Universiteit Leiden

Master Computer Science

Exploring the Bouba-Kiki Effect: Cross-Modal Associations in Vision-and-Language Models

Name:	Kiana Shahrasbi
Student ID:	s3662055
Date:	10/12/2024
Specialisation:	Artificial Intelligence
1st supervisor:	Dr. Tessa Verhoef
2nd supervisor:	Tom Kouwenhoven

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Contents

Acknowledgements

I want to begin by expressing my heartfelt gratitude to my first supervisor, Dr. Tessa Verhoef. Her invaluable guidance, encouragement, kindness, and thoughtful feedback have been a constant source of support throughout this journey. Her calm and reassuring approach has always helped me stay focused, even during the most challenging times. I have learned so much from her expertise and insights, which have greatly shaped my understanding and growth.

I am also deeply grateful to my second supervisor, Tom Kouwenhoven, for his invaluable advice and support. His expertise, thoughtful input, and readiness to provide guidance whenever needed have been incredibly important throughout this process. I truly appreciate the time and effort he dedicated, which have contributed significantly to this thesis.

My family and friends, thank you for your patience, understanding, and encouragement. Your belief in me has been my greatest motivation, and I could not have done this without your support.

Abstract

Cross-modal associations are defined as the ability to connect information across different sensory modalities. For example, linking linguistic features (sound or words) to corresponding visual attributes (images). This thesis investigates whether Vision-and-Language Models (VLMs), specifically Vision Transformer-based and ResNet-based CLIP models, show robust cross-modal associations between linguistic inputs (words) and visual shapes (curved and jagged images). These associations are important as they reveal how closely AI models reflect human perception and can enhance human-machine interactions through shared preferences. The CLIP model was chosen for this thesis because of its strong performance in aligning visual and textual representations. Additionally, prior research has shown that CLIP exhibits patterns of human-like associations in specific contexts, such as connecting linguistic features to visual attributes. This study explores the presence of cross-modal associations in VLMs through four experiments: probability comparisons, image-to-text matching, phonetic component analysis, and attention pattern evaluation. Overall, the findings reveal that cross-modal associations are not consistently present in VLMs and depend heavily on the word type, the model architecture, and the specific task. While familiar word types like English synonyms produce strong associations, more abstract or complex pseudowords reveal significant limitations. Future work should explore several key directions, such as examining other cross-modal associations to provide a broader perspective on how different sensory modalities interact in VLMs. Moreover, investigating models with diverse architectures shows how architectural differences influence cross-modal associations in VLMs.

Keywords: Cross-Modal Associations, Bouba-Kiki Effect, Vision-and-Language Models, CLIP

1 Introduction

Understanding the connection between different types of information (such as sound and vision) and human perception is a fascinating area of study in cognitive science. Links and relationships between different sensory experiences, such as linking sound to another modality, such as vision, are defined as cross-modal associations. One of the most famous examples in this context is the bouba-kiki effect, where most people tend to associate the word "kiki" with jagged and sharp shapes and the word "bouba" with soft and round shapes. In 1929, köhler revealed that people consistently associate the pseudoword "takete" with jagged, angular shapes and "maluma" with soft, rounded shapes [Köhler, 1929, 1947]. This effect, which has since then been widely studied, shows that humans share strong preferences for mapping auditory information to visual shapes.

Cross-modal associations have been investigated in both human studies and Artificial Intelligence. Human studies indicate that certain sounds can be consistently linked to specific shapes [Ramachandran and Hubbard, 2001, Lockwood and Dingemanse, 2015, Cuskley and Kirby, 2013, Nielsen and Rendall, 2012]. Besides human studies, the advancement of AI has provided researchers with new opportunities to investigate these associations in artificial systems. Recent work by Alper and Averbuch-Elor [2023] on VLMs explored whether AI models show the bouba-kiki effect. Their findings show strong evidence of cross-modal associations in VLMs. Another study by Verhoef et al. [2024] further examined the bouba-kiki effect in VLMs. Unlike Alper and Averbuch-Elor [2023], who reported strong evidence of cross-modal associations in VLMs, Verhoef et al. [2024] found only limited evidence, with outcomes heavily dependent on model architecture and task design.

Understanding whether cross-modal associations exist in VLMs like CLIP is important for assessing how closely these AI models reflect human perception. Such associations are fundamental to how humans interpret and engage with the world. These associations support key aspects of human cognition, such as learning language by linking abstract symbols to sensory experiences and communicating effectively through multiple types of information. Examining whether VLMs replicate these associations allows us to evaluate how well these models align with human-like representations. Such alignment enhances interactions between machines and humans by enabling the development of shared preferences and creating a common ground for communication Kouwenhoven et al. [2022]. Furthermore, understanding cross-modal associations in VLMs increases the development of AI systems that perform effectively and replicate how humans process and convey meaning.

In order to understand whether human-like cross-modal associations exist in VLMs, this thesis builds upon the work of Verhoef et al. [2024] and extends their work in several key ways. First, we incorporate a wider range of words to explore the robustness of cross-modal associations across different types of words. Then, we employ a comprehensive analysis of probability experiments to examine how VLMs like CLIP align with human cross-modal preferences. Second, we extend the analysis by utilizing different CLIP [Radford et al., 2021] architectures, such as Vision Transformer and ResNet-based version of CLIP, which allow us to assess the generalizability of the findings across different model structures. Lastly, we analyze the attention patterns of these models on images containing both curved and jagged regions when paired with a label to examine whether their focus aligned with the relevant region, similar to human perception. As shown by Taubert et al. [2011], human perception works holistically by integrating visual information to focus on relevant features within an image.

Our findings confirm the work of Verhoef et al. [2024], demonstrating that cross-modal associations in VLMs highly depend on the specific stimuli, the model architecture, and the task design. This variability emphasizes that cross-modal associations do not consistently emerge across different contexts, inputs, or model designs.

The remainder of this thesis is structured as follows: Section 2 reviews the related literature on cross-modal associations in both humans and VLMs. Section 3 discusses the methodology and explains the process of image generation, pseudoword selection, probability-related experiments, attention pattern analysis, and other analytical techniques. Section 4 presents the results and their findings. Section 5 provides a discussion of the findings. Section 6 concludes with a summary of the contributions. Finally, Section 7 offers suggestions for future research directions.

2 Background

Cross-modal associations are defined as the non-arbitrary relationships between different sensory experiences, such as linking sounds with visual shapes. These associations have been investigated in both human studies and artificial intelligence. The following sections will discuss the relevant literature from both domains.

2.1 Cross-Modal Associations in Humans

Cross-modal associations play an important role in human language processing and language evolution. These associations challenge the traditional belief that language is completely arbitrary and show that there may be patterns or meanings behind how words and sounds are connected [Lockwood and Dingemanse, 2015, Cuskley and Kirby, 2013]. More specifically, Lockwood and Dingemanse [2015] defines sound symbolism as an intuitive connection between a word's sound and its meaning. They argue that sound symbolism is not a random language feature but a key part of a more extensive system where the brain links different senses and experiences. In these cases, the brain processes sound symbolic words more effectively because there is a clear link between how the word sounds and what it represents. Furthermore, Cuskley and Kirby [2013] show that sound symbolism and cross-modal associations play an essential role in language evolution. They propose that iconic form-meaning mappings, like sound symbolism, offer a natural starting point for language development. These connections make communication easier because of common cognitive tendencies. Moreover, Nielsen and Rendall [2012] show how sound symbolism and cross-modal associations influence the language learning process because humans tend to learn and retain words more effectively when there is a meaningful, non-arbitrary relationship between the sound of a word and its meaning. They emphasize that these meaningful connections make it easier to learn a language, especially during the earliest stages when humans first start to develop ways to communicate. Over time, these connections have shaped how new languages have formed and evolved. Blasi et al. [2016] have conducted a thorough study and examined sound-meaning correspondences in a large representative subsection of all existing languages. Their findings show that specific sounds mainly correspond to particular meanings, even across unrelated languages. This finding indicates that some words might originate from natural cognitive connections. This shows that sound symbolism is not just a rare occurrence but a key feature embedded in the core vocabulary of languages.

The bouba-kiki effect, investigated by [Köhler, 1929, 1947], is a key part of sound symbolism and has been studied and replicated extensively in research across different cultures, languages, and sensory experiences [Maurer et al., 2006, Westbury, 2005]. Ramachandran and Hubbard [2001] highlight the importance of the bouba-kiki effect by linking it to larger cognitive processes. Their research suggests that these sound symbolic associations show deeper connections in human cognition between sensory inputs and meanings. This finding suggests that these cross-modal mappings are not random but rooted in how we perceive and process information neurologically.

Furthermore, Cuskley et al. [2017] examine how the shape of written letters, called orthography, can evoke cross-modal associations. For instance, some letter shapes might be seen as "rounded" or "jagged." They also show that this effect holds even when the shapes of the words are absent and only the sounds of the letters and words are present, which is aligned with the "bouba-kiki" effect.

Nielsen and Rendall [2013] also further explore the "bouba-kiki" effect. They aim to analyze how consonants and vowels influence cross-modal associations between sounds and shapes. Their experiments use novel pseudowords and newly generated curved and jagged images. They ask participants to create two-syllable words from consonant and vowel options to match curved or jagged shapes. They found that participants tend to associate rounded vowels and sonorant consonants with curved images and non-rounded vowels and plosive consonants with jagged images. Their study shows that both consonants and vowels affect the "bouba-kiki" effect.

Besides the bouba-kiki effect, many other cross-modal associations have been observed. A study by Hubbard [1996] explores how visual lightness relates to auditory pitch. This study shows that people tend to associate lighter images with higher-pitched sounds and darker images with lower-pitched ones. This built on earlier work by Marks [1974], which found that melodic intervals are connected to visual lightness, meaning that higher-pitched sounds usually go along with lighter visuals. In contrast, lower-pitched sounds are linked to darker visuals. Hubbard [1996] also mentions that the background color affects these associations. More specifically, these associations are more substantial with a black background. Parise and Spence [2009] also explore how people inherently associate high-pitched sounds with smaller or brighter objects, while low-pitched sounds are linked to larger or darker ones.

2.2 Cross-Modal Associations in VLMs

In addition to human studies, some research has explored how cross-modal associations function in artificial systems. For example, Kann and Monsalve-Mercado [2021] examine the link between character embeddings in neural networks and a phenomenon called grapheme-color synesthesia, where people see specific colors when they look at letters. Based on data from this grapheme-color synesthesia, the researchers measure letter similarities based on these color associations. Then, they compare these similarities to character embeddings derived from various neural architectures. More specifically, they trained models on different tasks. They found that models focused on tasks involving the relationship between letters and their sounds aligned most closely with human perceptions, as letter sounds often influence synesthetic color associations. Their research shows how understanding cross-modal associations, for example, focusing on sound-letter relationships, can help improve AI systems and bring their representations closer to human cognition.

Despite advancements in AI models and their ability to process multimodal data, these models still face limitations.

Thrush et al. [2022] state that even advanced models struggle with tasks requiring visio-linguistic compositional reasoning. This involves understanding complex relationships between images and text, such as distinguishing how word order changes the meaning of captions. The poor performance of AI models on the *Winoground* dataset highlights these challenges. Unlike humans, these models fail to distinguish subtle changes in word order and match captions to images correctly. The study by Diwan et al. [2022] shows that the failure of VLMs on the Winoground dataset arises from factors such as reasoning about complex relationships, identifying fine-grained visual details, challenges with unusual or visually difficult data, and, most importantly, the difficulty in robustly aligning textual semantics with visual elements. Kamath et al. [2023] also indicate that VLMs struggle with basic spatial reasoning tasks, such as understanding the difference between "on" and "under." These failures arise from spatial relationships being rarely labeled in training datasets and a bias toward common object arrangements. The study by Kamath et al. [2023] also confirms the challenges VLMs face, as seen in the Winoground dataset, which results from their inability to accurately align semantic and spatial relationships.

Furthermore, [Jabri et al., 2016, Goyal et al., 2017, Agrawal et al., 2018] highlight another limitation of VLMs. Their studies show that these models rely heavily on textual patterns instead of precisely analyzing visual content. Therefore, these models struggle to integrate textual and visual data meaningfully. For example, these models tend to answer "yes" to questions without basing their answers on the image. The work by Goyal et al. [2017] shows that even when these models are trained on a balanced dataset, they face challenges connecting textual input with subtle visual details.

These examples show that poor visio-linguistic compositional reasoning, difficulty understanding spatial relationships, and challenges in integrating textual and visual data may cause AI models to struggle to replicate human-like cognitive patterns.

A recent study by Alper and Averbuch-Elor [2023] explores whether VLMs like CLIP and Stable Diffusion show sound-symbolic patterns, similar to the bouba-kiki effect. Specifically, they used Stable Diffusion to generate images of pseudowords carefully designed to reflect phonetic properties linked to sharp or round shapes. These images were then projected into CLIP's shared semantic space, which enables a comparison of the generated images' visual properties with the pseudowords' linguistic properties. Then, by learning a projection in CLIP's semantic space, they determined whether the generated images aligned with the sharp or round associations of the pseudowords. Their findings indicate strong evidence for the existence of cross-modal associations in VLMs. More precisely, they show that VLMs tend to associate sharp-sounding pseudowords with jagged and sharp visual elements and round-sounding pseudowords with round and soft visual elements. This finding is surprising because, as mentioned earlier, VLMs still face limitations. These limitations suggest that even if VLMs can extract sound-symbolic information from the texts they have been trained on, they are likely to have difficulty associating this information with visual features.

Unlike Alper and Averbuch-Elor [2023], the study by Verhoef et al. [2024] tests various VLMs, including CLIP, BLIP2, ViLT, and GPT-4o, using newly generated curved and jagged images paired with novel pseudowords adapted from Nielsen and Rendall [2013] 's work. They used a far more direct approach than the study of Alper and Averbuch-Elor [2023] by analyzing the probabilities generated by VLMs for matching specific pseudowords with curved and jagged images. Their findings show limited evidence for the bouba-kiki effect, with CLIP and GPT-4o showing some alignment with human-like associations. The study concludes that cross-modal associations in VLMs are highly dependent on factors such as model architecture, training data, and the specific text prompts used.

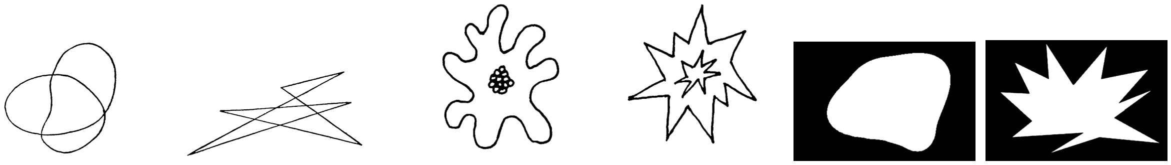


Figure 1: Original pairs of curved and jagged images used in previous studies [Köhler, 1929, 1947, Maurer et al., 2006, Westbury, 2005]

This thesis, building upon the work by Verhoef et al. [2024], delves deeper into these associations by using a broad set of words and different CLIP architecture and also analyzing the attention patterns to assess the alignment between model focus and human perceptual tendencies. This study shows that while VLMs may exhibit certain human-like cross-modal tendencies, their performance highly depends on the evaluation approaches and model-specific characteristics.

