Vision-Language Model through the lens of Linguistic Relativity

Rahul Singh Maharjan & Angelo Cangelosi
Centre of Robotics and AI
University of Manchester, United Kingdom
rahulsingh.maharjan@manchester.ac.uk

Abstract

Linguistic relativity proposes that the categorical structure of a language systematically biases how continuous perceptual domains are partitioned. While this effect is well studied in human color perception, it remains unclear whether vision–language models (VLMs) exhibit a language-conditioned categorical structure analogous to it. In this work, we treat VLM as an *artificial subject* and adapt psycholinguistic paradigms to test for Whorfian effects in a multimodal neural network. Using a tightly controlled set of color-discrimination triads that vary only in OKLab lightness, we quantify model decisions across English and Russian lexical frames. For a contrastive dual-encoder VLM, we observe a robust shift in the internal decision boundary between light and dark blue when the language of the textual input changes, despite identical visual input. This boundary displacement disappears in vision-only embeddings, demonstrating that the effect is not perceptual in origin but emerges from vision–language alignment. In contrast, when languages share the same color lexicon (e.g., English and French), category boundaries largely overlap, and when generic labels are used, no stable boundary emerges. Together, these findings show that a computational analogue of linguistic relativity can arise in artificial systems, but only when visual representations are explicitly aligned with language. Our results establish a framework for testing cognitive theories in multimodal models.

Keywords: linguistic relativity; categorical perception; color semantics; vision–language models; multimodal cognition; computational psycholinguistics

Introduction

Linguistic Relativity (Sapir, 1929) is the principle that the grammatical structures, semantic categories, and habitual expressions of a language systematically bias how its speakers perceive, categorize, and remember experience, especially in domains where the world is continuous or ambiguous (such as color, time, and space). As such, contemporary work on linguistic relativity holds that language does not strictly determine cognition, but it systematically biases patterns of attention, categorization, and discrimination (Boroditsky, 2001; Lucy, 1992; Lupyan et al., 2020; Sapir, 1929; Winawer et al., 2007).

This effect has been demonstrated most clearly in the domain of color perception. Winawer et al. (2007) compared Russian and English speakers in a color discrimination task. Russian divides the color blue into two basic lexical categories - голубой *(goluboy)* (light blue) and синий *(siniy)*(dark blue) - whereas English uses a single term, *blue*. In Winawer et al. (2007), participants were asked to identify which of the two bottom color squares matched a top target square. Although the physical color distance was controlled, Russian speakers were significantly faster when the two choices crossed their language's categorical boundary (голубой *(goluboy) vs.* синий *(siniy)*) than when both belonged to the same lexical category. English speakers showed no such difference. Crucially, when Russian participants were given a concurrent verbal interference task, the categorical advantage disappeared, indicating that linguistic resources were actively engaged during perception (Gilbert et al., 2006). Similarly, Forder and Lupyan (2019) argues that hearing a color word can immediately bias visual color discrimination.

Parallel effects have been found outside of color. Boroditsky (2001) showed that speakers of English and Mandarin, which differ in how they spatially metaphorize time (horizontal vs. vertical), display corresponding differences in temporal reasoning. Similarly, Levinson (2003) demonstrated that speakers of languages that use absolute spatial frames of reference (e.g., north/south rather than left/right) exhibit superior orientation abilities and encode spatial relations differently in memory. Together, these findings support a moderate version of linguistic relativity: language does not impose rigid cognitive limits, but it guides default interpretive strategies by highlighting some distinctions while backgrounding others. The influence of language is therefore probabilistic and context-sensitive rather than deterministic.

Recent advances in artificial intelligence have introduced a new class of models called Vision-Language Models (VLMs), which jointly learn from visual and textual data to perform tasks that require multimodal reasoning, such as image captioning, visual question answering, and cross-modal retrieval (Jia et al., 2021; Li et al., 2023; Liu et al., 2023; Radford et al., 2021). This training paradigm implicitly introduces linguistic structure into the visual representation itself: objects, attributes, and relations are encoded not only by visual similarity but also by their lexical and semantic categories. Consequently, VLMs do not possess a purely perceptual vision system; rather, their internal representations reflect language-conditioned perceptual organization (Bisk et al., 2020; Goh et al., 2021). Furthermore, Marjieh et al. (2024) argued that Large-Language Models (LLMs) encode substantial structure about human perception despite the lack of direct sensory experiences.

