EVALUATING LANGUAGE-BIASED IMAGE CLASSIFICATION BASED ON SEMANTIC COMPOSITIONALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans show language-biased image recognition for a word-embedded image, known as picture-word interference. Such interference depends on hierarchical semantic categories and reflects that human language processing highly interacts with visual processing. Similar to humans, recent artificial models jointly trained on texts and images, e.g., OpenAI CLIP, show language-biased image classification. Exploring whether the bias leads to interferences similar to those observed in humans can contribute to understanding how much the model acquires hierarchical semantic representations from joint learning of language and vision. The present study introduces methodological tools from the cognitive science literature to assess the biases of artificial models. Specifically, we introduce a benchmark task to test whether words superimposed on images can distort the image classification across different category levels and, if it can, whether the perturbation is due to the shared semantic representation between language and vision. Our dataset is a set of word-embedded images and consists of a mixture of natural image datasets and hierarchical word labels with superordinate/basic category levels. Using this benchmark test, we evaluate the CLIP model. We show that presenting words distorts the image classification by the model across different category levels, but the effect does not depend on the semantic relationship between images and embedded words. This suggests that the semantic word representation in the CLIP visual processing is not shared with the image representation, although the word representation strongly dominates for word-embedded images.

1 INTRODUCTION

Language is not only a fundamental tool enabling humans to communicate with each other: it also shapes other aspects of perception and cognition, ranging from influencing visual processing to enabling abstract thinking (Vygotsky, 1934; Miller, 1951; Waxman & Markow, 1995; Colas et al., 2021). Analogous to this human ability, recent artificial intelligence technology has also utilized language to drive the training of visual skills in machines. For instance, the OpenAI team introduced CLIP (Contrastive Language–Image Pre-training), consisting of joint learning of language and vision (Radford et al., 2021). The CLIP model efficiently learns visual concepts from natural language supervision and can be applied to various visual tasks in a zero-shot manner. Joint learning of language and vision has also been utilized in the Visual Question Answering (VQA) literature, where artificial models are asked to answer questions about visual content (Antol et al., 2015; Lu et al., 2016; Fukui et al., 2016).

While joint training with language effectively fosters the acquisition of general abilities for visual recognition/understanding, its abstraction can produce biased recognition for either humans or machines. The Stroop effect known in human cognitive science literature is one of such biases (Stroop, 1935). When human participants observe a mismatched word, where the word "red" is printed with the "blue" color, and are asked to answer the "color" of the word, their response to the word can be slower than when they see the word "red" with the "red" color. In addition to the color-word interference, it has been shown that language interferes with human image category recognition, known as picture-word interference (Rosinski, 1977; Lupker, 1979). When a participant observes an image coupled with an incongruent category word (e.g., Figure 1, right), the image categorization speed can be delayed. Similar to the effect in humans, recent deep learning systems jointly trained on language and vision, e.g., CLIP, also show object recognition interference when words are embedded

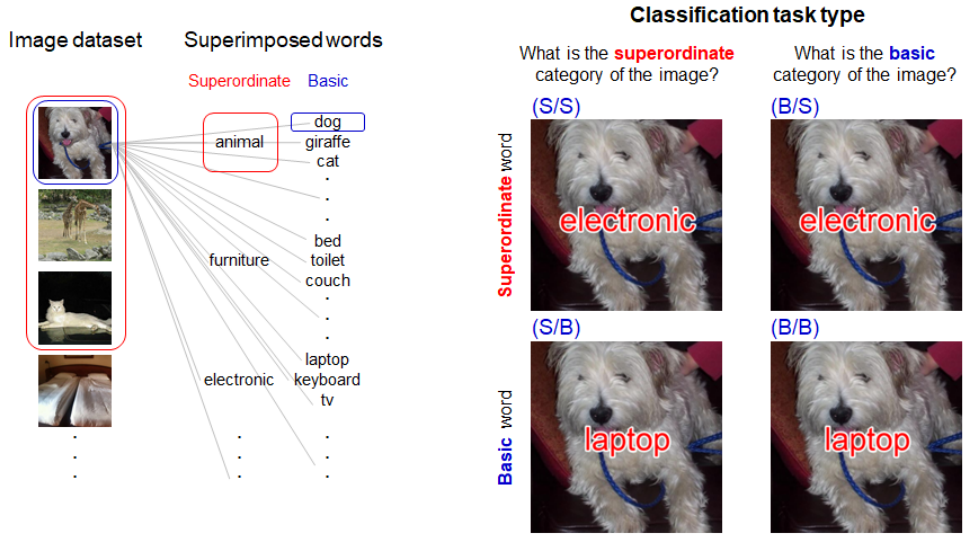

Figure 1: Overview of our benchmark test. Our dataset is a set of word-superimposed images consisting of the natural image datasets and hierarchical word labels. The classification tasks are divided into four conditions: superordinate image classification for the image with the superordinate category word (S/S, left top), superordinate image classification for the image with the basic word (S/B, left bottom), basic image classification for the image with the superordinate category word (B/S, right top), basic classification for the image with basic category word (B/B, right bottom).The classification examples by the CLIP model are shown in Appendix A1.

in images (Goh et al., 2021). Specifically, by simply adding a handwritten word to an image, the object recognition can be biased to the superimposed word.

A critical aspect of picture-word interference in humans is that interference depends on the semantic relationship between images and superimposed words. For instance, the reaction time on the image classification for word-superimposed images is slower when the image category is semantically similar to the word one (Rosinski, 1977). Based on the accumulated findings, previous works commonly suggest that at least two different processes mediate the effect in humans, as introduced in detail in Section 2.2. Specifically, when a participant observes a word-superimposed image, an *activation process* synthesizes semantic representations corresponding to the superimposed word and image (Figure 2, activation process). Based on the representations, it has been assumed that a *selection process* decides which possible activation is the answer for the current task.(Figure 2, selection process). They suggest that the activated representation is shared across visual word forms and images, and the dual activation by words and images confuses the decision in the selection process.

These human cognitive findings raise the question of whether interferences in artificial models have some similarities with those observed in humans. Indeed, some artificial models show language-biased image classification, but one possibility is that the bias might emerge irrespective of semantic interaction between words and images. When we consider a model with text and image encoders for picture-word interference (Figure 2), we can regard the functional role of the image encoder as synthesizing semantic representations for word-superimposed images. Although picture-word interference in humans depends on shared semantic representations of written words and images, this is not the only mechanism that could produce interference. For instance, the image encoder of the model might assign the written word ”dog” a different semantic representation than the visual image ”dog” and show strong preferential activation to written words while ignoring image contents. Since joint learning of language and vision is expected to enable machines to acquire highly interacted representations of texts and images to solve various general tasks, observing such independent representations and a bias for one of them could be problematic and show a limit of the joint training process.

