

Institute for Visualization and Interactive Systems  
Department of Human-Computer Interaction and Cognitive Systems  
University of Stuttgart  
Pfaffenwaldring 5a  
70569 Stuttgart

Bachelor Thesis

**The Need to Explore Different  
Perspectives:  
Visual Question Answering and its  
Variability in Replies Caused by  
Mental States**

Mah-Rukh Sarfaraz Chaudhry, 3337107  
st156852@stud.uni-stuttgart.de  
Matr. Nr.: 3337107

**Course of Study:** Software Engineering

**Examiner:** Prof. Dr. Andreas Bulling

**Supervisor:** Susanne Hindennach, M.Sc.

**Commenced:** November 10, 2022

**Completed:** May 10, 2023

## **Abstract**

Most subjective reasoning is influenced by mental states. It facilitates communication between people and enhances interpersonal relationships. It can be difficult to duplicate using machine learning since complicated questions with subjective and challenging themes allow for multiple correct answers rather than just one. In my thesis, I mainly explored the kinds of diverse replies in Visual Question Answering and what elements could cause mental state references to signify the need for further exploration. Therefore, I created a user study to collect the rationales for answers given to question-image pairs. The survey study's results offer a data set with several justifications that include mental states.

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Problem and Goals . . . . .	5
1.1.1	Problem Statement . . . . .	5
1.1.2	Mandatory and Optional Goals . . . . .	6
<b>2</b>	<b>Related Work</b>	<b>7</b>
2.1	VQA . . . . .	7
2.2	Existing Studies of Mental State and Theory of Mind . . . . .	8
2.3	Studies about Automated Rationale . . . . .	8
<b>3</b>	<b>Methods</b>	<b>10</b>
3.1	Image and Question Pair - Data Sets . . . . .	10
3.1.1	Defining Agreement Rate . . . . .	12
3.2	Mental State References Study . . . . .	12
3.3	Defining Mental State References . . . . .	14
3.4	Calculations of the Similarity of Click Pattern . . . . .	15
3.5	Participant Demographic . . . . .	16
<b>4</b>	<b>Results</b>	<b>17</b>
4.1	Agreement Rate and Confidence Rate . . . . .	17
4.2	Mental State References Analysis . . . . .	17
4.3	Hypothesis: Mental State Reference Depends on a Specific Type of Question . . . . .	19
4.4	Hypothesis: Agreement Rate Depends on the Area a Participant Looked on an Image . . . . .	20
4.5	Hypothesis: The Lower the Agreement Rate, the Higher the Mental State Reference Rate . . . . .	20
<b>5</b>	<b>Discussion</b>	<b>21</b>
5.1	Meaning of the Results . . . . .	21
5.2	In Comparison to Previous Studies . . . . .	22
5.3	Limitations of this Study . . . . .	22
5.4	Potential Future Research . . . . .	23
<b>6</b>	<b>Conclusion</b>	<b>23</b>
<b>A</b>	<b>Appendix</b>	<b>28</b>
A.1	Click Pattern Map on Stimuli Images and their Click Pattern Similarity Score . . . . .	28
A.2	Code to calculate Click Pattern Similarity Score . . . . .	29
A.3	Selected Question Image Pair Stimuli . . . . .	30
A.4	Master List of Verbs of Mental State References . . . . .	32

## List of Figures

1	Amount of Questions in each Annotator Agreement Rate . . . . .	11
2	User Study Screenshots . . . . .	13
3	Participant Demographic - Ethnicity and Educational Background Pie Charts . . .	16
4	Bar Chart of Mental State Reference Rate in each Question-Image Pair . . . . .	18
5	Scatter Plot Between Agreement Rate and Mental State Reference Rate . . . . .	19
6	Sample of Summarized Click Patterns Placed on their Image . . . . .	20
7	Sample of Placed Click Patterns from Question-Image Pair ID 1 . . . . .	21
8	Python Code for Calculating PCC of Every Click Pattern . . . . .	29

## List of Tables

1	Data of Agreement Rate, Mental State Reference Rate and Confidence Rate . . . .	17
2	Summarized Click Pattern . . . . .	28

# 1 Introduction

Mental state[16] is a term that covers the psychological state behind most cognitive reasoning, like beliefs, desires and emotions. Mental state reference depicts the usage of referring to one's own mental state and others. Theory of Mind[33, 27] covers part of it, specifically the area where one assesses the mental states of another person, whether in real life, in an image, or in a video. It is a skill used to convey or understand a perspective. It helps in dissolving differences of perspective in case several people desire to coexist better and it helps in predicting another person's actions[16].

It is necessary to understand how users function and explain their actions to replicate them. And also to understand how much of our behaviour is influenced by our mental state.

Programs function and react within defined variables. Human explanations that involve abstract terms like beliefs, desires and emotions are harder to define or reproduce. The need to explore this topic further needs to be made, especially, on the topic of recognising questions which have diverse answers made by humans and delivering possible solutions which are more than one.

Typical in trying to replicate such an explanation is that one seeks one true defined answer. VQA, Visual Question Answering[8], is an example where it prioritises finding common sense answers when it comes to questions about images. Their main objective is to automatize answers to visual questions in a natural language to aid the visually impaired. However, because they were written or spoken by humans, questions in natural language also contained words that were ambiguous and subject to interpretation. There was a lot of room for interpretation when it came to questions that touched on our emotions, beliefs, and opinions. These questions frequently had a wide array of responses since humans, unlike machines, can recognise objects and humans in an image through their experience, memories and even imagination differently[20, 22].

These are difficulties in VQA that are challenging to solve because there is still research being done on how to evaluate questions with several probable answers besides disregarding them to focus on their actual goal of automating answers to questions that have one answer [15, 43].

Therefore, I needed more context from the human perspective to make a connection with mental state references and the variety of opinions. By proving a significant correlation, there could be a possible solution on where to research, specifically by being able to automate answers that have a high variety of opinions on how an image is perceived. I have the desire to move in the direction of a model that can reflect the variety of answers, in which there is not one true answer but a multitude of perspectives that lead to different conclusions. This can be regarded as a suggestion system that would generate natural-language explanations of different perspectives. Gaining insight into another perspective leads to improved human-to-human interaction.

This is why I created this thesis to investigate the situation by creating a user study for human-computer interaction, so I was capable of collecting explanations of answers and classifying whether they contained mental state references and potential different factors on why an image was perceived differently.

## 1.1 Problem and Goals

This section describes the problem statement and the hypotheses the thesis was dealing with. After this, the next subsection describes the goals of this thesis.

### 1.1.1 Problem Statement

At the moment, the problems in resolving disagreements in image recognition due to the ambiguity of the visual questions remain due to a lack of exploration. Human recognition is strongly affected by our memories and beliefs [22] and has shown that there are situations where there is not one right answer to questions. This is why I sought to find its involvement in Visual Question Answering to understand and gather the variability in answers and tested it against different potential factors. For this reason, I put up the following hypotheses that had been investigated:

- If the questions possess a variety of answers, the explanations for them contain a higher number of mental state references.
- Mental state references depend on the type of question.
- If people do not look in the same area on the image, there is a lower agreement rate.

These hypotheses' variables are explained in more detail and are defined in the method section.

### **1.1.2 Mandatory and Optional Goals**

A mandatory goal for this bachelor thesis was to seek out VQA samples that had a wide array of different amounts of conflicting answers when it came to describing visuals. A user study was designed that prompted the participants to explain their reasons, involving their states of mind, for answering the visual questions. The second mandatory goal was gathering these explanations and classifying and analysing them with the help of an automatic natural language tool. The goal of the analysis was to analyse the hypotheses and determine whether or not they were true.

## 2 Related Work

The associated work with the thesis is explained in this part. It begins with the first segment, Visual Question Answering and its earlier research aimed at addressing disputes in replies and their limitations. Then, related literature on the value of discussing mental states follows. Finally, earlier initiatives compile justifications for automating natural texts.

