What does Kiki look like? Cross-modal associations between speech sounds and visual shapes in vision-and-language models

Tessa Verhoef*, **Kiana Shahrasbi, Tom Kouwenhoven***

Leiden Institute of Advanced Computer Science
Leiden University, The Netherlands
t.verhoef@liacs.leidenuniv.nl, k.shahrasbi@umail.leidenuniv.nl,
t.kouwenhoven@liacs.leidenuniv.nl

Abstract

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.

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 revolutionize 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 ac-
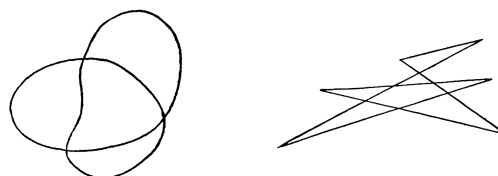

Figure 1: Which of these two shapes is Kiki? Images from Köhler (1929, 1947)

tively investigated (Alper et al., 2023; Kamath et al., 2023; Zhang et al., 2024b; Karamcheti 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 Köhler (Köhler, 1929, 1947), using the two images shown in Figure 1 with the labels *maluma* and *takete*. Since then it has been widely studied (as reviewed in Section 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, 2016; 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.

---

*Equal contribution

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 (Kouwenhoven et al., 2022). 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., 2023; Kouwenhoven et al., 2024). 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., 2022; Lian et al., 2023). 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, 2016; Tamariz et al., 2018; Little et al., 2017), the phenomenon is still 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 to find 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 this previous finding, that the bouba-kiki effect does not occur consistently in VLMs, and the presence of this cross-modal preference may depend on the way it is tested and properties like model architecture, attention mechanism, and training details.

2 Background

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 called 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 rated high in iconicity as well (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 analyzed 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 (Simner et al., 2010; Ramachandran and Hubbard, 2001; 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 1.

While the bouba-kiki effect may be the most famous example of a universal cross-modal association, many other cognitive biases in cross-modal perception have been reported. For example, 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 paper only scratch the surface of what is possible in this domain.

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 auditory properties of speech sounds, has also been studied (Cuskley et al., 2017; Bottini et al., 2019). Across set-ups, non-arbitrary preferences are found robustly 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.

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 behaviors 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 have trouble integrating 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.

Their approach involved generating two large sets of pseudowords, where one set was more likely associated with round shapes (examples: *bodubo,*

Model	Train objective	Architecture	Attention	#Params	#imgs,caps (M)
CLIP	CON	Dual-Stream	Modality-specific	151.3M	400, 400
ViLT	ITM&MLM	single-stream	Merged	87.4M	4.10, 9.85
BLIP2	CON&IGTG&ITM	Dual-stream	Q-Former	~3.8B	129, 258
GPT-4o	Unknown	Unknown	Unknown	Unknown	Unknown

Table 1: Overview of the models. Objectives are Image Text Matching(ITM), Masked Language Modelling(MLM), Image-grounded Text Generation(IGTG), or Contrastive Learning(CON). Numbers are millions(M) or billions(B).

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 analyzed 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 text-to-image generation (with Stable Diffusion) or 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. 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 suggest that dataset diversity and scale are the primary drivers of alignment to human representations (Conwell et al., 2023; Muttenthaler et al., 2023). We compare four models here, with different architectures, attention mechanisms, and training objectives.

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 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 modeling 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 paper are shown in Table 1. They are different in the above aspects, allowing investigation into the effect of their designs and input data on the bouba-kiki effect. In addition, we include GPT-4o; even though no information is available for this model, its generative performance is unprecedented.

3 Methods

To test for the presence of a bouba-kiki effect in VLMs we employ previously used as well as newly generated images (§3.1) and use a method for constructing pseudowords (3.2) that is directly borrowed from Nielsen and Rendall (2013). Probing (§3.3) was used to obtain image-text scores and responses were analyzed in two ways (§3.4).

3.1 Image selection and generation

The original set of images used by Köhler (1929, 1947), as shown in figure 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 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 Appendix A.

3.2 Pseudoword generation

Following the experiment conducted by Nielsen and Rendall (2013) with human participants, 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 effect, 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 set-up and our work, is that their participants also listened to a spoken version of the pseudowords, while 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).

