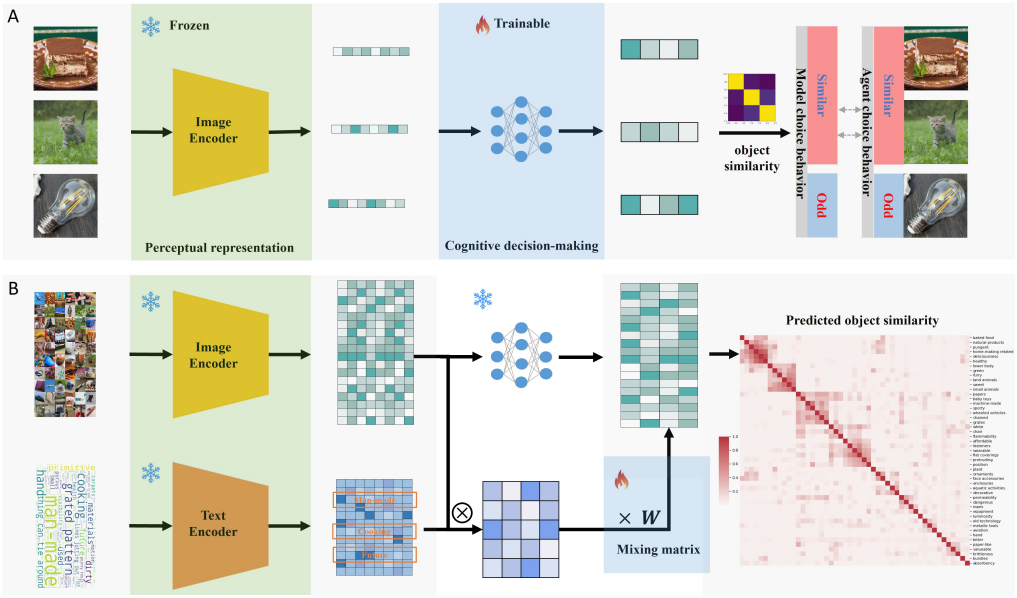

Figure 1: A model simulating the process from visual stimuli to behavioral decision-making is established and applied to the odd-one-out task. A. Starting from the left of the figure, we input a triplet of image data, which is processed by a frozen pre-trained visual encoder as the perception module, yielding three sets of visual features. These features are then input into a learnable cognitive decision module, resulting in three sets of decision representations. By calculating the similarity among these three representations, we derive the decision outcome. Comparing the model's results with the choices of human participants, we calculate the loss and perform backpropagation to train our cognitive decision module. B. After training the cognitive decision module, we can obtain the behavioral encoding of all visual objects, which we refer to as the codebook. With the codebook, we can predict the representational similarity matrix for all objects, enabling us to perform the decision task on unseen image triplets. To interpret the features in the codebook, we input a large number of potential descriptive words (properties) of visual objects into a text encoder. Then, by matching text features and image encoding, we obtain the property encoding of all objects. Through a simple linear model, we establish a linear mapping relationship (mixing matrix) between property encoding and the codebook. Based on this, we can reveal the connection between features in the codebook and semantic properties.

3 Method

3.1 Task description

The triplet task, also known as the triple odd-out task, is used to discover object concept embeddings through similarity judgments on a set of m different objects. These judgments were collected from human participants who were given queries consisting of triple objects. Participants were asked to consider three pairs in a three-pair group and decide which pair was most similar, leaving the third pair as an odd-one-out.

3.2 Multimodal Interpretable Semantic Embeddings (MISE)

Learn semantic representation constraint by behavior We use a multimodal model (such as CLIP) to build a model simulating the process from visual stimuli to behavioral decision-making and apply it to the odd-one-out task. We call it Multimodal Interpretable Semantic Embeddings (MISE).