### 2.1 VQA

The most common way for a system to replicate natural explanations was to find a defined true answer and conclusion. The "VQA: Visual Question Answering" [14, 41, 21] study and project aimed to provide natural-language answers to questions about a given image. Their motivations involved helping the visually impaired, so they gained context about images through open-ended questions reflecting realistic scenarios. These questions had been gathered through a set-up, according to the author [14], where participants in the experiment posed questions about an image by "trying to outsmart a tiny robot, an alien or a toddler". The accuracy of answers was graded by how many people had the same opinion on a question. Therefore, answers made by humans that were conflicting and had low accuracy were regarded as incorrect or not usable since the aim was to seek one common-sense answer. Their large data set of visual questions and answers was gathered through the recruitment of Amazon Mechanical Turk, a crowd-sourcing marketplace[1]. This project naturally led to challenges and other projects trying to create a deep-learning model that was capable of answering natural language questions with visuals trained by the collected and given answers by humans.

The emphasis was laid on image recognition[13, 40] in seeking clearly defined answers and an understanding of the image with complex reasoning and binding it with natural language. The project posed challenges from 2015 to today for researchers to create systems, and they regarded their accuracy in answering the open-ended questions in the context of the images. The results of the researchers' method were, in most cases, publicly posted on their site's challenger leaderboard [9]. The project was built by seeking a common-sense perspective, with which the majority agreed.

To improve the answer accuracy, studies have been done to consider dynamic word usage[44] by taking the tone and situation of the image into context. This study pointed out that whenever the answer was evaluated in VQA in most instances, the words were compared to see whether or not they matched the answer.

In another following study, "Perception Matters: detecting perception Failures of VQA models using metamorphic testing" [43] attempted to overcome any ambiguity in a question by trying to create a meta-model of the thought process for a more complex question. Before it came to the actual logical conclusion to be able to answer it, it checked with smaller questions if certain objects were present in the image associated with the question and used this information to answer the question.

While these mentioned studies did showcase ways to improve in helping answer difficult logical questions and giving a foundation on decreasing the amount of invalid ambiguous questions, these studies still had a limit to the unsolved problems of questions that weren't low-quality or vaguely phrased.

However, ambiguous and subjective questions that allowed for many possible answers remained a challenge. Only in a recent paper, "Why does a Visual Question Have Different Answers?" [15] the limitation was further addressed by trying to understand the causes of different opinions in answers. In this paper, they proposed the idea that data sets with visual questions should be filtered further by differentiating whether certain questions have a possible technical answer. They checked whether an image was of bad quality, such as badly lit or blurry images, or whether it was a subjective or ambiguous question that requested an opinion. The study did note their limitations on how to approach the questions, which can have valid different answers, or how to train a model of complex reasoning like common sense.

## 2.2 Existing Studies of Mental State and Theory of Mind

This is the subsection where the research now moves further into mental state references, as it covers most of the subjective reasoning and even our common sense. The introduced related works showcase attempts to define the mental state references and also to see their potential and benefit by indulging in exploring mental states further.

The paper "Why talk about mental states? the significance of children's conversations with friends, siblings, and mothers" [16] mentioned the importance of why people shared their mental state in social conversation. The result showed that children reveal their mental state to convey their perspective for their friends to understand or even modulate their assertion. By sharing their mental states, their motivation was to prolong their time for shared activity and resolve conflicting views while playing.

With this paper, I learned that understanding another perspective could lead to the improvement of human-to-human interactions and is very vital in social settings. To understand a different perspective, one needs to understand how it is shaped.

A related study involving mental state references, specifically Theory of Mind, "Spontaneous mentalizing captures variability in the cortical thickness of social brain regions" [34] examined a group of people watching naturalistic videos. The participants were prompted to make spontaneous descriptions of the characters they perceived in the video. Their discovery provided new insight into the neural bases of variability in naturalistic mentalizing performance. It further showcased the usage of mental state references for an individual to grasp the reason and perspective of the figures they are perceiving.

To be able to differentiate what is a mental state reference in collected explanations, the study "Representing other minds: Mental state reference is moderated by group membership" [28] had a good foundation for word usage. This research collected explanations from the participants and analysed them for potential mental state references in the logged conversations of their experiments. With their results, they were capable of defining explanations for whether they refer to a mental state or not.

## 2.3 Studies about Automated Rationale

The natural explanation is important in situations where human operators work alongside autonomous and semi-autonomous systems because it can help build understanding between the agent and its operator. In health care, [17] natural explanation was necessary for conveying information more efficiently than structured data, as it had been shown that communication was very vital.

The paper "Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions" [19], moved in this direction. This was an approach for real-time explanation generation whereby a computational model learned to translate an autonomous agent's internal state and action data representations into natural language.

In this paper, they collected a corpus of explanations with the goal of using it by desiring to train a neural rationale generator to produce different rationales and also checked how people perceive these rationales. The study was broader when it came to capturing human rationale as it involved the possibility of AI and how well they were able to convey human likeness by the use of a simpler game. A similar project that moved in the direction of automating explanation is "Generating Visual Explanations" where people could request an explanation to a visual question but not the reasons for the thought processes.

In this thesis, I expanded on the research of Visual Question Answering by focusing on questions and images with several different amounts of opinions. Although the reason for different answers is known, the actual cause needs to be further investigated. In this case, my desire was to connect the research to mental states and the Theory of Mind that could help further indicate what people use to answer complex, subjective, and ambiguous questions.

Inspired by the mentioned experiments of prior research in automated rationale, I desired to create a survey in which participants regarded the collected images from VQA and explained



their rationale when they gave an answer to a question. The purpose was to gather data on natural language explanations that contained mental state references and demonstrated different numbers of views to assist in understanding another person's conclusion and resolving conflicts of opinion. The data collection could provide more context for the VQA model's responses to open-ended questions and a need to explore further mental states to automate questions that may have more than one correct answer.

### 3 Methods

This section explains the approach used for several tasks in this thesis. The first sub-section explains the selection of the image and question pairs and their criteria. Then I showcase the user study that used the selected question image pairs as stimuli and how I gathered the desired data I needed for my calculation. After this, I defined the variables needed for my hypotheses and how I used the gathered data from the user study to properly evaluate them in the analysis.

#### 3.1 Image and Question Pair - Data Sets

The main source of data set for this thesis was the VQA v2.0 data set. The images and questions presented for the user study were derived from the data set from Visual Question Answering [8] and COCO, Common Objects in Context[26]. The data set was the most up-to-date and accessible to use. The real-life images were common cases of wanting a description in daily life. Furthermore, the validation data set and test data set have been used most by VQA studies for their models to learn [18, 42]. However, since I only required a small sample for a large data set, I mostly focused on analysing the validation data set.

The official VQA site provided annotation files about the collected data of their natural-language questions, where people had regarded specific images and answered questions in their context of them. It contained information on the image ID, question ID, and ten answers to said question-image pair. Alongside it, each answer contained a confidence value between "yes", "maybe", and "no". The annotation files were stored as large JSON files[3].

One of my first goals was to seek out the information that told which question-image pair contained a variety of different responses to certain questions. To be able to differentiate the largest amount of similar opinions, there was in each question an already expected answer under the tag "multiple.choice.answer". The value in this tag told me what was the most mentioned answer in each question, but not how many out of the ten annotators agreed with said answer.

$$Annotator_{agreement} = \frac{\text{Amount of answers agreeing to the value of "multiple\_choice\_answer"}}{10} \tag{1}$$

Therefore, to seek out a rough number, I coded an annotation file parser in Python that went through the answers to each question. It compared the answer to the value in the tag "multiple.choice.answer". If the answers were regarded as matching, the parser would count up. At maximum, the parser could only count up to ten because there were only ten answers to compare to the expected answer. To cover the cases in which some answers were similar to each other, words that were synonyms were also regarded through the usage of nltk[4]. Specifically, the package WortNet [10, 5] and NLTK.corpus as it is a lexical database about the English language with a heavy focus on context and making natural language processing in Python simpler. The list of synonyms was created by the value of "multiple.choice.answer" and that list would check whether the provided answers from the annotator matched with any of the words in the list.

