Collaborative meaning-making: the emergence of novel languages in humans, machines, and human-machine interactions

Kouwenhoven, T.

4

Kiki or Bouba?

Humans have clear cross-modal preferences when matching certain novel words to visual shapes. Evidence suggests that these preferences play a prominent role in our linguistic processing, language learning, and the origins of signal-meaning mappings. With the rise of multimodal models in AI, such as vision-and-language (VLM) models, it becomes increasingly important to uncover the kinds of visio-linguistic associations these models encode and whether they align with human representations. Informed by experiments with humans, we probe and compare four VLMs for a well-known human cross-modal preference, the bouba-kiki effect. We do not find conclusive evidence for this effect, but suggest that results may depend on features of the models, such as architecture design, model size, and training details. Our findings inform discussions on the origins of the bouba-kiki effect in human cognition and future developments of VLMs that align well with human cross-modal associations.

Originally published as: Tessa Verhoef*, Kiana Shahrasbi, and Tom Kouwenhoven*. 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., Oseki, Y., editors, *In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 199–213, Bangkok, Thailand. Association for Computational Linguistics. (*denotes equal contribution.)

4.1 Introduction

The development of machine understanding and generation of natural language has benefited immensely from the introduction of transformer-based architectures (Vaswani et al., 2017). These architectures have since then been adapted and extended to handle multimodal data, leading to the creation of various types of multimodal models, including vision-and-language models. These models can potentially revolutionise how AI systems understand the world and interact with humans. However, we lack direct access to the exact representations and associations they encode. How VLMs integrate representations in the two modalities and whether associations between modalities are made in a human-like way is still being actively investigated (Alper et al., 2023; Kamath et al., 2023; Zhang et al., 2024c; Karamcheti et al., 2024; Jones et al., 2024).

Here, we use a well-known paradigm from the field of cognitive science to probe into a specific cross-modal association between speech sounds and visual shapes: the bouba-kiki effect. When humans see two figures, one with jagged and one with smooth edges, and are told one is a Kiki and the other a Bouba, 95% will name the jagged figure Kiki (Ramachandran and Hubbard, 2001). This effect was initially discovered and described anecdotally by Wolfgang Kóhler (Köhler, 1929, 1947), using the two images shown in Figure 4.1 with the labels *maluma* and *takete*. Since then, it has been widely studied (as reviewed in Section 4.2), and expanded with many other cross-modal preferences in human processing of (speech) sounds and visual imagery. Moreover, a wealth of evidence suggests that such preferences widely influence patterns we see in human languages (e.g., Ramachandran and Hubbard, 2001; Cuskley and Kirby, 2013; Imai and Kita, 2014; Verhoef et al., 2015, 2016a; Tamariz et al., 2018). Even though non-arbitrariness in language is often still regarded as an exception in some disciplines, in fields such as language evolution and sign language linguistics, iconic form-meaning mappings are considered omnipresent (Perniss et al., 2010). Given the central role cross-modal preferences play in human visio-linguistic representations and their effects on language, it is pertinent to investigate whether VLMs associate non-words and visual stimuli in a human-like way.

Examining universal human cross-modal preferences in VLMs can help us gain key insights across disciplines. First, it may reveal whether VLMs process multimodal information in a human-like way and whether similar biases drive their understanding of visual-auditory form-meaning mappings. Overlap in cognitive biases can potentially increase mutual understanding and improve interactions between humans and machines (Chapter 1). Second, it may help pinpoint what is missing to make VLMs more suitable for realistic simulations of human language emergence. Increasingly, VLMs are used in emergent communication settings, where agents communicate with each other and develop a novel language (Bouchacourt and Baroni, 2018; Mahaut et al., 2025). These models are used to improve machine understanding of human language (Lazaridou and Baroni, 2020; Lowe et al., 2020; Steinert-Threlkeld et al., 2022; Zheng et al., 2024), but also to simulate and study human language evolution processes (Galke et al.,

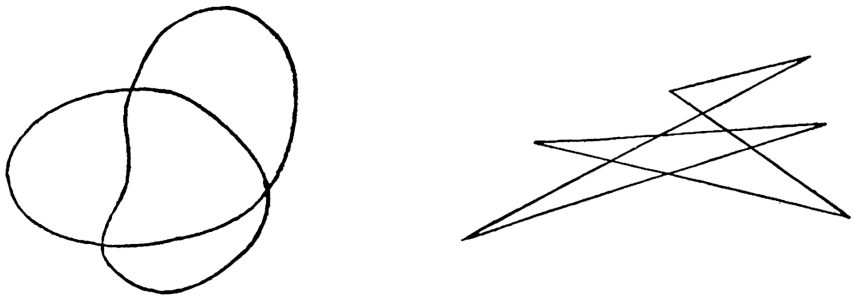

Figure 4.1: Which of these two shapes is Kiki? Images from Köhler (1929, 1947).