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[1]. Importantly, only the labels differ between inferences such that variance in the probability given an image is caused by the label only. 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 B).

3.4 Analysis

All findings were analyzed for statistical significance using Bayesian models with the *brms* package (Bürkner, 2021) in R (R Core Team, 2023). To analyze 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

---

[1]Additional analysis revealed that the overall results remain consistent even when only the label is provided.
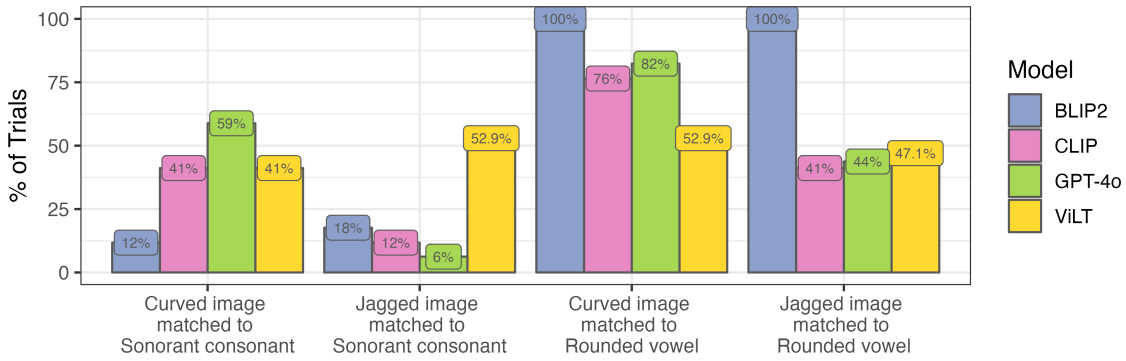
Figure 2: Percentages of trials in which selected syllables contain sonorant consonants or rounded vowels, separated by image shape (Jagged or Curved) for all four VLMs

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 the upper bounds of the CI have to be either both positive or both negative. All plots were created in ggplot2 (Wickham, 2016).

4 Results

The findings are analyzed 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 analyzed 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 7 in Appendix A) was flagged as 'content that is not allowed by our safety system'.

4.1 Single syllable selection

VLMs were first probed using single syllables, here we are interested to see if the models predominantly pair Jagged images with P-NR and Curved images with S-R syllables, as was found with humans. Figure 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 where 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 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 $[-3.66, -0.36]$).

4.2 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 analyze 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 3 shows the probabilities for the pseudoword 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
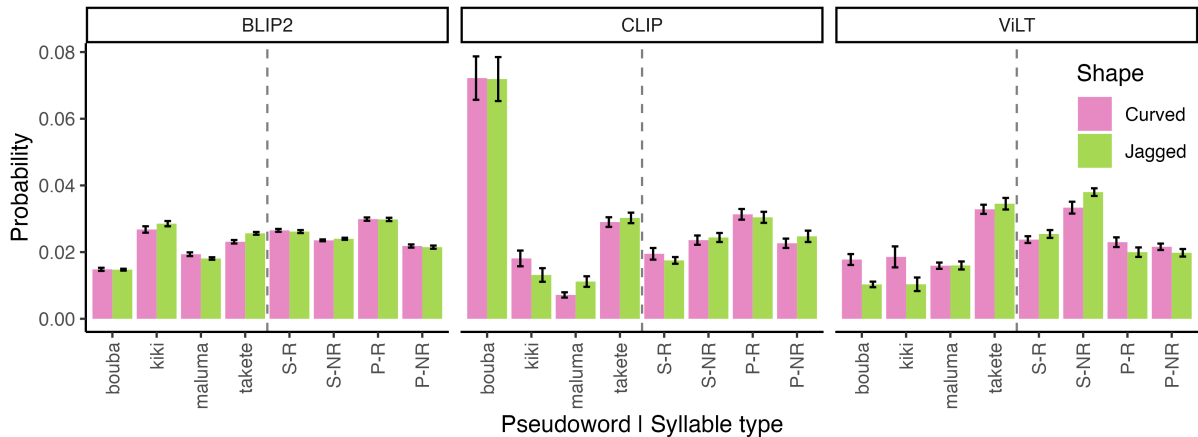
Figure 3: Probability scores for the original pseudowords (bouba, kiki, takete and maluma), as well as for the four different generated syllable types: Sonorant-Rounded (S-R), Sonorant-Non-Rounded (S-NR), Plosive-Rounded (P-R) and Plosive-Non-Rounded (P-NR), paired with two types of shapes (Jagged or Curved) for three VLMs

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 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.

