Universiteit Leiden

Perception of Ambiguous Tangram Shapes by Vision-Language Models

Sofie de Graaf (s3698300)

Supervisors:
Derya Soydaner & Mike Preuss

Bachelor Thesis– Informatica

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 1, 2026

Abstract

Vision-Language Models (VLMs) have achieved strong performance in a wide range of vision-language tasks, such as image captioning and visual question answering. However, it is still not entirely clear how skilled they are at interpreting abstract visual representations, and whether their visual reasoning aligns with that of humans.

In this thesis, we investigate how VLMs perceive ambiguous tangram shapes under different visual conditions, including rotations, colour variation, and variations in structural representation (tangrams as full silhouettes versus having visible pieces). We compare their outputs to human annotations from the KiloGram dataset to reason about human similarity.

We evaluate three VLMs (BLIP-2, LLaVA 1.6, and Qwen3-VL) using measures of exact match consistency and semantic similarity based on sentence embeddings, as well as qualitative analyses of interpretation categories.

Our results show that VLMs are sensitive to visual transformations, in particular rotation, and show model-specific patterns of stability and variation. While interpretations of the abstract tangrams sometimes align with human annotations in human-like shapes, performance decreases when the shapes become more abstract. Overall, the findings highlight differences between the interpretations of ambiguous shapes by humans and VLMs, and show that model behaviour is strongly influenced by the visual structure and representation of the tangrams.

Contents

1 Introduction

Vision-Language Models (VLMs) show strong performances across a wide range of vision-language benchmarks, including image captioning, visual question answering, and multimodal reasoning tasks. These models combine visual encoders with large language models, enabling them to process textual as well as visual information.

Despite their success on various benchmarks ([14, 15]), several studies suggest that VLMs still struggle with abstract visual reasoning and spatial cognition tasks. In particular, tasks that require interpretations of simplified or non-realistic visual input often reveal limitations in the way these models reason about visual structure.

One type of visual input that is well-suited for studying such limitations is the tangram puzzle. Tangrams consist of simple geometric shapes that can be combined into a wide range of ambiguous figures, which makes them a useful and flexible tool for investigating visual perception and interpretation in both humans[4] and AI systems[10].

In this thesis, we investigate how VLMs perceive tangram shapes under different visual conditions, including rotations, colour variation, and changes in their internal structure. We further compare model outputs to human annotations from the KiloGram dataset[10] in order to evaluate how similar VLM interpretations are to human perception.

We carry out a series of experiments using three vision-language models: BLIP-2, LLaVA 1.6, and Qwen3-VL. Our analysis focuses on: similarity to human interpretations, robustness to visual transformations, and the influence of visual structure on model outputs.

The results provide insight into how VLMs interpret abstract shapes and highlight differences in how models respond to changes in visual input. Overall, this work aims to better understand the capabilities of current VLMs in abstract visual reasoning tasks.

1.1 Research Question

In order to get a better understanding of the perception of ambiguous abstract shapes by VLMs, this research aims to answer the question:

> *In what ways do Vision-Language Models perceive ambiguous tangram shapes similarly to humans?*

With the following subquestions:

SQ1. Do VLMs perceive tangram shapes in a similar way to humans?

SQ2. Do rotations affect the consistency of VLM interpretations of tangram shapes?

SQ3. Does color influence the perception of a tangram shape by VLMs?

SQ4. Does the visibility of the individual tangram pieces change the VLMs perception?

1.2 Thesis overview

Following this introduction, Section 2 provides background on VLMs and tangram puzzles. Section 3 discusses literature related to the visual reasoning capabilities of VLMs. Section 4 describes the used dataset, models and the setup of the experiments. Section 5 describes the individual experiments and their outcomes. Section 6 discusses the results from the previous section and relates them to the subquestions. Section 7 concludes the findings and suggests future work.
This bachelor thesis is carried out under the supervision of Dr. Derya Soydaner and Dr. Mike Preuss, as part of the Bachelor Thesis Project of the Computer Science programme at LIACS.

2 Background

2.1 Vision-Language Models

Vision-Language Models (VLMs) are AI models that are not restricted to one type of input, such as text, but are capable of processing both text and image input at the same time. Since they process more than one type of information, they are a type of Multimodal Large Language Model (MLLM). They are generally made up of a Large Language Model (LLM) together with a Vision Encoder[2]. This setup allows them to generate text responses to given prompts, causing them to be used in tasks such as image captioning and visual reasoning.
There are several surveys that analyse existing research on VLMs. For example, Zhang et al.[22] conducted a survey focusing on the application of VLMs for various visual recognition tasks, such as image classification and object detection, while a survey by Danish et al.[5] reviews the progress of VLM optimisation.

2.2 Tangram

The tangram puzzle, originally from China, is an abstract geometric puzzle. It belongs to the type of dissection puzzles, which have the goal of putting tiles together to form geometric shapes. Tangram consists of seven pieces: two large triangles, one medium triangle, two small triangles, a square, and a parallelogram. Following the same principle as dissection puzzles, the goal of tangram is to put these seven pieces together to create a full shape. But, for tangram, the full shape is already given as a silhouette, which the player has to completely fill with the pieces, without any overlap between them. An example of a tangram is given in Figure 1.

The abstract nature of tangram figures means that they can be interpreted in a variety of ways, without having a true 'correct' interpretation. As a result, tangrams are often used to study a range of cognitive processes. For example, Clark & Wilkes-Gibbs[4] studied how two individuals can align their perceptions of abstract shapes by having them describe tangram shapes, observing how their conceptualisation of the shapes evolved when describing it to the other person. Hawkins et al.[7] continued this study on a larger scale to study how human brains decide on naming shapes, in order to create a dataset for AI models. Another study by Hu et al.[9] used tangrams to study the influence of creative thinking and spatial reasoning on brain activity. Furthermore, Judd & Klingberg[11] studied whether training childrens' spatial cognition would improve their mathematical abilities. They used tangram tasks to evaluate the mental rotation and spatial

transformation abilities of the children, by having them rotating, flipping and sliding the tangram shapes.

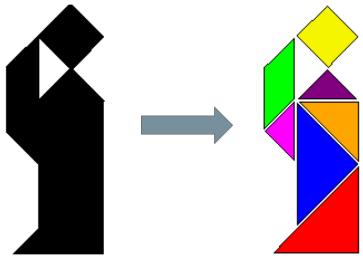


Figure 1: Example of a tangram puzzle (tangram shape originally taken from the KiloGram dataset[10] and modified to create this example). On the left is the empty silhouette that has to be filled with the pieces as shown on the right.

3 Related Work

3.1 Visual Reasoning in VLMs

The rapid development of VLMs goes hand in hand with an increasing amount of research on evaluating their visual reasoning capabilities. Although models show strong performance on tasks such as image captioning and visual question answering, their capabilities seem to be limited on abstract visual tasks.

When it comes to abstract visual reasoning tasks, studies show that VLMs have difficulty giving accurate answers. Cao et al.[3] report that, in comparison to human reasoning, MLLMs lack a similar level of intuitive reasoning in abstract visual tasks, while Ji et al.[10] observe comparable difficulties when evaluating models using tangram puzzles. Both studies suggest that current vision-language models rely heavily on concrete pre-learned associations, rather than actual human-like intuitive reasoning.

Several studies suggest biases as explanations for these limitations of VLMs. Hemmat et al.[8] find that VLMs often display a local-focus bias. So unlike a human, who would look at the full picture, VLMs tend to focus on individual components, causing them to miss the full shape. Similarly, Rudman et al.[18] state that VLMs are 'shape-blind', relying on memorized associations rather than deliberate reasoning. Vo et al.[20] address bias in VLMs, after observing that VLMs refuse to change their descriptions after being shown an altered image. However, they find that there is a slight reduction in bias when removing the background of a shape.

There are also multiple studies that investigate the influences of transformations on the VLMs visual perceptions. Eppel et al.[6] show that transformations, such as rotation, worsen the accuracy in abstract shape recognition. Panagopoulou et al.[17] find that image manipulations can increase variance in model responses, even when they had a strong preference for a certain interpretation

before. Yang et al.[21] look further into the reasons why, and show that VLMs are sensitive to pixel-level perturbations and that they often resort to hallucinations, by giving a confident but incorrect answer.