2022; Lian et al., 2023b). While the influence of cross-modal associations on the emergence of language has been studied extensively in language evolution experiments with humans (Verhoef et al., 2015, 2016a; Tamariz et al., 2018; Little et al., 2017), this phenomenon remains absent from current emergent communication paradigms. Evidently, cognitively plausible VLMs are more suitable for simulating aspects of the evolution of meaning in language. Finally, the actual origin of the bouba-kiki effect is still being debated within cognitive science and linguistics, with proposed explanations ranging from attributing it to similarities between shape features and features of either orthography (Cuskley et al., 2017), acoustics and articulation (Ramachandran and Hubbard, 2001; Maurer et al., 2006; Westbury, 2005), 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). If the bouba-kiki effect can be reproduced in a VLM, it can help reveal the crucial ingredients for this effect, potentially leading to models better aligned with human representations.

To the best of our knowledge, only one previous paper discussed the bouba-kiki effect in VLMs. Alper and Averbuch-Elor (2023) tested two models, CLIP (Radford et al., 2021) and Stable Diffusion (Rombach et al., 2022), and reported finding strong evidence for the effect in these models. This is somewhat surprising given the way these models are trained and the absence of relevant data sources, such as auditory information and experience with physical object properties. Therefore, we introduce nuance in this discussion and show, contrary to the previous finding, that the bouba-kiki effect does not occur consistently in VLMs and that the presence of this cross-modal preference may depend on the way it is tested, as well as properties such as model architecture, attention mechanism, and training details.

4.2 Background

4.2.1 Sound-symbolism and cross-modal associations in language and cognition

When Hockett (1960) listed a set of design features deemed essential to natural human language, 'arbitrariness' was included. This feature refers to the arbitrary/unmotivated mapping between words and their meanings. However, when exploring beyond Indo-European languages, non-arbitrary form-meaning mappings appear to play a significant role in many languages (Imai et al., 2008; Perniss et al., 2010; Dingemanse, 2012). Most obviously, perhaps, sign languages are rich in non-arbitrary 'iconic' mappings, with articulators that lend themselves particularly well to representing meanings by mimicking, for example, shapes or actions. However, some spoken languages also have specific classes of words where characteristics of the meaning are mimicked or iconically represented in the word. Examples have been identified as 'ideophones,' 'mimetics', or 'expressives,' and this phenomenon is often referred to as sound-symbolism (Imai et al., 2008; Imai and Kita, 2014; Dingemanse, 2012). Even in languages not typically considered rich in sound symbolism, such as English and Spanish, vocabulary items from specific lexical categories, like adjectives, are also rated high in iconicity (Perry et al., 2015). Perhaps the most overwhelming evidence for the widespread importance of sound-symbolism in human languages comes from a study by Blasi et al. (2016), who analysed vocabularies of two-thirds of the world's languages and found evidence for strong associations between speech sounds and particular meanings across geographical locations and linguistic lineages. Consequently, non-arbitrariness is an important property of all languages.

In addition, human language learning, processing, and evolution are affected by cross-modal associations. Sound-symbolic mappings help young children acquire new words (Imai et al., 2008), and iconic words are learned earlier in child language development (Perry et al., 2015). Furthermore, parents use sound-symbolic words in their infant-directed speech more often than in adult-to-adult conversations (Imai et al., 2008). In a novel word learning task, participants trained on a mapping congruent with a known cross-modal association performed better than participants in an incongruent condition (Nielsen and Rendall, 2012). Sound-symbolic mappings in language have been connected to cross-modal mappings in the human brain (Ramachandran and Hubbard, 2001; Simner et al., 2010; Lockwood and Dingemanse, 2015) and processing of sound-symbolic words is less affected by aphasia (language-affecting brain damage after left-hemisphere stroke), than arbitrary words (Meteyard et al., 2015). It is also argued that universally shared cross-modal biases play an essential role in the evolution of language by bridging the gap between sensory input and meaning by providing a basis for linguistic conventions (Ramachandran and Hubbard, 2001; Cuskley and Kirby, 2013; Imai and Kita, 2014). Shared biases can help to create mutual understanding because communicative partners will automatically understand what is meant when a word like 'kiki' is used for the

first time in a context like the one shown in Figure 4.1.

While the bouba-kiki effect may be the most famous example of a universal cross-modal association, numerous other cognitive biases in cross-modal perception have been reported. For instance, non-arbitrary associations exist in human processing between high pitch sounds and light shades (Marks, 1974; Melara, 1989; Ward et al., 2006), light shades with rising intonation (Hubbard, 1996), graphemes and colours (Cuskley et al., 2019), vowel height and lightness (Cuskley et al., 2019), small size and high pitch (Evans and Treisman, 2010; Parise and Spence, 2009) and vowel openness and visual size (Schmidtke et al., 2014). Therefore, the findings presented in this chapter only scratch the surface of what is possible in this domain.