4.3 Two-syllable pseudoword selection

Although the results in Nielsen and Rendall (2013) were analyzed 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 two of the pos-

sible 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 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, independently from 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 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 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 most right pseudoword
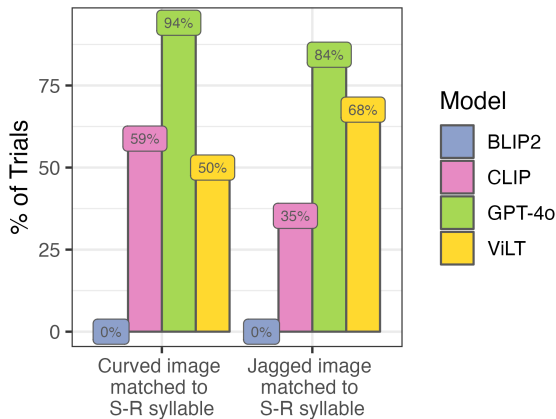
Figure 4: Percentages of trials in which Jagged or Curved visual shapes were matched to Sonorant-Rounded (S-R) syllables embedded in two-syllable pseudowords for all VLMs. Here 0% for S-R syllables implies a 100% preference for P-NR syllables.

combines two S-R syllables and should evoke higher probabilities for Curved shapes. 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 9 in Appendix C shows this result.

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 results were analyzed. 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.

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). We use a different method, focusing on image-to-text probabilities, which is more similar to how the effect has been tested with humans. We show that it is too early to conclude that VLMs understand sound-symbolism or map visio-linguistic representations in a human-like way since the results depend heavily on which specific model is tested and how the task is formulated. 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 artifacts 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 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.

Regarding the design features of the models we tested, we see 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 on, not including GPT-4o). This finding aligns with previous work showing that dataset properties affect
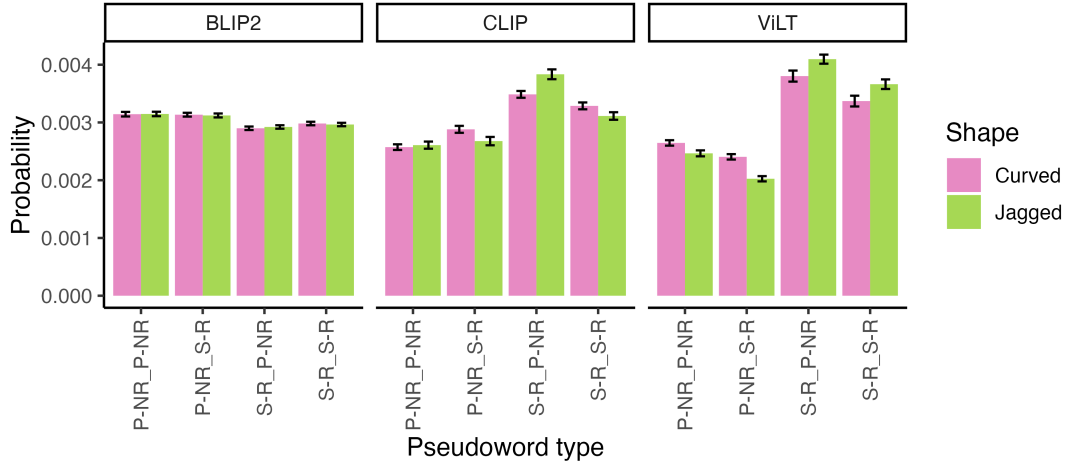
Figure 5: Probability scores for four pseudoword types, combining Sonorant-Rounded (S-R) and Plosive-Non-Rounded (P-NR) syllables, paired with two types of shapes (Jagged or Curved) for three VLMs

alignment with human representations (Conwell et al., 2023; Muttenthaler et al., 2023). However, despite having much 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. More aligned models show more robust few-shot learning (Sucholutsky and Griffiths, 2023) and promote more natural interactions between humans and machines (Kouwenhoven et al., 2022). 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.