3.2 Previous Tangram-based Evaluations

Since tangram puzzles have a history of being used in evaluations of cognitive processes (see Section 2.2), Ji et al.[10] used them to evaluate abstract visual reasoning in MLLMs. They created a large set of tangrams (the *KiloGram* dataset) and tested models by giving them a textual description of a tangram, for which they had to choose the correct matching tangram. As previously mentioned, they observed that VLMs struggle when they cannot rely on their pre-learned associations to relate to the tangram shapes.
Likewise, Junker et al.[12] used tangrams to reason about perception of abstract shapes, but changing it slightly to include visual contexts. They added backgrounds to tangram shapes and evaluated whether background changes influenced the VLMs perception. However, they found that VLMs lacked the variability to adapt to different scenes. They observed background-specific answers and hallucinations from the VLMs.

3.3 Thesis Relation

These works show that VLMs struggle with abstract visual reasoning and are quick to display biases or hallucinations. They have primarily focused on overall recognition accuracy and the influences of contextual information, giving less attention to the stability of VLM perceptions under controlled transformations, such as rotations, changes in colour and visibility of individual tangram pieces. Therefore, this thesis further explores these aspects by comparing three VLMs on their perceptions of tangrams under several transformations.

4 Methods

We test the selected VLMs against several tangram shapes that vary in their visualizations, after which we compare their performances.

4.1 Dataset

In order to create our dataset, we take tangrams from the existing KiloGram dataset[10]. This dataset contains more than 1000 tangram figures, and each has at least 10 annotations made by different people. This provides us with multiple human-made descriptions per figure.

For this research, we select 20 tangrams from the KiloGram dataset to create our own dataset. We manually divide these tangrams into two subsets of 10 tangrams each: a *human*-like subset and an *abstract* subset. The sets are separated on *human*-like tangrams and *abstract* tangrams. This separation is based on the content of the annotations per tangram. Tangrams in the *human* subset mainly have descriptions referring to human figures, whereas the tangrams in the *abstract* subset have much more variation in their descriptions, without a dominant overall theme. We create these

two subsets so that we can analyse the performance of VLMs on tangrams that have a common semantic interpretation against tangrams without it. We randomly select 20 tangrams from the KiloGram dataset, as long as they fit into one of these subsets.

From these 20 tangrams, we create six dataset variations. First, we focus on geometric transformations and apply a 90° and 180° rotation to each tangram. Second, we create a coloured variant by assigning different colours to each individual tangram piece. We only do this for the original tangram, so without being rotated. Finally, we make silhouette versions of the tangrams, by removing the internal piece boundaries. This means that we fill up the white lines between the pieces to make it one solid figure. The silhouette versions are made for the original, as well as the rotated black tangrams.
By using these specific variations, we create a way to test the VLMs on the features that our subquestions mention (transformations, colour and piece visibility). In total, our resulting dataset contains 140 tangram figures, including the different variants.
An overview of all dataset variants is shown in Figure 2.

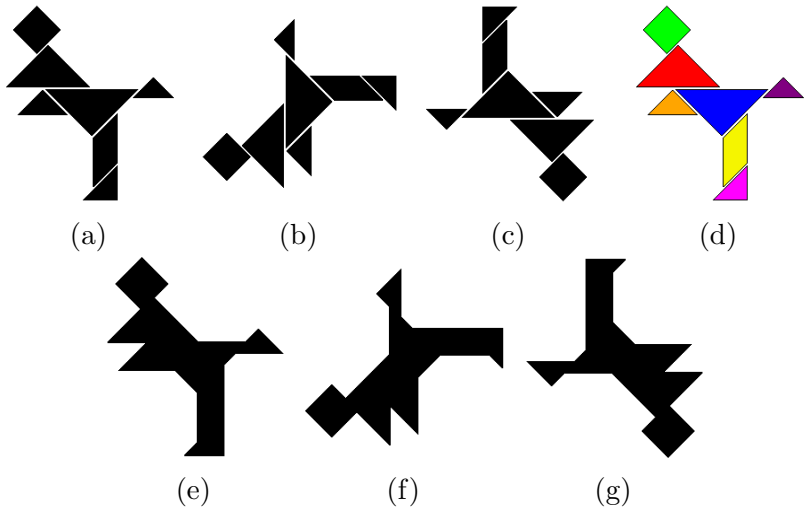


Figure 2: Example of a tangram in all variations. Figure (a) shows the original tangram, taken from the KiloGram dataset. Figures (b) and (c) show the 90° and 180° rotations, respectively. Figure (d) shows the colour variation. Figures (e)-(g) show the full silhouette versions of Figures (a)-(c).

The complete dataset can be found in Appendix A.1, with the *humans* subset in A.1.1 and the *abstract* subset in A.1.2.

4.2 Models

The three models used in this research are: BLIP-2[13], LLaVA[16] and Qwen[1]. Each is an open-source model that combines a vision encoder with a large language model, allowing them to process visual inputs and generate textual outputs.

- **BLIP-2.** BLIP-2 (Bootstrapping Language-Image Pre-training)[1] was introduced by Salesforce

[1]We used BLIP-2 through Hugging Face, specifically `Salesforce/blip2-opt-2.7b`: https://huggingface.co/Salesforce/blip2-opt-2.7b

in 2023. It connects a frozen vision encoder to a frozen large language model by using a Querying Transformer (Q-Former). The Q-Former is a small transformer that selects the most important parts of an image and gives them to the LLM. The frozen components ensure that BLIP-2 can only train the Q-Former, which is more efficient than when it would be able to change all components. This design also allows BLIP-2 to use strong existing models without having to train a lot. Previous work shows that models of the BLIP family show strong preferences for certain interpretations and low sensitivities to image transformations[17].

- **LLaVA 1.6.** LLaVA (Large Language and Vision Assistant)[2], introduced in 2023 and updated to version 1.6 in 2024, is a VLM that combines a CLIP-based vision encoder with a large language model to create visual understanding. Version 1.6, also known as LLaVA-Next, improves the visual reasoning capabilities compared to earlier versions. LLaVA was selected because it is one of the most widely used open-source vision-language models in visual reasoning research. Previous work shows that it is relatively sensitive to image transformations and produces diverse interpretations of ambiguous images [12, 17].

- **Qwen3-VL.** Qwen3-VL[3], released in 2025, is the latest vision-language model developed by Alibaba Cloud and is designed for multimodal reasoning, visual understanding, and image-based instruction following. Compared with earlier Qwen-VL models, it shows strong performance on a wide range of vision-language benchmarks and supports long context multimodal reasoning. Qwen3-VL was included because it represents a recent state-of-the-art open-source VLM. However, previous work indicates that, despite its overall capabilities, it still struggles with abstract visual reasoning tasks such as interpreting abstract shapes [18].

Although each of these models has a temperature variable that allows a user to control the randomness and creativity of their answers, we decided to keep it at the default temperature settings. This reflects their standard behaviour and avoids adding additional experimental variables through manually changing the temperature. We want to compare the models as they are typically used, which is why we do not make model-specific changes to the temperatures. This also means that we do not train or fine-tune the models on the tangram data. We conduct the experiments using the pre-trained models to evaluate their perception abilities.

4.3 Experimental Setup

First of all, we gather outputs from the models for every tangram in all variations. We do this by using the following prompt on the VLMs:

> *"This is a tangram, consisting of 7 black pieces. Describe what the full shape represents. Answer with a single word or short phrase, don't give more explanation."*

The prompt has slight variations depending on which variation is being used, meaning that *"black pieces"* becomes *"coloured pieces"* for the colour variant, and for the silhouettes the prompt begins with *"This is a tangram, without distinctions between the 7 black pieces."*. We use the outputs on

[2]We used the LLaVA 1.6 model (also known as: LLaVA-Next) available through Hugging Face: https://huggingface.co/llava-hf/llava-v1.6-34b-hf

[3]We used the 8-billion-parameter Qwen3-VL model through Hugging Face: https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct.

these prompts in five experiments. In order to study different aspects of VLM shape perception, each experiment focuses on a specific factor to determine if it has a significant influence on the interpretation of the tangrams. This approach aligns with the setup of our subquestions, as each examined factor relates to a subquestion.