4.2.2 Testing the bouba-kiki effect in humans

After its initial discovery, the bouba-kiki effect has been studied increasingly rigorously, extending the initial pair of two images with more possible pairs (Maurer et al., 2006; Westbury, 2005) and even randomly generated ones to control for biases related to deliberate selection by the researchers (Nielsen and Rendall, 2011, 2013). In addition, various sets of labels and pseudowords have been contrasted and compared to study the relative importance of vowels versus consonants in the labels (Westbury, 2005; Nielsen and Rendall, 2011, 2013). The role of orthography, in addition to the auditory properties of speech sounds, has also been studied (Cuskley et al., 2017; Bottini et al., 2019). Across setups, non-arbitrary preferences are found to be robust across varying cultures and writing systems (Ćwiek et al., 2022). Remarkably, to some extent, this can even be found in blind individuals who undergo a haptic version of the bouba-kiki task (Bottini et al., 2019).

Most experiments in this domain are conducted using a two-alternative forced choice design, where two contrasting images are shown side by side (one jagged and the other curved), and two possible labels are offered, asking participants to make the "correct' mapping. However, it has been argued that this is an anti-conservative method in the sense that the concurrent presentation of two images that differ along one dimension and two labels that also differ along one dimension strongly primes participants to match the two, noticing their similarities. Nielsen and Rendall (2013) therefore introduced a different method, in which images are presented independently, and participants are asked to generate novel pseudowords to match the images. Here, we adopt their approach as a stringent method for probing VLMs for the bouba-kiki effect.

4.2.3 Vision-and-language models

Despite recent advances in multi-modal models (Zhang et al., 2024a) using transformer architectures, they remain poorly understood and often show unwanted behaviours such as poor visio-compositional reasoning (Thrush et al., 2022; Diwan et al., 2022) or spatial reasoning skills (Kamath et al., 2023). In addition, in the visual question-answering domain, it is a well-known

problem that models often lack visual grounding and struggle to integrate textual and visual data (Goyal et al., 2017; Jabri et al., 2016; Agrawal et al., 2018). This makes it perhaps even more puzzling that Alper and Averbuch-Elor (2023) found strong evidence for a bouba-kiki effect in CLIP and Stable Diffusion. Even if these models are able to extract sound-symbolic information in the absence of auditory data, they will likely struggle to actually associate that information with visual properties.

The approach taken by Alper and Averbuch-Elor (2023) involved generating two large sets of pseudowords, where one set was more likely associated with round shapes (examples: *bodubo, gunogu, momomo*) and the other set would evoke associations with jagged shapes (examples: *kitaki, hipehi, texete*). The CLIP embedding vector space was used to define a visual semantic dimension that best separates two sets of pre-selected adjectives (various synonyms of round and jagged). Within this space, pseudoword properties could reliably predict adjective type (round or jagged), and geometric properties associated with those adjectives could predict the category of pseudowords. With Stable Diffusion, novel images were generated based on pseudowords and analysed by embedding them using CLIP and through human evaluation. Both methods revealed evidence for the presence of sound symbolic mappings in these models (Alper and Averbuch-Elor, 2023).

While their methods mainly involved generating images from text (with Stable Diffusion) or investigating text-to-text mapping (with CLIP embeddings), we focus on image-to-text classification. We use images previously used in experiments with humans, as well as novel images generated following a procedure previously used to generate items for human experimentation. This approach provides an additional way of testing for cross-modal associations in VLMs and yields data that can be more directly compared to human data from studies into the bouba-kiki effect. If VLMs indeed learned human-like associations between visual and textual modalities, these should show robustly across multiple experiments that test the same associations differently. Moreover, Alper and Averbuch-Elor (2023) did not explicitly compare different VLMs (Stable Diffusion also uses CLIP). However, it would not be surprising if properties relating to the architecture, for example, affect the presence of this effect since these properties directly determine how the modality gap is bridged. Previous findings also suggest that dataset diversity and scale are the primary drivers of alignment to human representations (Conwell et al., 2023; Muttenthaler et al., 2023). As such, we compare four models here, each with a different architecture, attention mechanism, and training objective.

While many different architectures exist, they typically use single or dual-stream architectures. Either combining the inputs from two modalities and encoding them jointly (single-stream) or encoding them by two separate modality-specific encoders (dual-stream). Single-stream architectures typically use merged attention, where the language and visual input attend to both themselves and the other modality. Dual-stream architectures often use some form of cross-model attention, like co-attention and modality-specific attention, in addition to merged attention. Recently, Li et al. (2023) introduced a lightweight Querying Transformer (Q-Former) to

Model	Objective	Architecture	Attention	#Params	#imgs,#caps
CLIP	CON	Dual-stream	Mod-spec	151.3M	400, 400M
ViLT	ITM, MLM	Single-stream	Merged	87.4M	4.10, 9.85M
BLIP2	CON, IGTG, ITM	Dual-stream	Q-Former	∼3.8B	129, 258M
GPT-4o	Unknown	Unknown	Unknown	Unknown	Unknown