6 Conclusion

Given the pervasive role 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 and 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.

7 Limitations

Our work has a few notable limitations. First, we used synthetic images that were 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 tokenization of the textual input. While humans in the experiment evaluate pseudowords as a whole, the tokenization 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 curved). Despite being a fundamental difference, the primary goal of this work was to assess the preferences of VLMs in their most basic form. Further work should investigate whether tokenization affects results and identify whether there may be model-specific cross-modal associations on a token instead of 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 available experimental stimuli from human studies. 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, though. The set of generated images can easily be expanded in future work. However, given the current pattern of results, this is not expected to lead to a more robust bouba-kiki effect in most models.

References

Ekaterina Abramova and Raquel Fernández. 2016. Questioning arbitrariness in language: a data-driven study of conventional iconicity. In *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.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 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.

Morris Alper and Hadar Averbuch-Elor. 2023. Kiki or bouba? sound symbolism in vision-and-language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 78347–78359. Curran Associates, Inc.

Morris Alper, Michael Fiman, and Hadar Averbuch-Elor. 2023. Is bert blind? exploring the effect of vision-and-language pretraining on visual language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6778–6788.

Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. 2018. Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, 14(12):1–43.

Benjamin K. Bergen. 2004. The psychological reality of phonaesthemes. *Language*, 80(2):290–311.

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

Roberto Bottini, Marco Barilari, and Olivier Collignon. 2019. Sound symbolism in sighted and blind. the role of vision and orthography in sound-shape correspondences. *Cognition*, 185:62–70.

Diane Bouchacourt and Marco Baroni. 2018. How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 981–985, Brussels, Belgium. Association for Computational Linguistics.

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

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. 2023. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*.

Christine Cuskley, Mark Dingemanse, Simon Kirby, and Tessa M Van Leeuwen. 2019. Cross-modal associations and synesthesia: Categorical perception and structure in vowel–color mappings in a large online sample. *Behavior research methods*, 51:1651–1675.

Christine Cuskley and Simon Kirby. 2013. 869Synesthesia, Cross-Modality, and Language Evolution. In *Oxford Handbook of Synesthesia*. Oxford University Press.

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

Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus, Gary Lupyan, Grace E. Oh, Jing Paul, Caterina Petrone, Rachid Ridouane, Sabine Reiter, Nathalie Schümchen, Ádám Szalontai, Özlem Ünal-Logacev, Jochen Zeller, Marcus Perlman, and Bodo Winter. 2022. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B*, 377(1841):20200390.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2024. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*.

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

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karla K Evans and Anne Treisman. 2010. Natural cross-modal mappings between visual and auditory features. *Journal of vision*, 10(1):6–6.

Mathilde Fort and Jean-Luc Schwartz. 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.

Lukas Galke, Yoav Ram, and Limor Raviv. 2022. Emergent communication for understanding human language evolution: What's missing? In *Emergent Communication Workshop at ICLR 2022*.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2021. Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems*, volume 34, pages 23885–23899. Curran Associates, Inc.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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.

Katherine Hermann, Ting Chen, and Simon Kornblith. 2020. The origins and prevalence of texture bias in convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 19000–19015. Curran Associates, Inc.

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

Timothy L Hubbard. 1996. Synesthesia-like mappings of lightness, pitch, and melodic interval. *The American journal of psychology*, pages 219–238.

Matsumi Imai, Sotaro Kita, Miho Nagumo, and Hiroyuki Okada. 2008. Sound symbolism facilitates early verb learning. *Cognition*, 109(1):54–65.

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

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

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore. Association for Computational Linguistics.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic VLMs: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

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

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

Tom Kouwenhoven, Max Peeperkorn, Bram van Dijk, and Tessa Verhoef. 2024. The curious case of representational alignment: Unravelling visio-linguistic tasks in emergent communication. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kouwenhoven, Tessa Verhoef, Roy De Kleijn, and Stephan Raaijmakers. 2022. Emerging grounded shared vocabularies between human and machine, inspired by human language evolution. *Frontiers in Artificial Intelligence*, 5:886349.