A brief overview of the experiments and their focus (they will be explained in more detail in their respective sections):

1. **Human similarity** compares the model outputs for the original tangrams to the human annotations from the KiloGram dataset. It looks at how often models give the same interpretations as humans and how much they agree with each other.

2. **Rotation robustness** tests whether the models' view of the tangram remains consistent under different rotations (90° and 180°). We evaluate the individual robustness and see how this differs between the models.

3. **Colour influence** inspects whether changing the colours of the individual tangram pieces has an effect on the output. We determine the effect by comparing the outputs for the colour variant to the outputs for the original tangrams.

4. **Visible pieces versus silhouettes** compares the model outputs for tangrams with visible piece boundaries against full silhouette versions. With this experiment, we want to discover whether the visibility of the pieces actually distracts them from viewing the full tangram as one shape.

5. **Pose Descriptions** is an additional experiment that tests the effect of including more context in the prompt. Here, we tell the VLMs in advance that the tangrams resemble human figures and ask them to describe their poses. We consider it an additional experiment because it does not have a direct relation to our subquestions.

5 Experiments & Results

5.1 Experiment 1 - Human Similarity

5.1.1 Setup

This experiment inspects whether VLMs interpret tangram shapes in a way similar to that of humans. We compare the original tangrams of both subsets (*human & abstract*) with the human annotations from the KiloGram dataset. Afterwards we also compare the model outputs to each other.

The KiloGram set contains at least ten annotations per tangram, which are generally more descriptive than the model outputs, because we gave the VLMs a prompt that asked them to answer in just a single word or short phrase. Additionally, the fact that the participants of the KiloGram dataset were asked to relate their descriptions to the individual pieces of the tangram contributes to the reason why their descriptions are longer and more detailed.
Therefore, a direct comparison between the both is not straightforward. So in order to compare the

VLM outputs to the human annotations, we define and assign general categories to each tangram, to avoid comparing a single word to very detailed descriptions. This generalization is done by going through each annotation per tangram and selecting the most commonly used descriptions. From this selection, a general category is decided that encapsulates the majority of annotations of the tangram. Figure 3 shows an example of the categorization of the annotations.
In the case where we could not find a term that corresponded to the majority of the annotations, the most common annotation is selected as category instead.

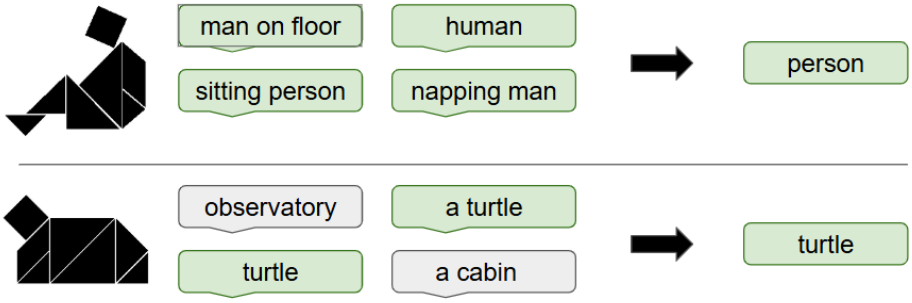


Figure 3: The categorization of the human annotations. We take the most commonly used descriptions, and decide the general theme that they all belong to (above). Otherwise, when there is no general term, the most common annotation is chosen instead (below).

5.1.2 Results

The complete comparison of the human categories against the model outputs is shown in Table 1 for the human set, and Table 2 for the abstract set. Figure 4 summarizes the percentage of outputs that match the human categories, together with the percentage of outputs that exactly match one of the original human annotations.

The most prominent result is the difference between the two tangram subsets. All three models show substantially higher agreement with the human categories for the *human-like* tangrams than for the *abstract* tangrams. For the human subset, BLIP-2 and Qwen match the human category for 9 out of the 10 tangrams, while LLaVA matches 5 out of 10. This is a large contrast compared to the abstract subset, where the agreement with the human categories almost completely disappears: BLIP-2 and Qwen do not match any categories, whereas LLaVA matches only one. These results indicate that all three VLMs are better at matching the human interpretations when a tangram resembles a human figure than when it represents a more abstract shape.

The models exhibit similar interpretation styles for the human tangrams, with LLaVA as an exception. We see that BLIP-2 and Qwen are highly consistent in their outputs, almost always describing the tangram simply as a *person*. However, LLaVA shows more variety in its outputs, and is more often descriptive in its answers than BLIP-2 or Qwen. This is seen in outputs such as *person sitting*, *abstract design*, and *man holding fist up*. The additional variety in LLaVA's responses results in lower agreement with the human categories.

Tangram	Human Category	BLIP-2	LLaVA 1.6	Qwen3-VL
	person	person	Figure	person
	person	person	Person sitting	sitting person
	person	person	Bird	dog
	person	triangle	Tree	person
	person	person	Person	person
	person	person	Abstract design	person
	person	person	Man holding fist up	person
	person	person	Robot	person
	person	person	Man	person
	person	person	Humanoid	person

Table 1: Model outputs for each tangram in the *human* subset, next to the selected category for the human annotations. BLIP-2 and Qwen show similar behaviour and often match the human category, while LLaVA has more diversity in its outputs.

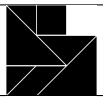
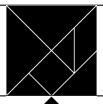
Tangram	Human Category	BLIP-2	LLaVA 1.6	Qwen3-VL
	chair	square	Rectangle	square
	snake	bird	Vase	swan
	envelope	square	Triangle	square
	spaceship	triangle	Person	house
	turtle	dog	Dog house	dog
	tank	mountain	Puzzle	square
	tree	triangle	Words	bird
	scythe	question mark	Pipe	question mark
	foot	triangle	Cube	house
	House	triangle	House	N

Table 2: Model outputs for each tangram in the *abstract* subset, next to the selected category for the human annotations. All models show much variation and have almost no overlap with the human category.

Similarity to human annotations	BLIP-2	LLaVA 1.6	Qwen3-VL
Human (category)	90%	50%	90%
Human (exact)	40%	10%	30%
Abstract (category)	0%	10%	0%
Abstract (exact)	10%	10%	20%

Figure 4: Figure showing the level of agreement between the model outputs and the human categories *(category)*, as well as the amount of exact matches of model outputs to any of the human annotations for a tangram *(exact)*. For the human tangrams, the model outputs are much more similar to the categories and have more exact matches than for the abstract tangrams.

For the abstract tangrams, all models produce a much wider variety of interpretations than for the human tangrams, which results in less agreement with the human categories, rather than more. We see that the models appear to favour different types of descriptions. BLIP-2 frequently describes the tangrams as geometric shapes (*square* and *triangle*), while Qwen also mentions more animals in its descriptions (*swan*, *dog*, *bird*). LLaVA, again, has the most diversity in its outputs, describing the tangrams largely as different objects (*vase*, *words*, *cube*). These differences suggest that the models rely on different strategies when interpreting abstract shapes that lack a clear human-like structure (as in the *human* subset).

Finally, for the human subset, exact matches with the original human annotations (from the KiloGram dataset) occur less frequently than category matches. BLIP-2 had four exact matches, Qwen three, and LLaVA one. This is a large drop from their category matches. For the abstract tangrams, we observe a different pattern. Although BLIP-2 and Qwen do not match any of the selected human categories, both models produce outputs that exactly match some less common human annotations. Consequently, their number of exact matches is slightly higher than the number of category matches. For example, both BLIP-2 and Qwen describe one tangram as a *question mark*, which corresponds to one of the original human annotations, but not to our selected category used in this evaluation. So, while the models mostly do not agree with the human categories for the abstract set, they occasionally produce interpretations that match less common human annotations.