This is how I defined the annotator agreement rate, and equation 1 showcases the definition. For each question-image pair, the annotator agreement rate was counted up from a value of one to ten whenever an answer from an annotator matched it. The validation file was sorted according to its annotator agreement rate and split into ten files.

To note in Figure 1 and its amounts, there was still a chance that some of the question-image pairs in the sorted files weren't, for example, exactly 5 out of 10 people in agreement but potentially had a higher agreement. The explanation for this was that the parser didn't consider the context of how words were used, like analysing the image the question was paired with. The parser only matched the answers with the provided words from the synonym list and the value of the multiple-choice answer.

In some cases, if the expected answer was, for example, "mix of red and white", the parser

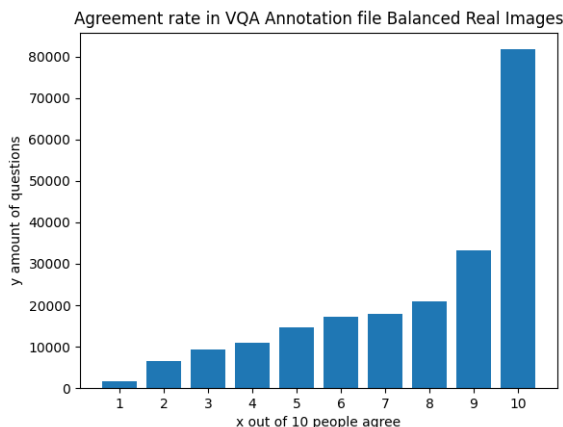


Figure 1: This bar chart showcases the number of question-image pairs for each annotator agreement rate.

wouldn't see the answer "pink" as the same opinion because the word wasn't in the list. Some answers, while using different vocabulary, could imply the same opinion, but may not be covered by the NLTK synonym database. Since the files were too large with 214,354 questions, it was inefficient to be able to check through the annotator agreement manually. If the annotation parser was much more accurate, then the graph would still resemble a function that exponentially grew. The question and image IDs in the separated ten files and associated questions had been identified for each agreement rate file by comparing the IDs in the question data set, also provided on the official VQA project site. In these split files, I noted that the questions with higher agreement rates contained the most question types, which involve colour questions, counting questions, and questions that contain only "yes" and "no" answers. The lowest level of agreement they could get before one of the responses ends up being higher would be 50% annotator agreement rate, especially because "yes" and "no" questions were binary. For the files that had a low annotator agreement rate, they contained more complex questions that would require specific expert knowledge, vague questions that left too many possible answers, and even too subjective questions where one should guess a person's name.

I wanted to select question-image pairs that had different annotator agreement rates. To have a range, I selected question-image pairs that had an annotator agreement rate of 20%, 40%, 60%, 80% and 100%.

Out of the sorted annotation files, specific types of questions were sought from each mentioned annotator agreement rate.

For the specific types of reasoning questions, I decided on sentiment recognition, scene recognition, and object recognition. I decided on them since I wanted to pick a type of question where I didn't expect any mental state references and one where I did. Sentiment recognition questions demanded the need to evaluate the emotional state, specifically the mental state of the depicted person in an image. This was the type of question I expected to have the most mental state references if someone had to explain their thought process to answer it. Scene recognition questions were picked because I had the assumption that the people who would answer these types of questions would have to rely on their experiences and memories. But my assumption was that it wouldn't evoke as many mental state references as sentiment recognition. The least amount of mental state references I would expect in object recognition was that it purely relied on the visuals of an image. The type of question was defined as follows: Emotional recognition questions contained adjectives or verbs that describe sentiments like "smiling, sad, upset" etc. Object recognition was a question that identified the object in an image. I relied on how COCO images were tagged, which described the content of the images [26] to specify how an object was.

Scene recognition was defined, in comparison to object recognition, to recognise the context, state, or semantic relationship of objects depicted in the image. But to simplify, any question that questioned the functionality and state of an object in the context of the image layout, like a wall, sky, or ground, was seen as a scene recognition question. To evaluate how confident people were in answering a question, I calculated a mean value between the ratings of "yes", "maybe" and "no". The rating "yes" received a value of 1, the rating "maybe" received a value of 0.5 and the rating "no" received a value of 0.

The selection of the questions had to fulfil the following criteria:

1. The selected questions' answers must have a confidence rate of at least over 0.7 to avoid question-image pairs that have bad quality.
2. To select a sentiment recognition question, it must contain an adjective or verb that indicates sentiments.
3. To ask an object recognition question, it must only prompt the participant to identify the object in the image.
4. For scene recognition questions, the question must ask about the function or state of an object or scene in the context of its full environment depicted in the image.

To pick out a selection of questions of specific question types, I used keywords to search for the types of questions in the different agreement rates. An example keyword search for a scene recognition question involved words like "weather, function, purpose, wall, sky". I also specifically regarded questions with the format "What is... for?" because they ask about the functionality of a depicted object. Then I checked if the question fulfilled the criteria mentioned prior and selected it if it was fulfilled.

In the picked-out selection, most of the sentiment recognition questions depicted women and children. The decision to select them was made without focusing particularly on gender but rather on whether or not they fulfilled the criteria belonging to specific types of questions.

### **3.1.1 Defining Agreement Rate**

Since the question-image pair was used in a user study and therefore did not have a predefined answer that defined the majority opinion like in the annotation files, I defined the agreement rate similarly to the prior evaluation when I was dividing the annotation file by their agreement rate. The agreement rate of the answers to a question was defined as the most commonly appearing answer, divided by the number of answers given by participants. This gave me a value between 1 and 0. To be able to decide whether one answer was the same as another answer but with different word choices, synonyms were considered by checking dictionaries like thesaurus[7, 2]. Since the selected data set was much smaller than the annotation file, I was capable of manually analysing the agreement rate by comparing answers. To note, I wouldn't be taking into account the context given by the explanations to an answer, as I was evaluating similarly to how VQA proceeded by having only the context of the short answer given by an annotator.

## **3.2 Mental State References Study**

This subsection explains the setup of the user study to gather the desired data I required for my hypotheses. The premise of a survey required no expert knowledge or experience aside from one's own biased perspective and common sense.

The survey was structured as a questionnaire; therefore, LimeSurvey was used as it covered most functions and was readily available and provided. The user study took place locally at the University of Stuttgart with the set-up of a mouse and laptop in a laboratory. The participants had to be there in person to participate. This decision had been made for the sake that the quality of data was higher with fewer disturbances in comparison to participating in the user

**First part:**

- Answer the question based on what is depicted in the image
- Your answer should be a brief phrase (not a complete sentence)
  - "It is a kitchen" -> "kitchen"
- For yes/no questions, please just say yes/no.
- If you need to speculate (e.g., "What just happened?"), provide an answer that most people would agree on.
- If you don't know the answer (e.g., specific dog breed), provide your best guess.
- Respond matter-of-factly and avoid using conversational language or inserting your opinion.



**What is wrong with his face?**

Example Answer: rolling his eyes

**When answering the prior question which area did you focus on?**

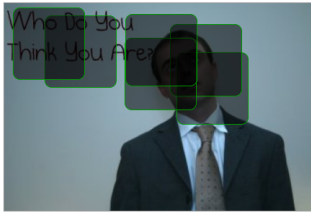


Figure 2: Screenshots of the user study's first part sans the confidence rating. The second image portrays the second part of the user study with placed semi-transparent squares.

study at home since online recruitment in comparison to Amazon Mechanical Turk had a lower quality of data[32].

