Universiteit Utrecht

Artificial Intelligence master thesis

Do more 'humanlike' vision-language models perform better on grounding challenges? An attribution-based study on the VALSE image-caption alignment benchmark

First examiner:

Dr. Pablo Mosteiro Romero

Candidate:

Eduard Saakashvili (1066102)

Second examiner:

Dr. Albert Gatt

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in Artificial Intelligence in the Department of Information and Computing Sciences at Utrecht University.

March 14, 2024

Abstract

Vision-language models (VLMs) are increasingly successful, but questions remain about the extent and nature of their grounding in the visual modality. Many prior approaches to this question tend to focus on either performance-based measures of grounding (*what* can a model do?) or comparisons between a model's internal representations and a normative human baseline (is a model doing things in a *humanlike* way?). This study tests whether the results of each of these two approaches are correlated with one another in the context of a benchmark specifically designed to measure grounding. I design a human experimental environment to extract human saliency maps for a subset of the VALSE grounding benchmark. I also generate attribution maps for four VLMs for the same stimuli. My analysis creates a "humanlikeness" similarity metric for visual model attribution maps, and finds that model attribution maps are detectably "humanlike" on average. However, the degree of attribution humanlikeness does not correlate with model performance on the VALSE benchmark, either *between* or *within* models. The utility of this attribution-based humanlikeness metric as a complement to performance-based benchmarks remains unclear.

Acknowledgements

A big thank-you to my supervisor, Pablo Mosteiro Romero, for the deep engagement and incisive feedback throughout each stage of the thesis process, from initial formulation to the smallest details of the final text. Pablo and second reader Albert Gatt, whose feedback was also indispensable, were patient and flexible throughout the process as the project evolved. Our weekly meetings made the entire project feel truly collaborative. Fellow master's student Vincent Quantmeyer, who also attended these meetings, gave helpful questions and suggestions as well. All three also assisted materially in the initial stimuli filtering process.

Early in the thesis timeline, attending the Diverse Intelligences Summer Institute in St. Andrews, Scotland was a pivotal experience that helped motivate my research focus. Productive discussions with fellow attendees Skyler Wang and Nuhu Osman Attah made me realize the centrality of the concept of "humanlikeness" in my project and clarified the philosophical stakes of the study.

I am grateful to my wife, who gave crucial feedback on the intellectual underpinnings of the study. She, my parents, and other family members tolerated my mental absence in the final weeks of the writing process, and I thank them for their patience.

Finally, I want to thank Ohøj Coffee Roasting, the best coffee outlet in Utrecht, for fueling the week-long manic state that got this project across the finish line.

Contents

1. Introduction

Recent advances in AI have included the development of vision-language models or VLMs. Though architecturally similar to their large language model (LLM) cousins, their advantage is that as multimodal models, they are able to interpret multiple modalities and, in some cases, create joint representations. Though the performance of these and related model architectures continues to reach once-unattainable heights, this has only intensified long-running debates about the opaque internal worlds of these "exotic, mind-like entities" (Shanahan 2023, p. 11). In an echo of Senator Howard Baker's famous Watergate line, AI researchers keep asking, in various formulations: *What* does the model know and *how* does it know it?[1]

A prominent line of inquiry focuses on whether models' representation of the world is *grounded*, and grounded *correctly*, in the physical environment. (Grounding is a theoretically specific concept that I will discuss later.) Many approaches to answering these questions rely on designing challenges that would presumably require a model to be grounded to succeed at them. Such challenges focus on *what* a model can do as a measure of whether it is grounded. An example of such a challenge set is VALSE (Parcalabescu, Cafagna, et al. 2022), which we will explore in more detail later. This benchmark includes images such as those shown in Figure 1.1, each paired with a correct and incorrect caption (called the foil). The results for four models on whether they could pick the correct caption are given in Figure 1.1; the idea is that the more of these challenges each model gets right, the more grounded a model is by this or that criterion.

But there is another way to investigate whether a model is seeing the world in a desirable way: that is to ask not *what* a model can do, but *how*

[1] "What did the president know, and when did he know it?", repeatedly asked during the 1973-74 Watergate hearings that terminated the Nixon presidency (Cass 2014).

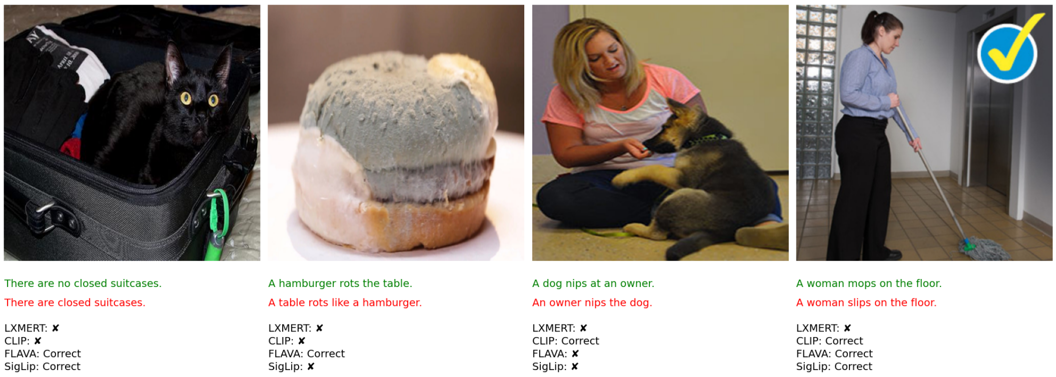


Figure 1.1: Four randomly-selected images and captions/foils from the VALSE benchmark, along with information on whether each of four vision-language models was able to choose the correct caption. (The images have been resized to square to display here; the typo "hamburger rots the table" occurs in the original dataset.)

the model does it. Explainable AI (XAI) tools are one way to explore this aspect of model behavior. Importantly, XAI can tell us whether models are succeeding at tasks in a way that is humanlike or, alternatively, *non*-humanlike and even incomprehensible to human observers. To those who see human ways of thinking as the goal, a humanlike model is "right for the right reasons" (Selvaraju et al. 2019). Different approaches to model evaluation reflect different normative goals: some frame the goal of AI models as performance by any means; others see AI as a way of emulating specifically humanlike ways of thinking.

When models *are* more humanlike in what they pay attention to, are they also *better* at grounding challenges? To explore this question, I design and implement an experimental study on a specific dataset and a number of VLMs that asks whether a *model's attribution maps'* similarity to *human saliency maps* predicts the model's performance on grounding challenges.

How do we get this information about models and people? For AI models, information about what the model finds important in an image is generated with model-agnostic visual *attribution* methods. Humans' importance distribution over each image will be called *human saliency maps*, and they will be gleaned from human subjects in an online experimental environment.

While not exhaustive, this study's results can help set the agenda for future research about the relationship between humanlikeness and performance in AI.

1.1 Key terms and variables

- **Vision-Language Model** (VLM): An AI model that incorporates both text and image inputs.
- **Stimulus** (s): One of the 99 data points (image, caption, foil) selected from the VALSE dataset for use in this study.
- **Validated Subset of VALSE**: Parts of the VALSE dataset that are considered validated by the validation criteria of the original VALSE authors.
- **Model** (m): Each model m in the study is an implementation of LXMERT, CLIP, FLAVA, or SigLip, all of which are VLMs.
- **Model Output Score** ($f_m(s)$): I define and implement this single scalar output for each model m where it takes stimulus s as input. Positive outputs indicate the model has correctly responded to a given stimulus. Also referred to as the **prediction difference**, it is typically the model's score for the correct caption minus the model's score for the incorrect caption.
- **Human Saliency Map** ($H_{s,\text{down}}$): The aggregate human saliency map for each stimulus s, a downsampled (to a 4x4 matrix) average of importance scores given to each image pixel by human subjects in the online experimental environment.
- **SHAP/Shapley Attribution Map** ($\Phi'_{m,s}$): A 4x4 matrix representing the importance score assigned to each image region in stimulus s by the SHAP method, for model m. The prime symbol ($'$) denotes that it has been normalized and set to all-positive values before being used in the analysis. The term "attribution map" is used rather than "attention map" to avoid confusion with other concepts in machine learning.

- **Humanlikeness Score** ($\mathrm{RC}_{m,s}$): This metric, also referred to as the **human-SHAP similarity metric**, represents the similarity (rank correlation) between the human saliency map $H_{s,\mathrm{down}}$ and SHAP attribution map $\Phi'_{m,s}$ for a given stimulus s and model m.

2. Background and motivation

This chapter will provide the background that motivates this line of inquiry, as well as the related work and considerations that inform the methodology. Finally, it will formulate the research questions of the project in more detail.

The first section, 2.1 will give background on the current state of VL models. Next, section 2.2 will frame the AI grounding debate's stakes and discuss what it means to ask whether a VLM is grounded.

Next, section 2.3 will discuss two ways of measuring the grounding of models: performance-based (section 2.3.1) and interpretability-based (section 2.3.2). For interpretability-based evaluation, models' comparison to ground-truth saliency maps is also discussed, as well as the usefulness of this method for the present study. Section 2.3.2.2 explores the pragmatic side of generating both XAI attribution maps and human saliency maps, outlining some of the ways prior researchers have generated and used such maps. This includes a discussion of studies that extract saliency maps directly from human subjects, and the usefulness of this method. On the topic of human saliency maps, section 2.3.2.4 clarifies the relationship between model grounding and humanlike attribution maps, asking: are humanlike attribution maps a good measure of model grounding in the first place?

Finally, having outlined both the overall conversation around VLM grounding and the motivations for focusing on XAI and human saliency maps as a way of exploring this question, we proceed to the final research questions in section 2.4.

2.1 Current state of VLMs

New neural architectures have enabled an ongoing "boom of Transformer-based universal multimodal encoders pretrained on several multimodal tasks"

(Bernardi and Pezzelle 2021, p. 7). Such models are categorizable into single- and two-stream variants which, respectively, treat the visual and language input as either two concatenated pieces of data or first create separate representations of each medium which are then often combined into a shared representation (Bugliarello, Cotterell, et al. 2021).

Though VLMs tend to achieve impressive performance compared with earlier computer-vision algorithms, there remains significant room for improvement. As recent study of VLMs (Bugliarello, Sartran, et al. 2023, p. 9) concludes:

> While recent pretrained VLMs achieve impressive performance on various downstream benchmarks (such as visual question answering and image retrieval), recent benchmarks have highlighted that they still **struggle with tasks that require fine-grained understanding**—where a model needs to **correctly align various aspects of an image to their corresponding language entities**. [emphasis added]

This assessment, based on testing several noteworthy recent VLMs on four benchmarks (including the VALSE benchmark used in the present study), reflects a key challenge in producing multimodal intelligence: the challenge of *grounding*.

2.2 What grounding means for VLMs

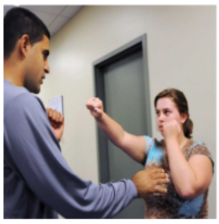
Current discussions of grounding originate from the "symbol grounding problem" that emerged in philosophy of mind discussions in the 20th century: the contention that symbols acquire true semantic meaning only in relation to real-world objects. This argument has long bedeviled AI researchers, who have in various forms had to confront the question: "How is symbol meaning to be grounded in something other than just more meaningless symbols?" (Harnad 1990, p. 340). Searle 1980 seminally explored this question in his "Chinese Room" thought experiment describing a system of rules (i.e. a computer program) that a human being could execute to produce speech in Chinese without themselves understanding the language. Searle generalizes this state to computers; he proposes that “[t]he fact that the programmer and the interpreter of the computer output use the symbols to stand for objects in the world is totally beyond the scope of the computer. The computer, to repeat, has a syntax but no semantics” (Searle 1980, p. 423).

The arrival of connectionist architectures (of which modern transformers are a subset) seemed initially promising, considering such models have a more “bottom-up” relationship with input data. Even so, philosophers of mind continued to challenge the meaningfulness and groundedness of connectionist models’ internal representations (Christiansen and Chater 1993).

In contemporary AI research, the grounding problem has frequently taken on a less philosophical and more empirical flavor, including in research on the grounding of multimodal models. In technical research, the question is often less whether the internal representations of the VLM have any unambiguous semantic meaning in theory. Rather, the focus lies on ability: do models evince grounding in the success with which they are able to handle specific challenging inputs? As Bernardi and Pezzelle 2021 write: “Answering a question that is grounded in an image is a crucial ability that requires understanding the question, the visual context, and their interaction at many linguistic levels” (p. 1).

While their discussion focuses on visual question-answering (VQA), it is broadly applicable to VLMs which are tasked with simultaneously attending to visual and textual information. For this study, we can frame the symbol grounding problem in that more limited way: the symbols in the textual data acquire meaning in relation to elements of the visual data. If this relationship is not correctly understood by the model, that indicates a lack of grounding. To make it more specific, take the example from VALSE in Figure 2.1. To solve challenges like these reliably, the model should be able to identify the action in the photo and the female subject as the actor, while also understanding on some level that the caption has the correct subject-verb relationship for the scene.

But we should pause on the word “understand”. The extent to which



caption: A woman punches a man.
foil: A man punches a woman.

Figure 2.1: An image drawn from the VALSE dataset, with caption and foil.

current models can "understand" something at all remains a subject of increasingly active discussion. One recent paper (Mitchell and Krakauer 2023, p. 1) notes that,

> [u]ntil quite recently there was general agreement in the AI research community about machine understanding: while AI systems exhibit seemingly intelligent behavior in many specific tasks, they do not understand the data they process in the way humans do.

Note how this quote holds "the way humans do" as a kind of gold standard, as opposed to "seemingly intelligent" machines. More recently, the anthropocentric consensus on intelligence has been challenged, including in the popular press (Johnson and Iziev 2022). Even so, Shanahan cautions against applying "understand" and other anthropomorphic terms to models. Instead, he argues that a model does not "know" or "believe" a piece of information; rather, it "contains" it, like an encyclopedia (Shanahan 2023, p. 5).

Sidestepping the ambiguous word "understand", Yuksekgonul et al. 2022, p. 2 uses the verb "represent" to describe what a grounded model must be able to do to the relation between text and image:

> Whereas humans effortlessly parse natural scenes containing rich objects in relation to one another, it is unclear whether machines understand the complexity of these scenes. To do so, models must be able to correctly represent objects, their attributes, and the relations between objects.

We need not endorse the anthropomorphic "understand" but can stick with the more neutral “represent”. We are concerned with the question of whether models represent the relationship between image and text in a multimodal setting in a way that is desirable. But how does one measure grounded representation empirically? And what role does the concept of "humanlikeness" play in the measurement of grounding?

2.3 Quantitative evaluation of model grounding

There are two main ways of quantitatively testing whether a model is grounded, and the present study engages with both. This section will describe various approaches to measuring grounding in the literature.

One approach is a more performance-focused understanding (section 2.3.1). Performance-based challenges can, for instance, be drawn from the VALSE benchmark, which was specifically designed to measure grounding in the visual modality (Parcalabescu, Cafagna, et al. 2022).

The second approach (section 2.3.2) focuses on *how* a model accomplishes its tasks, rather than *whether* it does. In the visual modality, this is often done by comparing what the model looks at in an image to what humans look at., i.e. taking humanlikeness as a measure of grounding.

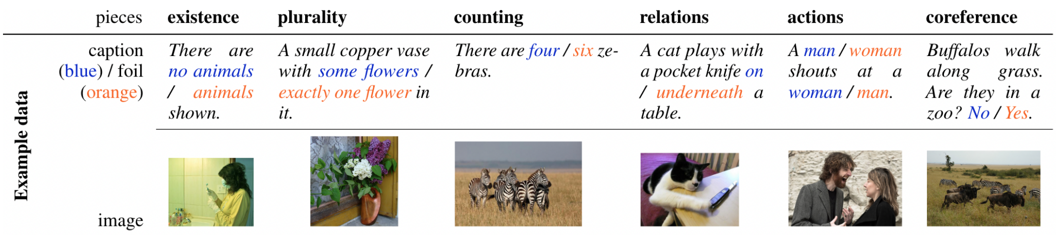
2.3.1 Performance-based evaluation of VLM grounding

A dominant approach to probing the groundedness of current VLMs in the visual modality is the development of benchmarks which specifically target one or several types of grounding (Bugliarello, Sartran, et al. 2023). The concern these benchmarks address is that without specifically designing challenges that require grounding, models are “prone to exploiting shortcut strategies” (Yuksekgonul et al. 2022, p. 6). Such shortcut strategies are those that use “spurious” correlations in the large datasets which are difficult to detect because “[t]ypically such correlations are not apparent to humans performing the same tasks” (Mitchell and Krakauer 2023, pp. 3–4). The benchmarks are designed to make it more difficult for models to solve challenges using such shortcuts.

Benchmarks tend to focus on one or several parts of speech or linguistic phenomena, and include challenges that test a model's ability to use this part of speech or concept. Major benchmarks include **SVO-Probes**, which tests VLMs on verb understanding by making models distinguish correct and incorrect images which can differ by verb, subjects, or objects (Hendricks and Nematzadeh 2021). Another benchmark, **VSR** (F. Liu, Emerson, and Collier 2023), focuses on spatial reasoning; it consists of a series of images and candidate captions for each image, such as "The hair dryer is facing away from the person"—a VLM is then tasked with determining whether or not each caption is correct, and the authors shows performance for the VLMs they tested significantly lags behind humans. The **Winoground** benchmark (Thrush et al. 2022) contains pairs of images with corresponding pairs of captions; both captions contain the same words, but in a different order, and models are tasked with matching captions with images within each pair of pairs; the researchers write: "To perform well on Winoground, models must not only encode text and images well [...], but they also must be able to synthesize information across the two modalities" (Thrush et al. 2022, p. 1)

A benchmark which specifically probes the visual grounding of VLMs' linguistic competence, is the aforementioned **VALSE** (Vision And Language Structured Evaluation) benchmark introduced in Parcalabescu, Cafagna, et al. 2022. VALSE's focus on testing the grounding of linguistic competence in the visual modality in a variety of linguistic phenomena makes it well-suited to exploring the questions raised in this paper, and the VALSE dataset is the basis of the methodology developed here. Another reason for choosing VALSE is the variety of types of challenges compared to other benchmarks (Bugliarello, Sartran, et al. 2023), which affords more flexibility in data selection throughout the methodology development process.

VALSE comprises a suite of tests which target the visio-linguistic grounding of pre-trained VLMs, for six distinct linguistic phenomena. As noted in the VALSE paper, "Since most V&L models are pretrained on some version of the image-text alignment task, it is possible to test their ability to distinguish correct from foiled captions (in relation to an image) in a zeroshot



	pieces	existence	plurality	counting	relations	actions	coreference
Example data	caption (blue) / foil (orange)	*There are no animals / animals shown.*	*A small copper vase with some flowers / exactly one flower in it.*	*There are four / six zebras.*	*A cat plays with a pocket knife on / underneath a table.*	*A man / woman shouts at a woman / man.*	*Buffalos walk along grass. Are they in a zoo? No / Yes.*
	image						

Figure 2.2: This image adapted from the VALSE paper depicts the 6 linguistic phenomena used, with an example each (Parcalabescu, Cafagna, et al. 2022).

setting. The construction of foils can serve many investigation purposes. With VALSE, we target the linguistic grounding capabilities of V&L models" (Parcalabescu, Cafagna, et al. 2022, p. 2). This approach uses a previously proposed foiling technique (Shekhar et al. 2017), providing the model with a correct or incorrect caption for the same image, and asking it to provide a likelihood score for each caption-image pair. Several techniques are used in generating the VALSE dataset to minimize the possibility of spurious associations and confounds that could enable the model to pick the correct caption without grounding its decision in the image. Figure 2.2 shows instances of each linguistic phenomenon in the VALSE dataset.

The more reliably a model picks the caption over the foil, the better the model is assumed to be at grounding its interpretation of linguistic structures in the visual modality. For instance, choosing the correct caption between "A woman shouts at a man" vs. "a man shouts at a woman" would presumably require a correct representation of subject-object relations in the visual modality; merely identifying the presence of "man", "woman", and "shouting" in the image is not enough to reliably solve this challenge (Parcalabescu, Cafagna, et al. 2022, p. 3).

2.3.2 XAI-based evaluation of VLM grounding

Though performance on fine-grained benchmarks is one way to probe the groundedness of models, there are metrics other than performance which can shed light on whether models represent or encode the desired relationship between text and image. One way has involved explainable AI (XAI) techniques, which create a representation of *how* or *why* a model arrived at

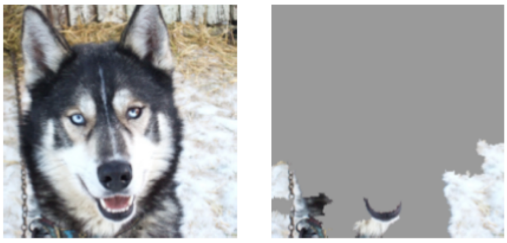


Figure 2.3: A figure from Ribeiro, S. Singh, and Guestrin 2016 uses LIME to show that a wolf vs. husky classifier is incorrectly using snow in the background as a major reason for its decision – a spurious correlation deliberately inserted into the training data by the researchers.

a certain output. This section will summarize different directions in this space and go into detail on *visual* XAI methods and their role in grounding research.

In many cases, XAI-based comparisons between models and humans come with a normative assumption that *humanlike* ways of solving problems are preferable. An early and well-known example is a 2016 paper which used the LIME explainability method to generate visual evidence that a model in the study was using "spurious correlations" in the training data to classify photos as "wolf' or "husky", relying on the presence of snow in the background rather than features of the animal (Ribeiro, S. Singh, and Guestrin 2016). Implicit in this approach is the assumption that some models not only get tasks wrong, but can do a task in an *undesirable* way by, in this instance, focusing on apparently irrelevant parts of an image. See Figure 2.3 for that paper's illustration of how XAI exposed this incorrect approach. In this example, the model misclassifies the husky as a wolf, and commits the dual sin of being both wrong *and* wrong for the 'wrong reasons'. But even if the image had shown a wolf, the model would have been *right for the 'wrong reasons'*, because its decision was caused by an undesirable part of the image—at least if we follow a human-understanding-based concept of model grounding.

On the other hand, XAI can also show that models *do* represent information in a way researchers deem desirable. Some researchers have used localization-based interpretability techniques to find evidence that large language models seem to represent core syntactic structures and other linguistic constructs in specific parts of their architecture (Tenney, D. Das, and Pavlick 2019; Htut et al. 2019). In one case, this led to the conclusion that “[models] are representing language in a satisfying way” (which is itself a human-normative assessment) and are learning to represent “syntactic and semantic abstractions” (Tenney, D. Das, and Pavlick 2019, pp. 4593–97). XAI tools can also be used to assess the overall way a model processes its inputs, like whether multimodal models like VLMs are actually attending to both modalities at all, and to what extent (Hessel and L. Lee 2020; Parcalabescu and Frank 2023). This last approach is helpful for checking whether multimodal (vision + language) models are actually relying on both modalities, or suffering from "unimodal collapse" (an issue explored in Yuksekgonul et al. 2022).

On a more granular level, Cao et al. 2020’s paper "Behind the Scene" uses probing techniques to determine to what extent visual and linguistic knowledge is encoded inside the attention layers of several VLMs. This approach is notable for combining model dissection (an XAI technique) with performance-based measures, using a benchmark to assess the usefulness of latent knowledge in the model’s internal representations. Thus, this last approach could be seen as a hybrid, XAI + performance approach to evaluating grounding.

In the visual modality, XAI can be combined with ground-truth attribution maps to evaluate grounding. This is explored in the next section.

2.3.2.1 The use of ground-truth maps + XAI to measure grounding

A notable way XAI has been used in grounding-related research is by comparing a model’s XAI explanations to “ground-truth” explanations, such as a map of human saliency over an image (i.e. a numeric representation of which parts of the image humans look at, and the relative important of each pixel or region). An implicit assumption behind such measurement is that a model whose attention pattern is more *humanlike* is likely to be a better model, or at least more desirable by a specific criterion.

Ground-truth explanations are sometimes used not just to evaluate models, but to improve them. Human-generated explanations have been used to validate XAI techniques themselves (Rao et al. 2021; Park et al. 2018), as well as to better train models by forcing their attention or attribution patterns to be more humanlike (Selvaraju et al. 2019; Sood et al. 2023). Such work has a longer history in the machine-learning research community; for instance, Donahue and Grauman draw on the NLP-derived method of "annotator rationales" to force a non-neural model architecture to more closely align with what humans report as important areas of an image (Donahue and Grauman 2011).

Other studies have used XAI to probe to what extent object detection models are looking at the correct annotated regions of the image input (Y. Liu and Tuytelaars 2020; Xu et al. 2020). This last type of paper uses ground-truth saliency not for performance enhancement, but to answer a question about a model: is the model interpreting the visual modality correctly?

For my project, the most relevant subset of prior XAI studies are those papers that directly use ground-truth annotations of images to probe whether a *multimodal* VL model is correctly attending to the relationship between text and image in a task.

C. Liu et al. 2016, for instance, use ground-truth saliency maps to evaluate the "correctness" of the attention of an image-captioning model, using object locations as the basis for the ground truth. The authors define "correctness" as "the consistency between the attention maps generated by the model and the corresponding region that the words/phrases describe in the image" (C. Liu et al. 2016, p. 1). Wu and Mooney take a similar approach, mathematically comparing human-annotated maps to explanations generated by a model by vectorizing each explanation and taking a cosine similarity (Wu and Mooney 2019).