Angeliki Lazaridou and Marco Baroni. 2020. Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yuchen Lian, Arianna Bisazza, and Tessa Verhoef. 2023. Communication drives the emergence of language universals in neural agents: Evidence from the word-order/case-marking trade-off. *Transactions of the Association for Computational Linguistics*, 11:1033–1047.

Hannah Little, Kerem Eryılmaz, and Bart De Boer. 2017. Signal dimensionality and the emergence of combinatorial structure. *Cognition*, 168:1–15.

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

Ryan Lowe, Abhinav Gupta, Jakob Foerster, Douwe Kiela, and Joelle Pineau. 2020. On the interaction between supervision and self-play in emergent communication. In *International Conference on Learning Representations*.

Matéo Mahaut, Francesca Franzon, Roberto Dessì, and Marco Baroni. 2023. Referential communication in heterogeneous communities of pre-trained visual deep networks. *arXiv preprint arXiv:2302.08913*.

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

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

Robert D Melara. 1989. Dimensional interaction between color and pitch. *Journal of Experimental Psychology: Human Perception and Performance*, 15(1):69.

Lotte Meteyard, Emily Stoppard, Dee Snudden, Stefano F Cappa, and Gabriella Vigliocco. 2015. When semantics aids phonology: A processing advantage for iconic word forms in aphasia. *Neuropsychologia*, 76:264–275.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. 2023. Human alignment of neural network representations. In *The Eleventh International Conference on Learning Representations*.

Alan Nielsen and Drew Rendall. 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.

Alan Nielsen and Drew Rendall. 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.

Alan Nielsen and Drew Rendall. 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.

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

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

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

Tiago Pimentel, Arya D. McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764, Florence, Italy. Association for Computational Linguistics.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

David S Schmidtke, Markus Conrad, and Arthur M Jacobs. 2014. Phonological iconicity. *Frontiers in psychology*, 5:80.

Julia Simner, Christine Cuskley, and Simon Kirby. 2010. What sound does that taste? cross-modal mappings across gustation and audition. *Perception*, 39(4):553.

Shane Steinert-Threlkeld, Xuhui Zhou, Zeyu Liu, and C. M. Downey. 2022. Emergent communication fine-tuning (EC-FT) for pretrained language models. In *Emergent Communication Workshop at ICLR 2022*.

Ilia Sucholutsky and Tom Griffiths. 2023. Alignment with human representations supports robust few-shot learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 73464–73479. Curran Associates, Inc.

Mónica Tamariz, Seán G Roberts, J Isidro Martínez, and Julio Santiago. 2018. The interactive origin of iconicity. *Cognitive Science*, 42(1):334–349.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 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.

Nora Turoman and Suzy J Styles. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tessa Verhoef, Simon Kirby, and Bart De Boer. 2016. Iconicity and the emergence of combinatorial structure in language. *Cognitive science*, 40(8):1969–1994.

Tessa Verhoef, Seán G Roberts, and Mark Dingemanse. 2015. Emergence of systematic iconicity: Transmission, interaction and analogy. In *37th Annual Meeting of the Cognitive Science Society (CogSci 2015)*, pages 2481–2486. Cognitive Science Society.

Jamie Ward, Brett Huckstep, and Elias Tsakanikos. 2006. Sound-colour synaesthesia: To what extent does it use cross-modal mechanisms common to us all? *Cortex*, 42(2):264–280.

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

Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.

Yi Zhang, Ce Zhang, Yushun Tang, and Zhihai He. 2024b. Cross-modal concept learning and inference for vision-language models. *Neurocomputing*, 583:127530.

Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13785–13795.