From the methods by which I picked out question-image pairs for certain criteria, depending on question types and annotator agreement rate, fifteen question-image pairs had been selected as stimuli. The full list of stimuli can be found in the appendix.

I decided to present the stimuli with the lowest annotator agreement rate at the beginning and the highest at the end of the questionnaire. The justification for this was the assumption, based on a small pilot test among volunteers, that people would be too focused on explaining their thought process if they were confronted with question-image pairs with low agreement rates from the start, as they tend to evoke the need to explain.

The study was split into three parts for each question-image pair presented.

The first part's purpose was to replicate the setup of the Visual Question Answering project. This was to garner comparable results to the original annotation data sets to validate my selection of stimuli. The participant had to answer the visual question presented. The criteria were to give an unbiased answer or an answer most people would agree on. They also had to keep the answer short and precise. After their answer, they had to rate how confident the participant was about their given answer by choosing between "yes", "maybe" and "no". The values to pick from for the confidence rating were also taken from the VQA annotation files.

The second part of the user study involved clicking on the area the user focused on when answering the prior question. They placed a semi-transparent black square of the size of 50 pixels onto the image. The highest number of squares that could be placed onto the image was eight, due to the limitations the survey provided. The coordinates of said placed squares were recorded to later help in making a click pattern, which was used to calculate later as described in section 3.4.

Before the participants proceeded with the study, they were able to read through an example. Figure 2 visualised the example every participant had to read through and showcased the two parts described above.

The third part requested the thought process behind the answer given to the question-image pair. The criteria and conditions, in this case, were to write in full sentences and explain their reasons for giving a certain answer to a visual question. This was for the sake of having a proper text analysis later on for variables like mental state reference, which is defined in the upcoming subsection.

### 3.3 Defining Mental State References

To approach the upcoming text analysis and to differentiate whether a mental state reference had been made in given explanations, I researched studies about their methods.

I focused on verb usage in the explanation rather than adjectives alone since most studies point in the direction of relying on verbs. Examples of it were the automatic text analysis study about the Linguistic Category Model [38, 39] that argued verb usage was the most reliable in analysing text.

To differentiate what a mental state reference is, I regarded papers that had done analysis about the language used between humans when they revealed a mental state. Certain words could indicate mental state references by having an explanation involving the participant as the object of the sentence. The basis for the classification of verb terms that insinuated mental state references belonged to the study[37], already referred to in the related work section. This study provided a master list of mental state verbs that had been collected in a user study involving group discussions. For further extension, newer studies like [31, 12] had been regarded as extending the master list as they pointed out further verbs and what most people regarded as verbs that could describe a mental state.

The full list of mental state verbs used in the quantitative analysis can be found in the appendix. For images that involved the depiction of humans, I considered the categories of Theory of Mind, and Spontaneous Theory of Mind, STOMP[34]. STOMP was defined as an unconscious and automatic understanding of others' mental states. To be able to differentiate it from the usage of Theory of Mind in the analysis, I needed to check whether the participant attempted to understand the mental state of the figure depicted in an image without the question prompting it. So to find out whether STOMP was present, I would disregard any sentiment recognition questions, as they asked for an intentional understanding of the emotional state of a person depicted in an image.

To verify whether Theory of Mind had been used, I must verify that the referenced subject and the known mental state reference verbs in a participant's explanation were about the person depicted in the images and not the participant themselves. Theory of Mind and STOMP was taken into consideration in order to determine whether they might be present as the variable "mental state references rate", which would be used in subsequent calculations described in the result section, served as an umbrella term that included references to Theory of Mind, STOMP, and the mental state of the participants. For easier identification in the upcoming analysis, I refer to the mental state reference only directed at the participant's mental state as MSR. This meant the variable mental state references rate covered Theory of Mind, STOMP and MSR.

Naturally, it was to be expected that not every explanation, as I gathered, would refer to a mental state. The sentences that described the exterior or situation of the image were kept in the discussion.

It was to be expected that some terms were not included in the master list that covered the entire usage of making mental state references, so the list of terms that referred to mental state references was extended due to the result given from the experiment. This is explained in detail in the quantitative analysis of the mental state references subsection.

The second method of analysing the mental state references relied not solely on the mental state references master list but also on manually analysing the noun and language usage in the

given explanation to check whether a mental state reference had been made. To differentiate later for the analysis, I called the mental state reference results that solely relied on the master list "Automatic MS" and the second method "Manual MS". When an explanation given by a participant contained a mental state reference, I gave it a value of 1, otherwise, it would have a value of 0. Then I calculated the mean of how many participants gave a mental state reference for one question-image pair. The resulting value was my mental state reference rate.

### 3.4 Calculations of the Similarity of Click Pattern

This section discusses the reasoning behind the methodologies used for comparing the click patterns acquired by the user study. Then I explain the ensuing calculation and definition of the pattern similarity score. The defined similarity score helped in answering the third proposed hypothesis.

Comparing two images about their similarities by using Pearson correlation was common practice in studies [35, 23, 29]. The most notable paper whose methods I based on was the study "BubbleView: an interface for crowdsourcing image importance maps and tracking visual attention" [23]. The researchers tracked the areas where the participant clicked on an image and turned it into maps. These maps were images compared to a similar map that visualised where the participant actually looked. The paper also used Pearson correlation; therefore, I proceeded with the equation too. This was how I approached the resulting maps I received from my user study. From the study in which the participant placed squares on the images in task two, I created a black-and-white map that visualised the click pattern of the participant. The placed squares were black on the image, while the remainder of the image that wasn't covered by any squares remained white. An example is displayed in figure ??imgex]. A sigma of one degree was used to blur the patterns to allow for error. To receive the similarity score of two images, I performed the Pearson correlation coefficient[30]. It gave me a result between minus one and one. Equation 2 [30] describes the formula that performs it on two images.

$$r = \frac{\sum_i (x_i - x_m)(y_i - y_m)}{\sqrt{\sum_i (x_i - x_m)^2} \sqrt{\sum_i (y_i - y_m)^2}} \quad (2)$$

In this equation x were the pixels in image 1 and y the pixels in image 2.  $x_i, y_i$  described the intensity of the  $i^{th}$  pixels of the first and second image. Intensity[36] can be understood as the pixel value between 1, which is seen as black, and 256 which is the colour white. In bits, the grey scale starts with 0 and goes until 255.

The  $x_m, y_m$  described the mean intensity of the pixels of their respective images. To note, both images needed to have the same length and width for the comparison to be able to succeed.

To calculate the image similarities between two images in Python, the image was exported as a Mat object[24] which was an n-dimensional array. To be able to use the readily available function in Numpy [6] that calculated the Pearson correlation coefficient, the image was flattened as a one-dimensional array[11]. If r gave us the result "0", the images were not similar at all. For result "1", the images were identical. In the case of the result of "-1", the two images were negatives of each other. Because the Pearson correlation coefficient was only capable of doing pairwise comparison results, I received a nxn matrix for n numbers of click pattern maps as I needed to compare them with one another. The values were combined and divided by their entire quantity to obtain one mean value. The numbers that contained the picture comparison with itself were removed from the calculations because the r result is naturally 1. Because each value below the matrix's diagonal had identical values in the upper triangle, only the values above the diagonal were added together to save complexity.

If the mean score was less than 0.5, they were considered unrelated. In the appendix, the figure 8 displays the code for what was previously described.