A highly relevant paper which directly uses human subject-derived ground-truth maps to evaluate a model is "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?" A. Das, Agrawal, et al. 2017. In this study, the researchers use an online survey to extract saliency maps from human subjects for visual question-answering (VQA) tasks; I take this study as the primary inspiration for the present study's human data collection. The authors compare these human saliency maps to a machine-produced attribution map for each question-image pair, which is similar to my study's ultimate approach. The authors find that both models they studied produce "attention maps [that] are positively correlated with human attention maps" (A. Das, Agrawal, et al. 2017, p. 7).

A. Das, Agrawal, et al. 2017 compare their human saliency maps to machine attribution maps using mean rank-correlation coefficients across examples for a given model. Other metrics that can be used to quantify similarity are EMD (earth-mover's distance) and IoU (intersection over union), among others (Rodis et al. 2023, pp. 18–19). Saliency/attribution map comparison metrics will be discussed further in chapter 3.

One ambiguity that can emerge in the XAI research space is what ground-truth comparison even *tells* us about machine-produced attribution maps. In some cases, researchers use such comparisons as a way of evaluating not the quality of the model, but the quality of the XAI method itself (see Rodis et al. 2023, pp. 19–20 for examples). In one study, we see a hybrid approach where the XAI method's quality is evaluated against human explanations, while also confirming the fact that the XAI method's output is still generally faithful to the model's actual attention (Wu and Mooney 2019). In this study's case, I take attribution maps output by the chosen XAI method (SHAP) as reflective of the importance assigned by the model to various parts of a given image. I use the comparison to human saliency maps to draw conclusions about the humanlikeness of a model's relationship with the visual modality.

2.3.2.2 How other studies generate human saliency maps

How do prior studies create the human saliency maps they use as ground-truth? There are those which use researcher-annotated objective ground truth like bounding boxes of objects for object detection, where the annotation is straightforward and the choices seemingly obvious. However, even for object detection there are more subtle cases; for instance, it may be reasonable to use tropical flora in the background to help identify a species of bird. In such cases, the ground truth of "correct" human saliency is less obvious, and the non-trivial nature of generating ground-truth maps becomes clearer. A recent survey on multimodal explainability identifies the generation of ground-truth saliency maps as a much-needed area of further research (Rodis et al. 2023, p. 20), which the present study contributes to.

Many previous approaches to generating such saliency maps use eye-tracking technology (e.g. Yang et al. 2022), using human eye-tracking data for a given image as the ground-truth attention distribution; since this is not feasible for the present study, this overview focuses on cheaper and less labor-intensive methods.

In one approach, Park et al. 2018 use object segmentation and then human input as the basis for saliency maps. The researchers pre-segment images into objects and ask human subjects to report the most relevant segments of an image for a given question-answer pair (the correct answer is provided). This self-reported human importance is then aggregated into saliency maps over each image.

In Jiang et al. 2015, by contrast, the experimental environment lets subjects explore images *without* a specific task. This paper's innovation lies in providing a robust alternative to eye-tracking technology that still does not rely entirely on self-reporting by human subjects. The researchers create an interface that initially blurs the image, and lets users move their mouse to unblur the parts they find salient, a way of presenting the image that is inspired by how the human gaze glances over a visual scene region by region. An example of this process (from the original paper) is shown in Figure 2.4. The mouse movement pattern is recorded for each subject, and used to approximate the task-neutral ("task-free") saliency of various parts of each image.

The aforementioned paper A. Das, Agrawal, et al. 2017 takes Jiang et al. 2015's methodology as a starting point, creating a "game-like" inter-

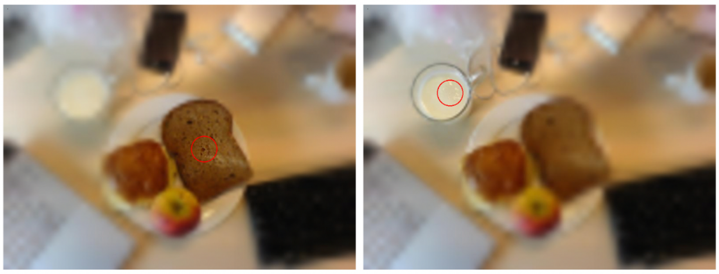
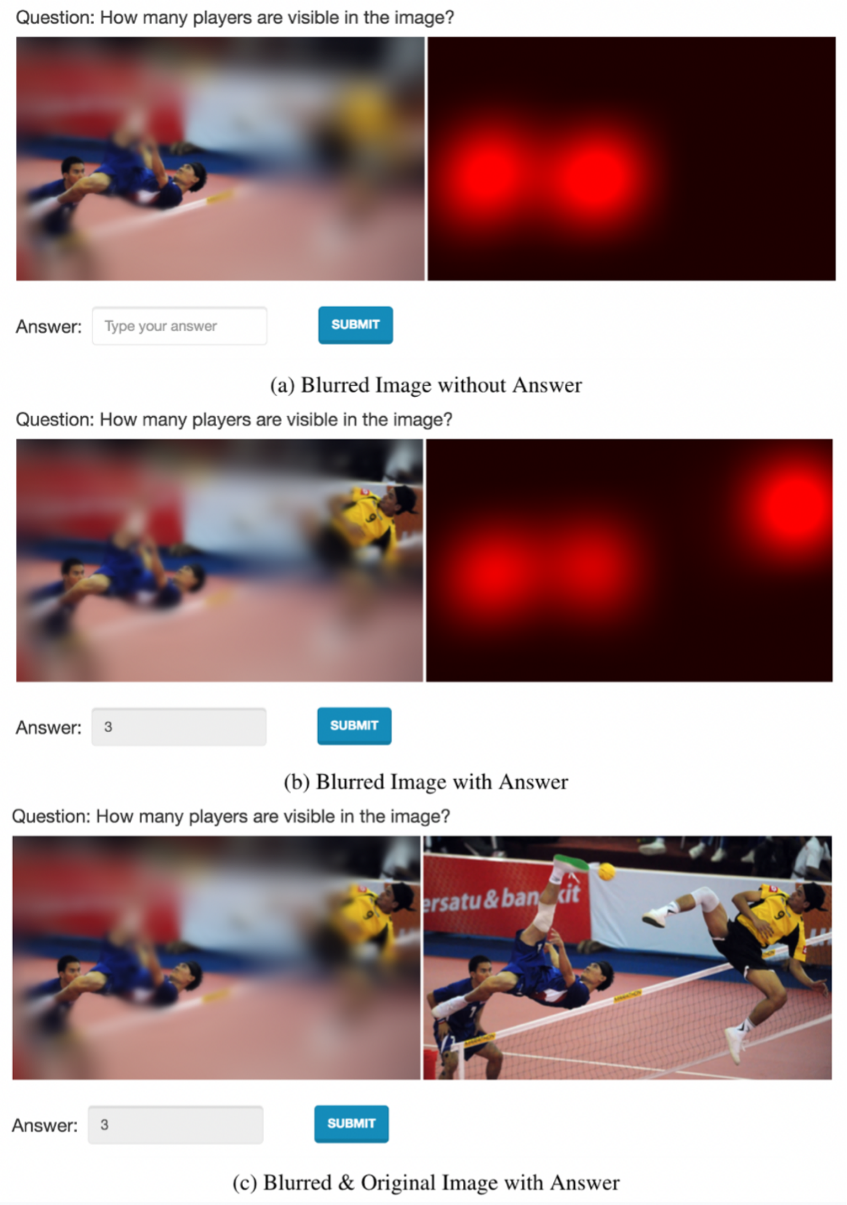


Figure 2.4: This image from the Jiang et al. 2015 paper visualizes their experimental interface with an example: the red circle represents the location of the mouse. Near the mouse's location, the image is unblurred in a way that emulates the way the human eye runs over a scene.

face. However, rather than letting subjects freely explore an image, the researchers prompt them with specific question-answering tasks, allowing them to click-to-unblur parts of the image in the context of the VQA task. The researchers trial three different variants of the interface in which, respectively, (a) the blurred image is provided with a question but without the correct answer, which the subjects have to provide, (b) the blurred image is provided with the correct answer, and the subjects need only unblur the relevant portions to the question-answer pair, and (c) the blurred image is provided alongside the answer and the unblurred image. The three variants of the interface are depicted in Figure 2.5.

To evaluate the quality of the human saliency maps gathered from each variant of the interface, the researchers created a new set of saliency-sharpened images for each interface. In these variants, the parts selected by that interface's human subjects are unblurred. The researchers then feed these selectively-sharpened images to *other* human subjects and evaluate their performance on the VQA task. The assumption is that the more accurate the saliency map, the higher the accuracy of the subjects will be on the VQA task, because the unblurred parts will provide the information needed to answer correctly. This led the researchers to the outcome that the blurred-image-with-answer (interface b) led to the highest performance (78.7% accuracy), followed by the blurred image *without* answer (75.2%), shown in the first image in Figure 2.5.

Based on this finding, they deploy the interface b) in the final study,



Question: How many players are visible in the image?

Answer: Type your answer

(a) Blurred Image without Answer

Question: How many players are visible in the image?

Answer: 3

(b) Blurred Image with Answer

Question: How many players are visible in the image?

Answer: 3

(c) Blurred & Original Image with Answer

Figure 2.5: The three variants of the interface implemented in A. Das, Agrawal, et al. 2017, as depicted in the original paper.

which provides the blurred image along with the question and correct answer. The researchers use this method to create the VQA-HAT dataset of human saliency maps, subsequently used by other researchers as ground-truth saliency maps (e.g. Selvaraju et al. 2019). My approach to human data collection takes A. Das, Agrawal, et al. 2017's method as a starting point, since it is a method that extracts human saliency maps in an "objective", task-oriented way but is far cheaper and easier to deploy than eye-tracking technology.

2.3.2.3 Model attribution maps

I have just discussed techniques in the literature for creating ground-truth human saliency maps to compare AI models' behavior to. But equally important is the generation of *AI* attribution maps: *what* do we compare to the human maps?

The approach used to generate AI attribution maps should ideally be model-agnostic, meaning it can work for any model; this facilitates the scalable calculation of humanlikeness metrics across and between models. There are different levels of the model at which attribution can occur. Some researchers use model internals like attention (a specific term within attention-based architectures) as a way to get at which parts of an image a model finds important (Sood et al. 2023). Others, however, caution that it is important to show not just *whether* a model looked at a given part of the image, but whether that region meaningfully *influences* the model's *output* (Sun et al. 2020; Selvaraju et al. 2019).

These considerations make SHAP (Lundberg and S.-I. Lee 2017) an appropriate XAI method: it is both model-agnostic and focuses on cause and effect, empirically estimating how a given part of the data (in this case, region of an image) influences the output of a model. This causal approach is why I call the XAI maps in my study *attribution* maps rather than attention maps.

2.3.2.3.1 SHAP SHAP is a method for approximating Shapley values. Shapley values, introduced in 1953, can be used as the basis for model-agnostic, local explanations; that is to say, they can provide an explanation connecting a specific model output to specific parts of a specific input, for any model for which the input data is available. Segmenting the input into a discrete number of features, Shapley values are a representation of each feature's contribution to model m's output value f_m (which should be a single scalar value) (Molnar 2022). This relationship is captured in equation 2.1 (adapted from Molnar 2022):

$$\phi_{m,s}(i) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{i\}} \frac{|S|!(p - |S| - 1)!}{p!} \Big(f_m(S \cup \{i\}) - f_m(S) \Big) \tag{2.1}$$

Here, Shapley value $\phi_{m,s}(i)$ reflects the numeric average contribution of feature i to the model m's output value $f_m(s)$ for stimulus s. This contribution is obtained by summing over S, the set of all possible coalitions of non-i features (i.e., sometimes each feature is included, sometimes excluded from a coalition). p is the number of features of each stimulus s. For each coalition, we take the difference of the value of the output variable f_m with feature i included in the feature coalition and the output value without feature i included. In the case of images, features are image regions (superpixels), and the omission of a feature means masking out the corresponding region. To summarize, we calculate the difference of the output value for the input with and without the target feature i, and average this difference over all possible coalitions of other features. This is a local feature contribution value reflecting only a specific image and a specific superpixel. The f_m values are calculated by repeatedly modifying the image by masking out superpixels excluded from a given coalition. A Shapley value $\phi_{m,s}(i)$ for feature i is a *local* value, pegged to a specific model and input set of features $1 \ldots p$.

When we finally apply the estimated Shapley value of each region to an image, we get a saliency map: positive $\phi_{m,s}(i)$ values suggest that superpixel i of stimulus m on average *increases* a given model output $f_m(s)$ by that

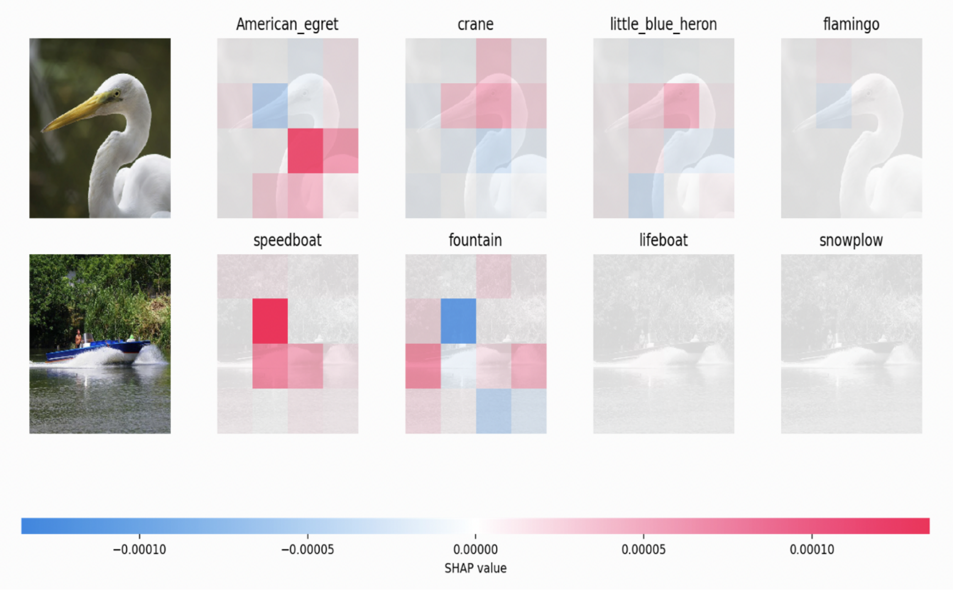


Figure 2.6: The SHAP values for different superpixels for image classifiers; each classification label gets its own model output and corresponding SHAP values (source: *Explain ResNet50 ImageNet classification using Partition explainer — SHAP latest documentation* 2024).

$\phi_{m,s}(i)$ value; negative values suggest that region on average *decreases* the output by that much. With some mathematical processing on both ends, this distribution can be numerically compared to a human saliency map. Figure 2.6 shows some examples of SHAP attribution maps for two images; each map corresponds to a single per-label output value for a classifier, and red superpixels are those with positive SHAP values, i.e. those which are estimated to increase the output value.

The SHAP method approximates Shapley values (the overwhelming number of coalitions makes exhaustive Shapley calculations computationally unaffordable). Lundberg and S.-I. Lee 2017, introducing the method, show that the theoretical underpinnings of SHAP satisfy several important requirements for a robust attribution model and, in fact, encompass several previously-extant XAI methods. SHAP's theoretical robustness, model-agnosticity, intuitive appeal, easy implementation, and wide usage make it a suitable candidate for generating VLM attribution maps that can be compared to human saliency maps. My study's implementation of SHAP is described in detail in chapter 3.

2.3.2.4 A key question: Is humanlike attribution a measure of grounding?

I began this overview with a discussion of the grounding question (section 2.2), and then the assertion that XAI techniques and ground-truth human maps can be used to evaluate the groundedness of a model in some sense (section 2.3).

But is groundedness necessarily related to humanlikeness? Are more humanlike models in some way "better"?

At least since the Turing Test was proposed, many thinkers have assumed a link between humanlikeness and true intelligence in the behavior of machines (Brynjolfsson 2022). In the 1970s, AI pioneer Marvin Minsky said “I draw no boundary between a theory of human thinking and a scheme for making an intelligent machine” (Lake et al. 2016). This is an assumption implicitly or explicitly reflected in AI research that uses human saliency as a model for AI attention (for examples see section 2.3.2.1). Those who favor humanlike intelligence explicitly support their approach with reasoned arguments. Lake et al. 2016, for instance, distinguishes between outcome-oriented “statistical pattern recognition” (how machine learning generally learns) and understanding-focused “model-building” (what they believe humans do), implicitly positioning the latter as the preferable goal for truly intelligent AI. This reflects the views of those who see human saliency maps as the gold standard for intelligent vision models. In this view, humanlikeness is not just indicative of a more grounded model, but rather almost a prerequisite.

However, exponential performance gains among machine-learning models whose behavior is largely unintelligible to human observers have prompted some to reevaluate this anthropocentric view and grant more credence to another view: that non-humanlike pattern recognition is a meaningful, and perhaps equally legitimate path to intelligence (Mitchell and Krakauer 2023). If this view is correct, humanlikeness is more of a subjective measure with limited practical utility, and its relationship to model quality or grounding becomes more dubious.

As it stands, the AI research community continues to, on the one hand, emphasize performance-based measures of grounding (benchmarks like VALSE) while often implicitly assuming, in XAI studies, that some attention patterns are better than others—better because they are more humanlike. The ongoing ambiguities around the concept of "humanlikeness" and its role in AI research are core to the motivation of the present study.

My research centers this dichotomy: between emphasizing performance and emphasizing humanlikeness in normative assessments of AI. I try to find an empirical relationship between the two metrics: 1) a metric of humanlikeness and 2) a performance metric on a grounding benchmark (VALSE). The implications of this will now be discussed.

2.4 Research questions

The previous sections have discussed performance-based measures of grounding as well as XAI-based measures of humanlikeness. In this study, I design an experimental methodology and set of statistical tests to shed light on the relationship between humanlikeness and model performance, specifically model performance on a dataset (VALSE) that is itself created to test for model grounding. In the A. Das, Agrawal, et al. 2017 study my methodology is inspired by, a slight relationship is found between model humanlikeness and performance, though it is a secondary finding and the effect is not very strong. I do expect, as a result, that a similar relationship will be found here.

If it turns out that when their attribution is more humanlike, models do *better* on performance-based grounding challenges, that would validate approaches to AI research that center humanlikeness as a desirable outcome. It would suggest that human-machine attribution comparisons could serve as complements to performance-based benchmarks as another part of the same story, a slightly different angle on grounding. It would imply that the two approaches—humanlikeness and performance—converge to some degree.

If, on the other hand, a VLM's benchmark performance has *no* statistical relationship with how humanlike its attribution is, this puts into question whether at least this particular attribution map to human comparison technique has anything meaningful to say about a model's performance and, by extension, its groundedness (assuming performance on VALSE benchmark correlates to some degree with grounding). Such an outcome would favor understandings of model quality that are centered on performance, and call into question the utility or wide applicability of humanlikeness metrics in model evaluations for visio-linguistic grounding.

However, before we can even begin studying the relationship between performance metrics and humanlikeness (research question 2), it is important to establish whether there is a statistically measurable relationship between human saliency maps and model attribution maps in the first place. This is the focus of research question 1. After all, if the human saliency maps and AI attribution maps in the study are statistically unrelated, the humanlikeness metrics in RQ2 are unlikely to be statistically meaningful.

This study asks two quantitatively testable research questions:

1. **RQ1**: For each of the VLMs considered, is the average similarity between the model's XAI attribution maps and the corresponding human saliency maps statistically significantly higher than the average similarity expected by random chance? (More simply: are the attribution maps for each model on average significantly humanlike?)
    - Null Hypothesis (**H01**): The average similarity between each model's XAI attribution maps and corresponding human saliency maps is not significantly different from the average similarity that would be expected by chance.
    - Alternative Hypothesis (**HA1**): The average similarity between each model's XAI attribution maps and the corresponding human saliency maps is significantly greater than what would be expected by chance.
2. **RQ2**: Does the similarity between a model's XAI attribution maps and

human saliency maps correlate with the model's performance, both on individual images for each model and in aggregate across models, on the VALSE benchmark? (More simply: do models tend to do better on VALSE when their attribution maps are more humanlike?)

- Part 1: Within-model test
  - Null hypothesis (**H02.1**): There is no statistically significant correlation between a model's per-image similarity to human saliency maps and its performance on individual images from the VALSE benchmark.
  - Alternative Hypothesis (**HA2.1**): There is a positive correlation between a model's per-image similarity to human saliency maps and its performance on individual images from the VALSE benchmark.
- Part 2: Between-model test
  - Null hypothesis (**H02.2**): There is no statistically significant correlation between a model's aggregate similarity to human saliency maps and its overall performance on the VALSE benchmark.
  - Alternative Hypothesis (**HA2.2)**: There is a positive correlation between a model's aggregate similarity to human saliency maps and its overall performance on the VALSE benchmark.

The implications of **RQ1**'s alternative hypothesis being confirmed would be that the SHAP-human comparison methodology developed in this study successfully detects an expected similarity between human and XAI maps for the models considered. This is an important part of verifying the validity of the SHAP-human comparisons; it also contributes to understanding the relationship between human-labeled ground truth attribution maps and XAI methods. Beyond this study, it would also indicate the more generalizable usefulness of the human saliency map generation technique developed in my study, helping remedy what Rodis et al. 2023 calls a shortage of studies developing ground-truth human saliency maps.

The implications of **RQ2** are potentially more impactful: it helps us learn more about whether we can measure model grounding by evaluating its attribution maps, and what role a "humanlikeness" metric can play in evaluating model grounding. Confirming either alternative hypothesis for RQ2 could be suggest humanlikeness metrics as a useful complement to performance-based measures of grounding, perhaps even to be used in conjunction with a benchmark like VALSE, telling a different "side" of the model grounding story.

Another benefit of finding a humanlikeness metric that is validated in this way is that it would be useful to situations where "correct" model outputs are less obvious, so performance-based benchmarks become less applicable. In such contexts, researchers may still want to know to what extent a model's performance-agnostic internal behavior aligns with expectations.

3. Methodology

The methodology, which passed an Utrecht University ethics QuickScan with no flagged issues[1], is sketched out in the diagram in Figure 3.1. It can be divided into 5 major steps, which are labeled as such in the diagram.

The first three steps were data collection and generation steps. The first step was to select what subset of the VALSE benchmark to collect saliency and attribution maps for, and generate a set of stimuli (section 3.1.1). The second step was to generate human attribution maps by running a human experiment exposing human subjects to the selected stimuli, and processing the results (section 3.1.2). The third step was to implement four VLMs (LXMERT, CLIP, FLAVA, SigLip) and generate model attribution maps (with SHAP), as well as model outputs for the *same* stimuli and across the broader VALSE dataset — in order to have performance data to analyze (see section 3.1.3).

The fourth and fifth steps were data analysis steps. The fourth step was to compute similarity metrics between the human and SHAP attribution maps (section 3.2.1). In the fifth step, these comparison metrics, as well as the outputs for each model, were used to answer the two research questions (section 3.2.2).

3.1 Data collection

To generate the data used in the final analysis, I selected a subset of the VALSE dataset and made slight modifications to it (section 3.1.1). I then fed these stimuli along with correct and incorrect captions to both human sub-

[1] The Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing Sciences was conducted (see Appendix C). This research project was classified as low-risk without a fuller ethics review or privacy assessment required.

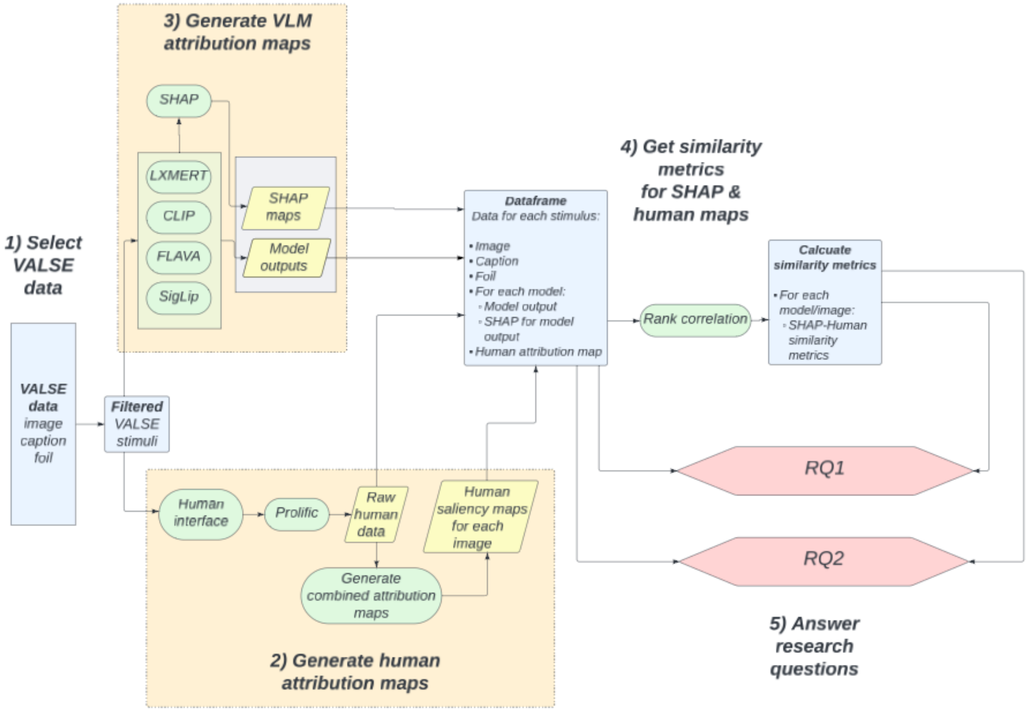


Figure 3.1: A sketch of the overall methodology in five steps.

jects and AI models. For the humans, I used a data collection interface to generate attribution maps for each stimulus (section 3.1.2). For the AI models, I used the model-agnostic XAI method SHAP to generate attribution maps (section 3.1.3).