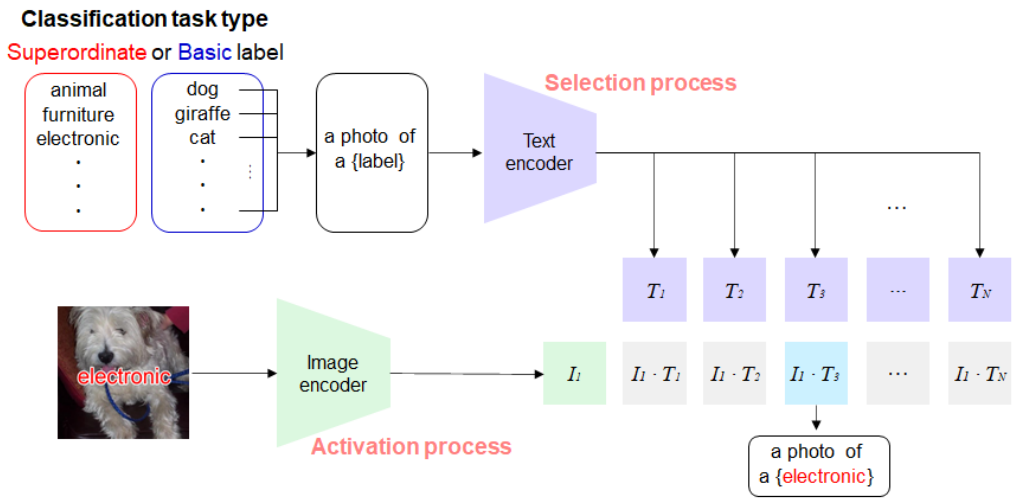

Figure 2: Process diagram of the CLIP model for our task. This figure is created based on Radford et al. (2021). The list of superordinate or basic word labels was used for the classification depending on the task type. We regard the image/text encoder as the selection/activation process in human picture-word interference as discussed in the main text.

In this study, we import methodological tools from the cognitive science literature to assess biases of artificial models in image classification for word-superimposed images. Specifically, we introduce a paradigm to evaluate picture-word interference, in which an agent has to predict the image superordinate/basic category for images with superordinate/basic words (Figure 1). Our benchmark dataset is a set of images with superimposed word and consists of a mixture of natural image datasets (Cichy et al., 2016; Mohsenzadeh et al., 2019) and hierarchical word labels (Lin et al., 2014; Krizhevsky et al., 2009). Our benchmark task aims to test artificial models with a text encoder and a image encoder (Figure 2). The task enables us to test 1) whether language-biased decisions (through the text-encoder) happen across different category levels and 2) the extent to which picture-word interference in artificial models depends on the semantic similarity between superimposed words and images. Furthermore, by importing a methodological tool from the neuroscience literature, called representational similarity analysis (RSA) (Kriegeskorte et al., 2008), we can also test 3) whether the image encoder has a common semantic representation for superimposed words and images. Out of various joint learning models of language and vision, the present study evaluated the CLIP model using the dataset because the model consists of dual processing of visual and text encoders and was developed to apply to various general visual tasks without fine-tuning to the dataset. To anticipate the result, we show that 1) presenting words disturbed the CLIP image classification even across different category levels, 2) the effect did not depend on the semantic relationship between images and words, and 3) the superimposed word representation in the CLIP image encoder is not shared with the image representation. The present study provides open-source code to reproduce the results and the full dataset.

2 Related Works

Joint learning of Language and vision in artificial models: Some recent works in machine learning have reported language-biased image recognition for artificial models trained with joint learning of language and vision. One example is the CLIP model (Radford et al., 2021). The CLIP architecture consists of an image encoder and a text encoder (Figure 2). The image encoder is based on a ResNet (He et al., 2016) or Vision Transformer (Dosovitskiy et al., 2020), and the text encoder is a Language Transformer model (Vaswani et al., 2017). The model is trained on a large dataset of image-text pairs with contrastive objectives, where it learns to align text and image representations for each pair. The pre-trained model can be used for various visual tasks in a zero-shot manner. For instance, when the model is applied for the image classification task, the learned text encoder synthesizes a zero-shot linear classifier for the descriptions, including classification labels, e.g., "a

photograph of [label]" (Figure 2). The learned image encoder outputs the representation for an input image, and the text classifier decides which label representation is the closest to the image representation. Language-biased image classification is reported in the CLIP image classification task as a typographic attack (Goh et al., 2021). As in Figure 1, when a written word is superimposed on an image, the CLIP misclassifies the image category as the superimposed word.

Another example of language-biased image classification is reported in the context of the Visual Question Answering (VQA) task (Antol et al., 2015; Lu et al., 2016; Fukui et al., 2016). This task goal is to answer questions about visual content, e.g., a text question "What color is the dog?" for a dog image. The task also needs joint learning of language and vision, i.e., an image encoder to recognize what the visual content is and a text encoder to understand the question. However, joint learning tends to be biased by superficial correlations between texts and images in the training data and ignore visual input to solve a task (Agrawal et al., 2016). For instance, even when a model sees a green (untypical) banana and is asked about its color, the model tends to answer "the color is yellow" while ignoring actual image content. Based on the observations, recent VQA models try to remove the biases (Agrawal et al., 2018; Ramakrishnan et al., 2018).

Picture-word interference in humans: It is essential to understand how the language-biased processing in artificial models is related to the human mechanisms discussed in cognitive science literature. This subsection reviews the functional connection between text/image encoders in artificial models and picture-word interference in humans (Figure 2). Picture-word interference refers to language-biased recognition for word-superimposed images, and previous works in cognitive science have investigated this effect using various experimental paradigms (reviewed by Bürki et al. (2020)). Here, we focus on interference when a participant observes a word-superimposed image and answers the image category for the image to compare the underlying cognitive mechanisms with artificial models (Rosinski, 1977).

The discussion about what kind of mechanisms mediate picture-word interference is still controversial in cognitive science literature, especially at which stage the interference emerges. However, we can summarize common functional mechanisms shared across different theoretical views and ask whether artificial models acquire these fundamental components in the current study. First, most views assume that visual words and images activate common semantic representations, i.e., a written word "dog" activates similar semantic representations to an image "dog." Second, they commonly assume that interference emerges due to some kinds of selection process. The type of selection process is extensively debated and different depending on the views. One is the lexical selection hypothesis, and it works to compare the semantic representation competitively (Levelt et al., 1999; Roelofs, 1992). This selection relies on the relative activation in semantic representations, and picture-word interference emerges because the activation by words induces misselection when it is semantically close to the one by images. However, it has also shown that the lexical selection is inconsistent with some behavioral findings, e.g., the activation amplitude does not always correspond to the effect size of interference. Another theoretical view assumes the response selection, which is closer to the decision stage than the lexical selection (Finkbeiner & Caramazza, 2006). Since the response selection assumes additional processing for relative activated representations, it can explain the phenomena inconsistent with the lexical selection assumption.

Our question is first to ask whether one observes patterns of interference at the behavioral level that are similar to humans and then to study more precisely the learned representations to see if mechanisms proposed to explain human interference could also be at play in the artificial model. To investigate machine mechanisms in terms of the cognitive science findings, we regard an image encoder in the artificial model as the processing to activate semantic representations from images (Figure 2). We also consider a text encoder as the selection process, which tries to find the answers according to the task objectives (superordinate/basic selection).