If we look at the outputs of all models for the abstract set, we see that, even though the models do not agree with the human categories, they do have some agreements between themselves. For example, while the human category is *chair*, or *envelope*, BLIP-2 and Qwen agree that it is a *square*, while LLaVA mentions *rectangle* and *triangle*. Another example is where the category is *turtle*, and all models give a *dog*-related answer. This shows that the models sometimes have a similar way of interpreting, even when this differs from human perception.

Overall, the results show a clear trend. The VLMs are capable of matching human interpretations when a tangram resembles a clear figure, such as a person. As the tangram shape becomes more

abstract, the models have less agreement with human interpretations and among themselves.

5.2 Experiment 2 - Rotation Robustness

5.2.1 Setup

This experiment evaluates whether the VLMs produce consistent interpretations when a tangram is presented under different rotations. We compare the tangrams in three different rotations: 0°(original), 90°, and 180°. We measure rotation robustness using two metrics: an exact matching score and a semantic similarity score.

- **Exact matching score.** This metric measures how often a model produces identical outputs for all three rotations of the same tangram. The exact matching score is calculated for each tangram and then averaged across all tangrams in the set to get the final overall score. This method is based on a metric used in previous works, where it is used to compare similarity in answers.[4] We adapted it to fit this specific situation.

- **Semantic similarity score.** The exact matching score does not account for outputs that have different wording but similar meanings (such as *human* and *person*). Therefore, we additionally measure the semantic similarity between outputs. We use a pre-trained Sentence Transformer model (`all-mpnet-base-v2`[19]) to compute the embeddings of the model outputs. Between the embeddings for the three rotations, we calculate the pairwise cosine similarities, which tell us how closely related the outputs are. We obtain the final score by averaging these values across all tangrams in the set.

Using both of these metrics provides a more complete evaluation of rotation robustness. The exact matching score measures strict output consistency, whereas the semantic similarity score captures whether changes in wording correspond to actual changes in interpretation.

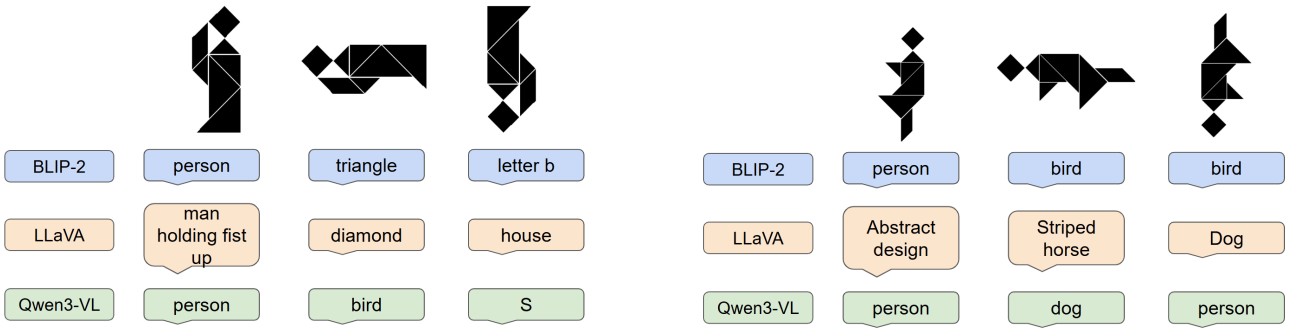
5.2.2 Results

Figure 5 illustrates examples in which the outputs of the three VLMs are given for all rotations of two tangrams. One for which the interpretations change considerably after each rotation, and one where only LLaVA shows a lot of change. Figure 6 summarizes the robustness of the three VLMs to rotations using the exact matching score, and the semantic similarity score.

For the exact matching measure, visible in Figure 6a, the plot shows that BLIP-2 and Qwen achieve similar scores on the dataset, while LLaVA consistently scores lower, except on the *human silhouette* subset. This indicates that LLaVA is quicker to change its interpretation of a tangram under a different orientation. Because the exact matching score cannot distinguish between a change in wording and a change in interpretation, we use the semantic similarity scores to determine whether these output changes also correspond to semantic changes. While BLIP-2 and Qwen score similarly to each other, BLIP-2 has slightly higher scores more often than Qwen. The highest score is obtained by BLIP-2. Therefore, it seems that overall, BLIP-2 is the most consistent in its answers, LLaVA the least, and Qwen is positioned in between. The only exception to this are the results

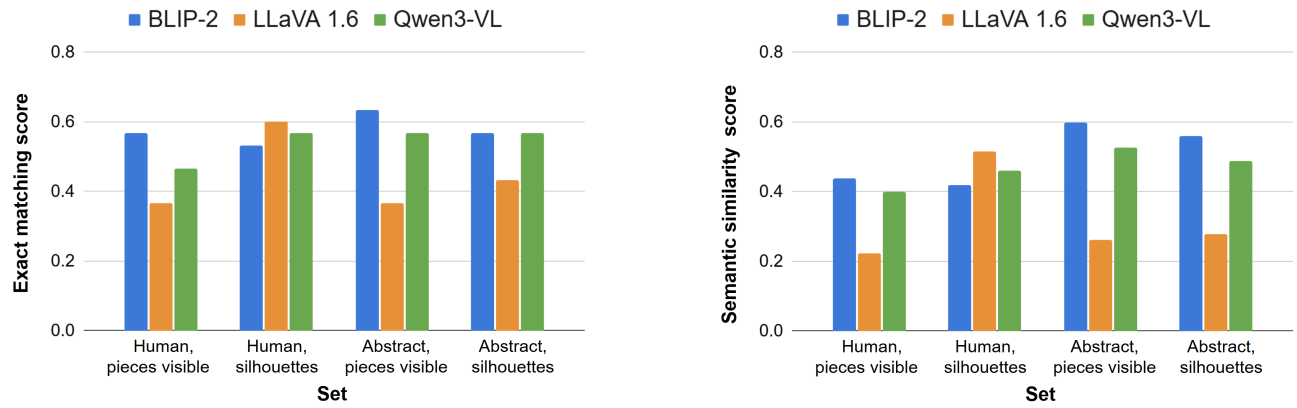
[4] It is based on the Shape Naming Divergence (SND) score used in work by Ji et al.[10].

for the *human silhouette* set, which has a sudden spike in LLaVA's scores and a drop in those of BLIP-2.



(a) (b)

Figure 5: Example of a tangram in different rotations (0°, 90°, 180°) and the answers of the VLMs. For each model the interpretation changes a lot when rotated (a). There are also situations where LLaVA changes a lot, while the other models stay rather consistent (b).



(a) (b)

Figure 6: Figure showing the exact matching scores (a) and the semantic similarity scores (b) for the different sets (*human & abstract*). We see that BLIP-2 and Qwen are generally more robust under rotations, whereas LLaVA is much less robust for both sets, with a single exception for the silhouette variants of the human tangrams.

Figure 6b shows the semantic similarity scores. The general trend agrees with that of the exact matching scores: BLIP-2 and Qwen are semantically more stable in their outputs than LLaVA. This suggests that, even when the models change their wording of a description, BLIP-2 and Qwen remain in related categories for their output, while LLaVA diverts further away. By inspecting the outputs we see situations where, for example, LLaVA goes from a human figure to a bird, vehicle or entirely different object in every rotation, while BLIP-2 and Qwen remain more consistent. An example of this is illustrated in Figure 5b. The same exception occurs as in the exact matching results, namely that the score ratios turn around for the *human silhouette* subset.

When comparing the scores of the human set with those of the abstract set, the figures show that, on average, the scores for the abstract set are slightly higher. This is the case for both score metrics, although the difference is slightly larger for the semantic scores. Inspection of the model outputs suggests that this could be caused by the nature of their descriptions. Abstract tangrams more often receive a simple geometric description from the VLMs (such as *triangle, square*), which stays stable under rotation, or changes to another geometric shape. Changes such as these stay semantically close to the previous description, resulting in higher robustness scores than when they describe a new, unrelated object each time, which happens more often for the human tangrams.
Often, for the human set as well as the abstract set, BLIP-2 and Qwen reinterpret the tangrams as letters after rotation. This shows that the models also switch from objects or shapes to symbol recognition.