The key task that both humans and AI models were tasked with is this: presented with an image and a caption (correct description) and foil (incorrect description), they were asked to choose which is correct. Data from this interaction was used to generate attribution maps, which show which regions of the image contributed more or less to the output.

3.1.1 Creating a set of experimental stimuli

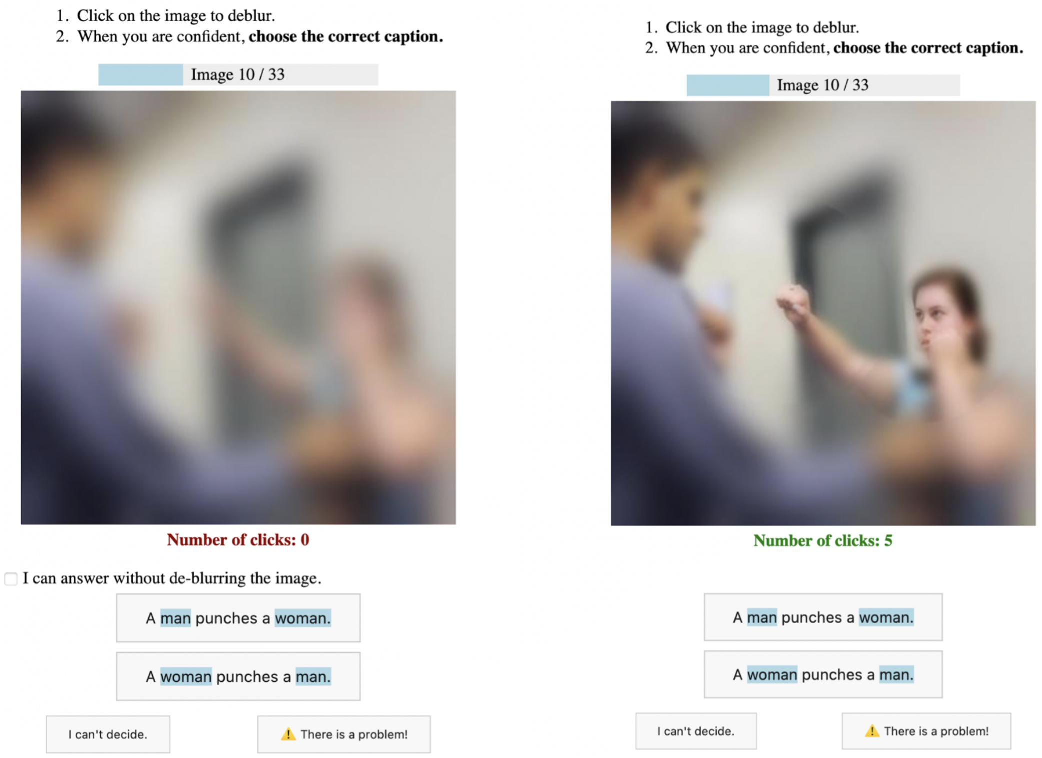
I selected a subset of the VALSE dataset to present to human subjects as well as to use as the basis for the SHAP attribution maps. I refer to the elements of this dataset as the **stimuli** – with each stimulus consisting of an image and a caption (correct) and foil (incorrect) that goes with it. This section describes how the stimuli were selected and prepared.

The images in VALSE are themselves drawn from various pre-existing datasets, namely the MSCOCO 2014 (Lin et al. 2014), MSCOCO 2017 (*COCO - Common Objects in Context* 2017), VisDial 1.0 (A. Das, Kottur, et al. 2016), SWiG (Ocampo and Bather 2022), and Visual7w (Zhu et al. 2015) datasets. Many images in some of the non-MSCOCO datasets are originally from MSCOCO 2014 or 2017 as well. I loaded the images from the source datasets and aligned them with the filenames present in the VALSE dataset's online repository, thus creating a copy of the original VALSE dataset, as instructed by VALSE's online data repository (Parcalabescu, Cafagna, et al. 2022).

VALSE contains a number of distinct linguistic phenomena that models are tested on in the original VALSE paper. These are "existence", "plurality", "counting", "relations", "actions" and "coreference." (Examples can be seen in section 2.3.1.) Some of these “pieces” are more appropriate than others to deploy in an experimental setting with humans to collect saliency maps.

Existence, for instance, requires models to evaluate the presence of an object in the image and is thus related to object detection; in this way, it is often obviously localizable to a specific point in an image. Similarly, *relations* and the subject-object interchange of *actions*, while more complex, are reasonably localizable. *Coreference*, *plurality*, and *counting*, however, may pose the risk of having more complex saliency considerations. Ultimately, I made the decision to limit the study to the first three phenomena mentioned here (actions, relations, existence), largely out of a desire to have significant results with a modest number of samples, since I expect there to be less divergence between different human subjects on which regions matter.

In the end, I generated 33 samples from VALSE for each of the three pieces: *existence*, *relations*, and *actions*, for a total of 99 experimental stimuli. Within each linguistic phenomena, I aimed for a balance between different levels of difficulty. This is accomplished through three steps: first, initially sampling a subset of the VALSE dataset; then, manually evaluating the quality of the stimuli; and finally, sampling the final set of 99 stimuli in a balanced way. The exact way this was done is likely to hold little theoretical interest and is thus laid out in detail in Appendix B.



1. Click on the image to deblur.
2. When you are confident, **choose the correct caption.**

Image 10 / 33

Number of clicks: 0

□ I can answer without de-blurring the image.

1. Click on the image to deblur.
2. When you are confident, **choose the correct caption.**

Image 10 / 33

Number of clicks: 5

Figure 3.2: The human data collection interface, initially (left) and after 5 clicks (right).

In the rest of this paper, when I refer to "stimuli" or "experimental stimuli", it refers specifically to this set of 99 stimuli which have been edited and balanced before being presented to both human subjects and AI models.

3.1.2 Generating human saliency maps

I used an online, custom-built experimental environment[2] to collect data from human subjects (Figure 3.2). I then used the collected data to create averaged human saliency maps for each stimulus. The score each point on an image gets this way is meant to represent the following intuition: **The higher the score that a part of the image gets, the greater its interest to human subjects as they explored the image to determine the correct caption.**

[2] The code for the interface is available on GitHub at https://github.com/skshvl/thesis-webapp-public

3.1.2.1 Considerations for the interface design

My approach to collecting human saliency maps takes as its starting point the data collection methodology of a previous paper (A. Das, Agrawal, et al. 2017), which is discussed at some length in section 2.3.2.2. After testing several iterations of their interface, the authors finally selected one in which human subjects were presented with the image, a question *and* the correct answer. They then had to unblur the most relevant regions of the image. This was found to produce the highest quality attention maps. However, in my study an approach *without* the answer pre-given is used. In this interface variant, humans see both the caption and the foil and can reason about which is correct by deblurring the image, without already knowing the right answer. The following considerations support this design decision:

1. The quality difference between the provided answer and no-provided-answer variants of the interface was not very large in the A. Das, Agrawal, et al. 2017 study (78.7% vs 75.2% by their quality metric for attribution maps).
2. Unlike the straightforward question-answering (VQA) task from A. Das, Agrawal, et al. 2017, choosing a correct caption between two highly similar captions is *not* an intuitive and familiar task to most people. As a result, asking subjects to reason *hypothetically* about what regions would help them choose the correct answer when they already *know* what the answer poses a risk to the quality of the data. Turning the task into a "real" task in which the answer is unknown and must be found through deblurring the image would address this concern.

I deployed an interface online, collecting several human attention maps per image based on where people clicked; the contributions of people who picked the correct caption were then converted to aggregate attention maps for each image.

The methodology for converting clicks into attention maps is a modified version of very similar methodologies already present in the literature (A.

Das, Agrawal, et al. 2017; Jiang et al. 2015)[3], discussed in detail in section 2.3.2.2.

3.1.2.2 Interface design

I built an interface using the Flask framework in Python, and deployed it to a GDPR-compliant web server provided by the cloud computing service PythonAnywhere.

The interface consisted of introductory pages to get consent and explain the experiment, a training step, and finally a series of images presented in the interface shown in Figure 3.2. For each image, a user can click one of two candidate captions. They are encouraged to first deblur the relevant areas of the picture, but if the answer is already clear, they can check "I can answer without de-blurring" and choose a caption. They also have the option of saying they "can't decide" or to report a problem. Initially (left in Figure 3.2), the image is fully blurred. After several clicks on relevant parts of the image, those areas are de-blurred (right). Information about where the user clicks is stored to generate the aggregate human attention map.

The 99 stimuli were randomly divided into three groups of 33 each, with equal representation of *actions*, *relations*, and *existence* stimuli in each subgroup.

3.1.2.3 Turning clicks into saliency maps

The technical details of how human clicks were collected and converted into an attention map for each subject and picture are given here. The math of the de-blurring and mask generation process is based on an adapted version of the technical implementation of the A. Das, Agrawal, et al. 2017 study.

Each **image was resized to 400 by 400** pixels before being presented to experimental subjects. This both keeps attribution map sizes consistent between images and matches the 1:1 aspect ratio presented to the VLM models in generating their attribution maps. The de-blurring of each image was

[3] An author on A. Das, Agrawal, et al. 2017 provided me with the files to their original online interface implementation.

governed by a **de-blurring mask** with the same size (400 x 400). Each pixel in the de-blurring mask could have a score between 1 and 255, and is initialized to 1.0, which we can represent with the variable mask(x, y) where x and y are the coordinates of the pixel.

A mask(x, y) value of 1.0 corresponds to a maximally blurred pixel, while 255 corresponds to a fully unblurred pixel (i.e. the original image). Meanwhile, a value exactly in the middle (128) corresponds to a medium-blurred-pixel. The three main blur intensities and their corresponding mask values are as follows:

1. No blur: mask$(x, y) = 255.0$
2. Gaussian blur with kernel size (33,33): mask$(x, y) = 128.0$
3. Gaussian blur with kernel size (99,99): mask$(x, y) = 1.0$

When $m(x, y)$ for a pixel lies between the three mask values given in this list, the pixel value is determined by **linear interpolation** between the two adjacent blur levels. So for instance if a pixel's mask$(x, y) = 64.5$, which lies at the midpoint between 1 and 128, the pixel's RGB color value will be an exact mean of the medium (33,33) and maximally (99,99) blurred versions of the pixel value.

Each time a user **clicks** on the image, the coordinates of the click (X, Y) are passed to the de-blurring mask, and the value of the mask at each point within the mask, mask(x, y) is then updated by calculating a mask update variable $\Delta_{\text{mask}}(x, y)$.

$\Delta_{\text{mask}}(x, y)$ can have a nonzero value for all points (x, y) in the mask within the radius **BrushRadius** of 100 (outside this radius, the mask is not updated). Equation 3.1 shows how $\Delta_{\text{mask}}(x, y)$ is determined.

$$\Delta_{\text{mask}}(x, y) = \begin{cases} 100 \times \exp\left(-\frac{d(x,y)^2}{0.4 \times \text{BrushRadius}^2}\right) & \text{if } d(x, y) \leq \text{BrushRadius} \\ 0 & \text{if } d(x, y) > \text{BrushRadius} \end{cases} \tag{3.1}$$

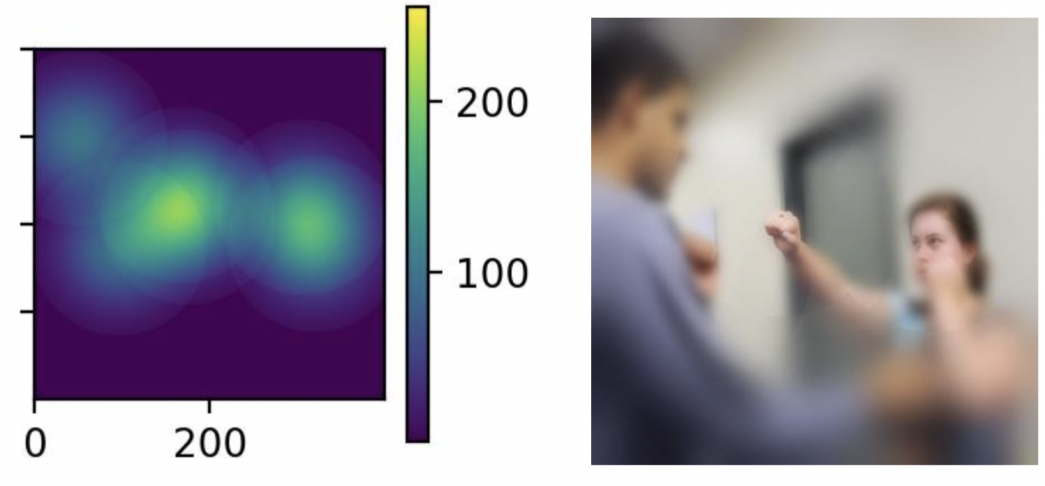


Figure 3.3: The mask (left) depicting the value of $m(x,y)$ at each point, and the correspondingly unblurred image (right) after a user has clicked to unblur six times.

where $d(x,y)$ is the Euclidean distance between any given point (x,y) in the mask and the user click location (X,Y). This equation is the same as used in the implementation of A. Das, Agrawal, et al. 2017, except for the BrushRadius value differing and the maximum $\Delta_{\text{mask}}(x,y)$ value per click of 100 being significantly higher in this case than used in the original study. (In the original study, users could "stroke" across an area, encompassing multiple clicks in a single mouse movement, hence the Δ_{mask} for each click was lower.)

With each user click, $\Delta_{\text{mask}}(x,y)$ is calculated for each point and used to update the mask value for each point, mask(x,y), up to 255, based on Equation 3.2.

$$\text{mask}(x,y) = \min\Big(\text{mask}(x,y) + \Delta_{\text{mask}}(x,y), 255\Big) \tag{3.2}$$

After each click, this updated mask was then used to update the display of the image to the user. Figure 3.3 depicts the mask for an image after six clicks (left, with lighter areas having higher mask values), and the corresponding unblurred version of the image (right).

Once a user submitted an answer, the final version of the mask was recorded as the saliency mask corresponding to that interaction.

3.1.2.4 Human data collection timeline

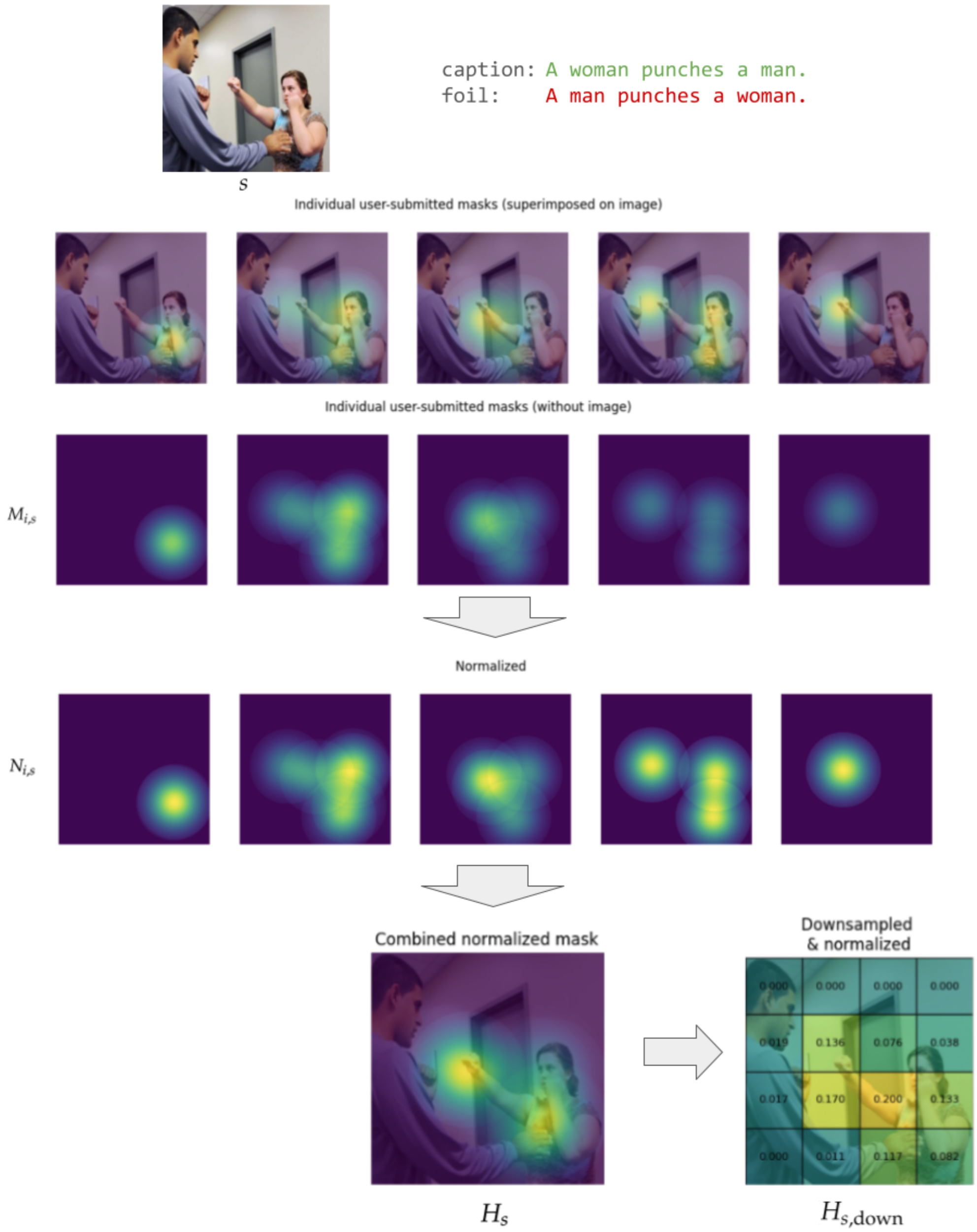
Between February 1 and 9, 2024, a group of 17 experimental subjects recruited through the service Prolific were each shown a set of stimuli. First, 15 experimental subjects saw 33 stimuli each, leading to 495 user-stimulus interactions. Each individual stimulus was seen by 5 subjects. Human interactions with each stimulus were then validated according to the following criteria: 1) the user correctly identified the caption and 2) the user clicked to deblur the image at least once.

The rationale for these validation criteria, respectively, is that 1) we are interested in interactions where humans explored the image sufficiently to produce a correct answer, which is presumed to be trivial for a sufficiently-informed human subject (otherwise they can select that they cannot answer) and 2) a response in which nothing is deblurred cannot contribute to making a deblurring-based attribution map.

After this initial group of 15 subjects was shown a set of 33 stimuli each, this generated 337 validated user-stimulus interactions (out of 495 total interactions). However, **only 78 out of 99 stimuli** corresponded to at least 3 such validated user interactions – 3 was the minimum I had decided on for generating an aggregate attribution map for each stimulus.

To increase the number of stimuli for which human saliency maps were available, I presented each of the 17 stimuli which only produced 2 validated interactions in the initial data collection to two *additional* experimental subjects on Prolific. The goal of this was to increase the number of validated interactions for these stimuli, so that more aggregate human saliency maps would be created.

After collecting more data, the number of stimuli with 3 or more validated responses increased to **92 out of 99 stimuli.** (Now, out of 529 total interactions, 357 were validated, and of those, 348 belonged to a stimulus with at least three validated interactions.) Thus, aggregate human saliency



```
caption: A woman punches a man.
foil:    A man punches a woman.
```

0.000	0.000	0.000	0.000
0.019	0.136	0.076	0.038
0.017	0.170	0.200	0.133
0.000	0.011	0.117	0.082

Figure 3.4: A diagram depicting the normalization, averaging, and downsampling of human attention maps.

maps could be generated for 92 stimuli.

3.1.2.5 Creation of aggregate human saliency maps

On a stimuli-per-stimuli basis, the masks from validated user responses were 1) normalized, 2) averaged on a pixel-by-pixel basis across masks from the same stimulus, and 3) downsampled to a 4x4 array. This whole process is visualized for the "A woman punches a man" stimulus in Figure 3.4.

Assume each stimulus s each validated user mask is represented by the matrix $M_{i,s}$, where i represents the user and each element consists of the corresponding $\text{mask}(x, y)$ value for that stimulus s and human subject i. I first normalized each of these masks by dividing each element by the element-wise sum of the matrix to produce the normalized mask $N_{i,s}$, according to Equation 3.3, where $\text{mask}(x, y)$ is any given element of $M_{i,s}$.

$$N_{i,s} = \frac{M_{i,s}}{\sum M_{i,s}} \tag{3.3}$$

Then, for each stimulus s (among stimuli for which at least 3 validated responses were available) for which there are three or more validated normalized masks each represented as some matrix $N_{i,s}$, I generated **aggregate human mask** H_s for that stimulus (s), as shown in Equation 3.4, where n is the number of validated normalized masks for a given stimulus s (never less than 3), and i represents each human-generated index from 1 to n.

$$H_s = \frac{1}{n} \sum_{i=1}^{n} N_{i,s} \tag{3.4}$$

$$H_{s,\text{down}} = \text{Downsample}_{4\times4}(H_s) \tag{3.5}$$

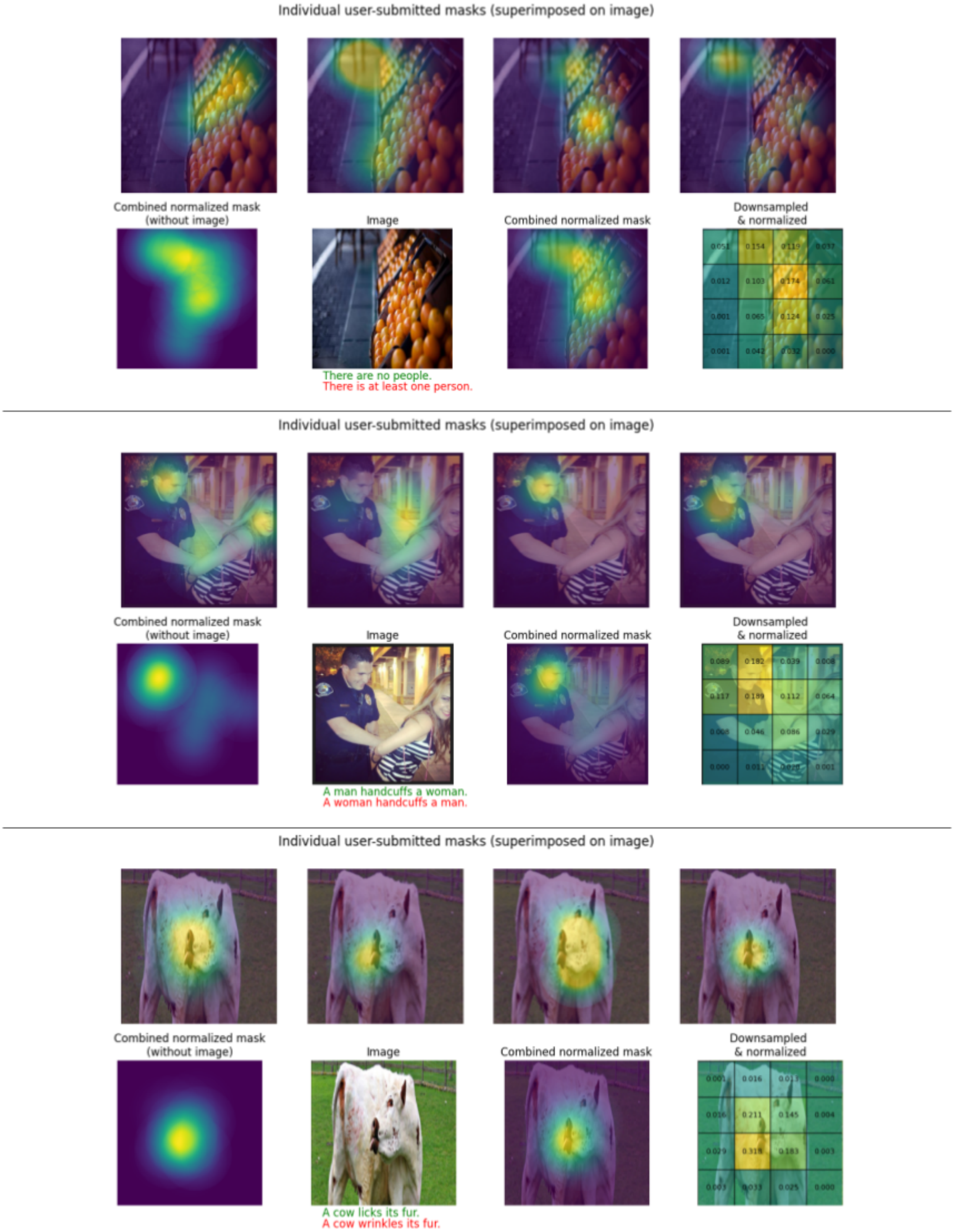


Figure 3.5: For each image, the top row visualizes four human-generated masks, while the second row visualizes what the aggregate saliency map looks like before and after downsampling.