Fig1A illustrates a model simulating the process from visual stimuli to behavioral decision-making, applied to an odd-one-out task. We start by inputting a triplet of image data into a pre-trained visual encoder (serving as the perception module). This fixed encoder extracts three sets of visual features. These features are then fed into a learnable cognitive decision module, resulting in three sets of decision representations. By calculating the similarity between these three representations, we arrive

at the decision outcome. Comparing the model's results with the choices of human participants, we calculate the loss and perform backpropagation to train our cognitive decision module. The loss function we use is the same as the SPoSE [33] algorithm used in previous research.

$$h_I = E_I(I), h_B = E_B(h_I), L = L_{\text{SPoSE}}(h_B) \tag{1}$$

Where E_I is the pre-trained visual encoder, E_B is the learnable cognitive decision module, L_{SPoSE} is the function used to calculate the loss, I, h_I, and h_B are the image, visual features, and behavioral features, respectively.

Although we could train a code for each image rather than each concept, to accommodate the training data that only recorded concept indices and to facilitate comparison with previous experiments, we still learn representations based on concepts. However, our approach can also be applied directly to decoding individual images.

Map features to semantic propreties Fig 1B depicts the process followed after training the cognitive decision module. We can compute the behavioral encoding for all visual objects, creating what we refer to as a "codebook". This codebook allows us to predict the representational similarity matrix for all objects, thereby enabling us to carry out the odd-one-out decision task on image triplets that the model has not previously encountered.

Based on prior research[33, 13, 23], the features within the 'codebook' often contain rich semantic characteristics. When they satisfy sparsity and non-negativity, the model's performance tends to be closer to human behavior. Therefore, interpreting these features helps us further understand the process of human visual decision-making. To interpret the features present in the codebook, we input a multitude of potential descriptors (properties) for visual objects into a text encoder. By matching these text features with image encodings, we are able to obtain property encodings for all objects. Here, other language models can also be used to generate property encodings.

Subsequently, we employ a simple linear model to establish a linear mapping relationship, referred to as a "mixing matrix", between the property encodings and the codebook. This process enables us to uncover the relationship between the features stored in the codebook and the semantic properties they represent.

4 Experiments

We used three datasets: a human subject behavioral dataset, an AI-agent subject behavioral dataset, and a visual property words dataset. These datasets aim to help us understand the impact of visual stimuli on human and AI behavioral cognitive decision-making.

For the human subject behavioral dataset, we chose the THINGS dataset[7], which is a publicly available dataset widely used in visual perception and cognition research. The THINGS dataset contains 1854 diverse object concepts, and random samples of all possible triplet combinations of different objects. There are no duplicate data, that is, each triplet has only one human response. About 10% of the triplets are assigned to a predefined validation set. Both test sets contain results of 1000 random triplets that are not included in the training data, and each triplet has 25 repetitions. We divided this dataset into training set (90%) and test set (10%), with no overlap between triplets.

At the same time, we also generated an AI-agent subject behavioral datasets, which can help us better understand and compare the behavioral patterns of humans and AI. Specifically, we used the image encoder of CLIP to do the triplet odd-one-out task. Similar to the humans subject on triplet odd-one-out task, we input the THINGS triplet images as visual stimuli to the AI-agent and obtained its visual features. We used cosine similarity to calculate the similarity between pairs of images, and obtained the most similar two images, leaving the remaining one as odd.

Finally, we collected descriptive words from previous research and the Internet, and established a visual feature descriptive word library, covering the target semantic features required for understanding visual objects and performing visual tasks. We also verified other ai-agents (Alexnet, Resnet, VIT [18, 11, 4]) in supp.

Table 1: Comparison of different models for behavioural prediction. The accuracy in the test set for each model is listed. MISE variants and a MISE AI-agent are compared to a baseline (Chance) and an upper bound (noise ceiling).