In general, the two evaluation measures lead to the same conclusion. BLIP-2 is the most rotation-robust, closely followed by Qwen, whereas LLaVA is more sensitive to changes in orientation. However, higher robustness does not automatically equal better visual understanding.
LLaVA scores higher on the full silhouette tangrams, with a spike for the human silhouettes. So, LLaVA becomes more rotation robust when internal piece boundaries disappear.
Overall, the results demonstrate that the perceptions of tangrams by VLMs are not invariant to rotations. Although BLIP-2 and Qwen maintain more consistent descriptions than LLaVA, all three models frequently change their descriptions when the orientation of a tangram changes.

5.3 Experiment 3 - Colour Influence

5.3.1 Setup

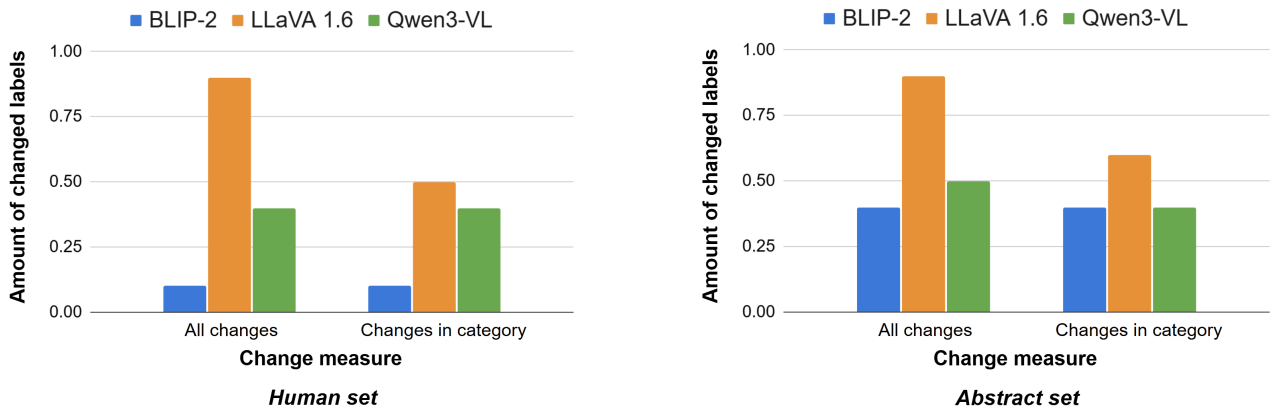
This experiment investigates whether colouring the individual tangram pieces influences the interpretations generated by the models. For each tangram, we compare the model output for the original image with the output for the corresponding coloured version. The colour assigned to each tangram piece is not fixed across tangrams. Since the objective of this experiment is to determine whether the presence of colour influences the interpretation, rather than the effect of a specific colour assignment, we do not expect this variation to affect the analysis.

To assess the influence of colour on the interpretations, we first determine whether the interpretation changes after colouring the tangram. Then we calculate the percentage of tangrams for which the interpretation changes for each model.

Because this experiment focuses on identifying interpretation changes and not the consistency in output, the outputs are analysed manually instead of using score metrics. Following the categorisation procedure described in Experiment 1, outputs that belong to the same general semantic category are considered equivalent (such as *human* and *figure*) and are therefore not counted as changes. Outputs belonging to different semantic categories are recorded as changes and analysed.

5.3.2 Results

The measured amount of changes per subset can be seen in Figure 7, which puts the total amount of label changes (left side) next to the amount that remains when we remove all of the changes that stay in the same category (right side). Figure 7a shows these results for the human set, Figure 7b shows them for the abstract set.



(a) (b)

Figure 7: The amount of description changes for coloured tangrams. (a) shows the amount of changes for the human set, while (b) shows them for the abstract set. Both have the total amount of changes on the left side, and the amount of changes when considering category (so only the label changes that do not stay in the same category) on the right. BLIP-2 and Qwen have little to no changes in output after addition of colour, while LLaVA has much more. Overall, the models have more changes in output for the abstract tangrams.

The most prominent result is that colouring the individual tangram pieces has only a limited effect on BLIP-2, while LLaVA and Qwen are more influenced by it. Across both sets, BLIP-2 shows the least amount of changes, followed by Qwen, with LLaVA changing the most frequently, indicating that it is the most sensitive to colour.

After grouping the outputs into semantic categories, LLaVA is the only model that shows significant reductions in the number of output changes. This shows that many of its modified outputs actually remain closely related to the originals. For example, it often showed changes including *man*, *figure*, and *human*, which all refer to the same object category. This is as opposed to BLIP-2 and Qwen, which, except for one abstract tangram for Qwen, show no reduction in the number of output changes, indicating that they generally change not only their wording, but also the actual object that they perceive.

The distributions of the models are the same for both the human and abstract tangrams, even though the numbers are slightly higher for the abstract set. This suggests that the addition of colour has a larger influence on tangrams that already lack a clearer shape interpretation.

By inspecting the outputs, we see slight differences in the manner that the models change their descriptions. BLIP-2 generally retains its original interpretation after colouring, but when it changes its description, it does not stay close to the original. LLaVA often replaces a description with one that still belongs to the same semantic category. And Qwen most frequently changes to a different object altogether. Overall, the results suggest that changing the colours of the tangram pieces has only a limited influence on VLM perception.

5.4 Experiment 4 - Visible Pieces versus Silhouettes

5.4.1 Setup

In this experiment, we investigate whether the presence of visible internal piece boundaries influences the interpretations generated by the VLMs. We compare the model outputs for tangrams with visible pieces to the outputs for the silhouette tangrams. This comparison is performed for both the original and rotated tangrams in the *human*-like and *abstract* subsets. Although these sets were also used in some of the previous experiments in both their silhouette and piece variants, those experiments focused on other influences on outputs, and did not explicitly isolate the effect of removing internal shape boundaries.

We use the sets of the original tangrams and their rotated versions for both the *human* and *abstract* sets in their separated and full silhouette forms. We specifically look at whether there is a consistent change of perception in a model after removing the shape boundaries.

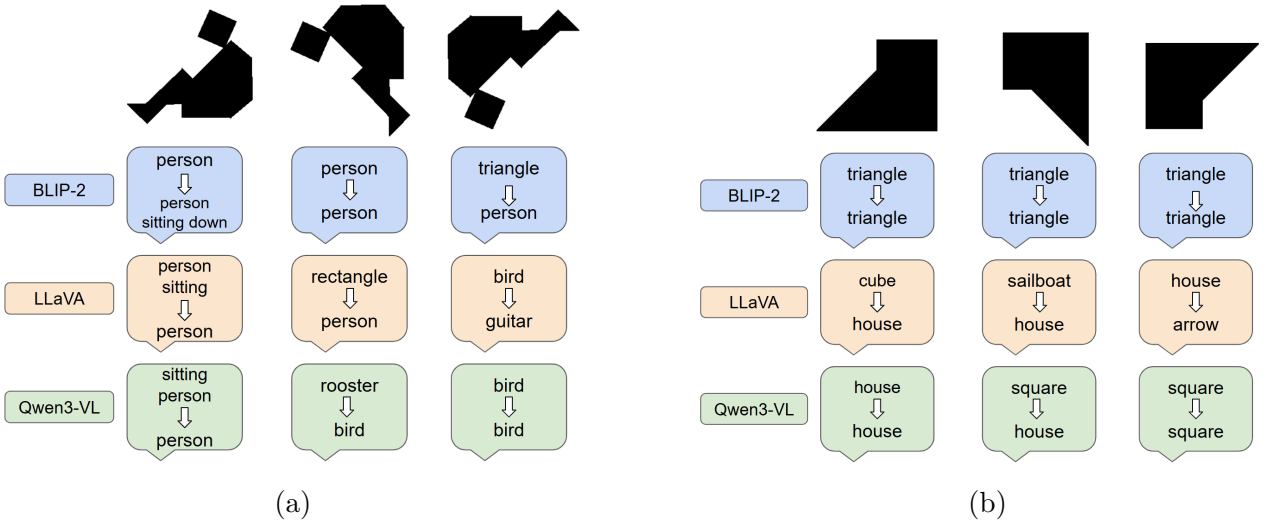
5.4.2 Results

Figure 8 presents two examples comparing the outputs for tangrams with visible pieces and their corresponding silhouette versions.