This design choice invites a deeper inquiry: **Do VLM ex-**

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

hibit a form of linguistic relativity analogous to that observed in humans? If human perception is biased by the categorical structure of language, then VLMs—whose visual representations are optimized to align with linguistic tokens—may display similar category-dependent biases. Since color is both perceptually continuous and linguistically discrete (Berlin & Kay, 1991), it provides an ideal testbed for examining whether linguistic categories shape VLM behavior in a manner consistent with the linguistic relativity hypothesis.

Despite growing interest in the interpretability and cognitive plausibility of multimodal models, relatively little work has examined whether their internal representations exhibit language-specific categorical effects rather than merely reflecting visual similarity. Recent computational studies suggest that neural models can develop categorical color representations that resemble human color categories (Akbarinia, 2025; Arias et al., 2025; Yuksekgonul et al., 2022), but it remains unclear whether such effects in VLMs arise from vision-only structure or from language-conditioned category boundaries. Distinguishing between these possibilities is crucial for determining whether any observed categorical behavior constitutes a genuine analogue of linguistic relativity or merely an artifact of visual clustering. Investigating linguistic relativity in VLMs is therefore important for three reasons. First, it provides a framework for analyzing how language shapes internal perceptual representations in artificial systems. Second, it allows direct comparison between human and model cognition using an established psycholinguistic paradigm. Third, it clarifies whether multimodal models encode meaning in a fundamentally symbolically biased manner rather than in a perceptually grounded manner.

In the present study, we adapt classic color-categorization paradigms from the linguistic relativity literature (following Winawer et al. (2007)) to a controlled multimodal evaluation of VLM. By testing whether model performance differs across language-specific color boundaries (e.g., Russian голубой *(goluboy)* / синий *(siniy)* vs. English *blue*) under matched perceptual conditions, we assess whether VLMs exhibit language-conditioned categorical biases analogous to those found in human cognition.

We test whether category responses differ across languages for identical stimuli, and whether any such effect disappears in vision-only embeddings. In this study, we restrict our analysis to a contrastive dual-encoder model, using Jina-CLIP v2 Koukounas et al. (2024) as a representative model, to test whether language-conditioned category structure emerges directly in aligned embedding spaces.

Methods

Stimuli Generation

Color stimuli were generated in OKLab color space (Ottosson, 2020) to ensure approximately perceptual uniformity across lightness steps. Patches were sampled from a restricted blue region of color space, with lightness varied systematically and hue and chroma constrained to narrow ranges

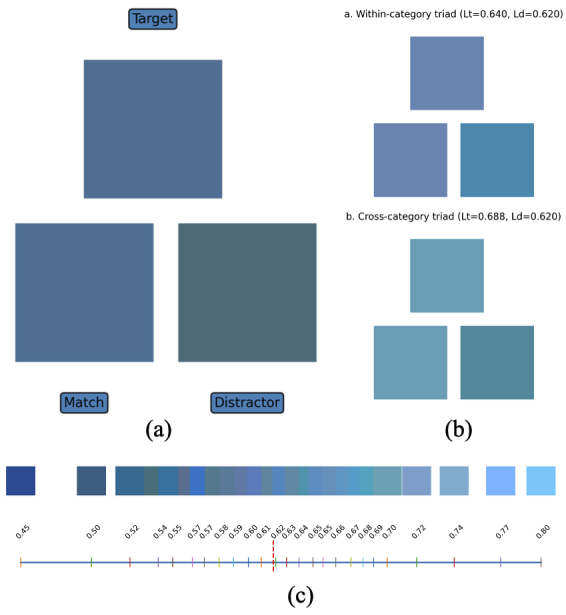

Figure 1: (a) Each trial presents a target (top), an identical match (bottom-left), and a distractor (bottom-right) differing only in OKLab lightness. (b) Example within-category (top) and cross-category (bottom) triads with equal physical lightness difference. Only the cross triad spans the reference boundary. (c) Lightness continuum (L^* = 0.45–0.80) used for stimulus sampling. The dashed line marks the reference boundary for labeling; it was not shown to the models.