To visualise the maps of every participant in one image, I converted the black-and-white map into a matrix. Since I only wanted 1 and 0 values in the matrix, I divided the intensity of the pixels by 255 to keep the image on a black-and-white scale. I added the click patterns of other

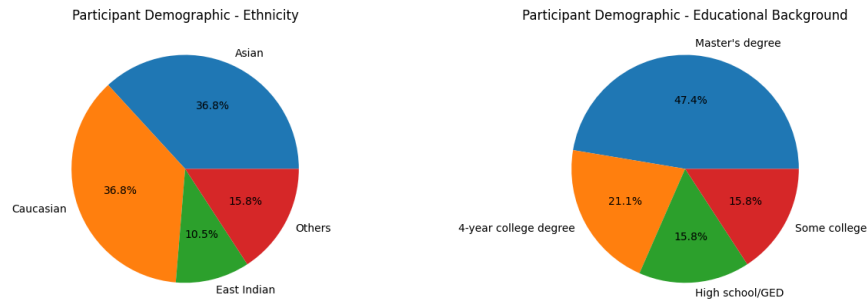


Figure 3: Ethnicity and educational background pie charts from the demographic user study

participants from the same question-image pair to each other and divided them by the number of click patterns that were available in one image, which was 19 to return to the range of values between 1 and 0. This would lead me to an image that had varying tones of grey. This would indicate that the darker the grey areas are in the resulting image, the more people clicked on the same area as others.

This map in grey scale had a filter performed on it, which I coded. It extracted any values of grey and made it have different values of opacity depending on how dark the grey value was. Pixels with the color black had for example an opacity of 0. The map was placed onto the stimuli image to visualise which area the participants most focused on and can be regarded in the appendix.

### 3.5 Participant Demographic

The people that were recruited for the user study were gathered through the university forums and platforms. The number of samples that were collected was nineteen. The majority of the participants in the study were male. Out of 19 people, 15 identified as male and four as female. Almost half of the participants had an educational background of a Master's degree as visualised in figure 3. Seven of the participants' ages ranged from 18 to 25, and twelve participants' ages ranged from 25 to 35. Most participants were Caucasian or Asian. The right image in figure 3 displays it. In general, the demographic wasn't diverse. The participant's time spent finishing the user study was approximately 45 minutes, with a standard deviation of  $\pm 14.81$  minutes.

Participants were compensated for their time because the expected duration of the experiment was around an hour, with 10 euros.



## 4 Results

This section showcases the results of the calculations of the defined values described in the method sections. Then it describes the results of the hypotheses by performing tests to determine whether they were viewed as true or disregarded.

Question-Image Pair ID	Agreement Rate	Automatic MS	Manual MS	Confidence
1	0.42	0.74	0.84	0.95
2	0.42	0.47	0.58	0.55
3	0.26	0.47	0.52	0.63
4	0.37	0.78	0.89	0.76
5	0.37	0.26	0.31	0.92
6	0.37	0.58	0.68	0.55
7	0.78	0.84	0.94	0.82
8	0.78	0.26	0.47	0.94
9	0.37	0.47	0.52	0.69
10	1	0.89	1	0.82
11	0.84	0.21	0.42	0.84
12	0.63	0.21	0.37	0.84
13	1	0.94	1	0.94
14	0.94	0.15	0.15	0.92
15	1	0.15	0.15	0.97

Table 1: Data collected from mental state reference analysis, agreement and confidence rate

### 4.1 Agreement Rate and Confidence Rate

The variable agreement rate was evaluated as described in subsection 3.3. In the evaluation of the agreement rate of the question-image pairs, there were differences between the expected agreement rate that was calculated through the answers found in the annotator file of VQA and the calculated agreement rate from the answers collected in the user study.

I performed a Mann-Whitney U-test to check whether the differences between the calculated agreement rate from my user study and the agreement rate I calculated from the answers in the annotator file had a significant difference. The p-value of the test was over 0.05, indicating there were no significant differences between my calculations and my prior calculations.

The standard deviation of the agreement rate was 0.279.

There were confidence rates that were lower than the posed criteria in the stimuli selection, which was at least over 0.7 when calculating the mean value. In table 1, the question-image pair with ID 2 and ID 6 had the lowest confidence rate, which was 0.55.

### 4.2 Mental State References Analysis

In the quantitative analysis, I performed the automatic MS method that regarded the verb used with the help of the master list that contained verbs that indicated mental state references as previously defined in the section 3.3.

Explanations that did contain a verb from the master list, like "be excited" appearing in the text shown below, were counted as mental state references. The following explanation is one of the explanations given to answer the question image pair ID 13 "Is the kid crying?".

"I can't see any tears, and the child seems to be excited about the umbrella it holds."

In the analysis where I manually searched the mental state references, the ratio differed by two more participants making mental state references without using the identified verbs. In these instances, I accounted for sentences that mentioned "in my experience" or "I have seen other people do", making a mental state reference to their experience and memory. A sample, from the

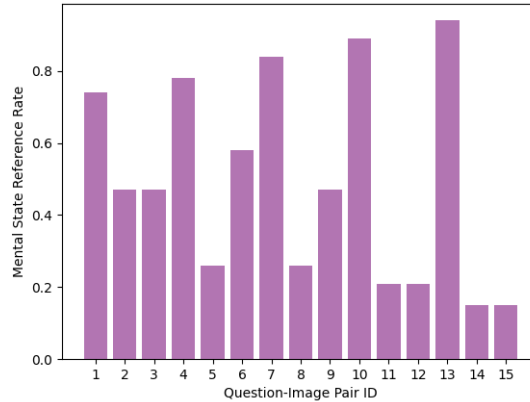


Figure 4: Bar chart depicts mental state references in each question-image pair.

question image pair ID 4, of an explanation with mental state references without the verbs from the master list looked like this:

”She stares very intently, her head slightly nicked to the side, with her mouth lacking and large expressions. Other people I’ve seen concentrating look like that.”

The participant made a note of how they saw other people make a similar face to justify their answer ”concentrated” when they replied to the question. This explanation contained an assessment of the emotional state of the person depicted by mentioning that ”she is staring intently”. The participant made also a mental state reference about their mental state, MSR, by using the words ”other people I have seen... look like that”.

The mental state references, which were identified as Theory of Mind were only mentioned in the sentiment recognition questions. Their value was greater than the number of identified MSR references because it was made with the association of the Theory of Mind. The explanation sample shown prior was a prime example that contained MSR and Theory of Mind.

Participants directed their assumption to make a mental state reference about the figure depicted in the image.

If I counted only the numbers of MSR appearing in explanations as my mental state reference rate, I would receive approximately 0.1 for each sentiment recognition question in table 1 with the Manual MS method. Around two participants explicitly stated their experience to be counted as MSR. Theory of Mind appeared 100% in the identified mental state reference from the Automatic MS method and Manual Ms method.

Any finding of STOMP outside the sentiment recognition questions where the participant assessed the mental state of a figure depicted in the image without being prompted to barely appeared unless one took into account assessing the emotional state of a bird in the question-image pair IDs 12 and 15 of the people holding a surfboard.

”She is looking at food, it seems; her posture is very calm and focused.”

The explanation above is from question-image pair ID 12.

”Looks like they are having fun on the beach and are trying to surf with their surfboards on a nice sunny day in Australia or somewhere where the climate is similar.”

This explanation is from question-image pair ID 15. There could be no large value supplied to STOMP as a mental state reference rate other than 0.05 for question-image pairs ID 15 and 12. It was part of one-third of the discovered 0.15 mental state reference rate in the question-image pair ID 15 in the table 1.