In general, the removal of internal piece boundaries has little influence on the interpretations of the original tangrams. Looking back at the results of Experiment 1, the level of agreement between the model outputs and the human annotations does not seem to be greatly affected by turning the tangrams into silhouettes. For both sets, the model outputs for the full silhouettes remain relatively similar to those for the original tangrams. This indicates that removing the piece boundaries does not significantly improve, or change, the models' ability to recognize the tangram shape.

We observe a clearer effect when considering the rotated tangrams. Just as in Experiment 2, LLaVA becomes more robust to rotations when the piece boundaries are removed, especially for the human tangrams. This is unlike BLIP-2 and Qwen, which show minor changes in their robustness for the silhouette variants. This suggests that the internal boundaries have a greater influence on LLaVA's interpretation than on the other two models.

When comparing the output changes, we see that LLaVA frequently replaces diverse descriptions with more consistent labels, such as *person* or *house*. Qwen's changes in output generally stay close to its original interpretations, or change to a closely related description, for example, *rooster* to *bird*, or



(a) (b)

Figure 8: Example of model outputs for silhouette variants. Each bubble has an arrow pointing from the original output (with visible pieces) to the output for the silhouette version. (a) shows a human example, while (b) shows an abstract example. We see that for the silhouette versions, LLaVA becomes generally more consistent, while the other models show less changes.

S to *5*. However, the changes in the outputs of BLIP-2 are generally entirely different interpretations.

Overall, the silhouette variants show us that the removal of internal piece boundaries affect the models differently. It has little effect on BLIP-2 in terms of overall changes, but significantly reduces the variation in LLaVA's outputs. This indicates that LLaVA relies more strongly on the internal structure caused by the separated tangram pieces, whereas Qwen appears to generate its descriptions with a focus on the global outline of the shape.

5.5 Additional Experiment - Pose Descriptions

5.5.1 Setup

The previous experiments showed that the models, especially BLIP-2 and Qwen, frequently described the human-like tangrams with a simple *person* label, or something similar. However, this does not reveal whether the models actually recognize the pose of the figure, or solely recognize that the overall shape resembles a human figure. Therefore, we conduct this additional experiment in which we prompt the VLMs on a slightly different question. Instead of only mentioning that it is a tangram and asking about its shape, we tell the models that it is a tangram that represents a human figure and ask them what kind of pose they see in it: *"It represents a human figure, describe its full pose."*. We tested on the *human* set, on the original, rotated, colour and silhouette versions. This experiment investigates whether providing additional context encourages the models to give more detailed interpretations.

5.5.2 Results

Figure 9 shows an example of the outputs by the three VLMs for the modified prompt about pose descriptions. The results demonstrate clear differences in the way that the models respond to the additional context.

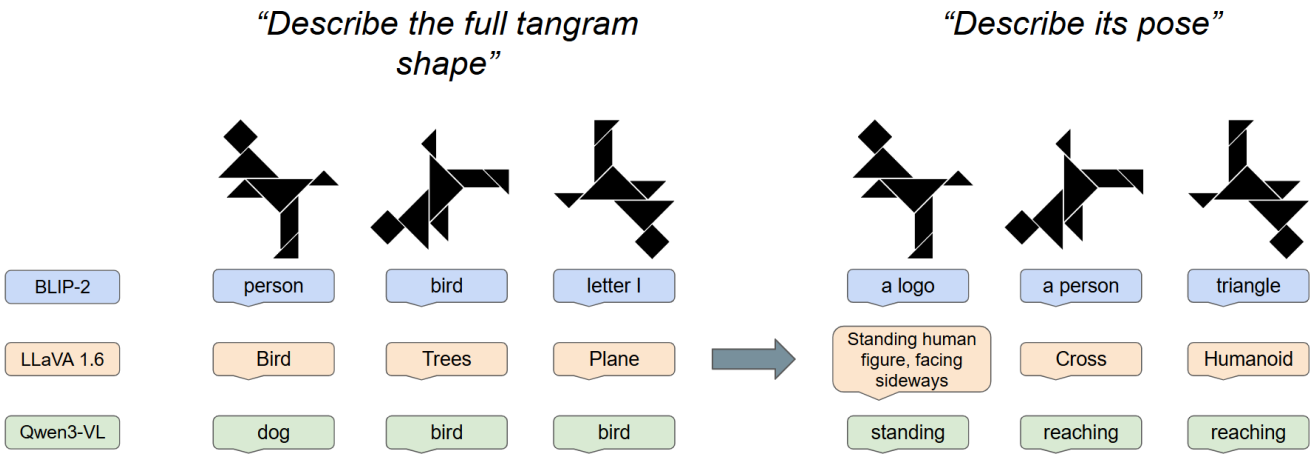


Figure 9: Example of VLM outputs for the pose description prompt, compared to the outputs for the original prompt, for all rotations of a tangram. Here, BLIP-2 does not generate any pose description, LLaVA gives only gives one detailed description, while Qwen manages to give a pose description for all of the rotations.

BLIP-2 is generally unable to provide pose descriptions. Only once does it give a very clear pose description: *sitting down*, and twice it answers with *dancer*. However, the rest of its answers ignore the requested pose description altogether and revert to a general *human figure*, or completely unrelated outputs such as *a logo*, *the image*, and *triangle*. This indicates that BLIP-2 is largely unaffected by the request for more descriptive responses.

LLaVA demonstrates more sensitivity to the modified prompt. It occassionaly generates detailed pose descriptions, such as *walking, man walking with stick, man with arms crossed,* and *arms spread wide*. Nonetheless, many of its responses remain generic (*human, man*) or indicate that it could not determine a pose (*none, not possible*). Thus, although LLaVA is able to give more descriptive outputs than in the previous experiments, it is still unable to consistently produce clear pose descriptions.

Qwen responds most consistently to the modified prompt. For every tangram and its variations it manages to generate a pose-related description. Noticeable is that Qwen also shows a difference in output between the regular and silhouette versions. For the tangrams with visible pieces, many of its responses are simply *standing*, with occasionally *running* or *reaching*. For the silhouette variants, this variety increases a lot, with outputs such as *jumping, walking, arms raised, dancing,* and *leaning* also getting mentioned. This behaviour suggests that Qwen is sensitive to prompt changes and that removal of the internal piece boundaries helps it to focus on finding a pose in the tangram shape.

In general, the provision of additional information and requests in the prompt does not substantially improve the descriptive capabilities of BLIP-2 or LLaVA. Only Qwen is consistent in identifying poses in the tangram shapes, although the descriptions often remain relatively generic. These results suggest that additional information in the prompt alone is not enough to obtain more detailed interpretations from these VLMs, although there are visible differences in the way that they incorporate the added information.

5.6 Summary of Experimental Results

Here is a brief overview of the conclusions per experiment.

- **Experiment 1 (Human Similarity)** shows that all three VLMs produce descriptions that are much more similar to the human annotations when the tangrams resemble human figures, rather than when they are more abstract shapes. BLIP-2 and Qwen demonstrate the highest agreement with the human categories, while LLaVA contains more variety in its outputs.
- **Experiment 2 (Rotation Robustness)** shows that none of the models is invariant to rotations, often changing not only their wording, but also the actual interpreted object. BLIP-2 is most robust to rotation, followed by Qwen, while LLaVA is the least consistent.
- **Experiment 3 (Colour Influence)** shows that colouring the individual tangram pieces affects the models differently. BLIP-2 and Qwen are least affected, whereas LLaVA shows a larger amount of output changes. However, a lot of these changes stay related to the same object category.
- **Experiment 4 (Visible Pieces versus Silhouettes)** shows that removing the visible piece boundaries does not improve the similarity to human annotations. Instead, the models show different behaviour: LLaVA becomes more consistent across rotations, Qwen changes relatively little, and BLIP-2 shows large inconsistent changes in its descriptions.
- **Additional Experiment (Pose Descriptions)** shows that providing additional context in the prompts does not consistently lead to more detailed outputs. Qwen shows the most pose-related variation, LLaVA occasionally produces more pose-related outputs, while BLIP-2 rarely adapts its outputs to the new prompt.