($L^* \in [0.45, 0.80]$, $C \in [0.06, 0.16]$, $h \in [210°, 265°]$). Hue and chroma were randomly jittered within this subspace to avoid pixel-level artifacts, but only lightness (L^*) was used for analysis. This design isolates categorical effects along a single continuous perceptual axis while holding other visual properties approximately constant. Each trial consisted of a triad (Figure 1A) comprising a *target*, a physically identical *match*, and a *distractor* differing from the target primarily in lightness. Perceptual distance between target and distractor was calibrated using ΔE_{2000} (Fairchild, 2013), and all patches were embedded on a uniform neutral white background to minimize contrast effects.

To construct category conditions, we defined a reference lightness $L_{\text{ref}} = 0.62$ approximating the Russian blue boundary (Winawer et al., 2007). This boundary was used only for labeling, not imposed on the model: triads were labeled *within* if L_t and L_d fell on the same side of L_{ref}, and *cross* otherwise. The difference in physical lightness was held constant across conditions. The full stimulus set comprised 600 triads per language condition, rendered as 224×224 RGB images.

This controlled sampling strategy ensures that any observed categorical effects cannot be attributed to low-level color statistics or clustering in hue–chroma space, but must arise from how the model maps continuous visual input onto language-conditioned semantic categories.

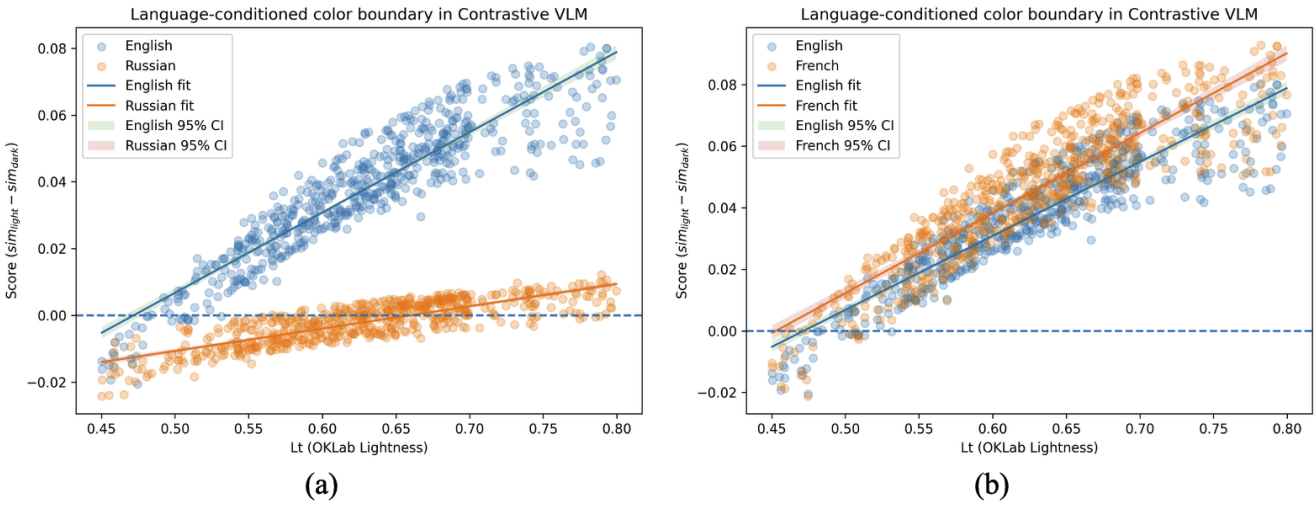

Figure 2: **Language-conditioned color boundaries in a contrastive VLM.** Category score as a function of OKLab lightness for identical stimuli under different language framings. Solid lines show linear fits; shaded bands indicate 95% confidence intervals; the dashed line marks the zero-crossing boundary. **(a)** English vs. Russian: the Russian boundary occurs at higher lightness. **(b)** English vs. French: boundaries largely overlap, showing no shift.

