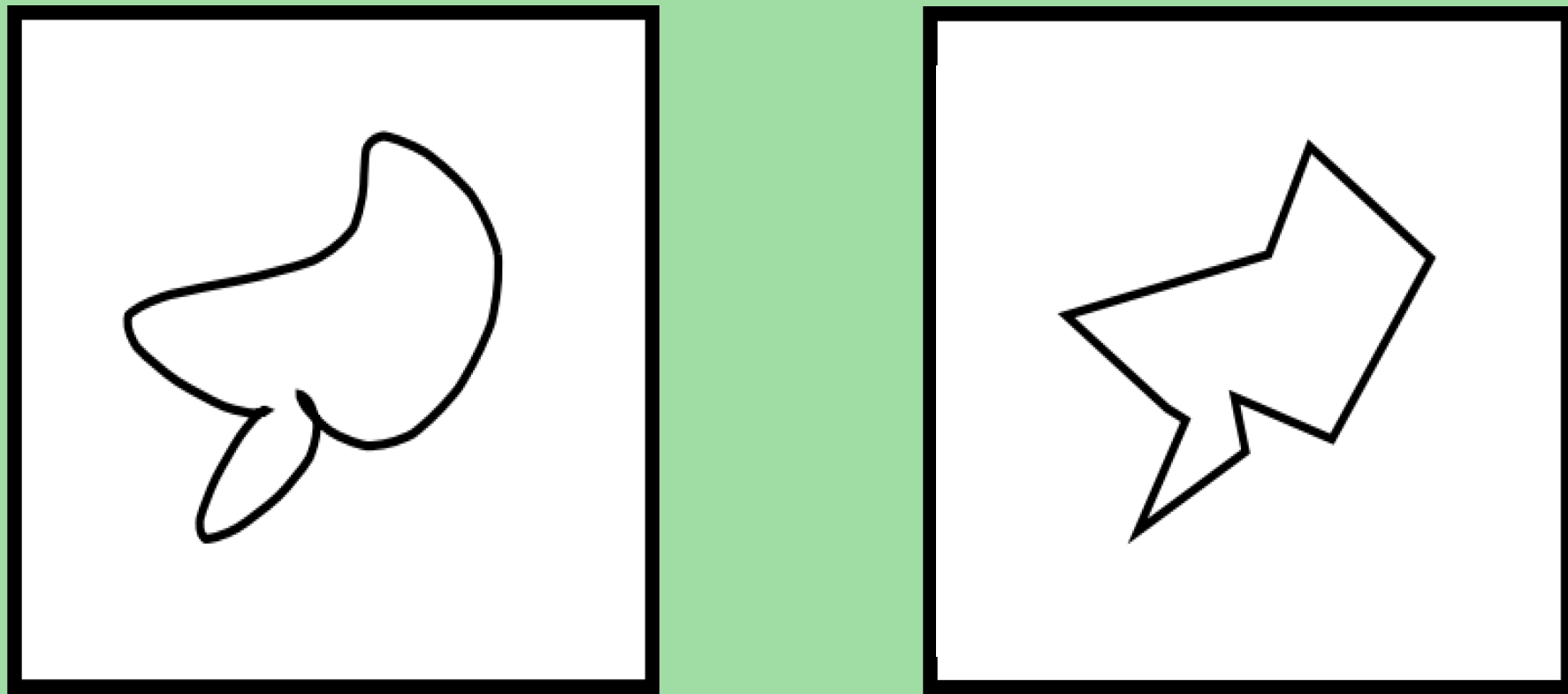


Tom Kouwenhoven, Kiana Shiarasbi, Tessa Verhoef

Let’s play a game



Which of these images is called ‘kiki’?

Motivation

Humans consistently integrate visual and auditory modalities. The most well-studied test case in the domain of **cross-modal associations** is the **bouba-kiki** effect. Humans reliably associate pseudowords like ‘bouba’ with round shapes and ‘kiki’ with jagged ones. Training models to develop human-like prior expectations could improve learning efficiency and enable natural interactions. Given recent advances in multimodal vision and language models (VLMs) and mixed evidence in prior studies [1,2], this work revisits the domain of cross-modal associations and asks:

- » Do vision and language models exhibit a human-like bouba-kiki effect?
- » Why do vision and language models perform the way they do?

Contributions

- » Tested as **close** as possible to **human experiments**, we find that **models** (two CLIP versions), compared to humans, **do not make consistent cross-modal associations**.
- » **Two methods** to test cross-modal associations, like the bouba-kiki effect, in VLMs. Enabling **robust direct** comparisons between our results and human findings.
- » A novel test using Grad-CAM [3], to unravel the **visual processing** in the bouba-kiki context. Revealing VLMs do **not attend** to bouba-kiki related **shape-specific features**.