3 LANGUAGE-BIASED IMAGE CLASSIFICATION (L-BIC) TEST

Our dataset was a set of word-superimposed images. We used the two image datasets from the cognitive neuroscience literature on object recognition (Cichy et al., 2016; Mohsenzadeh et al., 2019). These datasets included 118 and 156 images with natural object categories, and previous works recorded the spatiotemporal neural activity using fMRI and MEG measurements for these images. The first dataset is annotated with 118 basic categories (Cichy et al., 2016) and the second

one is with five superordinate categories (faces, bodies, animals, objects, scenes) (Mohsenzadeh et al., 2019). For the word labels, we extracted the superordinate category words from the MS-COCO dataset and the basic category words from MS-COCO and CIFAR-100 datasets (Lin et al., 2014; Krizhevsky et al., 2009). The MS-COCO dataset had the "person" label for superordinate and basic categories. We used this label only for the superordinate category because the label was a higher category than other person-related basic labels in CIFAR-100. We removed multi-word labels and duplicates (including the plural of an already existing label) when compounding the MS-COCO and CIFAR-100 basic word label sets. The number of superordinate and basic word labels was 12 and 138, respectively. We superimposed all the labels on each image, and thus our word-superimposed image dataset includes 41,100 images in total. The color of superimposed words was red and bordered by white lines to enhance the word visibility considering future usage in human experiments. Since the image dataset is supposed to observe objects in the foveal vision, we put the words in the center of the image with a size that largely does not hide the original object.

The classification task was to answer the image category for word-superimposed images. The classification category level to answer was superordinate or basic. The lists of superordinate and basic categories were the same as those of words superimposed on images. Since the superimposed words were superordinate or basic, the stimulus conditions were divided into four, two classification types and two superimposed word categories (Figure 1). The former index S of "S/S" in Figure 1 indicates the classification task level (superordinate or basic), and the latter shows the word category level on the images (superordinate or basic). In addition to these conditions, we used no-word images to evaluate the original prediction by the model.

4 Experiments and results

4.1 Experimental setup

The image classification by the pre-trained CLIP model needs the input of a list of classification labels for the text encoder and an image for the image encoder (Figure 2). We used a sentence of "a photo of a [label]" for the text encoder. These two encoders output the representations of texts and images. We computed the similarities of text and image representations for a set of labels, passed them through the softmax function to get the probability of each label for a given input image, and extracted the label with the highest value as the classified one. (Radford et al., 2021). We show the results of the pre-trained CLIP image encoder using a Vision Transformer (ViT-B/32) in the main text and show the results of other image encoders in Appendix C.

4.2 Misclassification rates

To test whether presenting the word affects the classification, we fisrt evaluated the misclassification rate. We define the **label-switched image** as the word-superimposed image for which the word attachment changed the image classification for the no-word image. The misclassification rate refers to the ratio between the label-switched images and all images in each condition.

Table 1 shows the results of the misclassification rates for each condition. When the classification task type is consistent with the superimposed word category (i.e., the S/S and B/B conditions), the choices of the classification labels include the same word as that superimposed on the image. For these conditions, the CLIP model switched the original classification with high rates. These findings suggest that the model can match the visual word form with the classification label while detecting the form superimposed on the image and that the language-biased classification highly dominates the original visual prediction. These are consistent with the previous work reporting the language-biased image classification of the CLIP model (Goh et al., 2021).

In contrast, when the classification type is different from the superimposed word category (i.e., the S/B and B/S conditions), the classification word labels include the semantically related words with the superimposed one but not the same word. Even under these conditions, the CLIP model showed the label-switching for some of the word-superimposed images. The findings suggest that the model can match the meaning of the visual word form extracted from the image encoder with the label category from the text encoder across different levels. However, it is still unclear that the image encoder in the CLIP model acquires a common semantic representation for the images and words. There is also a possibility that the semantic word representation of superimposed words

is separated from the image representation in the CLIP image encoder (e.g., a superimposed word ”dog” is differently represented from an image ”dog”). To understand the semantic relationship, we evaluated the following semantic similarity between images and superimposed words for the label-switched and unswitched images.

	Superordinate prediction	Basic prediction
Superordinate word	(S/S) 92.77 %	(B/S) 41.24 %
Basic word	(S/B) 28.29 %	(B/B) 73.97 %

Table 1: CLIP’s misclassification rates for word-superimposed images.

4.3 Semantic similarity

We evaluated the semantic similarity between the original prediction labels for no-word images and the superimposed words to test the semantic dependency in picture-word interference. Specifically, we calculated the vector representations of these labels using a pre-trained Word2Vec model (Mikolov et al., 2013b;a) and then measured the cosine similarity between two vector representations. If picture-word interference in the CLIP depends on the semantic similarity between superimposed words and images, the similarity in the label-switched condition would be higher than the unswitched condition. Figure 3a shows the semantic similarity distribution for the label-switched and unswitched images in the CLIP model. The results show that for all task type and word category conditions, the semantic similarity distributions on the label-switched condition were similar to those on the unswitched condition. This finding suggests that picture-word interference in the CLIP model is irrespective of the semantic relationship between images and superimposed words, suggesting that the CLIP image encoder does not acquire a common semantic representation across superimposed words and images.

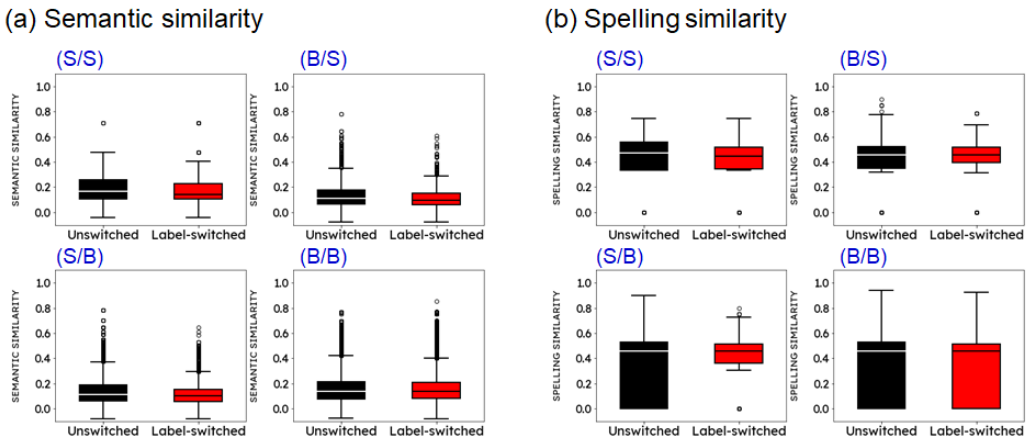

Figure 3: Results of the (a) semantic and (b) spelling similarity analyses.

4.4 Spelling similarity

In addition to the semantic similarity, we evaluated the spelling similarity using the Jaro-Winkler metric (Winkler, 1990) to test how the superimposed word is related to the image category. If the CLIP image encoder acquires a common representation of superimposed words and images based on the edit distance, this metric should change with the label-switched/unswitched condition. Figure 3b shows the result of the CLIP model on the spelling similarity for the label-switched and unswitched images. As in the semantic similarity, the distribution median for the task conditions was always similar between the label-switched and unswitched conditions.