Formalization of model scores and category boundaries

We denote a visual stimulus by x and a language context by L (e.g., English or Russian). If linguistic relativity holds for a VLM, then its response function satisfies:

$$f(x, L) \neq f(x, L'), \tag{1}$$

for two distinct languages $L \neq L'$ given identical visual input x.

To test linguistic relativity in VLM, we require a way to convert each model's response to a color stimulus into a continuous scalar that reflects its categorical bias. Rather than relying solely on discrete labels, we define a category score that measures the model's relative preference for competing lexical categories (e.g., light vs. dark). This score serves as a psychometric analogue of response probability in human experiments and allows us to estimate an internal decision boundary.

By modeling the relationship between category score and physical lightness, we obtain a continuous psychometric function whose zero-crossing defines the model's implicit category boundary. This procedure mirrors psychophysical approaches in which perceptual thresholds and category transitions are inferred from fitted response curves rather than from single-trial judgments (Gescheider, 2013; Hautus et al., 2021). Similar curve-based boundary estimation has been widely used in studies of categorical perception of color and speech (Goldstone, 1994; Harnad, 2003; Winawer et al., 2007). This formulation enables direct quantitative comparison of category boundaries across languages.

The same framework can be applied at multiple representational levels—joint vision–language embeddings and vision-only embeddings —allowing us to distinguish language-conditioned categorical structure from intrinsic perceptual organization. The formalization provides a bridge between human psychophysics and artificial cognition.

Stimulus categories. Each trial is a triad with a target patch of OKLab lightness L_t, a physically identical match, and a distractor patch of lightness L_d. We define *within* vs. *cross* categories relative to a *reference* boundary L_{ref} that is used *only* to label stimuli for analysis and visualization; it is not provided to the model and does not constrain the model's fitted boundary. This mirrors standard categorical-perception designs in which a hypothesized category boundary is used to construct matched stimulus pairs, while the observer's boundary is inferred from responses (Gescheider, 2013; Harnad, 2003; Winawer et al., 2007). Formally,

$$\text{condition} = \begin{cases} \text{within}, & (L_t \geq L_{\text{ref}}) = (L_d \geq L_{\text{ref}}), \\ \text{cross}, & \text{otherwise}. \end{cases} \tag{2}$$

This construction is motivated by the human color studies, which show that discretizing a continuous color space into lexical categories (e.g., basic color terms) can induce categorical advantages around language-specific boundaries (Berlin & Kay, 1991; Winawer et al., 2007). Recent machine learning studies also suggest that neural models can develop category-like color representations, and that these representations shift when the models are aligned across modalities (Akbarinia, 2025; de Vries et al., 2022).

Contrastive model score. For a contrastive VLM (CLIP-style), we embed the stimulus image x into an image vector e_x and embed language-specific labels into text vectors $e_{\text{light}}^{(L)}$

and $e^{(L)}_{\text{dark}}$. Contrastive VLMs are trained to align matched image–text pairs in a shared embedding space using cosine similarity, making differences of cosine similarities a natural decision variable (Radford et al., 2021). We define a category score:

$$s(x, L) = \cos\left(e_x, e^{(L)}_{\text{light}}\right) - \cos\left(e_x, e^{(L)}_{\text{dark}}\right). \quad (3)$$

Positive values indicate that the image embedding is closer to the *light* label than the *dark* label in the joint space, yielding a graded analogue of a psychometric response function. Using a continuous score draws on psychophysics practice: thresholds and category transitions are inferred from response curves, which are more stable and comparable across conditions than single-trial decisions (Gescheider, 2013).