	Behavioural prediction
Chance*	33.33%
MISE + MLP	63.41%
MISE + mixing	58.35%
MISE + softplus	61.15%
MISE + L1(0.001)	62.77%
MISE	62.76%
noise ceiling*	66%
MISE AI-agent	65.56%

4.1 Model validation

Table 1 shows the impact of different model details on prediction accuracy. 'Chance' represents the accuracy of random guessing, while 'noise ceiling' represents the highest accuracy estimated due to human decision noise. In the MISE model, we used a single-layer linear model. To avoid the problem of neuron death caused by the ReLU function[8], we used LeakyReLU[22] as the activation function to ensure that the values in the codebook are mainly positive. We used L1 regularization to make the codebook sparse, and the regularization coefficient was set to 0.01. However, our experimental results show that the model performs well in terms of the stability of this coefficient. In MISE+MLP, we used a three-layer fully connected network; in MISE+mixing, we directly used property encoding to predict behavior; in MISE+softplus, we used the softplus activation function. The results show that our model performs stably under various model variants and training methods. We used the Adam optimizer, and the learning rate was set to 0.001. As shown in Fig2, after training, we can obtain the behavioral encoding of all objects to form a codebook (lower left). Based on these encodings, we can calculate the representational similarity matrix for behaviors (right), which allows us to perform decision-making tasks on unseen image triplets (upper right). In addition, we can further verify the predictive performance and accuracy of the model by comparing it with the choices of human participants.

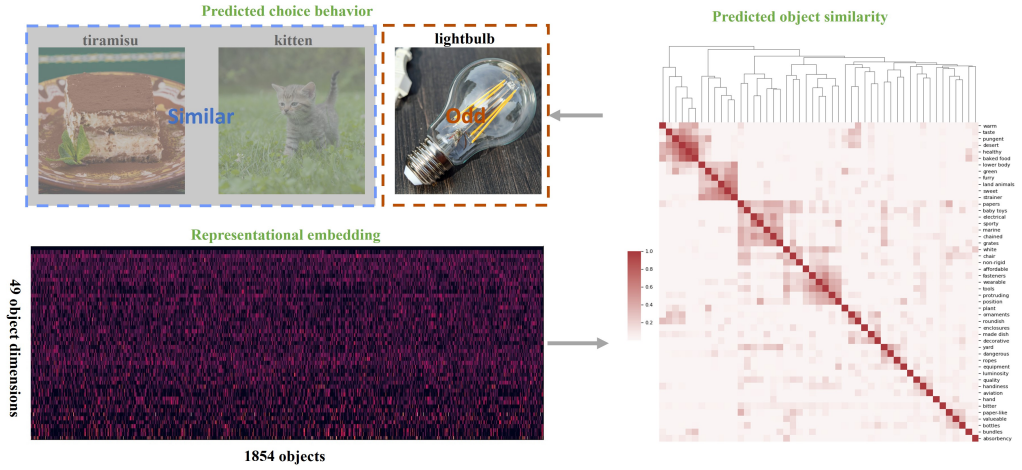

Figure 2: An illustration of how the codebook simulates human decision-making behavior. The codebook simulates human decision-making behavior. Once the cognitive decision module is trained, we obtain the behavioral encoding for all objects, forming the codebook (bottom left). Based on these encodings, we can compute the representational similarity matrix for behaviors (right), which in turn facilitates decision-making tasks (top right).

4.2 Interpretability of semantic features

In this section, we validate the interpretability of our model. As shown in the word cloud illustrations(Fig3), our semantic features are interpretable [13]. We generate word clouds based on the properties features, which are derived from a CLIP-based text encoder and decoded through a properties mixing matrix. The size of each word in the cloud corresponds to its proportion in the feature, as determined by the associated weights in the properties mixing matrix. For visual effect, the size of the largest word has been adjusted. Alongside these word clouds, we provide example images that represent the highest weights along these semantic features. These word clouds serve to visualize the semantic interpretations of these features, offering insights into their semantic meanings.

Moreover, we depict example objects and their corresponding semantic features through bar graphs (Fig4). The height of each bar indicates the extent to which a semantic feature is expressed for a given object image. For clarity in visualization, features with small weights are not labeled. These visualizations further enhance our understanding of how the model interprets and represents these objects, underscoring the model's interpretability.