6 Discussion

This chapter discusses the main findings of the experiments in Section 5 in relation to the research question (*In what ways do Vision-Language Models perceive ambiguous tangram shapes similarly to humans?*) and subquestions:

SQ1. Do VLMs perceive tangram shapes in a similar way to humans?

SQ2. Do rotations affect the consistency of VLM interpretations of tangram shapes?

SQ3. Does color influence the perception of a tangram shape by VLMs?

SQ4. Does the visibility of the individual tangram pieces change the VLMs perception?

6.1 Human-like tangrams lead to greater agreement with human perception

The strongest finding of this thesis is the clear distinction between the human and abstract tangram subsets. All three models show substantially greater agreement with human annotations when the tangrams strongly resemble human figures than when they are more abstract (as decided by the human annotations). This indicates that the VLMs are able to recognise familiar shape structures even in ambiguous tangrams, but struggle to agree with human annotations when the structures become more abstract.
A possible reason for this could be that *humans* are popular shapes to form with tangrams and that humans are common in VLM training, making it easier to recognise a human shape than a general abstract object. The fact that people are common in tangram shapes could cause the VLMs to associate a tangram quickly with a human figure. Although, the abstract tangrams receive many different interpretations, which is similar to human behaviour on those tangrams, although they almost never agree on the same interpretation.
We did not expect there to be such a clear difference in behaviour between the two sets. Previous work generally did not make a comparison between two sets such as these, and focused more on evaluating a full set of ambiguous shapes. So, we did not see such a distinction in interpretations between two types of ambiguous shapes before.
In response to SQ1, we see that the VLMs perceive the tangrams in a similar way to humans when there is a strong familiar structure present, but that this human-like interpretation decreases when the level of abstraction increases. The level of detail, however, is generally much lower for the VLMs, when compared to humans.

6.2 The three VLMs exhibit different interpretation strategies

We noticed several consistent patterns in the behaviours of BLIP-2, LLaVA and Qwen while analysing the variation in outputs. BLIP-2 tends to go for geometric interpretations or individual letters, when not sticking to a general theme such as *person*. It is generally robust in its answers, since it often mentions the exact same descriptions. The variation in the outputs of Qwen often remain closely related in category. Qwen produces many animal-related descriptions and is responsive to prompt changes. LLaVA's outputs are more descriptive than those of BLIP-2 and Qwen, cover a

wide range of objects, are sensitive to changes, and almost never stick to one specific interpretation. These patterns suggest that the models rely on different strategies when processing visual shape ambiguity, which could also be indications of biases in their models.
These differences in interpretation strategies between the models also provide additional context for SQ1. They show that the degree of agreement with the human annotations depends not only on the level of abstraction of the tangram, but also on the interpretation tendencies of the VLM itself. We expected beforehand that LLaVA would be more diverse in its answers and that BLIP-2 would be more consistent with strong preferences, since previous work has also shown this (see [17]). Since Qwen is the latest model that we evaluated, we also expected it to be stable and descriptive, although we expected it to be slightly more descriptive. Instead, Qwen gave many general outputs similar to BLIP-2, while we had expected it to behave more similarly to LLaVA and provide more detail.

6.3 Shape transformations influence VLM perception

Factors such as rotation, colour and silhouettes visibly influence the VLMs perception.

6.3.1 Rotation changes model interpretations

We see that rotation of the tangrams cause much variation in the model outputs. The perceptions of the VLMs are therefore visibly not invariant to geometric transformations, answering SQ2. This is especially visible in LLaVA, which shows large shifts in interpretation across rotations. BLIP-2 and Qwen perform more stable, suggesting that they have a stronger reliance on the global shapes than changes in orientations. However, both of these models were not fully invariant and still showed sensitivity to the tangram orientations.
This invariance was generally expected from previous work on ambiguous shapes, although we did not expect Qwen to be more similar to BLIP-2 than LLaVA.

6.3.2 Colour has only a limited influence

Colour changes have a relatively small effect on interpretation compared to geometric transformations, directly answering SQ3. This suggests that appearance features such as colour are not the main priority when the VLM evaluates a tangram shape. Instead, it appears that the VLMs base their answers more on the global shape of the tangrams.

6.3.3 Removing piece boundaries does not have a universal effect

Removing internal tangram piece boundaries affects the models differently, rather than causing a uniform change in perception, answering SQ4.
LLaVA becomes more stable when the tangram is presented as a silhouette, suggesting that the visible internal pieces caused variability in its interpretations. Qwen remains largely stable for silhouettes, implying that it already relied on the full shape rather than evaluating the tangram pieces. BLIP-2 shows inconsistent changes, which does not indicate that it relied on pieces more than global shape, or the other way around.
We had not expected LLaVA to become more stable for the silhouette tangrams. Since the overall

shape remains ambiguous and abstract, we expected all models to behave more like BLIP-2; changing their output but not in a clear or consistent way.

6.4 Providing additional context has limited influence on pose understanding

The pose experiment shows that adding additional context to the prompt does not consistently improve tangram shape understanding in the VLMs.
Qwen demonstrates the strongest responsiveness to the prompt change and produced more varied pose descriptions, especially for the silhouette variants. However, BLIP-2 and LLaVA largely tend to answer with generic pose labels, such as *standing*, with little variation. This indicates that they are either insensitive to the prompt change, or that they are not able to further recognise the poses of the human tangrams.
Although this observation does not directly address one of our subquestions, it provides more insight into the main research question. We expect that explicitly mentioning that the tangram resembles a human figure and asking about its pose would encourage humans to provide a more detailed pose description (which we know would be possible, since the human annotations in the KiloGram dataset were often more detailed). So, if VLMs truly perceive the tangrams in a human-like way, they would show the same behaviour as humans. Since this only occurred for Qwen, while BLIP-2 and LLaVA were largely unaffected, the results suggest that changing the prompt alone is insufficient to compensate for limitations in the VLMs perception of ambiguous tangram shapes.

6.5 Limitations

There are several limitations that influence our results. First of all, our dataset is very small: only twenty tangrams were selected, while the original KiloGram dataset contained more than a thousand tangrams. We manually put the tangrams together to represent a 'human' and an 'abstract' set, but they were randomly selected based on the human annotations per tangram. Thus, the tangrams were not evaluated or chosen based on specific shape features, which could have allowed for other types of observations and experiments. Furthermore, we only evaluate outputs of three VLMs, which we left at their default temperatures and did not train. We used one type of prompt, and evaluated one reply of the VLM, which is not always the same if the VLM is prompted several times. Thus, this single output may not have been representive of the VLMs perception.
Finally, the way of evaluating the performances is a limitation, since a lot of it was based on qualitative measures and subjective generalisation.

7 Conclusions and Future Work

This thesis investigates how vision-language models perceive ambiguous tangram shapes under variation in rotation, colour, and structural visibility, and if this is similar to human perception.

The experiments show that VLMs demonstrate only partial agreement with human perception. The biggest finding is that when tangrams resemble human figures, all models generally produce interpretations that correspond to the human categories. However, as the tangram becomes more abstract, this agreement significantly decreases and the interpretations become very diverse and model-specific. This suggests that the VLMs rely strongly on familiar shape structures, but have no shared human-like way of interpreting more ambiguous tangram shapes.

The experiments further demonstrate that VLM perception is not invariant to transformations. Rotations have a large influence on the outputs, often completely changing the interpretation of a tangram. Colours have a smaller effect, indicating that the global shape plays a bigger role in the VLMs perception. Removal of the visible piece boundaries does not consistently improve the agreement with human perception, but instead affects each model differently, implying that the models do not all rely on global outlines and internal tangram structure in the same way.

Although all three models demonstrate these general trends, they also have some specific individual differences in behaviour. BLIP-2 tends to go for geometric descriptions, LLaVA has very diverse outputs, and Qwen is relatively robust and often stays close in category. These differences show that perceiving ambiguity in the shapes is very model-specific, rather than a common visual reasoning strategy between the models.

