Cross-modal Associations in Vision and Language Models: Revisiting the Bouba-Kiki Effect

Tom Kouwenhoven*, **Kiana Shahrasbi, Tessa Verhoef***
Leiden Institute of Advanced Computer Science
Leiden University, The Netherlands
{t.kouwenhoven, k.shahrasbi, t.verhoef}@liacs.leidenuniv.nl

Abstract

Recent advances in multimodal models have raised questions about whether vision-and-language models (VLMs) integrate cross-modal information in ways that reflect human cognition. One well-studied test case in this domain is the bouba-kiki effect, where humans reliably associate pseudowords like 'bouba' with round shapes and 'kiki' with jagged ones. Given the mixed evidence found in prior studies for this effect in VLMs, we present a comprehensive re-evaluation focused on two variants of CLIP, ResNet and Vision Transformer (ViT), given their centrality in many state-of-the-art VLMs. We apply two complementary methods closely modelled after human experiments: a prompt-based evaluation that uses probabilities as a measure of model preference, and we use Grad-CAM as a novel approach to interpret visual attention in shape-word matching tasks. Our findings show that these model variants do not consistently exhibit the bouba-kiki effect. While ResNet shows a preference for round shapes, overall performance across both model variants lacks the expected associations. Moreover, direct comparison with prior human data on the same task shows that the models' responses fall markedly short of the robust, modality-integrated behaviour characteristic of human cognition. These results contribute to the ongoing debate about the extent to which VLMs truly understand cross-modal concepts, highlighting limitations in their internal representations and alignment with human intuitions.

1 Introduction

Recent advances in multimodal models that integrate vision and language have brought artificial intelligence a step closer to understanding the world in ways that resemble human experience and cognition. These models, which learn from vast amounts of paired visual and textual data, have demonstrated impressive capabilities in tasks such as image captioning, visual question answering, and cross-modal retrieval (Radford et al., 2021; Li et al., 2022, 2023). Yet it remains unclear whether these models integrate visual and linguistic information in ways that parallel human cognitive processes. In this paper, we investigate whether VLMs exhibit human-like patterns of association between abstract visual shapes and unfamiliar words. Particularly, we focus on one of the most widely studied cross-modal test cases in human cognition, the bouba-kiki effect, in which people consistently associate pseudowords like 'bouba' with round shapes and those like 'kiki' with jagged shapes (Ramachandran and Hubbard, 2001; Maurer et al., 2006a; Ćwiek et al., 2022). These associations have recently become a test case for evaluating whether language models trained on large-scale data show similar patterns (Alper and Averbuch-Elor, 2023; Verhoef et al., 2024; Loakman et al., 2024; Iida and Funakura, 2024). Results, however, have been mixed. While Alper and Averbuch-Elor (2023) found overwhelming evidence for a bouba-kiki effect in CLIP and Stable Diffusion, other

*Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

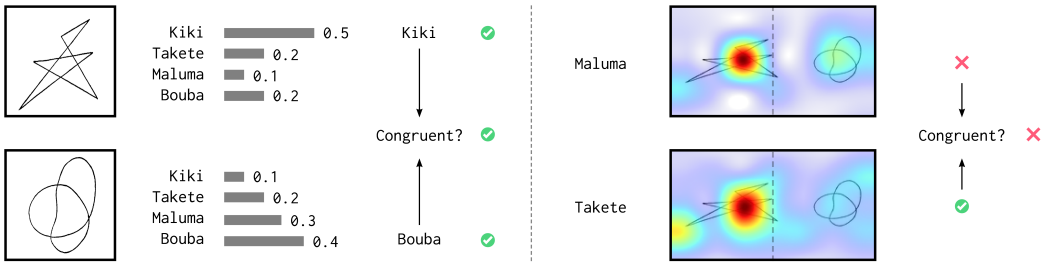

Figure 1: An overview of the two complementary methods used. On the left, we calculate the probabilities for each label across the four original pseudowords (note that the number of labels varies per label source) for each image shape and select the label with the highest probability (values are exemplary). On the right, we use concatenated image pairs and their labels as targets to calculate attention patterns with Grad-CAM and select the shape with the highest sum of attention.

studies have found less convincing patterns across different VLMs, testing methods and datasets (Verhoef et al., 2024; Loakman et al., 2024; Iida and Funakura, 2024).

This work therefore revisits this effect by thoroughly testing two versions of CLIP (Radford et al., 2021): ResNet (He et al., 2016) and ViT (Dosovitskiy et al., 2021). We focus on CLIP because, among four models (CLIP, BLIP2, ViLT, and GPT-4o) tested, Verhoef et al. (2024) found that CLIP demonstrated the most promising alignment with a human-like bouba-kiki effect. This aligns with previous work demonstrating that CLIP outperforms other models in capturing human-like decision patterns (Demircan et al., 2024). Perhaps most importantly, CLIP often acts as a foundational model in state-of-the-art VLMs (see section 3). If a 'base' model does not exhibit human-like associations, it is difficult to imagine that a model using that base model as a backbone will show human-like preferences without explicit fine-tuning on relevant cross-modal associations. Unravelling how these 'base' models represent cross-modal information additionally benefits our understanding of limitations in, for example, spatial reasoning (Thrush et al., 2022), (in-context) shape-colour biases in VLMs (Allen et al., 2025), and emergent communication setups (Kouwenhoven et al., 2024).

By probing these models' preferences for shape-word association in novel contexts, we aim to shed light on the nature of their internal representations and their alignment with human intuitions. When these differ, training models to develop human-like prior expectations or instilling them into models may help and could even improve learning efficiency (Lake et al., 2017). Doing so is essential, since a flexible, grounded understanding of this kind could enable conceptually fluent, natural human-machine interactions, in which machines intuitively understand what we mean, even in unfamiliar settings. To this end, we employ two complementary methodological approaches that address the same cross-modal associations from different perspectives, thereby minimising the limitations of relying on a single evaluative framework. Despite this comprehensive strategy, we do not find compelling evidence for the bouba-kiki effect. This work contributes in the following ways:

(1) We test for the bouba-kiki effect in VLMs in ways that are as close as possible to how it has been tested with humans in the past. This enables direct comparisons between our results and human findings, building on and expanding the work by Verhoef et al. (2024). Tested comprehensively, we find that models, compared to humans, do not make consistent cross-modal associations.

(2) We introduce a novel way to test models using Grad-CAM (Selvaraju et al., 2020), a method from the model interpretability literature, to look more closely at visual processing in the bouba-kiki context. Strengthening the robustness of our results, it reveals that the models do not explicitly focus on bouba-kiki related shape-specific features.

2 Related work

First, we discuss how prior work on cross-modal associations in human language and cognitive processing has shown that non-arbitrariness is both pervasive and affects how we learn, process and develop language. We then present previous studies that have tested for the bouba-kiki effect in VLMs.

2.1 Cross-modal associations in human language

Traditionally, it has been argued that mappings between words and their meanings are largely arbitrary. Hockett (1960) uses the words 'whale' and 'microorganism' as an example: 'whale' is a short word for a large animal, while 'microorganism' is the reverse. However, growing evidence from fields like cognitive science, language evolution and (sign language) linguistics suggests non-arbitrary form-meaning mappings are more widespread in human language than initially thought. This suggests that it should be considered a general property of human language (Perniss et al., 2010), shaped by cognitive mechanisms similar to those involved in the bouba-kiki effect. Especially when looking beyond Indo-European languages, 'iconic' mappings, where word forms seem to resemble their meanings, appear to play a significant role in many languages (Perniss et al., 2010; Dingemanse, 2012; Imai and Kita, 2014). Some languages have specific word classes where characteristics of the meaning are mimicked or iconically represented in the word. For example, Japanese ideophones allow speakers to depict sensory information through word forms, such as saying *nuru nuru* when describing something as 'slimy' and *fuwa fuwa* when it is 'fluffy' (Dingemanse et al., 2016). Similarly, in Siwu, *pimbilii* means 'small belly', while *pumbuluu* refers to 'fat belly', and non-speakers of those languages can typically guess the meanings of such words (Dingemanse et al., 2016). Even in languages not considered rich in sound symbolism, such as English and Spanish, vocabulary items from specific lexical categories, such as adjectives, are rated relatively high in iconicity (Perry et al., 2015).