4.5 PREDICTION PROBABILITY

The prediction probability analysis tells us how the label switching relies on the prediction confidence, i.e., the probability of the classified label in the label list, for no-word images and how the original prediction confidence decreases by presenting words. Figure 4 shows the results of the prediction probability analysis. Different panels show different stimulus conditions, as in Figure 1. In each panel of four, the prediction probabilities of the CLIP model for original images (no-word images) and word-superimposed images are shown. We separate the probabilities into label-switched/unswitched conditions for the original prediction and only show the probabilities on the label-switched images for word-superimposed images.

When we divided the original prediction probabilities into the label-switched and unswitched images, the probabilities for the label-switched images were lower than the unswitched images. The trend was notable when the task type condition was inconsistent with the superimposed word category (the B/S and S/B conditions). These results suggest that picture-word interference is more effective when the original prediction is not confident. In addition, by comparing the probabilities of the initial prediction label between original and word-superimposed images (two red plots in each panel of Figure 4), we can evaluate how the initial prediction confidence decreases. The results showed that the initial prediction probabilities drastically decreased when the CLIP model predicted the word-superimposed images for all stimulus conditions. However, it is crucial to be aware that a label prediction probability depends on other label predictions and that it can be lower if the other label prediction is high. We show how the original image representation is changed/unchanged by presenting words in the next section. Furthermore, the prediction probability analysis showed that the new prediction confidence for the label-switched images (the blue plot in Figure 4) was high when the classification task type was consistent with the superimposed word category (the S/S and B/B conditions). The trend is consistent with the results of the misclassification rate analysis.

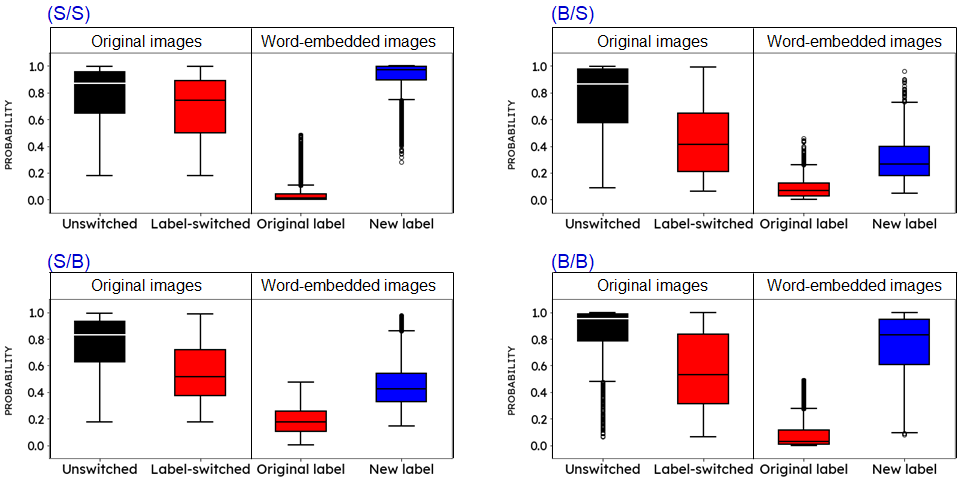

Figure 4: Results of the prediction probability analysis.

4.6 REPRESENTATIONAL SIMILARITY ANALYSIS (RSA)

We also utilized a methodological tool from cognitive neuroscience literature, called representational similarity analysis (RSA) (Kriegeskorte et al., 2008), to understand what is represented in the image encoder for word-superimposed images. RSA refers to image-by-image similarity assessment of intermediate representations in a brain or model. By computing representational dissimilarity matrices (RDMs) for no-word images and word-superimposed images, we can evaluate how the representation for the no-word images in the CLIP image encoder is affected by adding words on the image. We input one of two datasets (156 images) (Mohsenzadeh et al., 2019) for this analysis because it consisted of the superordinate categories defined by the previous work. In addition to the CLIP model, we evaluated the two feed-forward convolutional networks (i.e., VGG-19 and ResNet-

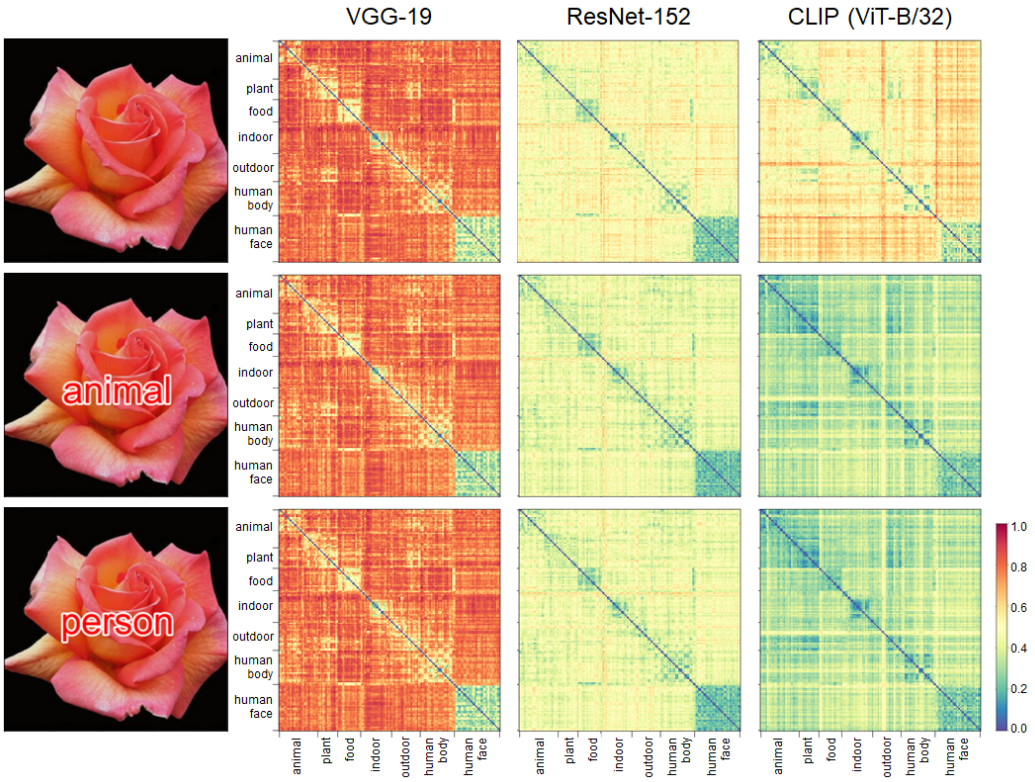

Figure 5: Results of the representational similarity analysis. See also Appendix B1.

152) trained for object category classification using ImageNet without joint learning of language and vision to confirm whether presenting words as image manipulation could affect the image encoder without joint learning.