Finally, this aggregate mask for each stimulus was downsampled as in Equation 3.5 by averaging across 16 square patches arranged in a 4 x 4 structure, leading to a final $H_{s,\text{down}}$ with shape (4,4); this downsampling by averaging does not affect normalization.

This **aggregate downsampled human saliency map** $H_{s,\text{down}}$ was used as the human saliency map for each of the 92 stimuli s for which it was generated. The entire process for a single example with 5 validated user attention masks is illustrated in Figure 3.4. Three more examples of this process are illustrated in more compact form in Figure 3.5.

3.1.3 Generating model attribution maps

I used the model-agnostic SHAP method (described in 2.3.2.3.1) to generate attribution maps for each stimulus-model combination for four models. Separately, I also recorded the output of each model on a large portion of the VALSE dataset as well as the experimental stimuli. The code for each model's implementation and the generation of attribution maps can be found on this project's data analysis GitHub repository[4].

3.1.3.1 Model selection and implementation

The aim of this study is not to give a comprehensive overview of the state of the art of VLMs in terms of how humanlike their attribution maps are. Rather, I use a mix of models to inject variety into the results of the study, and aim to use models with some variety of architectures and expected performances on the VALSE dataset. Another consideration in model selection is ease of implementation for the **zero-shot image-text alignment task**. All models were loaded through the Huggingface interface (*Hugging Face - Documentation* 2024), and in the case of CLIP and LXMERT I repurposed some of the implementation code from the VALSE paper's GitHub repository (Parcalabescu, Cafagna, et al. 2022).

All images were resized to square before being passed into the models

[4] Data analysis repository for this project: https://github.com/skshvl/thesis-data-public/

Model (year)	**Dual encoder**	**Cross attention layer**	**Multimodal layer**	**Joint representation**	**Separate image/text encodings**	**Scalar Output Value for Stimulus** s	**Overall VALSE Accuracy**
LXMERT (2019)	✓	✓		✓		$f_m(s) = P_{\text{caption}}(s) - P_{\text{foil}}(s)$	53.3%
CLIP (2021)	✓				✓	$f_m(s) = \text{score}_{\text{caption}}(s) - \text{score}_{\text{foil}}(s)$	71.6%
FLAVA (2022)	✓		✓	✓		$f_m(s) = \text{score}_{\text{caption}}(s) - \text{score}_{\text{foil}}(s)$	67.2%
SigLip (2023)	✓				✓	$f_m(s) = \text{score}_{\text{caption}}(s) - \text{score}_{\text{foil}}(s)$	66.8%

Table 3.1: A summary of the models used in the study. $f_m(s)$ represents the single scalar output value of model m for stimulus s. The overall VALSE accuracy is averaged between the accuracy on the validated actions, existence, and relations examples in the entire VALSE dataset.

to generate output scores as well as for SHAP attribution maps.

The models are summarized in Table 3.1. The models used are all dual encoders, but generate one of two types of final representations: a joint image/text representation (LXMERT and FLAVA) or separate encodings of image and text (CLIP and SigLip). The former involves a cross-modal or multimodal layer to create a joint representation, while the latter does not. Each either by design or with a little final processing allows users to generate a similarity score between the image and text modalities, represented in a single scalar output. This scalar output f_m for each model is used as the basis for SHAP attribution maps, but also to generate performance data for each model in this study.

The Shapley value (defined in Equation 2.1) estimate can be represented as $\phi_{m,s}(i, j)$, which represents the impact of image patch with coordinates (i, j) for a model m and stimulus s on the output variable $f_m(s)$. (s refers

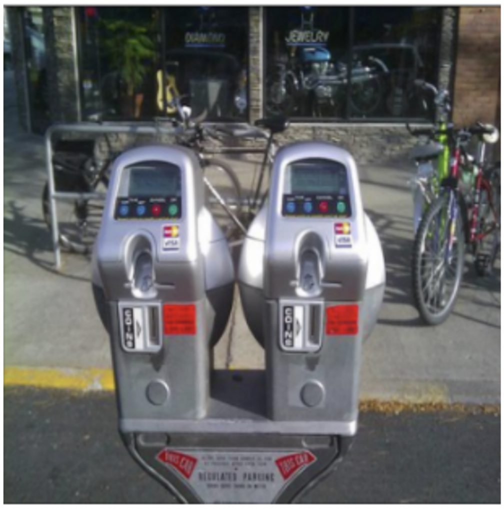


Figure 3.6: SigLip, when shown this image with the caption "There is at least one bike on the rack" and the foil "There are no bikes on the rack", assigns them each the following image-text similarity scores, respectively: -4.9801 (caption) and -4.7209 (foil) and the caption-foil difference is -0.2592. The model thus gets it *wrong*.

to both the image and the caption/foil of each stimulus.) We define the scalar variable f_m for each model so that positive values correspond to the correct choice (caption) and the negative values to the incorrect choice (foil). Moreover, the higher the magnitude of the value, the more confident the model is. The output value defined here for each model will be used in this study whenever a single scalar output for a model is required. Table 3.1 contains the scalar output formula for each model.

For example, the model SigLip, when shown Figure 3.6 with the caption "There is at least one bike on the rack" and the foil "There are no bikes on the rack", assigns them each the following image-text similarity scores, respectively: -4.9801 (caption) and -4.7209 (foil) and the caption-foil difference is -0.2592. The caption-foil score difference will be $f_m(s)$, where s identifies the stimulus for which f_m is the single scalar output for model m.

In this case, $f_m(s)$ is negative: -0.2592. This shows the model making a mistake: it assigns a higher score to the foil than to the caption. The caption-foil difference $f_m(s)$ is the representation of the model's final answer on which text is correct: positive means caption, and negative means foil.

Further technical details of each model and its implementation are discussed in Appendix A.

3.1.3.2 Generating SHAP attribution maps

We generate attribution maps for each image-model pair by using the SHAP package. The SHAP package, as explained in section 2.3.2.3.1, approximates Shapley values. The entire process for generating model attribution maps in my study is illustrated in Figure 3.7.

To run my 99 experimental stimuli through the SHAP process, I resized each image to a square and divided it into 16 equally sized square patches, arranged 4x4. I then defined a 4 x 4 matrix that identifies whether any given patch is visible or masked. In its starting position, the image matrix A looks like this, with all patches set to visible:

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}$$

We also provide SHAP with a "background matrix" B which has the value:

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

We then use the SHAP package to generate a set V of 172 variants of matrix A that “mask out” different areas using the `shap.Explainer()` class. For any matrix A' in set V, one or more patches may be replaced with the 0 background value from background matrix B. An example can be seen in part (a) of Figure 3.7.

We also pass SHAP a prediction function which takes this masked version of the matrix, A', converts it to a masked image we can call s' (part (b) of Figure 3.7), and generates a model output score $f_m(s')$ for the modified image s' (part (c) of the figure). When converting the matrix to an image, patches with a 0 value get replaced with a blurred version, generated with a Gaussian Blur function with kernel size (99, 99) — the same as the highest blur level in the human study. Two examples of such partly-blurred images are shown in part (b) of Figure 3.7.

In the SHAP package's approximation algorithm, the model outputs for these 172 variant matrices A' are generated and compared, to estimate the contribution of each of the 16 square patches of the image to the output score, represented in a Shapley value $\phi_{m,s}(i,j)$ for each patch row i and column j in the 4x4 arrangement of patches. A high positive $\phi_{m,s}(i,j)$ value is indicative of an important region that increases the correctness of the model's output by increasing the caption-foil difference. But what of a negative value?

If we consider what a negative Shapley value $\phi_{m,s}(i)$ means, a negative-valued patch is one that lowers the model output score, i.e. decreases the caption-foil difference and pushes the model towards the wrong answer. However, we are most interested in the *saliency* of the region; that is, we want to know whether it has a big impact on the model's assessment, because an impactful region for the model is likely to be important for human subjects, too, if they see the picture in a similar way. Thus, we take the magnitude of the SHAP values of each patch, and normalize them to add up to 1 (see part (e) of Figure 3.7), creating a new set of SHAP values $\phi'_{m,s}(i,j)$ for each patch (i,j), where m is the model and s is the stimulus for which the SHAP values are being generated:

$$\phi'_{m,s}(i,j) = \frac{|\phi_{m,s}(i,j)|}{\sum_{i=1}^{4}\sum_{j=1}^{4}|\phi_{m,s}(i,j)|}$$

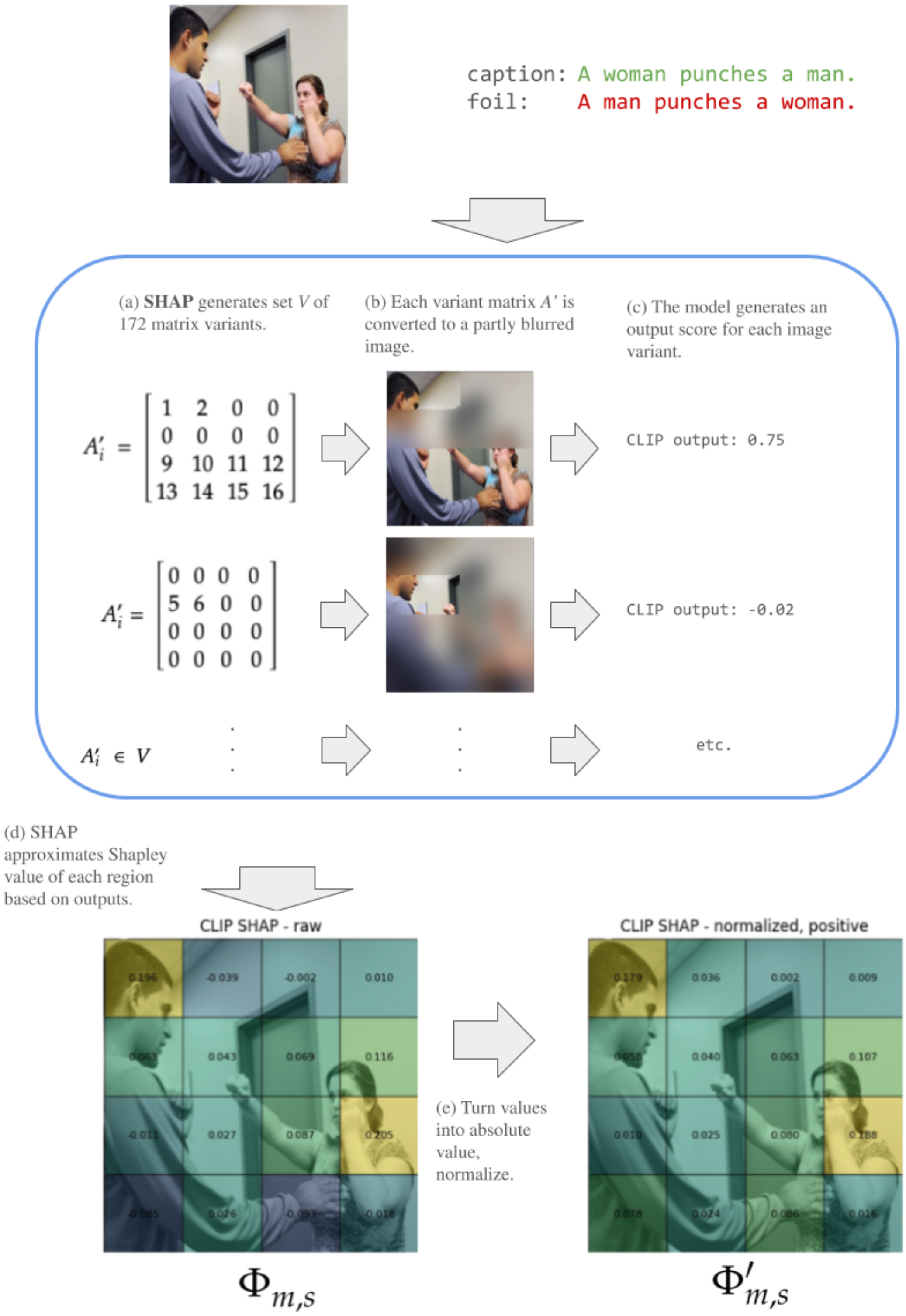

caption: A woman punches a man.
foil: A man punches a woman.
(a) SHAP generates set V of 172 matrix variants.
(b) Each variant matrix A' is converted to a partly blurred image.
(c) The model generates an output score for each image variant.
$A'_i = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}$
CLIP output: 0.75
$A'_i = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 5 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
CLIP output: -0.02
$A'_i \in V$
etc.
(d) SHAP approximates Shapley value of each region based on outputs.
CLIP SHAP - raw
0.196
-0.039
-0.002
0.010
0.063
0.043
0.069
0.116
-0.011
0.027
0.087
0.205
-0.085
0.026
-0.093
-0.018
$\Phi_{m,s}$
(e) Turn values into absolute value, normalize.
CLIP SHAP - normalized, positive
0.179
0.036
0.002
0.009
0.058
0.040
0.063
0.107
0.010
0.025
0.080
0.188
0.078
0.024
0.086
0.016
$\Phi'_{m,s}$

Figure 3.7: A graphical representation of the generation of a normalized SHAP map for CLIP. (Other models go through the same process.)

The 4x4 normalized, positive SHAP matrix of ϕ' values for each model m and stimulus s can be represented as $\Phi'_{m,s}$ as seen in Equation 3.6.

$$\Phi'_{m,s} = \begin{bmatrix} \phi'_{m,s}(1,1) & \dots & \phi'_{m,s}(1,4) \\ \vdots & \ddots & \vdots \\ \phi'_{m,s}(4,1) & \dots & \phi'_{m,s}(4,4) \end{bmatrix} \tag{3.6}$$

This process was applied to all 99 experimental stimuli (s) for all four models m: LXMERT, CLIP, FLAVA and SigLip, resulting in a total of 396 model attribution matrices $\Phi'_{m,s}$. Figure 3.8 shows the SHAP maps generated for the same stimulus for four different models, as well as the model (m) output $f_m(s)$ for each unmodified stimulus s (positive f means correct).

3.1.3.3 Generating model outputs

Each model m's output for the edited set of 99 final stimuli s, represented by variable $f_m(s)$, was also saved. These will be important for the comparison between humanlikeness metrics and model performance data.

Separately, I ran each model on the entire validated VALSE dataset for the *existence*, *relations*, and *actions* pieces, leading to 2637 output data points $f_m(v)$ for each model m and VALSE data point v (I do not call v a "stimulus" because I reserve that term for the 99 stimuli in the human study).

Note that this process used the original content of the VALSE dataset, leaving out the edits made in the data generation stage of my experiment. This data is not used to directly compare behavior between the model and humans, but rather as reference data for the overall performance of each model on the relevant parts of VALSE.

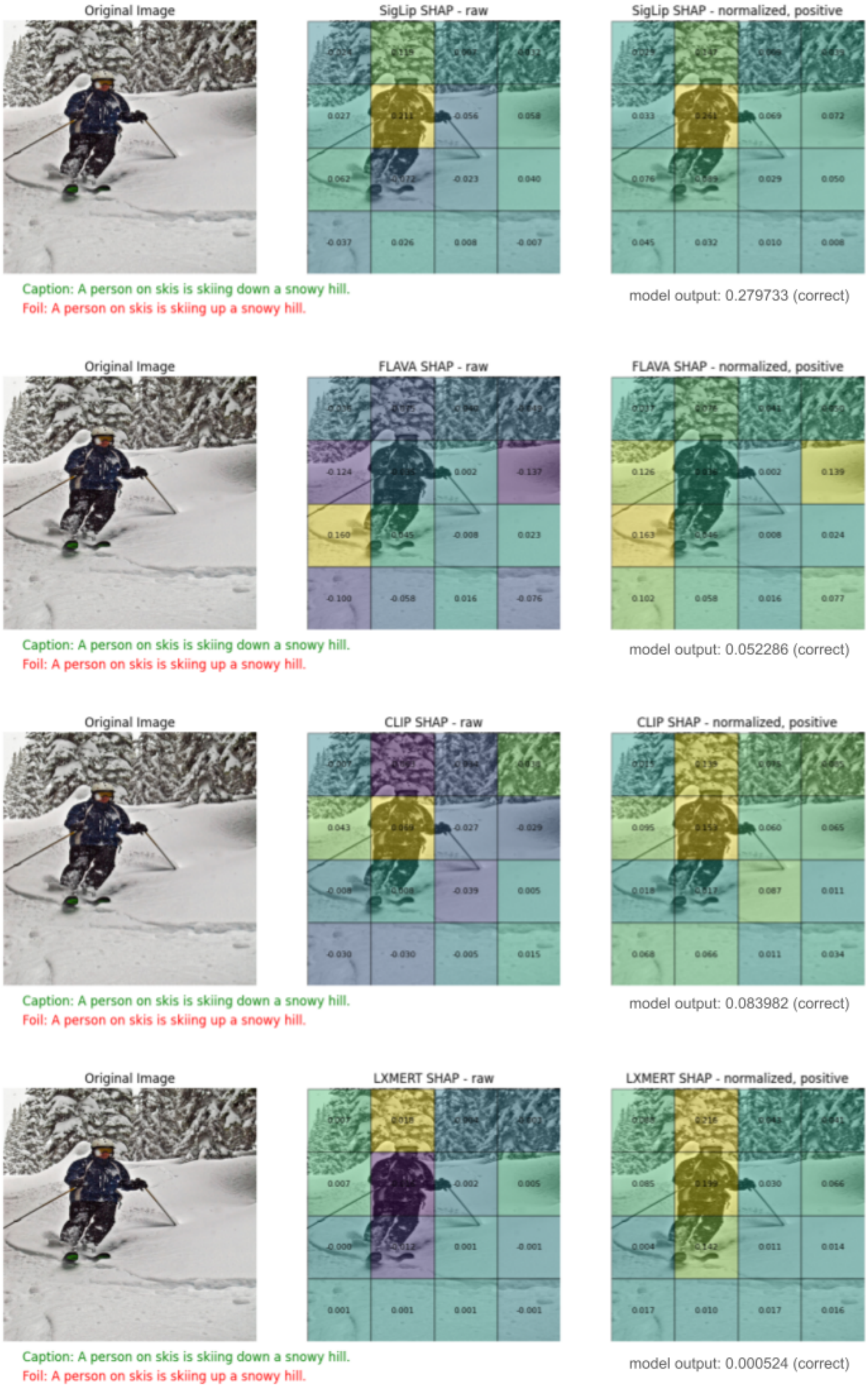


Figure 3.8: Raw and normalized SHAP attribution maps for four models on the same stimulus, as well as the model output for each stimulus.

3.1.4 Summary of data collection

To summarize, as a result of the data collection steps described in sections 3.1.1, 3.1.2, and 3.1.3, I have created the following data structures:

1. A set of 99 edited stimuli drawn from the VALSE dataset, including an image, caption, and foil.
2. For each stimulus s for which enough human data is available (92 out of 99 stimuli), an aggregate downsampled human saliency map $H_{s,\text{down}}$, represented as a 4x4 matrix.
3. For each stimulus s and model m (among CLIP, LXMERT, FLAVA, SigLip):
   (a) A model output score for that stimulus, $f_m(s)$.
   (a) A normalized, positive-valued SHAP attribution map represented as matrix $\Phi'_{m,s}$ which estimates each stimulus image region's impact on the model output for that stimulus, $f_m(s)$
4. For each validated data element v in the *actions*, *relations*, and *existence* pieces of the VALSE dataset, a model output score $f_m(v)$ for each model m.

For the rest of the analysis, I will use 92 rather than 99 stimuli in the analysis, as these are the ones for which human maps are available.

3.2 Data analysis

This section discusses final data processing by generating comparison metrics and then describes the methodology for answering the research questions quantitatively.

3.2.1 Calculation of comparison metrics

Before running the final data analyses, I calculated individual similarity metrics between each model's SHAP attribution maps and their corresponding human saliency maps, each of them being normalized arrays of shape

(4,4).

There is **one SHAP attribution map**, $\Phi_{m,s}$ for each combination of model m and stimulus s, as defined in Equation 3.6. There is **one aggregate human attribution map**, $H_{s,\text{down}}$, for each stimulus s, as defined in Equation 3.5.

The metric comparing each model-stimulus combination's $\Phi'_{m,s}$ and $H_{s,\text{down}}$ should be a metric that would be able to detect correlation between the two maps, if it exists.

3.2.1.1 Metrics considered

In Rodis et al. 2023's review of multimodal explainability, a list of metrics for comparing attention maps to ground-truth maps is given; among these, Earth Mover's Distance (**EMD**) and **Rank Correlation** (referred to later as **RC**) are particularly suitable for this context. RC was also used in two closely related precursor studies: A. Das, Agrawal, et al. 2017 and Selvaraju et al. 2019.

3.2.1.1.1 Earth Mover's Distance (EMD) The intuition behind EMD is that it is the amount of "work" required to turn one distribution (say, the human saliency) into another (the SHAP attribution map) if we imagine each as a pile of earth (hence the "earth" mover). The earth mover's distance between the two matrices in this case can be represented by the following equation:

$$\text{EMD}_{s,m}(\vec{a},\vec{b}) = min_F \sum_{ij} f_{ij} d_{ij}$$

where$\vec{}$and $\vec{b}$ are **flattened** versions of the 4x4 SHAP attribution map $\Phi_{m,s}$ (for a given model and stimulus) and 4x4 aggregate downsampled human attribution map $H_{s,\text{down}}$ (for a given stimulus). d_{ij} is the Euclidian distance between elements i and j of the first and second matrices[5], respectively, and

[5] The distance is based on the Euclidian distance between the two entries in the orig-

F represents the set of all possible transportation arrangements of "mass" between the distributions, where each element f_{ij} is an element representing that transport between elements i and j.

I implemented EMD using the Python Optimal Transport library (Flamary et al. 2021).

Note that, unlike Rank Correlation, for our purposes the numeric value of EMD has little direct meaning except relative to other EMD values. Smaller EMD values indicate greater closeness between two distributions.

3.2.1.1.2 Rank Correlation (RC) as a similarity metric The rank correlation, also called Spearman Rank Correlation (*scipy.stats.spearmanr — SciPy v1.12.0 Manual* 2024) approach first converts each two-dimensional distribution to a flat vector, say $\vec{x}$ and $\vec{y}$, and then replaces values at each index with the rank of that value in the vector by applying a **Rank** function.

In our case, for each vectorized 4 x 4 attribution/saliency map, the lowest value cell gets the rank 16, while the highest value gets rank 1. This creates two new vectors, $\vec{rx}$ and $\vec{ry}$ with 16 elements each.

We can arbitrarily say that the first of these vectors is a flattened and ranked version of a given stimulus s's human saliency map:

$$\vec{rx} = \text{Rank}\Big(\text{Flatten}(H_{s,\text{down}})\Big)$$

While the second vector is a flattened and ranked version of a given stimulus s and model m's SHAP attribution map $\Phi'_{m,s}$:

$$\vec{ry} = \text{Rank}\Big(\text{Flatten}(\Phi'_{m,s})\Big)$$

inal two-dimensional matrixes, treating the row and column of each element as a spatial location.

The order of which vector is which has no impact on the final RC value.

We then calculate d_i for each i going from 1 to 16:

$$d_i = rx_i - ry_i$$

The equation for **RC** is given in Equation 3.7, where $n = 16$ in our case and RC$_{m,s}$ represents the Rank Correlation between the corersponding human and SHAP maps for model m and stimulus s. The Spearman correlation RC can go from -1 to 1, with 0 representing no correlation, 1 a perfect correlation, and -1 a perfect inverse correlation.

$$\text{RC}_{m,s} = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{3.7}$$

Intuitively, RC represents how similar the two vectors are in how the elements are ranked. In a perfectly correlated set of vectors, the values may differ, but each element i has the same ranking in both vectors and thus each $d_i = 0$, leading to a perfect RC $= 1$.