Figure 3: Illustration of semantic features using word clouds for interpretability. Our word clouds are generated based on the properties features derived from a CLIP-based text encoder, and decoded through a properties mixing matrix. The size of each word in the cloud corresponds to its proportion in the feature, determined by the associated weights in the properties mixing matrix. For visual effect, the size of the largest word has been adjusted. The images represent the highest weights along these semantic features. The word clouds visualize the semantic interpretations of these features, offering insights into their semantic meanings.

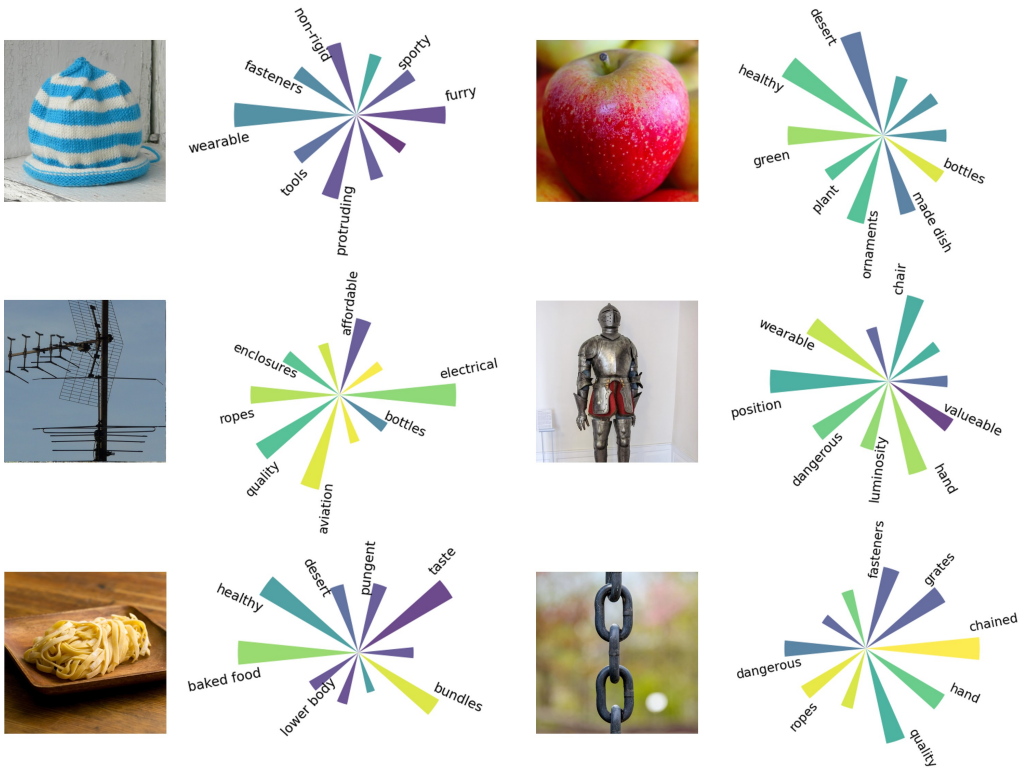

Figure 4: Depiction of example objects and their corresponding semantic features, represented through bar graphs. The height of each bar indicates the extent to which a semantic feature is expressed for a given object image. For clarity in visualization, features with small weights are not labeled.

Table 2: most significant features in human and AI-agent

features	importance of human features	importance of ANN features
top 1	mammal (0.40)	natural products (0.40)
top 2	healthy (0.34)	baked food (0.37)
top 3	baked food (0.34)	equipment (0.37)
top 4	green (0.32)	land animals (0.35)
top 5	electrical (0.31)	wheeled vehicles (0.34)
top 6	wearable (0.31)	healthy (0.34)
top 7	marine (0.30)	warm clothes (0.34)
top 8	sweet (0.30)	red (0.34)
top 9	taste (0.30)	ties (0.33)
top 10	quality (0.30)	sweet (0.33)

4.3 Comparison human with AI agent

In this section, we validate the interpretability of our model. One way to demonstrate this is by examining the semantic representation changes in the codebook between human subjects and AI-agents. As shown in the Fig5, the left side presents the features of human subjects, while the right side displays those of AI-agents [12].