Boundary estimation. To estimate the model's *implicit* category boundary for each language, we model the relationship between score and physical lightness with a simple parametric psychometric approximation:

$$s(x, L) = a_L L_t + b_L, \quad (4)$$

and define the model's boundary as the point of indifference (score zero),

$$L^*_L = -\frac{b_L}{a_L}. \quad (5)$$

This approach mirrors how category boundaries are estimated in human perception: the boundary is taken to be the point where the stimulus provides equal support for each category (often where a logistic fit crosses $P = 0.5$) (Gescheider, 2013; Harnad, 2003). The language-conditioned boundary shift is then:

$$\Delta L^* = L^*_{\text{ru}} - L^*_{\text{en}}. \quad (6)$$

A non-zero ΔL^* provides a simple quantitative test of linguistic relativity: identical stimuli x gives rise to different category transitions when the linguistic label changes. This parallels language-conditioned boundary effects in human color categorization (Winawer et al., 2007) and echoes recent work on emergent categorical structure in neural networks (Akbarinia, 2025; de Vries et al., 2022).

Vision-only discrimination margin. To separate language-conditioned effects from intrinsic visual structure, we evaluate the same triads using only the vision encoder. We compute target–match similarity and target–distractor similarity in the *visual* embedding space and define a discrimination margin:

$$m(x) = \left|\cos(e_t, e_m) - \cos(e_t, e_d)\right|, \quad (7)$$

where e_t, e_m, and e_d denote target, match, and distractor visual embeddings. This margin is a machine analogue of discriminability: a larger $m(x)$ indicates that the distractor is less confusable with the match given the target. A categorical-perception-style prediction would be that cross-category pairs become more separable than within-category pairs at matched physical distance. The corresponding null hypothesis is:

$$\mathbb{E}[m \mid \text{cross}] \approx \mathbb{E}[m \mid \text{within}]. \quad (8)$$

This control is motivated by both human work distinguishing perceptual from lexical contributions (Winawer et al., 2007) and computational findings that category boundaries can either (i) emerge within vision-only representations or (ii) be strengthened by task/label structure (Akbarinia, 2025; de Vries et al., 2022).

Model

All experiments in this work were conducted using a single contrastive VLM: **Jina-CLIPv2** (Koukounas et al., 2024). Jina-CLIPv2 follows the CLIP-style dual-encoder model (Radford et al., 2021), jointly trained to align image and text embeddings in a shared semantic space. We selected Jina-CLIPv2 because it combines multilingual text encoders with a high-capacity visual backbone and has demonstrated strong performance on zero-shot image retrieval in multilingual settings. This makes it particularly well-suited for testing language-conditioned category structure across different lexical systems. Both image and text embeddings were ℓ_2-normalized prior to all similarity computations. This study focuses exclusively on *contrastive* vision–language representations. We do not evaluate generative vision–language models in the present work, in order to isolate language-conditioned structure in joint embedding spaces rather than prompt-dependent generation behavior.

For each language L, category scores were computed by comparing each image embedding to a fixed set of language-specific text prompts representing the *light* and *dark* categories. The prompts were chosen to reflect natural, commonly used color terms in each language.

For English, we used *light blue* for the light category and *dark blue* for the dark category. For Russian, we used голубой for light blue and синий for dark blue to account for grammatical gender. For French, we used *bleu clair* and *bleu foncé* for the light and dark categories, respectively.

All text prompts were encoded with the corresponding multilingual text encoder of Jina-CLIPv2. No other parameters or inputs were changed across languages.

Results

Language-conditioned category boundaries in a contrastive VLM

Figure 2 visualizes how a contrastive VLM partitions the same continuous color continuum under different linguistic framings. Each point corresponds to one stimulus triad, and each curve summarizes the model's graded preference for the *light* versus *dark* label under a given language. The visual stimulus is identical across language conditions; only the textual inputs differ. Therefore, any systematic shift in the curves cannot be attributed to changes in perceptual input, but must arise from the interaction between visual representations and language-specific semantic structure.