Evidence for strong associations between speech sounds and particular meanings has been found in a broad sample of vocabularies from two-thirds of the world's languages Blasi et al. (2016). Iconic mappings are not only widespread in the world's languages, they also help both children Imai et al. (2008); Perry et al. (2015, 2018) and adults Nielsen and Rendall (2012) learn new words more easily. They, moreover, allow us to communicate successfully even when a shared language is absent or existing vocabulary is insufficient, because cross-modal associations lay the groundwork for the negotiation of novel words and their meanings (Ramachandran and Hubbard, 2001; Cuskley and Kirby, 2013; Imai and Kita, 2014). Shared cross-modal representations enable people to instantly interpret unfamiliar words even on first exposure by directly linking sensory experiences with meaning. Indeed, experiments with humans that studied the influence of cross-modal preferences on the emergence of novel vocabularies show that iconic strategies are frequently adopted when word forms and meanings can be intuitively mapped, and they help to communicate successfully (Verhoef et al., 2015, 2016b,a; Tamariz et al., 2018). An understanding of such mappings, however, requires human-like integration of multi-sensory information or an awareness of common cross-modal associations. While multimodal computational models have demonstrated remarkable capabilities across a range of tasks, the internal mechanisms through which these systems form and link representations remain opaque. In particular, it is still an open question whether these models integrate visual and linguistic information in ways that mirror human cognitive processes.

2.2 Bouba-kiki effect in VLMs

The bouba-kiki effect, as a specific example of visuo-linguistic processing, has been studied in VLMs before; however, the results so far seem to be conflicting. One key study in this space is the innovative work by Alper and Averbuch-Elor (2023), who convincingly demonstrated that patterns aligning with the bouba-kiki effect are reflected in the embedding spaces of CLIP and Stable Diffusion. They used Stable Diffusion to generate images based on pseudowords that were carefully designed to reflect phonetic properties associated with sharp or round shapes. Specifically, they used CLIP's text encoder to embed prompts containing either pseudowords or descriptive adjectives, while the images generated from pseudoword-based prompts were passed through CLIP's vision encoder. This setup allowed both text and image representations to be analysed within the same multimodal embedding space. The embedding similarity between the pseudowords and the adjective or image representations was then used to compute geometric and phonetic scores, indicating alignment with sharp or round associations. They concluded that their findings indicate strong evidence for the existence of cross-modal associations in VLMs. Typical explanations given for the presence of a bouba-kiki effect in humans include experience with acoustics and articulation (Ramachandran and Hubbard, 2001; Maurer et al., 2006b; Westbury, 2005a), affective–semantic properties of human and non-human vocal communication (Nielsen and Rendall, 2011), or physical properties relating to audiovisual regularities in the environment (Fort and Schwartz, 2022). This renders the conclusion by Alper and Averbuch-Elor (2023) somewhat surprising, as these explanations all involve situated, real-

world experience with a body and the environment, something these models entirely lack. Moreover, limitations of VLMs in visual grounding have been observed in many other domains (Thrush et al., 2022; Diwan et al., 2022; Kamath et al., 2023; Jabri et al., 2016; Goyal et al., 2017; Agrawal et al., 2018; Jones et al., 2024), suggesting that these models do not integrate textual and visual data in a manner that is human-like. Nonetheless, this work presented an innovative method for testing the bouba-kiki effect in VLMs. It convincingly demonstrated that VLMs encode relationships between word forms and semantic concepts related to roundness and jaggedness. As described previously, these associations are indeed abundantly present in human languages, and prior work with text-only language models has also shown that these models can detect such regularities (Abramova and Fernández, 2016; Pimentel et al., 2019; de Varda and Strapparava, 2022; Marklová et al., 2025).

Interestingly though, Iida and Funakura (2024) replicated Alper and Averbuch-Elor's experiments for Japanese and found that Japanese VLMs did not exhibit the expected bouba-kiki effect, even though Japanese is a language rich in sound-symbolism (Dingemanse, 2012), and speakers of Japanese display the strongest bouba-kiki effect compared to speakers of 25 other languages in a study by Ćwiek et al. (2022). These findings suggest that Alper and Averbuch-Elor's method may not capture true sensory mappings, but instead detects regularities between word forms and meanings that are language-specific rather than universal. For example, sounds like */p/*, which are typically linked to sharpness according to the bouba-kiki effect, are strongly associated with roundness in ideophonic Japanese words, like *pocha-pocha* 'chubby', *puyo-puyo* 'fat', and *puku-puku* 'puffing up' (Iida and Funakura, 2024). This suggests that the method used to disambiguate sharp and round pseudowords and images may pick up on relationships between semantic concepts and word forms—being heavily entangled with the choice of ground-truth adjectives—rather than capturing true sensory mappings in languages. Moreover, the vectors used to assign a geometric or phonetic score to a pseudoword or image must be sufficiently dissimilar, for which Iida and Funakura (2024) reported that this was not the case. The contradictory findings between Japanese and English VLMs highlight the need for a different approach that aligns with human experiments on cross-modal associations.

A key ingredient of human bouba-kiki experiments is that tests are centred around specifically designed pairs of visual images that minimally differ, but with one more rounded and one more jagged version, as shown in Figure 1. For example, Maurer et al. (2006a) presented a pair of jagged/round images along with a pair of bouba/kiki-like words and participants were asked to pair them up in the most fitting way. Other studies employed more stringent tests, as in Ćwiek et al. (2022), where only the images were presented side by side, and participants had to choose the best-fitting image after listening to only one of the spoken words at a time. Finally, Nielsen and Rendall (2013) presented single images and asked participants to generate novel pseudowords to match the shapes. In all of these cases, humans exhibit a strong bouba-kiki effect. Inspired by these studies, Verhoef et al. (2024) used images from existing human experiments and explored four prominent VLMs on carefully designed image-to-word matching tasks. They directly used model probabilities generated by VLMs to match specific pseudowords with images as a measure of preference and found limited evidence for the bouba-kiki effect. Two out of four models (CLIP and GPT-4o) exhibit moderate alignment with human-like associations, but only in some of the tests they conducted, and not consistently. The study concludes that cross-modal associations in VLMs are highly dependent on factors such as model architecture, training data, and the specific test used.

Others have also investigated the bouba-kiki effect and other cross-modal associations, such as a relation between perceived size and vowels (Loakman et al., 2024) and understanding shitsukan terms (a Japanese concept that captures the sensory essence of an object) (Shiono et al., 2025). However, none of these consistently find a resemblance between the human and model associations. Tseng et al. (2024) used the embedding method from Alper and Averbuch-Elor to test sensitivity to sound-symbolic associations in audio-visual models and report that these models capture sound-meaning connections akin to human language processing. Yet, this method again seems to rely on the relationships between semantic concepts and word forms, as was mentioned before. Given the conflicting evidence in this domain, we revisit the bouba-kiki effect through a thorough investigation using two versions of CLIP and a wide variety of prompts. This approach introduces a novel method for testing it using visual interpretability methods and directly compares the results with similar findings from human studies.

3 Methodology

Cross-modal associations are assessed by prompting two different versions of CLIP (Radford et al., 2021), specifically, a ResNet-50 and a ViT version. We deem this as reasonable since state-of-the-art VLMs such as Molmo (Deitke et al., 2024), LLaVA (Liu et al., 2023), BLIP2 (Li et al., 2023), and InternVL (Chen et al., 2024) commonly use pre-trained frozen vision models and train lightweight alignment layers to align visual features with existing linguistic embeddings present in (large) language models (Liu et al., 2024). The frozen ViT version of CLIP is particularly often the basis for many current state-of-the-art VLMs (BLIP2, Molmo, InternVL, LLaVa, i.a.). If CLIP, as a backbone, does not consistently exhibit human-like cross-modal associations, it is difficult to imagine how additional alignment layers can capture human-like representations without extensive fine-tuning. Moreover, CLIP enables us to extend existing approaches with an interpretability-based methodology that cannot be applied to proprietary models.[2]

Linguistic inputs Models are probed through combining images and labels, the latter of which are pseudowords originating from four different sources. First, two sets of 'original' labels are used, which have been traditionally used the most in human studies (*bouba-kiki* and *maluma-takete*), allowing for explicit comparison of human results with those of VLMs. Second, we borrow English adjectives from Alper and Averbuch-Elor (2023) that are 10 synonyms of 'sharp' and 'curved'. These serve as a baseline informing us whether the models can, in principle, make correct cross-modal associations. This differs from Alper and Averbuch-Elor (2023), who use adjectives to create a vector to calculate a geometric score. Third, two-syllable labels were constructed following Nielsen and Rendall (2013), using previously established cross-modal patterns in English. Sonorant consonants (M, N, L) and rounded vowels (OO, OH, AH) tend to match curved shapes, while plosive consonants (T, K, P) and non-rounded vowels (EE, AY, UH) align with jagged shapes. Combining these yields 36 syllables (e.g., loo, nah, kee, puh), categorised into four types: sonorant-rounded (S-R), plosive-rounded (P-R), sonorant-non-rounded (S-NR), and plosive-non-rounded (P-NR). These were paired into two-syllable pseudowords, loosely replicating the task humans did in Nielsen and Rendall (2013). For analysis, we focus on 'pure' pseudowords that are either fully round-associated (S-R-S-R) or fully sharp-associated (P-NR-P-NR), yielding 162 labels. Fourth, VLMs are tested on pseudowords generated using the method from Alper and Averbuch-Elor (2023). Each pseudoword follows a three-syllable structure, where consonants (P, T, K, S, H, X, B, D, G, M, N, L) are combined with vowels (E, I, O, U, A).[3] Pseudowords repeat the first syllable at the end (e.g., 'kitaki', 'bodubo'). Only 'pure' items—composed entirely of syllables from a single class—are used, excluding mixed forms like 'kiduki'.