3 Methodology

In order to explore the existence of cross-modal associations in VLMs, this thesis focuses on the CLIP model because it demonstrates the best performance in the work of Verhoef et al. [2024] and is further supported by Demircan et al. [2024], who show that CLIP outperforms other models in capturing human-like decision patterns. As mentioned, we use the Vision Transformer (ViT) and ResNet-based versions of the CLIP model. The ViT version used in this study is *clip-vit-base-patch32*, while the ResNet-based version corresponds to the *RN50* model. The methodology begins with creating and defining the newly generated images and word sets used in the experiments. Following this, we explain the methods used to investigate the existence of cross-modal associations in CLIP. This project is primarily implemented using Python and its various libraries, which provide the necessary tools for processing data, generating embeddings, and visualizing results.

3.1 Visual Shapes

In this thesis, a combination of curved and jagged images is used. Some of these images are sourced directly from previous works [Köhler, 1929, 1947, Maurer et al., 2006, Westbury, 2005], while others are specifically generated for this study, inspired by the methods described in Nielsen and Rendall [2013]. More specifically, one set of curved and jagged images is taken from Köhler [1929, 1947], four sets are obtained from Maurer et al. [2006], and four additional sets are sourced from Westbury [2005], which feature white shapes on a black background. Figure 1 shows three pairs of these original curved and jagged images, presented from left to right, corresponding to [Köhler, 1929, 1947], Maurer et al. [2006], and Westbury [2005].

In addition to these original images, we generate new curved and jagged images following the methodology outlined in Nielsen and Rendall [2013]. We first create random points uniformly distributed within a circle with one radius to generate these images. More precisely, the points are arranged in polar coordinates, with random angles sorted in ascending order to ensure a sequential path and random radii sampled between zero and the circle's radius to vary the point distribution. These points are then connected to form closed shapes, with the method differing for curved and jagged images.

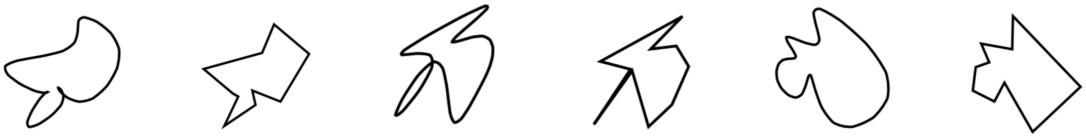


Figure 2: Newly generated curved and jagged images

To create curved images, we connect the points using cubic spline interpolation. This method produces smooth, continuous contours by fitting a periodic spline through the points, ensuring that the ending and starting points are connected. We add extra interpolated points along the curves to make the shapes even smoother. In contrast, straight-line segments connect the points directly in jagged images. The points are linked sequentially to form sharp, angular edges, with the final segment connecting the last point back to the first to complete the closed shape. Figure 2 shows three pairs of newly generated curved and jagged images used in this study.

Eight pairs of newly generated images and original curved and jagged shapes have been selected for the experiments, totaling 17 pairs of images used in the study. Appendix A shows all the images described in this section.

3.2 Linguistic Inputs

This study uses a diverse set of English adjectives and pseudowords that conform to the phonotactic rules of English. English adjectives serve as a baseline for analyzing associations within established language use. Pseudowords, however, are included to investigate how specific linguistic features influence cross-modal associations without the constraints of pre-existing semantic meanings. By doing this, we ensure that the observed associations are based on the structure and syllables of a pseudoword and not on any arbitrary semantic associations that may have been learned during the VLM training. This approach differentiates our work from the study of Alper and Averbuch-Elor [2023] because, in this thesis, we focus only on the relationships between linguistic features and visual representations. However, in Alper and Averbuch-Elor [2023]'s work, they used CLIP's shared semantic space to analyze embeddings, which may include associations influenced by patterns or biases in the training data.

The pseudowords used in this thesis include the common initial words used in cross-modal association research, including *bouba, kiki, maluma, and takete*. Additionally, it contains the pseudowords introduced in Alper and Averbuch-Elor [2023]'s study as well as newly generated one-syllable and two-syllable pseudowords presented in the work of Verhoef et al. [2024]. These pseudowords are constructed following the methodology outlined in Nielsen and Rendall [2013] for generating novel pseudowords. In the following sections, we will explain each word set in detail.

Adjectives We select English synonyms for the adjectives to represent curved and jagged shapes. The synonyms for curved shapes are: *curved, round, circular, soft, wavy, oval, smooth, plush, arc-shaped, and rotund*. For jagged shapes, the synonyms are: *jagged, spiky, sharp, uneven, angular, serrated, edgy, pointed, prickly, and rugged*. These words are chosen to capture a variety of terms commonly associated with curved and jagged features, resulting in 10 curved and 10 jagged synonyms.

Alper and Averbuch-Elor [2023]' Word Set This thesis uses the round and sharp pseudowords generated by Alper and Averbuch-Elor [2023]. These pseudowords are specifically designed to capture phonetic properties linked to roundness and sharpness. The set includes 324 round words, such as *baloba*, and 324 sharp words, such as *kitiki*.

Novel Generated Words In the study by Nielsen and Rendall [2013], participants constructed novel words to match curved or jagged shapes. They were provided with specific consonants and vowels categorized by their phonetic characteristics. Participants selected syllables from these groups to form two-syllable words that they felt best matched the visual properties of an object, such as its curviness or jaggedness.

For consonants, sonorants *(/m/, /n/, /l/)* were associated with curved shapes, while plosives *(/t/, /k/, /p/)* were linked to jagged shapes. For vowels, rounded vowels *(/oo/, /oh/, /ah/)* were associated with curved shapes, and non-rounded vowels *(/ee/, /ay/, /uh/)* with jagged shapes.

Inspired by this method, we use the same set of consonants and vowels to construct novel words for our study. A key reason for using this set is that it has been used in elaborate experiments with humans. This provides the opportunity to directly compare the VLMs' performance to human behavior, which is not possible by using the Alper and Averbuch-Elor [2023]'s word set. For one-syllable words, we create categories such as *s_r* (sonorant consonants with rounded vowels) for curved shapes and *p_nr* (plosive consonants with non-rounded vowels) for jagged shapes. Additionally, we create combinations like *s_nr* (sonorant consonants with non-rounded vowels) and *p_r* (plosive consonants with rounded vowels) to examine mixed associations. This process results in 9 words for each group of one-syllable words, totaling 36 one-syllable words. We extend this approach for two-syllable words to produce fully curved words *s_r_s_r* and fully jagged words *p_nr_p_nr* align with prior findings [Nielsen and Rendall, 2013, Verhoef et al., 2024] on their relevance to the bouba-kiki effect. Other combinations, such as *s_r_p_nr* and *p_nr_s_r*, are also included to examine the roles of consonants and vowels in our experiments, similar to the study by Verhoef et al. [2024]. This approach results in a total of 324 two-syllable words.

In our study, both adjectives and pseudowords are incorporated into the sentence structure: *The label for this image is [insert word].*, where [insert word] represents the selected label. This approach aligns with how VLMs are typically used and trained in practice. Moreover, incorporating these labels into a structured sentence helps the model process the pseudowords more effectively within a context.

3.3 Embeddings

Before analyzing the cross-modal associations in CLIP, we extract CLIP vision and text embeddings and project these into a two-dimensional space using the t-SNE technique [van der Maaten and Hinton, 2008]. By doing this, we can observe patterns and clusters within the data, which helps us to identify potential patterns or clusters that reflect how the model organizes curved and jagged images and different word types in its feature space.

Figure 3 shows the result of the t-SNE technique applied to our curved and jagged images. The left plot represents embeddings from the ViT-based model, while the right plot shows the embeddings for the ResNet-based model. In these images, the blue dots show the jagged images, and the orange dots show the curved images. The smaller dots correspond to images with black backgrounds, as used in Westbury [2005] 's study, while the larger dots represent shapes with white backgrounds. Figure 3 shows distinct clusters for curved and jagged images and images with black and white backgrounds. While there is some overlap between curved and jagged points in both models' embeddings (corresponding to images from [Köhler, 1929, 1947] and some of Maurer et al. [2006] 's datasets), most points form well-defined clusters. In the ViT-based embeddings (left plot), a diagonal line could largely separate curved and jagged images. In contrast, a horizontal line could achieve similar separation in the ResNet-based embeddings (right plot). This observation indicates that even without considering the labels, both models can distinguish between curved and jagged images and identify shape-specific differences in their representations.

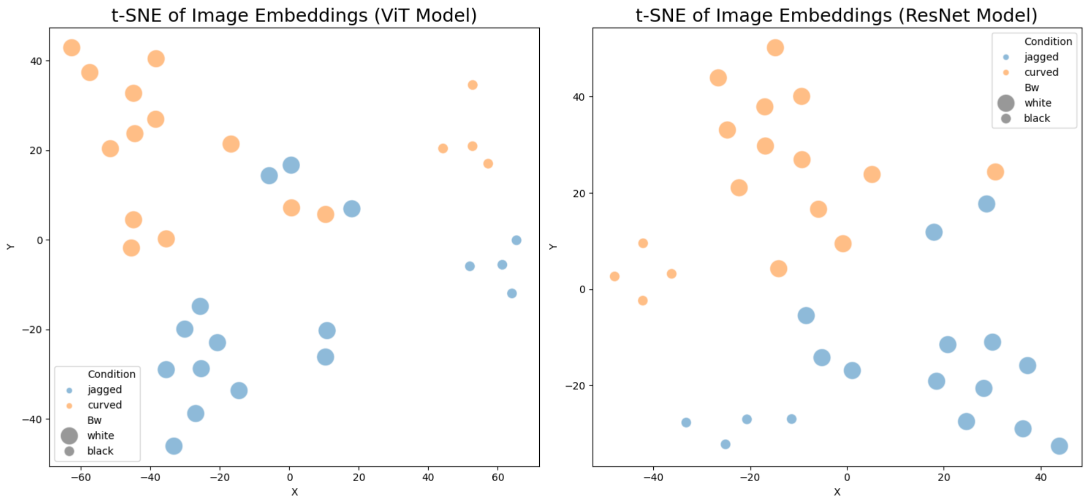


Figure 3: Image embeddings

On the other hand, figure 4 shows the result of the t-SNE technique applied to the different word sets we use in this thesis. In this figure, the left plot shows embeddings from the ViT-based model, and the right plot shows the embeddings for the ResNet-based model. We observe distinct clusters for certain word categories, such as the round and sharp words from the Alper and Averbuch-Elor [2023]' study and English adjectives. However, there is a noticeable overlap between some clusters, particularly those representing one-syllable words (s_r and p_nr) and two-syllable words (s_r_s_r and p_nr_p_nr). The distinct clustering of round and sharp pseudowords from Alper and Averbuch-Elor [2023]'s word set in the word embeddings could potentially help explain why they found strong evidence of cross-modal associations in their results. Their sharp and round pseudowords show more distinct clusters than the other curved and jagged pseudowords. This clearer separation of clusters likely made it easier for the model to capture cross-modal associations. In contrast, the overlapping clusters of curved and jagged words in one-syllable and two-syllable pseudowords may have contributed to the weaker cross-modal associations observed in Verhoef et al. [2024]'s study.

3.4 Probability Analysis

In this analysis, we use the probabilities generated by different CLIP architectures to analyze cross-modal associations in VLMs. In the first experiment, we focus on directly analyzing the probability distributions produced by the models for different word-image pairs. In the second experiment, we use these probabilities to determine the label that best matches each image

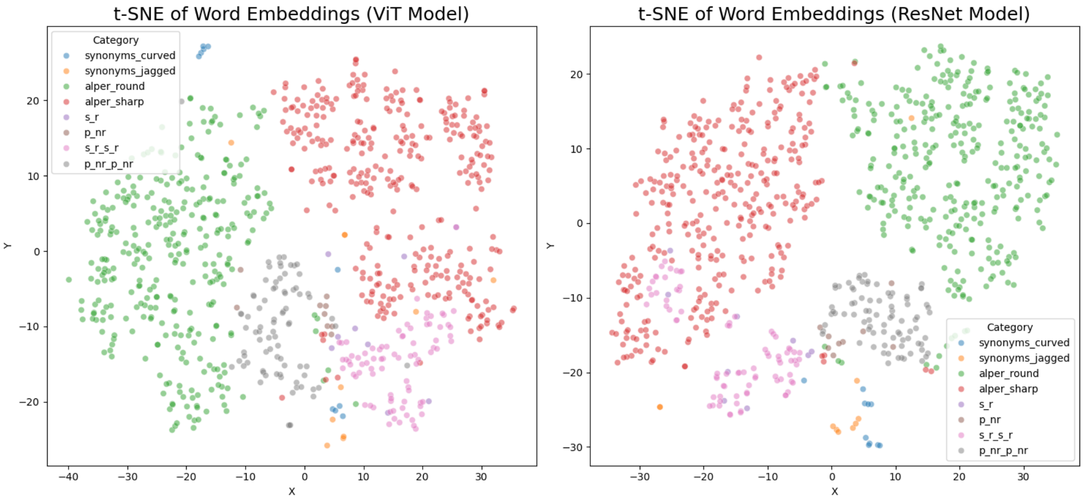


Figure 4: Word embeddings

and generate results that can be compared more directly to human data. The following sections explain the process in detail.

3.4.1 Probability Comparison

As mentioned, for this experiment, we directly use the probabilities generated by the CLIP model to analyze the association between textual inputs and visual shapes. This process is conducted separately for each word set and ensures that all the images within a given word set are evaluated against the corresponding textual prompts. More specifically, we compute logits for each image, representing the raw scores and indicating how well the image aligns with each textual prompt. These logits are then converted into probability distributions using a softmax function. This process is repeated for all the images, and word types. We can analyze the relationships between textual inputs and visual features modeled by the different CLIP architectures by calculating these probability distributions for each word set and its associated images. Then, by comparing these probabilities, we can determine if the model prefers to match curved images with curved labels and jagged images with jagged labels. This approach is similar to what Verhoef et al. [2024] conducted using different VLMs. This comparison allows us to explore the consistency of cross-modal associations in the CLIP model and evaluate its ability to align linguistic and visual features in a human-like way.

3.4.2 Image-To-Text Matching Task

In the second experiment, for each word set, we identify the label with the maximum probability for each image. This label, which has the highest probability, is considered the best match for the corresponding image. This approach is similar to what Nielsen and Rendall [2013] conducted with human participants and what Verhoef et al. [2024] implemented using different VLMs. This experiment, like the probability comparison 3.4.1, helps us to examine whether the model aligns linguistic inputs with visual inputs in a manner consistent with human-like cross-modal associations.

We analyze the results of this experiment in two ways. First, we examine whether the model correctly matches the correct labels to the images based on the complete word structure. For example, for one-syllable and two-syllable pseudowords, we expect that fully curved labels (s_r, s_r_s_r) would be matched to curved images, and fully jagged labels (p_nr, and p_nr_p_nr) would be matched to jagged images correspondingly. As another example, for the Alper and Averbuch-Elor [2023] 's word set, we expect round words to align with curved images and sharp words to align with jagged images. Second, we analyze the effect of consonants and vowels in one-syllable and two-syllable pseudowords. We examine whether labels containing sonorant consonants or rounded vowels tend to align with curved images and whether labels containing plosive consonants or non-rounded vowels align with jagged images.