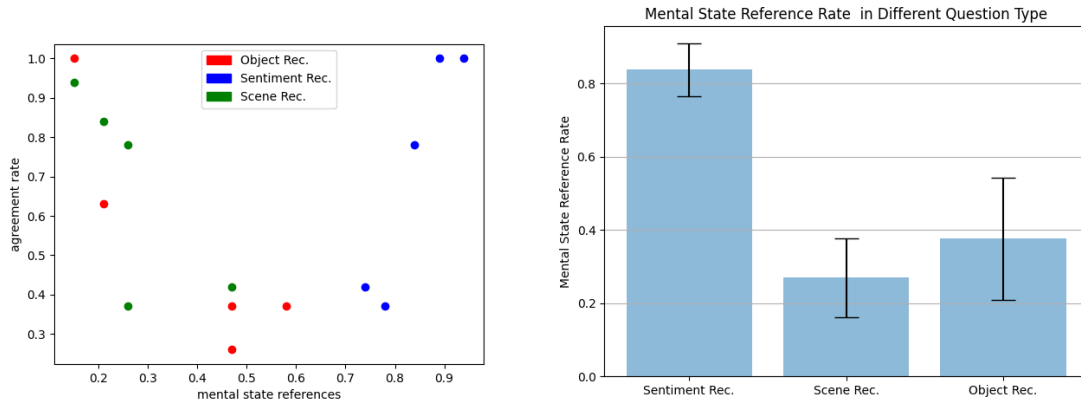


Figure 5: The image on the left displays a scatter plot between agreement rate and mental state references with the values from the table 1. The right graphic showcases the average amount of mental state references with an error bar in each question type

The most common verb used was "think" with 79 instances and "assume" with around thirty instances in the collected explanations from the user study. The number of mental state references varied throughout the question-image pairs, as figure 4 shows. The standard deviation of the results in the Automatic MS method is 0.28 and 0.27 for the Manual MS method. The sentiment recognition questions have a higher value for mental state references. The values are depicted in table 1. The significant differences will be proven in the next subsection.

The number of mental state references varied throughout the question-image pairs. The standard deviation of the results in the Automatic MS method is 0.28 and 0.27 for the Manual MS method. The sentiment recognition questions have a higher value for mental state references. The values are depicted in table 1. The significant differences will be proven in the next subsection.

### 4.3 Hypothesis: Mental State Reference Depends on a Specific Type of Question

For the hypothesis to showcase what type of questions had the most mental state references, I performed a Friedman Test as the data wasn't independent due to participants answering in a certain order of questions. The mean value of how likely a participant made a mental state reference for the question type was calculated as a vector. For the three vectors of sentiment-, scene-, and object recognition, I received a p-value below 0.05,  $2.07 * 10^{-6}$  to be exact. Therefore, I performed a post hoc test through the Wilcoxon signed-rank test. The results from the test showed that the comparison between the sentiment recognition type of questions and object recognition had a p-value of 0.0026 which was lower than 0.05, making the result significant that there was a difference between the data. A similar result of this happened between the sentiment question and the scene recognition question, with a p-value of  $1.14 * 10^{-5}$ . The comparison between the scene recognition and object recognition questions had a p-value of 0.16, indicating no significant difference. The figure 5, specifically the right image, showcases the differences between each question type.

#### 4.4 Hypothesis: Agreement Rate Depends on the Area a Participant Looked on an Image



Figure 6: Samples of summarized click patterns placed on the original image. The left image portrays the click pattern of the question image pair ID 6. The right one portrays question image pair ID 15

Prior discussed in the method section 3.4, I calculated the values on how similar the click patterns were to other participants for each question-image pair and received values between 0 and 1. To see an example of a click pattern, figure 7 visualises two. The full click pattern similarity score can be viewed in the appendix, specifically the table 2. Due to the result of how there was a potentially significant difference between the types of questions, I performed the Pearson correlation coefficient tests for each type of question separately to check if there were any differences between them. For the image click similarity data against the agreement rate data, I received the following results: For the results without regard to the types of questions, I received a  $r$  result of 0.15 and a p-value of 0.58. For the emotional recognition question image pairs, I received a p-value of 0.057, scene recognition a p-value of 0.71, and object recognition a p-value of 0.34. None of the results of the tests indicated any significant relationship between agreement rate and click pattern similarities, as the results were above the value of 0.05.

#### 4.5 Hypothesis: The Lower the Agreement Rate, the Higher the Mental State Reference Rate

Finally the main hypothesis of this thesis, I performed the Pearson correlation coefficient on the data of mental state references and agreement rate, visualised in table 1. Similar to testing the prior hypothesis, I performed the Pearson correlation coefficient once without regard to the type of questions and also another time by disregarding the sentiment recognition questions since it had shown that they evoked the most mental state references, regardless of agreement rate.

In the results of the whole data set of values between mental state references and agreement rate, I received a correlation of -0.23 and a p-value above 0.05. Regarding the agreement rates and mental state references that excluded the values from the sentiment recognition questions, I received a higher correlation of -0.843 and a p-value of 0.0022, which is below 0.05 with the data from the automatic MS method, and a p-value of 0.017 with the data from the manual MS method. Both results show a significant relationship between the variables for the case regarding only the object and scene recognition question image pair. The scatter plot in figure 5 visualises the correlation between the agreement rate and mental state references if one disregards the blue dots.



Figure 7: Sample of placed click patterns from question image pair ID 1

## 5 Discussion

This section discusses the meaning of the results from the prior section and also what findings I didn't expect from the user study and my calculations. Then I compare my findings to the related studies mentioned in the related work and my limitations in this study. At the end of the section, I discuss the potential of future work.

### 5.1 Meaning of the Results

For easier identification in this section, I refer to the agreement rate that was calculated from the given answers in the annotation file of the VQA v2.0 data set as the annotator agreement rate, like the one described in section 3.1. From the calculations of the agreement rates, I noted that the question-image pair, which had an annotator agreement rate of 0.2, had a higher agreement rate in comparison. The question image pair with IDs 1 and 2 had an agreement rate of 0.42, which is showcased in the table 1.

The biggest difference in agreement rate that was lower than the annotator agreement was with the question-image pair ID 9 because the annotator agreement rate of this question was 0.6 while it ended up with an agreement rate of 0.37.

For this result of differences, I acknowledged that the number of participants to calculate the agreement rate was 19 and not ten as in the annotation file in VQA. The demographic of the participants was different as the user study took place in Germany, while the majority of the demographic of their annotators in VQA was in the United States. However, despite the differences, the differences weren't significant through the Mann-Whitney U test.

For the lack of any STOMP usage in the explanations provided by the participants in the mental state reference analysis, the reason could be that only two images had humans depicted outside of the sentiment recognition questions.

For the reason that Theory of Mind consisted 100% of the identified mental state references in the sentiment recognition question, could be explained that the skill was asked to be used since the participant had to assess the emotional state of the person depicted in the image.

The small percentage of identified MSR in sentiment recognition questions was because they were only counted when the participant explicitly mentioned they based it on their own experience as the assumption was mostly about the figure depicted in the image.

I noted that by participants trying to keep their answers short and precise as instructed by the user study, there was a potential risk that some answers wouldn't be viewed as the same opinion. The most visual example to showcase that risk was the question image pair ID 1 "Why is the girl smiling?". The provided answers to this, like "heard good news", "talking with family", and "having a phone call" could be considered the same opinion if I took into account the explanation the participants provided because most participants indicated in their explanation further context that they meant the phone call.

Out of my three put-up hypotheses, only two had shown any significance. Mental state references depend on the type of question and it had shown that sentiment recognition questions evoked the most mental state references in comparison to object and scene recognition. The main hypothesis of the relationship between agreement rate and mental state references showed significance if I disregarded the sentiment recognition questions.

This indicated that the questions, which had the potential for a low agreement rate and didn't directly ask for the participant to evaluate the mental state of a figure depicted in the image, evoked mental state references, specifically mental state references about the participant's experience and memories, to answer these types of questions. This could showcase the potential of people using mental state references to answer more difficult questions.

The hypothesis that checked the correlation between the agreement rate and the image click pattern score had no significance and therefore was not linear. The only special case in creating the click pattern map was the question image pair ID 6 with the question "What's in the vase?". The term "vase" was very ambiguous in the context of the image, leaving room to interpret for people what they considered a vase because there was no traditional-looking vase in the image. The mean click pattern of this image showcased no clear focal point, as the image covered by the click pattern appeared blurry in figure 6 in comparison to the click pattern of the question image pair with ID 15 that focused on the surfboards in the image.

This was one of the click maps where people clicked at other areas of the image and came to different conclusions. The other example was the question image pair ID 9, where most users were unsure if the question referred to the inside of the church or the outside.