The linguistic inputs are comprised of a label of interest—which should induce certain preferences—and a prompt ('The label for this image is <label>') such that embedding the label happens at the sentence level and is closer to the models' natural objective. Importantly, for each image, we *only* differ the pseudowords in question, so variation in probability must be a result of the pseudoword. Provided that the pseudowords are generated anew, it is unlikely that the models encountered them during training. Ten different prompts are used (see Appendix A) to ensure that the results are robust and not an artefact of peculiarities in the prompts. These prompts are sourced from Verhoef et al. (2024), Alper and Averbuch-Elor (2023), or are newly created such that half of the labels appear as a noun, and the other half as an adjective.

Visual inputs The images fed to the models are either curved or jagged. Some of these images are sourced directly from previous works involving human participants (Köhler, 1929, 1947; Maurer et al., 2006a; Westbury, 2005b), while others are specifically generated to test cross-modal associations in VLMs. These were inspired by the method described in Nielsen and Rendall (2013) and already used by Verhoef et al. (2024). The generated images were created by randomly distributing points within a circle and then connecting them sequentially using curved lines for curved images and straight line segments for jagged images. Hence, this method generated image pairs that subtly differ *only* in features that are seemingly important for cross-modal effects. This differs from Alper and Averbuch-Elor (2023), who generated images using Stable Diffusion, leaving less experimental

[2]The source data and code are available at `https://osf.io/gqsv6/`

[3]The letter A is included per Alper and Averbuch-Elor (2023) despite that it is typically not regarded to evoke cross-modal associations in humans.

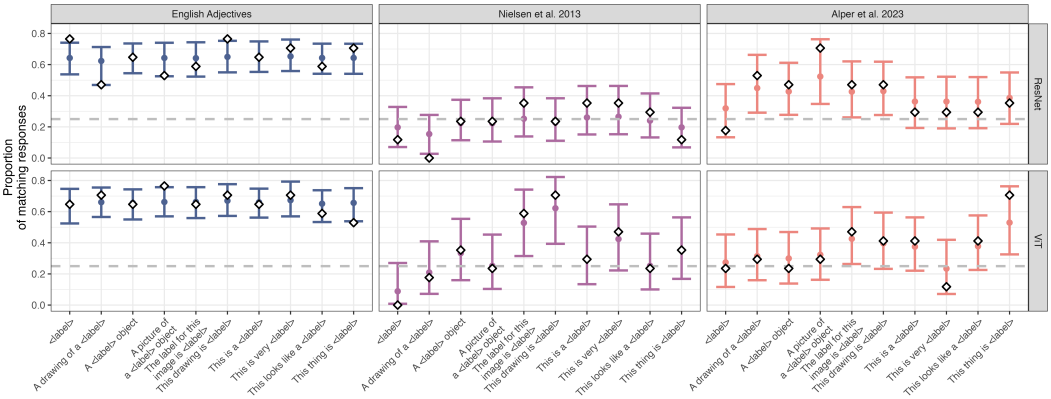

Figure 2: The proportion of congruent responses for matching both images of an image pair correctly ($Match = 1$). A result in which models consistently match images above chance (the grey dashed line) across prompts would suggest the presence of cross-modal associations. This is only the case for the English adjectives, which function as a baseline. Model: $Match \sim 1 + Word\_type + (1 + Word\_type|Prompt)$. Diamonds are descriptive means, and the dots are posterior means.

control and potentially distorting the understanding of which features cause the observed effects. An overview of all image pairs is displayed in Appendix B.

Analyses All our analyses use Bayesian Regression Models as implemented in the *brms* package (Bürkner, 2021) in R (R Core Team, 2024). We fit models (using 4 chains of 4000 iterations and a warm-up of 2000) to predict the proportion of correct guesses given a *Word_type* with fixed effects for *model*, *prompt*, *image pair*, or *label pair*. The exact model formulas are displayed under each figure. The interpretation of our visualisation is straightforward: an effect is significant if the posterior means *and* their credible intervals are above the chance level.

4 Probing through probabilities

To assess the preferences of the ViT and ResNet versions, we extract probabilities for all possible labels (i.e., syllables and pseudowords) conditioned on an image. For each image, we consider the label with the highest probability to be the model's preference (shown on the left in Figure 1). This is comparable to how Nielsen and Rendall (2013) tested for human preferences. While using the probability data for all labels would have been possible, previous work demonstrated that using only the best-matching label yielded a small bouba-kiki effect, whereas using all probabilities did not (Verhoef et al., 2024); as such, we use the most promising method. Importantly, we follow Ćwiek et al. (2022), who rightfully treated human responses as bouba/kiki-congruent when participants matched bouba to a round shape, and crucially, also matched kiki to a spiky shape. In our work, this means that an image pair of a sharp *and* round shape must be matched with a congruent label.

Results We begin by considering the English adjectives, which serve as a baseline and do not inform us about the bouba-kiki effect. Figure 2 reveals that these are successfully matched to images with round or sharp features. This result is consistent across prompts and is also somewhat expected. Yet, a performance of roughly 80% across prompts and models could also be considered low given the clear-cut distinctions between images and adjectives. The primary labels of interest (i.e., the pseudowords) can not be robustly and correctly matched to images with corresponding features by either model version. Despite some variability across prompts and models, with some combinations leading to above-chance performance (e.g., ViT, Nielsen and Rendall's pseudowords, and *This drawing is <label>* or ResNet, Alper and Averbuch-Elor's pseudowords, and *A picture of a <label> object*), the descriptive means are mostly at chance level. Hence, we conclude that neither model displays clear cross-modal preferences. Interestingly, the individual modalities are, in principle, separable by both models (A.1 and B.1), suggesting that the difficulty lies in combining information from multiple modalities. This is also visible when looking into which labels are most commonly chosen, revealing that model predictions do not differ much even though they are presented

with different images (Appendix C). For the experimental pseudowords, models seem to rely on a few frequently selected labels (the percentage of unique labels for ResNet and ViT is $\approx 35\%$). This strengthens our observation that no robust syllable-level cross-modal associations influence predictions.

5 Beyond behavioural observations

In addition to using probability data, we are interested in knowing whether the predictions are wrong for the right reasons, i.e., do they fail to attend to curved or jagged regions when presented with our images? As such, we further analyse the behaviour and preferences of models when associating an image with a label, using Grad-CAM, a technique from the model interpretability literature (Selvaraju et al., 2020). Grad-CAM offers a visual explanation of a model's prediction by calculating the gradients of the target class score. This involves using the cosine similarity between the label and image embeddings in relation to the feature maps from the final convolutional layer or the last attention block. In the case of ViT, we applied Grad-CAM to the last attention layer, which retains spatial information via its attention heads. We specifically focused on the attention from the [CLS] token to the image patches, similar to the approach in Caron et al. (2021) and following public implementations of Grad-CAM (Zakka, 2021; Mamooler, 2021; Chefer et al., 2021). Specifically, to identify which image regions contributed most to the decision, we computed the gradients of the target score with respect to the attention weights. These are averaged across heads, and we then extract the attention weights from the [CLS] token to the image patches by removing the [CLS] column. This is reshaped into a 2D spatial grid that acts as a feature map. The values in these feature maps (typically displayed as a heatmap) can be interpreted as the importance or contribution of a specific region in the image to the model's prediction (see Appendix D for some visual examples).