Table 4.1: Overview of the models used in this chapter. The training objectives are Image Text Matching (ITM), Masked Language Modelling (MLM), Image-grounded Text Generation (IGTG), or Contrastive Learning (CON). Mod-spec refers to modality-specific attention. Numbers are millions (M) or billions (B).

bridge the modality gap between any arbitrary pre-trained frozen vision model and a language model, resulting in BLIP2. Frequently, image text matching and masked language modelling are used as learning objectives (e.g., ViLT; Kim et al., 2021), but some methods use a contrastive learning objective (e.g., CLIP) or use image-grounded text generation loss (e.g., BLIP, BLIP2). The models used in this chapter are shown in Table 4.1. They are different in the above aspects, allowing investigation into the effect of their designs and training data on the cross-modal associations that are potentially learned. In addition, we include GPT-4o; even though no information is available for this model, its generative performance is unprecedented.

4.3 Methods

To test for the presence of a bouba-kiki effect in VLMs, we employ previously used as well as newly generated images (Section 4.3.1) and use a method for constructing pseudowords (4.3.2) that is directly borrowed from Nielsen and Rendall (2013). Probing (Section 4.3.3) was used to obtain image-text scores, and responses were analysed in two ways (Section 4.3.4).[1]

4.3.1 Image selection and generation

The original set of images used by Köhler (1929, 1947), as shown in figure 4.1, has been expanded in subsequent experiments. Maurer et al. (2006) for example introduced additional line drawings and Westbury (2005) used images with white shapes on a black background. Here we use the original pair and the two sets of four image pairs by Maurer et al. (2006); Westbury (2005). In

[1] All code, materials, and data are available on OSF: `https://osf.io/3w7k9/`.

addition, we generated new random curved and jagged images using a method inspired by Nielsen and Rendall (2013). We generated 10 uniformly distributed points within a circle with a radius of 1. These points were connected with either smooth curves or straight lines. For curved images, we generated curves that pass through the given points such that they form a closed path. Jagged images were generated by connecting the ordered points with straight lines, also forming a closed path. All images are displayed in Section A.1.

4.3.2 Pseudoword generation

Following the experiment with human participants conducted by Nielsen and Rendall (2013), we present the VLMs with a constrained set of syllables that can be used to construct novel pseudowords. Based on previously established cross-modal association patterns, Nielsen and Rendall (2013) selected sets of vowels and consonants that were expected to evoke a sense of correspondence with either jagged or curved visual shapes. We adopt exactly their set here, consisting of sonorant consonants M, N and L and rounded vowels OO, OH and AH, expected to match to curved shapes, and plosive consonants T, K and P and non-rounded vowels EE, AY and UH, expected to match to jagged shapes. Syllables were created by making consonant-vowel combinations. In total, 36 different syllables (e.g., loo, nah, kee, puh) can be constructed in this way, with nine different versions of each syllable type: sonorant-rounded (S-R), plosive-rounded (P-R), sonorant-non-rounded (S-NR) and plosive-non-rounded (P-NR).

In addition to single syllables, we generated pseudowords by concatenating two syllables, as this was exactly the task human participants were asked to complete in the experiment (Nielsen and Rendall, 2013). However, since we are not primarily interested here in distinguishing the separate roles played by consonants versus vowels in the bouba-kiki effect, and Nielsen and Rendall (2013) demonstrated that both have an impact, we limit the set of possible syllables in two-syllable probing to combinations of S-R syllables and P-NR syllables.

An important difference between the human setup and our work is that their participants also listened to a spoken version of the pseudowords, whereas our models are only exposed to the written form. Since the bouba-kiki effect is most often assumed to integrate vision and sound, this may influence the result. However, the relation between orthographic shapes and the sounds they represent is not arbitrary either and has presumably been shaped by human iconic strategies in their development and evolution (Turoman and Styles, 2017). This perhaps also explains why a role for English orthography has been demonstrated in the bouba-kiki effect for humans (Cuskley et al., 2017), while at the same time it is robust across different writing systems (Ćwiek et al., 2022).

4.3.3 VLM probing

To assess the preferences of BLIP2, CLIP, and ViLT, in each trial, we extract probabilities for all possible labels (i.e., syllables and pseudowords) conditioned on an image. Instead of only

embedding the label, each label is fed in a sentence ("The label for this image is {label}") such that embedding the textual input is closer to the models' natural objective[2]. Importantly, only the labels differ between inferences such that variance in the probability given an image is only caused by the label of interest. Where Alper and Averbuch-Elor (2023) use an *indirect* metric by embedding the inputs in CLIP space, our method uses the model probabilities as a more *direct* measure of how well a given syllable or pseudoword matches a novel image. For GPT-4o, we prompt the model to generate a label and use its probability directly (Appendix A.2).