A Full set of images

This appendix presents the full set of images with visual shapes that were used in the experiments. Besides the original image pair from Köhler (1929, 1947) which was shown in Figure 1, we used four image pairs from Maurer et al. (2006), displayed in figure 6, four from Westbury (2005), displayed in figure 7, and 8 additional pairs we newly generated using a method inspired by the one described by Nielsen and Rendall (2013), displayed in Figure 8. For each image pair, the Curved version is displayed on the left and the Jagged version on the right.
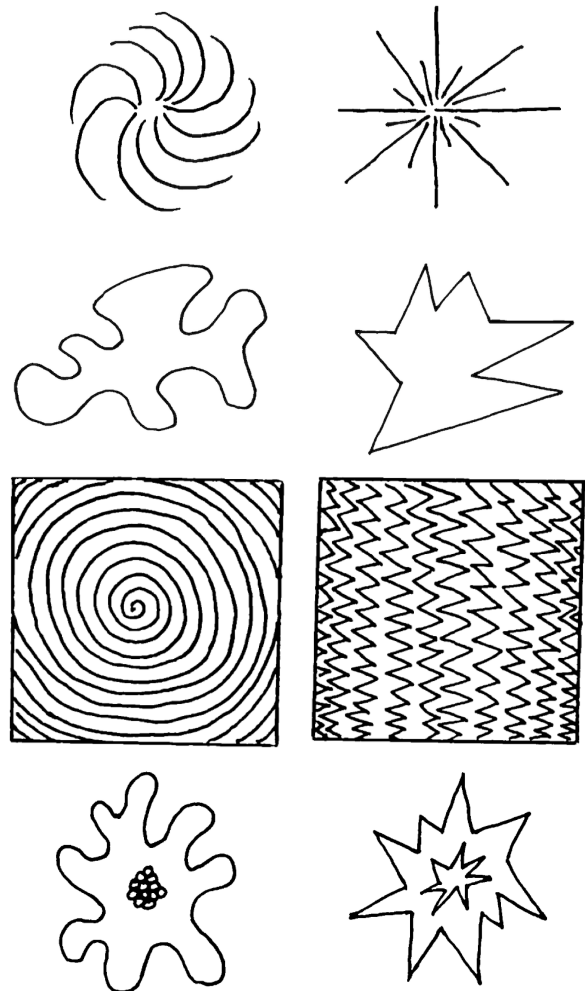

Figure 6: Images from (Maurer et al., 2006)

B GPT-4o prompting

Image-label matching is not directly possible for GPT-4o since the probabilities of the input tokens cannot be accessed. We therefore prompt (B.1) this model, with the temperature being 0.0, to generate a syllable or pseudoword given an image and use the log probabilities of the generated tokens to calculate the probability for a label conditioned on an image. Just like in the sentence setup used in the other models, our interest lies not primarily in the variability that may arise from using different prompts but rather focuses on the influence of the image on the predictions by using a simple and effective prompt that is identical for each image. Doing so allows us to use the resulting probabilities as a gauge for the models' preference of a label for a given image.
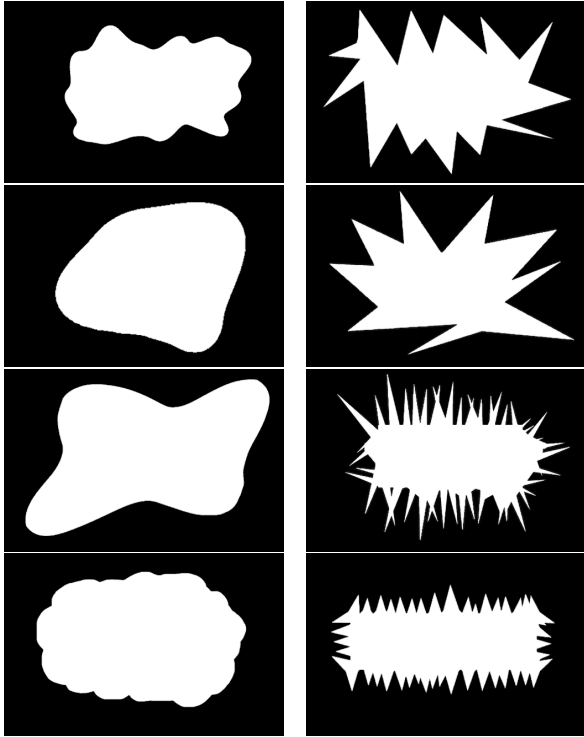
Figure 7: Images from (Westbury, 2005)

```
You are given an image for which you
need to assign a label. Use {one/two} of
 the following labels: {possible_labels
}. Only respond with the label.
```

Prompt B.1: The exact prompt used to obtain GPT-4o probabilities. *possible_labels* corresponds to the syllables of interest.