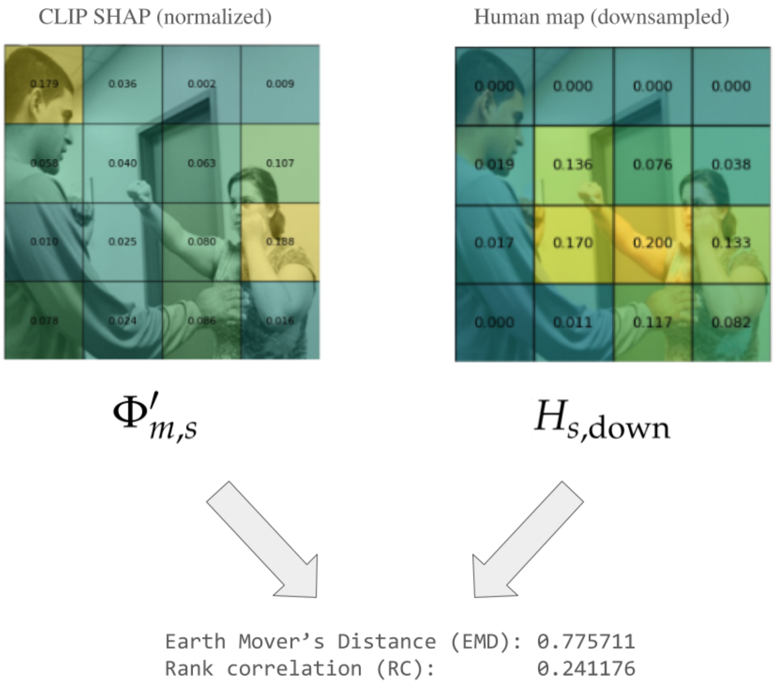
I implemented RC with SciPy's Spearman implementation `scipy.stats.spearmanr`[6].

3.2.1.1.3 Other metrics considered Two further metrics considered were Chi-Squared [7] and Kullback-Leibler divergence [8]. However, the first produced no discernible quantitative difference between randomly shuffled maps and related maps, while the second led to cases with infinite values. Both were discarded.

[6] *scipy.stats.spearmanr — SciPy v1.12.0 Manual* 2024

[7] *scipy.stats.chisquare — SciPy v1.12.0 Manual* 2024

[8] *scipy.stats.entropy — SciPy v1.12.0 Manual* 2024



```
Earth Mover’s Distance (EMD): 0.775711
Rank correlation (RC):        0.241176
```

Figure 3.9: An example of an EMD and RC calculation between a downsampled human saliency map and a normalized SHAP attribution map for CLIP.

3.2.1.2 Evaluation of metrics

To evaluate the appropriateness of each metric, RC and EMD, density distributions were created *for each model*, with two variants of the metric:

1. Similarity metric between SHAP and human map for *same* stimulus (this is how the metric is meant to work)
2. Similarity metric between a) the SHAP map for a given stimulus and model and b) a human map drawn from a random stimulus (after the human maps list was shuffled). Thus, each SHAP map is paired with most likely an **unrelated** human map.

The rationale for evaluating these two variants for each metric is as follows: for the metric to be practically useful, variant (1) should produce values notably different from variant (2). Otherwise, it would appear that the metric is either not measuring the correlation properly, or there is no correlation at all between *corresponding* human and SHAP attribution maps, since a randomly paired set of SHAP-human maps produces the same average SHAP-to-human correlation. [9]

Figures 3.10, 3.11, 3.12 and 3.13 show the distributions of the two metric variants above for both EMD and RC for each model's SHAP attributon maps and human saliency maps. We see that both EMD and RC produce, to varying degrees, a clear distribution that is distinct from random (note that for EMD smaller values imply greater similarity; for RC, greater values imply greater similarity).

A visual inspection of these graphs leads to the conclusion that EMD distances are generally *closer* to the randomized metric's distribution, while RC shows both a clear positive correlation between human and SHAP maps *and* is visibly distinct from its shuffled variant.

Whether the difference between the RC to human maps and the RC to *shuffled* human maps is statistically significant is the subject of RQ1, so we will not rigorously answer it here; regardless, a visual inspection of the graphs in Figures 3.10, 3.11, 3.12 and 3.13 suggests RC as the most promising metric for identifying similarities between SHAP and human maps. Thus, in the final analysis I generated a Spearman RC score between each SHAP map $\Phi'_{m,s}$ for stimulus s and each model m and its corresponding human saliency map $H_{s,\text{down}}$.

An $\mathbf{RC}_{m,s}$ score was generated for each stimulus s for which a human saliency map was available, so for 92 out of 99 stimuli. This score $\mathbf{RC}_{m,s}$ represents the *humanlikeness* of the attribution map for model m's output for stimulus s.

[9] A reader may object: Isn't one of the research questions *whether* there is a correlation at all between SHAP maps and human maps? Isn't choosing a metric based on whether it shows a correlation cherry-picking? My answer would be that not every metric is suitable to detect a difference, and I use visual intuition at this stage, in addition to a motivated choice of metric based on prior research, to determine which metric is most appropriate to detect a correlation.

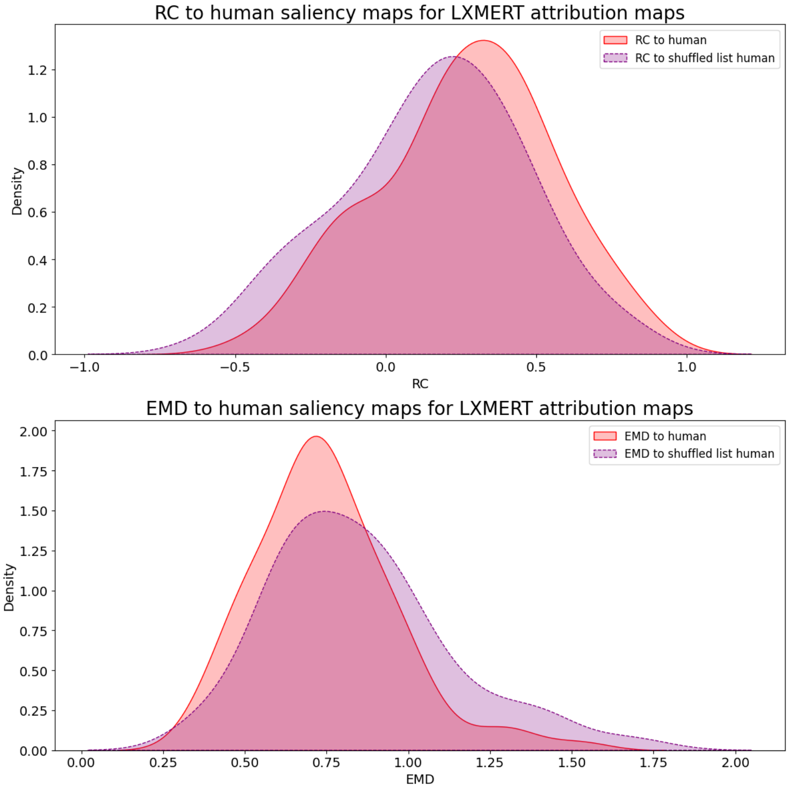


Figure 3.10: The distribution of RC and EMD variant values for LXMERT.

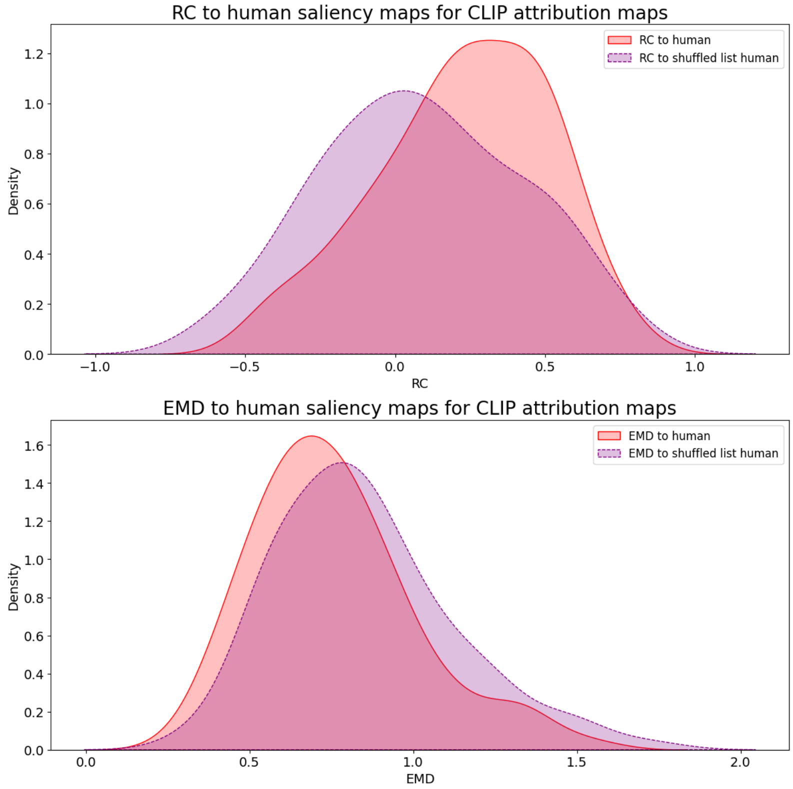


Figure 3.11: The distribution of RC and EMD variant values for CLIP.

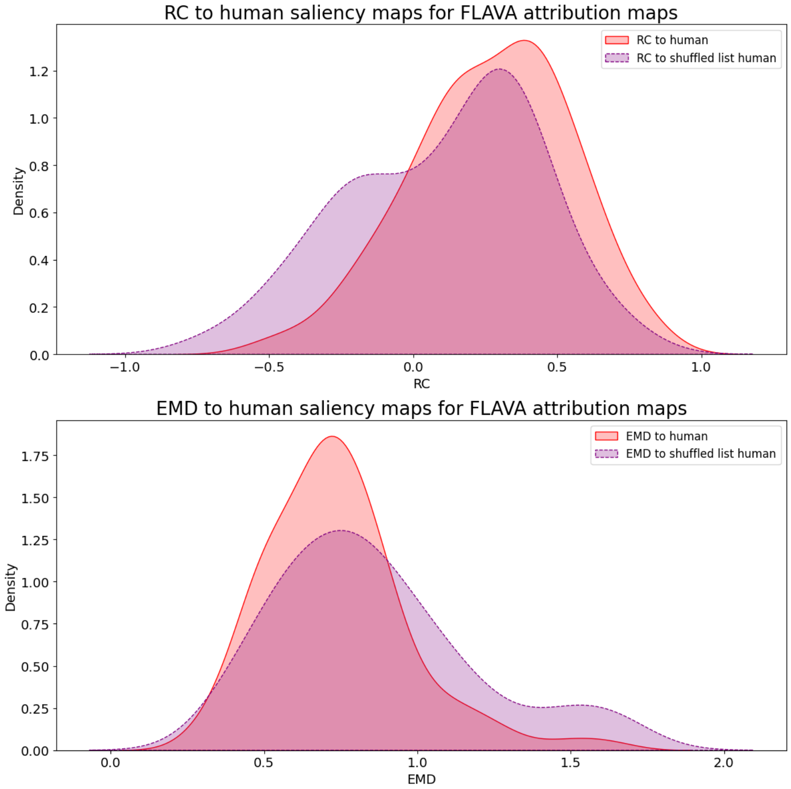


Figure 3.12: The distribution of RC and EMD variant values for FLAVA.

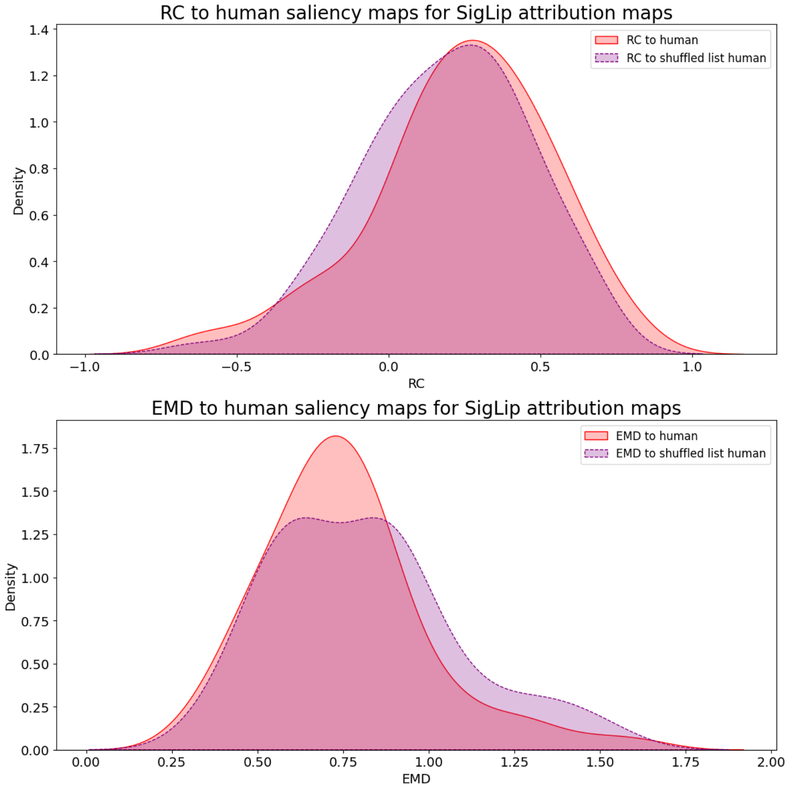


Figure 3.13: The distribution of RC and EMD variant values for SigLip.

3.2.2 Answering the research questions

What follows are the main research questions (posed in section 2.4), this time restated including the variables developed in the methodology so far:

1. **RQ1:** For each of the four models, is the average (across stimuli s) similarity $RC_{m,s}$ between each model m's attribution map $\Phi'_{m,s}$ and the corresponding human saliency map $H_{s,\text{down}}$ statistically significantly higher than the average similarity expected by random chance?
   - Null Hypothesis (**H01**): The average similarity $RC_{m,s,\text{AVG}}$ between each model's attribution maps and corresponding human saliency maps is not significantly greater than the average similarity that would be expected by chance.
   - Alternative Hypothesis (**HA1**): The average similarity $RC_{m,s,\text{AVG}}$ between each model's XAI attribution maps and the corresponding human saliency maps is significantly greater than what would be expected by chance.
2. **RQ2:** Does the similarity between a model m's XAI attribution maps and human saliency maps, $RC_{m,s}$, correlate with the model's performance, both on individual stimuli s for each model and in aggregate between models, on the VALSE benchmark?
   - Part 1: Within-model test
     - Null hypothesis (**H02.1**): There is no statistically significant correlation between a model's per-image similarity to human saliency maps $RC_{m,s}$, and its output on individual stimuli, $f_m(s)$.[10]
     - Alternative Hypothesis (**HA2.1**): There is a positive correlation between each model m's per-image similarity to human saliency maps, $RC_{m,s}$, and its output on individual stimuli, $f_m(s)$.

[10] In this part of the study, I take output score $f_m(s)$as a direct and more continuous indicator of performance than binarizing the performance into correct/incorrect.

- Part 2: Between-model test
  - Null hypothesis (**H02.2**): There is no statistically significant correlation between a model m's distribution of similarity scores $RC_{m,s}$ to human saliency maps and its overall accuracy on the VALSE benchmark.
  - Alternative Hypothesis (**HA2.2**): There is a positive correlation between a model m's distribution of similarity scores $RC_{m,s}$ to human saliency maps and its overall accuracy on the VALSE benchmark.

What follows is a discussion of the methodology for answering each research question.

3.2.2.1 RQ1 methodology

From the distribution of rank correlation (RC) scores visualized in section 3.2.1.2, it *appears* that the average rank correlation $\mathrm{RC}_{m,\mathrm{AVG}}$ between SHAP and equivalent human maps, averaged across stimuli s for a given model m, is significantly 1) greater than zero and 2) greater than the average correlation when the list of human maps is shuffled. This seems evident in each graph, but visual inspection does not prove significance. There are two tests I do to demonstrate the significance.

1. To test whether the average attribution humanlikeness scores ($\mathrm{RC}_{m,\mathrm{AVG}}$) for each model significantly deviates from zero, suggesting a meaningful similarity between SHAP and human-generated maps, we conducted a one-sample t-test against the null hypothesis that $\mathrm{RC}_{m,\mathrm{AVG}} = 0$ for each model.
2. However, even if we prove that the $\mathrm{RC}_{m,\mathrm{AVG}}$ for each model is (statistically significantly) greater than zero, there is a possible confound: it could be that on average *any* human map is, equally correlated with *any* SHAP map due to, for instance, a center bias in both types of maps. To rule this out, I also ran a kind of permutation test: For each model, I generated 10,000 shuffled variants of the dataset that shuffled the pairing of SHAP and human maps without changing the maps them-

selves. Here, I hypothesize that $\text{RC}_{m,\text{AVG}} > (\text{RC}_{m,\text{AVG}})_{\text{AVG, 10k variants}}$; that is to say, $\text{RC}_{m,\text{AVG}}$ for each model is significantly greater than the average such value across all permutations of the dataset.

The following are the exact steps I took for this second analysis:

1. Determine the average rank correlation between the 92 SHAP maps and their corresponding human maps, for each model m across stimuli s (this average was also needed for the one-sample t-test described above):

$$\text{RC}_{m,\text{AVG}} = \frac{\sum_{s}^{92\text{ stimuli}} \text{RC}_{m,s}}{92}$$

2. Generate 10,000 variants indexed by $i \in \{1, ..., 10,000\}$ of the dataset in which the order of the human maps is randomly shuffled in each case, such that a given human saliency map $H_{s,\text{down}}$ is replaced with some human map $H_{s(i),\text{down}}$, where $s(i)$ is the stimulus which replaces stimulus s in given dataset variant i.
3. For each shuffled dataset variant i, compute the $\text{RC}_{m,\text{AVG},i}$ value.
4. Calculate the proportion of $\text{RC}_{m,\text{AVG},i}$ values that is greater than the actual average rank correlation for the model, $\text{RC}_{m,\text{AVG}}$.

$$p = \frac{\#\{i | \text{RC}_{m,\text{AVG},i} \geq \text{RC}_{m,\text{AVG}}\}}{10,000}$$

This is a p-value representing the likelihood that the average human-likeness metric for model m, $\text{RC}_{m,\text{AVG}}$, came from the distribution of randomly shuffled datasets RC scores. If it is low enough, I can reject the null hypothesis and take it as significant that $\text{RC}_{m,\text{AVG, original dataset}} > (\text{RC}_{m,\text{AVG}})_{\text{AVG, 10k variants}}$.

3.2.2.2 RQ2 methodology

For this part of the analysis, I normalized the output scores $f_m(s)$ for each model by dividing it by that model's output's standard deviation $\sigma(f_m)$ on all of VALSE (validated actions, relations, existence pieces only). This leads to normalize output score $f'_m(s)$ for a given stimulus s and model m:

$$f'_m(s) = \frac{f_m(s)}{\sigma(f_m)}$$

This normalization has no impact on the statistical significance of the findings for each model, but is done to make comparisons between models more intuitive, since each architecture produces a wider or narrower spread of output values.

3.2.2.2.1 RQ2 part 1: Within-model test I evaluate the null hypothesis that *within* each model's performance data on the set of 92 stimuli (each denoted by s), the model attribution's humanlikeness scores (RC$_{m,s}$) correlate with the normalized model outputs for the same stimuli, $f'_m(s)$. The output $f'_m(s)$ is taken as representative of performance: the greater the value, the more confidently accurate the model is.

For each model, a separate rank-correlation calculation was performed between the list of RC$_{m,s}$ scores and the list of equivalent $f'_m(s)$ scores, generating a rank-correlation score ρ. Note that this is a completely separate use of rank-correlation as a statistical analysis metric, rather than a map similarity metric as earlier. The repeated use of rank correlation in the study for two *different* purposes is incidental and has no methodological import.

The reason for this choice of statistical test is that we are not interested in the numerical details of *how* RC$_{m,s}$ correlates with $f'_m(s)$ for a given model. Rather, we want to know *whether* it does. Rank correlation focuses on identifying a relationship without worrying about whether the relationship is linear, quadratic, or otherwise.

To test for statistical significance of the correlation for each model m between $\mathrm{RC}_{m,s}$ and $f'_m(s)$ across stimuli s, I conduct a Spearman rank correlation test with the SciPy library[11], which automatically generates a p value.

3.2.2.2.2 RQ2 part 2: Between-models test For this part of the analysis, I am interested in determining whether the distribution of $\mathrm{RC}_{m,s}$ humanlikeness scores for each model across stimuli has a relationship with the model's overall accuracy on the *actions*, *relations* and *existence* phenomena in the validated VALSE dataset.

I treat the $\mathrm{RC}_{m,s}$ values as a distribution, one for each model. I use a one-way ANOVA test to determine whether the mean rank correlations are significantly different *between* models. Here, a sufficiently low p value is required to reject the null hypothesis that the rank correlations are not different, on average, among the four models considered. If they *are* significantly different, the relationship between their average values and the aggregate accuracy of each model will be explored.

3.2.2.3 Significance threshold

To set the significance thresholds for this study, I divide the standard $\alpha = 0.05$ by the total number of statistical tests performed. I then apply the updated α uniformly to each statistical test.

I count the number of statistical tests as follows:

- For each model m in RQ1, I perform 2 statistical tests to see whether (1) the average humanlikeness $\mathrm{RC}_{m,\mathrm{AVG}} > 0$ and (2) the average humanlikeness $\mathrm{RC}_{m,\mathrm{AVG}}$ is significantly greater than the average across 10,000 permutations of the dataset, $(\mathrm{RC}_{m,\mathrm{AVG}})_{\mathrm{AVG,\ 10k\ variants}}$. This makes for a total of **8 statistical tests**.
- For each model m for RQ2.1, I perform a single rank-correlation test between its output values $f_m(s)$ and its humanlikeness scores $\mathrm{RC}_{m,s}$. This makes for a total of **4 statistical tests**.

[11] *scipy.stats.spearmanr — SciPy v1.12.0 Manual* 2024

- For RQ2.2, I perform a single one-way ANOVA test against the null hypothesis that all models m have statistically the equivalent $\mathrm{RC}_{m,\mathrm{AVG}}$.This makes for **1 statistical test**.[12]

Thus, I have a total of 13 statistical tests, meaning $\alpha = \frac{0.05}{13} = 0.0038$.

[12] If the reader is confused, I can "spoil" here that because no significant difference is found by this test, no additional tests are performed on individual differences between models, hence only one statistical test is done.

4. Results

The experimental results for each research question are provided below.

4.1 Research question 1

For **RQ1**, I found with high significance that we can reject the null hypothesis and conclude that the **average** humanlikeness score, $\mathrm{RC}_{m,\mathrm{AVG}}$, for each model m is:

1. significantly greater than 0 (no correlation), and
2. significantly greater than $\mathrm{RC}_{m,\mathrm{AVG}}$ for randomly shuffled variants of the dataset, which we can represent as $(\mathrm{RC}_{m,\mathrm{AVG}})_{\mathrm{AVG,\ 10k\ variants}}$.

The p values for each confirmed hypothesis per model are in Table 4.1.

The findings are also visualized in Figure 4.1: the left side of each graph shows the distribution of $\mathrm{RC}_{m,s}$ values for the unpermuted dataset for each model, as well as the mean value $\mathrm{RC}_{m,\mathrm{AVG}}$ as a dashed vertical line.

The right side of each graph shows the distribution of $\mathrm{RC}_{m,i,\mathrm{AVG}}$ across 10,000 permuted datasets (where i is one of these permutations), as well as

model	$\mathrm{RC}_{m,\mathrm{AVG}}$ (± standard error of means)	p-value for finding: $\mathrm{RC}_{m,\mathrm{AVG}} > 0$	p-value for finding: $\mathrm{RC}_{m,\mathrm{AVG}} > (\mathrm{RC}_{m,\mathrm{AVG}})_{\mathrm{AVG,\ 10k\ variants}}$
LXMERT	0.263 ± 0.030	< 0.0001 (*t*-statistic: 8.663)	<0.0001
CLIP	0.227 ± 0.030	< 0.0001 (*t*-statistic: 7.601)	<0.0001
FLAVA	0.266 ± 0.028	< 0.0001 (*t*-statistic: 9.354)	<0.0001
SigLip	0.236 ± 0.031	< 0.0001 (*t*-statistic: 7.622)	0.0005

Table 4.1: p-values for the two statistical tests in RQ1.

where the original $\mathrm{RC}_{m,\mathrm{AVG}}$ falls on this distribution. It is visually obvious that the $\mathrm{RC}_{m,\mathrm{AVG}}$ of the original dataset falls on the far edge of the distribution, which explains why the p-values are so low for each model.

Note that the distribution of humanlikeness scores for each stimulus can vary quite a lot, and many stimuli produce *negative* humanlikeness scores. My study only finds with high significance that the *average* humanlikeness for each model is significantly (1) greater than zero and (2) greater than what could be expected for a randomly shuffled dataset. The focus on ***averages*** leads to a significant result despite the wide spread of individual $\mathrm{RC}_{m,s}$ values.