Conducting these experiments enables us to evaluate the model's ability to capture holistic word structures and its sensitivity to fine-grained phonetic properties. Additionally, it enables us to explore whether the observed patterns differ across different word sets, such as adjectives, pseudowords, and novel syllables.

3.5 Attention Pattern Analysis

In the second part of this study, we want to understand which parts of the image the CLIP model focuses on when seeing a label. We use the Grad-CAM Selvaraju et al. [2020] technique to do this. Grad-CAM (Gradient-weighted Class Activation Mapping) generates visual explanations by highlighting regions in the input image that the model considers important for its predictions. More precisely, Grad-CAM computes the gradient of a target class score (for example, the probability of a cat in an image classification task) with respect to the feature maps in the last convolutional layer. These gradients show how much each neuron in that layer contributes to the target. After that, these gradients are averaged to determine the importance of the weights for each feature map. Then, by combining these feature maps with their corresponding weights, Grad-CAM generates a numerical map that shows the importance or contribution of each specific location in the image to the model's prediction. This map is typically visualized as a heatmap, overlaid onto the image, with colors indicating regions with higher or lower importance for the model's prediction of the target class Selvaraju et al. [2020].

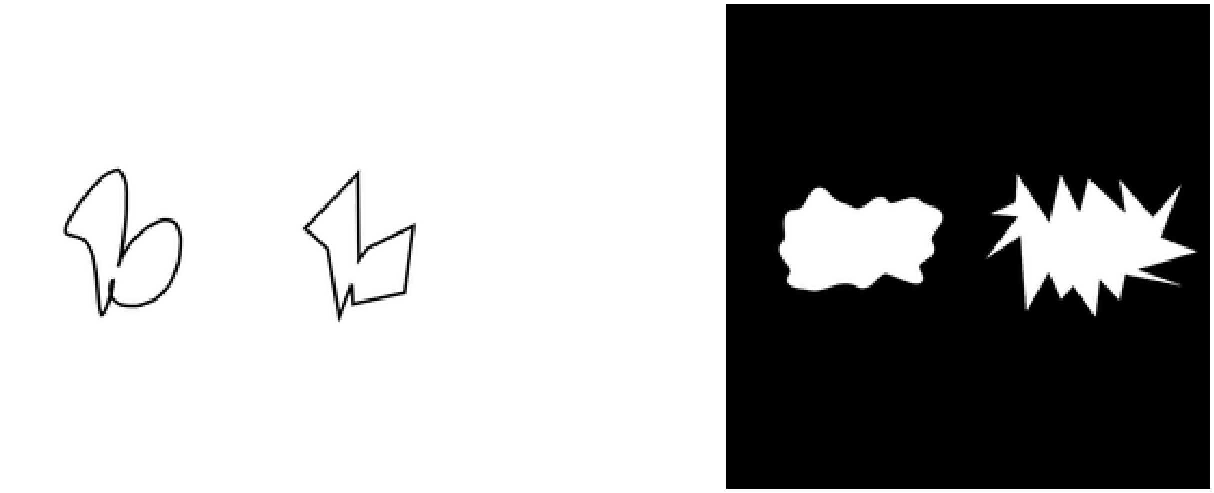
In our study, we apply Grad-CAM to both ViT and ResNet-based versions of CLIP and set the target as the cosine similarity score between text and image embeddings. This choice is made because the similarity score shows how well the model associates a given text label with an image, which is suitable for analyzing cross-modal attention patterns.

We use the last convolutional layer as the target layer to adapt Grad-CAM for ResNet-based CLIP. A forward hook is attached to this layer to capture the activations and gradients during the forward and backward passes. The importance weights are computed by averaging the gradients across the spatial dimensions. Then, these averaged gradients are combined with the activations to generate a relevance map. For ViT-based CLIP, which lacks convolutional layers, we modify the Grad-CAM implementation to work with the last attention block. In order to do this, we define hooks to capture both the attention probabilities during the forward pass and their corresponding gradients during the backward pass. The attention probabilities and gradients are then combined and averaged across attention heads to compute a relevance map. It is important to mention that for both ViT and ResNet, we clamp the relevance maps to remove negative values because we only want to highlight the regions that positively impact the model's predictions.

3.5.1 Preprocessing The Images

For the attention pattern experiments, we combine pairs of curved and jagged images into single images to analyze attention patterns and better understand how the model distinguishes between these shapes when given a label. Each pair is concatenated in two configurations: one with the curved image on the left and the jagged image on the right, and the other with the jagged image on the left and the curved image on the right, to prevent any unwanted effect of a potential positional bias in the model's attention. Positional bias, as highlighted by Wang et al. [2024], is an issue in both language and vision-and-language models. VLMs may correctly identify objects in certain positions but fail when the objects are positioned elsewhere Wang et al. [2024].

After concatenating these square images, we obtain rectangular images, which, if passed directly to the model, would become squeezed during the processing steps. We add a consistent background to address this issue by resizing and padding the images to 224x224 pixels. Specifically, black backgrounds are used for images with black background Westbury [2005], and white backgrounds for images with white backgrounds. This preprocessing step avoids unnecessary attention to the background. Figure 5 shows an example of how the combined images appear after concatenation and background adjustment.



(a) Image with white background

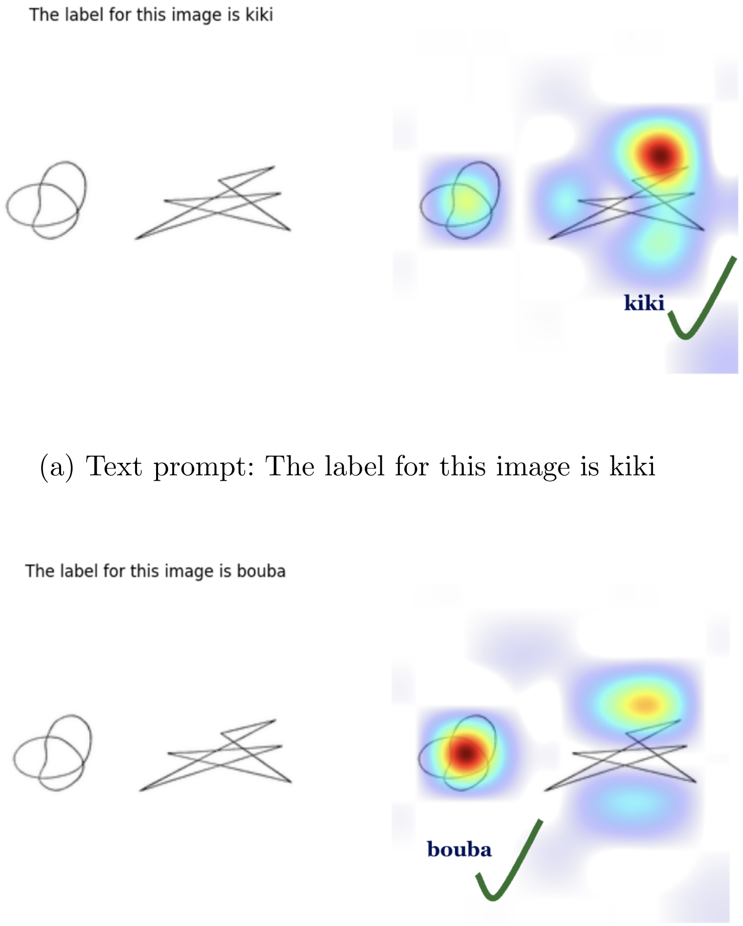
(b) Image with black background

Figure 5: Concatenated images used in attention pattern experiments

3.5.2 Quantifying The Attention Pattern

As mentioned, Grad-CAM allows us to generate heatmaps that highlight the regions of an image where the model focuses when given a specific label. Figure 6 provides two examples of visualizing the attention pattern for an input image with different text prompts. In these examples, the left image shows the input image and its corresponding text prompt, while the right image displays the resulting attention pattern for this input image and text prompt. In the right image, the green checkmark indicates the region expected to receive more attention for that specific label. Additional visualizations of attention patterns can be found in Appendix B.

While these heatmaps visually indicate which parts of the image are most important to the model, we need a way to quantify this attention in a measurable form. As mentioned, the output



(a) Text prompt: The label for this image is kiki

(b) Text prompt: The label for this image is bouba

Figure 6: Visualizing the attention patterns for the ResNet-based version of CLIP for different text prompts

of the Grad-CAM is essentially a map containing importance weights for different regions, so metrics are required to analyze and compare the model's attention patterns. We explored several methods for this purpose, including the sum of intensities (which measures the total attention allocated to each part of the image), maximum intensities (which identifies the single most focused point), average intensities (which shows the mean attention value), attention ratio (which compares the distribution of attention across different regions of the image), standard deviation (which indicates the variation in attention values), and peak attention density (which measures the density of attention within a small patch surrounding the most focused area).

We decided to use the sum of intensities as the primary metric for quantifying attention patterns because it captures the total attention distributed across the image, similar to how humans perceive images. Taubert et al. [2011] suggest that humans perceive images holistically by combining visual details to understand the overall context and identify the most relevant features within an image. The sum of intensities provides a precise measure of how much

attention the model allocates to different regions of the image. The sum of intensities also helps us understand the model's attention as a whole rather than focusing on isolated areas.
We divide each combined image into curved and jagged regions to quantify the attention patterns using the sum of intensities. For each text label, we calculate the sum of intensities in both regions. By comparing the sum of intensity values, we evaluate whether the model allocated more attention to the expected region corresponding to the given label. It should be noted that we conducted this experiment using other metrics as well. The results show that, besides the advantages mentioned for the sum of intensities, it performs better and provides more stable results than the other metrics, making it the most reliable choice for analyzing attention patterns. Performance of the other metrics can be seen in Appendix C.

3.6 Analyzing the Results

In this study, we assess the statistical significance of our findings using Bayesian models with the brms package in R. This method enables us to evaluate the reliability of the results for the observed patterns.
We model the generated probabilities using a Gaussian distribution for the experiments in Section 3.4.1. Based on the definition of our data, we use the formula shown in Equation 1. In this formula, *Condition* (curved and jagged images) and *Category* (different word sets) are the fixed effects. In contrast, *Label* and *Image* are included as random effects to account for variability across different labels and images. The model is run with 4 chains, 4000 iterations, and 2000 warmup steps.

$$\text{Probability} \sim \text{Condition} * \text{Category} + (1 + \text{Condition} \mid \text{Label}) + (1 \mid \text{Image}) \tag{1}$$

We use a binomial logistic regression model for the experiments in Section 3.4.2, where the outcome is binary (correct match or not). In other words, we are quantifying the probability of correctly matching curved and jagged images to their corresponding labels. The formula can be seen in Equation 2. In this formula, *Condition* (curved or jagged) serves as the predictor, and the response variable *Occurrence* represents the number of correct matches within a given sample size. The model is configured with 4 chains, 1000 iterations, and 500 warmup steps.

$$\text{Occurrence} \mid \text{trials(Sample\_Size)} \sim \text{Condition} \tag{2}$$

Finally, for the experiments in Section 3.5, we again work with binary outcomes that indicate whether the higher sum of intensities is observed in the expected region. In other words, in these experiments, we need to model the proportion of times attention is higher in the expected region for each word type. Given the structure of this data, we employ a binomial logistic regression model with the formula defined in Equation 3. Here, *Label* serves as the predictor, while the response variable *Occurrence* represents the occurrence of correct matches relative to the sample size. The model is run with 4 chains, 1000 iterations, and 500 warmup steps.

$$\text{Occurrence} \mid \text{trials(Sample\_Size)} \sim \text{Label} \tag{3}$$

4 Result

In this section, we present the results obtained from each of the conducted experiments for both versions of the CLIP model. First, we compare the probability scores for curved and jagged images across different word sets. Next, we analyze the percentage of cases where the selected label, which has the highest probability, aligns with the image, considering both the whole word structure and its individual components (consonants and vowels). Finally, we evaluate the quantified results of the attention maps and analyze the percentage of cases where, for each label, the sum of intensities is higher in the expected region, indicating that the model allocates more attention to the correct region.

As mentioned in Section 3.6, in addition to visually analyzing the results, we use the Bayesian significance test to analyze the results' significance statistically. Tables 1, 2, 3, and 4 summarize the details of these tests.

It should be noted that, in addition to the standard CLIP models, we also use the specific version of the CLIP model from Alper and Averbuch-Elor [2023]'s study, which is *laion/CLIP-ViT-H-14-laion2B-s32B-b79K*. The results of this experiment are presented in Appendix D.

4.1 Probability Comparison

Figures 7, 8, 9, 10, and 11 show the average probabilities for curved and jagged shapes across different word types. The x-axis represents the word types, while the y-axis shows the average probability scores assigned to curved and jagged shapes. The red bars show the results for the ViT-based version of CLIP, with lighter red for curved shapes and darker red for jagged shapes. Similarly, the blue bars represent the result for the ResNet-based version of the CLIP model, with lighter blue for curved shapes and darker blue for jagged shapes. In these plots, the error bars show the standard error of the mean for the average probability scores and represent the confidence in the mean estimates.

In these plots, we expect higher probabilities for curved shapes when paired with curved-related words and higher probabilities for jagged shapes when paired with jagged-related words. This pattern aligns with the sound-shape associations observed in human cognition.

Initial Words Figure 7 presents the results for the initial words *kiki, bouba, takete, maluma*. As mentioned, we expect curved images to achieve higher probabilities when associated with the curved words *bouba* and *maluma*, and jagged images to achieve higher probabilities when associated with the jagged words *kiki* and *takete*. In the ViT model, for *bouba*, we observe high probabilities for both curved and jagged images. However, no notable pattern differentiates the two, indicating that the model does not strongly associate this word with curved images as expected. Additionally, for *takete*, the average probabilities for jagged images are slightly higher than those for curved images. We do not observe notable expected patterns for the other initial words in the ViT model. The results from the Bayesian model show that *takete* has a small positive effect on its association with jagged images. However, the model indicates no statistically significant interactions between curved and jagged shapes and the initial words in the ViT model.

The ResNet model shows clearer patterns than the ViT model for the initial word set. Specifically, for the words *kiki*, *bouba*, and *takete*, the expected patterns are visually apparent. However, the results of the Bayesian model indicate no statistically significant effects, which suggests that the ResNet model does not consistently align with our expectations.