4.3.4 Analysis

All findings were analysed for statistical significance using Bayesian models with the *brms* package (Bürkner, 2021) in R (R Core Team, 2023). To analyse VLM probability scores, we fitted Bayesian multilevel linear models (4 chains of 4000 iterations and a warmup of 2000, family = *gaussian*) to predict probability with image shape (Jagged versus Curved), consonant (plosive or sonorant) and vowel (rounded or non-rounded) categories ($Probability \sim shape * (consonant + vowel)$). For all models of this type, the random effects structure consists of varying intercepts for image and label with by-label random slopes for shape. When comparing proportions of vowels, consonants, or selected pseudoword types, we fitted Bayesian logistic models (4 chains of 1000 iterations and a warmup of 500, family = *binomial*) to test whether shape predicts the occurrence of particular vowels, consonants or pseudoword types ($Occurrence|trials(SampleSize) \sim Shape$). Effects are considered significant when the computed 95% Credible Interval does not include 0, i.e., the lower and upper bounds of the CI must be either both positive or both negative. All plots were created in ggplot2 (Wickham, 2016).

4.4 Results

The findings are analysed in two ways. First, we compare the results of VLM probing to the performance of human participants (Nielsen and Rendall, 2013). For BLIP2, CLIP and ViLT this means we first only consider the syllable or pseudoword with the highest probability for each image. These are then analysed similarly to those selected by humans or generated by GPT-4o. Second, we examine the probabilities for *each* possible syllable or pseudoword from BLIP2, CLIP and ViLT, to obtain a more comprehensive measure of cross-modal associations. For the GPT-4o results reported below, one image in the Jagged shape condition is consistently missing since it (top right image in Figure A.2 in Section A.1) was flagged as 'content that is not allowed by our safety system'.

[2]Additional analysis revealed that the overall results remain consistent even when only the label is provided.

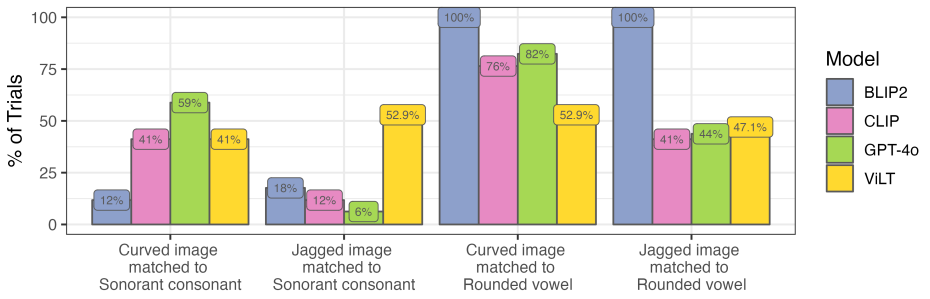

Figure 4.2: Percentages of trials in which selected syllables contain sonorant consonants or rounded vowels, separated by image shape (Jagged or Curved) for all four VLMs. Human percentages as reported by Nielsen and Rendall (2013) are (from left to right): 52.4%, 45.1%, 56,9%, and 48.3%.

4

4.4.1 Single syllable selection

VLMs were first probed using single syllables; here, we are interested in seeing if the models predominantly pair Jagged images with P-NR and Curved images with S-R syllables, as was found with humans. Figure 4.2 shows these results as the percentage of trials (where each individual image of the set of 17 pairs forms a trial) in which model probabilities were highest for sonorant consonants or rounded vowels with either Curved or Jagged shapes. A result that fits the expected human pattern would show higher bars for the Curved than for the Jagged shapes in both sets. The only models where this seems to go in the right direction are CLIP and GPT-4o. BLIP2 mostly displays a general preference for P-R syllables, without considering the shape, and ViLT does not display any clear preference. To test whether the differences in percentages for CLIP and GPT-4o are significant, we use Bayesian logistic models (as described in Section 4.3.4). For both models, Jagged images are paired with sonorant consonants significantly less often than Curved images (CLIP: b = -1.79, Bayesian 95 % Credible Interval $[-3.86, -0.05]$, GPT-4o: b = -3.51, 95 % CI $[-6.69, -1.37]$) and Jagged images are paired with rounded vowels significantly less often than Curved images (CLIP: b = -1.62, 95 % CI $[-3.06, -0.19]$, GPT-4o: b = -1.97, 95 % CI $[-3.66, -0.36]$).