## 5.2 In Comparison to Previous Studies

In contrast to previous studies mentioned in the related work section, most VQA studies tried to lead solutions by trying to improve the agreement rate from the answers they collected by applying them to different evaluation methods. Either by filtering the question's data to be more unbiased and more precisely worded or by using simpler-worded questions to answer a more complex question[43].

Only one of the recent studies investigated the potential reasons for different answers, however, it suggested a system to differentiate questions that had the potential for one true answer and one that didn't. But the suggestion only showed how they could separate the types of questions, not necessarily a solution to how to overcome them. My results, especially with my first hypothesis, showcased that mental state references were used in questions with low agreement rates. Mental state references appeared in a subjective type of questions like emotions[25]. This added a step on how to solve the problem by considering creating data sets about mental state references and the rationale to answers for a visual question to teach deep-learning algorithms about dealing with questions that allowed for more answers.

## 5.3 Limitations of this Study

The limitation of the study was potentially the number of participants due to the hypothesis, which involved image comparison, as it would likely have had an accurate result if more participants were involved. However, for the main hypothesis like the correlation of agreement rate and mental state references, it was sufficient, especially in comparison to how VQA only had ten annotators to potentially decide the answer to an image-question pair when excluding the sentiment type of questions. Another limitation could be the potential for further categorization of images by putting them in further distinct types, which involved terms like "difficult," "ambiguous," and "subjective," similar to the previous work, which was about "Why does a visual question have different answers" for each question type.

The demographic of the participant wasn't diverse, so there was a chance of cultural differences, especially when the recruitment in VQA through Amazon Mechanical Turk have annotators who were native to the United States and answered these questions in their free time while doing other activities.

## 5.4 Potential Future Research

With the results and the mentioned limitations, there could be future research that focused solely on "ambiguous" and "subjective" question image pairs, already identified in the study "Why does a visual question have different answers?" [15]. This could create a potential data set of explanations that relied on the subjective perspective of a person and be used as a training data set to be automated through a deep-learning model that handled questions with the possibility of several different answers.

In the proposal of the bachelor thesis, there was originally an extra feature in which the user interacts with a "robot" who reveals their mental state with the expectation that the user will react in a similar context. This was to reflect the improvement of human and computer interaction and would be closely tied to the research mentioned in related work about the Theory of Mind. This future research would showcase the benefits of exchanging answers and explanations.

## 6 Conclusion

The intended outcome of this bachelor thesis was to collect various perspectives and reasoning behind the participants' recognition. Through the classification of the analysis, I sought a clearer context on why different answers had been given and also to recognise the involvement of our mental states in our actions and judgement by analysing the explanations participants provided in the user study. This helped in explaining why question image pairs had been perceived differently and also provided a data set of explanations for answers to the visual questions selected out of VQA. The several hypotheses I put up to check the several factors that potentially caused mental state and potentially influenced the agreement rate, specifically in the VQA scenario, showed only two of them had significance according to the calculations to validate the hypotheses.

We learned that the questions that evoked the most mental state references, especially those involved in the Theory of Mind due to the participants being asked directly to evaluate the mental state of the person depicted in the image, were the sentiment-recognition questions.

Because of the results of this hypothesis, the main question was whether there was a significant correlation between participants relying on their mental state to overcome questions that had a high variety of answers; in other words, a low agreement rate was calculated once with and once without the sentiment recognition questions. Without the sentiment recognition questions, there was a correlation between the agreement rate and the number of mental state references, telling me that even without the questions prompting a participant to evaluate someone's mental state in images, participants referred to their own mental state to answer the visual question, especially through their experience and assumptions on what they understood and believed. The significance of the hypothesis indicated the potential benefit of exploring mental state reference in the context of VQA, especially when it involved questions with a low agreement rate.

## References

- [1] Amazon mechanical turk. <https://www.mturk.com/>. Accessed: 2023-02-10.
- [2] Cambridge thesaurus. <https://dictionary.cambridge.org/thesaurus/>. Accessed: 2023-04-12.
- [3] Json file type. <https://fileinfo.com/extension/json>. Accessed: 2022-09-10.
- [4] Natural language toolkit. <https://www.nltk.org/>. Accessed: 2022-08-10.
- [5] Nltk wordnet. <https://www.educba.com/nltk-wordnet/>. Accessed: 2023-5-1.
- [6] Numpy. <https://numpy.org/>.
- [7] thesaurus. <https://www.thesaurus.com/>. Accessed: 2023-04-12.
- [8] Visual Question Answering. <https://visualqa.org/>. Accessed: 2022-09-10.
- [9] VQA leaderboard challenge 2020. [https://visualqa.org/roe\\_2020.html](https://visualqa.org/roe_2020.html). Accessed: 2022-09-30.
- [10] Wordnet Boot. <https://www.holisticseo.digital/python-seo/nltk/wordnet>. Accessed: 2023-05-1.
- [11] Flatten a matrix in python using numpy. <https://www.geeksforgeeks.org/flatten-a-matrix-in-python-using-numpy/>, Aug 2020.
- [12] JUAN E. ADRIAN, ROSA A. CLEMENTE, LIDON VILLANUEVA, and CAROLIEN RIEFFE. Parent-child picture-book reading, mothers' mental state language and children's theory of mind. *Journal of Child Language*, 32(3):673–686, 2005.
- [13] I. Aleksander, W. V. Thomas, and P. A. Bowden. Wisard: a radical step forward in image recognition. *Sensor Review*, 4(3):120–124, Jan 1984.
- [14] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015.
- [15] Nilava Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4270–4279, 2019.
- [16] J R Brown, Nancy S Donelan-McCall, and Judy Dunn. Why talk about mental states? the significance of children's conversations with friends, siblings, and mothers. *Child development*, 67 3:836–49, 1996.
- [17] Alison J. Cawsey, Bonnie L. Webber, and Ray B. Jones. Natural Language Generation in Health Care. *Journal of the American Medical Informatics Association*, 4(6):473–482, 11 1997.
- [18] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2022.
- [19] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. *CoRR*, abs/1901.03729, 2019.



- [20] Susan R. Fussell and Robert M. Krauss. Accuracy and bias in estimates of others' knowledge. *European Journal of Social Psychology*, 21(5):445–454, 1991.
- [21] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, 2017. Language in Vision.
- [22] Boaz Keysar, Linda E. Ginzel, and Max H. Bazerman. States of affairs and states of mind: The effect of knowledge of beliefs. *Organizational Behavior and Human Decision Processes*, 64(3):283–293, 1995.
- [23] Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(5):36, 2017.
- [24] Maruthi Krishna. Explain the mat class in java opencv library. <https://www.tutorialspoint.com/explain-the-mat-class-in-java-opencv-library>.
- [25] Joseph E LeDoux and Stefan G Hofmann. The subjective experience of emotion: a fearful view. *Current Opinion in Behavioral Sciences*, 19:67–72, 2018. Emotion-cognition interactions.
- [26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.
- [27] Fernando Lizcano-Cortés, Jalil Rasgado-Toledo, Averi Giudicessi, and Magda Giordano. Theory of mind and its elusive structural substrate. *Frontiers in Human Neuroscience*, 15, 2021.
- [28] Jennifer Susan McClung and Stephen David Reicher. Representing other minds: Mental state reference is moderated by group membership. *Journal of Experimental Social Psychology*, 76:385–392, 2018.
- [29] Sonisilpa Mohapatra and James C. Weisshaar. Modified pearson correlation coefficient for two-color imaging in spherocylindrical cells. *BMC Bioinformatics*, 19(1):428, Nov 2018.
- [30] A. Miranda Neto, A. Correa Victorino, I. Fantoni, D. E. Zampieri, J. V. Ferreira, and D. A. Lima. Image processing using pearson's correlation coefficient: Applications on autonomous robotics. In *2013 13th International Conference on Autonomous Robot Systems*, pages 1–6, 2013.
- [31] Ram Isaac Orr and Michael Gilead. Development and validation of the mental-physical verb norms (mpvn): A text analysis measure of mental state attribution. *Behavior Research Methods*, Jul 2022.
- [32] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4):1643–1662, Aug 2022.
- [33] David Premack and Guy Woodruff. Does a chimpanzee have a theory of mind. *Behavioral and Brain Sciences*, 1:515 – 526, 12 1978.
- [34] Katherine Rice and Elizabeth Redcay. Spontaneous mentalizing captures variability in the cortical thickness of social brain regions. *Social cognitive and affective neuroscience*, 10 3:327–34, 2015.