Language-dependent decision geometry. In Figure 2 (a), the English and Russian curves cross the indifference line

at markedly different lightness values. For a wide region of the continuum, the same visual stimulus is categorized as *light* under English but *dark* under Russian. This demonstrates that the model's internal decision geometry is not fixed in visual space, but is instead reparameterized by language. In other words, the linguistic frame shifts how the model interprets the same visual axis, producing different category partitions without altering the underlying visual encoding.

Why this constitutes a Whorfian effect. Although the fitted functions for all languages remain monotonic in L_t (Eq. 4), their intercepts differ by language, shifting the location of the zero crossing. This indicates that language shifts where category boundaries fall along a continuous perceptual axis, without changing the underlying geometry. The same x can yield different signs of $s(x, L)$ across languages, directly satisfying Eq. 1.

Specificity to linguistic category structure. Figure 2 (b) shows that when two languages (English and French) share the same color lexicon, the fitted functions yield overlapping L_L^* values under Eq. 5. Consequently, ΔL^* tends to be very small in Eq. 6, confirming that the boundary shift depends on genuine semantic differences rather than token variation.

Re-anchoring rather than distortion. While the slope a_L in Eq. 4 remains smooth and monotonic across languages (English, Russian, and French), the intercept b_L varies, shifting the zero crossing L_L^*. Thus, language adjusts the categorical reference point while leaving the visual continuum unchanged.

Implications for artificial categorical perception. The non-overlapping confidence intervals around the zero crossings indicate that the differences in L_L^* are statistically reliable, yielding a non-zero ΔL^* in Eq. 6. This establishes that the model encodes language-conditioned category structure rather than a single, language-invariant boundary.

Figure 3 shows the same effect from another angle. The English curve shifts from low to high values at much lower lightness levels than the Russian curve. As a result, for intermediate colors, the model consistently assigns different category labels based solely on the linguistic context.

Boundary displacement. The midpoint of each curve corresponds to the zero crossing of the score function (Eq. 5). The separation between these midpoints matches the displacement measured by ΔL^* (Eq. 6), confirming that the two analyses capture the same underlying boundary shift.

Categorical sharpening. The steeper transition of the Russian curve reflects a larger absolute slope a_L in Eq. 4, meaning that small changes in L_t lead to larger changes in $s(x, L)$ near the boundary. This corresponds to stronger categorical commitment around L_L^*.

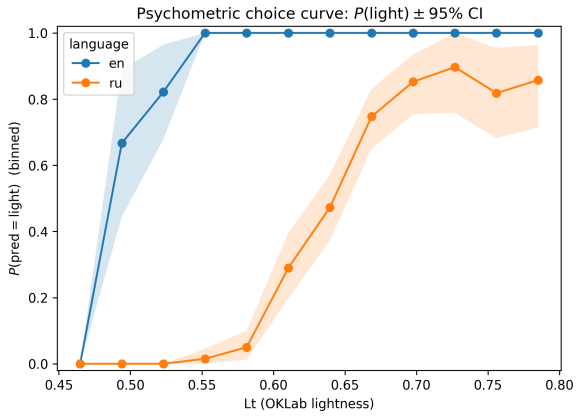

Figure 3: **Psychometric choice curves for a contrastive VLM.** Binned probability of predicting the *light* label as a function of target lightness for English and Russian prompts (95% CIs). The Russian curve is shifted rightward and is steeper, indicating a later and more categorical boundary than English for identical stimuli.

Language-dependent classification of identical stimuli. Within the intermediate region, the same x yields opposite signs of $s(x, L)$ across languages. By definition of the decision rule tied to Eq. 3, this means that category labels differ solely because L differs, while x is fixed.

Lexical specificity. From Figure 4, we can see that, when generic labels are used (light vs. dark, without providing *blue* label), the fitted function never satisfies $s(x, L) = 0$, and thus no L_L^* exists under Eq. 5. In contrast, Russian anchors yield a valid zero-crossing. This demonstrates that category structure arises only when the language supplies distinct semantic reference points.

Vision-only control: absence of categorical sharpening

Figure 5 evaluates the null hypothesis in Eq. 8. Because within- and cross-category margins overlap across L_t, we fail to reject the null, showing that categorical sharpening does not arise from the visual encoder alone.