4.4.2 Probability scores for novel syllables

While GPT-4o only selects the best-fitting syllable out of all options for each image, CLIP, BLIP2, and ViLT provide probability scores for each possible syllable, yielding more comprehensive data. Here, we therefore also analyse the probability scores for these three models to investigate whether higher scores occur when pairing S-R syllables with Curved images than with Jagged images and vice versa for P-NR syllables. Figure 4.3 shows the probabilities for the pseudoword

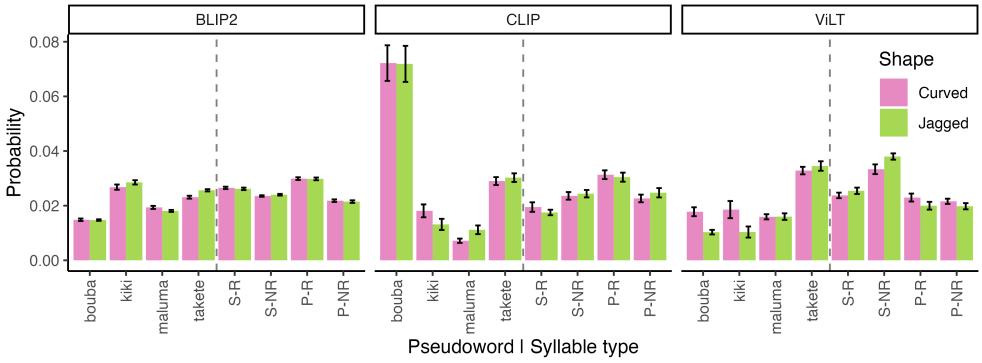

Figure 4.3: Probability scores for the original pseudowords (bouba, kiki, takete and maluma), as well as for the four different generated syllable types: Sonorant-Rounded (S-R), Sonorant-Non-Rounded (S-NR), Plosive-Rounded (P-R) and Plosive-Non-Rounded (P-NR), paired with two types of shapes (Jagged or Curved) for three VLMs.

pairs that were used in the classic experiments with humans (bouba & kiki, takete & maluma) and the four different syllable types (S-R, S-NR, P-R, P-NR).

Looking at the original pseudowords, none of the models display a clear bouba-kiki or takete-maluma effect. Probabilities for the different words differ overall (with a curiously high probability for 'bouba' in CLIP), but this does not seem modulated by the visual shape. For the syllables, BLIP2 shows no shape-modulated variation at all, and ViLT displays contradictory patterns (e.g. higher probability scores for S-NR than S-R syllables with Curved shapes and higher scores for S-NR with Jagged than with both P-R and P-NR). Only CLIP gets close to the expected pattern, with equal scores for the ambiguous syllable types (S-NR and P-R) but slightly higher scores for P-NR with Jagged and S-R with Curved. Yet, no significant effects are found when testing whether CLIP shows a pattern of preferring the expected consonants and vowels with their associated shapes using a Bayesian multilevel linear model (as described in Section 4.3.4). For ViLT, we find one (tiny) interaction between shape and consonants in the opposite direction of what is expected, where scores for Jagged shapes are significantly higher when paired with sonorant versus plosive consonants (b = .0056, 95 % CI $[.0001, .0112]$). For BLIP2, we find a significant overall preference for rounded vowels (b = 0.0055, 95 % CI $[.0019, .0091]$), but no other effects.

4.4.3 Two-syllable pseudoword selection

Although the results in Nielsen and Rendall (2013) were analysed by looking at single syllables, the actual task human participants performed involved creating novel pseudowords consisting of two syllables. We therefore also used our VLMs to generate (GPT-4o) or provide probability scores (CLIP, BLIP2 and ViLT) for two-syllable pseudowords that were created by concatenating

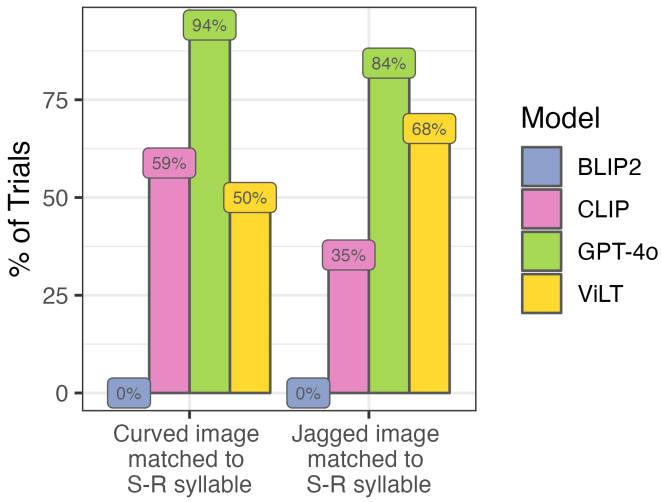

Figure 4.4: Percentages of trials in which Jagged or Curved visual shapes were matched to Sonorant-Rounded (S-R) syllables embedded in two-syllable pseudowords for all VLMs. Here 0% for S-R syllables implies a 100% preference for P-NR syllables.

two of the possible syllables from the set of S-R (most Curved) and P-NR (most Jagged) syllables resulting in 324 words. For CLIP, BLIP2 and ViLT, we first look at the 'preferred' pseudowords by only considering the option with the highest probability score for each image. Figure 4.4 shows the percentages of trials in which S-R syllables were matched to either Curved or Jagged images, counting each one of the two syllables in a word separately. BLIP2 never used S-R syllables and only selected pseudowords that contained two P-NR syllables, regardless of which image was shown. Both CLIP and GPT-4o show a higher percentage of Curved matched to S-R compared to Jagged, but GPT-4o seems to mostly just prefer S-R syllables overall. A manual inspection of GPT-4o's generated pseudowords revealed that in 25 out of 33 trials, the word 'nohmoh' was used, 12 times for Jagged and 13 times for Curved images. For ViLT, if a preference is present, it is in the wrong direction. In the case of CLIP, we find that Jagged images are indeed paired with S-R syllables significantly less often than Curved images (b = -1.00, 95 % CI $[-2.04, -0.04]$).