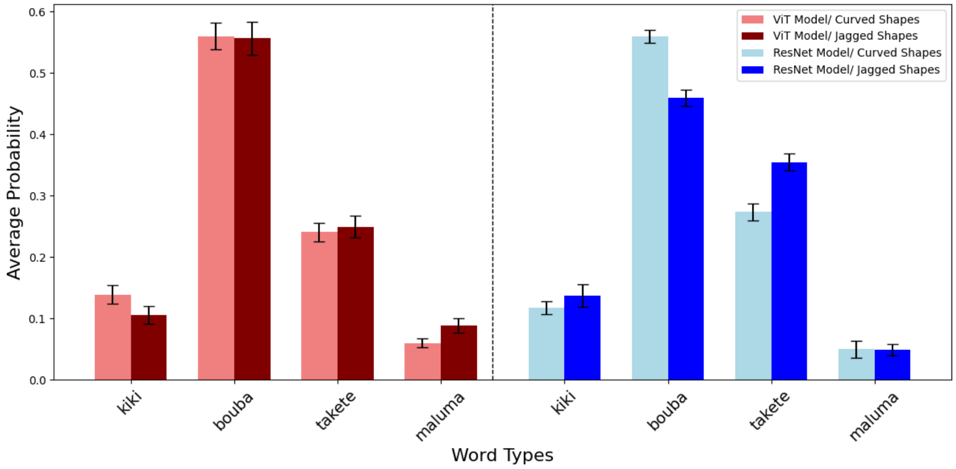


Figure 7: Average probability results for initial words

Adjectives Figure 8 represents the average probabilities for curved and jagged images for the English synonyms. As this word set consists of adjectives, it is our benchmark. The red bars show a clear, curved, and jagged synonym pattern. More specifically, for curved synonyms, the model assigns higher probabilities to curved images; for jagged synonyms, the model assigns higher probabilities to jagged images. This result aligns well with our expectations. The Bayesian significance test supports these observations and indicates that the ViT model significantly assigns higher probabilities to jagged shapes when associated with jagged synonyms. Similarly, the Bayesian significance test shows that the ViT model assigns higher probabilities to curved shapes when associated with curved synonyms.
Like the ViT model, the ResNet model shows a strong and clear pattern. The Bayesian significance test also confirms the observed pattern in this plot. It shows that the ResNet model reliably assigns higher probabilities to jagged shapes when associated with jagged synonyms and significantly assigns higher probabilities to curved shapes when associated with curved synonyms.

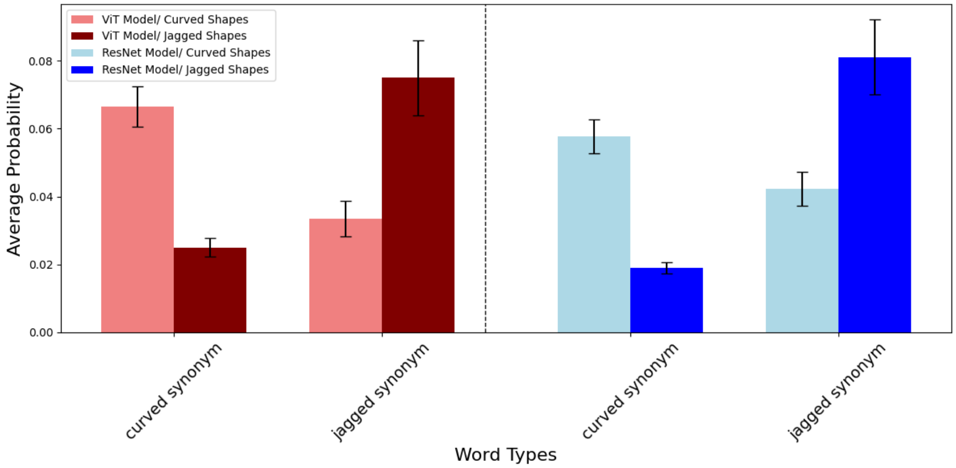


Figure 8: Average probability results for adjectives

Alper and Averbuch-Elor [2023]'s Word Set Figure 9 shows the average probability scores for the Alper and Averbuch-Elor [2023] 's word set. For the ViT model, we can see that the curved images achieve slightly higher probabilities for round words than jagged images. For sharp words, the jagged images achieve higher probabilities than curved images. However, the difference is minimal. The error bars also show that the variability is very small. The results of the Bayesian significance test show that there are no meaningful effects, as the coefficients for all terms, including the interaction between the shape condition and word category, are effectively zero. This result is reasonable, as the probabilities for this word set are very small due to the large number of words (648), which causes the model to distribute its predictions evenly across all categories. When using a Bayesian significance test with the Gaussian family on such small probabilities, the coefficients naturally remain close to zero, making it challenging to detect meaningful effects.

Although the ResNet model shows more pronounced visual patterns than the ViT model, the Bayesian significance test similarly reports no statistically significant effects. The interaction term remains near zero, and the confidence intervals are narrow, which indicates that the observed patterns are not strong enough to reach statistical significance. It is worth noting that the original study by Alper and Averbuch-Elor [2023] also used binary outcomes and applied mixed-effects logistic regression to model participants' correct answers for their significance testing and did not use probability scores.

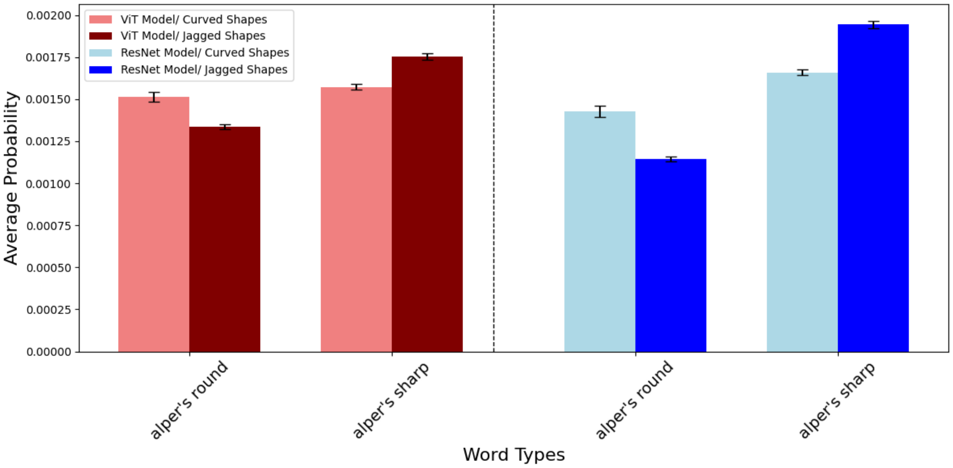


Figure 9: Average probability results for Alper and Averbuch-Elor [2023]'s words

One-syllable Pseudowords The results can be seen in figure 10 for one-syllable words. The results for the ViT model show that for *s_r* syllable, curved shapes achieve slightly higher probabilities than jagged shapes. Moreover, for the *p_nr* pseudowords, we can see slightly higher probabilities for jagged images compared to the curved images. These patterns are in line with our expectations. For *s_nr and p_r* pseudowords, probabilities are very similar between curved and jagged shapes. However, the Bayesian significance test results for the ViT model reveal a small positive effect for curved images paired with p_nr pseudowords, which is opposite to our expectations. Most interaction effects between word categories and shape conditions are insignificant, as the estimates are near zero, with narrow credible intervals including zero. The results of the ResNet model are very similar to the ViT model, with slightly higher probabilities for curved images given *s_r* pseudowords and higher probabilities for jagged images when given *p_nr* pseudowords. Bayesian significance test shows a small positive association between curved images and *s_r* pseudoword. However, it is not significant. Furthermore, it shows that there is also a slightly positive effect for curved images and p_nr pseudowords, which is opposite to our expectations.

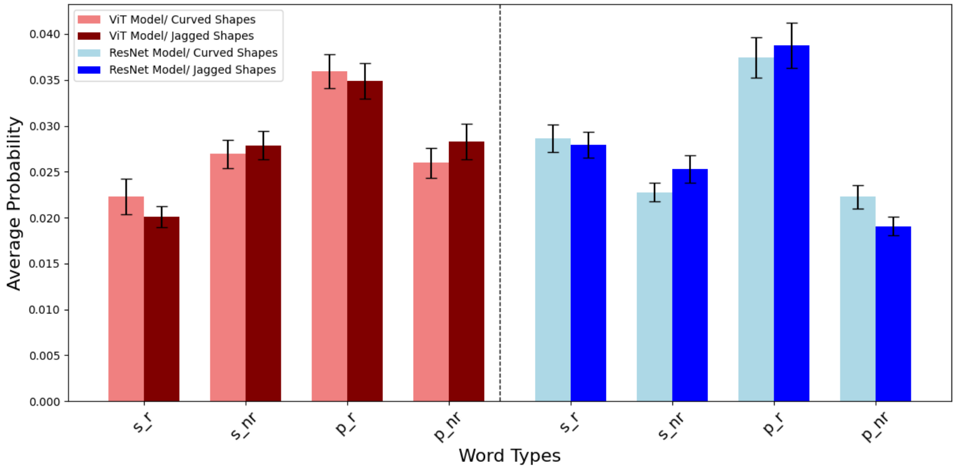


Figure 10: Average probability results for one-syllable pseudowords

Two-syllable Pseudowords Figure 11 shows the result for two-syllable pseudowords. For the ViT model, we can see that the probabilities for curved and jagged shapes are relatively close across all categories. Only for *s_r_s_r* pseudowords, probabilities are slightly higher for curved shapes, which are as expected. However, the results of the Bayesian model show that there are no statistically significant effects for any category.

For the ResNet model, although we can see higher probabilities for curved shapes when having *s_r_s_r* pseudowords and higher probabilities for jagged shapes when having *p_nr_p_nr* pseudowords, again, with credible intervals that include zero, significance test confirms that no statistically significant differences exist between curved and jagged shapes across the two-syllable word categories.

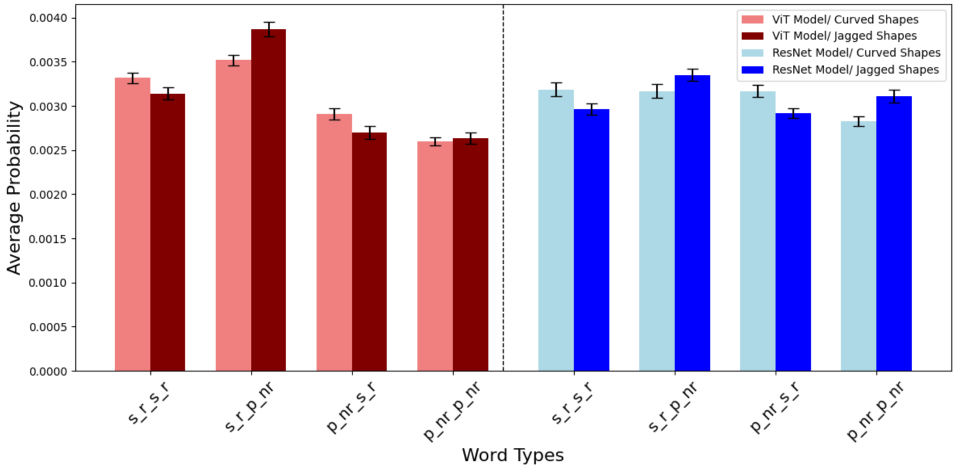


Figure 11: Average probability results for two-syllable pseudowords

Table 1 shows the credible intervals from the Bayesian significance test for different word types in this experiment.

Table 1: Details of the Bayesian significance test for experiment 4.1

Word Type	Model	Association(Image-Label)	Bayesian Significant Test	Statistics
Initial Words	ViT	Jagged-takete	Not Significant	b = 0.39, 95% CI = [-8.12, 9.19]
	ResNet	Curved-bouba	Not Significant	b = 3.66, 95% CI = [-5.43, 18.40]
	ResNet	Jagged-kiki	Not Significant	b = 2.07, 95% CI = [-5.57, 11.66]
	ResNet	Jagged-takete	Not Significant	b = 2.49, 95% CI = [-4.84, 17.65]
Adjectives	ViT	Curved-Curved	Significant	b = 0.07, 95% CI = [0.03, 0.10]
	ViT	Jagged-Jagged	Significant	b = 0.08, 95% CI = [0.02, 0.15]
	ResNet	Curved-Curved	Significant	b =0.06, 95% CI = [0.03, 0.09]
	ResNet	Jagged-Jagged	Significant	b = 0.08 95% CI = [0.01, 0.14]
Alper's Words	ViT	-	Not Significant	-
	ResNet	-	Not Significant	-
One-Syllable Words	ViT	Curved-Jagged	Significantly Opposite	b = 0.03, 95% CI = [0.01, 0.04]
	ResNet	Curved-Jagged	Significantly Opposite	b = 0.02, 95% CI = [0.01, 0.03]
	ResNet	Curved-Curved	Not Significant	b = 0.01, 95% CI = [-0.01, 0.02]
Two-Syllable Words	ViT	-	Not Significant	-
	ResNet	-	Not Significant	-

4.2 Image To Text Matching Task

In this experiment, we identify each image's label with the highest probability. The goal is to analyze how often the label with the highest probability aligns with the expected associations. This experiment is conducted in two ways, which will be explained below.

4.2.1 Based on Complete Word Structure

This experiment examines how effectively the CLIP models associate specific words with the expected visual shapes (curved and jagged) by selecting the label with the highest probability for each image. If the bouba-kiki effect exists in VLMs, there should consistently be higher percentages of matches between the predicted labels and the expected shapes. For instance, curved-related labels should align more frequently with curved images, and jagged-related labels should align more frequently with jagged images. It should be noted that in this experiment, the word selection is based on the complete form of the words.
Figure 12 shows the percentage of cases where the selected label aligns with the expected association for each word type and image shape. The x-axis represents the different word types, and the y-axis indicates the percentage of correct label assignments. The blue bars show the results for the ViT-based model, with darker blue indicating jagged images and lighter blue indicating curved images. Similarly, the green bars represent the results for the ResNet version of CLIP, with darker green indicating jagged images and lighter green indicating curved images.

Initial Words For the initial words, in the ViT model, 100% of curved images are matched to curved initial words, while only 6% of jagged images are associated with jagged words. The Bayesian analysis supports this result and shows a strong association between curved images and curved initial words. The significant test also shows that jagged images might increase the likelihood of association with jagged initial words, but this effect is not significant. Similarly, in the ResNet model, the results for curved initial words are similar to those of the ViT model, with the Bayesian analysis confirming this strong association. However, it can be seen that the

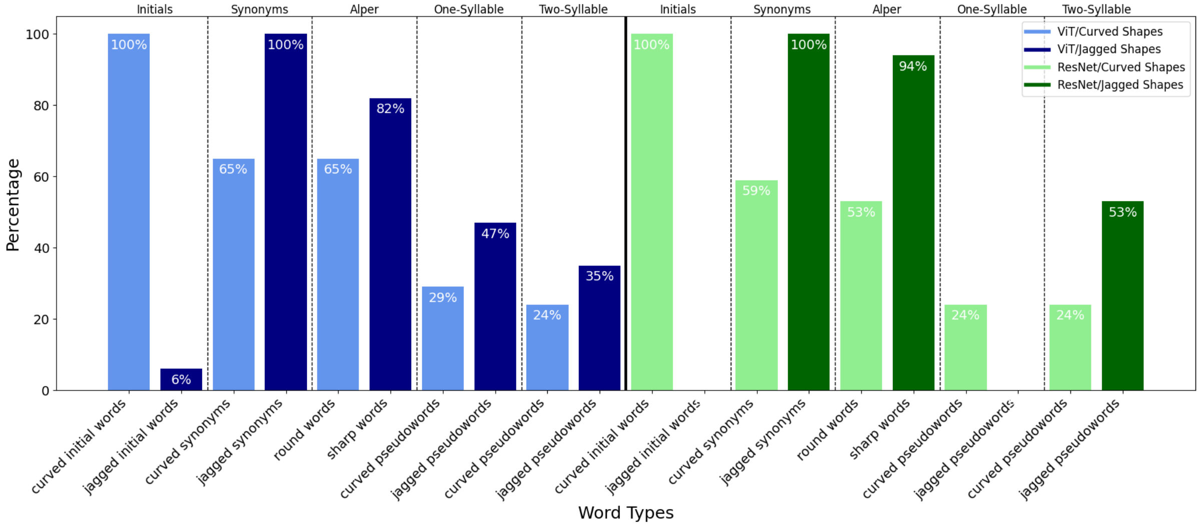


Figure 12: Percentage of selecting the expected label for each image across different word types

jagged images are not strongly associated with jagged words. The Bayesian significance test also confirms that the results for jagged images and initial words are insignificant.