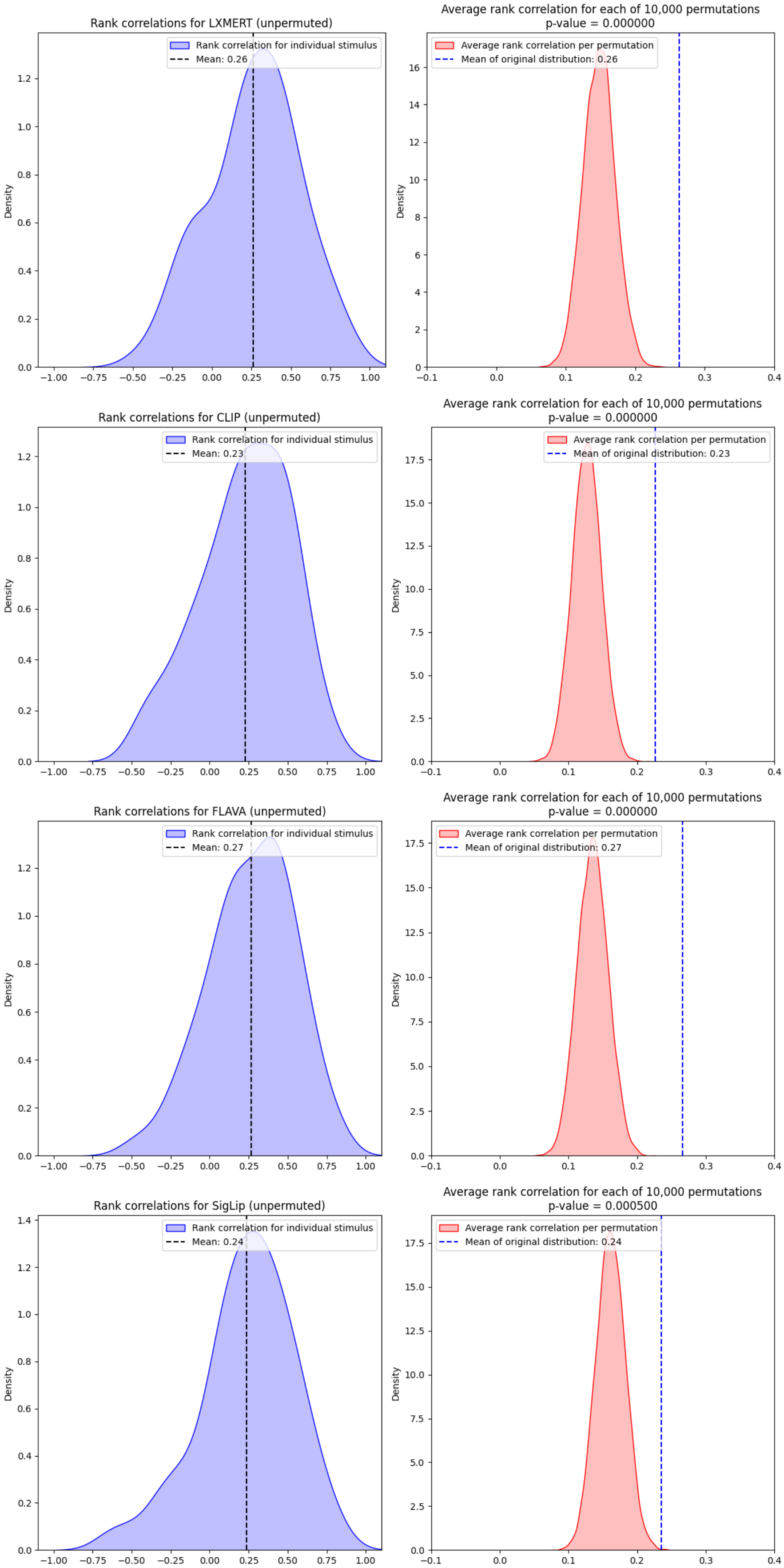


Figure 4.1: The left side of each graph shows the value density of $\text{RC}_{m,s}$ for the unshuffled dataset for each model, as well as the mean value $\text{RC}_{m,\text{AVG}}$ as a dashed vertical line. The right side of each graph shows the distribution of $\text{RC}_{m,i,\text{AVG}}$ across 10,000 permuted datasets, as well as again $\text{RC}_{m,\text{AVG}}$ as a dashed line.

4.2 Research question 2

For RQ2, the study failed to reject either null hypothesis (H02.1 and H02.2).

4.2.1 RQ2: Part 1

The outcome of the rank correlation tests between model attribution humanlikeness per stimulus s (RC$_{m,s}$) and model output for stimulus s ($f'_m(s)$, σ-normalized) is seen in Figure 4.2, including rank-correlation coefficients (ρ, referred to in the graphs as "Spearman correlation") for each model and the associated p-value. (Again, note that this "Spearman correlation" is completely distinct from the rank correlation equation previously used to compute the human-model attribution similarity scores, and serves a distinct purpose.

There was no statistically significant positive rank correlation within any of the four models between the model humanlikeness (RC$_{m,s}$) and the model output ($f'_m(s)$) across stimuli. For SigLip, there is a *negative* correlation with p value 0.02, but with the Bonferroni-corrected α of 0.0038, that p-value is too high to be statistically significant.

4.2.2 RQ2: Part 2

The distribution of RC$_{m,s}$ scores on a model-by-model basis is seen in Figure 4.3, which on the x-axis depicts the model's overall average accuracy (a straight average between the accuracy over the three linguistic phenomena).

The visual intuition that these distributions are not significantly different from one another is confirmed by a one-way ANOVA test, which leads to a p-value of 0.73, meaning it is highly likely that the four RC$_{m,\text{AVG}}$ values come from the same population.

Without significant differences in RC$_{m,s}$, or its average RC$_{m,\text{AVG}}$, between models, analyzing the relationship between RC$_{m,\text{AVG}}$ and overall model quality is pointless.

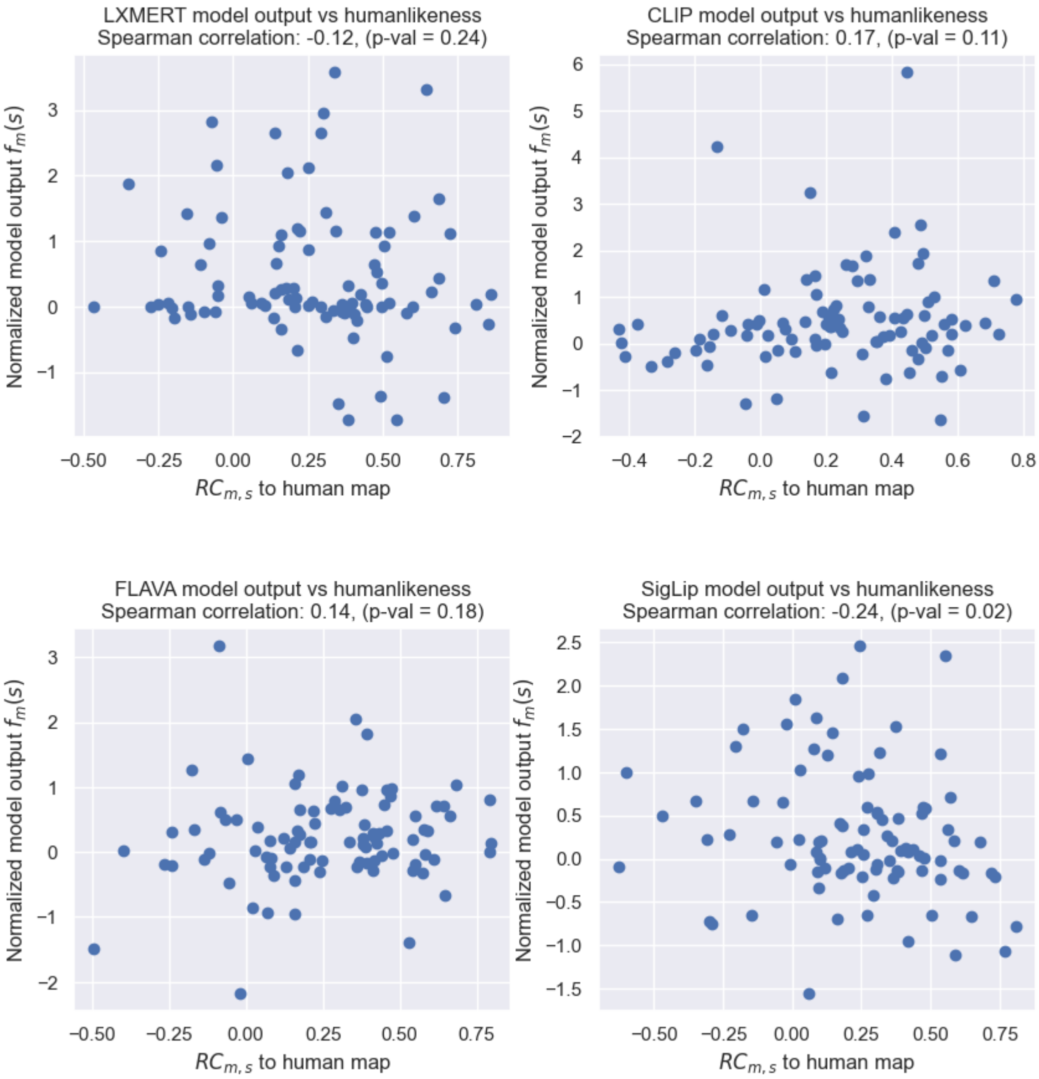


Figure 4.2: The Spearman rank-correlation coefficients (plus p-value) for each model are printed above the scatter-plot. This represents the rank correlation between the x-axis $\mathrm{RC}_{m,s}$ similarity score and the y-axis normalized model output ($f'_m(s)$) for each stimulus s.

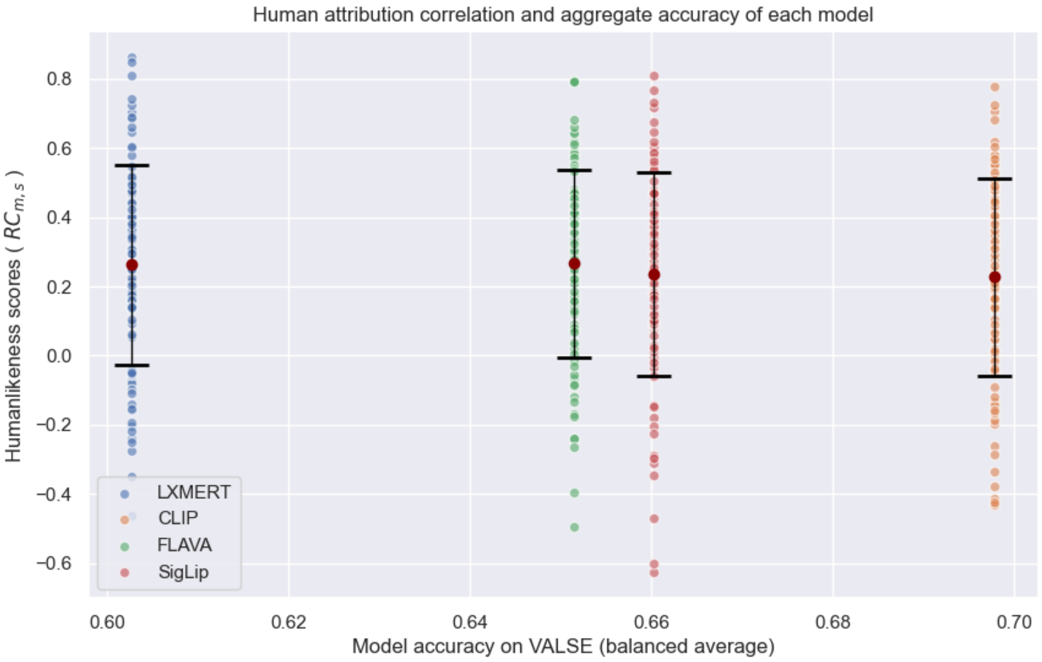


Figure 4.3: The distribution of $\mathrm{RC}_{m,s}$ for each model m is plotted in this scatterplot, where the x-axis represents each model's overall accuracy on the three linguistic phenomena used from VALSE.

5. Discussion

5.1 What the results say

The statistical confirmation of the hypothesis for **RQ1** allows me to conclude that the average humanlikeness score $\mathrm{RC}_{m,\mathrm{AVG}}$ between SHAP and human maps for each model is the result of a statistically significant relationship indexing SHAP maps to human maps, which is virtually impossible to be replicated by shuffling the pairings of the maps randomly.

An additional vindication of the RC similarity scores calculated here is that they closely track the rank correlations found in A. Das, Agrawal, et al. 2017 for two VLM models on another task: there, the models had average XAI-human map rank correlations ranging from 0.249 to 0.264. (Mine range from 0.227 to 0.266.)

This means the outcome of RQ2 becomes more meaningful as well, because the humanlikeness metric used in that analysis, $\mathrm{RC}_{m,s}$, is on average indicative of a genuine, non-random alignment between the models' attribution maps and the corresponding human saliency maps.

The fact that the $\mathrm{RC}_{m,s}$ metric appears to be meaningful strengthens the validity of the results for **RQ2**, which lead to the conclusion that at least this method of measuring model humanlikeness (comparing SHAP attribution maps to human saliency through rank correlation) does *not* correlate with performance on VALSE, whether across stimuli for a single model, or between models.

5.2 What the results mean

My findings vindicate my A. Das, Agrawal, et al. 2017-inspired method of generating task-specific human saliency maps. However, the intuition that

that a humanlikeness metric based on these saliency maps should itself correlate with performance on VALSE has not been confirmed.

Had it been confirmed, it may have suggested, on a more conceptual level, that there is some kind of quality called "grounding" that is related to, and affects, *both* performance and humanlikeness. As it stands, this remains speculative in the context of this experiment.

Ultimately, this experiment and its results demonstrate some of the shortcomings of a subjective measure like "attribution humanlikeness" as in this case. On an ontological level, it is far from guaranteed that downsampled SHAP maps and human saliency maps from my web interface are an "apples-to-apples" comparison, and that the humanlikeness score this produces measures the kind of grounding that VALSE also gets at.

In the end, when we compare this humanlikeness score with a more performance-based evaluation approach like the VALSE benchmark, a kind of epistemological hierarchy emerges. A correlation with VALSE performance would have vindicated the humanlikeness score in this study. Now, the absence of a correlation leaves its status indeterminate, while not meaningfully undermining the validity of VALSE as a benchmark.

The strength of a benchmark like VALSE lies in the fact that it is in some sense not "just" a measure of raw performance; its challenges are designed to match human-defined categories of comprehension. Thus, without directly using XAI methods, this granular performance-based approach gets at some understanding of what human concepts the model represents. This implies that benchmarks like VALSE might still offer indirect insights into models' conceptual alignment with human thought, despite not directly drawing on humanlikeness metrics derived from attribution maps or other XAI methods.

5.3 Study limitations

I did my best in the methodology design stage to avoid some of the pitfalls of this study, such as an unbalanced or poor-quality dataset, or a meaningless human-AI comparison metric. Even so, some limitations remain:

- Despite the VALSE filtering process documented in Appendix B, some of the final stimuli that "made the cut" continue to have obvious issues. The stimulus whose foil is "A toilet pees on a woman" is certainly not helping the quality of either the human or model attribution data. A more rigorous data filtering might have improved this issue somewhat, though most stimuli are unproblematic and this is unlikely to be the decisive factor leading to a lack of significance in RQ2.
- The model-agnostic SHAP approach reduced the question of attribution to a single 4x4 matrix for each model and stimulus; the lack of granularity and the one-size-fits-all nature of this approach, while helping make the methodology scalable, has obvious drawbacks for the quality of the output data.
- The human data collection process was open to abuse by bad-faith actors who just wanted to finish the study quickly; subjectively, there were individual human saliency maps I saw which struck me as nonsensical on a case-by-case basis, but there was no way to "prove" they should be discarded. More subjects in the future could cancel out this issue.

5.4 Future work

The humanlikeness metric $\mathrm{RC}_{m,s}$ is drawn from a very specific kind of attribution map, which looks only at the visual modality. Other, more model-specific methods could shed more light on different and more subtle ways a model may or may not correspond to "humanlike" cognition (e.g. Cao et al. 2020).

Attention-based mechanisms which examine VLM attention in the visual modality (rather than model-agnostic attribution) could be a useful, more dissection-based XAI counterpart to the human data collection interface as I developed it here. Sacrificing model-agnosticity and scalability for a more granular approach to one or two models could end up being a more fruitful approach.

Ultimately, creating a rigorous, empirically validated metric to evaluate the groundedness of *how* a model does what it does, regardless of performance, continues to be a worthwhile pursuit.

A. Appendix: Details of models and their implementation

More details about the four models used in this study are given in this appendix. Their properties are summarized in Table 3.1.

A.1 LXMERT

Published in 2019, the LXMERT model represented a significant step forward in multimodal representation. Its architecture first encodes the text and image input separately with transformer architectures, and then passes the resulting representations through a cross-modal encoder. It was trained on five tasks that incorporate both modalities to varying extents: "(1) masked cross-modality language modeling, (2) masked object prediction [...], (3) masked object prediction via detected-label classification, (4) cross-modality matching, and (5) image question answering" (Tan and Bansal 2019, p. 2). A key feature is that at the image encoding stage, it represents the image as a set of detected objects and their positions, while text is represented through "index-aware word embeddings".

LXMERT was used as one of the key models used in both the original VALSE study (Parcalabescu, Cafagna, et al. 2022) and the more recent **bugliarellf_measuring_2023** benchmarking study, where LXMERT achieved the worst performance on VALSE (though still better than chance) out of all models tested, with an accuracy of 59.6%.

Given an image resized to a square (as in all models in this study) and a caption and foil, LXMERT outputs, among other quantities, two "cross relationship scores" for each text-image pair: one for the likelihood of the text and image matching, and the other for the likelihood of the text and image *not* matching. We pass these scores through a softmax filter per caption to

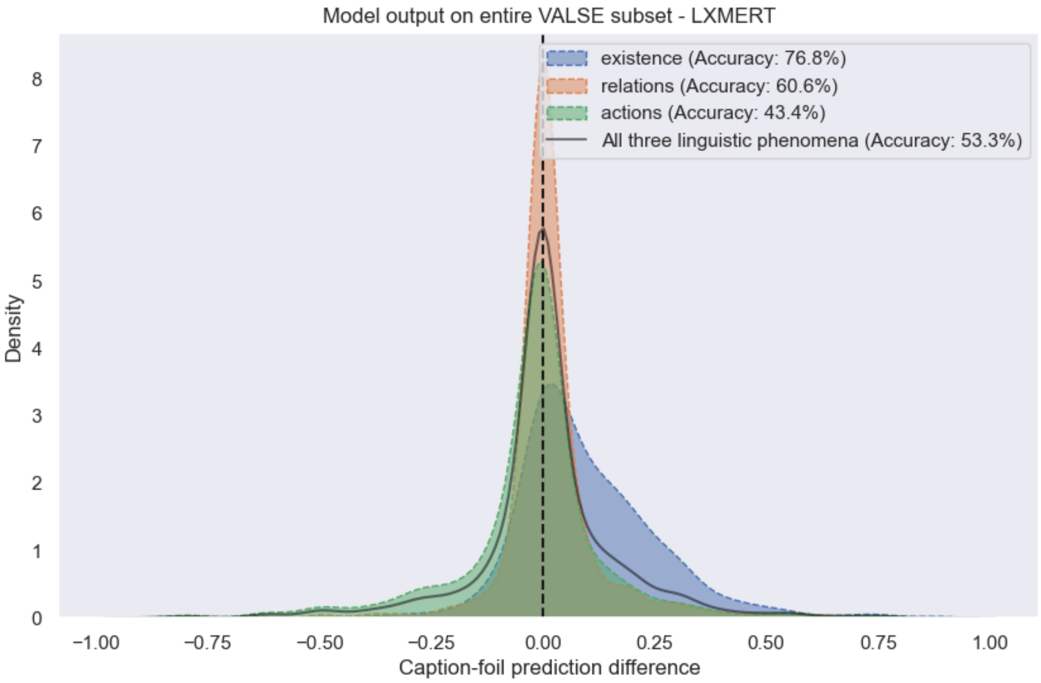


Figure A.1: Model output f (on the x axis) distribution for LXMERT.

generate the following quantities:

- Probability of caption being correct, probability of caption being incorrect, adding up to 1: $P_{\text{caption}}, P_{\text{caption incorrect}}$
- Probability of foil being correct, probability of foil being incorrect, adding up to 1: $P_{\text{foil}}, P_{\text{foil incorrect}}$

I then take the difference between the likelihood of the caption P_{caption} and the foil P_{foil} as the output variable $f_m(s)$ for stimulus s. If this value is positive, the model gets it “right.” If this value is greater in magnitude, the model is *more* confident in its choice.

In Figure A.1, we can see the density distribution of the scalar LXMERT output $f_{\text{LXMERT}}(s)$ (on the x axis) for the stimuli s in the validated subset of VALSE (so not just the selected stimuli) for each of the three linguistic phenomena in our study and overall, as well as the accuracy rate for each. Values where the x-axis is above 0 represent correct judgements.

A.2 CLIP

Published by OpenAI in 2021 (Radford et al. 2021), CLIP is a widely-used VLM featured in the original VALSE study (Parcalabescu, Cafagna, et al. 2022). CLIP uses two separate transformer-based architectures to encode, respectively, images and captions, creating a vector representation for each. Rather than using cross-attention directly, CLIP was trained with a contrastive objective, aiming to keep correct image-caption pairs close by in the vector space, while incorrect pairs were kept more distant; however, this objective is accomplished without creating a joint representation.

Aside from the VALSE study, CLIP also featured in the more recent Bugliarello, Sartran, et al. 2023 paper that included the VALSE benchmark; there, it achieved a middle-of-the-road overall VALSE accuracy of 64.0%.

Given an image and two captions, CLIP produces a raw logit output for each image-caption pair, which represents the image-text similarity score for the caption and foil. We take the difference between these values as the scalar output $f_{\text{CLIP}}(s)$; if it is positive, the caption has a higher score and the model gets it right; greater difference magnitudes suggest increased confidence.

Figure A.2 shows the density distribution of the scalar CLIP output for the validated subset of VALSE for each of the three linguistic phenomena in our study and overall, as well as the accuracy rate for each.

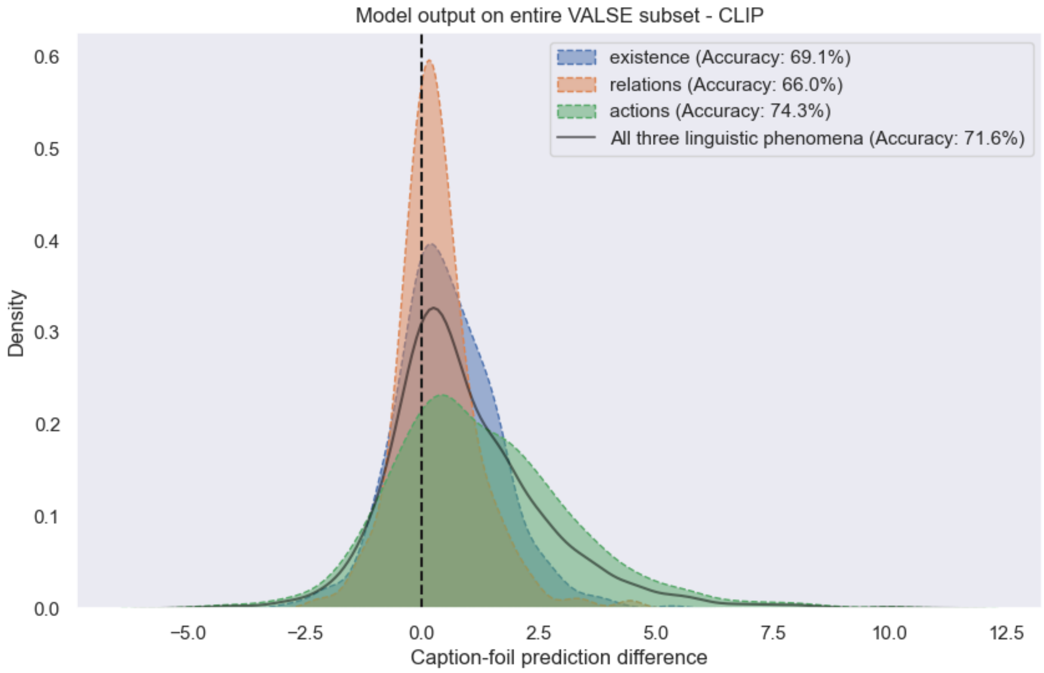


Figure A.2: Model output f (on the x axis) distribution for CLIP.

A.3 FLAVA

Published in 2022, the FLAVA model aims to combine the strengths of contrastively-trained dual-encoder models like CLIP and cross-modal fusion models like LXMERT. It uses pretraining objectives generally associated with both of these categories. It is also designed to be useful for unimodal contexts in addition to the multimodal tasks for which CLIP and LXMERT were built. The researchers behind FLAVA aim for "a foundational language and vision representation that enables unimodal vision and language understanding as well as multimodal reasoning, all within a single pre-trained model" (A. Singh et al. 2022).

Architecturally, FLAVA resembles LXMERT more than CLIP: It first encodes the text and image modalities separately, then passes these encodings into a multimodal encoder that is used as the basis for a final representation. However, unlike LXMERT, FLAVA directly calculates "contrastive logits" between the input text and images, which are equivalent to the CLIP output logits, with higher values representing greater similarity.