4.4.4 Probability scores for novel two-syllable pseudowords

We obtained probability scores for all possible two-syllable pseudowords when paired with each image for CLIP, BLIP2 and ViLT. Figure 4.5 shows these results by plotting probabilities for four different pseudoword types. The pseudoword on the left combines two P-NR syllables and is therefore expected to result in higher probabilities for Jagged shapes. Conversely, the

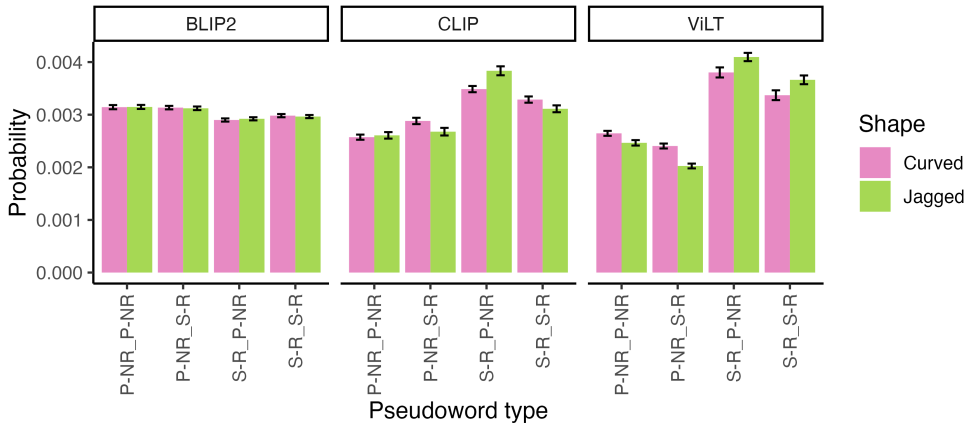

Figure 4.5: Probability scores for four pseudoword types, combining Sonorant-Rounded (S-R) and Plosive-Non-Rounded (P-NR) syllables, paired with two types of shapes (Jagged or Curved) for three VLMs.

most right pseudoword combines two S-R syllables and should evoke higher probabilities for Curved shapes. In the latter case, a pattern in which pink (Curved) bars rise while green (Jagged) bars fall would therefore reflect evidence for the bouba-kiki effect. None of the tested VLMs fit this pattern. Since GPT-4o generated 'nohmoh' (and similar variants like 'moomoh') almost exclusively when given the freedom to select two syllables from the full set of Jagged-associated and Curved-associated syllables, we also independently obtained probabilities for both syllable types. For this, we asked GPT-4o to generate a pseudoword for each image twice, once when given only the set of Jagged-associated syllable options, and once with only the Curved-associated syllables as options. Yet again, no significant effect of shape on probability scores for different syllable types was found. Figure A.4 in Section A.3 shows this result.

4.4.5 Summary

In summary, the bouba-kiki effect appeared absent for BLIP2 and ViLT, while for CLIP and GPT-4o, the results varied depending on how the effect was tested and the results were analysed. When asking the model to select one best-fitting syllable, CLIP and GPT-4o both display the effect in the expected direction. However, this pattern disappears when looking at a richer dataset of probability scores (from CLIP, BLIP2, and ViLT) for each possible syllable. In the case of two-syllable words, GPT-4o results no longer display significant evidence for a bouba-kiki effect.

4.5 Discussion

Our findings partly contradict previous work, which found that sound-symbolic associations are present in CLIP and Stable Diffusion (Alper and Averbuch-Elor, 2023). A possible reason for this could be that we use a different method, focusing on image-to-text probabilities, which is more similar to how the effect has been tested with humans. If VLMs indeed learned human-like cross-modal associations, we should be able to observe them in various experimental setups, i.e., the results should be robust. Given the contradicting findings, we suggest that it is too early to conclude that VLMs understand sound-symbolism or map visio-linguistic representations in a human-like manner, as the results depend heavily on which specific model is tested and how the task is formulated.