Overall interpretation. Taken together, these results demonstrate that the contrastive VLM does not simply encode a continuous perceptual color axis and then read out labels in a language-agnostic manner. Instead, the same visual representation is *re-anchored* to different lexical reference points depending on the language context, yielding distinct category boundaries for identical stimuli. Crucially, this boundary shift is not accompanied by any categorical sharpening in the vision-only control, indicating that the visual encoder itself preserves a smooth, continuous representation of lightness. The categorical structure, therefore, arises only at the level of vision–language alignment.

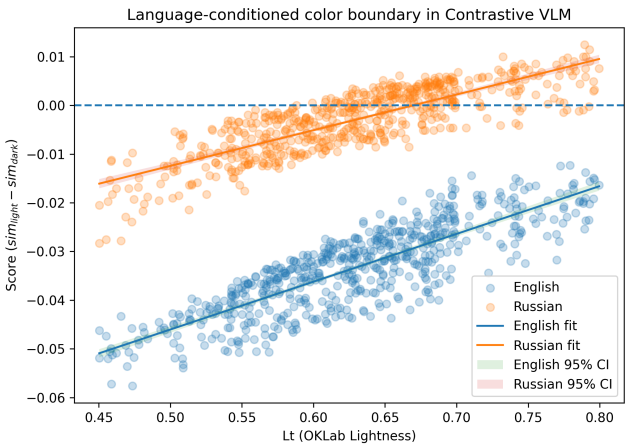

Figure 4: Language-conditioned color boundary with generic vs. categorical anchors. English uses generic labels (*light/dark*) and shows no zero crossing, while Russian uses lexical blue categories (*goluboy/siniy*) and exhibits a clear category boundary.

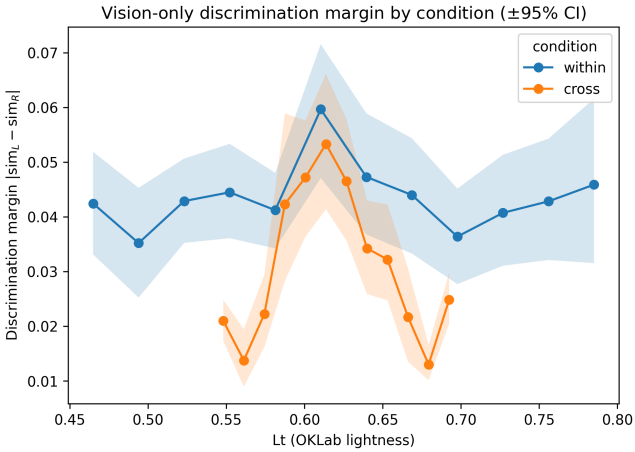

Figure 5: **Vision-only control: no categorical sharpening.** Discrimination margin as a function of target lightness for within- and cross-category triads. Cross-boundary pairs are not more discriminable than within-category pairs, indicating that categorical effects do not arise from vision-only embeddings.

The observed effect satisfies the criterion of linguistic relativity expressed in Eq. (1): for fixed visual input x, the model's categorical response function differs across linguistic contexts, $f(x, L) \neq f(x, L')$. At the same time, the null result in the vision-only margin analysis supports Eq. (8), showing that intrinsic visual geometry alone does not impose discrete boundaries. The model's behavior thus parallels the classic human dissociation in which categorical perception disappears under verbal interference (Winawer et al., 2007).

We therefore interpret the boundary shift not as a change in perception, but as a change in the *decision geometry* induced by language. In this sense, the model implements a computational analogue of moderate linguistic relativity: language does not determine what is seen, but it systematically reshapes how continuous perceptual information is partitioned into categories. This demonstrates that Whorfian-like effects can emerge as a property of multimodal alignment itself, rather than from symbolic rules or hand-coded semantics.