This technique enables us to very closely mimic the cross-cultural study conducted by Ćwiek et al. (2022) in which human participants were shown the two classical bouba-kiki shapes and listened to the spoken words bouba or kiki. Hereafter, they selected which of the two shapes they thought corresponded to the word. To simulate this experiment with VLMs, we concatenate all image pairs into a single image containing a curved and jagged shape (n=17). We then identify, for each label—which defines the expected target—which image region (left or right) receives the most attention from a model using Grad-CAM (see Figure 1). While the generated heatmaps visually indicate attention patterns, we quantify attention by summing the attention allocated to each curved and jagged part of the image. Taking the total attention allocated to each part, rather than focusing on very specific regions, aligns with how humans perceive faces, objects, and words holistically (Taubert et al., 2011; Zhao et al., 2016; Wong et al., 2011). Nevertheless, we also experimented with entropy and centroid-of-attention as quantifying measures, but none of these changed the results meaningfully (Appendix D. We compare the sum of importance values in the expected region for a given pseudoword within a particular category with the sum of importance values in the non-expected region for each image and text label. Based on this comparison, we compute the percentage of trials in which the model consistently focuses more on the expected region than on the non-expected region (as shown on the right in Figure 1). The way images are concatenated is balanced such that the target appears eight or nine times on the left or right. Doing this eliminates a potential model bias toward objects on the left or right. Initial analyses across models, prompts, labels, and images revealed that both models are relatively consistent in their predictions, regardless of the target location, with ResNet being consistent for 77.0% of the predictions and the ViT variant 73.5% (Appendix D).

Comparing VLMs with humans Using Grad-CAM, we compare the performance of both models to that of English-speaking participants as reported by Ćwiek et al. (2022). While they only investigated the well-known label pair bouba-kiki, another pair, maluma-takete, combined with the two images shown in Figure 1, was first described by Wolfgang Köhler (Köhler, 1929, 1947) and sparked wider interest in cross-modal associations. We use only these two label pairs here, since the larger pseudoword set does not contain paired labels. However, our image generation process, which only varies how points are connected, allows evaluating across all image pairs, beyond just the original images. Specifically, we evaluate whether models attend more to the expected region (e.g., the sharp shape for 'kiki' and 'takete' or the round shape for 'bouba' and 'maluma'). A prediction is considered congruent only if both labels are correctly matched to their corresponding shapes. This setup closely mirrors human experiments and, for the bouba-kiki label pair, enables direct comparisons.

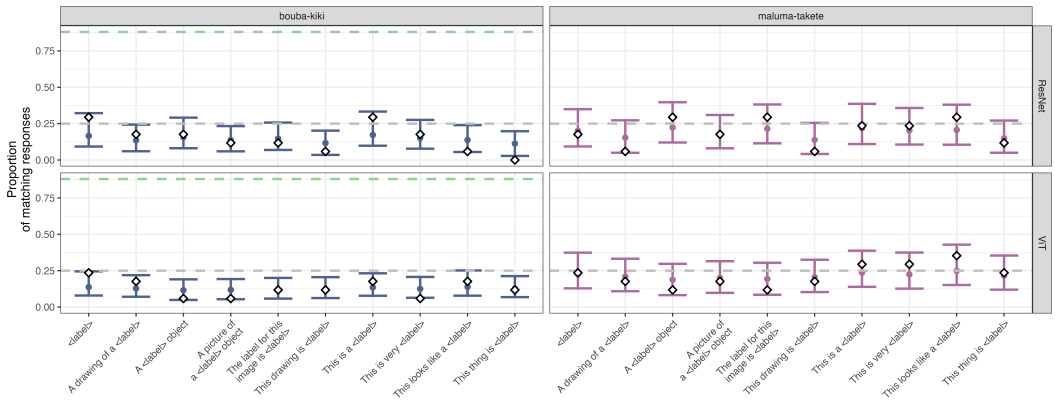

Figure 3: The proportion of correctly matched responses for both labels of a label pair ($Correct = 1$) given an image pair using Grad-CAM. The green line indicates human 'performance' reported in (Ćwiek et al., 2022). The grey line shows the chance level (25%) as the model must map 'bouba' *and* 'kiki' correctly. Model: $Correct \sim 1 + LabelPair + (1 + LabelPair|Prompt)$. Diamonds are descriptive means, and the dots are posterior means.

Strikingly, neither CLIP model consistently maps both labels to their intended shapes at a human-like level (Figure 3); in fact, performance does not reliably exceed chance. These results contradict previous claims that vision–language models exhibit human-like cross-modal associations (Alper and Averbuch-Elor, 2023), and instead reinforce earlier findings showing no such effect. If there is any situation in which an effect would be expected, it would be here since these classic word pairs are most dominantly present in the data. Yet, these results show that merely learning *about* an existing cross-modal effect from data distributions is different from having a mechanistic preference for cross-modal associations, which should not come as a surprise.

Analysing pseudowords Extending the analyses beyond the original word pairs, we now assess all pseudowords to examine whether CLIP displays a fundamental association between syllables and shapes. The proportion of pseudoword-label responses displayed in Figure 4 reveals that there is again some variability across prompts, but model preferences are relatively consistent. Using this method, the ResNet variant displays a general preference to attend to round shapes across all word types (note that not attending to a jagged shape if the label is of the jagged category means the model attends

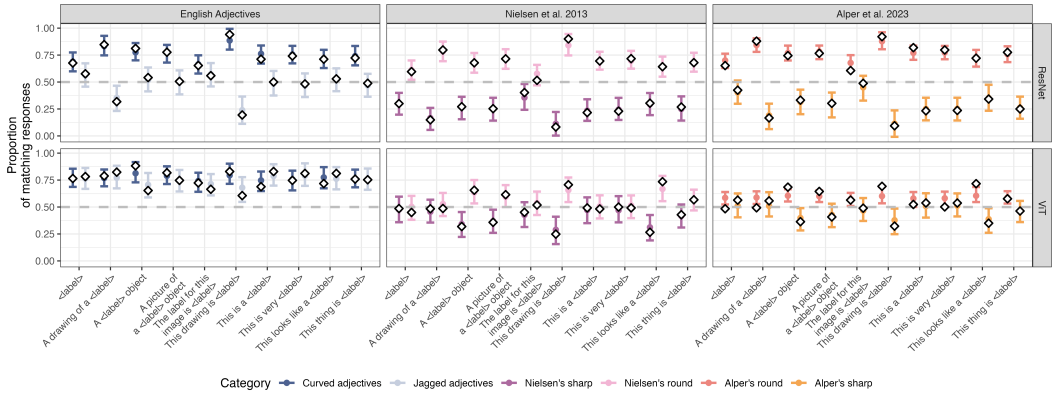

Figure 4: The proportion of correct matches (*CorrectProportion*) given a word type and category (i.e., curved or sharp). A cross-modal association is indicated when a model consistently matches images above chance (grey line) for both categories—observed only with ViT and the English adjectives used for comparison. Model: $CorrectProportion \sim 1 + Word\_type + Category + (1 + Word\_type + Category|Prompt)$. The diamonds are descriptive averages, and the dots are posterior means.

to the round shape). Provided that this is also the case for the English adjectives, it is unsurprising that the pseudowords do not display a bouba-kiki effect. The ViT variant is relatively consistent and can reliably distinguish shapes for English adjectives. Yet, using experimental pseudowords collapses performance to chance levels, indicating the absence of human-like associations between shape features and syllables or characters. A qualitative inspection of several attention patterns provides additional evidence against the presence of cross-modal associations in CLIP. This reveals that attention patterns primarily do not focus on sharp edges or round attributes of images, but instead mainly focus on the centres of shapes or background areas (Appendix D). The latter is similar to earlier findings by Darcet et al. (2024) who found that ViT networks create artefacts at low-informative background areas for computational purposes rather than describing visual information. Both observations are in contrast with what, at its core, is required for a bouba-kiki-like effect.

6 Discussion and conclusion

This work investigated whether vision-and-language models, specifically two versions of CLIP (ResNet and ViT), exhibit human-like patterns of cross-modal association, as reflected in the bouba-kiki effect. We aimed to rigorously evaluate the presence or absence of this phenomenon in model behaviour, using two methodological approaches closely aligned with human experiments and introducing the use of interpretability tools in this domain to probe internal representations. Crucially, we posit that we can only speak of a consistent model preference when results across both methodologies are congruent with human-like preferences. Our first experiment, which used image-label probabilities to gauge model preference, showed no clear cross-modal preference for either model. The second experiment utilised Grad-CAM to simulate experimental conditions in human studies. Here, the ResNet variant, which showed an overall preference for round shapes, did not demonstrate behaviour consistent enough to qualify as a bouba-kiki effect by human standards. The ViT-based version of CLIP, widely used as a foundational component in many current multimodal systems, proved capable of matching shapes and English adjectives, but failed to do so for pseudowords, which have even weaker alignment than ResNet. Contrary to prior work suggesting VLMs may encode bouba-kiki-like associations (Alper and Averbuch-Elor, 2023), our findings reveal little to no evidence that CLIP models, under these experimental conditions, exhibit consistent human-like mappings between pseudowords and visual shape features. This lack of alignment suggests that, despite their impressive performance across many downstream vision-language tasks, models like CLIP do not appear to represent cross-modal associations in a cognitively grounded manner internally. Together, this raises questions about how cross-modal grounding is encoded or inherited in many larger state-of-the-art architectures.