Figure 5 shows the results of the RSA. Each matrix indicates the image-by-image cosine dissimilarity between intermediate features for the paired images. For the features, we used the last fully connected layer for VGG-19/ResNet-152 and the visual encoder output for the CLIP model. The top, middle, and bottom rows show the results for the original images, the word-superimposed images with the ”animal” label, and the word-superimposed images with the ”person” label, respectively. Each left image is one example in the 156 images. The results of other word-embedding are shown in Appendix B1.

For VGG-19 and ResNet-152, we observed some representational clusters according to the superordinate categories, i.e., the squared clusters around the diagonal line in each panel with lower dissimilarity values (Figure 5, top, VGG-19 and ResNet-152). The result is consistent with the previous finding that the higher-order processing of feed-forward convolutional networks is correlated with human and monkey brain activities on the cortical area related to object recognition (Yamins & DiCarlo, 2016). Similar to these feed-forward models, the CLIP image encoder also showed representational clusters depending on the superordinate categories (Figure 5, top, CLIP), suggesting that the encoder has a hierarchical structure for visual images.

When we analyzed the dissimilarity matrices for word-superimposed images using VGG-19 and ResNet-152, presenting words had little effect on the representational matrices of the original images (Figure 5, middle and bottom, VGG-19 and ResNet-152). This confirmation is natural as these networks are not trained using joint learning of language and vision, but essential because it has been known that a small amount of image manipulation can drastically change the visual representation of these models (Szegedy et al., 2013). This confirmation suggests that our manipulation does not have the power to disturb the model representation without joint learning of lanugage and vision.

In contrast, the results of the CLIP model for word-superimposed images showed different representations from that of original images (Figure 5, middle and bottom, CLIP). First, the dissimilarity values for word-superimposed images were generally lower than the value for original images (i.e., the representations of all images are similar to each other). The results are consistent with the strong language bias of the CLIP model. As in the other analyses, presenting words on the image shift the CLIP image classification to the direction of the superimposed word. This means that when we put a single word on the different images, their representations can be similar.

Most importantly, our analysis showed that we could observe the original representational clusters even when the superimposed words changed the representation for original images. Specifically, we observed similar category-related squares, to the original one, around the diagonal line of each representational matrix for word-superimposed images (Figure 5, middle and bottom, CLIP). In addition, even when we put different word labels, "person" and "animal," these representational matrices are similar. If the superimposed words were similarly represented in the image encoder to the image category, presenting words on the image should selectively affect the original image category representation and depend on what words were superimposed. The current results are inconsistent with the hypothesis and suggest that the CLIP image encoder has different representations of word forms (i.e., superimposed words) from visual image representations.

5 Discussion and Conclusion

The present study developed a benchmark test for language-supervised visual models to evaluate the language biases in image classification. The CLIP evaluation using the test showed that the misclassification probabilities increased even across different categories. The finding suggests that the CLIP image encoder can recognize the visual word form as a word, and the model can match the visual word representation with the classification label representations across different category levels. However, the semantic similarity analysis showed that the language-biased image classification in the CLIP model was irrespective of the semantic/spelling relationship between image and superimposed word categories. This finding is different from human findings and suggests that the CLIP image encoder does not acquire the semantic word representation shared with the image representation. The RSA results supported the hypothesis by showing that presenting words on images did not affect the no-word image representation selectively to specific categories.

We consider multiple steps for an artificial model with joint learning of language and vision to acquire semantic compositionality for visual word and image categories. First, an image encoder needs to recognize the visual word form as a word. Next, the semantic representation for words and images needs to be shared in the image encoder. That is, the written word "dog" needs to have the same semantic representation as the image "dog." The present study showed that the CLIP model acquires the first step, but the second step is still challenging. Our language-biased image classification test enables researchers to evaluate the step-by-step semantic interactions in joint learning of language and vision.

A remaining question is how the CLIP model acquires the biased representations in the image encoder. The presented results may be improved by further studying them as a function of potential biases in the training data of CLIP itself: for example, it would be useful to understand in the first place how CLIP acquired the ability to read superimposed words and why we observed a bias favoring the superimposed text over the whole image. More specifically, we can ask whether written words are in the training data and if they are, whether there are adversarial examples in the training data like our word-superimposed images.

Picture-word interference in humans is a phenomenon showing a biased error in the image classification, but this can also be considered as a developmental product humans acquire, where such bias could have functionalities. Some human studies show that picture-word interference is inextricably linked to facilitation effects of language priming for image category recognition (Finkbeiner & Caramazza, 2006). The finding reminds us that acquiring human-like biased representations for artificial models can have a functional meaning to process world information efficiently in natural environments. The current efforts to manage language-biased image recognition in artificial models are mainly aimed to remove the biases (Agrawal et al., 2018; Ramakrishnan et al., 2018). However, one direction might be to make the biases more human-like by controlling joint learning curricula of language and vision (c.f., Sigaud et al. (2021)).

6 REPRODUCIBILITY STATEMENT

All the source codes are available to create the word-superimposed images from the image datasets (Cichy et al., 2016; Mohsenzadeh et al., 2019) and to reproduce the current analysis results, which are shared in the supplementary materials of the ICLR submission. The analysis relies on the pre-trained models shared in the previous studies (e.g., CLIP). We describe how to import them for our analysis using a python notebook and include the description in the supplementary materials.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*, 2016.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Audrey Bürki, Shereen Elbuy, Sylvain Madec, and Shravan Vasishth. What did we learn from forty years of research on semantic interference? a bayesian meta-analysis. *Journal of Memory and Language*, 114:104125, 2020.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Similarity-based fusion of meg and fmri reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, 26(8):3563–3579, 2016.

Cédric Colas, Tristan Karch, Clément Moulin-Frier, and Pierre-Yves Oudeyer. Language as a cognitive tool: Dall-e, humans and vygotskian rl agents, March 2021. URL `https://hal.archives-ouvertes.fr/hal-03159786`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Matthew Finkbeiner and Alfonso Caramazza. Now you see it, now you don't: On turning semantic interference into facilitation in a stroop-like task. *Cortex*, 42(6):790–796, 2006.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030. https://distill.pub/2021/multimodal-neurons.

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

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Willem JM Levelt, Ardi Roelofs, and Antje S Meyer. A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(1):1–38, 1999.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29:289–297, 2016.

Stephen J Lupker. The semantic nature of response competition in the picture-word interference task. *Memory & Cognition*, 7(6):485–495, 1979.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013b.

George Armitage Miller. Language and communication. 1951.

Yalda Mohsenzadeh, Caitlin Mullin, Benjamin Lahner, Radoslaw Martin Cichy, and Aude Oliva. Reliability and generalizability of similarity-based fusion of meg and fmri data in human ventral and dorsal visual streams. *Vision*, 3(1):8, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL `https://arxiv.org/abs/2103.00020`.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. *arXiv preprint arXiv:1810.03649*, 2018.

Ardi Roelofs. A spreading-activation theory of lemma retrieval in speaking. *Cognition*, 42(1-3): 107–142, 1992.

Richard R Rosinski. Picture-word interference is semantically based. *Child Development*, pp. 643–647, 1977.