Limitations

Our stimuli vary primarily along a single perceptual dimension (lightness), whereas human color categories are shaped by interactions among hue, chroma, and contextual factors. In addition, this study examines only one model and one semantic domain, so it is unclear how broadly the effect generalizes to other architectures or categories. Moreover, all results are based on a single contrastive vision–language model, so we cannot determine whether the observed effects are specific to this particular model or representative of contrastive VLMs more broadly. We also focus exclusively on a contrastive vision–language model; it remains an open question whether similar language-conditioned boundary shifts would emerge in large generative VLMs (Achiam et al., 2023; Alayrac et al., 2022; Li et al., 2023), where text–image alignment is mediated through autoregressive decoding and instruction tuning rather than a shared embedding space. Finally, the model's responses are deterministic and do not reflect the perceptual noise, uncertainty, or attentional dynamics that characterize human judgments, so the findings should be interpreted as properties of representational alignment rather than as evidence of human-like perception.

Conclusion

We showed that a contrastive vision–language model exhibits language-conditioned categorical perception. Using controlled color triads, we found that the model's light–dark boundary in OKLab space shifts systematically across prompts, despite identical visual inputs. The resulting psychometric functions display sigmoidal structure and boundary displacement, paralleling classic findings in human categorical perception. Although the polarity of the effect does not match human data, the presence of a language-dependent boundary demonstrates a computational analogue of linguistic relativity: language reshapes the model's perceptual decision structure without altering the visual representation itself.

Acknowledgments

This paper was funded by the ERC Advanced project eTALK (funded by UKRI).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S.,

Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akbarinia, A. (2025). Exploring the categorical nature of colour perception: Insights from artificial networks. *Neural Networks*, *181*, 106758.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: A visual language model for few-shot learning. *Advances in neural information processing systems*, *35*, 23716–23736.

Arias, G., Baldrich, R., & Vanrell, M. (2025). Color in visual-language models: Clip deficiencies. *arXiv preprint arXiv:2502.04470*.

Berlin, B., & Kay, P. (1991). *Basic color terms: Their universality and evolution*. Univ of California Press.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., et al. (2020). Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and english speakers' conceptions of time. *Cognitive psychology*, *43*(1), 1–22.

de Vries, J. P., Akbarinia, A., Flachot, A., & Gegenfurtner, K. R. (2022). Emergent color categorization in a neural network trained for object recognition. *Elife*, *11*, e76472.

Fairchild, M. D. (2013). *Color appearance models*. John Wiley & Sons.

Forder, L., & Lupyan, G. (2019). Hearing words changes color perception: Facilitation of color discrimination by verbal and visual cues. *Journal of Experimental Psychology: General*, *148*(7), 1105.

Gescheider, G. A. (2013). *Psychophysics: The fundamentals*. Psychology Press.

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the national academy of sciences*, *103*(2), 489–494.

Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal neurons in artificial neural networks. *Distill*, *6*(3), e30.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178.

Harnad, S. (2003). Categorical perception.

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide*. Routledge.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. *International conference on machine learning*, 4904–4916.

Koukounas, A., Mastrapas, G., Eslami, S., Wang, B., Akram, M. K., Günther, M., Mohr, I., Sturua, S., Wang, N., & Xiao, H. (2024). Jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*.

Levinson, S. C. (2003). *Space in language and cognition: Explorations in cognitive diversity* (Vol. 5). Cambridge University Press.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International conference on machine learning*, 19730–19742.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, *36*, 34892–34916.

Lucy, J. A. (1992). *Language diversity and thought: A reformulation of the linguistic relativity hypothesis*. Cambridge University Press.

Lupyan, G., Rahman, R. A., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in cognitive sciences*, *24*(11), 930–944.

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. L. (2024). Large language models predict human sensory judgments across six modalities. *Scientific Reports*, *14*(1), 21445.

Ottosson, B. (2020, December). *A perceptual color space for image processing* [Accessed: 2026-01-23]. https://bottosson.github.io/posts/oklab/

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

Sapir, E. (1929). The status of linguistics as a science. *language*, 207–214.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, *104*(19), 7780–7785.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., & Zou, J. (2022). When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.