The asymmetry between the results coming from our method and those of Alper and Averbuch-Elor (2023) implies that performance is influenced by the method used. But perhaps more urgently, there is also contradicting evidence *within* the same method. In a replication of Alper and Averbuch-Elor (2023)'s experiment for Japanese, Iida and Funakura (2024) found that Japanese VLMs did not exhibit the expected bouba-kiki effect, despite Japanese being a language rich in sound-symbolism (Dingemanse, 2012; Ćwiek et al., 2022). Hence, Kouwenhoven et al. (2025) suggest 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. This is unsurprising given that CNN-based models often classify based on superficial textural rather than shape features (Baker et al., 2018; Geirhos et al., 2019; Hermann et al., 2020) and, albeit less so, this texture bias is also present in vision transformers (Geirhos et al., 2021). Moreover, Darcet et al. (2024) identified that, during inference, ViT networks create artefacts at low-informative background areas of images that are used for computations rather than describing visual information. Both findings are in stark contrast with what, at its core, is required for sound symbolism. However, the fact that some evidence for a bouba-kiki effect could be found in two of the four models tentatively suggests that real-world physical experience with different object properties may not be needed to develop this cross-modal preference but that it can, to some extent, be learned from statistical regularities in data containing text and images.

Human language on its own already contains many non-arbitrary regularities between speech sounds and meaning (Blasi et al., 2016), and these regularities, like phonesthemes (Bergen, 2004), can be detected and interpreted by models such as word embeddings (Abramova and Fernández, 2016) and LSTM-based language models (Pimentel et al., 2019). No visual input is needed for this, and perhaps this is also what caused the appearance of the observed bouba-kiki effect in the work by Alper and Averbuch-Elor (2023). In our work, we gave more prominence to the visual input and found much less convincing evidence for the effect. With two complementary methods closely modelled after human experiments, Kouwenhoven et al.

(2025) attempted to interpret the visual attention patterns of two variants of CLIP in a shape-word matching task. Neither of the models showed performance in line with the expected associations. Direct comparison with prior human data on the same task, additionally, showed that the models' responses fall markedly short of the robust, modality-integrated behaviour characteristic of human cognition. Finally, qualitatively, they showed that both CLIP variants do not focus on sharp edges or round attributes of images, but instead mostly focus on the centres of shapes or background areas. Both observations are in contrast with what, at its core, is required for a bouba-kiki-like effect.

Regarding the design features of the models we tested, we observe that the model with the best bouba-kiki alignment to human preferences, CLIP, is also trained on the largest amount of data (comparing the three models we have information about, not including GPT-4o). This finding aligns with previous work showing that dataset properties affect alignment with human representations (Conwell et al., 2023; Muttenthaler et al., 2023). However, despite having many more parameters than CLIP, BLIP2 does not show the effect. In addition, while both BLIP2 and CLIP use dual-stream architectures, only CLIP, which uses modality-specific attention mechanisms, displays some evidence of a bouba-kiki effect. Despite impressive performance on vision-language tasks, the Q-Former in BLIP2 apparently does not promote sound-symbolic associations. This is important knowledge for developing models with vision-language representations that align with those of humans. Especially since more aligned models show more robust few-shot learning (Sucholutsky and Griffiths, 2023) and promote more natural interactions between humans and machines (Chapter 1). Although we find modest evidence for a bouba-kiki effect in GPT-4o, we cannot know the origin of this effect as model details are unknown.

4.6 Conclusion

Given the pervasive role that cross-modal associations play in human linguistic processing, learning, and evolution, we tested for the presence of a bouba-kiki effect in four VLMs that differ along various dimensions such as architecture design, training objective, number of parameters, and input data. Evidence for this effect is limited, but not entirely absent, in the tested VLMs. These findings inform discussions on the origins of the bouba-kiki effect in human cognition and future developments of VLMs that align well with human cross-modal associations.

4.7 Limitations

Our work has a few notable limitations. First, we used synthetic images that had been previously used in experiments with humans. Even though this makes our results easily comparable to those of human studies, there is a potential risk that these images are out-of-domain for models

that are predominantly trained on realistic images. In future extensions of this work, we therefore plan to include more naturalistic images.

A second limitation manifests itself in the tokenisation of the textual input. While humans in the experiment evaluate pseudowords as a whole, the tokenisation process in language models may split our syllables or pseudowords into tokens that would not necessarily evoke the expected cross-modal associations in humans either (e.g., a separate evaluation of H in OH may invite a jagged association instead of a curved one). Despite being a fundamental difference, the primary goal of this chapter was to assess the preferences of VLMs in their most basic form. Further work should investigate whether tokenisation affects results and identify whether there may be model-specific cross-modal associations on a token instead of a word level.

Third, the pseudowords we used were based on an experiment with humans but were different from those used by Alper and Averbuch-Elor (2023), who did find a strong bouba-kiki effect in CLIP embeddings. To allow for a better comparison with their findings, future work should also test our image-to-text approach with their set of pseudowords.

Finally, our experiments included a relatively small number of trials, limited by the availability of experimental stimuli from human studies. However, by combining images from several previous studies and augmenting this set with additional newly generated images, we used more trials than most studies conducted with humans. The set of generated images can easily be expanded in future work. But then again, given the current pattern of results, this is not expected to lead to a more robust bouba-kiki effect in most models.