Olivier Sigaud, Hugo Caselles-Dupré, Cédric Colas, Ahmed Akakzia, Pierre-Yves Oudeyer, and Mohamed Chetouani. Towards teachable autonomous agents. *arXiv preprint arXiv:2105.11977*, 2021.

J Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643, 1935.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Lev S Vygotsky. *Thought and language*. MIT press, 1934.

Sandra R Waxman and Dana B Markow. Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive psychology*, 29(3):257–302, 1995.

William E Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. 1990.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

A PREDICTION EXAMPLE FOR WORD-SUPERIMPOSED IMAGES.

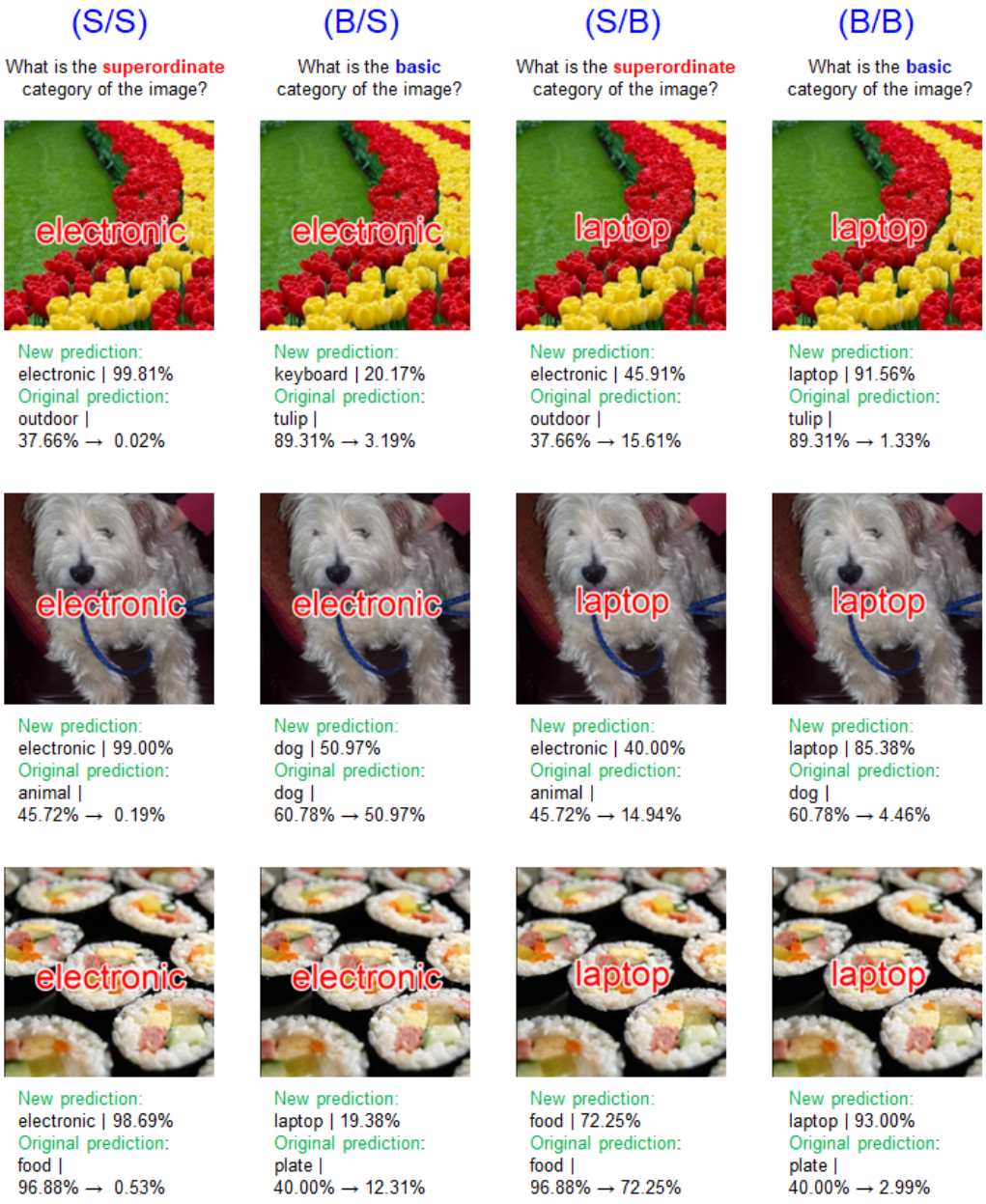

Figure A1: Prediction example for word-superimposed images using the CLIP model (ViT-B/32). For each stimulus condition (S/S, B/S, S/B, and B/B in Figure 1), the "new prediction" indicates the image classification probability for the word-superimposed image. The original prediction means the image classification probability for no-word images. We also show the extent to which the original prediction changed with the word attachment.

B REPRESENTATIONAL SIMILARITY ANALYSIS USING WORD-SUPERIMPOSED IMAGES.