Adjectives For adjectives, the results of the ViT model show that 100% of jagged images are matched to jagged synonyms, while 65% of the curved images are correctly matched to curved synonyms. The Bayesian significance test results support a strong match between jagged images and jagged words. However, a moderate association between curved images and curved synonyms can be seen, but it is not statistically significant. Similarly, in the ResNet model, 100% of the jagged images align with jagged synonyms, while only 59% of curved shapes align with curved synonyms. For jagged images, the Bayesian results confirm the observed pattern. However, as with the ViT model, the Bayesian results show a moderate association between curved images and curved synonyms, which is not significant.

Alper and Averbuch-Elor [2023]'s Word Set For Alper and Averbuch-Elor [2023] 's word set, the results of the ViT model show that 82% of the jagged images are matched to sharp words and 65% of the curved images are matched to round words. The Bayesian significant test results show that jagged shapes are more likely to be assigned to sharp words. Furthermore, assigning the curved images to round words has a positive effect, but it is not statistically significant. The Resnet model for this word set has a similar result to the ViT model but with different percentages: 94% of jagged images are associated with sharp words. In contrast, 53% of curved images are associated with round words. The Bayesian analysis confirms that jagged images are more likely to align with sharp words. While a slight positive association exists between curved images and round words, this effect is not statistically significant.
It should be noted that the observed significant patterns in this experiment, compared to the Bayesian significant results in Section 4.1, can be attributed to the nature of the test. Here, the analysis is based on binary outcomes (whether a word is correctly matched to an image) rather than directly comparing probability distributions. In other words, this approach is similar to what was done in the study of Alper and Averbuch-Elor [2023]. They used binary outcomes to model participants' correct answers.

One-syllable pseudowords For one-syllable words, the results of the ViT model show that 29% of the curved images are matched to curved labels (s_r) and 47% of the jagged images are matched to jagged words (p_nr_p_nr), which in both cases show a preference for the unexpected pattern. The Bayesian analysis confirms this finding and shows no significant difference between the predicted probabilities for curved and jagged shapes, as the credible interval includes zero. In ResNet, we can see that 24% of the curved images are matched to curved labels, and all the jagged images are also matched to curved labels, which is an unexpected pattern. The Bayesian analysis results show that jagged images are more likely to align with s_r words, and curved images are more likely to match with p_nr words, both of which are the opposite of our expectations based on the bouba-kiki effect.

Two-syllable pseudowords For two-syllable words and the ViT model, it can be seen that 24% of the curved images are matched to the curved labels (s_r_s_r), and 35% of the jagged images are matched to jagged labels (p_nr_p_nr). The significance test results show that jagged images are more likely to align with s_r_s_r words, and curved images are more likely to match with p_nr_p_nr words. These findings contradict the associations predicted by the bouba-kiki effect. Similarly, in the ResNet model, 24% of curved images are matched to curved labels, while 53% of jagged images are matched to jagged labels. The Bayesian analysis shows that curved images are more likely to align with p_nr_p_nr words, which is also contrary to expectations. Furthermore, no significant results are found for jagged images in the ResNet model.

Table 2 shows the coefficients related to the significance test in this experiment.

Table 2: Details of the Bayesian significance test for experiment 4.2.1

Word Type	Model	Association (Image-Label)	Bayesian Significant Test	Statistics
Initial Words	ViT	Curved-Curved	Significant	b = 8.17, 95% CI = [2.43, 22.39]
	ViT	Jagged-Jagged	Not Significant	b = 10.66, 95% CI = [-1.21, 72.30]
	ResNet	Curved-Curved	Significant	b = 12.82, 95% CI = [2.38, 54.16]
Adjectives	ViT	Jagged-Jagged	Significant	b = 9.31, 95% CI = [2.87, 29.64]
	ViT	Curved-Curved	Not Significant	b = 0.69, 95% CI = [-0.38, 1.78]
	ResNet	Jagged-Jagged	Significant	b = 8.37 95% CI = [2.57, 23.44]
	ResNet	Curved-Curved	Not Significant	b = 0.42 95% CI = [-0.54, 1.47]
Alper's Words	ViT	Jagged-Jagged	Significant	b = 2.28, 95% CI = [0.70, 4.04]
	ViT	Curved-Curved	Not Significant	b = 0.65, 95% CI = [-0.35, 1.75]
	ResNet	Jagged-Jagged	Significant	b = 3.26, 95% CI = [1.21, 6.12]
	ResNet	Curved-Curved	Not Significant	b = 0.17, 95% CI = [-0.78, 1.21]
One-Syllable Words	ViT	-	Not Significant	-
	ResNet	Jagged-Curved	Significantly Opposite	b = 9.82, 95% CI = [3.42, 28.97]
	ResNet	Curved-Jagged	Significantly Opposite	b = 1.35, 95% CI = [0.16, 2.74]
Two-Syllable Words	ViT	Jagged-Curved	Significantly Opposite	b = 1.89, 95% CI = [0.31, 3.51]
	ViT	Curved-Jagged	Significantly Opposite	b = 1.26, 95% CI = [0.23, 2.47]
	ResNet	Curved-Jagged	Significantly Opposite	b = 1.29, 95% CI = [0.12, 2.58]

4.2.2 Effect of Consonants and Vowels

This experiment is similar to the one described in Section 4.2.1. However, this experiment focuses on the effects of individual consonants and vowels instead of analyzing the complete word structure. Specifically, we calculate the percentage of cases where the selected label for each image aligns with expectations based on its structure. The results are analyzed for both the ViT and ResNet-based versions of CLIP.

Figure 13 shows the results for both models across different conditions, including rounded and non-rounded vowels, sonorant and plosive consonants for one-syllable words, and cases with at least one *s_r* or *p_nr* syllables for two-syllable words. The x-axis represents the conditions in these plots, while the y-axis indicates the percentage of shape-label matches. The blue bars represent the results for the ViT model, with light blue corresponding to the percentage of matches for curved shapes and dark blue for jagged shapes. Similarly, the green bars represent the results for the ResNet model, with light green indicating the percentage of matches for curved shapes and dark green for jagged shapes.

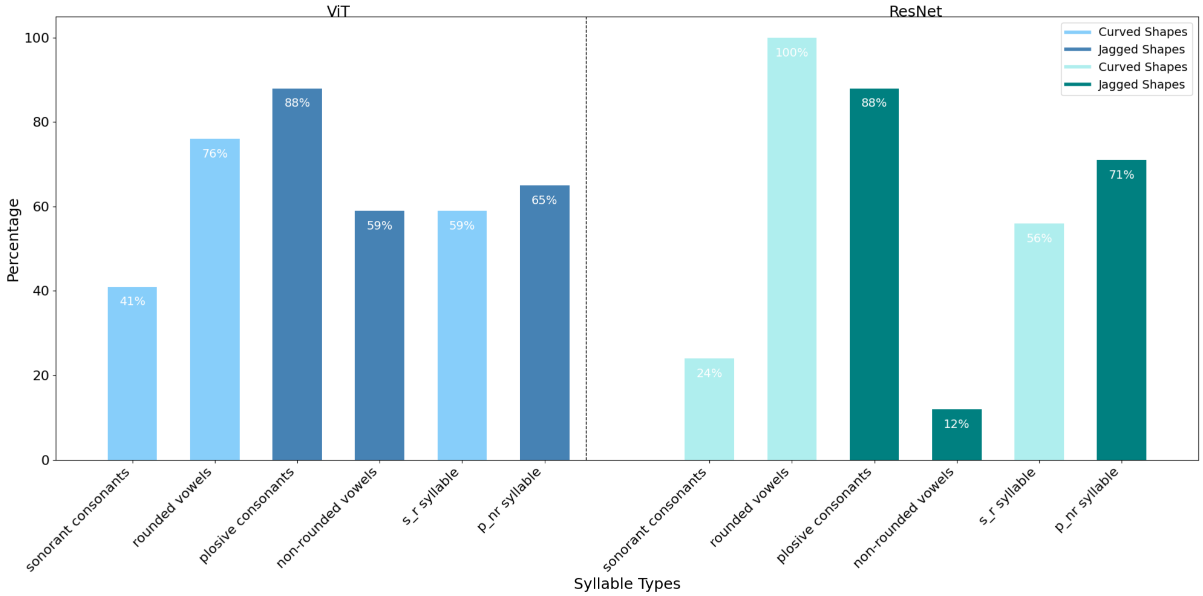


Figure 13: Percentage of selecting the expected label for each image based on consonants and vowels

Consonants Based on figure 13, we can see that in the ViT model, 41% of the curved images are matched to labels that have sonorant consonants. The results of the significance test confirm this pattern. It shows no significant effect of matching curved images with sonorant consonants. The ViT model also shows that 88% of jagged images are matched to labels that contain plosive consonants in their structure, which shows a strong association. The Bayesian results also confirm this and show that jagged shapes are significantly more likely to be associated with plosive consonants.
On the other hand, the ResNet-based model shows a lower percentage of 24% for sonorant consonants, which shows a weaker association of curved shapes with sonorant consonants compared to the ViT model. For plosive consonants, the ResNet model shows that 88% of the jagged images are matched with plosive consonants. The Bayesian analysis shows a positive effect of associating jagged images with plosive consonants; however, this effect is not statistically significant despite the observed pattern. Additionally, it reveals that curved images are statistically more likely to be associated with plosive consonants, contrary to our expectations.

Vowels For rounded vowels, we can see that in the ViT mode, 76% of the curved images are associated with labels that contain rounded vowels. The Bayesian analysis supports this observation and shows a significant association between curved images and rounded vowels. Moreover, for non-rounded vowels, the ViT-based model shows that 59% of jagged images

are associated with non-rounded vowels, and the result of the significance test also shows that jagged shapes are significantly more likely to be associated with non-rounded vowels.
In the ResNet-based model, all the curved images are associated with rounded vowels. The Bayesian significance test confirms this association and shows a strong association between curved images and rounded vowels. For non-rounded vowels, the ResNet-based model shows 12% of jagged images are matches to non-rounded vowels, which indicates a weak association. The Bayesian analysis confirms this observation and shows no significant association between jagged images and non-rounded vowels.

Syllables The last two bars in figure 13, which relate to two-syllable words, show the percentage of cases where the selected two-syllable label for each image contains at least one s_r (curved) or p_nr (jagged) component.
The plot for the ViT model shows that 59% of the curved images are associated with selected syllables containing at least one s_r syllable. The Bayesian analysis indicates a small effect of matching curved images with s_r syllables, although this effect is not statistically significant. Moreover, the ViT plot shows that 65% of the jagged images align with selected labels containing at least one p_nr syllable. The Bayesian analysis confirms that jagged images are significantly more likely to align with p_nr pseudowords.
On the other hand, the ResNet model shows that 56% of the curved images are associated with labels that contain at least one s_r syllable. The Bayesian analysis shows that curved images are likely associated with s_r syllables. However, this effect is not statistically significant. For p_nr syllables, 71% of the jagged images are matched to pseudowords that contain at least one p_nr syllables. The Bayesian analysis supports this finding and indicates that jagged images are significantly more likely to align with p_nr syllables.
Table 3 shows the details of the Bayesian significance test.

Table 3: Details of the Bayesian significance test for experiment 4.2.2

Word Type	Model	Association (Image-Label)	Bayesian Significant Test	Statistics
Consonants	ViT	Jagged-Plosive	Significant	b = 1.81, 95% CI = [0.19, 3.76]
	ResNet	Curved-Plosive	Significantly Opposite	b = 1.20 95% CI = [0.18, 2.41]
	ResNet	Jagged-Plosive	Not Significant	b = 0.94 95% CI = [-0.83, 3.00]
Vowels	ViT	Curved-Rounded	Significant	b = 1.25, 95% CI = [0.22, 2.51]
	ViT	Jagged-Non_rounded	Significant	b = 1.64, 95% CI = [0.16, 3.30]
	ResNet	Curved-Rounded	Significant	b = 7.86, 95% CI = [2.26, 24.74]
	ResNet	Jagged-Non_rounded	Not Significant	b = 5.76, 95% CI = [-0.28, 20.80]
Syllable	ViT	Jagged-p_nr	Significant	b = 0.99, 95% CI = [0.02, 1.90]
	ViT	Curved-s_r	Not Significant	b = 0.37, 95% CI = [-0.33, 1.10]
	ResNet	Jagged-p_nr	Significant	b = 1.15, 95% CI = [0.15, 2.13]
	ResNet	Curved-s_r	Not Significant	b = 0.27, 95% CI = [-0.46, 0.98]

4.3 Analyzing the Sum of Intensities

In this experiment, our goal is to determine whether attention is allocated more in the expected region of an image based on different word types. For example, for a curved-related word, we expect higher attention in the curved region of the image. For a jagged-related word, we expect higher attention in the jagged region. To do this, we begin by calculating the sum of intensities in each region to evaluate if the attention is higher in the expected region. Then, for each word type, we calculate the percentage of cases where the sum of intensities in the expected region is greater than the other region. By performing this analysis, we can assess whether the attention patterns of ViT-based and ResNet-based CLIP models align with expectations. Figure 14 shows the percentage of cases where the attention is higher in the expected region based on the given label across different word types and for both models. In this plot, the x-axis represents different word categories, and the y-axis shows the percentage of cases where the sum of intensities (attention) is higher in the expected region. The red horizontal line shows the chance performance (baseline), set at 50%, which indicates random alignment. The orange boxes show the result for the ViT model, and the blue boxes show the result for the ResNet model.

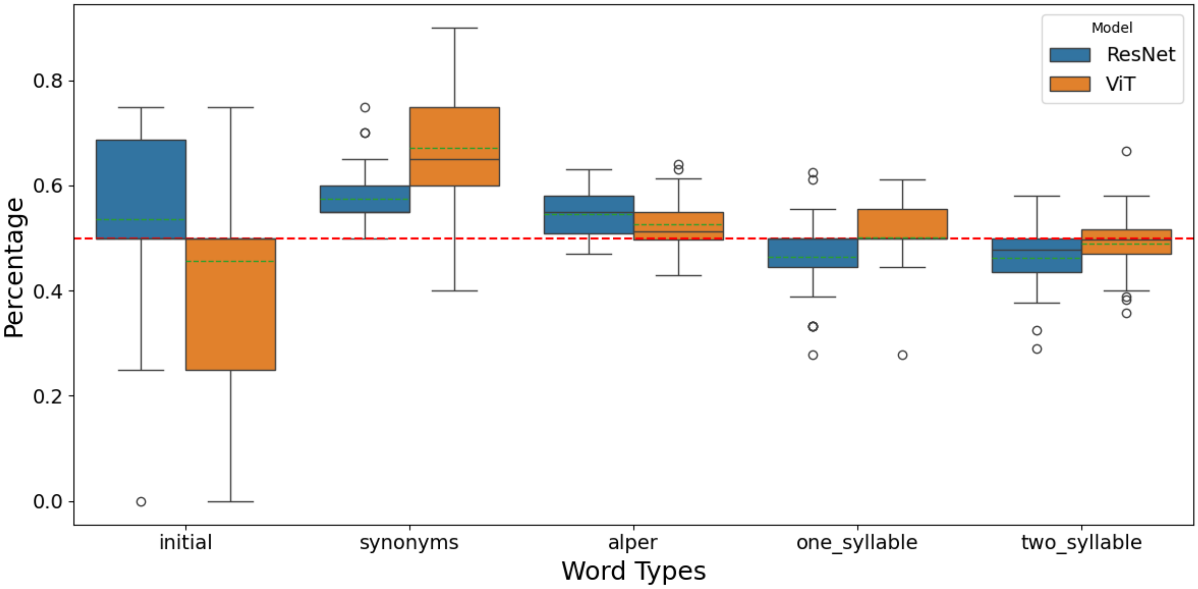


Figure 14: Percentage of achieving higher sum of intensities in expected regions across different word types for ViT and ResNet models

Initial Words The ViT model's result shows considerable variability, with percentages below and above chance, which indicate inconsistent attention alignment with the expected regions. The ResNet model's percentage for this word type is higher than chance, suggesting that ResNet focuses better on expected regions for initial words than ViT. However, the Bayesian significance test's results do not show any significant effect for this word type for both models, which shows that the alignment remains around chance.