Our findings complicate earlier claims about the presence of sound-symbolic associations in model embeddings. While Alper and Averbuch-Elor (2023) found strong evidence for a bouba-kiki effect using embedding similarity scores in CLIP and Stable Diffusion, Iida and Funakura (2024) used the same method with Japanese VLMs, but failed to pick up robust bouba-kiki-like mappings, even though Japanese is rich in sound-symbolism (Dingemanse, 2012). Instead, the embedding similarity scores seem to reflect language-specific regularities rather than general cross-modal preferences. Our approach deliberately avoided this by mirroring psycholinguistic testing paradigms with carefully controlled pairs of novel images. Corroborating other recent studies (Verhoef et al., 2024; Loakman et al., 2024), this empirically grounded probing method yielded results opposite to those of other studies, showing no evidence of a bouba-kiki effect. Whereas Verhoef et al. (2024) report limited, though not consistent, evidence for a bouba-kiki effect, our work differs since it uses a more comprehensive set of pseudowords, amending their analyses and suggesting that there is no general effect. We also ran the same tests using English adjectives, and in contrast to pseudowords, the models are able to make the expected mappings in that case. Furthermore, by employing Grad-CAM in the domain of cross-modal associations, we examined whether models visually attend to shape features that correlate with pseudoword form profiles. This enabled us not only to assess whether the models exhibit human-like cross-modal associations, but also to determine whether they do so for the right reasons by probing the decision-making pathways within the models. They largely did not attend to the expected shape features, which presents further evidence for the claim that VLMs may lack human-like cross-modal representations.

What causes this misalignment between humans and VLMs? The broader implication is that current VLMs, even those trained on massive paired datasets, lack a key component of human-like multisen-

sory understanding: grounded, flexible, and intuitive mappings between sensory modalities. This is perhaps not surprising given the lack of embodied interaction in their training and reliance on statistical co-occurrence rather than perceptual salience. Still, it can be argued that the co-occurrences in the training data are full of human-like preferences, including cross-modal associations. So while we may expect VLMs to pick up on these associations, their internal mechanisms inherently differ from those of humans and do not necessarily learn the same preferences from aligned image-text pairs. Previous work on visual grounding (Jones et al., 2024), and shortcut learning, i.e., solving a test using unexpected non-human-like shortcuts (Schwartz and Stanovsky, 2022; Mitchell and Krakauer, 2023) argued similarly. However, interestingly, we observed that the individual modalities are, in principle, separable. This suggests that CLIP's training objective (i.e., a sentence-level contrastive loss) and its method for aligning text and image modalities (i.e., cosine similarity) are the culprits. Neither seems to specifically promote learning relations between visual features and phonetic elements in language.

Another straightforward difference between humans and VLMs is tokenisation. Unlike humans, who perceive words holistically (Wong et al., 2011), tokenisation can distort word representations, reducing the potential for human-like associations with visual shapes. Inspection of the tokenised pseudowords, however, reveals that this does not occur frequently ('Bouba' becomes 'bou' and 'ba', 'kiki' is one token, 'sepise' becomes 'sep', 'ise', and 'kaykuh' becomes 'kay', 'ku', 'h'. See section A.2 for more examples.) Most phonological structures are preserved that could match shape features. However, tokenisation may still split some pseudowords in ways that wouldn't evoke the expected cross-modal associations in humans (e.g., 'H' in 'OH' might elicit jagged rather than curved associations). In the cases where tokenisation breaks pseudowords into syllables that also show no bouba-kiki effects in humans, the set of alternative pseudowords is large enough to contain options that could invoke these associations. If models had a genuine preference in the direction of a bouba-kiki effect, they would prefer those pseudowords that are tokenised as complete syllables, and retain the potential cross-modal association of interest, over those split into non-syllables in which the association may be corrupted. Our results show this is not the case. As such, the potential value of developing models with shared preferences between humans and machines remains significant. Establishing alignment in human and machine understanding of (visual-auditory) form-meaning mappings and mutual understanding can enhance their interactions (Kouwenhoven et al., 2022), helping in creating AI systems that resemble how humans process and communicate meaning. A potentially rich line of future work includes systematically exploring how design choices in VLMs affect the emergence of cognitively meaningful representations and consistently incorporating cross-linguistic studies to account for cultural variation (as demonstrated by Iida and Funakura (2024)).

Our work has some notable limitations. First, we only test on CLIP versions. We deliberately chose to focus on the CLIP model instead of more capable proprietary models (GPT-4o and Gemini2-Flash, i.a.), since they are not sufficiently transparent and thus do not help advance our understanding of their internal mechanisms. Nevertheless, it remains interesting to further unravel cross-modal associations in more contemporary models. Although they often use frozen ViT variants of CLIP, the alignment layer between CLIP's embedding and the language model's embedding may learn preferences that resemble those of humans. While the original bouba-kiki effect is rooted in sound symbolism, our work—and that of others (e.g., Alper and Averbuch-Elor, 2023; Loakman et al., 2024; Iida and Funakura, 2024)—relies on text, which may influence the outcomes. Yet, prior human experiments present pseudowords both acoustically and in written form (e.g., Nielsen and Rendall, 2013), making it challenging to disentangle orthographic from auditory contributions. Moreover, Cuskley et al. (2017) showed that auditory presentation cannot eliminate orthographic effects in literate participants, as characters are strongly associated with particular sounds. These associations are non-arbitrary: writing systems often employ iconic strategies, in which characters representing rounded sounds (rounded vowels, sonorants) tend to have curved shapes (Koriat and Levy, 1977; Turoman and Styles, 2017). Given this tight coupling between sound and orthography, correspondences between vision and spoken words should be highly correlated with vision-text correspondences. This makes it unlikely that audio-visual models would show more substantial bouba-kiki effects than text-based models. Nonetheless, exploring audio-visual models (Tseng et al., 2024) remains a valuable direction for understanding how different modalities contribute to cross-modal associations.

Overall, our results reinforce the importance of interpretability and cognitively inspired evaluation when assessing model performance in cross-modal reasoning. If VLMs are to serve as truly intuitive agents in real-world human-machine interactions, they must not only succeed on benchmark datasets but also exhibit a more human-like understanding of abstract, grounded concepts.

References

Abramova, E. and Fernández, R. (2016). Questioning arbitrariness in language: a data-driven study of conventional iconicity. In Knight, K., Nenkova, A., and Rambow, O., editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 343–352, San Diego, California. Association for Computational Linguistics.

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. (2018). 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.

Allen, K. R., Dasgupta, I., Kosoy, E., and Lampinen, A. K. (2025). The in-context inductive biases of vision-language models differ across modalities. In *Second Workshop on Representational Alignment at ICLR 2025*.

Alper, M. and Averbuch-Elor, H. (2023). Kiki or bouba? sound symbolism in vision-and-language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78347–78359. Curran Associates, Inc.

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., and Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823.

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5):1–54.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660.

Chefer, H., Gur, S., and Wolf, L. (2021). Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., and Dai, J. (2024). Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Cuskley, C. and Kirby, S. (2013). Synesthesia, Cross-Modality, and Language Evolution. In *Oxford Handbook of Synesthesia*. Oxford University Press.

Cuskley, C., Simner, J., and Kirby, S. (2017). Phonological and orthographic influences in the bouba–kiki effect. *Psychological research*, 81:119–130.

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., Lupyan, G., Oh, G. E., Paul, J., Petrone, C., Ridouane, R., Reiter, S., Schümchen, N., Szalontai, Á., Ünal-Logacev, Ö., Zeller, J., Perlman, M., and Winter, B. (2022). The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1841):20200390.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. (2024). Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*.

de Varda, A. G. and Strapparava, C. (2022). A cross-modal and cross-lingual study of iconicity in language: Insights from deep learning. *Cognitive Science*, 46(6):e13147.

Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J. S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., Lu, J., Anderson, T., Bransom, E., Ehsani, K., Ngo, H., Chen, Y., Patel, A., Yatskar, M., Callison-Burch, C., Head, A., Hendrix, R., Bastani, F., VanderBilt, E., Lambert, N., Chou, Y., Chheda, A., Sparks, J., Skjonsberg, S., Schmitz, M., Sarnat, A., Bischoff, B., Walsh, P., Newell, C.,

Wolters, P., Gupta, T., Zeng, K.-H., Borchardt, J., Groeneveld, D., Nam, C., Lebrecht, S., Wittlif, C., Schoenick, C., Michel, O., Krishna, R., Weihs, L., Smith, N. A., Hajishirzi, H., Girshick, R., Farhadi, A., and Kembhavi, A. (2024). Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models.