As in CLIP, we take FLAVA's scalar output f_{FLAVA} to be the difference

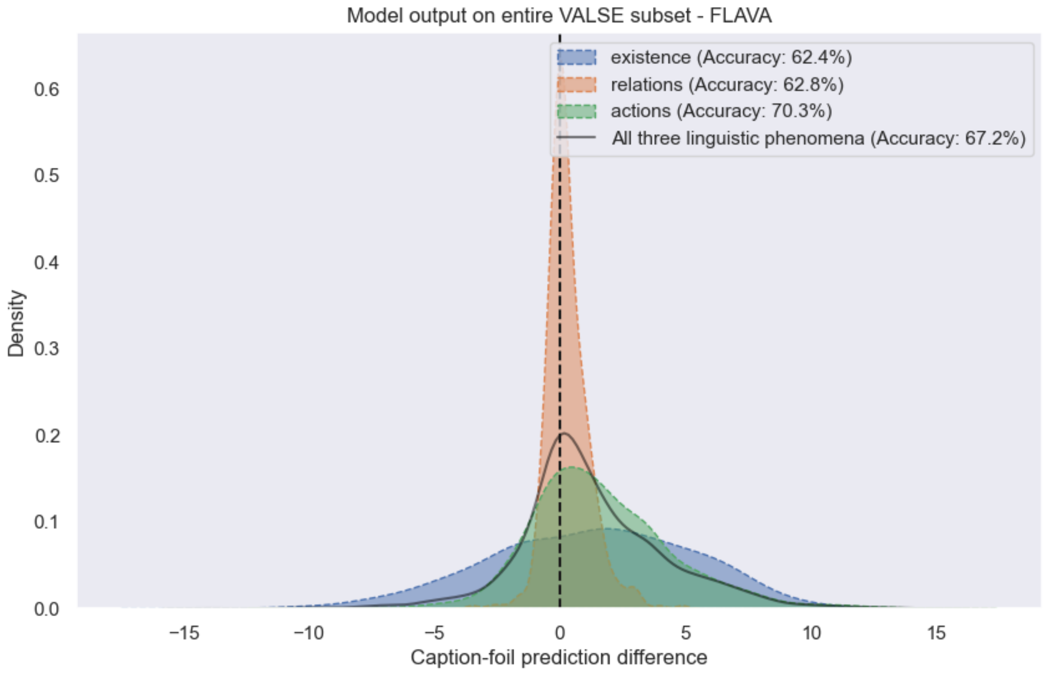


Figure A.3: Model output f (on the x axis) distribution for FLAVA.

between the similarity scores for a given stimulus s, the caption score minus the foil score. A positive value means the caption was chosen over the foil.

Figure A.3 shows the density distribution of the scalar FLAVA output for the validated subset of VALSE for each of the three linguistic phenomena in our study and overall, as well as the accuracy rate for each.

A.4 SigLip

Published by Google researchers in 2023, SigLip (Sigmoid Loss for Language Image Pre-Training) is the newest model used in my study (Zhai et al. 2023). It is CLIP-inspired in its architecture, but according to its authors achieves an improvement in performance while also employing a more efficient training strategy. Even though it still employs contrastive learning, it uses a sigmoid loss to assess similarity on the image-text pair level, and unlike the CLIP pretraining, does not normalize this similarity with softmax across the entire dataset.

As in CLIP, the scalar output f_{SigLip} is the difference in image-text similarity logit scores.

Note that on the three sub-parts of VALSE used in this study, SigLip actually performs worse than CLIP in *my* implementation. This is visible in Figure A.4, which depicts the density distribution of the scalar SigLip output for the validated subset of VALSE for each of the three linguistic phenomena in the present study and overall, as well as the accuracy rate for each. (Compare to CLIP's performance as depicted in Figure A.2.

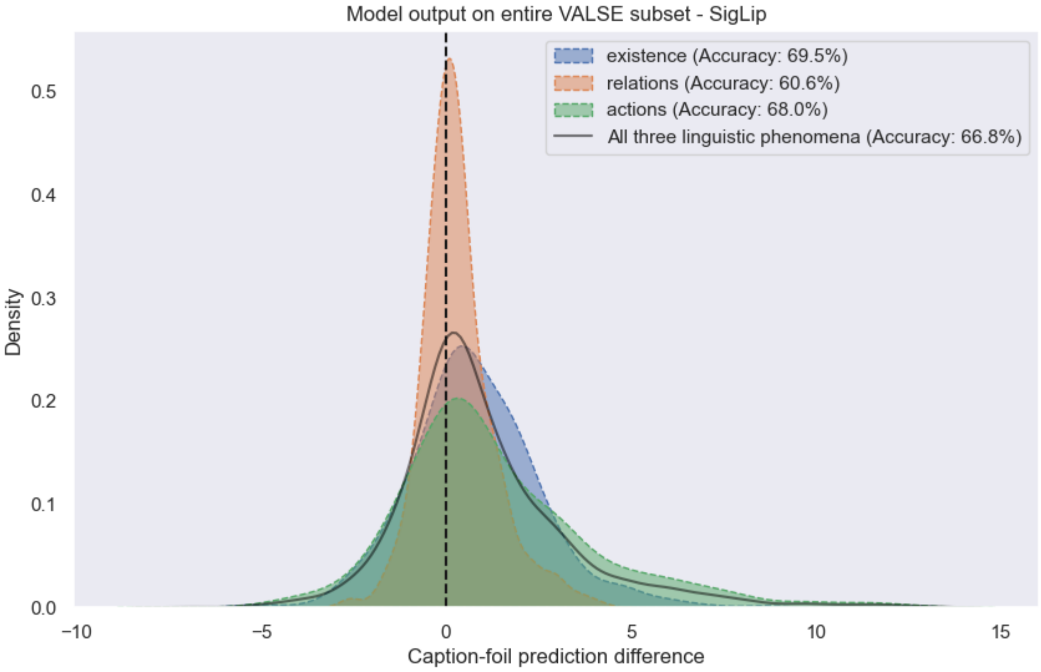


Figure A.4: Model output f (on the x axis) distribution for SigLip.

B. Appendix: Details of experimental stimuli generation

This appendix describes how the final set of 99 stimuli used to generate attribution maps for each model was selected and prepared for the experiment.

B.1 Initial sampling of the VALSE dataset

I restricted the study to those entries in the VALSE dataset where the majority of MTurk annotators to the original dataset successfully picked the caption as the correct label. This same validation criterion was used by the VALSE researchers to validate their data.

Another consideration is that I wanted to make sure the examples used in my study come from a range of difficulty levels. In order to do this, I first ran elements of the VALSE dataset through the CLIP model (Radford et al. 2021). This generated an image-text score for each image's caption and foil. The difference between these scores (the CLIP prediction difference) is taken as an indication of the difficulty of the question for CLIP. The higher the caption score is compared to the foil, the easier it was for the model to choose the correct caption. The graph in Figure B.1 shows a histogram of this score difference across the three "pieces" considered here. A scatterplot for the same data is shown in Figure B.2.

This part of the data sampling, using this distribution of performances, proceeded in the steps shown in Figure B.3, with the goal of producing a final subsample that was balanced between different CLIP output scores.

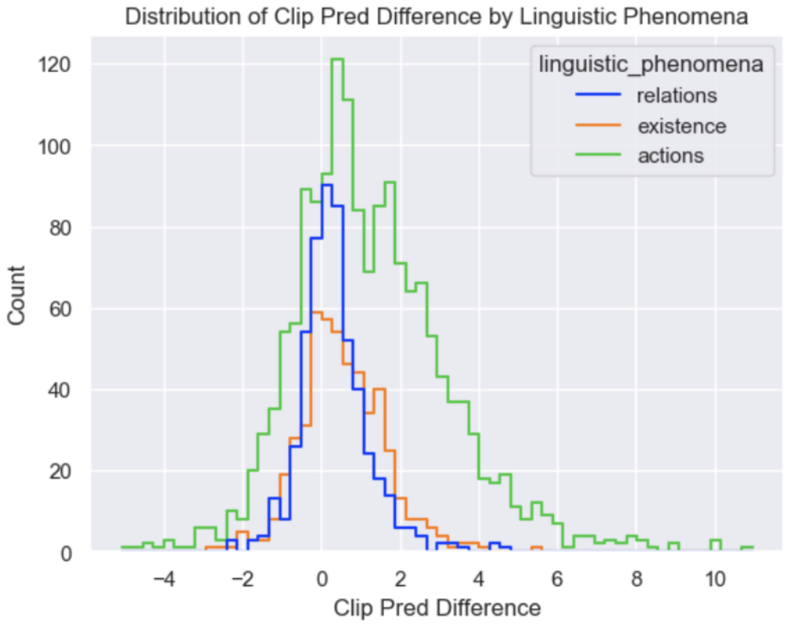


Figure B.1: CLIP prediction difference across validated instances of VALSE dataset for three linguistic phenomena.

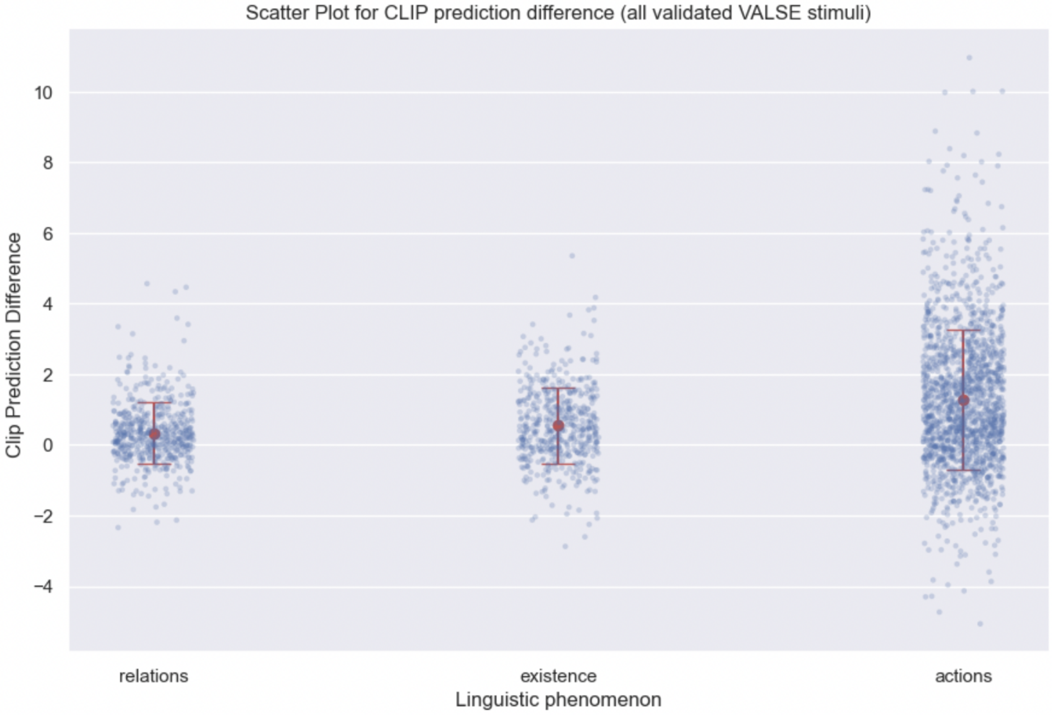
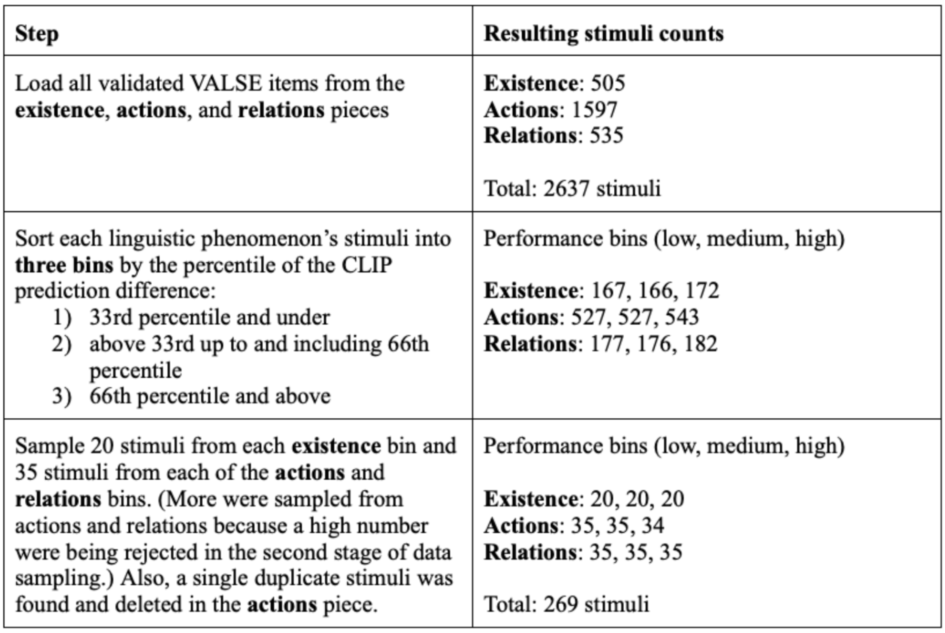


Figure B.2: A scatterplot (with false horizontal displacement) depicting the range of CLIP output values for each linguistic phenomenon considered in the study.



Step	**Resulting stimuli counts**
Load all validated VALSE items from the **existence**, **actions**, and **relations** pieces	**Existence**: 505 **Actions**: 1597 **Relations**: 535 Total: 2637 stimuli
Sort each linguistic phenomenon's stimuli into **three bins** by the percentile of the CLIP prediction difference: 1) 33rd percentile and under 2) above 33rd up to and including 66th percentile 3) 66th percentile and above	Performance bins (low, medium, high) **Existence**: 167, 166, 172 **Actions**: 527, 527, 543 **Relations**: 177, 176, 182
Sample 20 stimuli from each **existence** bin and 35 stimuli from each of the **actions** and **relations** bins. (More were sampled from actions and relations because a high number were being rejected in the second stage of data sampling.) Also, a single duplicate stimuli was found and deleted in the **actions** piece.	Performance bins (low, medium, high) **Existence**: 20, 20, 20 **Actions**: 35, 35, 34 **Relations**: 35, 35, 35 Total: 269 stimuli

Figure B.3: The initial stage of VALSE stimuli filtering. Note that the CLIP prediction difference is the f_{CLIP} variable defined in section A.2.

B.2 Sampling the final 99 stimuli

A number of captions and foils in the 269 initially sampled stimuli (see previous section) had issues that needed to be corrected before being usable in the final study. The edits made, and their rationales, are given in Figure B.4.

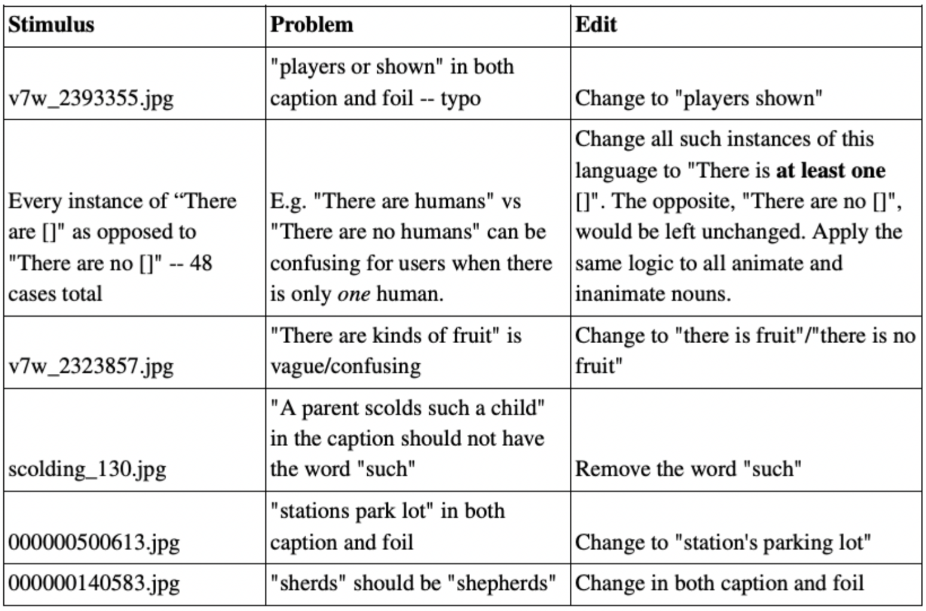
Having arrived at the set of 269 stimuli and edited the captions and foils as summarized in Figure B.4, I next eliminated those of the set that remained less suitable for the final experiment.

B.2.1 Eliminating unsuitable stimuli

Here, I took separate approaches for the actions/relations and existence pieces.

B.2.1.1 Actions and relations stimuli

For the actions and relations stimuli, the issues stem from the fact that the VALSE foils, which were generated by modifying the captions, are sometimes implausible or semantically invalid. This issue was widespread enough that in the initial data sampling, I sampled more stimuli from the actions



Stimulus	**Problem**	**Edit**
v7w_2393355.jpg	"players or shown" in both caption and foil -- typo	Change to "players shown"
Every instance of "There are []" as opposed to "There are no []" -- 48 cases total	E.g. "There are humans" vs "There are no humans" can be confusing for users when there is only *one* human.	Change all such instances of this language to "There is **at least one** []". The opposite, "There are no []", would be left unchanged. Apply the same logic to all animate and inanimate nouns.
v7w_2323857.jpg	"There are kinds of fruit" is vague/confusing	Change to "there is fruit"/"there is no fruit"
scolding_130.jpg	"A parent scolds such a child" in the caption should not have the word "such"	Remove the word "such"
000000500613.jpg	"stations park lot" in both caption and foil	Change to "station's parking lot"
000000140583.jpg	"sherds" should be "shepherds"	Change in both caption and foil

Figure B.4: How stimuli in the VALSE dataset were edited.

and relations pieces to ensure there would be enough validated stimuli left over.

For instance, in the actions piece, there is a stimulus with caption "A man chisels a metallic element" and matching foil "A metallic element chisels man." The foil here is so implausible as to be absurd, so neither human subjects nor AI models would really need to see the image to eliminate the foil from consideration. For the existence piece, the risk of such implausible foils is far lower, because all captions in the existence category are of the form "There are/is. . . " or "There are/is no. . . ". The presence and absence of some object in a picture are generally equally plausible.

To reduce implausible stimuli, I manually labeled all the sampled actions and relations stimuli with "approve", "reject", or "unsure." The rejections are meant to be obvious cases where the foil makes no sense, while the "unsures" capture those cases where the foil is implausible but not impossible. While the process relies partly on subjective judgment, labeling problematic captions can help improve the quality of the study, and the full list of stimuli marked in this way, including rejection rationales, is available in the

	Approve	Reject	Unsure	Total
relations	61	5	39	105
actions	41	27	36	104
Total	102	32	75	209

Table B.1: Actions and relations approve/reject/unsure records.

VALSE piece	CLIP output group	approval	#
actions	high_perf	Approve	8
		Unsure	4
	low_perf	Approve	8
		Unsure	4
	medium_perf	Approve	6
		Unsure	3
relations	high_perf	Approve	8
		Unsure	4
	low_perf	Approve	8
		Unsure	4
	medium_perf	Approve	6
		Unsure	3

Table B.2: Final counts for actions and relations pieces, 33 stimuli each.

project's GitHub repository for data[1]. The final counts of labeled stimuli by linguistic phenomenon are given in Table B.1.

To prevent this selection from biasing the data in a specific direction (e.g. making the challenges more difficult or easier for CLIP), I sampled from the labeled *relations* and *actions* examples based on the following requirements to produce 33 of each.

1. All "rejected" images were rejected
2. A 2:1 ratio was maintained between "approved" and "unsure" examples
3. Equal representation for high, medium, and low-performing examples (based on CLIP outputs calculated earlier)

This led to the counts in Table B.2.

[1] URL: https://github.com/skshvl/thesis-data-public/

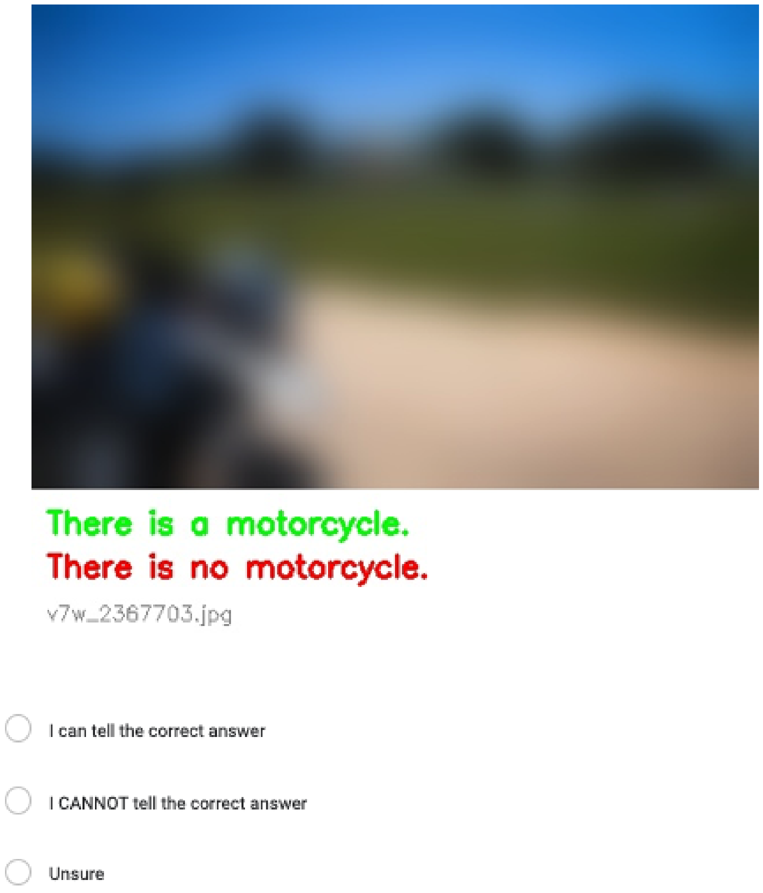
B.2.1.2 Existence stimuli

The 60 *existence* stimuli sampled in the initial sampling process from VALSE did not have plausibility issues on the captions or foils. They did, however, pose a different risk in light of the experimental design. This is because the human data collection relies on humans seeing blurred versions of each stimulus image, and selectively unblurring it to identify the correct caption. In discussions with my supervisor, the risk emerged that because existence stimuli only require identifying simply whether something is absent or present in the image, particularly many of the existence stimuli may be resolvable without un-blurring anything at all. This would likely lower the informativeness of the human saliency maps.

To reduce the risk of such too-easy (when blurred) stimuli in the existence piece, I created a simple survey showing a blurred version (the same blur level as in the final experiment) of each of the 60 existence stimuli sampled from VALSE, alongside the caption and foil. The survey then simply asked whether the user was able to identify the correct caption, including an option to answer “Unsure”. A screenshot of the survey is seen in Figure B.5.

A group of 5 adults including my research group and acquaintances completed this survey. The results were then sorted into bins by stimulus, to assemble the best possible quality group of 33 final *existence* stimuli. Figure B.6 summarizes the different groups of stimuli based on this survey, ranked from most to least desirable, including whether they were included in the final 33 or not.

This resulted in the distribution of "existence" stimuli by performance groups for CLIP output seen in Table B.3. While this distribution is less balanced than the final counts for *relations* and *actions* (Table B.2), it still shows a healthy mix of different difficulty levels for CLIP.



There is a motorcycle.

There is no motorcycle.

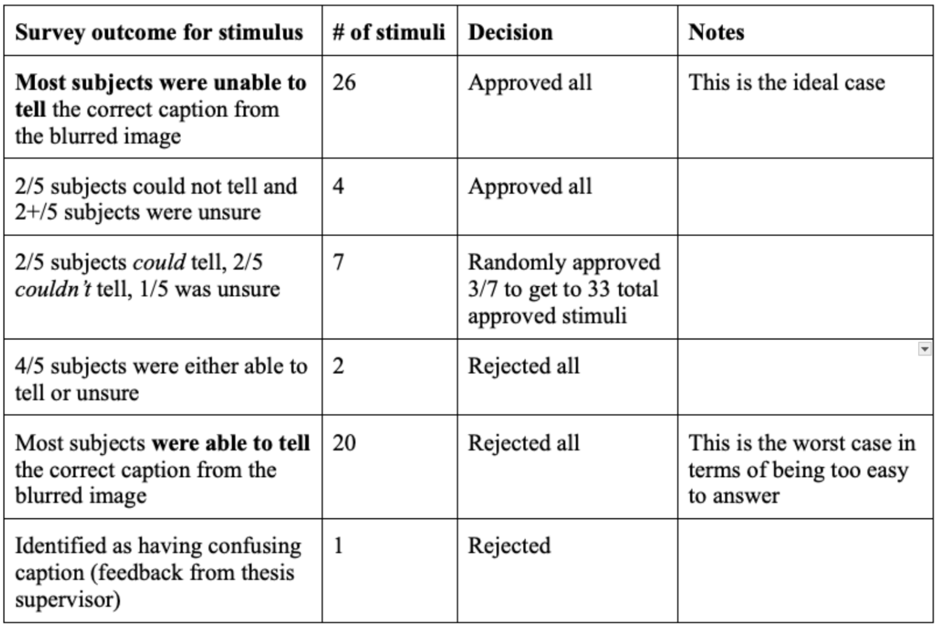
v7w_2367703.jpg

○ I can tell the correct answer

○ I CANNOT tell the correct answer

○ Unsure

Figure B.5: A screenshot from a small-scale survey on the *existence* piece.