Real-Words For this word set, both models show strong performance. The ResNet model performs above 50%, which shows attention alignment with expected regions. The ViT model performs even better, with a higher percentage of achieving aligned attention. For the ViT model, the Bayesian significance test confirms strong attention alignment with expected regions. For the ResNet model, the Bayesian significance test shows a positive effect. However, it is not significant.

Alper and Averbuch-Elor [2023]'s Word Set Both models perform near or slightly above the chance level for this word set. However, the significance test results show that the observed pattern is statistically significant for both models, indicating a significant alignment with higher attention in the expected regions for both models.

One-syllable Pseudowords For the ViT model, attention alignment for one-syllable words is slightly above chance. Still, the result of the Bayesian test shows that this is not statistically significant, indicating that the observed pattern may be due to random variation. For the ResNet model, attention alignment performs below chance, and the Bayesian test confirms that the result is statistically significant, showing that attention aligns less frequently with the expected regions for one-syllable words.

Two-syllable Pseudowords Both models perform close to chance for this word set, indicating weak and inconsistent alignment with expected regions. The Bayesian significance test shows that attention aligns significantly less frequently with the expected regions in both models in both models.
A finer-grained version of this experiment can be found in Appendix E, and the details of the Bayesian significance test are presented in Table 4.

Table 4: Details of the Bayesian significance test for experiment 4.3

Word Type	Model	Bayesian Significant Test	Statistics
Initial Words	ViT	Not Significant	-
	ResNet	Not Significant	-
Adjectives	ViT	Significant	b = 0.62, 95% CI = [0.46, 0.80]
	ResNet	Not Significant	b = 0.12, 95% CI = [-0.03, 0.28]
Alper's Words	ViT	Significant	b = 0.10, 95% CI = [0.07, 0.13]
	ResNet	Significant	b = 0.18, 95% CI = [0.15, 0.20]
One-Syllable Words	ViT	Not Significant	-
	ResNet	Significantly Opposite	b = -0.33, 95% CI = [-0.49, -0.18]
Two-Syllable Words	ViT	Significantly Opposite	b = -0.14, 95% CI = [-0.20, -0.08]
	ResNet	Significantly Opposite	b = -0.33, 95% CI = [-0.39, -0.27]

5 Discussion

This thesis explored whether VLMs, specifically ViT-based and ResNet-based CLIP models, display robust cross-modal associations between linguistic inputs and visual shapes. By examining probabilities, label assignments, and attention patterns across different word types, we aimed to evaluate how well these models show the same associations as observed in human cognition. The results provide insights into the models' ability to capture and represent connections between language and visual shapes. Table 5 summarizes the findings of each experiment.
In the **probability comparison** experiment, the expected pattern of achieving higher average probabilities for curved images paired with curved labels and jagged images paired with jagged labels was observed only for adjectives, which served as the benchmark. The plots and Bayesian significance tests confirmed this alignment, indicating that the models performed reliably with familiar and meaningful words. However, the expected patterns were not consistently observed for all other word types, including initial words, Alper and Averbuch-Elor [2023]' words, one-syllable pseudowords, and two-syllable pseudowords. Some cases (initial words and round and

sharp pseudowords) showed a positive effect in these word types, but the results were not statistically significant. Other cases (one-syllable and two-syllable pseudowords) deviated from expectations based on the bouba-kiki effect. This suggests that the models have difficulty creating strong associations for these word types.

In the **second experiment**, we calculated the percentage of cases where each image's selected label (based on the complete word structure) matched the expected association. The results of the Bayesian significance tests revealed that, overall, the expected patterns were not statistically significant across most word types. Furthermore, for certain word types such as initial words, adjectives, and Alper and Averbuch-Elor [2023]'s word set, significant results were found only for one condition, which shows the variability in the strength of the associations.

In the **third experiment**, which was similar to the approach used in the study by Verhoef et al. [2024], we analyzed the effect of consonants and vowels on the selected labels for each image. The Bayesian significance results showed that for the ViT model, jagged images were more likely to align with labels containing plosive consonants. However, there was no statistically significant effect between curved images and sonorant consonants. Moreover, curved images were more likely to align with labels containing rounded vowels, while jagged images were more likely to align with non-rounded vowels. These findings align with expectations and suggest that the ViT model can capture certain shape-related associations based on phonetic components, especially vowels. For the ResNet model, only curved images were more likely to align with round vowels, and all the other effects were not statistically significant. When analyzing the two-syllable pseudowords, we examined whether the selected labels contained at least one s_r syllable for curved images or at least on p_nr syllable for jagged images. The Bayesian analysis showed that jagged images were significantly more likely to align with labels containing at least one p_nr syllable in both models. In contrast, no significant effect was found for curved images paired with labels containing at least one s_r syllable in either model. These results suggest that both models, particularly ViT, show some sensitivity to individual phonetic components and syllable structures. However, the lack of consistency and significant effects in all conditions highlight their limitations in consistently capturing these associations.

In the **last experiment**, we calculated the percentage of cases where higher attention (measured as the sum of intensities) was observed in the expected regions for different word types. The Bayesian significance results showed that initial words did not produce significant results for either model, suggesting that attention was not consistently focused on the expected regions. The ViT model showed significant results in English adjectives, confirming that attention aligned well with the expected regions. However, although a positive alignment was observed for the ResNet model, it was not statistically significant, indicating weaker alignment compared to the ViT model. Both models showed significant results for the word set used in the work of Alper and Averbuch-Elor [2023], indicating that attention was consistently directed toward the expected regions for round and sharp words. In one-syllable words, no significant results were found for the ViT model. Moreover, attention aligned significantly less frequently with the expected regions for the ResNet model, suggesting a clear misalignment for this word type. Finally, both models demonstrated significant results opposite our expectation for two-syllable words, indicating that attention aligned significantly less frequently with the expected regions.

Table 5 demonstrates that the bouba-kiki effect is not consistently present in the VLMs analyzed in this thesis. Several factors could explain this. **First**, tokenization, which is an important step in how VLMs process text. Words may be broken down into smaller, often meaningless parts during tokenization. This process can distort word representations and weaken the semantic alignment with visual shapes. In contrast, humans do not process words in this fragmented manner. This mismatch between how humans and VLMs handle words helps explain why models often fail to replicate the intuitive sound-symbolic associations observed in humans. **Second**, we observed in our experiments that the ViT model performed better than the ResNet model, which might be related to differences in their architectures. Specifically, cross-modal associations require integrating global visual features with linguistic inputs. The ResNet model's early layers capture local features like edges and small patterns. Later layers progressively combine these features to form a broader understanding of the image. Therefore, the global context only emerges after several layers. Still, by that point, some of the finer relationships between the smaller parts of the image may have been lost or weakened. This architectural limitation makes it harder for ResNet to effectively capture cross-modal associations. In contrast, the ViT model processes images using a global self-attention mechanism, allowing it to consider the entire image from the beginning. This architecture of the ViT model makes it more suited for cross-modal association tasks. **Third**, the CLIP model is trained using a contrastive learning objective designed to align image and text embeddings. However, it does not focus on sound-symbolic associations because these associations are not part of CLIP's training objective. Moreover, the dataset used to train CLIP consists of commonly found language and image pairs. However, the pseudowords used in this thesis are unlikely to appear in the training data, meaning the model has no prior knowledge of their sound-symbolic properties or potential associations with visual features. Unlike humans, who intuitively associate these pseudowords with visual shapes due to cognitive biases, the model relies entirely on patterns within the training data, and these patterns are absent for pseudowords. Additionally, the dataset may introduce bias by including more examples of one type of word-image pairing (e.g. curved words with curved images) and fewer examples of the other. This imbalance leads to stronger associations for the more frequently represented pairings, while the associations for the less common pairings remain weaker due to limited exposure during training.

The overall finding suggests that cross-modal associations between linguistic inputs and visual shapes are not consistently present in VLMs such as ViT-based and ResNet-based CLIP models. The associations depend on multiple factors, including the model architecture, task design, the word type, and the images used. While our results suggest that cross-modal associations, such as the bouba-kiki effect, are not consistently present in VLMs, the potential value of shared preferences between humans and machines remains significant. Establishing shared preferences and common understanding between humans and AI can enhance their interactions. This helps in creating AI systems that work effectively and resemble how humans process and communicate meaning.

6 Conclusion

This thesis investigated whether VLMs, like CLIP, show consistent cross-modal associations between linguistic inputs (different word types) and visual shapes (curved and jagged images). The CLIP model was chosen due to its better performance and reported alignment with human-like associations in prior studies [Verhoef et al., 2024, Demircan et al., 2024]. To examine cross-modal associations in VLMs, we conducted different experiments: probability comparisons, image-to-text matching, and attention pattern analysis using the sum of intensities. Across these experiments, we observed that cross-modal associations depend on multiple factors, including the model architecture, the word type, and the specific experimental task. English synonyms, serving as the benchmark, showed stronger associations, while pseudowords

Table 5: Summary of ViT and ResNet model performance across experiments

Experiment	**Task/Word Type**	**ViT**	**ResNet**
Probability Comparison	Initial Words	Fail	Fail
	Adjectives	Pass	Pass
	Alper's Words	Fail	Fail
	One-Syllable Words	Fail	Fail
	Two-Syllable Words	Fail	Fail
Choosing the preferred label	Initial Words	Fail	Fail
	Adjectives	Fail	Fail
	Alper's Words	Fail	Fail
	One-Syllable Words	Fail	Fail
	Two-Syllable Words	Fail	Fail
Effect of Consonants and Vowels	Sonorant Consonants	Fail	Fail
	Plosive Consonants	Pass	Fail
	Rounded Vowels	Pass	Pass
	Non-Rounded Vowels	Pass	Fail
	s_r syllable	Fail	Fail
	p_nr syllable	Pass	Pass
Attention Pattern Analysis	Initial Words	Fail	Fail
	Adjectives	Pass	Fail
	Alper's Words	Pass	Pass
	One-Syllable Words	Fail	Fail
	Two-Syllable Words	Fail	Fail

demonstrated weaker and inconsistent patterns. The round and sharp pseudowords generated in prior research [Alper and Averbuch-Elor, 2023] performed well in certain tasks but did not show robust performance across all experiments. Moreover, while the ViT model performed better than ResNet in experiments analyzing the effect of consonants and vowels or attention alignment experiments, it still showed inconsistencies. The ResNet model was the most inconsistent and struggled to exhibit any clear patterns. These findings suggest that CLIP models can capture some shape-related associations in certain cases. However, their performance lacks robustness across different model architectures, experiments, and word types, highlighting their limitations in replicating the detailed cross-modal relationships observed in human cognition.

7 Limitations

This thesis provides insights into the cross-modal associations in ViT-based and ResNet-based CLIP models. However, its limitations provide opportunities for future research. First, this thesis focused only on linguistic-visual associations and did not examine other modalities. Future work should explore other types of cross-modal associations, such as sound-color, to provide a more comprehensive understanding of how different sensory modalities interact in VLMs. Exploring these additional modalities reveals new patterns in cross-modal associations in VLMs.

Second, we did not investigate the effects of different model architectures in detail. Examining models with diverse architectures would help clarify how architectural differences influence the emergence of cross-modal associations. Third, we used the standard tokenization method and did not evaluate alternatives that might preserve holistic word structures. Investigating tokenization methods that align more closely with how humans process words could potentially lead to stronger associations between linguistic and visual features. Fourth, we did not consider the effect of background color. Analyzing how different background colors influence cross-modal associations would provide valuable insights into how color affects the model's associations similar to human cognition Hubbard [1996]. Lastly, when generating the attention patterns using the Grad-CAM technique, we only focused on regions that positively affect the model's prediction. Analyzing regions that negatively contribute to the model's prediction provides insights into the model's preferences and decision-making processes.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018.

Morris Alper and Hadar Averbuch-Elor. Kiki or bouba? sound symbolism in vision-and-language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78347–78359. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f74054328beeb0c21a9b8e99da557f5a-Paper-Conference.pdf.

Damián E Blasi, Søren Wichmann, Harald Hammarström, Peter F Stadler, and Morten H Christiansen. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823, 2016.

Christine Cuskley and Simon Kirby. Synesthesia, Cross-Modality, and Language Evolution. In *Oxford Handbook of Synesthesia*. Oxford University Press, 12 2013. ISBN 9780199603329. doi: 10.1093/oxfordhb/9780199603329.013.0043. URL https://doi.org/10.1093/oxfordhb/9780199603329.013.0043.

Christine Cuskley, Julia Simner, and Simon Kirby. Phonological and orthographic influences in the bouba–kiki effect. *Psychological research*, 81:119–130, 2017.

Can Demircan, Tankred Saanum, Leonardo Pettini, Marcel Binz, Blazej M Baczkowski, Christian F. Doeller, Mona M. Garvert, and Eric Schulz. Evaluating alignment between humans and neural network representations in image-based learning tasks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=8i6px5W1Rf.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.143. URL https://aclanthology.org/2022.emnlp-main.143/.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

T L. Hubbard. Synesthesia-like mappings of lightness, pitch, and melodic interval. *American journal of psychology*, 109(2):219–38, 1996.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *The 2023 Conference on Empirical*

Methods in Natural Language Processing, 2023. URL https://openreview.net/forum?id=RN5KLywTll.

Katharina Kann and Mauro M. Monsalve-Mercado. Coloring the black box: What synesthesia tells us about character embeddings, 2021. URL https://arxiv.org/abs/2101.10565.

Wolfgang Köhler. *Gestalt Psychology*. New York: Horace Liveright, 1929.

Wolfgang Köhler. *Gestalt Psychology*. (2nd ed.) New York: Horace Liveright, 1947.

Tom Kouwenhoven, Tessa Verhoef, Roy De Kleijn, and Stephan Raaijmakers. Emerging grounded shared vocabularies between human and machine, inspired by human language evolution. *Frontiers Artif. Intell.*, 5:886349, 2022. URL http://dblp.uni-trier.de/db/journals/frai/frai5.html#KouwenhovenVKR22.

Gwilym Lockwood and Mark Dingemanse. Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in psychology*, 6: 145602, 2015.