Demircan, C., Saanum, T., Pettini, L., Binz, M., Baczkowski, B. M., Doeller, C. F., Garvert, M. M., and Schulz, E. (2024). Evaluating alignment between humans and neural network representations in image-based learning tasks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass*, 6(10):654–672.

Dingemanse, M., Schuerman, W., Reinisch, E., Tufvesson, S., and Mitterer, H. (2016). What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language*, 92(2):e117–e133.

Diwan, A., Berry, L., Choi, E., Harwath, D., and Mahowald, K. (2022). Why is winoground hard? investigating failures in visuolinguistic compositionality. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Fort, M. and Schwartz, J.-L. (2022). Resolving the bouba-kiki effect enigma by rooting iconic sound symbolism in physical properties of round and spiky objects. *Scientific reports*, 12(1):19172.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). 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.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203:88–96.

Iida, H. and Funakura, H. (2024). Investigating iconicity in vision-and-language models: A case study of the bouba/kikieffect in japanese models. In *Proceedings of the 46th Annual Conference of the Cognitive Science Society*, volume 46.

Imai, M. and Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical transactions of the Royal Society B: Biological sciences*, 369(1651):20130298.

Imai, M., Kita, S., Nagumo, M., and Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, 109(1):54–65.

Jabri, A., Joulin, A., and Van Der Maaten, L. (2016). Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer.

Jones, C. R., Bergen, B., and Trott, S. (2024). Do multimodal large language models and humans ground language similarly? *Computational Linguistics*, 50(3):1415–1440.

Kamath, A., Hessel, J., and Chang, K.-W. (2023). What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore. Association for Computational Linguistics.

Köhler, W. (1929). *Gestalt Psychology*. New York: Horace Liveright.

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

Koriat, A. and Levy, I. (1977). The symbolic implications of vowels and of their orthographic representations in two natural languages. *Journal of Psycholinguistic Research*, 6(2):93–103.

Kouwenhoven, T., Peeperkorn, M., Van Dijk, B., and Verhoef, T. (2024). The curious case of representational alignment: Unravelling visio-linguistic tasks in emergent communication. In Kuribayashi, T., Rambelli, G., Takmaz, E., Wicke, P., and Oseki, Y., editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 57–71, Bangkok, Thailand. Association for Computational Linguistics.

Kouwenhoven, T., Verhoef, T., De Kleijn, R., and Raaijmakers, S. (2022). Emerging grounded shared vocabularies between human and machine, inspired by human language evolution. *Frontiers in Artificial Intelligence*, 5:886349.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Liu, H., Li, C., Li, Y., and Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Loakman, T., Li, Y., and Lin, C. (2024). With ears to see and eyes to hear: Sound symbolism experiments with multimodal large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2849–2867, Miami, Florida, USA. Association for Computational Linguistics.

Mamooler, S. (2021). Clip explainability.

Marklová, A., Milička, J., Ryvkin, L., Ľudmila Lacková Bennet, and Kormaníková, L. (2025). Iconicity in large language models.

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

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

Mitchell, M. and Krakauer, D. C. (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Nielsen, A. and Rendall, D. (2011). The sound of round: evaluating the sound-symbolic role of consonants in the classic takete-maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 65(2):115.

Nielsen, A. and Rendall, D. (2012). 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.

Nielsen, A. and Rendall, D. (2013). 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.

Perniss, P., Thompson, R., and Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in Psychology*, 1(227).

Perry, L. K., Perlman, M., and Lupyan, G. (2015). Iconicity in english and spanish and its relation to lexical category and age of acquisition. *PloS one*, 10(9):e0137147.

Perry, L. K., Perlman, M., Winter, B., Massaro, D. W., and Lupyan, G. (2018). Iconicity in the speech of children and adults. *Developmental Science*, 21(3):e12572.

Pimentel, T., McCarthy, A. D., Blasi, D., Roark, B., and Cotterell, R. (2019). Meaning to form: Measuring systematicity as information. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764, Florence, Italy. Association for Computational Linguistics.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Schwartz, R. and Stanovsky, G. (2022). On the limitations of dataset balancing: The lost battle against spurious correlations. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2182–2194, Seattle, United States. Association for Computational Linguistics.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vision*, 128(2):336–359.

Shiono, D., Brassard, A., Ishizuki, Y., and Suzuki, J. (2025). Evaluating model alignment with human perception: A study on shitsukan in LLMs and LVLMs. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11428–11444, Abu Dhabi, UAE. Association for Computational Linguistics.

Tamariz, M., Roberts, S. G., Martínez, J. I., and Santiago, J. (2018). The interactive origin of iconicity. *Cognitive Science*, 42(1):334–349.

Taubert, J., Apthorp, D., Aagten-Murphy, D., and Alais, D. (2011). The role of holistic processing in face perception: Evidence from the face inversion effect. *Vision Research*, 51:1273–1278.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. (2022). 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.

Tseng, W.-C., Shih, Y.-J., Harwath, D., and Mooney, R. (2024). Measuring sound symbolism in audio-visual models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1165–1172.

Turoman, N. and Styles, S. J. (2017). Glyph guessing for 'oo'and 'ee': Spatial frequency information in sound symbolic matching for ancient and unfamiliar scripts. *Royal Society open science*, 4(9):170882.

Verhoef, T., Kirby, S., and De Boer, B. (2016a). Iconicity and the emergence of combinatorial structure in language. *Cognitive science*, 40(8):1969–1994.

Verhoef, T., Roberts, S. G., and Dingemanse, M. (2015). Emergence of systematic iconicity: Transmission, interaction and analogy. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 2481–2486. Cognitive Science Society.

Verhoef, T., Shahrasbi, K., and Kouwenhoven, T. (2024). What does kiki look like? cross-modal associations between speech sounds and visual shapes in vision-and-language models. In Kuribayashi, T., Rambelli, G., Takmaz, E., Wicke, P., and Oseki, Y., editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 199–213, Bangkok, Thailand. Association for Computational Linguistics.

Verhoef, T., Walker, E., and Marghetis, T. (2016b). Cognitive biases and social coordination in the emergence of temporal language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 38, pages 2615–2620.

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

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

Wong, A. C.-N., Bukach, C. M., Yuen, C., Yang, L., Leung, S., and Greenspon, E. (2011). Holistic processing of words modulated by reading experience. *PLOS ONE*, 6(6):1–7.

Zakka, K. (2021). A playground for clip-like models.

Zhao, M., Bülthoff, H. H., and Bülthoff, I. (2016). Beyond faces and expertise: Facelike holistic processing of nonface objects in the absence of expertise. *Psychological Science*, 27(2):213–222. PMID: 26674129.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: section 4 clearly shows that both models perform below chance level. section 5 directly compares computational results to work by (Ćwiek et al., 2022).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The final section includes some limitations to our methodology, and reliance on text-based models instead of audio-visual models. We, moreover, discuss for each experiment how consistent the models' predictions are, regardless of whether they make human-like associations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA] .

Justification: Our work does not focus on making a theoretical contribution.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The method section (section 3) clearly describes how our linguistic inputs are made, or where they originate from. Idem for the images used. The Bayesian Regression Models used are also clearly specified in each analyses including the parameters used to fit them.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our source code and the data will be published on OSF upon publication.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Similar to the result reproducibility, our methodology is clearly described in section 3. The methods to create the images and labels are also described in detail. The tests used to make our contributions are also clearly specified underneath each figure. Moreover, the source code will be available upon publication, exposing any other experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide credible intervals in each figure and explain how our analyses should be interpreted (section 3). Moreover, we explicitly state the models fitted and the parameters used to do the fitting.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Our experiments do not need heavy computing since we do not train new models or fine-tune anything. All the code and analyses ran on a single MacBook Air M1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do not violate the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

Justification: Our work does not discuss societal impact, as it primarily focuses on whether current CLIP models (already widely available) align with human cross-modal preferences. We do not develop a new model that may be used by society, deal with biased data that may result in discrimination, or any related issue. As such, we do not deem it necessary to discuss societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: There is no need for safeguards since we only evaluate already available models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the code-base we used for Grad-CAM (section 5), and properly mention the sources of the images and labels that we did not generate ourselves (section 3).

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not use human subjects but base our comparisons to human performance on other work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our work did not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

A Linguistic inputs

The experimental inputs are tested across ten different prompts to assess robustness. They originate from earlier investigations (Alper and Averbuch-Elor, 2023; Verhoef et al., 2024) and are extended with additional prompts such that the *label* occurs equally frequently as a noun and adjective (Table 1).