The strength of the connection between these features is determined by the Pearson correlation coefficient. In order to provide a clearer view, we have removed connections with a Pearson correlation coefficient less than 0.3 (Tab2). This visualization shows a strong correlation between the

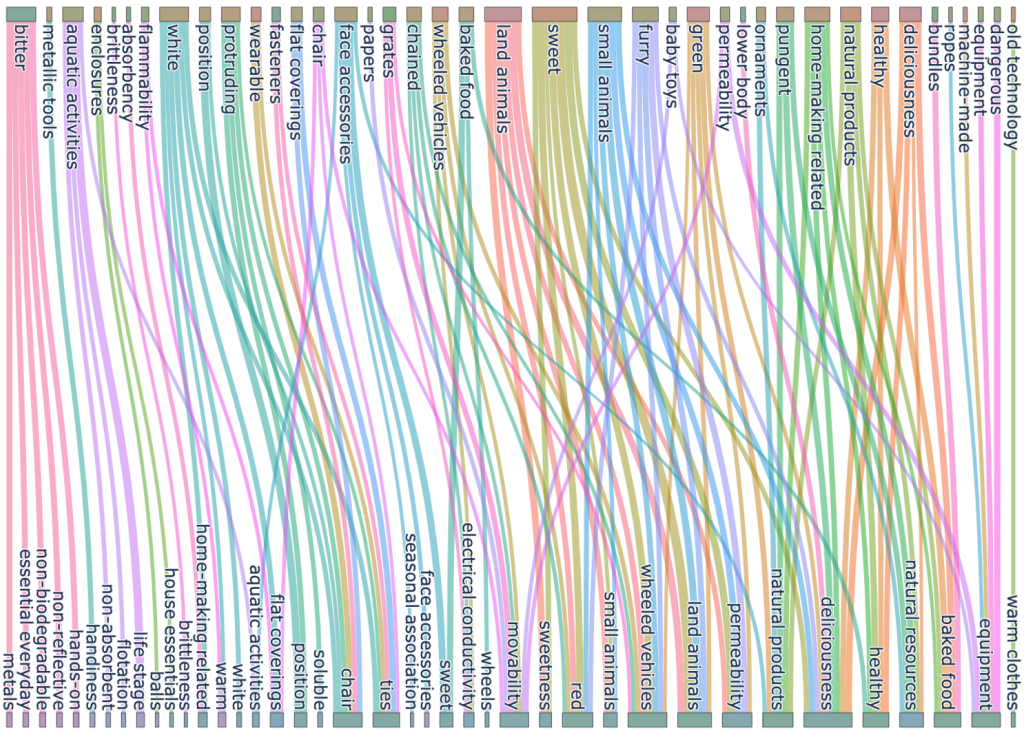

Figure 5: The change of codebook semantic representation between human subjects and AI-agents. The left side of the figure shows the features of human subjects, and the right side shows the features of AI-agents. The connection strength is determined by the Pearson correlation coefficient. For clarity, we removed connections with a Pearson correlation coefficient greater than 0.3. As can be seen, there is a strong correlation between human subjects and AI-agents. However, there are differences in some features.

features represented by human subjects and AI-agents, indicating that our AI model is able to capture human-like semantic features.

However, it is also important to note that there are certain differences in some features. These differences provide valuable insights into areas where the AI-agent deviates from human cognition, offering potential avenues for further improvement of the model.