Lawrence E. Marks. On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American journal of psychology*, pages 173–188, 1974.

Daphne Maurer, Thanujeni Pathman, and Catherine J Mondloch. The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental science*, 9(3):316–322, 2006.

Alan Nielsen and Drew Rendall. The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and cognition*, 4(2):115–125, 2012.

Alan Nielsen and Drew Rendall. Parsing the role of consonants versus vowels in the classic takete-maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 67(2):153, 2013.

Cesare Valerio Parise and Charles Spence. 'when birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PloS one*, 4(5):e5664, 2009.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Vilayanur S. Ramachandran and Edward M. Hubbard. Synaesthesia–a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12):3–34, 2001.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vision*, 128(2):336–359, February 2020. ISSN 0920-5691. doi: 10.1007/s11263-019-01228-7. URL https://doi.org/10.1007/s11263-019-01228-7.

Jessica Taubert, Deborah Apthorp, David Aagten-Murphy, and David Alais. The role of holistic processing in face perception: Evidence from the face inversion effect. *Vision Research*, 51: 1273–1278, 2011. URL `https://api.semanticscholar.org/CorpusID:16183020`.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248, June 2022.

L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9(nov):2579–2605, 2008. ISSN 1532-4435. Pagination: 27.

Tessa Verhoef, Kiana Shahrasbi, and Tom Kouwenhoven. What does kiki look like? cross-modal associations between speech sounds and visual shapes in vision-and-language models. In Tatsuki Kuribayashi, Giulia Rambelli, Ece Takmaz, Philipp Wicke, and Yohei Oseki, editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 199–213, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.cmcl-1.17. URL `https://aclanthology.org/2024.cmcl-1.17`.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach, 2024. URL `https://arxiv.org/abs/2407.01100`.

Chris Westbury. Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and language*, 93(1):10–19, 2005.

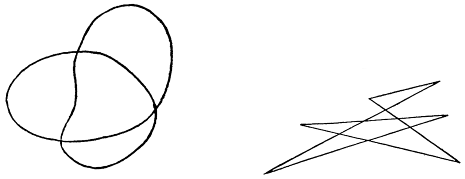


Figure 15: Images from [Köhler, 1929, 1947]

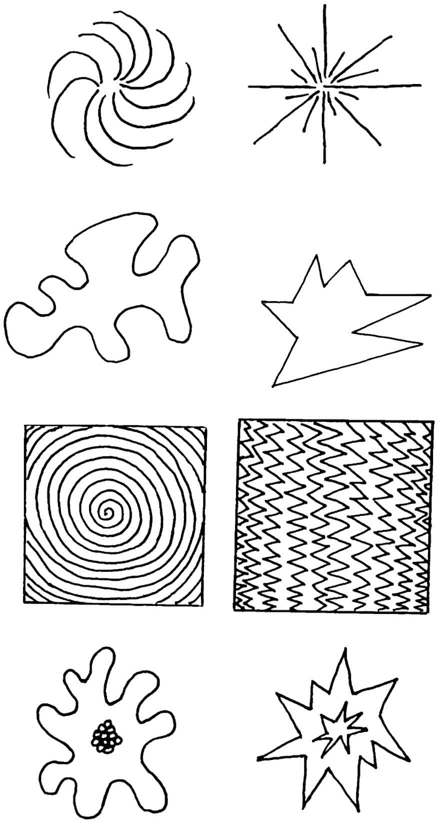


Figure 16: Images from [Maurer et al., 2006]

A Complete Set of Images Used in This Study

This section represents all the images used in this study. Figure 15 16, 17, 18 show the images.

B Examples of Attention Pattern Visualization

This section provides more examples of visualizing the attention pattern for different text prompts in the ViT and ResNet-based versions of the CLIP. The left image shows the input image with the text prompt above it, and the right image shows the resulting attention pattern. Figures 19 show these images.

C Performance of Different Metrics in Quantifying the Attention Pattern

This section shows the metrics' performance in quantifying the attention pattern. Figure 20 shows the performance and stability of different metrics across different word types for both models. As can be seen, the sum of intensities shows better performance and more stability compared to other metrics.

D Probability Experiments Using the Specific Version of CLIP Model Used in Alper and Averbuch-Elor [2023]

This section shows the results of the probability experiments (4.1, 4.2) using the specific version of the CLIP model used in the work of Alper and Averbuch-Elor [2023]. Figure 21 shows that patterns are similar to those observed in our ViT and ResNet versions of CLIP. In some cases, we observe stronger patterns with this specific model, which could be because of its larger architecture, finer granularity (patch size of 14x14), and training on a significantly larger dataset. However, the Bayesian model's result shows that most effects are not statistically significant. For the adjectives, a significant interaction effect (*b = 0.11, and 95% credible interval = [0.01,0.22]*) is observed, which shows that jagged images lead to higher probabilities for jagged synonyms compared to curved images. However, none of the fixed effects or interactions are significant for the initial, one-syllable, and two-syllable pseudowords. The Bayesian significant result for the word set used in the study of Alper and Averbuch-Elor [2023] is very close to zero, which is reasonable due to the nature of the dataset and the testing methodology. With 648 words in the word set, the probabilities for individual word-image pairs are inherently very small. When using Bayesian significance testing with a Gaussian distribution on such small probabilities, the coefficients approach zero, making it challenging to detect meaningful effects. This result aligns with the original study by Alper and Averbuch-Elor [2023], which did not rely on probability scores for statistical testing but instead used binary outcomes.

Figure 22 presents the experiment conducted in Section 4.2.1 using the specific version of CLIP employed in the study by Alper and Averbuch-Elor [2023]. The Bayesian significance results show that jagged images are significantly less likely to match with curved initial labels (*b = -10.45, credible interval = [-54.67, -1.0]*). Moreover, jagged images are significantly less likely to be associated with curved synonyms (*b = -2.82, credible interval = [-4.8, -1.02]*). For the word set used in Alper and Averbuch-Elor [2023], the Bayesian significance test does not show strong evidence of associations between round or sharp labels and curved or jagged images. Furthermore, the significance test indicates that jagged images are significantly more likely to match with curved-related one-syllable (s_r)(*b = 2.74, credible interval = [0.62, 5.63]*) and two-syllable pseudowords (s_r_s_r)(*b = 2.54, credible interval = [0.84, 4.53]*), which is opposite to our expectations.

Figure 23 shows the experiment conducted in Section 4.2.2 for this specific model. The results of the Bayesian significance test do not indicate any significant associations between curved or jagged shapes and sonorant or plosive consonants. However, the results show that jagged images are significantly more likely to match with rounded vowels (*b = 2.36, credible interval = [0.27, 5.15]*), which contradicts our expectations. Furthermore, no significant evidence exists for an association between curved and jagged images and labels containing at least one s_r or p_nr syllable.

E Sum of Intensities Across different Words Types

Figure 24 shows the percentage of cases achieving a higher sum of intensity in the expected region for each word type. The x-axis represents the different word types, and the y-axis shows the percentage. The blue bars represent the results for the ViT model, with lighter blue for curved regions and darker blue for jagged regions. The green bars show the ResNet results, with lighter green for curved regions and darker green for jagged regions.

For adjectives, the expected pattern appears for both curved and jagged labels, with percentages above 50%. In other word types, the percentage of correct matches for curved words is higher than that for jagged words and remains above 50%. This could indicate a bias toward curved regions, which may receive more attention than jagged regions.

To verify whether the attention pattern results are consistent, we checked each pair of combined images (one with the curved shape on the left, the jagged shape on the right, and the other with the jagged shape on the left and the curved shape on the right). Consistency is defined as having a higher sum of intensity in the same region (curved or jagged) across both configurations. If this condition is not met, it indicates inconsistency. Figure 25 shows the percentage of consistency results for the ViT model (blue bars) and the ResNet model (green bars). The x-axis shows the consistent (true) and inconsistent (false) cases. The y-axis represents the percentage of cases. The results show that while the ViT model shows relatively high consistency, there is a noticeable bias in the attention pattern results, particularly for the ResNet model.

Table 6 shows the details of the attention pattern result for each image and its swapped version. In this table, the first column indicates whether the attention pattern remains consistent across both versions of combined images. The second column shows the region with a higher sum of intensity when the curved region is placed on the left. The third column indicates the dominant region when the jagged region is on the left, and the last column provides the number of occurrences for each case. We observe a clear regional bias when examining the inconsistent cases (first two rows). More specifically, in the first 2 rows, the sum of intensity is higher in the left region. This bias highlights a strong preference for the left region. Calculating the proportions, 77.42% of the inconsistent cases show left-side, and only 22.58% show right-side dominance. This significant difference indicates a regional bias, where the left region consistently receives more attention than the right.

A different pattern can be seen for the consistent cases (last two rows). In these 2 rows, curved regions attract more attention overall compared to jagged regions, with 56.53% of the occurrences showing curved image dominance and 43.47% showing jagged image dominance. Although curved regions dominate slightly, the difference is less pronounced than the regional bias observed in the inconsistent cases. However, this slight preference for curved regions could explain the higher percentages for curved images in figure 24 across different word types.

Table 6: Consistency results for attention patterns

Consistent	Dominant Region(Curved_left)	Dominant Region(Jagged_left)	Count
False	left	left	5862
False	right	right	1710
True	left	right	12075
True	right	left	9287

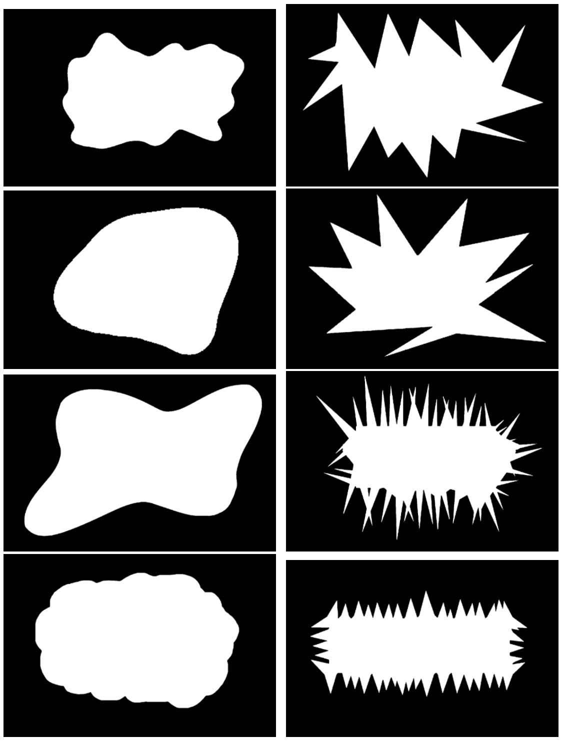


Figure 17: Images from [Westbury, 2005]

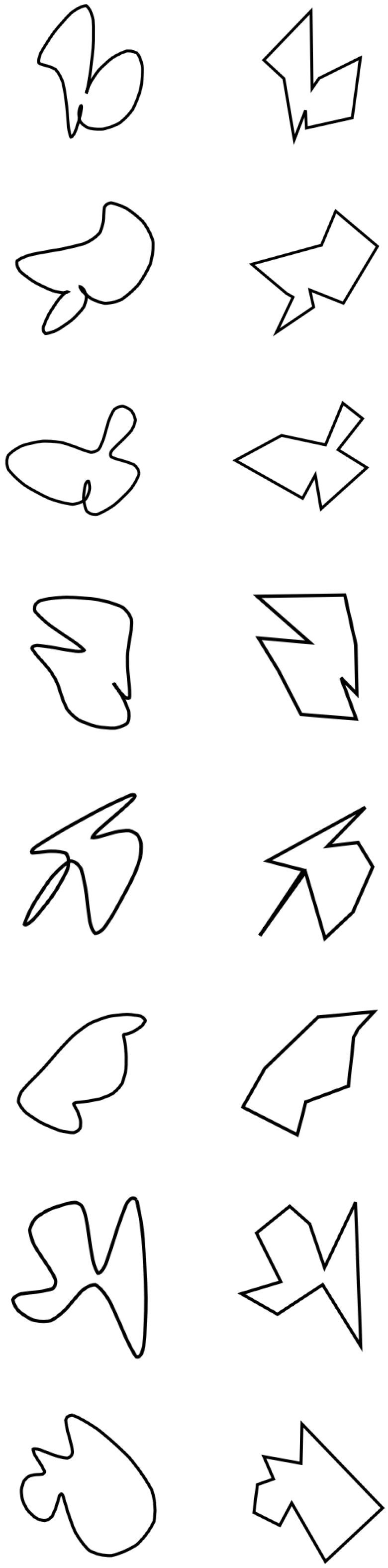
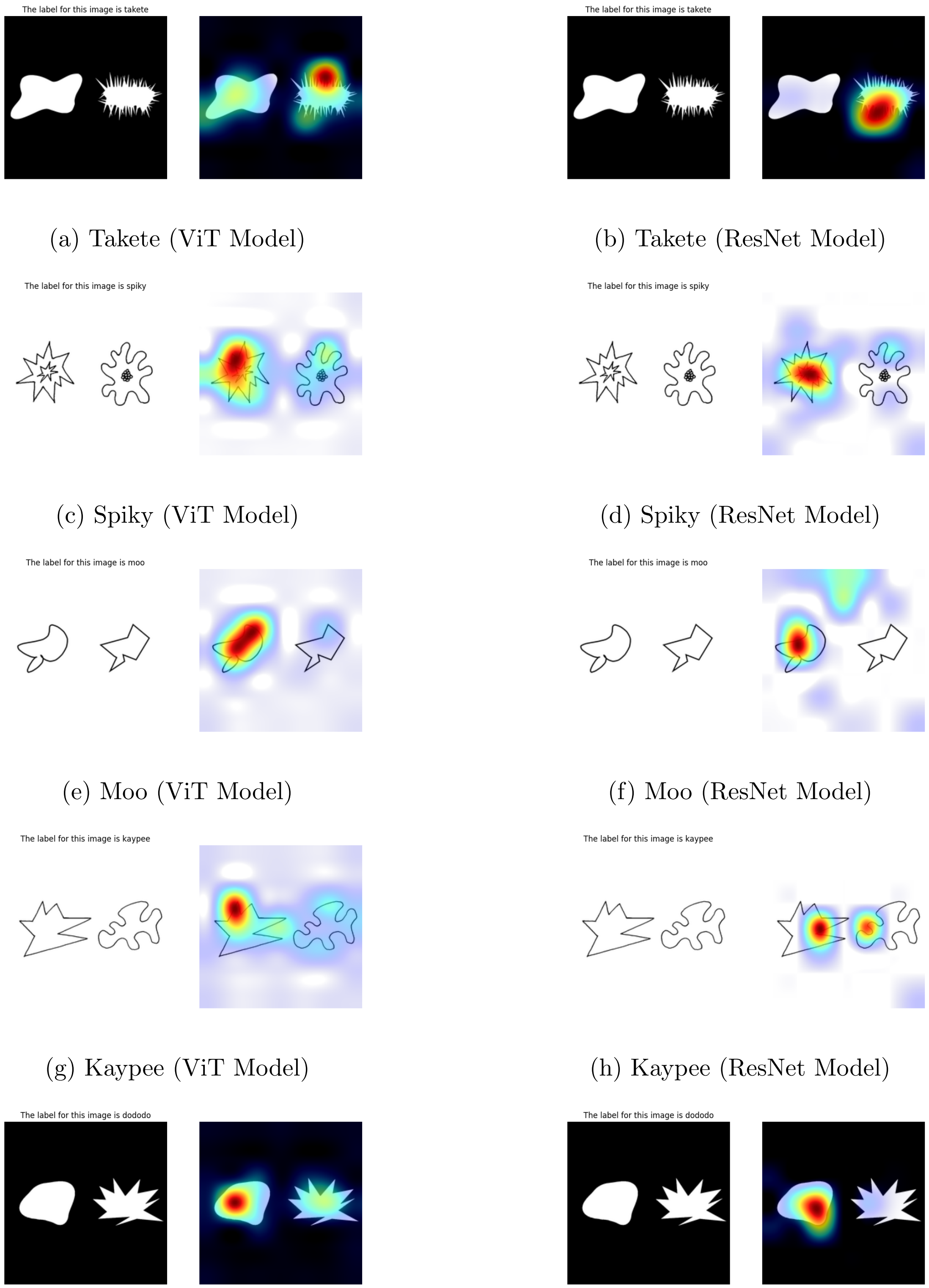