Prompt	Word type	Origin
The label for this image is <label>	Noun	Verhoef et al. (2024)
This is a <label>	Noun	new
A drawing of a <label>	Noun	new
This drawing is <label>	Adjective	new
This thing is <label>	Adjective	Alper and Averbuch-Elor (2023)
A <label> object	Adjective	Alper and Averbuch-Elor (2023)
A picture of a <label> object	Adjective	Alper and Averbuch-Elor (2023)
<label>	Noun	Alper and Averbuch-Elor (2023)
This looks like a <label>	Noun	new
This is very <label>	Adjective	new

Table 1: Different linguistic inputs are used to test our models for cross-modal associations.

A.1 Textual embeddings

To make correct associations, the models must, at a minimum, be able to disambiguate labels (real and non-words) from each other. If they fail to do so, it is presumably also impossible to link certain words and their linguistic features to shape-specific features. Figure 5 displays a t-SNE visualisation of the models' embeddings and reveals that they, in principle, should be able to disambiguate labels from different categories within a word type.

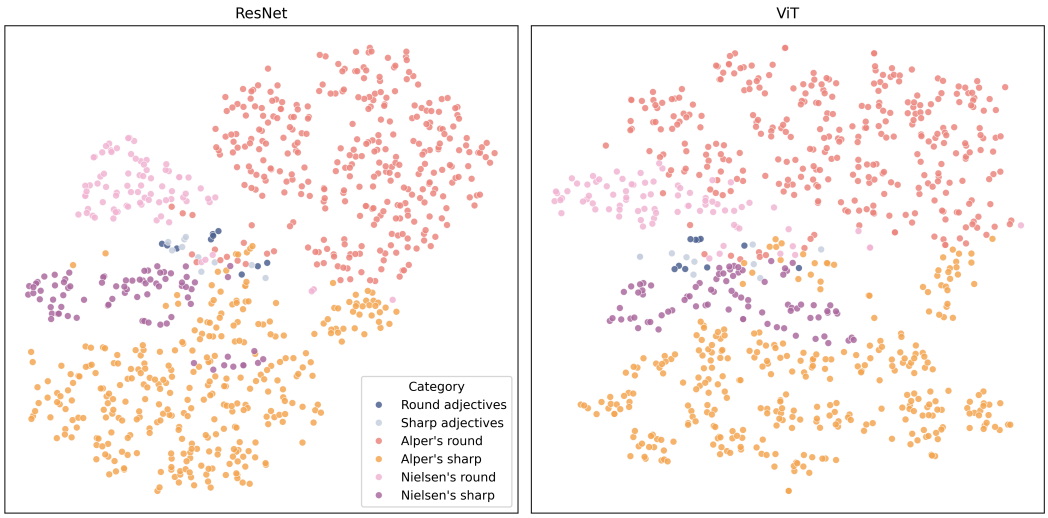

Figure 5: t-SNE plot showing how the language models of different CLIP variants interpret labels from different categories. The colour shades indicate which word type a label in a category belongs to. In order to correctly match labels to images with shape-specific features, a model must be able to discriminate word types between labels of the same category. This is clearly possible. This plot shows the embeddings for the prompt: *The label for this image is <label>*. Different plots result in similar distributions.

A.2 Tokenisation

To determine whether byte-pair encodings or our pseudowords break phonological structures, we qualitatively assess a set of tokenised examples. These pseudowords are parsed as elements of the

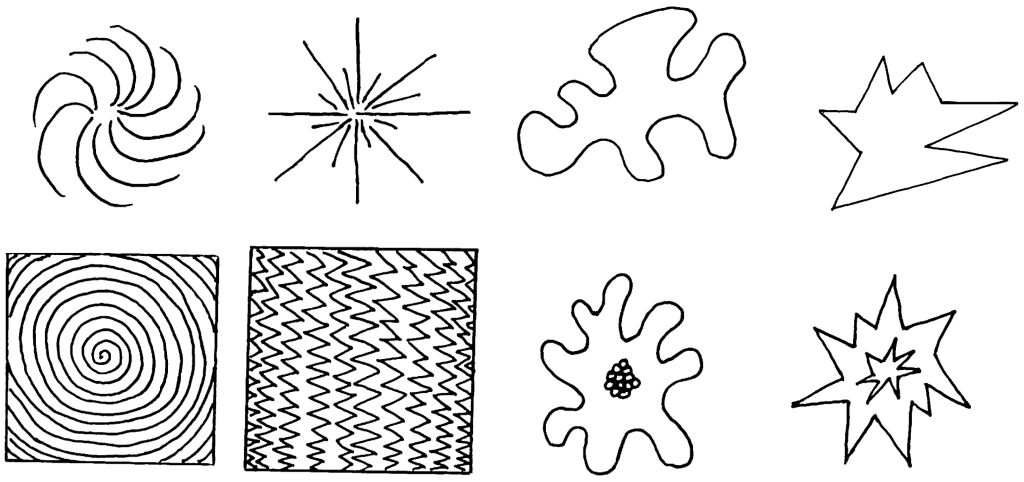

Figure 6: Image pairs from (Maurer et al., 2006b)

sentence prompts, but we only focus on the pseudowords themselves. The list below reveals that most of the target words are tokenised in a way that preserves at least some phonological structures that could be matched to shape features.

- Bouba → 'bou' and 'ba'
- Kiki → 'kiki'
- Takete → 'take', 'te'
- Maluma → 'mal', 'uma'
- Xehaxe → 'xe', 'ha', 'xe'
- Lohlah → 'loh', 'lah'
- Sepise → 'sep', 'ise'
- Kaykuh → 'kay', 'ku', 'h'
- Loomoh → 'loom', 'oh'
- Mohmah → 'moh', 'mah'

B Visual inputs

The full set of images with visual shapes that were used in the experiments is shown here. Besides the original image pair from Köhler (1929, 1947), we used four image pairs from Maurer et al. (2006b), displayed in Figure 6, four from (Westbury, 2005a, ; Figure 7), and eight generated pairs following the method described by Nielsen and Rendall (2013) and used in (Verhoef et al., 2024, ; Figure 8;). For each image pair, the Curved version is displayed on the left and the Jagged version on the right.

B.1 Visual embeddings

Similar to the textual embeddings presented before, Figure 9 displays a t-SNE visualisation of the models' visual embeddings. It is visible that the models should, in principle, be able to disambiguate our target images based on their condition, even when they only differ in how the randomly distributed points are connected.

C Unique labels

To test whether the models in the first experiment (section 4) change their predictions according to a set of images, we report the ratio of unique labels in Table 2. In this case, we are not concerned with

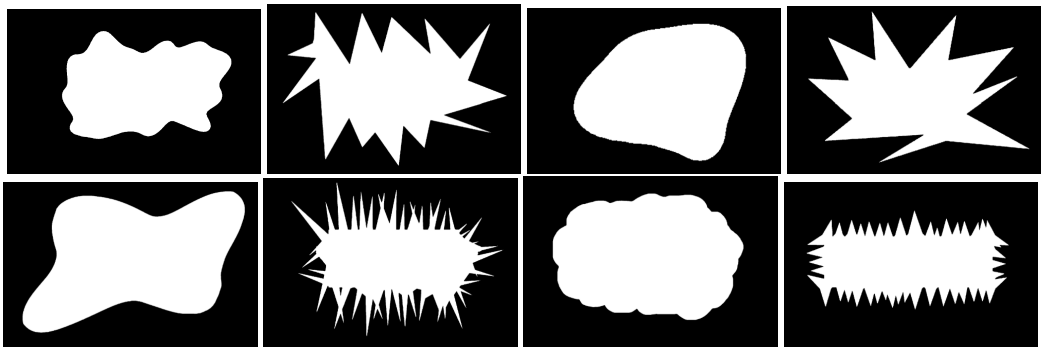

Figure 7: Image pairs from (Westbury, 2005a)