C   GPT-4o pseudoword probabilities

In section 4.4 we describe results for an experiment in which 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. Figure 9 shows the probabilities associated with these generated pseudowords. As concluded in the main text, no evidence for a preference to match P-NR syllables with Jagged shapes and S-R syllables with Curved shapes was found.
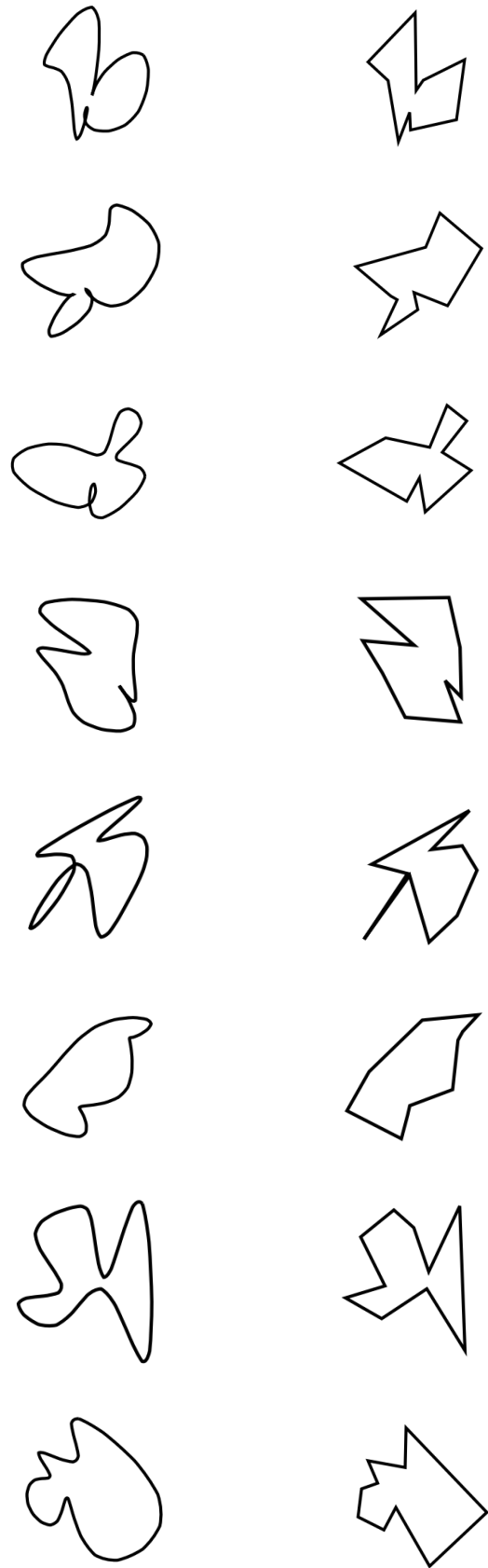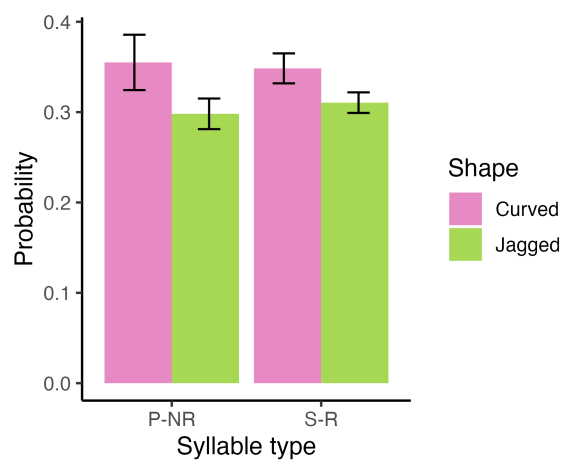

Figure 8: Newly generated images

Figure 9: Probability scores for GPT-4o when forced to generate a pseudoword for each image twice, once by combining two Jagged-associated syllables, and once with only the Curved-associated syllables as options.