Figure 18: Generated images in this study



(a) Takete (ViT Model)

(b) Takete (ResNet Model)

(c) Spiky (ViT Model)

(d) Spiky (ResNet Model)

(e) Moo (ViT Model)

(f) Moo (ResNet Model)

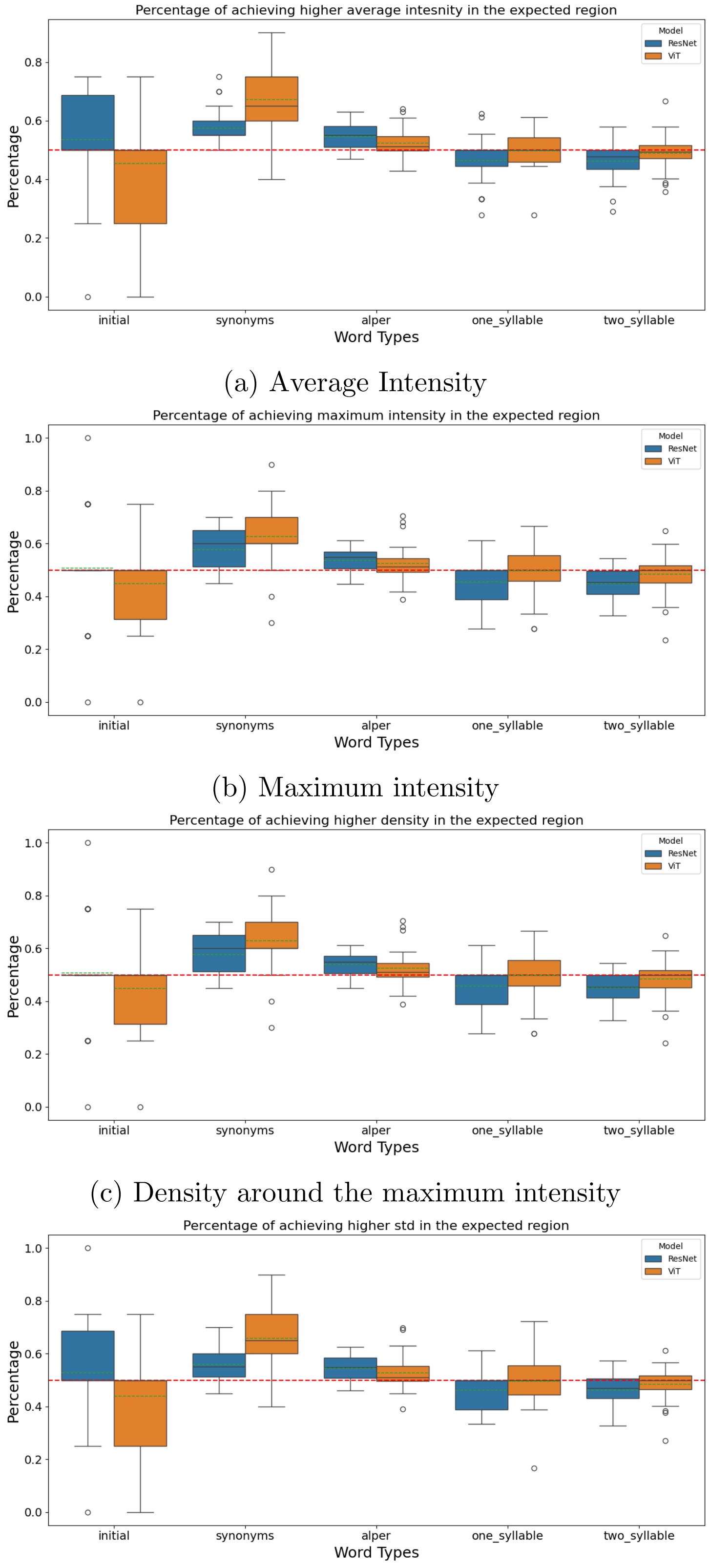
(g) Kaypee (ViT Model)

(h) Kaypee (ResNet Model)

(i) Dododo (ViT Model)

(j) Dododo (ResNet Model)

Figure 19: Visualizing the attention patterns for different text prompts. Each pair of subfigures shows the input image with the text prompt and the resulting attention pattern for ViT and ResNet models.



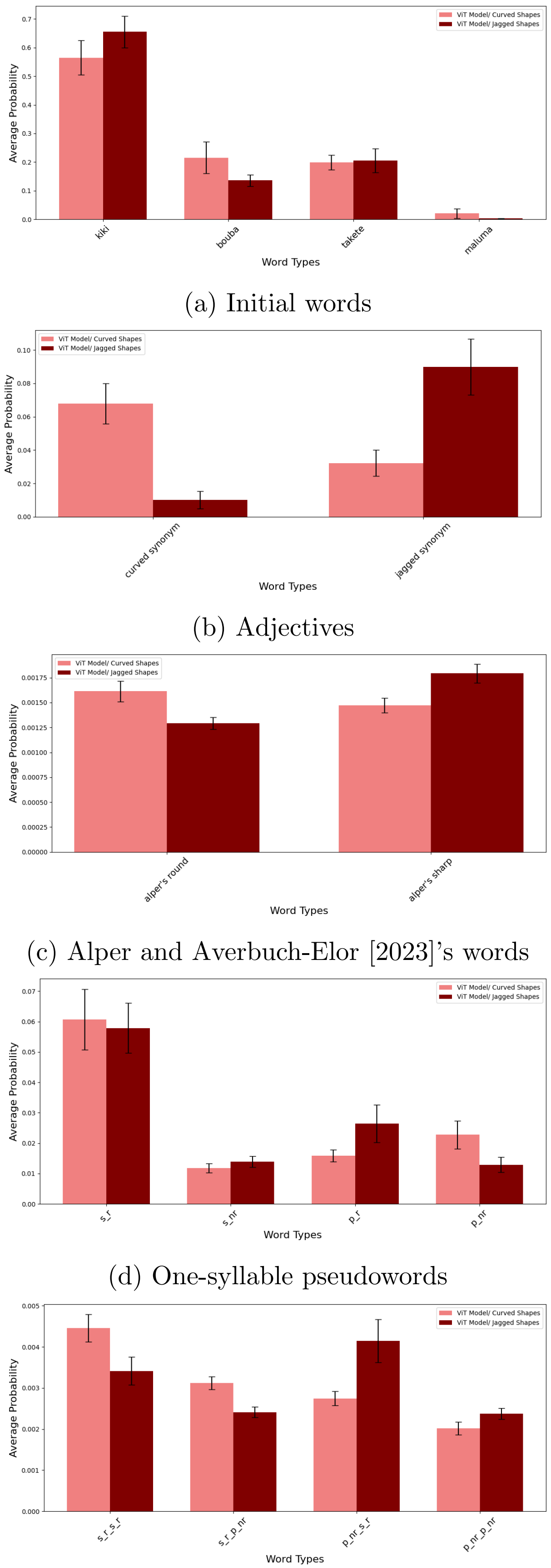
(a) Average Intensity

(b) Maximum intensity

(c) Density around the maximum intensity

(d) Standrad Deviation

Figure 20: Performance of ResNet- and ViT-based CLIP models across different word types using various attention metrics (average intensity, maximum intensity, peak attention density, and standard deviation). The boxplots show the percentage of cases where higher values are observed in the expected region (curved or jagged) for each label. The red dashed line represents the 50% baseline, and the orange and blue boxes correspond to ViT and ResNet models, respectively.



(a) Initial words

(b) Adjectives

(c) Alper and Averbuch-Elor [2023]'s words

(d) One-syllable pseudowords

(e) Two-syllable pseudowords

Figure 21: Average probability results using the specific version of CLIP model used in The study by Alper and Averbuch-Elor [2023].

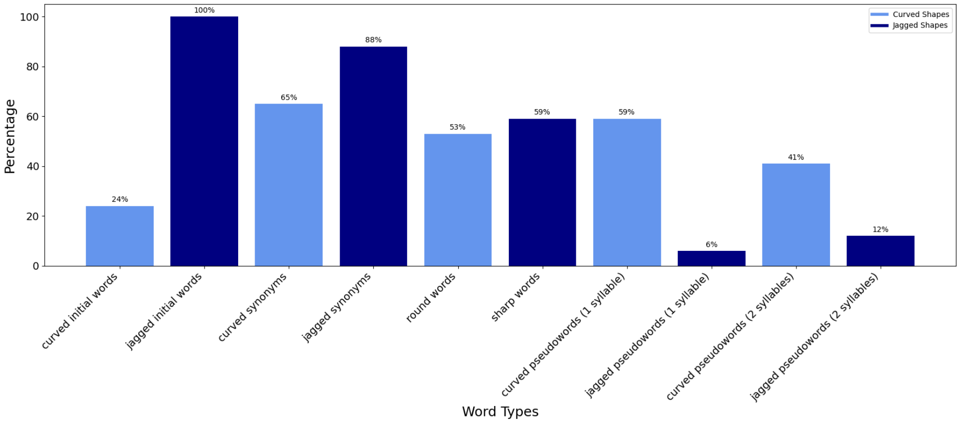


Figure 22: Percentage of selecting the expected label for each image across different word types for the specific CLIP model used in Alper and Averbuch-Elor [2023]

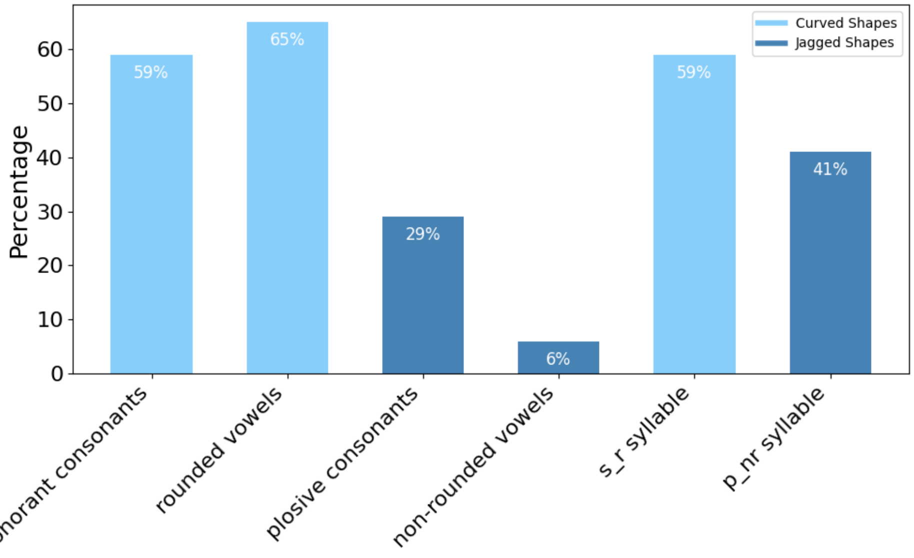


Figure 23: Percentage of selecting the expected label for each image based on consonants and vowels for the specific CLIP model used in Alper and Averbuch-Elor [2023]

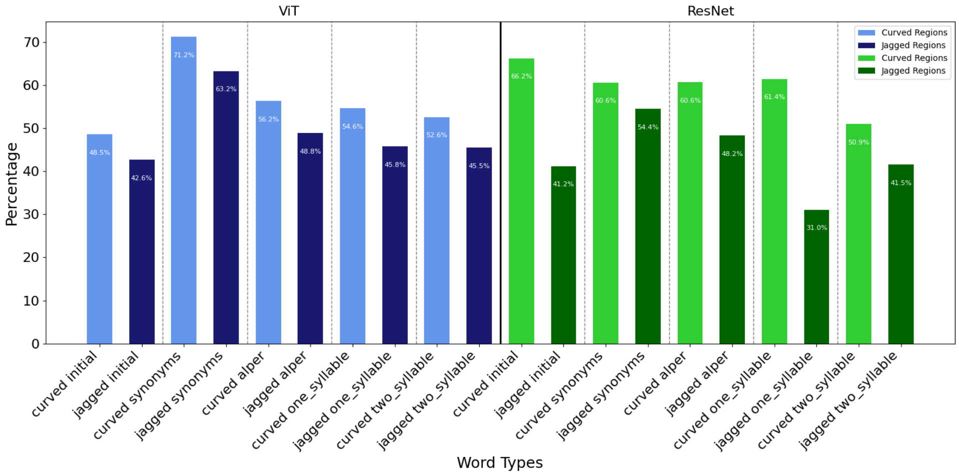


Figure 24: Percentage of achieving higher sum of intensity in the expected region across different word types

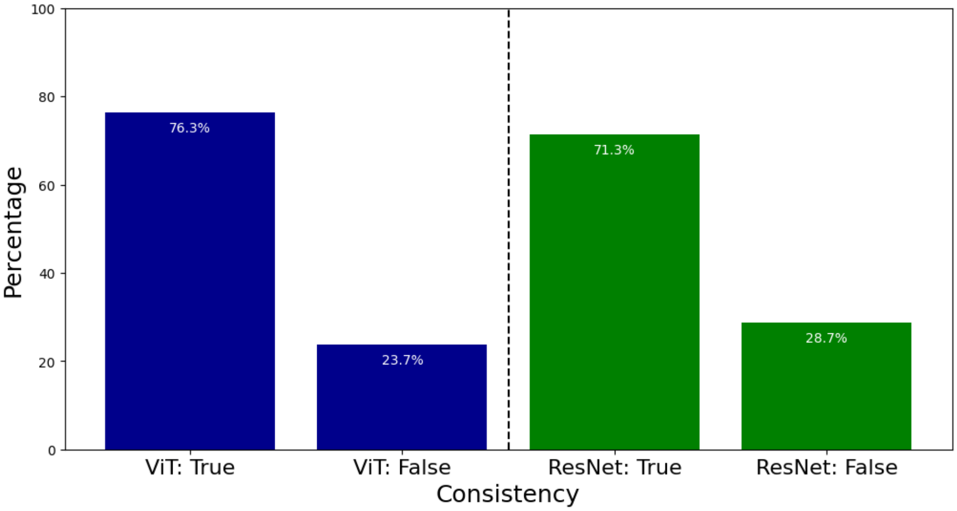


Figure 25: Consistency in attention pattern results