Survey outcome for stimulus	**# of stimuli**	**Decision**	**Notes**
Most subjects were unable to tell the correct caption from the blurred image	26	Approved all	This is the ideal case
2/5 subjects could not tell and 2+/5 subjects were unsure	4	Approved all	
2/5 subjects *could* tell, 2/5 *couldn't* tell, 1/5 was unsure	7	Randomly approved 3/7 to get to 33 total approved stimuli	
4/5 subjects were either able to tell or unsure	2	Rejected all	
Most subjects **were able to tell** the correct caption from the blurred image	20	Rejected all	This is the worst case in terms of being too easy to answer
Identified as having confusing caption (feedback from thesis supervisor)	1	Rejected	

Figure B.6: Outcome of the small-scale survey on "existence" stimuli.

linguistic phenomenon	performance group	Selection
existence	high_perf	9
	low_perf	10
	medium_perf	14

Table B.3

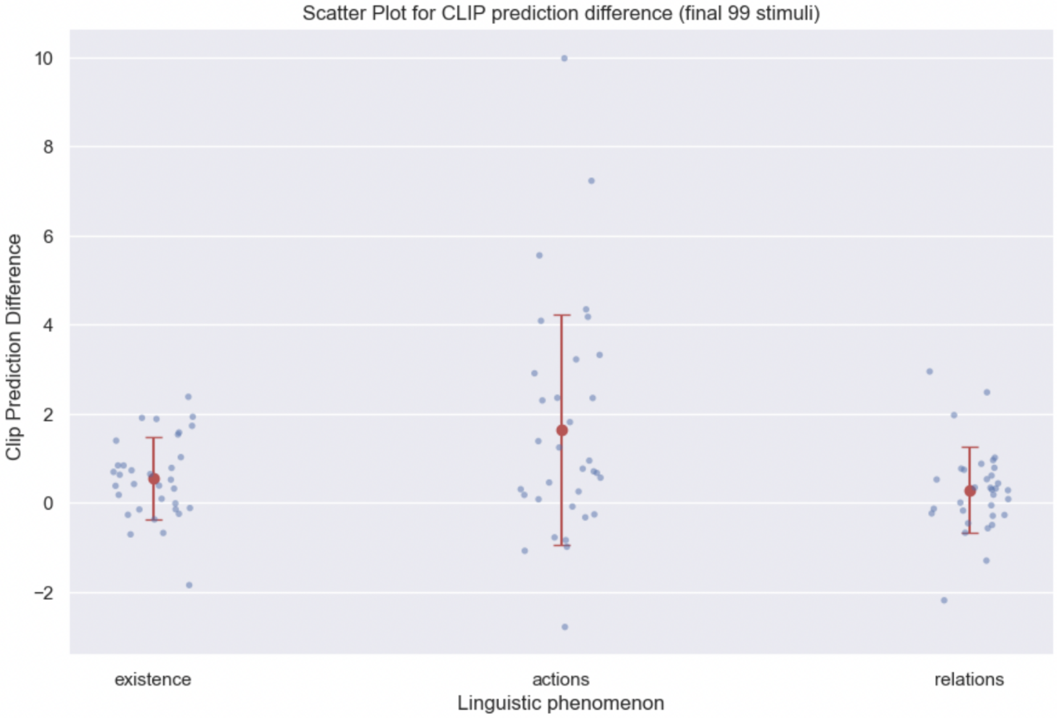


Figure B.7: A scatterplot of CLIP output values for the 99 stimuli. Compare to Figure B.2

.

B.3 Summary

For a final look at the sampled stimuli, 33 for each of the *actions*, *relations*, and *existence* linguistic phenomena, we can take a look at Figure B.7. It depicts the distribution of CLIP prediction difference values for each linguistic phenomenon, this time only for the final selected stimuli (including mean and standard deviation). A visual comparison to a similar graph for the entire validated VALSE dataset (Figure B.2) shows that our final sample is a good representation of the range of values in the unfiltered dataset.

The final number of stimuli (99) was chosen because it would result in close to this many aggregate human saliency maps, which could then be

Linguistic phenomenon	Performance group counts for final CLIP implementation (low, middle, high)
Existence	9, 13, 11
Relations	12, 9, 12
Actions	12, 9, 12

Table B.4: Performance group counts updated for CLIP implementation and outputs after dataset changes and re-implementation of CLIP.

used to calculate as many distance metrics to XAI attribution maps. We expect close to 100 individual distance metrics per AI model to be sufficient to test for a significant statistical relationship between these distance metrics and that model's output (as a measure of accuracy).

B.4 Revisiting dataset balancing after study conclusion

Note that the CLIP prediction difference used to balance the dataset came from an initial implementation of CLIP that was based on code found in the MMSHAP GitHub repository (Parcalabescu and Frank 2023). For the final analysis of SHAP maps and CLIP performance on each stimulus, a *different* implementation of CLIP was used. Additionally, some of the data was edited as described above, and these edits were made after the initial CLIP scores were generated. Likely due to small differences in implementation (such as resizing) and these differences in the input data, the scores in the output of my *final* CLIP implementation were *modestly* different from these initial outputs used to balance the 99-stimuli dataset. However, after re-running the same analysis as above with this final CLIP implementation, sorting the stimuli's CLIP outputs into performance groups based on updated thresholds derived from the larger VALSE dataset, the 99-stimulus dataset was still quite balanced. Table B.4 show the counts for each linguistic phenomenon according to the new analysis.

I report this for full transparency; the choice of CLIP output as a balancing metric is arbitrary in the first place. This updated analysis shows that the dataset remains well-balanced with a new implementation of CLIP.

C. Appendix: Utrecht University ethics check

The Utrecht University ethics QuickScan by the Research Institute of Information and Computing Sciences[1] produced the following result, meaning the research design was not flagged for ethics issues. The summary PDF is reproduced starting on the following page.

[1] https://www.uu.nl/en/research/institute-of-information-and-computing-sciences/ethics-and-privacy

Response Summary:

Section 1. Research projects involving human participants

P1. Does your project involve human participants? This includes for example use of observation, (online) surveys, interviews, tests, focus groups, and workshops where human participants provide information or data to inform the research. If you are only using existing data sets or publicly available data (e.g. from Twitter, Reddit) without directly recruiting participants, please answer no.

- Yes

Recruitment

P2. Does your project involve participants younger than 18 years of age?

- No

P3. Does your project involve participants with learning or communication difficulties of a severity that may impact their ability to provide informed consent?

- No

P4. Is your project likely to involve participants engaging in illegal activities?

- No

P5. Does your project involve patients?

- No

P6. Does your project involve participants belonging to a vulnerable group, other than those listed above?

- No

P8. Does your project involve participants with whom you have, or are likely to have, a working or professional relationship: for instance, staff or students of the university, professional colleagues, or clients?

- No

Informed consent

PC1. Do you have set procedures that you will use for obtaining informed consent from all participants, including (where appropriate) parental consent for children or consent from legally authorized representatives? (See suggestions for information sheets and consent forms on the website.)

- Yes

PC2. Will you tell participants that their participation is voluntary?

- Yes

PC3. Will you obtain explicit consent for participation?

- Yes

PC4. Will you obtain explicit consent for any sensor readings, eye tracking, photos, audio, and/or video recordings?

- Not applicable

PC5. Will you tell participants that they may withdraw from the research at any time and for any reason?

- Yes

PC6. Will you give potential participants time to consider participation?

- Yes

PC7. Will you provide participants with an opportunity to ask questions about the research before consenting to take part (e.g. by providing your contact details)?

- Yes

PC8. Does your project involve concealment or deliberate misleading of participants?

- No

Section 2. Data protection, handling, and storage

The General Data Protection Regulation imposes several obligations for the use of **personal data** (defined as any information relating to an identified or identifiable living person) or including the use of personal data in research.

D1. Are you gathering or using personal data (defined as any information relating to an identified or identifiable living person)?

- No

Section 3. Research that may cause harm

Research may cause harm to participants, researchers, the university, or society. This includes when technology has dual-use, and you investigate an innocent use, but your results could be used by others in a harmful way. If you are unsure regarding possible harm to the university or society, please discuss your concerns with the Research Support Office.

H1. Does your project give rise to a realistic risk to the national security of any country?

- No

H2. Does your project give rise to a realistic risk of aiding human rights abuses in any country?

- No

H3. Does your project (and its data) give rise to a realistic risk of damaging the University's reputation? (E.g., bad press coverage, public protest.)

- No

H4. Does your project (and in particular its data) give rise to an increased risk of attack (cyber- or otherwise) against the University? (E.g., from pressure groups.)

- No

H5. Is the data likely to contain material that is indecent, offensive, defamatory, threatening, discriminatory, or extremist?

- No

H6. Does your project give rise to a realistic risk of harm to the researchers?

- No

H7. Is there a realistic risk of any participant experiencing physical or psychological harm or discomfort?

- No

H8. Is there a realistic risk of any participant experiencing a detriment to their interests as a result of participation?

- No

H9. Is there a realistic risk of other types of negative externalities?

- No

Section 4. Conflicts of interest

C1. Is there any potential conflict of interest (e.g. between research funder and researchers or participants and researchers) that may potentially affect the research outcome or the dissemination of research findings?

- No

C2. Is there a direct hierarchical relationship between researchers and participants?

- No

Section 5. Your information.

This last section collects data about you and your project so that we can register that you completed the Ethics and Privacy Quick Scan, sent you (and your supervisor/course coordinator) a summary of what you filled out, and follow up where a fuller ethics review and/or privacy assessment is needed. For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information. Please see the guidance on the ICS Ethics and Privacy website on what happens on submission.

Z0. Which is your main department?

- Information and Computing Science

Z1. Your full name:

Eduard Saakashvili

Z2. Your email address:

e.saakashvili@students.uu.nl

Z3. In what context will you conduct this research?

- As a student for my master thesis, supervised by::
  Pablo Mosteiro Romero

Z5. Master programme for which you are doing the thesis

- Artificial Intelligence

Z6. Email of the course coordinator or supervisor (so that we can inform them that you filled this out and provide them with a summary):

p.j.mosteiroromero@uu.nl

Z7. Email of the moderator (as provided by the coordinator of your thesis project):

coordinator-ai-master@uu.nl

Z8. Title of the research project/study for which you filled out this Quick Scan:

Does VLM Humanlikeness Predict Performance on Grounding Challenges

Z9. Summary of what you intend to investigate and how you will investigate this (200 words max):

I intend to investigate to what extent the human likeness of vision language model attribution maps correlates with performance on grounding challenges. To investigate this, I collect anonymized human saliency maps from human participants on Prolific, which are then compared with model attribution maps.

Z10. In case you encountered warnings in the survey, does supervisor already have ethical approval for a research line that fully covers your project?

- Not applicable

Scoring

- Privacy: 0
- Ethics: 0

Bibliography

Bernardi, Raffaella and Sandro Pezzelle (2021). “Linguistic issues behind visual question answering”. In: *Language and Linguistics Compass* 15.6. _-eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12417, elnc3.12417. ISSN: 1749-818X. DOI: 10.1111/lnc3.12417. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12417 (visited on 09/19/2023).

Brynjolfsson, Erik (Jan. 11, 2022). *The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence*. DOI: 10.48550/arXiv.2201.04200. arXiv: 2201.04200[cs,econ,q-fin]. URL: http://arxiv.org/abs/2201.04200 (visited on 02/27/2024).

Bugliarello, Emanuele, Ryan Cotterell, et al. (May 30, 2021). *Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs*. DOI: 10.48550/arXiv.2011.15124. arXiv: 2011.15124[cs]. URL: http://arxiv.org/abs/2011.15124 (visited on 05/08/2023).

Bugliarello, Emanuele, Laurent Sartran, et al. (May 12, 2023). *Measuring Progress in Fine-grained Vision-and-Language Understanding*. DOI: 10.48550/arXiv.2305.07558. arXiv: 2305.07558[cs]. URL: http://arxiv.org/abs/2305.07558 (visited on 07/25/2023).

Cao, Jize et al. (2020). “Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models”. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 565–580. ISBN: 978-3-030-58539-6. DOI: 10.1007/978-3-030-58539-6_34.

Cass, Connie (June 26, 2014). *Remembering Howard Baker, whose famous question embodied the Watergate hearings*. PBS NewsHour. Section: Politics. URL: https://www.pbs.org/newshour/politics/remembering-howard-baker-whose-famous-question-embodied-watergate-hearings (visited on 09/22/2023).

Christiansen, Morten and Nick Chater (July 1, 1993). “Symbol Grounding-the Emperor's New Theory of Meaning”. In: pp. 155–160.

COCO - Common Objects in Context (2017). URL: https://cocodataset.org/#home (visited on 10/09/2023).

Das, Abhishek, Harsh Agrawal, et al. (Oct. 1, 2017). “Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?” In: *Computer Vision and Image Understanding*. Language in Vision 163, pp. 90–100. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2017.10.001. URL: https://www.sciencedirect.com/science/article/pii/S1077314217301649 (visited on 09/04/2023).

Das, Abhishek, Satwik Kottur, et al. (Nov. 26, 2016). *Visual Dialog*. arXiv.org. URL: https://arxiv.org/abs/1611.08669v5 (visited on 10/09/2023).

Donahue, Jeff and Kristen Grauman (2011). *Annotator rationales for visual recognition*. URL: https://ieeexplore.ieee.org/abstract/document/6126394/?casa_token=GwumiX7AORYAAAAA:-FQwOSnu5iAVSHcXWvPDiAWMYM7t-rw2GQkm5Z8HT8IJHQLvqyJSJvmYdYQp7GvR3LIcja5zoGtswg (visited on 10/12/2023).

Explain ResNet50 ImageNet classification using Partition explainer — SHAP latest documentation (2024). URL: https://shap.readthedocs.io/en/latest/example_notebooks/image_examples/image_classification/Image%20Multi%20Class.html (visited on 02/27/2024).

Flamary, Rémi et al. (2021). "POT: Python Optimal Transport". In: *Journal of Machine Learning Research* 22.78, pp. 1–8. ISSN: 1533-7928. URL: http://jmlr.org/papers/v22/20-451.html (visited on 02/25/2024).

Harnad, Stevan (June 1, 1990). "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1, pp. 335–346. ISSN: 0167-2789. DOI: 10.1016/0167-2789(90)90087-6. URL: https://www.sciencedirect.com/science/article/pii/0167278990900876 (visited on 09/19/2023).

Hendricks, Lisa Anne and Aida Nematzadeh (June 16, 2021). *Probing Image-Language Transformers for Verb Understanding*. DOI: 10.48550/arXiv.2106.09141. arXiv: 2106.09141[cs]. URL: http://arxiv.org/abs/2106.09141 (visited on 05/08/2023).

Hessel, Jack and Lillian Lee (Nov. 2020). "Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!" In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2020. Online: Association for Computational Linguistics, pp. 861–877. DOI: 10.18653/v1/2020.emnlp-main.62. URL: https://aclanthology.org/2020.emnlp-main.62 (visited on 08/14/2023).

Htut, Phu Mon et al. (Nov. 27, 2019). *Do Attention Heads in BERT Track Syntactic Dependencies?* DOI: 10.48550/arXiv.1911.12246. arXiv: 1911.12246 [cs]. URL: http://arxiv.org/abs/1911.12246 (visited on 05/02/2023).

Hugging Face - Documentation (2024). URL: https://huggingface.co/docs (visited on 02/27/2024).

Jiang, Ming et al. (2015). "SALICON: Saliency in Context". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1072–1080. URL: https://openaccess.thecvf.com/content_cvpr_2015/html/Jiang_SALICON_Saliency_in_2015_CVPR_paper.html (visited on 09/20/2023).

Johnson, Steven and Nikita Iziev (Apr. 15, 2022). "A.I. Is Mastering Language. Should We Trust What It Says?" In: *The New York Times*. ISSN: 0362-4331. URL: https://www.nytimes.com/2022/04/15/magazine/ai-language.html (visited on 09/19/2023).

Lake, Brenden M. et al. (Nov. 2, 2016). *Building Machines That Learn and Think Like People*. arXiv: 1604.00289[cs,stat]. URL: http://arxiv.org/abs/1604.00289 (visited on 10/03/2023).

Lin, Tsung-Yi et al. (2014). "Microsoft COCO: Common Objects in Context". In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 740–755. ISBN: 978-3-319-10602-1. DOI: 10.1007/978-3-319-10602-1_48.

Liu, Chenxi et al. (Nov. 23, 2016). *Attention Correctness in Neural Image Captioning*. DOI: 10.48550/arXiv.1605.09553. arXiv: 1605.09553[cs]. URL: http://arxiv.org/abs/1605.09553 (visited on 08/19/2023).

Liu, Fangyu, Guy Emerson, and Nigel Collier (2023). "Visual Spatial Reasoning". In: *Transactions of the Association for Computational Linguistics* 11. Place: Cambridge, MA Publisher: MIT Press, pp. 635–651. DOI: 10.1162/tacl_a_00566. URL: https://aclanthology.org/2023.tacl-1.37 (visited on 02/25/2024).

Liu, Yu and Tinne Tuytelaars (Jan. 1, 2020). "A Deep Multi-Modal Explanation Model for Zero-Shot Learning". In: *IEEE Transactions on Image Processing* 29, pp. 4788–4803. ISSN: 1057-7149. DOI: 10.1109/TIP.2020.2975980. URL: https://doi.org/10.1109/TIP.2020.2975980 (visited on 09/19/2023).

Lundberg, Scott and Su-In Lee (Nov. 24, 2017). *A Unified Approach to Interpreting Model Predictions*. DOI: 10.48550/arXiv.1705.07874. arXiv: 1705.07874[cs,stat]. URL: http://arxiv.org/abs/1705.07874 (visited on 09/22/2023).

Mitchell, Melanie and David C. Krakauer (Mar. 28, 2023). "The Debate Over Understanding in AI's Large Language Models". In: *Proceedings of the National Academy of Sciences* 120.13, e2215907120. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2215907120. arXiv: 2210.13966[cs]. URL: http://arxiv.org/abs/2210.13966 (visited on 08/19/2023).

Molnar, Christoph (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd. URL: christophm.github.io/interpretable-ml-book/.

Ocampo, Alex and Jemar R. Bather (June 2, 2022). *Single-World Intervention Graphs for Defining, Identifying, and Communicating Estimands in Clinical Trials*. arXiv.org. URL: https://arxiv.org/abs/2206.01249v1 (visited on 10/09/2023).

Parcalabescu, Letitia, Michele Cafagna, et al. (Mar. 14, 2022). *VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena*. DOI: 10.48550/arXiv.2112.07566. arXiv: 2112.07566[cs]. URL: http://arxiv.org/abs/2112.07566 (visited on 04/13/2023).

Parcalabescu, Letitia and Anette Frank (May 23, 2023). *MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks*. DOI: 10.48550/arXiv.2212.08158. arXiv: 2212.08158[cs]. URL: http://arxiv.org/abs/2212.08158 (visited on 08/14/2023).

Park, Dong Huk et al. (Feb. 15, 2018). *Multimodal Explanations: Justifying Decisions and Pointing to the Evidence*. DOI: 10.48550/arXiv.1802.08129. arXiv: 1802.08129[cs]. URL: http://arxiv.org/abs/1802.08129 (visited on 09/04/2023).

Radford, Alec et al. (Feb. 26, 2021). *Learning Transferable Visual Models From Natural Language Supervision*. DOI: 10.48550/arXiv.2103.00020. arXiv:

2103.00020[cs]. URL: http://arxiv.org/abs/2103.00020 (visited on 09/14/2023).

Rao, Varun Nagaraj et al. (Apr. 28, 2021). *A First Look: Towards Explainable TextVQA Models via Visual and Textual Explanations*. DOI: 10.48550/arXiv.2105.02626. arXiv: 2105.02626[cs]. URL: http://arxiv.org/abs/2105.02626 (visited on 09/04/2023).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (Aug. 9, 2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. DOI: 10.48550/arXiv.1602.04938. arXiv: 1602.04938[cs,stat]. URL: http://arxiv.org/abs/1602.04938 (visited on 02/26/2024).

Rodis, Nikolaos et al. (June 9, 2023). *Multimodal Explainable Artificial Intelligence: A Comprehensive Review of Methodological Advances and Future Research Directions*. DOI: 10.48550/arXiv.2306.05731. arXiv: 2306.05731[cs]. URL: http://arxiv.org/abs/2306.05731 (visited on 08/14/2023).

scipy.stats.chisquare — SciPy v1.12.0 Manual (2024). URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html (visited on 02/25/2024).

scipy.stats.entropy — SciPy v1.12.0 Manual (2024). URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.entropy.html (visited on 02/25/2024).

scipy.stats.spearmanr — SciPy v1.12.0 Manual (2024). URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html (visited on 02/25/2024).

Searle, John R. (Sept. 1980). "Minds, brains, and programs". In: *Behavioral and Brain Sciences* 3.3. Publisher: Cambridge University Press, pp. 417–424. ISSN: 1469-1825, 0140-525X. DOI: 10.1017/S0140525X00005756. URL: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A (visited on 10/05/2023).

Selvaraju, Ramprasaath R. et al. (Oct. 28, 2019). *Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded*. DOI: 10.48550/arXiv.1902.03751. arXiv: 1902.03751[cs]. URL: http://arxiv.org/abs/1902.03751 (visited on 09/19/2023).

Shanahan, Murray (Feb. 16, 2023). *Talking About Large Language Models*. DOI: 10.48550/arXiv.2212.03551. arXiv: 2212.03551[cs]. URL: http://arxiv.org/abs/2212.03551 (visited on 07/25/2023).

Shekhar, Ravi et al. (July 2017). "FOIL it! Find One mismatch between Image and Language caption". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2017. Vancouver, Canada: Association for Computational Linguistics, pp. 255–265. DOI: 10.18653/v1/P17-1024. URL: https://aclanthology.org/P17-1024 (visited on 08/01/2023).

Singh, Amanpreet et al. (Mar. 29, 2022). *FLAVA: A Foundational Language And Vision Alignment Model*. DOI: 10.48550/arXiv.2112.04482. arXiv: 2112.04482[cs]. URL: http://arxiv.org/abs/2112.04482 (visited on 02/24/2024).

Sood, Ekta et al. (June 2023). “Multimodal Integration of Human-Like Attention in Visual Question Answering”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, BC, Canada: IEEE, pp. 2648–2658. ISBN: 9798350302493. DOI: `10.1109/CVPRW59228.2023.00265`. URL: `https://ieeexplore.ieee.org/document/10208301/` (visited on 09/19/2023).

Sun, Jiamei et al. (2020). “Understanding Image Captioning Models beyond Visualizing Attention”. In.

Tan, Hao and Mohit Bansal (Dec. 3, 2019). *LXMERT: Learning Cross-Modality Encoder Representations from Transformers*. DOI: `10.48550/arXiv.1908.07490`. arXiv: `1908.07490[cs]`. URL: `http://arxiv.org/abs/1908.07490` (visited on 02/24/2024).

Tenney, Ian, Dipanjan Das, and Ellie Pavlick (July 2019). “BERT Rediscovers the Classical NLP Pipeline”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Florence, Italy: Association for Computational Linguistics, pp. 4593–4601. DOI: `10.18653/v1/P19-1452`. URL: `https://aclanthology.org/P19-1452` (visited on 05/02/2023).

Thrush, Tristan et al. (Apr. 22, 2022). *Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality*. DOI: `10.48550/arXiv.2204.03162`. arXiv: `2204.03162[cs]`. URL: `http://arxiv.org/abs/2204.03162` (visited on 02/25/2024).

Wu, Jialin and Raymond Mooney (Aug. 2019). “Faithful Multimodal Explanation for Visual Question Answering”. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. BlackboxNLP 2019. Florence, Italy: Association for Computational Linguistics, pp. 103–112. DOI: `10.18653/v1/W19-4812`. URL: `https://aclanthology.org/W19-4812` (visited on 08/19/2023).

Xu, Wenjia et al. (Mar. 2020). “Where is the Model Looking At?–Concentrate and Explain the Network Attention”. In: *IEEE Journal of Selected Topics in Signal Processing* 14.3, pp. 506–516. ISSN: 1932-4553, 1941-0484. DOI: `10.1109/JSTSP.2020.2987729`. arXiv: `2009.13862[cs]`. URL: `http://arxiv.org/abs/2009.13862` (visited on 09/19/2023).

Yang, Yi et al. (Oct. 14, 2022). “HSI: Human Saliency Imitator for Benchmarking Saliency-Based Model Explanations”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 10, pp. 231–242. ISSN: 2769-1349. DOI: `10.1609/hcomp.v10i1.22002`. URL: `https://ojs.aaai.org/index.php/HCOMP/article/view/22002` (visited on 02/27/2024).

Yuksekgonul, Mert et al. (Sept. 29, 2022). “When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It?” In: The Eleventh International Conference on Learning Representations. URL: `https://openreview.net/forum?id=KRLUvxh8uaX` (visited on 08/19/2023).

Zhai, Xiaohua et al. (Sept. 27, 2023). *Sigmoid Loss for Language Image Pre-Training*. DOI: `10.48550/arXiv.2303.15343`. arXiv: `2303.15343[cs]`. URL: `http://arxiv.org/abs/2303.15343` (visited on 02/24/2024).

Zhu, Yuke et al. (Nov. 11, 2015). *Visual7W: Grounded Question Answering in Images*. arXiv.org. URL: `https://arxiv.org/abs/1511.03416v4` (visited on 10/09/2023).