5 Conclusion

We proposed a novel computational model that simulates visual perception and cognitive decision-making based on visual semantic features. Our model leverages the powerful CLIP model to learn the semantic representation of visual stimuli and uses a linear mapping function to make decisions. We demonstrated the effectiveness and interpretability of our model through various experiments, such as similarity judgment and object evaluation. We also showed that our model can generalize well to different types of objects and images, and can align well with human and AI-agent judgments. Our model provides a new perspective and tool for studying the relationship between visual perception and cognitive decision-making, and opens up new possibilities for future research in this field.

References

[1] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94 2:115–147, 1987.

[2] J. R. Binder, L. L. Conant, C. J. Humphries, L. Fernandino, S. B. Simons, M. Aguilar, and R. H. Desai. Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33:130 – 174, 2016.

[3] B. Devereux, L. K. Tyler, J. Geertzen, and B. Randall. The centre for speech, language and the brain (cslb) concept property norms. *Behavior Research Methods*, 46:1119 – 1127, 2013.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[5] S. Edelman. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21:449 – 467, 1996.

[6] B. Fernando, H. Bilen, E. Gavves, and S. Gould. Self-supervised video representation learning with odd-one-out networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5729–5738, 2016.

[7] A. T. Gifford, K. Dwivedi, G. Roig, and R. M. Cichy. A large and rich eeg dataset for modeling human visual object recognition. *Neuroimage*, 264, 2022.

[8] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[9] R. L. Goldstone. The role of similarity in categorization: providing a groundwork. *Cognition*, 52:125–157, 1994.

[10] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, and S. Hoi. From images to textual prompts: Zero-shot vqa with frozen large language models. *ArXiv*, abs/2212.10846, 2022.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[12] M. N. Hebart, O. Contier, L. Teichmann, A. Rockter, C. Y. Zheng, A. Kidder, A. Corriveau, M. Vaziri-Pashkam, and C. Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12, 2023.

[13] M. N. Hebart, C. Y. Zheng, F. Pereira, and C. I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgments. *Nature human behaviour*, 4:1173 – 1185, 2020.

[14] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

[15] P. Kaniuth and M. N. Hebart. Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. *NeuroImage*, 257, 2021.

[16] M. A. Kramer, M. N. Hebart, C. Baker, and W. A. Bainbridge. The features underlying the memorability of objects. *Science Advances*, 9, 2023.

[17] N. Kriegeskorte and R. A. Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17:401 – 412, 2013.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

[19] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023.

[20] J. Li, D. Li, C. Xiong, and S. C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

[21] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He. Scaling language-image pre-training via masking. *ArXiv*, abs/2212.00794, 2022.

[22] A. L. Maas. Rectifier nonlinearities improve neural network acoustic models. 2013.

[23] L. Muttenthaler, C. Y. Zheng, P. McClure, R. A. Vandermeulen, M. N. Hebart, and F. Pereira. Vice: Variational interpretable concept embeddings. *ArXiv*, abs/2205.00756, 2022.

[24] R. M. Nosofsky. Attention, similarity, and the identification-categorization relationship. *Journal of experimental psychology. General*, 115 1:39–61, 1986.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[26] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.

[27] X. Ran, J. Zhang, Z. Ye, H. Wu, Q. Xu, H. Zhou, and Q. Liu. Deep auto-encoder with neural response. *arXiv preprint arXiv:2111.15309*, 2021.

[28] T. T. Rogers and J. L. McClelland. Semantic cognition: A parallel distributed processing approach. 2004.

[29] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[30] J. J. Singer, R. M. Cichy, and M. N. Hebart. The spatiotemporal neural dynamics of object recognition for natural images and line drawings. *The Journal of Neuroscience*, 43:484 – 500, 2022.

[31] A. M. H. Tiong, J. Li, B. Li, S. Savarese, and S. C. H. Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *ArXiv*, abs/2210.08773, 2022.

[32] J. Valenti and C. Firestone. Finding the "odd one out": Memory color effects and the logic of appearance. *Cognition*, 191:103934, 2019.

[33] C. Y. Zheng, F. Pereira, C. I. Baker, and M. N. Hebart. Revealing interpretable object representations from human behavior. *ArXiv*, abs/1901.02915, 2019.