- [35] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1), 1988.
- [36] Jonathan Sachs. Digital image basics. *Digital Light & Color*, 1996, 1999.
- [37] P J Schwanenflugel, R L Henderson, and W V Fabricius. Developing organization of mental verbs and theory of mind in middle childhood: evidence from extensions. *Dev Psychol*, 34(3):512–524, May 1998.
- [38] Yi-Tai Seih, Susanne Beier, and James W. Pennebaker. Development and examination of the linguistic category model in a computerized text analysis method. *Journal of Language and Social Psychology*, 36(3):343–355, 2017.
- [39] Yi-Tai Seih, Susanne Beier, and James W. Pennebaker. Development and examination of the linguistic category model in a computerized text analysis method. *Journal of Language and Social Psychology*, 36(3):343–355, 2017.
- [40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2014.
- [41] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017. Language in Vision.
- [42] Ming Yan, Haiyang Xu, Chenliang Li, Junfeng Tian, Bin Bi, Wei Wang, Weihua Chen, Xianzhe Xu, Fan Wang, Zheng Cao, Zhicheng Zhang, Qiyu Zhang, Ji Zhang, Songfang Huang, Fei Huang, Luo Si, and Rong Jin. Achieving human parity on visual question answering, 2021.
- [43] Yuanyuan Yuan, Shuai Wang, Mingyue Jiang, and Tsong Yueh Chen. Perception matters: Detecting perception failures of vqa models using metamorphic testing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16903–16912, 2021.
- [44] Ma Z, Zheng W, Chen X, and Yin L. Joint embedding VQA model based on dynamic word vector. *PeerJ Comput Sci.* 2021 Mar 3;7:e353. doi:, 10., 2021.



## A Appendix

### A.1 Click Pattern Map on Stimuli Images and their Click Pattern Similarity Score

CPS= 0.65 for ID 1	CPS=0.41 for ID 2	CPS= 0.57 for ID 2
CPS= 0.57 for ID 4	CPS=0.3 for ID 5	CPS= 0.11 for ID 6
CPS= 0.57 for ID 7	CPS=0.52 for ID 8	CPS= 0.36 for ID 9
CPS= 0.38 for ID 10	CPS=0.26 for ID 11	CPS= 0.34 for ID 12
CPS= 0.49 for ID 13	CPS=0.19 for ID 14	CPS= 0.57 for ID 15

Table 2: Table showcases the summarised click pattern on their image and Click Pattern Similarity CPS.



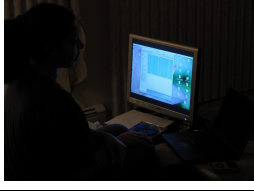
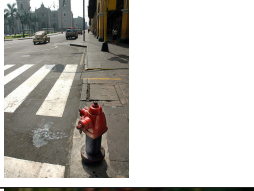
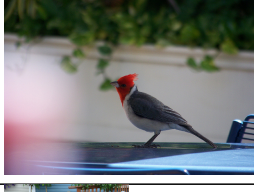
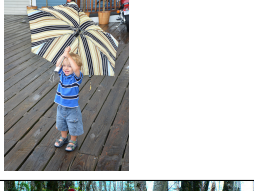


## A.2 Code to calculate Click Pattern Similarity Score

```
1 #calculate PCC by comparing every 19 images with each other
2 #saves the r results in a 19x19 matrix
3 def blurccorr(imgfolder):
4     fullmatrix=[]
5     for x in range(0,19):
6         array=[]
7         _str1=f"data/img/{imgfolder}/{x}.png"
8         OriImage = Image.open(_str1)
9         blurImage = OriImage.filter(ImageFilter.GaussianBlur).convert('RGB')
10        open_cv_image = np.array(blurImage)
11        open_cv_image = open_cv_image[:, :, :-1].copy()
12        for y in range(0,19):
13            _str=f"data/img/{imgfolder}/{y}.png"
14            OriImage2 = Image.open(_str)
15            blurImage2 = OriImage2.filter(ImageFilter.GaussianBlur).convert('RGB')
16            open_cv_image2 = np.array(blurImage2)
17            open_cv_image2 = open_cv_image2[:, :, :-1].copy()
18            cm = np.corrcoef(open_cv_image.flat, open_cv_image2.flat)
19            r=cm[0,1]
20            array.append(round(r, 3))
21
22        print(array)
23        fullmatrix.append(array)
24    return fullmatrix
25
26 def averagecorr(array:list):
27     #receive upper triangle of matrix
28     uppertriangle= list(array[np.triu_indices(19)])
29     while( 1 in uppertriangle): #remove self comparison values
30         uppertriangle.remove(1)
31     pearsonmean=np.mean(uppertriangle)
32     print(f'average of pearsoncorr: {pearsonmean}')
```

Figure 8: Python Code for calculating PCC of every click pattern

### A.3 Selected Question Image Pair Stimuli

ID	Image	Question	Annotator Agreement rate	Type
1		"Why is this girl smiling?"	20%	Sentiment
2		"What is the large pole sticking up for?"	20%	Scene
3		"What is the truck carrying?"	20%	Object
4		What is her expression?	40%	Sentiment
5		How is the weather?	40%	Scene
6		What's in the vase?	40%	Object
7		Is she sad?	60%	Sentiment

8		Has the bed been made?	60%	Scene
9		What is displayed under the triangular dome?	60%	Object
10		Is this woman happy?	80%	Sentiment
11		Does this hydrant look like it is in the wrong place?	80%	Scene
12		What is the bird perched on?	80%	Object
13		Is the kid crying?	100%	Sentiment
14		What season was this picture taken in?	100%	Scene
15		What are the people holding?	100%	Object

#### A.4 Master List of Verbs of Mental State References

"want"	"hope"	"wish"
"care"	"be pleased to"	"be tempted to"
"be interested in"	"be bothered to"	"be keen on"
"look forward to"	"be bored of"	"have a crush on"
"be mad about"	"desire"	"fancy"
"miss"	"need"	"enjoy"
"be fond of"	"hate"	"abhor"
"detest"	"loathe"	"hold in contempt"
"prefer"	"feel"	"be hurt"
"be angry"	"be happy"	"be excited"
"be sad"	"love"	
"like"	"dislike"	"be afraid"
"enjoy"	"have fun"	"be glad"
"be mad"	"be scared"	"be upset"
"be surprised"	"fear"	"be disgusted"
"worry"	"be anxious"	"be relieved"
"be shocked"	"be disappointed"	"be nervous"
"be sad"	"think"	"know"
"believe"	"wonder"	"remember"
"forget"	"guess"	"pretend"
"understand"	"expect"	"have a clue"
"be confused"	"notice"	"assume"
"find out"	"underestimate"	"agree"
"be sure"	"make sense"	"disagree"
"be able to relate"	"judge"	"be determined"
"be only joking"	"accept"	"mean"
"be serious"	"realise"	"recognise"
"learn"	"have an idea"	"be conscious of"
"imagine"	"reckon"	"fathom"
"figure"	"plan to"	"lie"
"be sorry"	"decide"	"choose"
"trust"	"look (serious/angry/sad)"	"seem (serious/angry/sad)"
"be intelligent"	"esteem"	"admire"