Stimuli

English Adjectives:
Sharp , Spikey, ... , Curved , Round

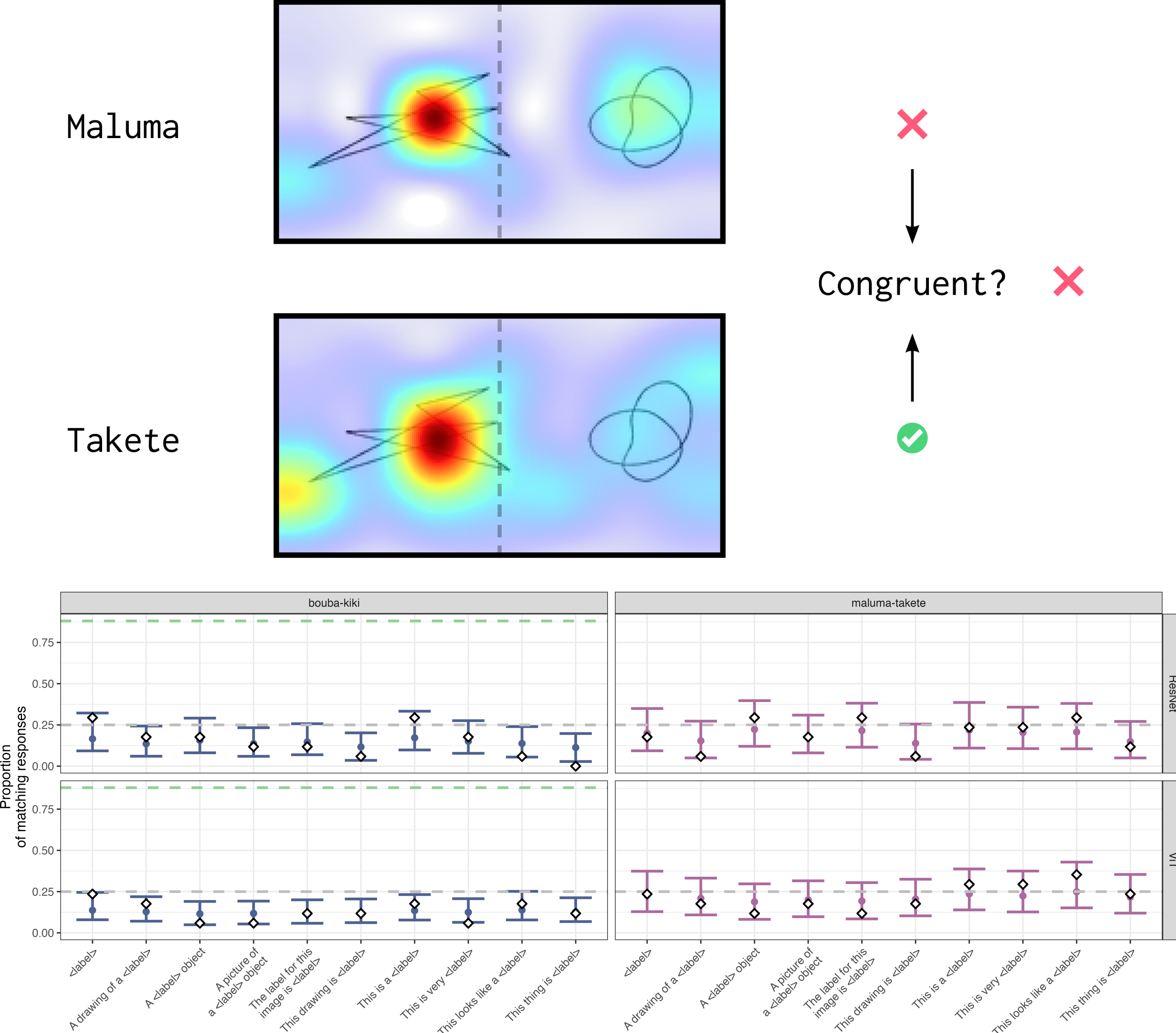
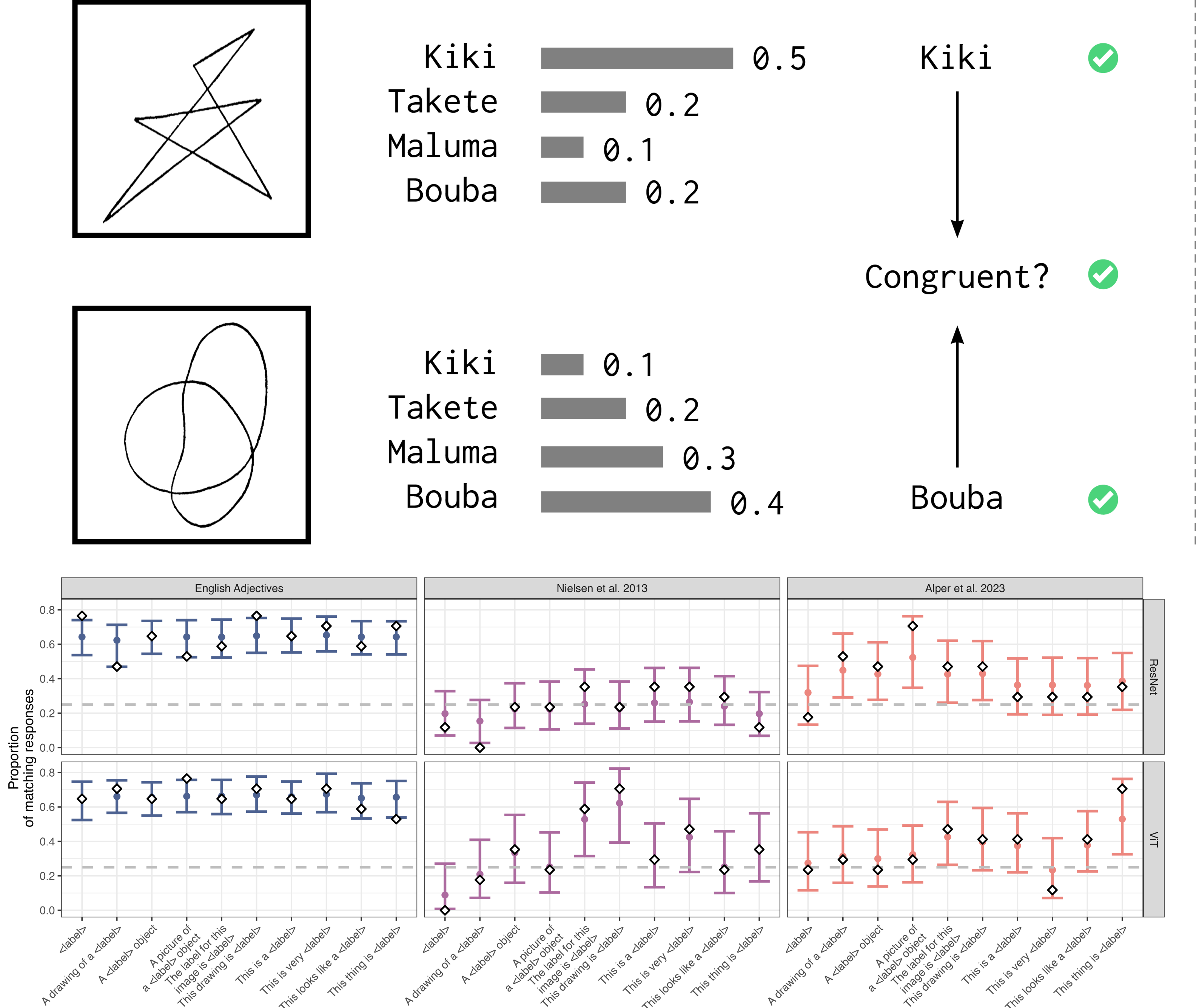
Pseudowords:
Nielsen et al. 2013:
‘kaytee’, ‘peetay’, ... , ‘lahmoh’, ‘noomah’

Alper et al. 2023:
‘kitaki’, ‘pikepi’, ... , ‘bodubo’, ‘numono’

Images (pairs) sourced from [1, 4, 5]:

Prompts:
‘The is a <label>’, ‘A drawing of a <label>’

Methods & Results



[1] Verhoef et al.,2024. What does Kiki look like? Cross-modal associations between speech sounds and visual shapes in vision-and-language models.
[2] Alper et al.,2023. Kiki or bouba? sound symbolism in vision-and-language models.
[3] Selvaraju et al.,2020. Grad-cam: Visual explanations from deep networks via gradient-based localization.
[4] Maurer et al.,2006. The shape of boubas: Sound-shape correspondences in toddlers and adults.
[5] Westbury et al.,2005. Implicit sound symbolism in lexical access: Evidence from an interference task.
[6] Cwiek et al.,2022. The bouba/kiki effect is robust across cultures and writing systems.

While English adjectives are classified correctly, CLIP **fails** to **robustly match pseudowords** associated with round or sharp shapes with respective **visual shape-features**.

Comparison to humans [6] shows CLIP does not make human-like associations. **Attention** patterns reveal labels **do not steer** models to **shape-features**. See our paper for additional results.