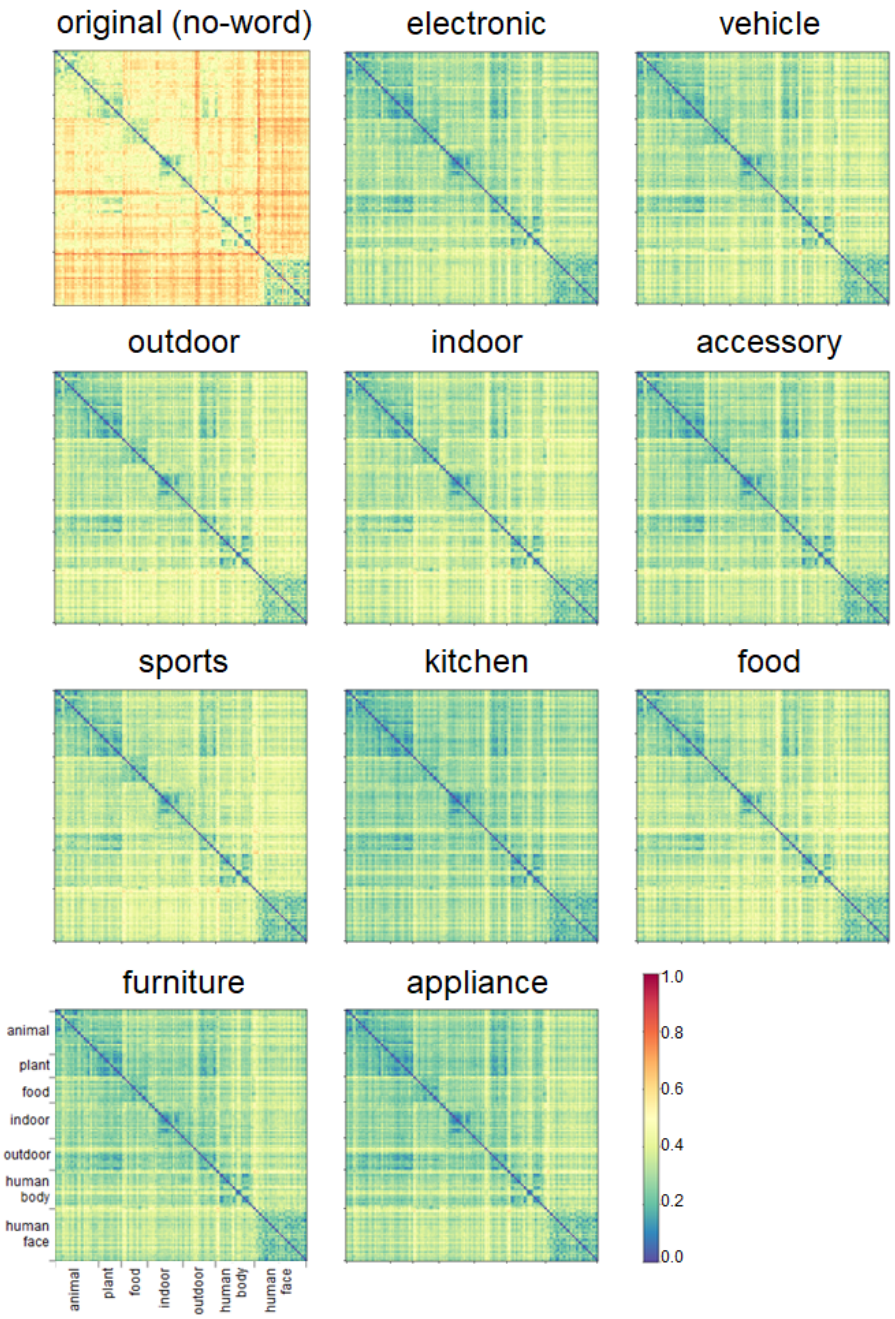

Figure B1: Results of the RDA. In addition to the labels "animal" and "animal" in Figure 5, we show the results of other words. Based on the CLIP image encoder (ViT-B/32) outputs, we calculated the image-by-image representation dissimilarity matrices. Similar to the results of the word labels "animal" and "person," presenting words on the images did not affect the no-word representation in a category-selective way. These results are consistent with the finding in the main text, suggesting that semantic word representations in the CLIP image encoder are different from image representations.

C EVALUATION OF DIFFERENT IMAGE ENCODERS OF THE CLIP MODEL

The image encoder of the CLIP pre-trained model is based on ResNets (He et al., 2016) or Vision Transformers (Vaswani et al., 2017). To explore the extent to which the language-biased image classification depends on the network architecture, we investigated the misclassification rate and the semantic relationship using various encoders in addition to a Vision Transformer model (ViT-B/32) used in the main text. Specifically, for the ResNets, we used a ResNet-50, a ResNet-101, and two models which follow EfficientNet-style model scaling (a ResNet 50x4 and a ResNet 50x16). For the Vision Transformer, we used a ViT-B/16 in addition to a ViT-B/32. We used the pre-trained CLIP networks using these architectures, which are shared by the previous work (Radford et al., 2021). The experimental procedure is the same as that of the main text.

Figure C1 shows the results of misclassification rates for the CLIP models using various image encoders. Even when the other architectures were used for the image encoder training, the results were consistent with the performance of the ViT-B/32 version. The misclassification rate was high when the classification type was consistent with the superimposed word level (S/S and B/B). In addition, the CLIP models showed significant misclassification across different category conditions under the other conditions (S/B and B/S).

The results of semantic similarity analysis and the RDA also did not depend on the architecture types (Figures C2 and C3, respectively). The semantic similarity analysis showed the median similarity of the label-switched images was not different from that of the unswitched images Figures C2 for all architectures. The RDA showed that the image encoder representation for no-word images changed with the architectures (Figures C3, first and fourth rows). However, the effect of attaching words did not change with the architecture. For all architectures, the overall dissimilarity across image-by-image pairs decreased for word-superimposed images, but the representation displacements were not selective to the original image representation.

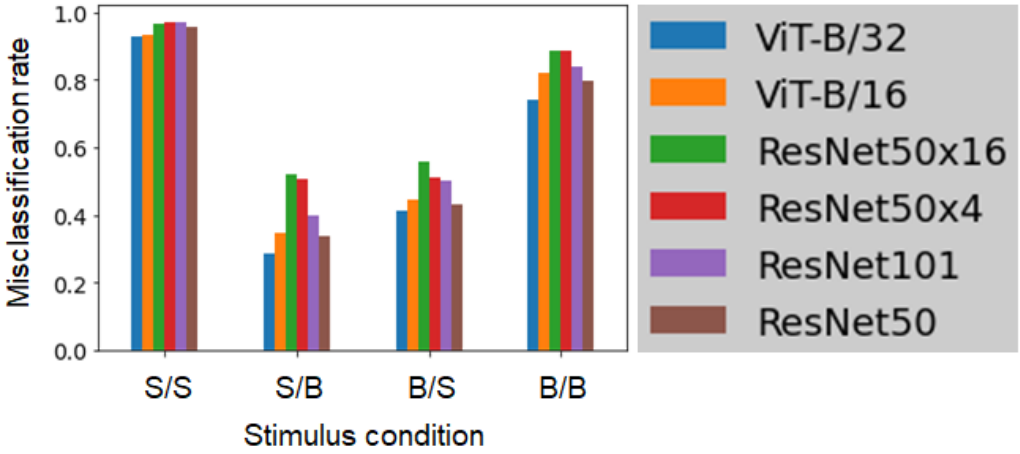

Figure C1: Results of misclassification rates for the CLIP models using various image encoders. The misclassification rate was plotted according to the stimulus condition (S/S, S/B, B/S, or B/B). Different color labels indicate different architectures.