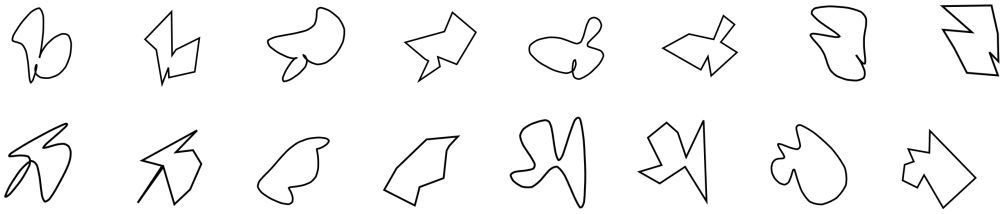

Figure 8: Newly generated image pairs

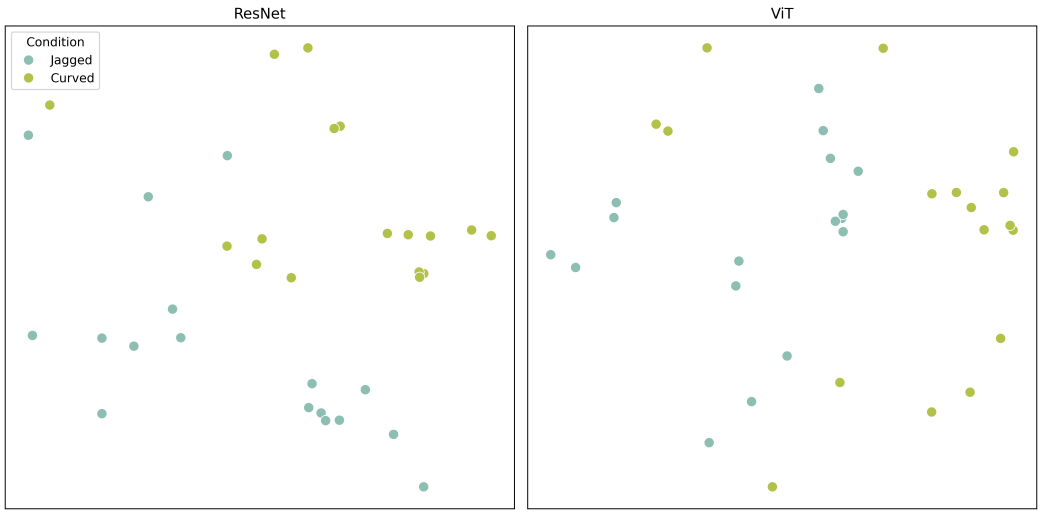

Figure 9: t-SNE plot showing how the vision models of different CLIP variants interpret images from conditions. The colour shades indicate which image condition an image belongs to. In order to correctly match labels to an image, a model must be able to discriminate the images based on their shape. This is clearly possible.

Model	Word type	Ratio
ResNet	Initial words	.675
ResNet	English adj.	.275
ResNet	Nielsen et al.	.329
ResNet	Alper et al.	.444
ViT	Initial words	.800
ViT	English adj.	.320
ViT	Nielsen et al.	.331
ViT	Alper et al.	.450

Table 2: The average ratio of unique labels chosen for each image (n=34) across different prompts (n=10) in different label sets. The latter means that we divide the unique labels by the length of the set of possible labels to gauge diversity. A high ratio indicates that the corresponding model assigned different labels to different images.

whether the models make predictions that align with the bouba-kiki effect, but rather with the diversity of their predictions. We calculate the ratio of unique labels for each word type across all images and the ten prompts. These ratios should, in a scenario where cross-modal associations are present, be high, and the number of correct trials, as seen in Figure 2, would be above chance. However, it is clear that the models somewhat change their predictions (mainly for the initial pseudowords) when they are conditioned on different images, but mostly *collapse* onto the same labels for different images. Table 2 confirms this by showing that there is only slight variation among picked labels for each prompt. Given that responses are only counted as congruent when both predictions (i.e. the prediction for a jagged and curved shape) are correct, this explains why we find that both models perform at chance levels. A qualitative example is provided in Table 3, which shows the unique labels and their ratios for a single prompt. The model and word type that are most affected by different images (i.e. have a high ratio of unique words) also exhibit the strongest bouba-kiki like effect Figure 2. This appears to happen even though, in the case of both ViT and ResNet, and Alper and Averbuch-Elor labels, the majority of the labels would be associated with sharp images by humans.

Model	Word type	Ratio	Uniquely chosen labels
ResNet	Initial words	.250	bouba
ResNet	English adj.	.300	angular, circular, curved, prickly, rotund, spiky
ResNet	Nielsen et al.	.412	kaykee, kuhpay, kuhpee, kuhpuh, lahmoo, lohloh, lohmah, lohmoo, mohmah, nahmoo, nohmoo, paykuh, peepay, teepee
ResNet	Alper et al.	.471	kehake, kehike, ladula, lunulu, malama, nulunu, paxapa, pihapi, pikepi, pikipi, pixapi, sepise, tatata, tekete, xaxixa, xehexe
ViT	Initial words	.500	bouba, takete
ViT	English adj.	.250	angular, circular, curved, rotund, spiky
ViT	Nielsen et al.	.412	keekay, kuhtay, lahmoo, lohlah, loomoh, mahloh, mahnoh, mohloo, mohnoh, nohloh, noonoo, taypay, taypuh, teepee
ViT	Alper et al.	.676	dododo, hexehe, kixiki, lamola, lubalu, lunulu, mamuma, monomo, mubomu, patipa, pisipi, pixapi, sahisa, satasa, tehite, texate, xahexa, xakixa, xasixa, xehexe, xikexi, xipaxi, xisixi

Table 3: The set of unique labels chosen across all images (n=34) and its ratio for an example prompt ('The label for this image is <label>'). The latter means that we divide the unique labels by the length of the set of possible labels to gauge diversity. A high ratio indicates that the corresponding model assigned different labels to different images.

D Grad-Cam visualisations, consistency, and additional results

Figure 10 provides example visualisations of the attention patterns for different randomly selected image pairs and randomly selected labels across both models. Here, we sampled different image pairs for the models as to give a complete view of the images used. The pseudowords are consistent for the columns. The images reveal that models mostly attend to the centres of shapes and/or focus on non-informative background areas. The latter is also described by Darcet et al. (2024). Although this is a small sample, it is clear that the target labels do not steer models towards shape-specific features.

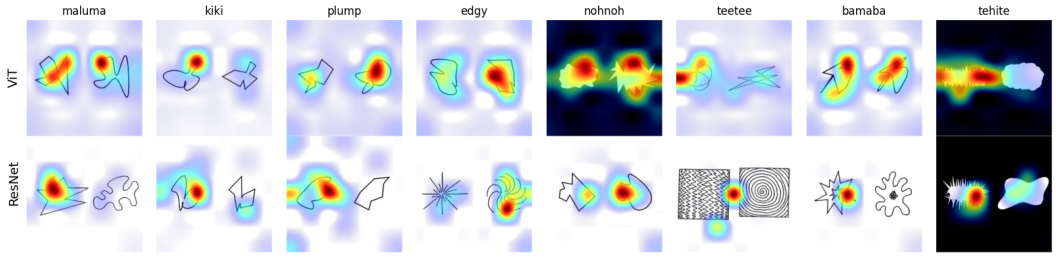

Figure 10: Visualising the attention pattern for the text prompt: 'The label for this image is <label>' in both models.

D.1 Prediction consistency

To test whether the models in the second experiment (section 5) change their predictions resulting from different shape positions, we report the ratio of consistency in Table 4. In this case, we are not interested in whether the models make predictions that align with the bouba-kiki effect, but rather in the consistency of their predictions. Both models are rather consistent in the mappings they make between labels and shapes.

D.2 Quantifying model preference

The results presented in section 5 utilise the sum of attention values to quantify the models' preferences, as this aligns with humans' holistic perception (Taubert et al., 2011; Zhao et al., 2016; Wong et al., 2011). Though artificial models may show their preference differently. To this end, we additionally experimented with the entropy and centroid-of-attention as quantifiers of model preference.

Comparing the predictions (averaged over all images, labels, and prompts) quantified by the centroid of attention with the sum of attention, we find that the predictions using the centroid of attention strongly overlap with those resulting from the sum of attention (ViT: 85.6% and ResNet: 88.0%). Given this considerable overlap, it is not surprising that the results remain highly similar when using

Model	Word type	Ratio
ResNet	Initial words	.799
ResNet	English adjectives	.761
ResNet	Nielsen et al.	.758
ResNet	Alper et al.	.765
ViT	Initial words	.723
ViT	English adjectives	.749
ViT	Nielsen et al.	.722
ViT	Alper et al.	.739

Table 4: The ratio of consistently attending to the same shape in a different position when presented with the same label. Values are averages across image pairs, prompts, and labels. A high ratio indicates that the corresponding model consistently focuses on the same shape, even when the target location is different.

Model	Sum	Entropy	Centroid
ResNet	.519	.488	.515
ViT	.522	.514	.515

Table 5: The correctness of all model prediction types using different quantifications averaged across all images, labels, and prompts.

this alternative method (Table 5. Yet, for the entropy of attention, the predictions (where the image half with the lower entropy acts as the model's preference) for ViT overlap strongly (80.2%) but not for ResNet (21.1%). This entropy method, despite yielding different predictions, still results in a slightly lower number of correctly predicted images, with only 48.8% of the shapes correctly identified, compared with 51.9% for the quantification method that uses the sum of attention. So if we were to use entropy-based predictions, the performance would be even worse than predictions based on the sum of attention.