Overall, this thesis shows that Vision-Language Models perceive ambiguous tangram shapes only partially in a human-like way. Their descriptions are highly influenced by shape familiarity and visual transformations, while they also show different interpretation strategies. These findings provide insight into their shape perception of not so standard image recognition tasks.

Future work could improve this study by using a larger and more diverse set of tangrams, as well as by evaluating additional and more recent vision-language models. Furthermore, collecting human annotations on the various tangram variations could improve the similarity measures. Finally, future work could prompt the VLMs several times for outputs to evaluate them, in order to create a better view of their actual perception.

A Appendix

A.1 Dataset

A.1.1 Human subset

Dataset: humans							
	With visible pieces				Only the silhouette		
Image	Original	Rotated 90°	Rotated 180°	Color	Full shape	Rotated 90°	Rotated 180°
page-A.png							
page-F.png							
page-I.png							
page-K.png							
page2-71.png							
page4-0.png							
page4-52.png							
page4-82.png							
page6-188.png							
page7-122.png							

Figure 10: Overview of all tangrams in the *human* subset.

A.1.2 Abstract subset

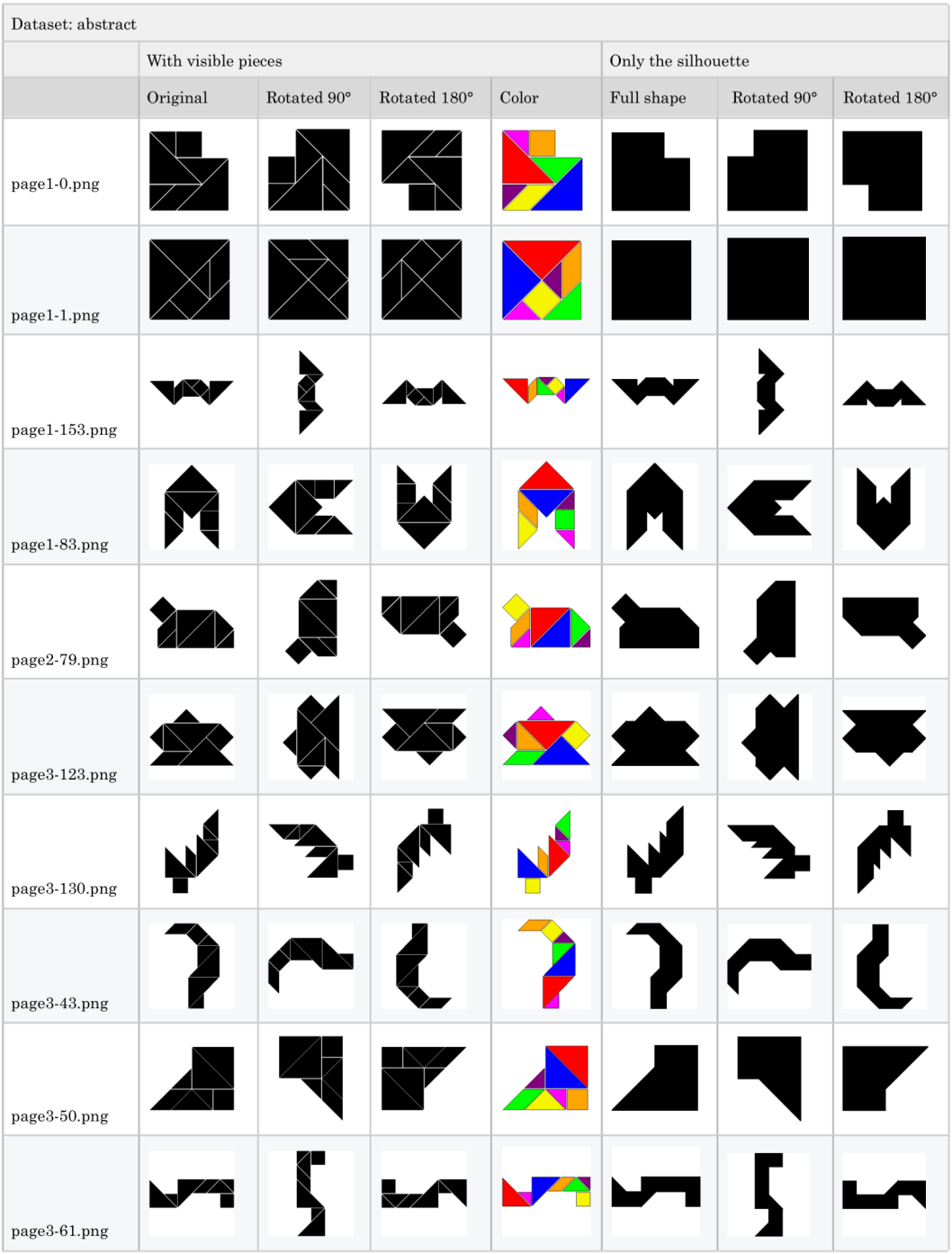


Figure 11: Overview of all tangrams in the *abstract* subset.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025.

[2] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

[3] Xu Cao, Yifan Shen, Bolin Lai, Wenqian Ye, Yunsheng Ma, Joerg Heintz, Jintai Chen, Meihuan Huang, Jianguo Cao, Aidong Zhang, et al. What is the visual cognition gap between humans and multimodal llms? *arXiv preprint arXiv:2406.10424*, 2024.

[4] Herbert H Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986.

[5] Sufyan Danish, Abolghasem Sadeghi-Niaraki, Samee Ullah Khan, L. Minh Dang, Lilia Tightiz, and Hyeonjoon Moon. A comprehensive survey of vision–language models: Pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets. *Information Fusion*, 126:103623, 2026.

[6] Sagi Eppel, Mor Bismut, and Alona Strugatski. Shape and texture recognition in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1839–1849, 2026.

[7] Robert D Hawkins, Michael C Frank, and Noah D Goodman. Characterizing the dynamics of learning in repeated reference games. *Cognitive science*, 44(6):e12845, 2020.

[8] Arshia Hemmat, Adam Davies, Tom A Lamb, Jianhao Yuan, Philip Torr, Ashkan Khakzar, and Francesco Pinto. Hidden in plain sight: evaluating abstract shape recognition in vision-language models. *Advances in Neural Information Processing Systems*, 37:88527–88556, 2024.

[9] Zhishan Hu, Keng-Fong Lam, and Zhen Yuan. Effective connectivity of the fronto-parietal network during the tangram task in a natural environment. *Neuroscience*, 422:202–211, 2019.

[10] Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert D. Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes, 2022.

[11] Nicholas Judd and Torkel Klingberg. Training spatial cognition enhances mathematical learning in a randomized study of 17,000 children. *Nature Human Behaviour*, 5(11):1548–1554, 2021.

[12] Simeon Junker and Sina Zarrieß. Scenegram: Conceptualizing and describing tangrams in scene context. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23976–23992, 2025.

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[14] Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. *arXiv preprint arXiv:2506.04633*, 2025.

[15] Daixian Liu, Jiayi Kuang, Yinghui Li, Yangning Li, Di Yin, Haoyu Cao, Xing Sun, Ying Shen, Hai-Tao Zheng, Liang Lin, et al. Tangrampuzzle: Evaluating multimodal large language models with compositional spatial reasoning. *arXiv preprint arXiv:2601.16520*, 2026.

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[17] Artemis Panagopoulou, Coby Melkin, and Chris Callison-Burch. Evaluating vision-language models on bistable images. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 8–29, 2024.

[18] William Rudman, Michal Golovanevsky, Amir Bar, Vedant Palit, Yann LeCun, Carsten Eickhoff, and Ritambhara Singh. Forgotten polygons: Multimodal large language models are shape-blind. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11983–11998, 2025.

[19] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020.

[20] An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.

[21] Jianyi Yang, Junyi Ye, Ankan Dash, and Guiling Wang. Illusions in humans and ai: How visual perception aligns and diverges. *arXiv preprint arXiv:2508.12422*, 2025.

[22] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024.