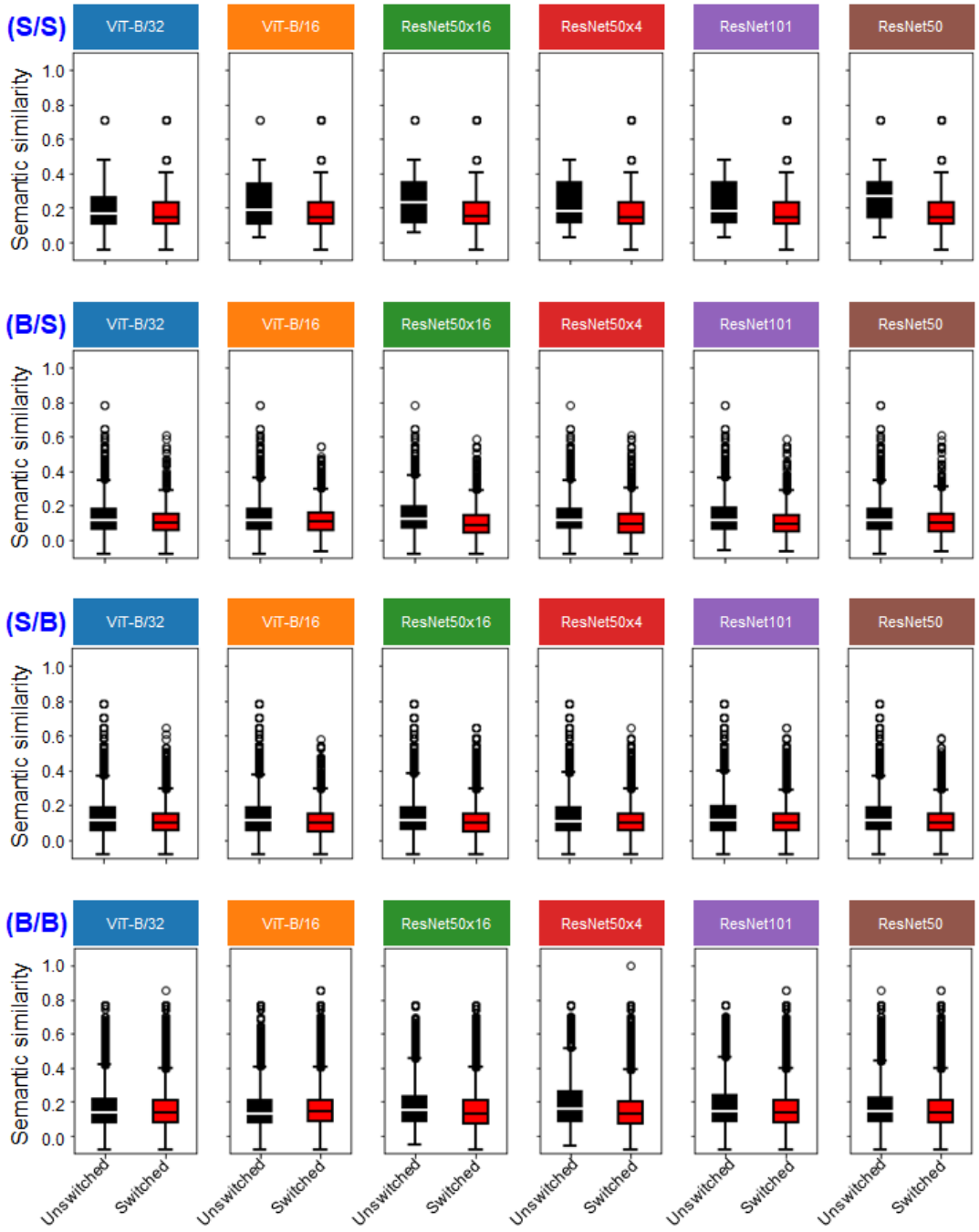

Figure C2: Results of semantic similarity analysis for the CLIP model using various image encoders. The semantic similarity analysis of Figure 3 was conducted for the CLIP models with the image encoder of a ViT-B/32, ViT-B/16, ResNet-50x16, ResNet-50x4, ResNet-101, and ResNet-59. The results of the ViT-B/32 image encoder were replotted from Figure 3. Different color labels indicate different architectures. Each row shows the stimulus condition (S/S, B/S, S/B, or B/B).

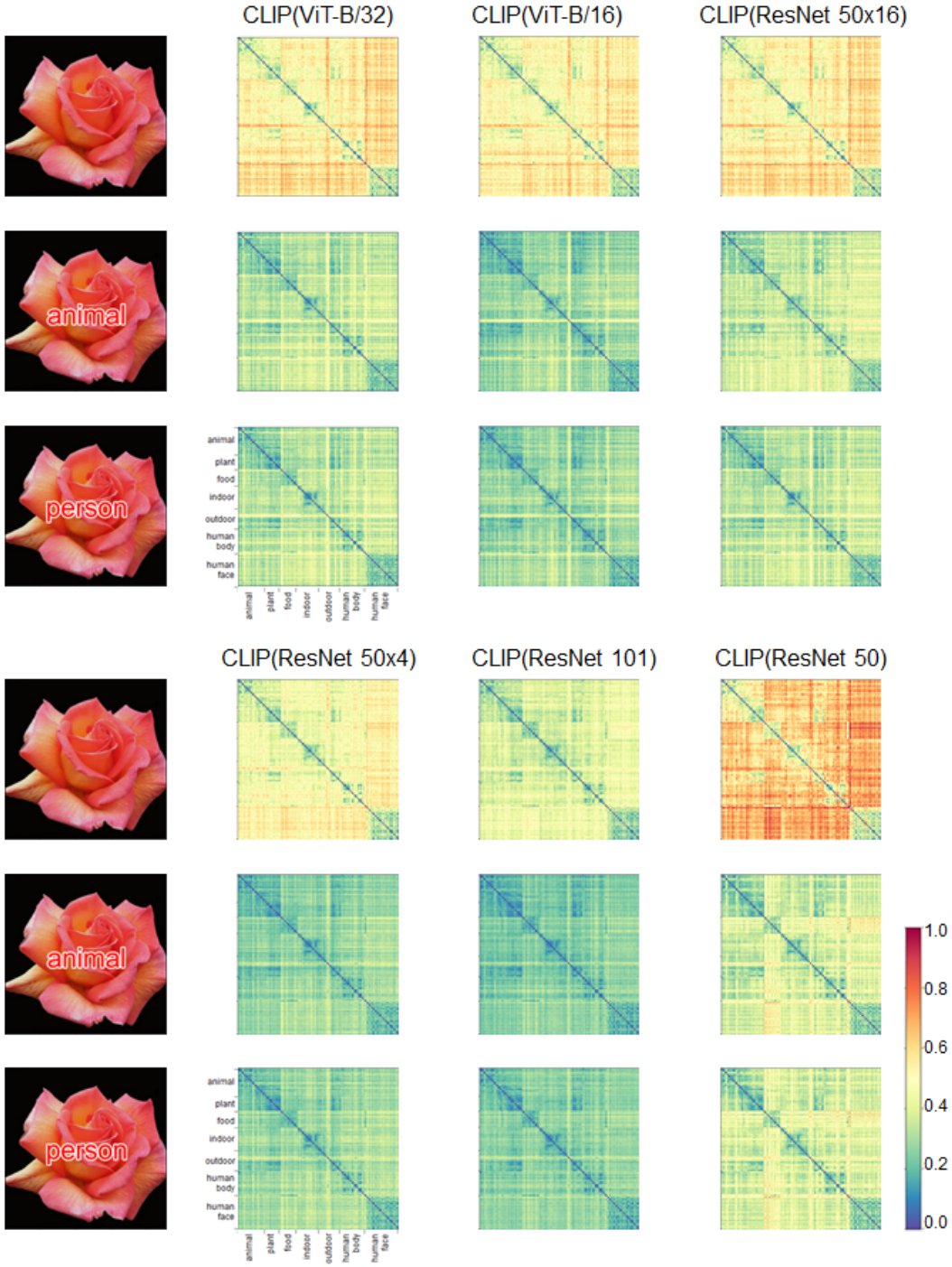

Figure C3: Results of the representational similarity analysis (RSA) for the CLIP model using various image encoders. The analysis method was the same as Figure 5, except that we also used the CLIP models with other image encoders than a ViT-B/32. The results of the ViT-B/32 image encoder were replotted from Figure 5.